A SYSTEMATIC APPROACH TO ELUCIDATE THE CONNECTION BETWEEN GENOMIC ALTERATIONS AND METABOLIC PATHWAYS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Konnor C. La

December 2020

A SYSTEMATIC APPROACH TO ELUCIDATE THE CONNECTION BETWEEN GENOMIC ALTERATIONS AND METABOLIC PATHWAYS

Konnor C. La, Ph.D.

Cornell University 2020

Systems biology provides a structure to elucidate complex biological networks from multi-omic measurements. However, a limitation of systems biology is the ability to create testable hypotheses for further experimentation. Here we present two system-based methods, a coessentiality network and single-cell RNA sequencing, that are aimed to uncover metabolic functionality for uncharacterized genes and mitochondrial DNA mutations.

Coessentiality mapping has been useful to systematically cluster genes into biological pathways and identify gene functions (Pan et al., 2018; Wainberg et al., 2019; Wang et al., 2017). Here, using the debiased sparse partial correlation (DSPC) method (Basu et al., 2017), we construct a functional coessentiality map for cellular metabolic processes across human cancer cell lines. This analysis reveals 35 modules associated with known metabolic pathways and further assigns metabolic functions to unknown genes. In particular, we identify *C12orf49* as an essential regulator of cholesterol and fatty acid metabolism in mammalian cells. Mechanistically, *C12orf49* localizes to the Golgi, binds membrane-bound transcription factor peptidase, site 1 (MBTPS1, site 1 protease) and is necessary for the cleavage of its substrates, including sterol regulatory element binding protein (SREBP) transcription factors.

The electron transport chain (ETC) activity in mammalian cells is necessary for survival and proliferation. The ETC is composed of ~100 subunits mostly encoded in the nuclear genome, but 13 essential subunits are in the mitochondrial genome (mtDNA). Accumulation of mutations in the mtDNA can lead to severe genetic defects and cell death. Interestingly, we and other groups have found the occurrence of loss-of-function (LOF) mtDNA mutations across a variety of cancer types at high heteroplasmy. Heteroplasmy is defined as the proportion of mtDNA copies with a specific mutation over the total number of mtDNA copies. Furthermore, there is an enrichment for these LOF mutations suggesting that they are positively selected. Despite their prevalence, it is unclear whether these mutations have functional roles in cancer progression or are simply passenger mutations as the study of mtDNA mutations is stymied by the lack of methods to genetically modify the mtDNA. Here, we have identified a novel method that combines single-cell RNA sequencing (scRNAseq) and fluorescence-activated cell sorting (FACS) in order to create a low and high heteroplasmic mixed population of cells. The high heteroplasmic populations have severe ETC dysfunction and are morphologically, transcriptomically, and metabolomically distinct from the low heteroplasmic cells. These differences result in the high heteroplasmy cells having elevated metastatic potential compared to the low heteroplasmy cells suggesting a role for LOF mtDNA mutations in cancer progression. Altogether, our findings reveal that a combination between single-cell RNAseq and FACS can produce distinct populations that correlate with LOF heteroplasmic mutations.

BIOGRAPHICAL SKETCH

The Human Genome Project was my first exposure to science. I felt the excitement as the world realized the potential for sequencing to improve our understanding of many biological disciplines. To this end, I started my research career studying genomics, but became interested in how biological pathways respond in the presence of genomic alterations. This interest progression began at UC Berkeley/Life Technologies, lead me to the NIH, and inspired my current research proposal in graduate school.

In my third year at UC Berkeley, I was an undergraduate researcher in Dr. Nipam Patel's lab studying evolutionary development. I became interested in how the Hox gene complex is conserved genomically, particularly in Parhyale hawaiensis and Drosophila melanogaster. I took this interest to the Ion Torrent platform at Life Technologies (now ThermoFisher Scientific) to work as a research and development intern. In addition to conducting my day-to-day responsibilities such as optimizing for sequence read- length, reducing homopolymer error, and improving pair-end sequencing, I was able to propose an additional project to sequence the full-genome of Parhyale. I was able to convince my mentor of the value of this project with my hard work, passion, and knowledge of the evolutionary importance of these Hox genes.

Returning to the Patel lab, I built on my work at Life Technologies and continued to annotate and assemble the Parhyale genome. Despite having little programming experience, I self-learned Perl to conduct a comparative genome analysis of the Hox

complex between the Drosophilia genome and the Parhyale bacterial artificial chromosomes. I found that there is genome size reduction from Parhyale to Drosophila, particularly in intron length in the Ultrabithorax to Deformed gene complex and contained more genes. From this, I concluded that during the evolution of the Hox complex there was a reduction in genome size without compromising complexity. The results of this work contributed to the publication of a manuscript (Serano et al.). My undergraduate career as a researcher and student is something that I am very proud of because of the hardships that I had to overcome. I had to balance my school work, while supporting myself through college and dealing with the death of my grandfather. Despite this, I was able to be the first in my family to graduate from college and those experience gave me the confidence to pursue a career in research.

The genomic region and gene relationship that I observed in the Hox complex led me to explore other factors that affect transcription and ultimately organismal phenotypes. This resulted in a computational study of non-coding DNA regions at the NIH's National Library of Medicine in Dr. Ivan Ovcharenko's lab as a post-baccalaureate IRTA fellow. We hypothesized that the spatial relationship between cooperative transcription factors (TFs) would affect gene expression. I used and developed statistical models to predict transcription factor binding sites on enhancers and determined significant cooperative TFs at a transcript level from publicly available data. We showed that TF pairs follow a spatial relationship that complements the helicity of double-stranded DNA, demonstrating that TF regulation cooperativity has a 3D spatial

relationship allowing the TFs to be spatially close despite being separated by large linear distances on the DNA strand.

Upon acceptance to the Tri-Institutional (Cornell University, Weill Cornell Medicine, and Memorial Sloan Kettering) Computational Biology and Medicine program, I intended to develop my research experience by studying how genomics and RNA expression cause a phenotypic change. To do so, I joined the laboratory of Dr. Kıvanç Birsoy and experimentally discovered that cell lines with increased expression of an aspartate- glutamate transporter (SLC1A3) were more tolerant to Electron Transport Chain (ETC) inhibition. This discovery led to a second author manuscript that is currently in review at Nature Cell Biology. Similarly, I learned that mutations in mitochondrial DNA, which largely encode for ETC proteins, can also lead to ETC inhibition. To explore this, I was also able to join the Dr. Andrew Clark's laboratory to use computational methods to determine the occurrence and selection for mutations in mitochondrial DNA.

This work is for my parents

虎父无犬子

ACKNOWLEDGEMENTS

During the 2020 closing of the Tri-Institutions due for the COVID-19 pandemic, I had a Zoom hangout session with two of my friends, Charlie Shi and Eric Bourgain-Chang. We have known each other since 2005 and were recapping our relationship. One of the questions that came up was “what are you most surprised by now that we have grown up?” I gave my answers for them and when it was their turn they both replied with “even throughout college, we never thought that you would do a PhD.” While, I was deciding whether I was just insulted, I realized that there is some truth to their statement. Things clearly have turned and I would like to acknowledge all of the people who have come through my life that have prepared me to embark on my PhD journey.

Thank you to my entire family starting from my parents to my sister Amenda, brother Casey and my cousins Nina, Steven, Calvin, and Lauren. All of you are my best friends and I could not do this without your support. Also thank you to my Los Angeles cousins Linda, Laura, Lisa, Binh, and Jane for reaching out to me throughout my graduate career. Lastly, I would like to thank my uncles and aunts from both sides of my family. My family is truly the most special thing in my life and they give me the acceptance necessary to try crazy new things!

I would like to thank my friends. Thank you to Jeff for making me feel like it was okay to ask for help. It is a lesson that I continue to practice. Thank you to Charlie and Eric who taught me how to be academically disciplined and how to lean on my

strengths. Thank you to Bobby for teaching me that limitations of the world are largely untrue and giving me confidence to pursue my goals.

Special gratitude to all of the collaborators that helped with both of these projects namely Sumanta Basu, Kara Karpman, and Junyue Cao. Your expertise in network analysis and single-cell sequencing were vital in the completion of both projects. I have learned so much from all of you and I hope that we can collaborate again in the future.

I would like to thank all of the Birsoy, Clark, and Schultz's lab members past and present for your help throughout this process. Thank you to Francisco for teaching me the basics of cancer genomics. Thank you to Manisha and Andrew for your company and statistical advice. Thank you to Javier for his consistent presence in troubleshooting and guiding my project as well as providing advice during the bumpy times. Thank you to Lou who really shaped up the lab to be a place where we could conduct science and for being my French tutor. Thank you to Tim for helping me with many of the functional mtDNA experiments. He brought a very different and complimentary background that proved to be invaluable throughout the mtDNA project. Thank you to Robbie, Roy, and Ross for being available to troubleshoot and discuss any obstacles in my projects.

Finally, I would like to thank my advisors Kıvanç Birsoy, Andrew Clark, and Nikolaus Schultz. Without them I truly would not be in this situation. Early in my graduate career, I asked each one of them what they feel like graduate students should learn. Niki advised me to be helpful and collaborative because you never know what skills and fields you may get exposed to. Because of his advice, I have had the fortunate to be a co-author on over 40 publications and learned the skills to do my

thesis work. Andy expressed that the most important thing that a graduate student needs to learn is how to figure things out on their own. Going into experimental biology from a computational background, I have certainly been in the position of having to adapt and learning things on the fly. If it were not for Andy's words, I would not have been able to learn how to think like a computationalist and experimentalist. This duality certainly came in handy when we were creating the single-cell and fluorescence based sorting method. Lastly, I want to express a great deal of gratitude to Kıvanç. He took a chance on me when I had no experimental experience and trained me to be a scientist and academic. He gave me a lot of attention to discuss how to think about science and the logic of experiments. Additionally, he taught me how to collaborate, present, and publish. Collectively, he trained me to be a well-rounded academic and I am truly indebted to his generosity and his words are some that I will always reflect on.

Table of Contents

BIOGRAPHICAL SKETCH iii

ACKNOWLEDGEMENTS vii

LIST OF ABBREVIATIONS xii

Chapter 1: Metabolic coessentiality mapping identifies C12orf49 as a regulator of SREBP processing and cholesterol metabolism 1

Introduction *1*

Results *4*

Establishing the "coessentiality" pipeline 4

Validating coessentiality networks 6

TMEM41a is associated with saturated fatty acid 9

C12orf49 is necessary for cholesterol synthesis and SREBP-induced gene expression in human cell 11

Discussion *16*

Materials and Methods *16*

Chapter 2: Developing a single-cell method to study the heteroplasmic effects of a ND4 loss-of-function mutation 25

Introduction *25*

Results *29*

Choosing a model experimental system 29

FTC133 and KHM_5M have opposing heteroplasmy distributions 32

Single-cell sequencing revealed a correlation between the ND4 11866 C insertion mutations and the whole-cell transcriptome in FTC133 35

Using transcriptomic differential gene expression as a strategy to isolate low and high heteroplasmic cells 37

High heteroplasmic cells are maintained in the parental cell line 51

High heteroplasmic cells display a more invasive phenotype 54

The alpha-ketoglutarate and succinate ratio may be the mechanistic link between the low and high heteroplasmy cells 59

Discussion *65*

Materials and Methods *67*

Chapter 3: Future directions and Perspectives 77

Improving the co-essentiality networks *77*

Generalizing the FTC133 heteroplasmic population isolation method *79*

Bibliography 82

LIST OF FIGURES

Figure 1: Coessentiality analysis pipeline……6

Figure 2: Genetic coessentiality analysis assigns metabolic function to uncharacterized genes……8

Figure 3: TMEM41a is associated with lipid saturation……11

Figure 4: C12orf49 is necessary for cholesterol synthesis and SREBP-induced gene expression in human cells……14

Figure 5: Thyroid cancer is enriched for mtDNA truncation mutations……..31

Figure 6: Location of mtDNA mutations for five thyroid cell lines……...32

Figure 7: scRNAseq shows intercellular heteroplasmic heterogeneity……..34

Figure 8: ND4 11866 C insertion heteroplasmic mutations correlates with whole-cell transcriptomic alterations……..37

Figure 9: Markers can be used to isolate different heteroplasmy cell populations……45

Figure 10: High ND4 11866 C insertion heteroplasmic mutations results in severe ETC dysfunction……..50

Figure 11: High heteroplasmy cells are significantly maintained in the parental cells... 53

Figure 12: High heteroplasmy cells have higher invasive potential than low heteroplasmy cells……..58

Figure 13: alpha-ketoglutarate to succinate ratio may be the mechanistic link between low and high heteroplasmy cells……..62

LIST OF ABBREVIATIONS

Acetyl-CoA carboxylase: ACC

Acetyl-coenzymeA: acetyl-CoA

Adenosine triphosphate: ATP

Alpha-ketoglutarate: αKG

American Type Culture Collection: ATCC

ATP citrate lyase: ACLY

Cancer Cell Line Encyclopedia: CCLE

Carbon dioxide: CO2

Clustered regularly interspaced short palindromic repeat: CRISPR-Cas9

Debiased sparse partial correlation: DSPC

Dimethyl-aKG: DM-aKG

Dimethyl-succinate: DM-Succ

Electron transport chain: ETC

Endoplasmic reticulum: ER

Epithelial mesenchymal transition: EMT

False discovery rate: FDR

Fatty acid synthase: FASN

Flavin adenine dinucleotide: FAD

Fluorescence-activated cell sorting: FACS

Gene Ontology: GO

Glutamic-Oxaloacetic Transaminase 1: GOT1

Guanosine triphosphate: GTP

Liquid chromatography–mass spectrometry: LC–MS

Loss-of-function: LOF

Low density lipoprotein receptor: LDLR

Mitochondrial genome: mtDNA

Monounsaturated fatty acids: MUFAs

Multiplicity of infection: MOI

Oxidative phosphorylation: OXPHOS

Oxygen consumption rate: OCR

Reactive oxygen species: ROS

Reduced nicotinamide adenine dinucleotide: NADH

RNA-sequencing: RNAseq

Short-tandem Repeat profiling: STR

Single-cell RNA-sequencing: scRNAseq

Solute carrier family 1 member 3: SLC1A3

Stearoyl-CoA desaturase: SCD1

Sterol regulatory-element bind protein: SREBPs

Succinate dehydrogenase: SDH

The Cancer Genome Atlas: TCGA

Transcription Factors: TFs

Tricarboxylic acid: TCA

Uniform Manifold Approximation and Projection: UMAP

Chapter 1: Metabolic coessentiality mapping identifies C12orf49 as a regulator of SREBP processing and cholesterol metabolism

Introduction

The Human Genome project has been monumental in thoroughly annotating the complete human genome sequence. The resulting genome has led to many efforts to annotate the genome for both coding and non-coding regions (Lander et al., 2001; Venter et al., 2001). The publicly available resource has further contributed to the development of many genomic technologies such as DNA mutation calling, proteomics, RNA-sequencing (RNA-seq), chromatin immunoprecipitation sequencing, and, as it relates to this chapter, clustered regularly interspaced short palindromic repeat (CRISPR-Cas9) based genetic screens (Cibulskis et al., 2013; Hood and Rowen, 2013; Shalem et al., 2014; Wang et al., 2014). More importantly, the data from these studies were made user-friendly through interfaces like cBioPortal, DepMap, UCSC Genome browser and many more (Cerami et al., 2012; Gao et al., 2013; Kent et al., 2002; Shimada et al., 2019). The impact of these data repositories and user-interfaces have significantly contributed to a bioinformatic revolution (Varmus, 2002). One instance is many computational and mathematical methods were developed in order to appropriately and effectively mine these large data sets.

The ultimate goal of mining these data sets is to discover important biological findings. These include and are not limited to, identifying conserved genes, determining altered gene regulation, and assigning biological function. We were most interested in

assigning biological function to poorly characterized and unknown genes. In order to do so, we turned towards combining both computational and experimental methods.

The use of computational methods provided a more effect method to reduce the total number of biological pathways to only those that were likely to be associated with uncharacterized genes. The number of potential functions that any poorly characterized gene may be associated with are innumerable and would be inefficiently identified using experimental techniques alone. Traditionally, computational biologists have created coexpression or coconversed networks to “group” genes (Pellegrini et al., 1999; Serin et al., 2016). The two assumptions being genes in related pathways will follow the “guilt-by-association principle,” which states that genes sharing the same function or that are involved in the same regulatory pathway will tend to present similar expression profiles and thus form groups, termed modules or clusters (Wolfe et al., 2005). Past studies have successfully leveraged these principles to make many important biological discoveries, for example identifying the uncharacterized gene *CCDC109A* as an essential component of the mitochondrial calcium uniporter (Baughman et al., 2011). Continuing with the progress made by these past studies, we aimed to apply these principles onto the single-gene loss-of-function cell line dataset produced by the Broad Institute (Achilles project) (Meyers et al., 2017).

Many studies, both in our lab and others, have shown that genetic dependencies vary significantly across cancer cell lines (Bertomeu et al., 2018; Blomen et al., 2015; Hart et al., 2015; McDonald et al., 2017; Wang et al., 2017). The Broad Institute took advantage of these differences by systematically conducting whole-genome CRISPR-

Cas9 genetic screens to assess genetic vulnerabilities across 789 cell lines of different tissues of origin and oncogenic alterations. We hypothesized that we can associate gene-to-gene interactions based on their effects across these cells lines to create new networks that would lead to novel gene-to-pathway associations. Here, the coessentiality analysis focused on the identification of poorly characterized genes to known metabolic pathways because metabolic pathways are generally well-annotated and perturbation of enzymes or regulatory units involved in the same metabolic pathway should display similar effects on cellular fitness across cell lines, suggesting that the correlation of essentiality profiles may provide a unique opportunity to identify unknown components associated with a particular metabolic function. As a result, we identified *TMEM41a* as being associated with saturated fatty-acid metabolism and *C12orf49* as a regulator of SREBP processing and cholesterol metabolism (Bayraktar et al., 2020).

Sterol regulatory-element bind proteins (SREBPs) are transcription factors that regulate the expression of many lipid-related genes such as *SCAP*, *MBTPS1*, and *SCD*. Their regulation controls for the organismal and cellular homeostasis and metabolism of lipids and are a key pathway for various physiological and pathophysiological processes (Shimano et al., 2017). The SREBP pathway is cross-regulated by other pathways in states of energy abundance (AKT-mTOR-SREBP) and lipid-mediated cellular stress (obesity). The interactions between SREBP with many biologically pathways suggest that further mechanistic understanding of SREBP regulation may have an outsized impact on basic cellular and organismal metabolism.

Results

Establishing the "coessentiality" pipeline

To generative putative networks of genes that have similar vulnerability effects across cells, termed coessential, for metabolic genes, we analyzed the genetic perturbation datasets from the Achilles project, also known as, the DepMap project collected from 558 cell lines (at the time of this work) (Figure 1a). Existing computational methods for constructing coessentiality networks primarily rely on Pearson correlation, which is not suitable for distinguishing between direct and indirect gene associations and leads to false positive edges in the network. This point is made explicitly by performing a comparative simulation between partial and Pearson correlation (Bayraktar et al., 2020). We used *E. coli* genetic networks as our simulation because their genetic networks are well-studied (Kim et al., 2015). Simply using Pearson correlation, results in a "hair-ball" network, where each node (gene) is interconnected with many other nodes. This multiplicity of interactions obfuscates clear signals that can be isolated for further experimentation. However, Gaussian graphical models calculate partial correlation and offer a unique advantage over commonly used Pearson correlation networks by automatically removing indirect associations among genes from the network, hence reducing false positives and producing a small number of high-confidence sets of putative interactions for experimental validation (Bayraktar et al., 2020; Krumsiek et al., 2011). The receiver operating characteristic curve summarizes the method comparisons and shows that over a 500-iteration sample, the partial correlation outperforms basic Pearson correlation (Bayraktar et al., 2020). We therefore applied DSPC, a Gaussian

graphical model technique, to measure associations between the essentiality scores of genes from human cancer cell lines. In prior work (Basu et al., 2017), we have successfully used DSPC to build networks among metabolites and have identified new biological compounds. Of note, this method, while useful for generating highly confident lists, does not account for dependence among cell lines, a key strength of previously published work (Kim et al., 2019; Wainberg et al., 2019). Lastly, we combined this statistical method with domain specific knowledge of metabolic networks. We removed networks with a large number of components (that is, the electron transport chain) because the majority of metabolic pathways do not have many key genes. Interactions between genes that were found to be within 1 cytogenic band were also removed because the knock-out of one neighboring gene could affect the transcription of the other and would thereby correspond to highly correlating genes that are not functionally related. We also focused on genes with a high Pearson correlation ($|r|>0.35$) and at least one of the 2,998 metabolism-related genes in the dataset (Figure 1a).

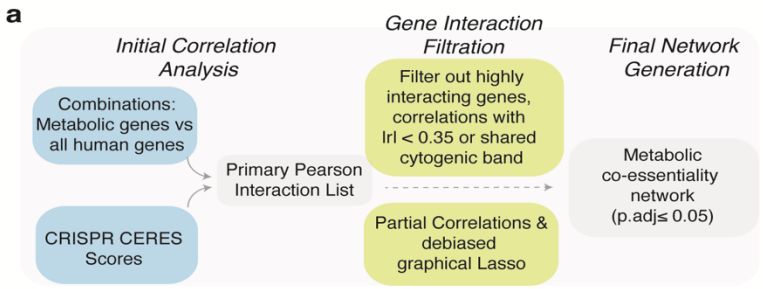

Figure 1: Coessentiality analysis pipeline a, Scheme of the computational steps to generate the metabolic coessentiality network.

Validating coessentiality networks

In order to make any claims on novel edges between poorly characterized genes and well-established pathways, we first determined whether our method predicts known metabolic pathways. Our analysis of positively correlated genes revealed a set of 202 genes organized into 35 metabolic networks, 33 of which we can assign a metabolic function using literature searches and the STRING database (Figure 2a) (Szklarczyk et al., 2019). Among these gene networks are glycolysis (*PGAM1*, *GPI*, *ENO1*, *HK2*, *PGP*), squalene synthesis (*FDPS*, *FDFT1*, *SQLE*), sialic acid metabolism (*SLC35A1*, *CMAS*, *GNE*, *NANS*), and pyruvate utilization (*MPC2*, *PDHB*, *DLAT*, *CS*, *MDH2*, *MPC1*) but also networks that are not part of a known metabolic pathway, suggesting the presence of unidentified metabolic pathways (Figure 2b). In addition to these pathways, we have also provided all remaining networks.

Our analysis also identified associations between genes of unknown function and those that encode components of well-characterized metabolic pathways. Interestingly, the functions of three of these genes have recently been discovered (Figure1b). UBIAD1, a prenyltransferase, has been shown to bind to HMGCR (the rate-limiting enzyme in cholesterol biosynthesis) to promote its degradation at the endoplasmic

reticulum (ER) in the presence of sterols (Schumacher et al., 2015). CHP1, which is associated with glycerolipid synthesis pathway in our analysis, binds to and is necessary for the function of the protein product of *AGPAT6*, the rate-limiting enzyme for glycerolipid synthesis (Zhu et al., 2019). In addition, a recent study identified *TMEM189*, a gene associated with plasmalogen synthesis, as the elusive plasmanylethanolamine desaturase (Gallego-García et al., 2019). Interestingly, squalene and mevalonate synthesis clustered into different networks, consistent with additional functions of the branches of cholesterol metabolism. Indeed, whereas loss of HMG-CoA synthase would decrease all intermediates as well as cholesterol, loss of squalene synthase or downstream enzymes would decrease cholesterol but increase upstream intermediates, hence leading to different cellular outcomes (Garcia-Bermudez et al., 2019). Finally, several genes of unknown function, such as *C12orf49* and *TMEM41A*, have correlated essentialities with those of genes that encode components of SREBP-regulated lipid metabolism, which raises the possibility that they may be involved in the regulation of SREBPs or their downstream targets (Figure 1d,e). Due to their strong correlation and unknown function, we focused our attention on these two genes.

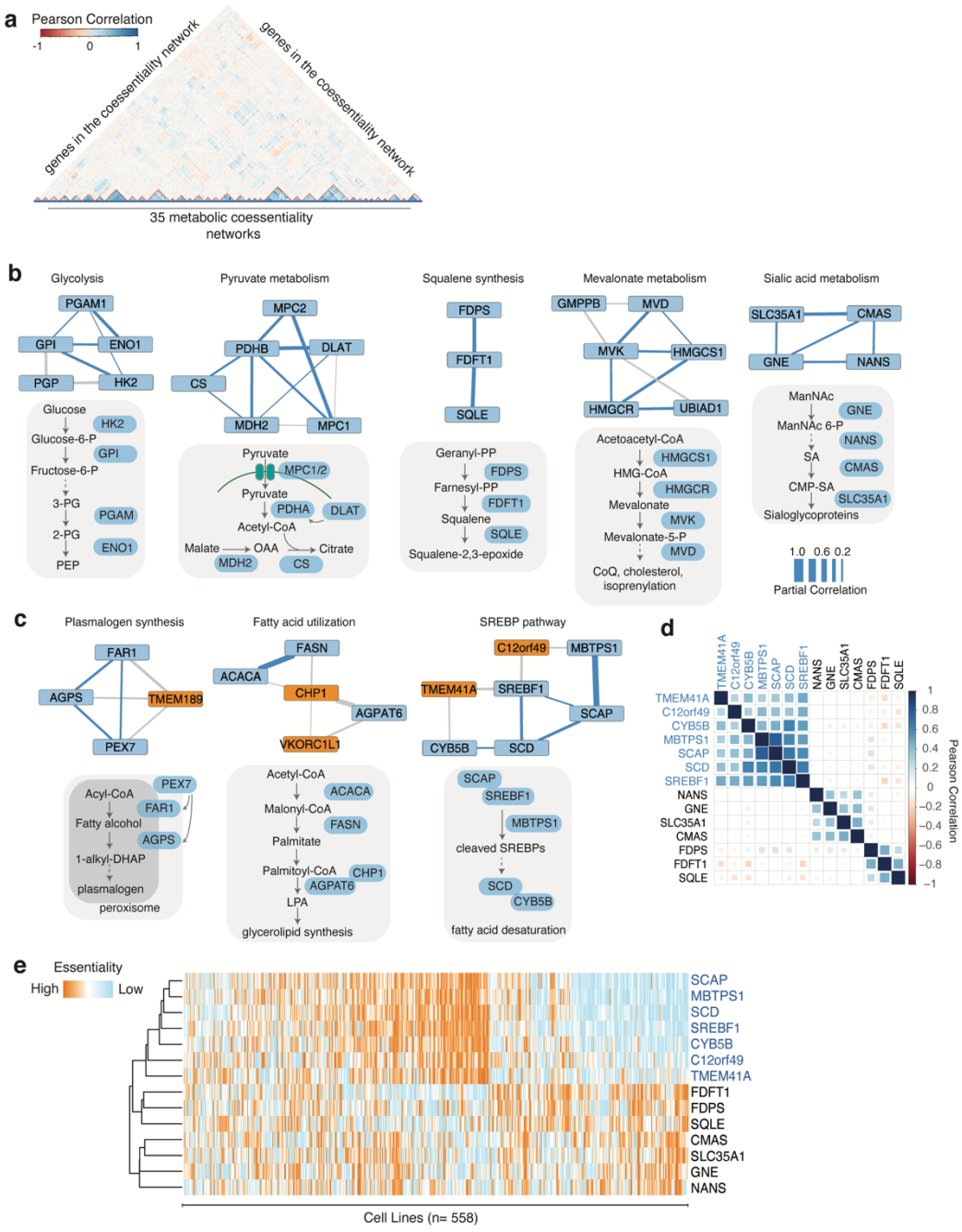
a
Pearson Correlation
-1
0
1
genes in the coessentiality network
genes in the coessentiality network
35 metabolic coessentiality networks
b
Glycolysis
PGAM1
GPI
ENO1
PGP
HK2
Glucose
HK2
Glucose-6-P
GPI
Fructose-6-P
3-PG
PGAM
2-PG
ENO1
PEP
Pyruvate metabolism
MPC2
PDHB
DLAT
CS
MDH2
MPC1
Pyruvate
MPC1/2
Pyruvate
PDHA
DLAT
Acetyl-CoA
Malate
OAA
Citrate
MDH2
CS
Squalene synthesis
FDPS
FDFT1
SQLE
Geranyl-PP
FDPS
Farnesyl-PP
FDFT1
Squalene
SQLE
Squalene-2,3-epoxide
Mevalonate metabolism
GMPPB
MVD
MVK
HMGCS1
HMGCR
UBIAD1
Acetoacetyl-CoA
HMGCS1
HMG-CoA
HMGCR
Mevalonate
MVK
Mevalonate-5-P
MVD
CoQ, cholesterol, isoprenylation
Sialic acid metabolism
SLC35A1
CMAS
GNE
NANS
ManNAc
GNE
ManNAc 6-P
NANS
SA
CMAS
CMP-SA
SLC35A1
Sialoglycoproteins
1.0 0.6 0.2
Partial Correlation
c
Plasmalogen synthesis
FAR1
AGPS
TMEM189
PEX7
Acyl-CoA
PEX7
FAR1
Fatty alcohol
AGPS
1-alkyl-DHAP
plasmalogen
peroxisome
Fatty acid utilization
FASN
ACACA
CHP1
AGPAT6
VKORC1L1
Acetyl-CoA
ACACA
Malonyl-CoA
FASN
Palmitate
Palmitoyl-CoA
CHP1
AGPAT6
LPA
glycerolipid synthesis
SREBP pathway
C12orf49
MBTPS1
TMEM41A
SREBF1
SCAP
CYB5B
SCD
SCAP
SREBF1
MBTPS1
cleaved SREBPs
SCD
CYB5B
fatty acid desaturation
d
TMEM41A
C12orf49
CYB5B
MBTPS1
SCAP
SCD
SREBF1
NANS
GNE
SLC35A1
CMAS
FDPS
FDFT1
SQLE
Pearson Correlation
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
e
Essentiality
High
Low
SCAP
MBTPS1
SCD
SREBF1
CYB5B
C12orf49
TMEM41A
FDFT1
FDPS
SQLE
CMAS
SLC35A1
GNE
NANS
Cell Lines (n= 558)

Figure 2: Genetic coessentiality analysis assigns metabolic function to uncharacterized genes
a, Heatmap depicting the partial correlation values of the essentialities of genes in the metabolic coessentiality pathways. b, Correlated essentialities of the genes that encode members of the glycolysis, pyruvate metabolism, squalene synthesis, mevalonate and sialic acid metabolism networks. The thickness of the lines indicates the level of partial correlation. d, Genetic coessentiality analysis assigns metabolic functions to uncharacterized genes. Orange and blue boxes show genes with unknown and known functions, respectively. The thickness of the line is indicative of partial correlation. e, Pearson correlation values of the essentiality scores of genes in the indicated metabolic networks. f, Unbiased clustering of fitness variation of indicated genes across 558 human cancer cell lines.

TMEM41a is associated with saturated fatty acid

Previous studies have thoroughly shown the mechanistic and physiological function of SREBPs in fatty acid and cholesterol metabolism (Brown and Goldstein, 1997). Given the strong coessentialities of *C12orf49* and *TMEM41A* with the SREBP pathway, we reasoned that these uncharacterized genes may be required for the activation of cholesterol and fatty acid metabolism. In order to determine whether there is an association between *TMEM41a* with either fatty acid or lipid metabolism, we generated a small CRISPR library consisting of 103 sgRNAs targeting genes involved in SREBP maturation and lipid metabolism (3–8 single guide RNAs (sgRNAs) per gene) (Figure 3a). Using this focused library, we performed negative selection screens for genes whose loss potentiates anti-proliferative effects of Stearoyl-CoA desaturase (SCD1) inhibitor, arachidonate (unsaturated acid), palmitate (saturated acid) and media with and without lipoprotein depletion (Figure 3,4).

After conducting our CRISPR-Cas9 genetic screen, we found that cells that are *TMEM41a*-NULL are sensitive to the inhibition of SCD1 and palmitate (Figure 3e). SCD1 is an ER enzyme that catalyzes the rate-limiting step in the formation of monounsaturated fatty acids (MUFAs), specifically oleate and palmitoleate from stearoyl-CoA and palmitoyl-CoA. Thus, the inhibition of SCD1 would result in the accumulation of saturated fatty acids. Building upon this finding, we also found that *TMEM41a* scores under palmitate treatment, but not arachidonate treatment further supporting the notion that TMEM41a is related to saturated fatty acid rather than unsaturated fatty acid treatment (Figure 3 c,d).

We decided to focus our attention on the palmitate sensitivity because of its physiological relevance and also to avoid the off-target effects of continuous use of a chemical inhibitor. Despite the genetic score showing striking sensitivity to palmitate addition, we needed to validate the screen and create an isogenic cell line model for further mechanistic study. Based on the genetic screen, we chose guides that scored the best and used them to knock-out the parental Jurkat cells (Figure 3e,f). After we knocked-out *TMEM41a* in Jurkat cells we chose two single cell clones that were sensitive to palmitate addition and became resistant after reintroducing *TMEM41a* (Figure 3g). The add-back of *TMEM41a* resulted in resistance to palmitate which provides a causal link between TMEM41a and palmitate sensitivity. Based on these data, we hypothesized that TMEM41a functions to alleviate saturated:unsaturated fatty acid imbalance. The combination of the palmitate and SCDi results also suggest that

TMEM41a acts downstream of palmitic acid and somehow the cells are unable to deal with this buildup.

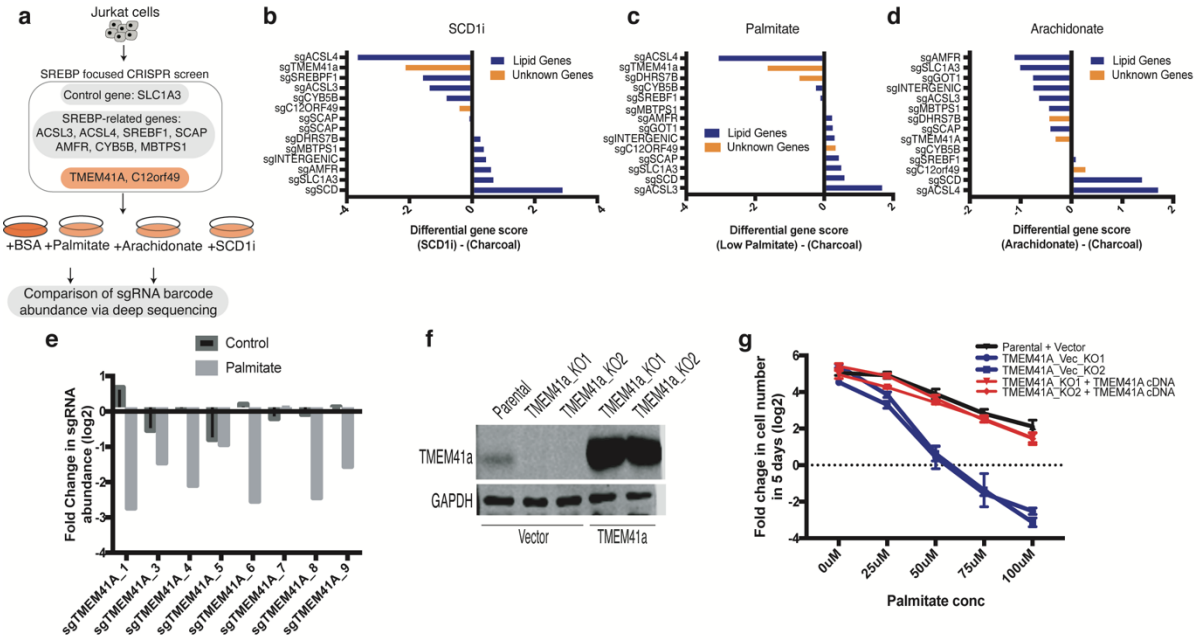

Figure 3: TMEM41a is associated with lipid saturation
a, Schematic for the focused CRISPR-Cas9-based genetic screen. b,c,d, The scores of each gene within the focused library as differentially required upon each treatment (top). Genes linked to lipid metabolism are colored in blue and unknown genes are in yellow. e, Changes in abundance in the primary screen of individual *TMEM41a* sgRNAs in the presence (gray) or absence (black) or palmitate. f, Immunoblot analysis of wild-type, *TMEM41a*-null, and rescued null cells. G, Expression of an sgRNA-resistant *TMEM41a* cDNA rescues palmitate sensitivity of the *TMEM41a*-null Jurkat cells. Fold change in cell number $\log_2$ of wild-type (black), *TMEM41a*-null (blue), and rescued *TMEM41a*-null (red) cells ager 5-day treatment with indicated palmitate concentrations (mean ± s.d., n = 3 biologically independent samples).

C12orf49 is necessary for cholesterol synthesis and SREBP-induced gene expression in human cell

SREBPs are transcription factors that regulate the transcription of genes that encode many enzymes in the cholesterol and fatty acid synthesis (Horton et al., 2002).

SREBPs are normally bound to ER membranes and are activated through a proteolytic cascade regulated by sterols (Brown and Goldstein, 1997; Wang et al., 1994). Cleaved SREBPs localize to the nucleus and induce expression of cholesterol synthesis genes, which enables cells to survive under sterol depletion (Sakai et al., 1996; Sakakura et al., 2001). Among the scoring genes were *MBTPS1* and *SCAP*, both of which are involved in SREBP processing (Hua et al., 1996; Matsuda et al., 2001; Yang et al., 2001), but also *C12orf49*, a gene of unknown function that has not been previously linked to cholesterol metabolism (Figure 4b). Consistently with the screening results, depletion of *C12orf49* strongly decreased proliferation of HEK293T, Jurkat and other cancer cell lines (U87 and MDA-MB-435) under cholesterol depletion, which indicates a generalized role for *C12orf49* in cholesterol homeostasis (Figure 4c,d). Importantly, expression of an sgRNA-resistant human *C12orf49* cDNA in the null cells, or free cholesterol addition, completely restores proliferation under lipoprotein depletion. None of the SREBPs scored, probably owing to highly complementary and redundant functions. Altogether, these results identify *C12orf49* and *TMEM41A* as major components of cholesterol and fatty acid metabolism.

We next sought to understand why cells require *C12orf49* to proliferate under cholesterol depletion. To first determine whether *C12orf49* is necessary for de novo cholesterol synthesis, we performed metabolite tracing experiments in Jurkat cells using 13C-acetate (Figure 4e). Whereas acetate contributes to cellular cholesterol under lipoprotein depletion, we observed significantly lower labelling in *C12orf49*-null cells, which indicates a problem in the synthesis (Figure 4e). Consistently with the

requirement of sterols for viral infection (Kleinfelter et al., 2015; Osuna-Ramos et al., 2018; Pombo and Sanyal, 2018), *C12orf49* loss also decreased Bunyamwera virus infectivity in mammalian cell lines and decreased total viral titer. As the cholesterol synthesis pathway comprises over thirty successive steps that are transcriptionally regulated (Edwards et al., 2000; Ericsson et al., 1996; Guan et al., 1997; Sakakura et al., 2001; Vallett et al., 1996), we considered that a dysfunction in gene expression might lead to defective synthesis and reliance on extracellular cholesterol. Indeed, *C12orf49*-null cells failed to induce expression of cholesterol metabolism genes under sterol depletion (Figure 4f,g). Furthermore, in line with the role of SREBPs in the transcription of cholesterol synthesis genes, loss of *C12orf49* reduced mature (cleaved) SREBP protein levels and blocked nuclear translocation of SREBPs (Figure 4h,i). In a similar manner, expression of other genes known to be induced by SREBPs, such as fatty acid synthase (*FASN*), low density lipoprotein receptor (*LDLR*), acetyl-CoA carboxylase (*ACC*) and ATP citrate lyase (*ACLY*) did not change in *C12orf49*-null cells (Figure 4f,g). Finally, SREBPs failed to induce the transcription of the reporter luciferase under the control of sterol regulatory elements in *C12orf49*-null cells. These results suggest that C12orf49, like SCAP and MBTPS1, is necessary for SREBP activation and subsequent regulation of its biosynthetic targets.

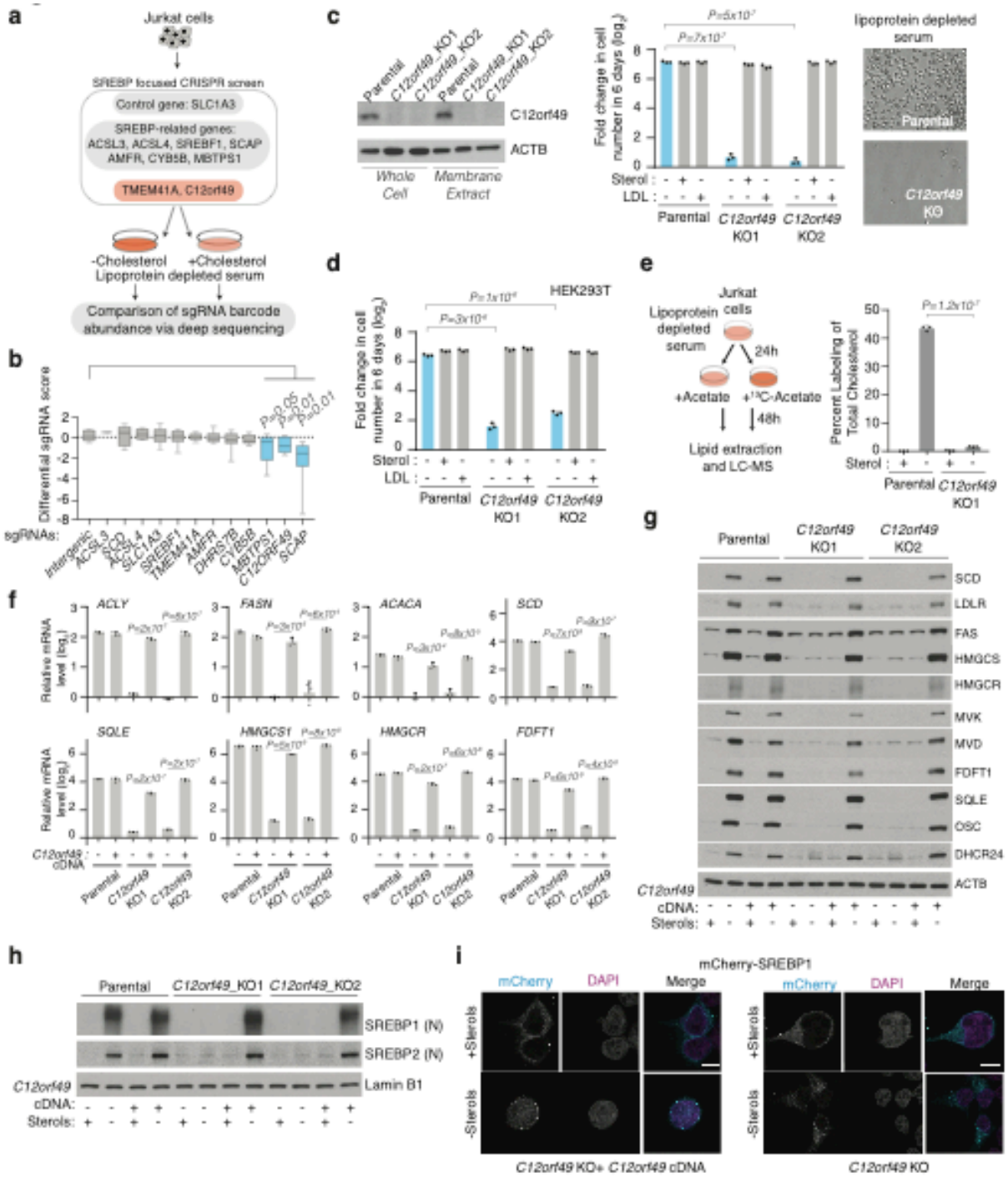
a
Jurkat cells
SREBP focused CRISPR screen
Control gene: SLC1A3
SREBP-related genes: ACSL3, ACSL4, SREBF1, SCAP AMFR, CYB5B, MBTPS1
TMEM41A, C12orf49
-Cholesterol
+Cholesterol
Lipoprotein depleted serum
Comparison of sgRNA barcode abundance via deep sequencing
b
Differential sgRNA score
sgRNAs:
Intergenic
ACSL3
SCD
ACSL4
SLC1A3
SREBF1
TMEM41A
AMFR
DHRS7B
CYB5B
MBTPS1
C12ORF49
SCAP
c
Parental
C12orf49_KO1
C12orf49_KO2
C12orf49
ACTB
Whole Cell
Membrane Extract
Fold change in cell number in 6 days (log2)
Sterol :
LDL :
Parental
C12orf49 KO1
C12orf49 KO2
lipoprotein depleted serum
Parental
C12orf49 KO
d
HEK293T
Fold change in cell number in 6 days (log2)
e
Jurkat cells
Lipoprotein depleted serum
24h
+Acetate
+13C-Acetate
48h
Lipid extraction and LC-MS
Percent Labeling of Total Cholesterol
f
ACLY
FASN
ACACA
SCD
SQLE
HMGCS1
HMGCR
FDFT1
Relative mRNA level (log2)
C12orf49 cDNA :
g
Parental
C12orf49 KO1
C12orf49 KO2
SCD
LDLR
FAS
HMGCS
HMGCR
MVK
MVD
FDFT1
SQLE
OSC
DHCR24
ACTB
C12orf49
cDNA:
Sterols:
h
Parental
C12orf49_KO1
C12orf49_KO2
SREBP1 (N)
SREBP2 (N)
Lamin B1
C12orf49
cDNA:
Sterols:
i
mCherry-SREBP1
mCherry
DAPI
Merge
+Sterols
-Sterols
C12orf49 KO+ C12orf49 cDNA
C12orf49 KO

Figure 4: C12orf49 is necessary for cholesterol synthesis and SREBP-induced gene expression in human cells

a, Schematic for the focused CRISPR-Cas9-based genetic screen. b, Differential sgRNA scores for the indicated genes. Blue bars indicate genes that are significantly and differentially essential under lipoprotein depletion. Boxes represent the median, and the first and third quartiles, and the whiskers represent the minimum and maximum of all data points. N = 8 independent sgRNAs targeting each gene except for previously validated sgRNAs for *ACSL3* (n = 3) and *ACSL4* (n = 4). c, Immunoblot of C12orf49 in the indicated cancer cell lines (left). Actin was used as the loading control. Log2(fold change) in cell number of Jurkat wild-type and C12orf49 KO cells following 6-d growth under lipoprotein depletion with the indicated treatment (mean ± s.d., n = 3 biologically independent samples) (middle). Representative images of indicated cell lines under lipoprotein depletion at the end of the experiment (right). LDL, low density lipoprotein. d, log2(fold change) in cell number of HEK293T wild-type C12orf49 KO cells following 6-d growth under lipoprotein depletion with the indicated treatments (mean ± s.d., n = 3 biologically independent samples). e, Mass isotopologue analysis of cholesterol in Jurkat wild-type and C12orf49 KO cells in the absence or presence of sterols after 48 h of incubation with 13C-acetate (mean ± s.d., n = 3 biologically independent samples). f, log2(fold change) in mRNA levels of SREBP target genes in indicated Jurkat cell lines following 8-h growth under lipoprotein depletion in the presence and absence of sterols (mean ± s.d., n = 3 biologically independent samples). g, Immunoblots of SREBP target proteins in indicated Jurkat cell lines following 24-h growth under lipoprotein depletion in the presence or absence of sterols. Actin was used as the loading control. h, Immunoblots of mature SREBP1 an dSREBP2 in indicated Jurkat cell lines following 24-h growth under lipoprotein depletion in the presence and absence of sterols. SREBP1 (N) and SREBP2 (N), mature SREBP N termini. Lamin B1 was used as the loading control. i, Localization of SREBP1 in *C12orf49*-null HEK293T cells expressing control or *C12orf49* cDNA under lipoprotein depletion in the presence or absence of sterols (scale bar, 8μM). The experiments were repeated independently at least twice with similar results. Statistical significance was determined by a two-tailed unpaired t-test.

Discussion

The metabolic coessentiality network offers an alternative method to discover unknown components of cellular metabolism and to functionally assign them to existing pathways. Using this method, we have identified *C12orf49* as an essential component of SREBP processing and cholesterol sensing in mammalian cells. It is not yet known precisely how C12orf49 contributes to the proteolysis of SREBPs but our findings suggest that its interaction with MBTPS1 is likely to be involved in the regulation of cholesterol metabolism. Remarkably, C12orf49 is highly conserved, even in lower organisms. As a subset of these organisms does not have an SREBP orthologue, yet harbours orthologues of C12orf49 and MBTPS1, the association between C12orf49 and MBTPS1 is likely to be relevant to cellular processes other than SREBP in these organisms. Interestingly, *C12orf49* is associated with hyperlipidemia, so additional work is required to understand whether this protein may be implicated in human disease or may have any clinical value. In conclusion, our work adds a new component to cellular cholesterol regulation and provides a platform by which to determine the function of other unknown metabolic components.

Materials and Methods

Metabolic coessentiality analysis

We adopted a three-step method to build a putative interaction network among genes on the basis of their coessentiality scores. In step one, we removed genes which were strongly correlated with a large number of genes because pathway analysis literature suggest that few proteins have many interaction partners. To do this, we

calculated a Pearson correlation network among all 17,638 genes with a threshold of $|r| = 0.25$. Then, we ranked the genes based on their degrees in this network and removed the top 10% from downstream analysis.

In steps two and three, we built partial correlation networks following the Correlation Analysis workflow proposed in section 3.1 of previous work (Basu et al., 2017). As the calculation of partial correlation among essentiality scores of many genes using fewer cell lines is computationally intensive, this workflow builds on a useful property of Gaussian graphical models that was previously established (Mazumder and Hastie, 2012). This property ensures that genes in different connected components of the partial correlation network are marginally uncorrelated. Therefore, we can first construct a network by applying a threshold on Pearson correlation, and then estimate partial correlation networks separately for each of its connected components.

In step two of our analysis, we built such a Pearson correlation network with a threshold $|r| = 0.35$. As we are only interested in finding novel genes that interact with metabolic genes, we removed all of the non-metabolic genes that are not connected to any metabolic genes in this network, using a curated metabolic gene set (Garcia-Bermudez et al., 2019; Possemato et al., 2011; Weber et al., 2020). Of note, we curated this metabolic gene set by exhaustive analysis of every known human gene combined with searches of the Kyoto Encyclopedia of Genes and Genomes database and literature verifying the known or proposed metabolic function of each gene (Possemato et al., 2011). We focused on positive Pearson correlations, which led to a network with

515 genes (275 metabolic genes and 240 non-metabolic genes), consisting of 55 components (component size varied between 3 and 20).

In step III, we calculated separate partial correlation matrices for each of these connected components and used statistically significant partial correlations (a false discovery rate (FDR) of <0.05) to construct the putative interaction network. We used the R function 'pcor' from the library 'ppcor' and debiased graphical lasso implemented in the DSPC software, as two different ways to calculate partial correlation networks. The debiased graphical lasso has an inbuilt regularization step and is particularly suitable when the number of genes in the network is high compared to the number of cell lines. As the Pearson network components were reasonably small, the results of the two methods were qualitatively similar and we reported the output from the pcor analysis in this paper. Finally, we removed interactions of genes within ±1 cytogenic band of each other in order to reduce false interactions, as CRISPR–Cas9 genome editing was reported to induce large truncations.

Cell lines

Cell lines HEK293T, Jurkat, MDA-MB-435, U-87 and BHK-21 were purchased from the ATCC cell lines were verified to be free of mycoplasma contamination and the identities of all lines were authenticated by short tandem repeat profiling.

Antibodies, compounds and constructs

Custom antibodies for C12orf49 and TMEM41A were designed and generated at YenZym Antibodies, using the following synthetic peptides; QEERAVRDRNLLQVHDHNQP (amino acids 37–56 of C12orf49) and

ETSTANHIHSRKDT (amino acids 251–264 of TMEM41A). Details of other antibodies, compounds, supplies, equipment, software, experimental models and constructs are provided in the supplementary files.

Cell culture conditions

Jurkat cells were maintained in RPMI medium (GIBCO) containing 2 mM glutamine, 10% FBS, penicillin and streptomycin. HEK293T, U87M and MDA-MB-435 cells were maintained in Dulbecco's modified Eagle's medium (DMEM) (GIBCO) containing 4.5 g l−1 glucose, 4 mM glutamine, 10% FBS, penicillin and streptomycin. All cells were maintained in monolayer culture at 37 °C and with 5% CO2.

Focused CRISPR-based genetic screen

The highly focused sgRNA library was designed by including representation of each gene within the SREBP module. For some of the genes, our sgRNAs have previously been published and validated (Zhu et al., 2019); we therefore used a smaller number of sgRNAs for particular genes. Oligonucleotides for sgRNAs were synthesized by Integrated DNA Technologies and were annealed before they were introduced in lentiCRISPR-v2 vector using a T4 DNA ligase kit (New England Biolabs), following the manufacturer's instructions. Ligation products were then transformed in stable competent E. coli cells (New England Biolabs) and the resulting colonies were grown overnight at 32 °C and plasmids were then isolated by the miniprep method (Qiagen). This plasmid pool was used to generate a lentiviral library containing five sgRNAs per gene target. This viral supernatant was titred in each cell line by infecting target cells at increasing amounts of virus in the presence of polybrene (8 µg ml−1) and by

determination of cell survival after 3 d of selection with puromycin. One million Jurkat cells were infected at a multiplicity of infection (MOI) of 1 before selection with puromycin for 3 d. An initial pool of one million cells was collected. Infected cells were then cultured for 14 population doublings in the lipoprotein-deficient serum (LPDS)-containing medium in the presence or absence of cholesterol, after which one million cells were collected and their genomic DNA was extracted by a DNeasy Blood & Tissue kit (Qiagen). For amplification of sgRNA inserts, we performed PCR using specific primers for each condition. PCR amplicons were then purified and sequenced on a MiSeq system (Illumina). Sequencing reads were mapped and the abundance of each sgRNA was measured. The sgRNA score is defined as the log2(fold change) in the abundance between the initial and final population of the sgRNA targeting a particular gene.

Isotope tracing experiments and lipid metabolite profiling

Jurkat cells were washed three times with PBS and plated as triplicates (1 × 106 cells per replicate) in 6-well plates using RPMI medium supplemented with 10% LPDS in the presence or absence of sterols (10 μg ml−1 cholesterol and 1 μg ml−1 25-hydroxycholesterol). After 24 h, medium was replaced with fresh medium containing sodium acetate (10 mM) or 13C1 sodium acetate (10 mM). Following an incubation of 48 h, cell pellets were washed twice with 1 ml of 0.9% NaCl (800g for 2 min) and resuspended in 600 μl of cold liquid chromatography–mass spectrometry (LC–MS) grade methanol. Non-polar metabolites were extracted by consecutive addition of 300 μl of LC–MS grade water followed by 400 μl of LC–MS grade chloroform. The samples

were vortexed (10 min) and centrifuged for 10 min at 20,000g at 4 °C. The lipid-containing chloroform layer was carefully removed and dried under liquid nitrogen. Dry lipid extracts were stored at –80 °C until further analysis.

Real-time PCR assays

Jurkat cells were washed 3 times with PBS and plated as triplicates (1 × 106 cells per replicate) in 6-well plates using RPMI medium including 10% LPDS supplemented with 50 μM compactin and 50 μM sodium mevalonate in the presence or absence of sterols (10 µg ml–1 cholesterol and 1 µg ml–1 25-hydroxycholesterol). After incubation for 8 h, RNA was isolated from cell pellets by a RNeasy Kit (Qiagen) according to the manufacturer's protocol. RNA was spectrophotometrically quantified and equal amounts were used for cDNA synthesis with the Superscript II RT Kit (Invitrogen). Quantitative PCR analysis was performed on an ABI Real-Time PCR System (Applied Biosystems) with the SYBR green Mastermix (Applied Biosystems). Primers for each target are provided in the supplementary files. Results were normalized to β-actin.

Analytical validation of method and comparison with alternatives

The Pearson correlation is the most commonly used method for building coessentiality networks among genes. Pan et al. have used genome-scale Pearson correlation networks to identify functional modules and protein complexes (Pan et al., 2018). However, gene networks based on statistically significant Pearson correlation tend to have many edges, including many false positives, which makes it difficult to identify suitable targets for novel gene interaction discovery and validation in the laboratory. Therefore, there is a need for computational methods with higher specificity

(lower false positives) that identifies fewer, but high-confidence, putative genetic interactions from data. In a recent study, Wainberg et al. proposed an alternative coessentiality network method based on generalized least squares, which explicitly accounts for non-independence of cell lines and reduces the number of false positives, and which has identified 93,575 significant coessential gene pairs (Wainberg et al., 2019). Although these comprehensive methods undoubtedly identified many novel gene functions, we wanted to create a conservative method that more easily allowed us to manually curate each individual network. As result, we looked towards alternative methods and filters that allowed us to short-list putatively novel gene interactions.

In essence, both of the methods described above measure pairwise association between two genes, without accounting for indirect or spurious effects due to their interactions with a third gene. Partial correlation, a canonical method in classical statistics, allows such indirect associations to be explicitly accounted for and produces a smaller, but high-confidence, set of putative interactions for follow-up validation in the laboratory. Whereas clustering on the basis of pairwise correlation allowed us to focus on a specific module of genes, the calculation of partial correlation among genes within the module helped us to focus on gene pairs that were more likely to interact directly. As a result, we were better equipped with a manageable list of gene interactions that could be studied on an experimental scale. This is in sharp contrast with the Pearson correlation-based methods described above, which only analyses the association between two genes at a time.

The principle of filtering out effects of other nodes in a network is at the core of graphical modelling literature in statistics and machine learning. Previous studies have successfully employed this idea to build metabolic networks (Basu et al., 2017; Krumsiek et al., 2011).

Lipotoxicity assays

Palmitic acid was conjugated to BSA. A 12 mM solution of the fatty acid was dissolved in 20 ml of 0.01 M NaOH and stirred for 30 min at 70 °C, followed by addition into a stirring 60 ml 10% BSA solution in PBS to make a final concentration of 3 mM. The solution was stirred for 1 h at 37 °C to allow fatty acids to conjugate with the BSA. Finally, the fatty acid–BSA solution was filtered through a 0.22 µm filter and stored in a glass container at 4 °C. Jurkat cells were cultured as triplicates in 96-well plates at 400 cells per well in a final volume of 0.2 ml of RPMI-1640 with increasing concentrations of palmitate. A duplicate plate without any treatment was used to determine initial luminescence on the day that plates were set up. To measure luminescence, 40 µl of CellTiter-Glo reagent (Promega) was added to each well according to the manufacturer's instructions and data were obtained using a SpectraMax M3 plate reader (Molecular Devices). Data are presented as relative fold change in luminescence of the final measurement to the initial measurement.

Statistical analysis

Sample size, mean and significance (P values) are indicated in the text and figure legends. Error bars in the experiments represent s.d. from either independent

experiments or independent samples. Statistical analyses were performed using GraphPad Prism 7 or as reported by the relevant computational tools.

Chapter 2: Developing a single-cell method to study the heteroplasmic effects of a ND4 loss-of-function mutation

Introduction

The role of the mitochondria in tumorigenesis and progression has had a gradual evolution over the past 90 years. The German physician Otto Warburg discovered that cancer cells produced excessive levels of lactate in the presence of oxygen also known as “aerobic glycolysis” or “the Warburg Effect” (Warburg, 1956; Warburg et al., 1927). This was counterintuitive in part because the energy requirements for cell proliferation would be more efficiently met by the complete catabolism of each glucose molecule through oxidative phosphorylation (Vander Heiden et al., 2009). Based on this observation, Warburg hypothesized that this increased dependence on glycolysis was due to dysfunctional mitochondria (Potter et al., 2016). Throughout the following decades, many studies tested Warburg’s hypothesis and found that in many cancer contexts the mitochondria are functional (Wallace, 2012).

From the time of Warburg, there has been a growing appreciation for mitochondrial function and their role in tumorigenesis. Many studies have shown that the inhibition of mitochondrial function leads to impaired cancer growth further signifying the role they serve in cancer progression (Cavalli et al., 1997; McBride et al., 2006; Morais et al., 1994; Pavlova and Thompson, 2016; Tan et al., 2015; Wallace, 2012; Weinberg and Chandel, 2015). Running parallel to these studies, many other groups have matured and evolved our understanding of mitochondrial biology. They have shown that the mitochondria are central organelles involved in cellular bioenergetics,

biosynthesis, and signaling. One of their most notable roles is the production of cellular energy in the form of adenosine triphosphate (ATP), leading to their popular nickname “the powerhouse of the cell.” However, the mitochondria are not limited to just energy production and are involved in the generation of reactive oxygen species (ROS), redox molecules and metabolites, regulation of cell signaling and apoptosis. This intersectionality positions the mitochondria to respond and adapt in order to promote tumor progression (Vyas et al., 2016).

The electron transport chain (ETC) and by extension the tricarboxylic acid (TCA) cycle are examples of mitochondrial functions that significantly impact the adaptability of tumor cells. The ETC is a central cellular process that regulates proper cellular metabolism. Specifically, the ETC is a series of protein complexes that transfer electrons from electron donors to acceptors and couples this with the transfer of protons into the inner membrane space, where oxygen is the final electron receptor. The transfer of protons into the inner membrane facilitates proper oxidative phosphorylation (OXPHOS) function and consistent ATP production. Proper ETC function also allows for the proper functioning of the TCA cycle. The TCA cycle is a key metabolic process because of the large number of substrates that feed into and out of it. It connects lipid, protein, and carbohydrate metabolism and normally produces acetyl-coenzymeA (acetyl-CoA), reduced nicotinamide adenine dinucleotide (NADH), and carbon dioxide (CO_2). The TCA cycle begins with a reaction to combine two-carbon acetyl-CoA that feed in from pyruvate, fatty acids, or amino acids, with oxaloacetate to citrate. Citrate is subsequently converted to its isoform isocitrate. Isocitrate proceeds to be

decarboxylated twice forming α-ketoglutarate, then succinyl-CoA, while forming two molecules of CO_2 and NADH. Succinyl-CoA is converted to succinate while creating guanosine triphosphate (GTP) which can be converted to ATP. The TCA cycle and ETC are linked by the conversion of succinate to fumarate by succinate dehydrogenase (SDH) because SDH also has the dual role of being the ETC complex II. In this step, two hydrogen atoms are transferred from flavin adenine dinucleotide (FAD) to $FADH_2$. Finally, fumarate gets converted into malate and eventually malate becomes oxaloacetate which restarts the cycle. Although the TCA cycle is described as a cycle, certain reactions can be acyclic and metabolite intermediates can be substrates for other reactions (Martínez-Reyes et al., 2020). These other reactions can be rewired in order to promote tumorigenesis and progression. For example, cancer cells upregulate the production of acetyl-CoA by increasing the activation of ATP-citrate lyase (ACLY) to increase their proliferative capacity. On a larger scale, intracellular α-ketoglutarate and succinate ratio has been shown to contribute to the regulation of cellular identity and have a role in the transcription and epigenetic state of stem and cancer cells (Carey et al., 2015; Morris et al., 2019). Taken together, the ETC and the TCA cycle provide cells the potential to adapt in different contexts.

The essentiality of the ETC for a cell's ability to adapt would thus suggest that any process that inhibits the ETC would be selected against. One would postulate that truncation mutations, also known as loss-of-function (LOF) mutations, would be strongly selected against in genes that encode for subunits of the ETC. As mentioned above, the ETC is a series of five protein complexes where each complex is made up of individual

subunits. Most of the ~100 subunits that comprise the ETC are encoded in the nuclear genome, but there are also 13 essential subunits encoded in the mitochondrial genome (mtDNA). While severe ETC dysfunction can lead to genetic diseases and cell death, we and other groups found the occurrence of truncation mtDNA mutations across a variety of cancer types (Brandon et al., 2006; Polyak et al., 1998). Surprisingly, the accumulation of truncation mutations in the nuclear encoded ETC genes are not as prevalent as ETC genes in the mtDNA (Gorelick et al., 2020). Despite their prevalence, it is unclear whether these mutations have functional roles in thyroid cancer progression or are simply passenger mutations as the study of mtDNA mutations is impeded by a lack of methods to genetically modify the mtDNA.

One reason for the lack of methods is the mitochondrion is a unique organelle in that they have their own mitochondrial localized genome distinct from the nuclear genome. The mtDNA is approximately 16,500 base pairs long and encodes for 37 genes, including 13 protein components of the ETC, two rRNA genes and the remainder being tRNA genes. Each cell contains multiple copies of mtDNA, so the variant allele frequency of any mtDNA mutation, termed heteroplasmy, may range from 0-1. Unlike the nuclear genome where any mutation either occurs as homozygous wild-type, heterogenous, or homozygous mutant. Thus, it is believed that there is a critical proportion, known as the threshold effect, of mtDNA that need to be mutated before a phenotypical effect can be observed and this threshold likely differs across cell types (Chinnery et al., 2001; Gorman et al., 2016; Jackson et al., 2020; White et al., 1999). This difference in heteroplasmy is further complicated by the fact that clonal cells may

have different heteroplasmy levels, a concept known as inter-cellular mtDNA heterogeneity (Aryaman et al., 2019). In order to study effects of cells with high heteroplasmy for a truncation mutation (increased ETC dysfunction) compared to cells with low heteroplasmy for a truncation mutation (decreased ETC dysfunction), one would need to create isogenic populations of cells where only the truncation mutation heteroplasmy is altered. These types of experiments have largely become routine with the advent of CRISPR-Cas9 gene editing, but effective gene editing techniques remain elusive for the mtDNA (Gammage et al., 2018). This is not due to a lack of effort as many techniques such as cytoplasmic hybridization, MitoTALEN, and bacterial cytidine deaminase are able to edit mtDNA, but are limited by their high selective pressure, efficiency, and mutation types, respectively (Bacman et al., 2013; Mok et al., 2020; Picard et al., 2014). Furthermore, none of these techniques explore a potential explanation for the heterogeneity of mtDNA mutations within a given population. Here, we aim to test the hypothesis that heteroplasmic LOF mutations have functionally critical roles in tumorigenesis and are a potential reason for why these cells are maintained in a population.

Results

Choosing a model experimental system

High heteroplasmy cells exists in different cancer types, levels, mutation types and positions, thus there is a wide range of possible model systems that could be chosen. We systematically identified the appropriate experimental model by first

choosing the mutation type, cancer type, and cell type. We decided to solely focus on out-of-frame truncation mutations because they are most likely to cause ETC dysfunction, whereas missense mutations carry the potential of having more complex functions. In the context of studying mtDNA mutations, we do not have a bias towards studying any particular cancer, thus we decided to focus on a cancer type that had a high occurrence of mtDNA truncation mutations that were enriched for high heteroplasmy. We took publicly available mtDNA mutation calls from The Cancer Genome Atlas (TCGA) whole-genome samples to determine which cancer type best fit our above criteria. We along with many other groups showed that thyroid, kidney, and colorectal cancer were the three cancer types that had the highest average for any type of truncation mutation (Figure 5a) (Gorelick et al., 2020; Grandhi et al., 2017; Yuan et al., 2020). Additionally, approximately 20% of samples for each cancer type have at least one truncation mutation in the mtDNA (Figure 5b). Based on these analyses, it was difficult distinguishing among those three cancer types. The distinction became clearer when we stratified the cancer types by heteroplasmy enrichment. We compared the heteroplasmy enrichment between truncation and silent mutations for each cancer type and found that thyroid cancer clearly has a high enrichment for heteroplasmic mutations compared to other cancer types (Figure 5c). The exact enrichment is further emphasized by looking at the median heteroplasmy level differences among the truncation, missense and silent mutations (Figure 5d). However, a limitation of TCGA thyroid samples is they are largely papillary thyroid cancer. To overcome this limitation, we examined off-target genomic sequence from MSK-IMPACT because of its coverage

of other thyroid cancer types (Figure 5e,f). We found that the majority of the thyroid cancer types that were profiled have at least 10% of samples with truncation mutations and on average their heteroplasmy levels are above 50%. The prevalent occurrence of

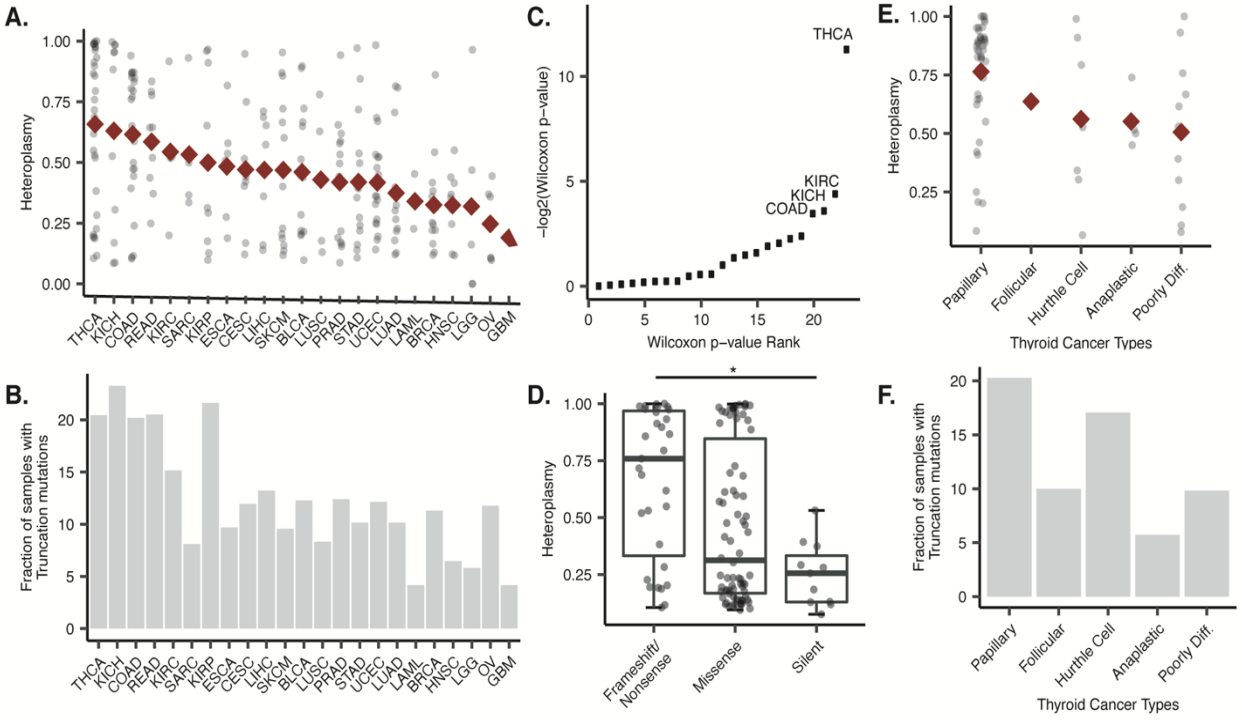

Figure 5: Thyroid cancer is enriched for mtDNA truncation mutations
a, Distribution of truncation mutations for each sample by cancer type. Red dot represents the average heteroplasmy b, The proportion of samples that have at least one truncation mutation by cancer type. c, A plot showing the significance between frameshift/nonsense and silent mutation for each cancer type in the TCGA. Abbreviations used are adopted from the TCGA. d, The boxplot displays the LOF heteroplasmy level for each thyroid cancer sample in the TCGA. e,f is the same as a,b but using the MSK-IMPACT data specifically on thyroid cancer subtypes.

mtDNA truncation mutations in different thyroid cancer types thereby provides the option to study any of these subtypes. Lastly, we needed to decide which was the most appropriate model to study heteroplasmy in thyroid cancer with some options being a thyroid tumor, mouse model, or cancer cell line. In order to study the effects of one mutation, we would need to create two populations that have an isogenic background

with the except of one mtDNA truncation mutation. Using a variety of different samples within a cancer type is complicated by differences in oncogenic driver mutations and sample-to-sample variability suggesting that intracellular comparison groups of high and low heteroplasmy levels are useful to study the role of LOF mtDNA mutations. With this consideration in mind, we decided to focus on cell lines because they are believe to be pure, genetically identical, easily propagated, and can be genetically manipulated (Saiselet et al., 2012). We searched through the Cancer Cell Line Encyclopedia (CCLE) to identify a thyroid cancer cell line with only one truncation mutation in the mtDNA. The analysis narrowed our study to the FTC133 and KHM_5M cell line because they fit the above criteria and have the same ND4 11866 C insertion mutation at different heteroplasmy levels (Figure 6).

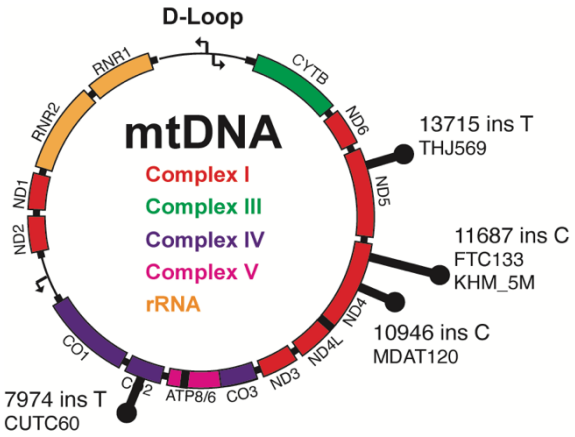

Figure 6: Location of mtDNA mutations for five thyroid cell lines Black bar represents the mutation position and height refers to the number of cell lines with that mutation.

FTC133 and KHM_5M have opposing heteroplasmy distributions

The CCLE provides many publicly available genomic data on over a thousand cell lines, including FTC133 and KHM_5M. However, bulk sequencing is unable to capture the intercellular heteroplasmy heterogeneity that we assumed was present across each individual FTC133 and KHM_5M cell. To this address this issue, we turned towards single-cell RNA-sequencing (scRNAseq). scRNAseq captures the

transcriptome of each individual cell providing higher resolution understanding for cell function. Traditionally, scRNAseq has been used to study the cell type heterogeneity in primary tumors because it is considered to be a diverse population of cells, whereas a cell line is considered to be clonal. Although cell line nuclear genomes are considered to be clonal, we hypothesize that the mtDNA heterogeneity is much more diverse.

We tested the intercellular heteroplasmy heterogeneity hypothesis by taking FTC133 and KHM_5M and running them through the 10x single-cell sequencer. 10x scRNAseq that we used is only able to sequence from the 3' end, which is mainly used to build the single-cell transcriptome. Fortunately, the ND4 11866 C insertion is located near the 3' end and that position is adequately sequenced. We were able to capture both the transcriptome and the ND4 heteroplasmy for each cell (Figure 7a). Before downstream analyses, we filtered out cells with low transcript coverage and cells with above 20% mitochondrial content. These two parameters are signs that cells have lysed during the 10x emulsion step and need to be removed. The remaining 1673 FTC133 cells and 1885 KHM_5M all contain the ND4 11866 C insertion at some heteroplasmy level and at sufficient mutation depth.

Based on the mutation calls, we explored the heterogeneity of mtDNA mutations within a clonal population. The mtDNA, relative to the nuclear DNA, has a higher mutation rate and undergoes asymmetric inheritance (Taylor and Turnbull, 2005). The concept of asymmetric inheritance is the idea that when cells divide their mitochondria are passed on in an unsystematic manner. Thus, it is likely that over time some cells will inherit a differing amount of mutated mitochondria, contributing to the overall

heterogeneity (Aryaman et al., 2019). We addressed this possibility by tabulating the

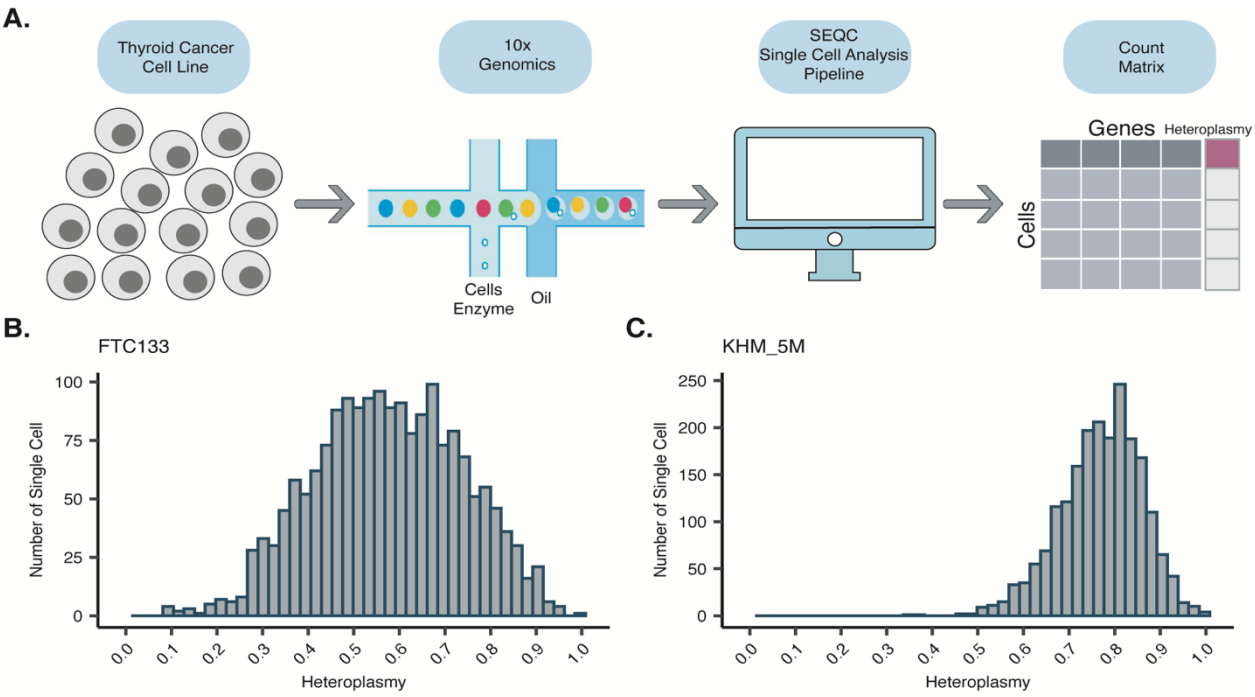

Figure 7: scRNAseq shows intercellular heteroplasmic heterogeneity
a, The scRNAseq and heteroplasmy calling workflow. Key steps are written in text. b, c Histogram of cells at each heteroplasmy interval between 0 and 1 and broken up by 0.1. The cell line is indicated above each histogram.

spread of heteroplasmy across both FTC133 and KHM_5M (Figure 7b,c). As we can see, these two cell lines tell a different story with FTC133 having a wide and almost normally distributed heteroplasmy distribution centered around 57%, while KHM_5M having a relative tight distribution centered around 77%. The wider FTC133 heteroplasmy range raises the possibility that there is biological significance associated with these cells, whereas the clonality of KHM_5M is more aligned with the assumption that cell lines are largely clonal. Regarding KHM_5M, the logical next question is to ask whether the heteroplasmy distribution is tight due to chance or selection, but currently this remains to be determined.

Single-cell sequencing revealed a correlation between the ND4 11866 C insertion mutations and the whole-cell transcriptome in FTC133

Although the distribution of heteroplasmy levels raises the possibility of biological significance, the presence of a mutation alone is not sufficient to result in phenotypic changes at the cellular level. Therefore, we determined whether this ND4 mutation results in a correlated change in the single-cell transcriptome and reasoned that these global transcriptomic changes cannot be easily rationalized through simple stochasticity of mutation acquisition because random alterations should not occur consistently. In order to determine this, we projected cells onto a Uniform Manifold Approximation and Projection (UMAP) in an unbiased manner and asked whether the occurrence of any clustering correlates to heteroplasmy changes (Figure 8a,b). The strength of the clustering approach is that it reveals consistent transcriptomic alterations within a population and is robust to noise that may be present at an individual cellular level. Here we can see that there are distinct clusters that form in FTC133, but not KHM_5M, which is consistent with the range in heteroplasmy distribution (Figure 7b, c). Due to the lack of clusters in KHM_5M, we turned our attention towards elucidating the changes that are found in FTC133.

Consistent with the visual UMAP representation, unbiased cluster assignment showed that there are four distinct clusters that have a corresponding increase in heteroplasmy as you move from top left to bottom right (8c,d). Off of first glance, it is tempting to jump to the conclusion that these clusters are solely explained by changes in heteroplasmy. Before we can confidently make this conclusion, we need to first rule

out the possibility that these clusters can be explained by other means such as, cell contamination, cell cycle, coverage, mitochondrial read percentage, and ND4 mutation depth (Figure 8e,f,g,h). We confirmed through short-tandem repeat (STR) analysis that our FTC133 cell is 100% similar to the American Type Culture Collection (ATCC) database proving that there is not another cell contamination. Prior studies have shown that cell cycle gene regulation may be a confounding variable in explaining the heterogeneity of a cell population (Buettner et al., 2015; McDavid et al., 2016). We used Seurat to calculate the cell cycle score of each cell and found that none of the four clusters are explained by any cell cycle phase (Figure 8e) (Butler et al., 2018; Kowalczyk et al., 2015). Within each cluster, however, we can see that cells are partially separated by the cell cycle suggesting that the cellular transcriptome is more strongly correlated to the ND4 heteroplasmy than it is to the cell cycle. Lastly, we largely see that there are not concerning differences in coverage, mtDNA read percentage, and mutation depth (Figure 8e,g,h). Cluster 3, albeit, is lower than the other clusters, but they still have high quality coverage.

Based on heteroplasmy levels, cluster 2 is the most similar to the parental cells at ~57%. Comparing cluster 2 to the others suggests that an absolute change $\geq$ 10% in either direction results in transcriptomic changes, lower in the case of cluster 1 and higher in the cases of clusters 3 and 4. Furthermore, the differences between cluster 4 and cluster 1 are approximately 20%, which is considered to be a significant

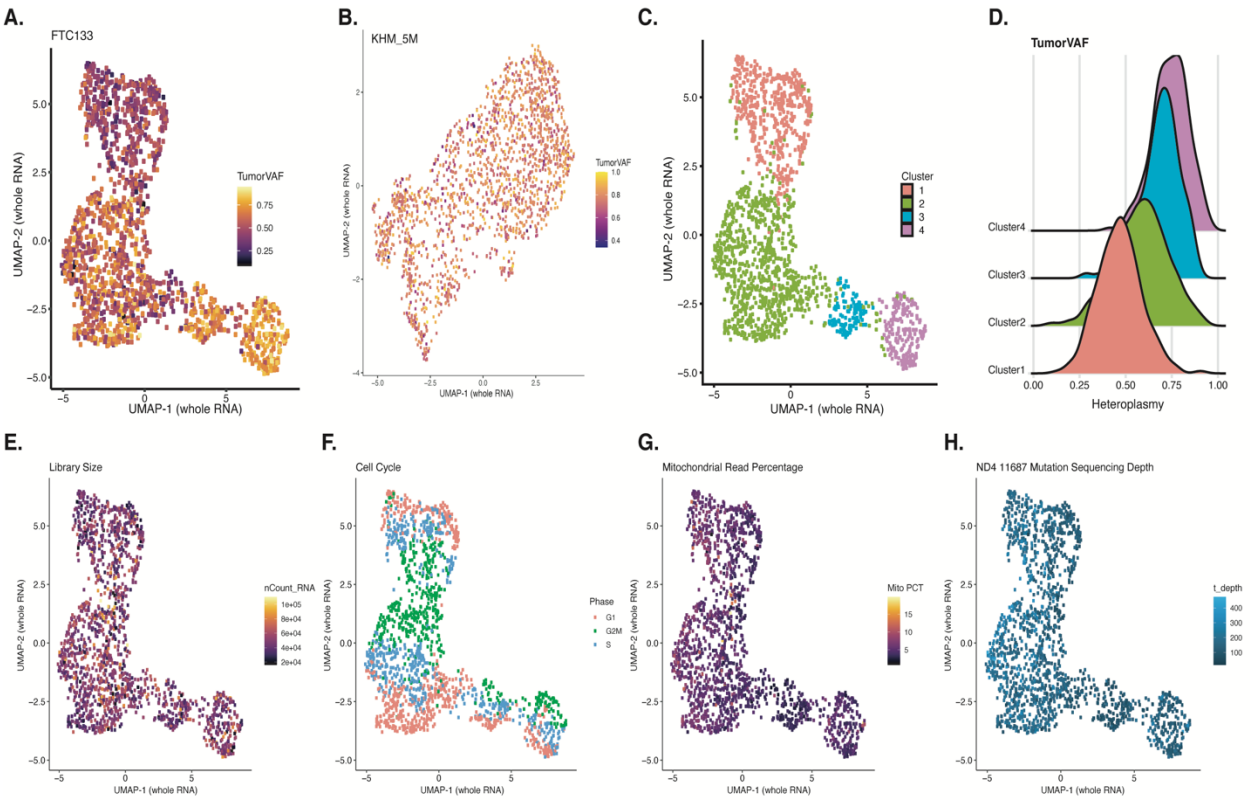

Figure 8: ND4 11866 C insertion heteroplasmic mutations correlates with whole-cell transcriptomic alterations
a,b UMAP visualization of FTC133 cells (n = 1,673) and KHM_5M (n = 1,885) based on their whole transcriptomes. c, Same as a, but colored by cluster ID from UMAP based on the whole-transcriptome. d, ND4 11866 C insertion heteroplasmy distribution by each cluster shown in c. e, Same as a, but colored by total library size. f, Same as a, but colored by normalized expression of G1/S/G2M marker genes by their overall expression levels. g, Same as a, but colored by the mitochondrial gene expression percentage. h, Same as a, but colored by ND4 11866 C insertion mutation depth.

heteroplasmic shift (Tasdogan et al., 2020). Yet in order to further study the differences between these two clusters, we needed to devise a new strategy to isolate cluster 1 and cluster 4 cells.

Using transcriptomic differential gene expression as a strategy to isolate low and high heteroplasmic cells

Originally, we devised a number of strategies to isolate low and high heteroplasmic cells including drug selection and single-cell cloning, but both did not produce large enough heteroplasmic changes and introduced additional biases that forced us to go into another direction.

We decided to think about our isolation method based on first principles. We asked ourselves what would we ideally like to see happen assuming we could identify or develop a technological approach. In other words, if we could make up how a technique would work, regardless of whether or not the technology exist, how would it best operate. In order for this to be possible there needs to be some identifier or marker to denote which cells are from cluster 1 or cluster 4. There is a distinction here that is made between cluster 1 and 4 cells compared to simply finding identifiers or markers that correspond to low or high heteroplasmy cells using the scRNAseq transcriptomic data.

Evident from figure 7, there are clearly many low heteroplasmy cells in cluster 4 and many high heteroplasmy cells in cluster 1. Therefore, simply taking high or low heteroplasmy cells may not be selecting for a consistent phenotype. It is a real possibility that these ND4 mutations in some cells are not functional, meaning that they do not elicit an ETC LOF phenotype. Another possibility is that individual cells that do in fact have LOF mutations in ND4 may have found unique and distinct pathway rewiring that has allowed them to adapt in a manner that circumvents the need for proper ETC function. There are many potential explanations in the cancer biology field for how cells are able to overcome dysfunctional ETC. One key metabolite that proteins, such as

Glutamic-Oxaloacetic Transaminase 1 (GOT1), are associated with the ETC is aspartate. It has been shown that chemical inhibition of the ETC Complex I with chemical inhibitors like phenformin or Piericidin A induces a reliance on the functioning of GOT1 to produce aspartate (Birsoy et al., 2015). Another example is under low-oxygen environments, also known as hypoxic environments, some cancer cells have been shown to upregulate solute carrier family 1 member 3 (*SLC1A3*). Low-oxygen environments are another form of ETC dysfunction because without oxygen there is not a final reducing agent at the end of the ETC. Under this different form of ETC stress, SLC1A3 transports aspartate to rescue cancer cell growth across various cancer types (Garcia-Bermudez et al., 2018). These are some established cellular reprogramming that have been shown in the literature, but there are many other possible adaptations that could occur, such as importing more glucose in order to increase glycolytic flux. There are many potential explanations as to why not all low heteroplasmy cells cluster together or why not all high heteroplasmy cells cluster together.

The clustering analysis comes with many benefits, but there remains a shortcoming that can be addressed with differential expression analysis. One caveat with clustering cells based on their single-cell transcriptome is that cells may have similarly expressing genes that are not related to the ND4 heteroplasmic mutation. Although, we have done our best to account for any technical noise surrounding the scRNAseq data, it still remains possible that these gene correlations are explained by other factors. To address this issue, we also conducted differential gene expression analysis using linear regression to correlate ND4 heteroplasmy levels with the

expression of each gene and a pseudotime analysis to determine the expression trajectory (Cao et al., 2019) (Figure 9a,b). The result of these analyses provides genes that most strongly correlated to ND4 mutations. The volcano plot stratifies the genes that are most differentially expressed by expression log fold-change and multiple hypothesis corrected significance. The pseudotime analysis shows many overlapping genes as the volcano plot, but provides additional information regarding the two states of cells within FTC133. The pseudotime analysis is broken into three parts, the "closing", "transient", and "opening". These terms refer to differentially significant genes upregulated in the low, intermediate, and high heteroplasmy cells, respectively. The green coloring symbolizes genes that are highly expressed and they only consistently appear on the low and high heteroplasmy ends. This suggests that there are not a large number of genes that are significantly differentially expressed in the intermediate heteroplasmic cells. This result came as a surprise to us because the majority of cells fall into this heteroplasmic level which raised the possibility that there could be some positive selection occurring and could be reflected in the transcriptome (Figure 7b). In light of these data combined with the 20% heteroplasmic mutation difference between cluster 1 and cluster 4, we were further encouraged to focus on the extreme ends of the heteroplasmic mutation range.

Overlapping the intersection between the clustering and the differential gene expression analysis provides the most robust gene list, which can be used to identify the most distinguishing gene markers. In the previous paragraph, I explained the process of identifying genes that most strongly correlate with the ND4 heteroplasmic

mutation. Here, I will describe the process for identify markers genes that are most closely representative of each cluster. For each cluster, we filtered these genes by only including the ones that were differentially expressed with an adjusted p-value of 0.05. We only considered genes that were most highly expressed in that cluster along with a minimum transcript per million threshold of 50. The most stringent filtration that we did was only consider genes that had a two-fold expression difference between the highest expressing cluster and the second highest expressing cluster. Then, the intersection between these genes along with the differentially expressed genes from the regression analysis were taken. Collectively, we felt that the resulting genes would most distinguish each cluster, while the intersection increases the confidence that these genes would be associated to the heteroplasmy mutation levels.

After taking the intersection, we noticed that there were many cell surface protein genes. The occurrence reminded us of how many immunologist utilize cell surface proteins in order to isolate key cell populations to conduct further experiments (Kalina et al., 2019). Based on these cell surface genes, we then hypothesized that these different heteroplasmic cell populations could be sorted through a fluorescence-activated cell sorting (FACS). The idea was simply to use fluorescence-conjugated antibodies that are sensitive to these cell surface proteins.

We further matured the strategy by systematically identifying the correct cell surface proteins. There were a few options, but the choices were limited to CD82, ITGA2 (CD49b), and CD300c to isolate cluster 4 cells and TSPAN8, CD24, and CD9 for the not high heteroplasmy cells (Figure 9c). Unfortunately, we were unable to identify

cell surface protein markers that are distinct to cluster 1, but there are a few markers that are not found in the high heteroplasmy clusters (cluster 3 and cluster 4). The trick that we developed was to mix the use of antibodies from each group. The purpose of this is to improve the mutual information gathered from each cell and this idea comes from probability theory and statistics (Vergara and Estévez, 2014). The concept is more simply stated as we want to be able to experimentally gather as much unique information about each cell as we can and that this information does not overlap into each other. From figure 9c, we observed that there are cells in cluster 1 that are positive for the cluster 4 markers and vice-versa. This clearly would complicate the FACS sorting because the cross-contamination of cells would dilute the isogenic populations. However, what we observed was most cluster 4 cells that were positive for their protein markers were also more likely to be negative for the cluster 1 protein markers. To capitalize on this observation, we created pairwise antibody mixtures, one from the cluster 1 marker group and one from the cluster 4 group (Figure 9d). We created three combinations: CD82 - TSPAN8, ITGA2 (CD49d) – CD9, and CD300c – CD24 and used them to optimize the best combination that will ultimately be carried toward future experimentation.

The parental FTC133 cell lines were stained with each of these combinations and proceeded to the FACS machine to be sorted. At the scRNAseq phase of the project, two legitimate concerns were that the increased expression would not lead to an increase in an upregulation of protein translation and that these proteins may not localize to the cell surface. If either of these two concerns were true, then the FACS-

based method would not work. However, during the analysis what we quickly observed were discrete clusters resembling the scRNAseq cluster analysis (Figure 8e). This was encouraging because the FACS analysis partially validates the technical considerations because some of these marker genes are expressed at the cell surface. Using the CD82-TSPAN8 as an example, we calculated a change in heteroplasmy with the $CD82^-$ and $TSPAN8^+$ (low heteroplasmy) population as 54%, while the high heteroplasmy population is 70% (Figure 9f). This result completely validates our scRNAseq and FACS-based method because we successfully isolated a low and high heteroplasmy population that is approximately 16% different. To ensure that our method was robust to any potential artifacts of the CD82 and TSPAN8, we continued to sort with the other two combinations. The CD49b and CD9 combination also work similarly well and proved that the isolation of these populations is not specific to CD82 and TSPAN8. The CD300c and CD24 combination did not work because the CD300c was not expressed on the cell surface if it is expressed at all. Revisiting the scRNAseq clusters, we calculated that the cluster 4 heteroplasmy is at 74%, while the cluster 1 heteroplasmy is at ~46%. This suggests that our sorting strategy has some further optimization that could be done to increase the heteroplasmic range.

In order to create ideal heteroplasmic populations, we sought to improve our sorting strategy by combing CD82, CD49b, and CD9. This combination proved to be the most effective in that the low heteroplasmy population decreased its heteroplasmic mutation level to 50%, while the high heteroplasmy cells increased to 74%. These heteroplasmy levels more closely resemble that of the scRNAseq calculations and

crosses the 20% heteroplasmic mutation threshold that is believed that be a significant change (Tasdogan et al., 2020). One potential concern with these newly sorted heteroplasmic populations is that the high heteroplasmy cells may have cellular stress from the increased ETC dysfunction. If this is the case, then the phenotypic manifestation would be that the high heteroplasmic cells may arrest or the heteroplasmy levels may quickly decrease over time. Counterintuitively, the high heteroplasmy cells maintain stable heteroplasmy levels, while also has a steady growth rate albeit slower than the low heteroplasmy population (Figure 9g).

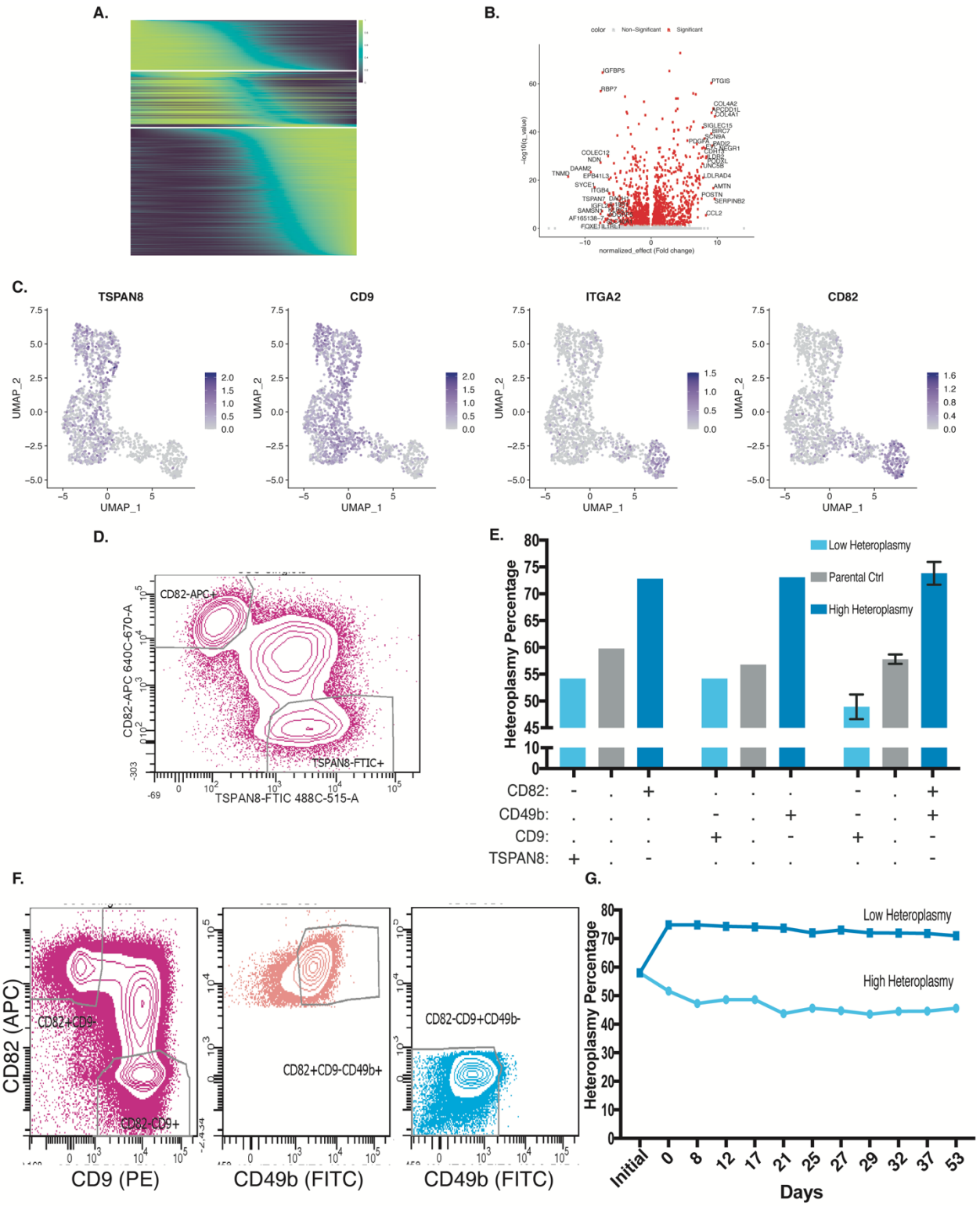
A.
B.
color Non-Significant Significant
JGFBP5
PTGIS
RBP7
COL4A2
ABCD1L
COL4A1
SIGLEC15
BIRC7
SCN9A
PDGFA
PADI2
EGLN3
NEGR1
CDH13
ADRB2
UNC5B
LDLRAD4
AMTN
POSTN
SERPINB2
CCL2
COLEC12
NDN
DAAM2
TNMD
EPB41L3
SYCE1
ITGB4
TSPAN7
DACH1
SFTA2
SAMSN1
AF185138-2
FOXE1
LTHL1
-log10(q_value)
normalized_effect (Fold change)
C.
TSPAN8
CD9
ITGA2
CD82
UMAP_1
UMAP_2
D.
CD82-APC+
TSPAN8-FTIC+
CD82-APC 640C-670-A
TSPAN8-FTIC 488C-515-A
E.
Low Heteroplasmy
Parental Ctrl
High Heteroplasmy
Heteroplasmy Percentage
CD82:
CD49b:
CD9:
TSPAN8:
F.
CD82 (APC)
CD82+CD9-
CD82-CD9+
CD82+CD9-CD49b+
CD82-CD9+CD49b-
CD9 (PE)
CD49b (FITC)
CD49b (FITC)
G.
Low Heteroplasmy
High Heteroplasmy
Heteroplasmy Percentage
Initial
0
8
12
17
21
25
27
29
32
37
53
Days

Figure 9: Markers can be used to isolate different heteroplasmy cell populations

a, Pseudotime analysis displaying gene enrichment (y-axis) as a function of heteroplasmy (x-axis). The color intensity is proportional to the level of gene expression. Lime-green refers to increased expression and purple refers to decreased expression. b, Volcano plot displaying linear regression differential gene expression as a function of heteroplasmy. Red dots are genes that are FDR ≤ 0.05, and gray dots are genes that are FDR > 0.05. c, UMAP visualization of FTC133 cells (n = 1,673 based on their whole transcriptomes, but colored by *TSPAN8*, *CD9*, *ITGA2*, and *CD82* normalized gene expression. d, CD82 and TSPAN8 fluorescence on the parental FTC133. e, Heteroplasmy percentage on the same day as FACS-based sorting using four different markers. The "+" and "-" refers to use of that antibody, where "+" and "-" symbolizes positive and negative expression, respectively. f, Same as d, but the fluorescence corresponds to CD82, ITGA2 (CD49b), and CD9. g, ND4 11866 C insertion heteroplasmy percentage for the low and high heteroplasmy cell population sorted from f.

Heteroplasmic ND4 11866 C insertion mutations cause severe ETC dysfunction

The stability of the heteroplasmy levels are reassuring when considering these populations are meant to be experimental models, but it raises the possibility that these mutations may not be functional. This possibility could be for a number of reasons, for instance, the mutations do not occur at a sufficient heteroplasmy level to elicit a phenotypic response. There are many phenotypic changes that could be explored that are related to ETC dysfunction, but we were able to shortlist these possibilities by leveraging past research on ETC dysfunction using chemical inhibitors and previous studies that explored the effects of mtDNA mutations (Jain et al., 2016; Quirós et al., 2017). Due to the difficulties with creating a mtDNA experimental model, the field has augmented their findings using chemical inhibitors like phenformin, piericidin A, antimycin, doxycycline, actinonin, carbonyl cyanide-4(trifluoromethyoxy)phenylhydrazone, and MitoBloCK-6.

Before moving into more complex phenotypic effects of ETC dysfunction, we first addressed whether the basal oxygen consumption rate (OCR) is affected. Oxygen is the final election receptor for the ETC and thus it is expected that without proper ETC function, oxygen would be consumed less. We can see that there is a clear decrease in OCR as each population increases in heteroplasmy level (Figure 10a). As we would expect, the low heteroplasmy population, also the population with the most ETC function, consumes the most oxygen while the opposite is true regarding the high heteroplasmy population. This result suggests that these mutations are in fact functional, but the level of dysfunction to which these mutations cause remains to be determined.

The severity of dysfunction of the high heteroplasmy population was determined based on phenotypic similarities that they shared with cells that are already known to have severe ETC dysfunction. The process of making rho0 cells or cells without mtDNA requires that these cells be grown in surplus amounts of pyruvate (King and Attardi, 1989). Some reasons for the proliferation rescue by pyruvate is due to restoration of the NADH:NAD^+ imbalance through lactate dehydrogenase and the production of aspartate (Birsoy et al., 2015; Han et al., 2008; Harris, 1980). These examples are by no means an exhaustive list of hypotheses to explore, however, it is not necessary for all hypotheses to be true for the cells to be consider under mild or severe ETC dysfunction.

We reasoned that under mild ETC dysfunction there would be a difference between the NADH:NAD^+ and potentially a relative aspartate abundance difference. Consistent with the differences in basal OCR, we would expect that there would be an

accumulation of NADH because the ETC would be less able to oxidize NADH back to NAD^+. We hypothesized that there would not be a significant difference between the low and high heteroplasmy population because these mtDNA mutations are examples of chronic ETC dysfunction. The examples that were previously cited are instances of acute mitochondrial stress meaning the cells have not experienced ETC dysfunction for many generations. This is particularly true in the studies where ETC chemical inhibitors are used and still true in the case of cybrid cells because they have only had dysfunctional mitochondrial for a relatively short time. In comparison, the FTC133 cell line has been growing for generation across various institutes and hands, so they are more likely to have adapted to the ETC dysfunction by any of the mechanisms that was discussed above. We tested this hypothesis by performing metabolomics on these isolated populations and quantified the abundance of NAD^+, NADH, and aspartate (Figure 10b,c). To our surprise, we found that there is a significant NADH:NAD^+ imbalance and it is alleviated in the presence of pyruvate. Additionally, we also observed a similar change in aspartate. Collectively, this would suggest that the high heteroplasmic cell population is at the very least mildly dysfunctional.

Building upon the NADH:NAD^+ and aspartate results, we next assessed the depth of the dysfunction. From the literature, we reasoned that rho0 cells, without any mtDNA, should be the most extreme version of heteroplasmic LOF mutations. Alternatively, Birsoy et al. have shown that only under excessive amounts of ETC chemical inhibitor would proliferation rescue by pyruvate supplementation work. Thus, if pyruvate supplementation is required for cell proliferation then by inductive reasoning

those cells would be considered to have high ETC dysfunction. We grew the low and high heteroplasmy cells in media without pyruvate and with 1.5mM pyruvate over 6-days (Figure 10d,e,f). We can see that both populations benefit from the supplementation of pyruvate, but there is approximately a six-fold increase in proliferation for the high heteroplasmy cells. The addition of Piericidin A is synergetic to the inhibition of cell growth in media without pyruvate (Figure 10f). Furthermore, over six days, we observed that the high heteroplasmy cells are unable to proliferate. Based on this, we concluded that the high heteroplasmic cells are severely dysfunctional.

The ETC is also a central metabolic process, so we next investigated global metabolic changes between the two populations. We observed that each TCA metabolite that was detected is reduced in the high heteroplasmy population in comparison to the low heteroplasmy population suggesting that the overall TCA activity is decreased (Figure 10g left). Aligned with this result, many glycolytic metabolites are upregulated (Figure 10g right). We concluded that the high heteroplasmy cells have severe ETC dysfunction and major global metabolic alterations. These experiments conclusively show that these populations are adequate models to study the effects of heteroplasmy LOF mutations.

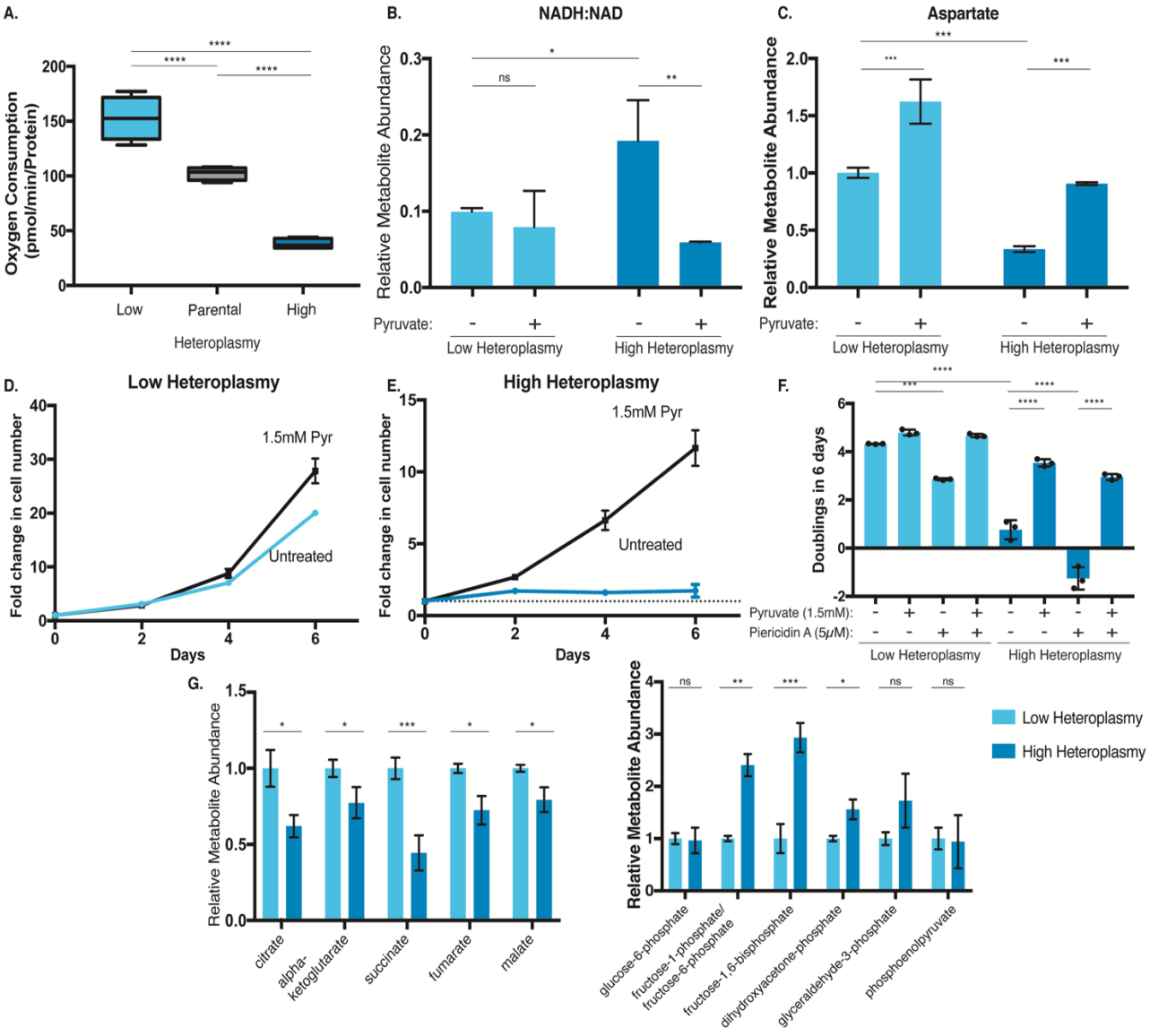

Figure 10: High ND4 11866 C insertion heteroplasmic mutations results in severe ETC dysfunction

a, Basal oxygen consumption normalized by protein concentration. b, NADH/NAD$^+$ ratio of corresponding FTC133 population. c) Aspartate relative metabolic abundance of corresponding FTC133 population. d,e Pyruvate-free media prevents high heteroplasmy cells from growing and it is rescued with 1.5mM pyruvate supplementation, whereas low heteroplasmy cells are more resistant. f, Piericidin A and without pyruvate synergistically sensitizes the corresponding cell lines. g, The relative glycolytic and TCA metabolite abundance in untreated low and high heteroplasmy cells. Mean ± SD., n=3 biologically independent samples and a one-way analysis of variance (ANOVA) followed by Tukey's multiple comparisons test was used for statistical analysis. ***$P < 0.001$, **$P < 0.01$, *$P \leq 0.05$, NS > 0.05.

High heteroplasmic cells are maintained in the parental cell line

The severe dysfunction can be observed through the decreased proliferation rate of the high heteroplasmy cells in comparison to the low and parental heteroplasmy population and cannot be explained by extracellular pyruvate abundance (Figure 11a,b). The observation may be expected based on the ETC dysfunction and it brings up another question of why these cells are not outcompeted by the low heteroplasmy cells (Gallaher et al., 2019). Intuitively, if there is a difference in proliferation rate, eventually the more proliferative cells will outcome the slower cells. This possibility is more realistic considering that the FTC133 cell line has been in culture for many passages before we began to use them. In order to address the proliferation differences, we created a proliferation model to determine at how many doublings the low heteroplasmy cells would outcome the high heteroplasmy cells. More formally speaking, our null-hypothesis is that the cells grow independently, while our alternative-hypothesis is that cell growth is dependent on other cells in the population.

The three variables that are present in our model are the 1) number of cells that are randomly chosen to make it the next generation, 2) proliferation fold-difference to be considered outcompeted, 3) growth rate, and 4) initial starting number of cells. The setting that the model is based off of is a standard tissue culture condition. Cell lines are grown in a tissue culture plate and when the cells grow to be confluent they are split at some fixed number or percentage into another plate. Based on the recommended ATCC tissue culture practice for FTC133, we randomly split back 500K cells after they reached a confluent 8.8M cell density (1). We next assumed that a 100 fold cell

proliferation change between the low and high heteroplasmy would be sufficient to determine that low heteroplasmy cells have outcompeted the high heteroplasmy cells because genetic drift would come into play (2). The growth rate was determined based on Figure 11a (3). Finally, the initial cell counts for each population had a range of possibilities to be considered. In an attempt to reduce the number of possible initial cell counts, we limited the model to only consider the lowest and highest cell count for the high heteroplasmy cells. The model is created to determine the number of low heteroplasmy cell doublings it will take for the high heteroplasmy cells to be outcompeted, thus a low initial high heteroplasmy cell count will result in the fewest doublings and a high initial cell count will lead to the most doublings. We determined the high heteroplasmy initial cell count as a percentage of 500K. We calculated that the

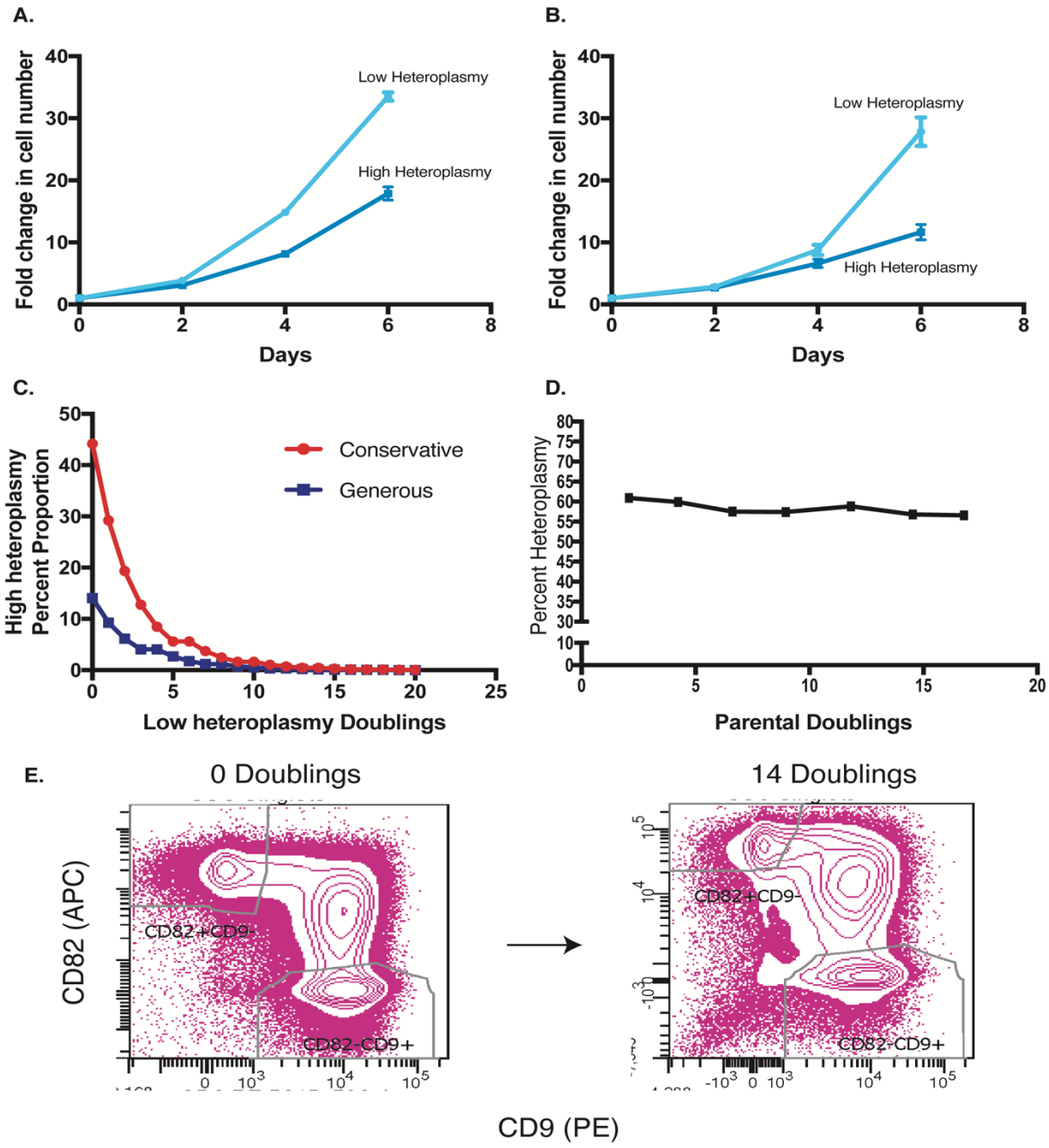

Figure 11: High heteroplasmy cells are significantly maintained in the parental cells

a, Cell proliferation in the ATCC recommended DMEM/F12 media in the corresponding heteroplasmic population. b, Same as a, but in both in the 1.5mM pyruvate. c, Projected number of doublings until high heteroplasmic cells becomes one percent of the population. The “conservative” red line refers to high heteroplasmy percent population based on FACS and the “generous” blue line refers to scRNAseq proportion. d, Heteroplasmy percent as a function of FTC133 parental doublings. e, CD82, and CD9 fluorescence on the parental FTC133.

percentage of cluster 4 in the scRNAseq resulted in the lowest cell count, while the ratio of high heteroplasmy cells compared to low heteroplasmy cells in the FACS analysis resulted in the highest cell count. Thus, we calculated that it would take between 9-12 low heteroplasmy cell doublings for there to be a 100x cell proliferation difference (Figure 11c).

We validated this model by tracking the presence of the high heteroplasmy cells in the parental cell line after 14 doublings, two doublings more than the most conservative estimate (Figure 11d). We reasoned that if the parental heteroplasmy level and the clustering by FACS are present after 12 doublings then we can reject the null-hypothesis and consider that the cells in this population are dependent on the other cells. At 14 doublings, we observed that both the parental heteroplasmy level and the high heteroplasmy cluster within the parental FTC133 cell line is maintained further suggesting that the high heteroplasmy cells are kept within the population (Figure 11e).

High heteroplasmic cells display a more invasive phenotype

The question still remains why do these cells accumulate dysfunctional mutations at such a high mutation burden. It is reasonable to assume that overtime any cells that are not beneficial to the overall cell population would be negatively selected. Due to the likelihood that the high heteroplasmy cells are somehow maintained in cell culture, we hypothesized that these cells confer some advantage in the context of cancer progression. The FTC133 is a metastatic follicular thyroid cancer that is homozygous for six oncogenes (*FLCN*, *MSH6*, *NF1*, *PTEN*, *TERT*, and *TP53*) (Corver et al., 2018). Due to the presumed homogeneity of the FTC133 cell line, the most parsimonious

explanation for the occurrence of the ND4 mutation is that the heteroplasmic range began to diversify after the oncogenic mutations. It is difficult to predict the order of the original somatic mtDNA or oncogene mutation, but since every cell is believed to have these oncogenic mutations, the change in heteroplasmy most likely occurred after tumorigenesis which raises the possibility that the increase in ND4 11866 C insertion mutation is associated with a tumorigenic advantage.

In order to shine light onto the potential functional changes that are present in the high heteroplasmy cells, we performed RNAseq on the low and high heteroplasmic populations. We then looked for the expression differences between the high and low heteroplasmy cells and found that a few notable gene signatures calculated from the GSEA enrichment analysis (Figure 12a). Using the Hallmarks of Cancer gene signatures, we find that the oxidative phosphorylation gene signature is enriched in the low heteroplasmy cells. This serves as a proof of concept since we have been able to experimentally validate the decrease in ETC function.

On the other end of the spectrum, we can see that among the top gene signatures are interferon alpha and gamma response and epithelial mesenchymal transition (EMT). In order to look for overlapping gene signatures, we also conducted Gene Ontology (GO) enrichment and found EMT and RNA transcriptional signatures. All of these signatures have been associated with mtDNA mutations or ETC dysfunction, therefore we considered that they may be existing within these populations (Gaude et al., 2018; Ishikawa et al., 2008; Münch et al., 2016; Quirós et al., 2017; West and Shadel, 2017). We first explored whether these mtDNA mutations associated with an

increase in interferon alpha or gamma response. The work from the Shadel lab has shown that mtDNA dysfunction leads to the mtDNA leakage into the cytosol, which activates the innate immune system through IRF3 phosphorylation. We check for the phosphorylation of IRF3 and did not find that there was any activation (data not shown). We next turned out attention towards the changes in RNA transcriptional signatures. Mitochondrial uncoupled protein response has been shown to be activated under mitochondrial stress in *Caenorhabditis elegans* through the activation of HSPD1 and HSPE1, but we do not see that these genes were differentially regulated between our populations (Münch et al., 2016). Lastly, we veered our focus towards the only signature that was shared between the GSEA and GO analysis, EMT.

The EMT is a process that allows epithelial cells that are normally interacting with the basement membrane undergo several biochemical changes that enables it to become a mesenchymal cell. The phenotypes of mesenchymal cells are enhanced migratory capacity, invasiveness, elevated resistance to apoptosis, increased components of the extracellular matrix, and a spindly morphology (Kalluri and Weinberg, 2009). Furthermore, there have been studies showing an association between ETC dysfunction and mtDNA mutations with metastatic potential (Beadnell et al., 2018; Guerra et al., 2017; Kenny et al., 2017; Vivian et al., 2017; Yuan et al., 2015). Observing the low, parental, and high heteroplasmy cells we see that there is a striking morphological difference among these three groups. The high heteroplasmy cells can be described as being spindly, while the low heteroplasmy cells are rounder (Figure 12b). This potentiates the possibility that the high heteroplasmic cells may be more

invasive compared to the low heteroplasmic cells. Because the FTC133 cell is derived from a metastatic nodule, we originally hypothesized that there would not be any differences between the two group. In order to test this, we assessed the cells' metastatic potential by determining the motility and invasion through the matrigel separated transwell invasion assay. During EMT, epithelial cells lose their junctions, reorganize their cytoskeleton, and change their cell shape, which increases the motility of individual cells (Lamouille et al., 2014). Consistent with this, we observed that the high heteroplasmy cells have an increased persistent motility compared to both the parental and low heteroplasmy cells (Figure 12c). This is also further supported by the transwell assay. The high heteroplasmy cells are able to migrate through the transwell more easily than the low heteroplasmy cells after 18 hours (Figure 12d top). Although visually apparent, we quantified this changed by counting the cell nuclei and found that there is a significant difference between the two groups (Figure 12d bottom). The evidence from both of these experiments prove that the high heteroplasmy cells have an increased invasive phenotype. Interesting to note is the parental population appears to be the most invasive suggesting that the low and high heteroplasmic cells being together have a higher metastatic potential than when apart.

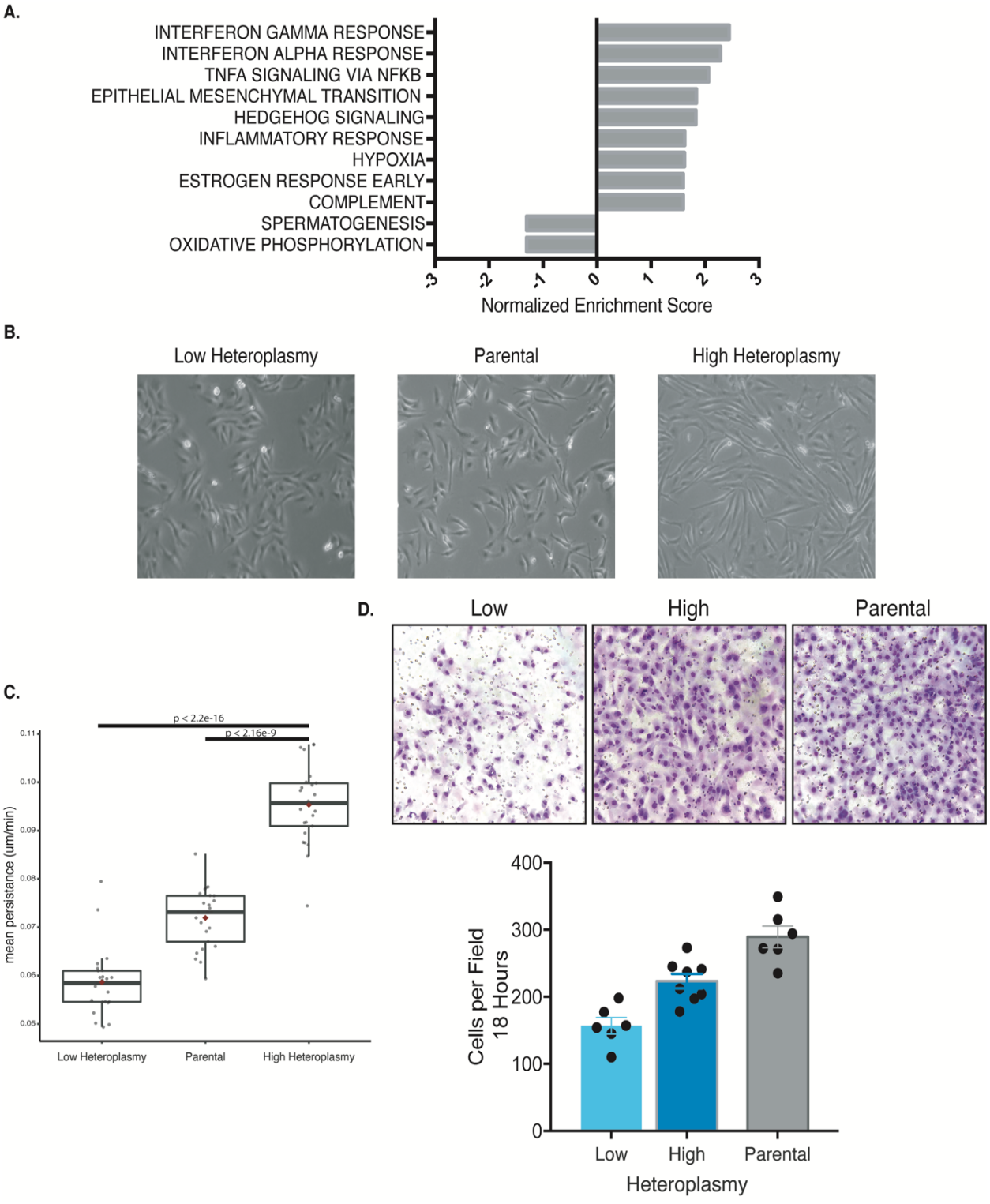
A.
INTERFERON GAMMA RESPONSE
INTERFERON ALPHA RESPONSE
TNFA SIGNALING VIA NFKB
EPITHELIAL MESENCHYMAL TRANSITION
HEDGEHOG SIGNALING
INFLAMMATORY RESPONSE
HYPOXIA
ESTROGEN RESPONSE EARLY
COMPLEMENT
SPERMATOGENESIS
OXIDATIVE PHOSPHORYLATION
-3
-2
-1
0
1
2
3
Normalized Enrichment Score
B.
Low Heteroplasmy
Parental
High Heteroplasmy
C.
p < 2.2e-16
p < 2.16e-9
mean persistance (um/min)
0.11
0.10
0.09
0.08
0.07
0.06
0.05
Low Heteroplasmy
Parental
High Heteroplasmy
D.
Low
High
Parental
Cells per Field
18 Hours
400
300
200
100
0
Low
High
Parental
Heteroplasmy

Figure 12: High heteroplasmy cells have higher invasive potential than low heteroplasmy cells
a, GSEA analysis using the Hallmarks of cancer gene signatures. b, Untreated images of the corresponding heteroplasmic cell population. c, Average persistence motility. Images were taking every eight minutes for 16-hours and persistence was determined by the final distance traveled at 16-hours. d, Representative images of transwell invasion (top) and nuclei count of top image (bottom).

The alpha-ketoglutarate and succinate ratio may be the mechanistic link between the low and high heteroplasmy cells

An outstanding question that exists between these two populations is the mechanistic link that allows one group of cells to become the other and vice-versa. There are many potential mechanistic links that relates to cancer and connects to mtDNA mutations. We limited our search towards changes that are associated to ETC dysfunction because of the nature of the ND4 mutation that is present in FTC133. The changes caused by ETC that we considered were reactive-oxygen species (ROS) and metabolite accumulation. In order to assess their effects on the cell populations we focused their ability to alter the cellular morphology. There are other things that we could assess such as proliferation, motility, invasion, transcription, heteroplasmy or metabolite accumulation or any other phenotypic difference that we have found between the low and high heteroplasmy cells. Although we do consider all of these potential differences, we prioritized morphology because it was the largest difference that we observed and may be an amalgamation of all the phenotypic changes that are caused by the ETC dysfunction.

The ETC dysfunction results in an accumulation of oxygen that may lead to a buildup of superoxide anion and hydroxyl radicals (Bhatti et al., 2017). We observed a modest difference in ROS by CellROX and MitoSox with an increase in ROS in the high heteroplasmy cells and could fit within the context of the large number of transcriptomic changes (Figure 13a) (Chandel et al., 1998). In order to reverse these effects, we added a variety of anti-oxidants (MitoQ, Trolox, Vitamin E, N-acetyl cysteine, and glutathione) over seven days to determine if there are any morphological changes. We treated the cells daily and noticed that there were no clear changes in morphology (Figure 13b). Out of the antioxidants, there were differences in proliferation rates under MitoQ treatment, which suggested that there may be some underlying effects. Although, there were not any changes in morphology we believed that the reduction of ROS by MitoQ may affect the cells' invasive potential (Figure 13c). The addition of MitoQ did not reverse any of the invasive potential, suggesting the difference in ROS may not be the primary driver for the transition between low and high heteroplasmy cells. This may suggest that the differences in ROS between the population may not be large enough.

We, next, turned our attentions towards metabolic changes. TCA intermediates, such as alpha-ketoglutarate (αKG), have been shown to affect chromatin modification (Baksh et al., 2020; Carey et al., 2015; Martínez-Reyes et al., 2020; Morris et al., 2019). αKG is required TETs and H3K27me3 demethylases JMJD3/UTX as cosubstrate raising the possibility that metabolites function in cell fate change (Baksh and Finley, 2020). The chromatin modification fits within the context of these populations because of the large number of transcriptomic changes seen when conducting the differential

gene expression. Looking back at the metabolomics data, we see that there is an approximate two-fold aKG:succinate ratio difference that is necessary for the effect of αKG on chromatin modification (Figure 13d). We tested this idea by adding dimethyl-αKG (DM-αKG) to the low heteroplasmy cells and dimethyl-succinate (DM-Succ) to the high heteroplasmy cells with the hopes that the morphology would flip. As a control for the ester groups, we also added DM-αKG on the high heteroplasmy cells and DM-Succ

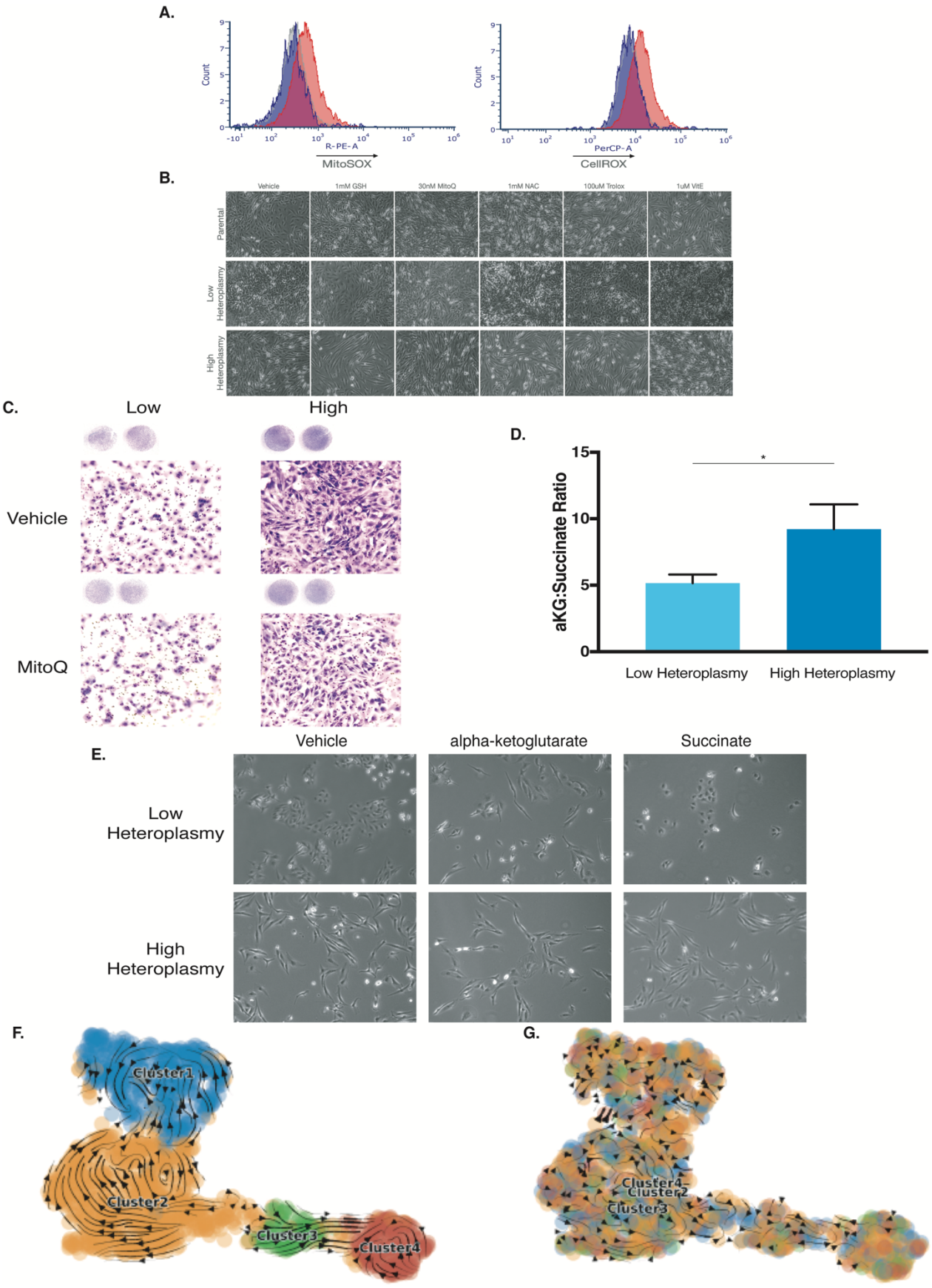
A.
Count
R-PE-A
MitoSOX
Count
PerCP-A
CellROX
B.
Vehicle
1mM GSH
30nM MitoQ
1mM NAC
100uM Trolox
1uM VitE
Parental
Low Heteroplasmy
High Heteroplasmy
C.
Low
High
Vehicle
MitoQ
D.
aKG:Succinate Ratio
15
10
5
0
*
Low Heteroplasmy
High Heteroplasmy
E.
Vehicle
alpha-ketoglutarate
Succinate
Low Heteroplasmy
High Heteroplasmy
F.
Cluster1
Cluster2
Cluster3
Cluster4
G.
Cluster4
Cluster2
Cluster3

Figure 13: alpha-ketoglutarate to succinate ratio may be the mechanistic link between low and high heteroplasmy cells

a, ROS distribution by corresponding treatment using FACS. Using low and high heteroplasmic population. b, Images of cells under various anti-oxidant conditions after seven-day days. c, Representative images of transwell invasion with and without MitoQ (30nM). d, alpha-ketoglutarate to succinate ratio of sorted FTC133 population. e, Images of each population with or with addition of di-methyl alpha-ketoglutarate or dimethyl-succinate. f, RNA velocity plot showing the cell-to-cell transition based on the UMAP clustering from figure 8. The arrows represent the directionality of the cellular transition. The tail of the arrow is the progenitor cell that potentially transitions into the nearby cell directed by the arrow head. g, The RNA velocity of the single-cell projected coordinates from the UMAP space from figure 8 was randomized. Mean ± SD., n=3 biologically independent samples and a one-way analysis of variance (ANOVA) followed by Tukey's multiple comparisons test was used for statistical analysis. *$P \leq 0.05$

on the low heteroplasmy cells. Here, we can see that there is a morphological change in the low heteroplasmy cell and they have adopted a spindlier phenotype, while the DMSO and DM-Succ control have maintained a similar shape (Figure 13e). On the contrary, the high heteroplasmy cells are not significantly altered. This may suggest that these cells have entered a fixed state that cannot be reverted. Lastly, knockdown of alpha-ketoglutarate dehydrogenase (increase the αKG:succinate ratio) and knockdown of SDH (decrease the αKG:succinate ratio) would be needed to further support these findings. Collectively, these results support further experimentation to determine if the low heteroplasmy cells are functionally similar to the high heteroplasmy cells.

In addition to exploring the potential mechanism that may lead to the differences between the low and high heteroplasmy states, we wanted to explore the possibility that a potential mechanistic perturbation would result in a transcriptional change. When we explored the possibility that changes in heteroplasmy levels may result in phenotypic alternations, we first observed the striking discovery that they correlate to a change in

the whole-cell transcriptome (Figure 8a). To be consistent with this observation, we wanted to determine if there was a cellular trajectory that resulted in low heteroplasmy cells becoming high heteroplasmy cells. More explicitly stated, we wanted to explore the possibility that there is a single cell trajectory that results in the transcriptome of the low heteroplasmy cells becoming more similar to the transcriptome of the high heteroplasmy cells or vice-versa. We hypothesized that if there was a trajectory between the low and high heteroplasmy cells, then there is a possibility that one heteroplasmy state may lead to the development of the other. To address this hypothesis, we conducted an RNA velocity experiment to predict the cellular trajectory of all individual cells in the FTC133 scRNAseq data. RNA velocity determines the cellular transcriptomic trajectory by comparing the ratio of unspliced to spliced transcript expression. The idea being that if a neighboring cell has an increase in this ratio then it is possible that the cells may transition between each other. Inherent to this analysis is a structure that positions these single cells based on their whole-cell expression similarity. The positioning provides the necessary information in determining which cells neighbor each other. We decided to provide this inherent structure using the scRNAseq UMAP cluster from figure 8a. RNA velocity analysis revealed that there are many cell-to-cell transition trajectories that exists within the individual FTC133 cells (Figure 13f). The arrows in this map represent the predicted cell-to-cell trajectories. The majority of the predicted cellular trajectories are seen to be either cycling within each cluster, between cluster 1 and cluster 2, or between cluster 3 and cluster 4. These trajectories are based on the cell positioning, and the randomization of cell coordinates on the UMAP space removes the

cell trajectory (Figure 13g). The cyclic trajectory found within each cluster is reminiscent of the cell cycle. In a previous analysis, we have shown that distinct cell cycle phases are present within each cluster suggesting that the cluster localized cyclic trajectory is likely explained by distinct expression profiles of each cell cycle phase (Figure 8f). Furthermore, we termed the arrows going between cluster 1 and cluster 2 or between cluster 3 and cluster 4 to be intraheteroplasmic cellular transitions. The term highlights that there is an expression fluidity that exists within these two groups, further highlighting the differences between the heteroplasmic states. Although, we do not observe a strong transition between cluster 2 and cluster 3 cells, there still exists a significant possibility that low heteroplasmic cells can transition into high heteroplasmic cells or vice-versa. The conclusive experiment to show that low heteroplasmy cells are able to become high heteroplasmy cells in a manner that reflects the RNA velocity analysis remains to be shown.

Discussion

The mtDNA field has been of interest for many decades due to their role in population genetics, disease, cancer, and metabolism. Despite this interest, the progress on the mtDNA mutations field is stymied by lack of methods to study them. Progress has certainly been made using cybrid-like models both in cell lines and in mice, but they come with additional biases. We feel like we have been able to contribute to these efforts by providing an alternative scRNAseq and FACS-based method for isolating cells with different levels of heteroplasmy. This method proved to be effective on many fronts both technically and biologically. The entire turnaround process can be

done within three months, which is approximately the length of time needed to create stable heteroplasmy populations, at least using FCCP and oligomycin. The sorting method on FTC133 have isolated two populations that are basically, but not technically, two distinct populations. These populations are partially consistent with what is discussed in the literature about mtDNA. This consistency makes FTC133 a useful model to be used for many subfields within the larger mtDNA field. The αKG:succinate ratio findings are personally intriguing because it postulates the idea that cells take on these mtDNA mutations in order to permanently elicit ETC dysfunction. As previously mentioned, genetic and functional experimentation is required to solidify the connection, but the possibility certainly highlights the dynamic roles that the mitochondria may play. These roles may also extend into determining the cellular trajectory or development over time under various cellular stimuli. These experiments would test the hypothesis that the low or high heteroplasmic cells may transition to the other if present under the correct stimuli, which would provide additional evidence that the mitochondria play an adaptive role under external stimuli by altering the cellular transcriptome through heteroplasmic mutations differences.

Furthermore, the effect of these populations has not been studied *in vivo* and additional experimentation is needed to show whether heteroplasmic mutations have clinical significance. If these results hold true *in vivo*, then it will increase the potential for ETC targeting therapeutics or targeting key TCA cycle components to reduce the metastatic potential of cells with LOF mtDNA mutations. In conclusion, we were the first to show a new method for isolating cells of high and low heteroplasmy that have

differences in ETC function and invasive potential. This sparks the possibility that other cell lines may benefit from this scRNAseq and FACS-based method.

Materials and Methods

Single-cell library preparation

Each respective cell line was counted and diluted in their respective media. The cells were then added into the Single Cell Master Mix (10x Genomics) and the cellular suspensions were loaded on a Chromium Controller targeting between 2,500-10,000 cells to generate single-cell 3' RNA-seq libraries. Single cell 3' RNA-seq libraries were generated follow the manufacturer's instructions (10x Genomics Chromium Single Cell 3' Reagent Kit User Guide v2 Chemistry).

Next-generation sequencing of single-cell libraries

After processing through 10x, the transcriptomic matrix was built using SEQC following their standard protocol (Azizi et al., 2018).

Single-cell RNAseq demultiplexing

The generated bam file has the following header: @<CELLBARCODE>:<UMI>, which was used to demultiplex each aligned read into individual cell bam files (**TABLE)**. Samtools was used to convert the bam file into a text file. The text file was then sorted into individual single cell files based on 100% barcode identity match. Lastly, the text file was reconverted into a bam file for downstream analyses using the Samtools -b command.

Single-cell RNA sequencing and quality control

Before proceeding with any further analyses, we filtered out any cells with too low coverage. The 10x single-cell sequencing process can be quite harsh resulting in many cells being lysed while in the droplet. This reality will result in both low cell coverage and a high percentage of mitochondrial genes. The cell coverage depends on each cell line, but can be quite clear when creating a histogram based on the log10 of the count sum for each cell. FTC133 and KHM_5M cells below 4.2 and 4.5, respectively were removed. In both cell lines, cellular integrity under the 10x process was further accessed by creating a mitochondrial read percentage threshold of 20%, which assumed that cells with greater than 20% were likely lysed during the process.

Single-cell RNA sequencing analysis

The normalization and processes were all done using the default Seurat pipeline (https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html). The main difference in our analysis is that we regressed the effect of the library size (nCount_RNA) using the "ScaleData" function. The library size correlation was slight, but the regression was done to ensure that these effects did not obfuscate the gene expression markers. The clustering and UMAP projection were created on Seurat by using the RunUMAP function and only including the dimensions one through ten. The following quality control analysis (cell cycle, library depth, mutation coverage) were created using the same UMAP coordinates and adjusting the color of each point.

RNA velocity

The raw FTC133 fastq were reran using Cell Ranger (https://support.10xgenomics.com/single-cell-gene-

expression/software/pipelines/latest/what-is-cell-ranger) in order to create the necessary output files that could be used by RNAVelocity. The following RNAVelocity procedure was completed using their default parameters found on their cite (La Manno et al., 2018). The UMAP coordinates from Seurat was used to read into RNAVelocity so that the trajectory would be mapped onto the same space. The loom file was created using the scVelo default parameters.

Antibodies and Reagents

The following antibodies were purchased from the following vendors. Cell Signaling Technologies: Vimentin D21H3 XP® Rabbit mAb (5741S), Phospho-S6 Ribosomal Protein Ser240/244 (2215S), S6 Ribosomal Protein 5G10 Rabbit mAb – (2217S), p70 S6 Kinase 49D7 Rabbit mAb (2708S), Phospho-p70 S6 Kinase (Thr389) (108D2) Rabbit mAb (9234S), Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) (D13.14.4E) XP® Rabbit mAb (4370S), Akt pan C67E7 Rabbit mAb (4691S), Phospho-Akt Ser473 D9E XP® Rabbit mAb (4060S). GeneTex GAPDH (GT239). Santa Cruz Biotechnology ERK 1/2 C-9 (sc-514302). Biolegend: APC anti-human CD82 Antibody (342114) and FITC anti-human CD49b (359306). Miltenyi Biotec: CD9 (130-103-988). The following secondary antibodies were used: HRP-conjugated horse anti-mouse IgG (Cell Signaling 7076), HRP-conjugated and goat anti-rabbit IgG (Cell Signaling 7074) The following reagents were purchased from the following vendors: Pierce BCA Protein Assay Kit, (Thermo Fisher), Seahorse XF Cell Mito Stress Test Kit (Agilent), Seahorse XF RPMI medium, pH 7.4 (Agilent), D-Glucose (VWR), FBS (Sigma), Dialyzed FBS (GIBCO), Fetal Bovine Lipoprotein Deficient Serum (Kalen Biomedical), HPLC Grade

Water (Fisher Scientific), HPLC Grade Methanol (Fisher Scientific), Sodium Pyruvate (Fisher Bioreagents), Mitoquinone mesylate (AbMole M9068), Piericidin A (Enzo Life Sciences - ALX-380-235-M002)

Cell Lines

The FTC133 and KHM5H cell lines were kindly provides by Dr. James Fagin (Memorial Sloan Kettering) and be purchased from ATCC. Cell lines were verified to be free of mycoplasma contamination and the identities of all were authenticated by Short-tandem Repeat profiling (STR).

Cell Culture Conditions

The FTC133 cell line was cultured in DMEM/F12 (1:1) (Gibco 11330-032) containing 2mM L-glutamine and 15mM HEPES, 10% feta bovine serum (Sigma), 1% penicillin and streptomycin (Invitrogen). For pyruvate supplementation rescue experiments, cells were grown in DMEM without pyruvate, 10% feta bovine serum (Sigma), 1% penicillin and streptomycin (Invitrogen), and supplemented with 1.5mM pyruvate (Fisher Bioreagents) or without. MitoQ supplementation is at 100uM in the previously mentioned DMEM/F12 conditions. In the aspartate rescue experiment, the FTC133 cells were cultured in RPMI-1640 (Gibco) containing 2mM glutamine, 10% fetal bovine. Serum (Sigma), 1% penicillin and streptomycin (Invitrogen) with or without the addition of 20mM aspartate (Sigma 07125-25G).

Primers

Primer Name	Sequence	Purpose
Human MT-ND4-F	CCTTTTCCTCCGACCCCCTAACA	PCR

Human MT-ND4-R	TAGCAGTTCTTGTGAGCTTTCTCGGT	PCR
Human MT-ND4-S	GGGCTTACATCCTCATTACTATT	PCR for Sanger Sequencing

Generation of various heteroplasmic populations

The generation of the low (~49% heteroplasmy), parental (~57% heteroplasmy), and high (~74% heteroplasmy) heteroplasmy levels was done using a fluorescence-activated Cell Sorting (FACS) based method. The parental FTC133 cell line was stained with three antibodies: CD82 (APC) (1:200), CD49b (FITC) (1:200), and CD9 (PE) (1:100) and isolated based on fluorescence. The cells that are APC^+, $FITC^+$, PE^-, and $DAPI^-$ were sorted for the high heteroplasmy population, while the cells that are APC^-, $FITC^-$, PE^+ and $DAPI^-$ were sorted for the low heteroplasmy population. The FACS control cells were sorted for $DAPI^-$ and were done after the low and high heteroplasmy populations were completed. The three populations were then culture in a 1:1 mixture of DMEM/F12 and conditioned media (from the parental FTC133 culture).

Generation of overexpression cell lines

The GFP-Luc and NDi1 overexpression vector can be found on the Addgene links below. The information on backbone, size, and selectable markers can also be found on the corresponding Addgene link.

Construct	Addgene link
NDi1	(Birsoy et al., 2014)
GFP-Luc	(Garcia-Bermudez et al., 2018)

Cell Proliferation Assays

The proliferation assays were seeded at 10,000 cells per well and in triplicate. They were seeded in DMEM/F12 overnight and the next day was prepared further for the proliferation assay. The DMEM/F12 was removed and undergone a 2x PBS wash in order to remove any nutrients from the DMEM/F12. Multiple plates were prepared for each timepoint with the same condition.

Transwell Invasion Assay

Cells were collected in the DMEM/F12 media without serum. Cells were counted using the Coulter Counter and 25,000 cells were seeded on 8μm pore size cell culture inserts for 24-well plates (Corning). Prior to cell seeding, the cell culture inserts were coated with growth factor reduced Matrigel (GIBCO) diluted in 1:100 PBS that was incubated at room temperature for two hours and removed. The complete DMEM/F12 media, including serum, was used in the lower chamber and the assays was terminated after 18 hours. Following incubation, the uninvaded cells were removed and the invaded cells were fixed, stained using the Hema 3 Manual Staining Stat Pack according to manufacturer's guidelines (Thermo Fisher Scientific), and placed on glass slides with mounting media (Permount).

Random cell migration assay

Each cell population was plated at 75,000 per well. Cells were imaged at 37^{o}C with 5% CO_2 with a 10x/0.45 NA objective on a Zeiss AxioObserver Z1 for 16 hours with 8-minute intervals and at 25 different positions in each well. The cell tracking of single

cells was performed using the TrackMate plugin in FIJI, in which cells are tracked based on cell contrast over time. Cells that experience cell division, cell death, a collision event, or migrated out of the field of view were excluded. To obtain velocity and persistence values, raw tracking data were analyzed using MATLAB.

Metabolite Profiling

For bulk polar metabolite profiling, 100K low and high heteroplasmy cells were seeding triplicate in 6 well plates and grown for 48hrs under indicated treatments. The cells were initially seeded the DMEM/F12 media and 24 hours before metabolite harvesting the cells undergone a 2x PBS wash followed by the addition of the RPMI 1640 plus or minus 1.5mM pyruvate. On the day of harvest, cells were washed 2x in cold 0.9% NaCl, extracted in 600ul 80% methanol containing 15N and 13C fully- labeled amino acid internal standards (MSK-A2-1.2, Cambridge Isotope Laboratories, Inc). Extracts were vortexed for 10 minutes, centrifuged at 20,000g for 10min to remove insoluble cell debris, nitrogen-dried, and stored at -80^{o}C until liquid chromatography-mass spectrometry (LC-MS). LC-MS was performed as previously described (Garcia-Bermudez et al., 2018) and relative quantification of metabolite abundances was performed using XCalibur QualBrowser 2.2 and Skyline Targeted Mass Spec Environment (MacCoss Lab) using a 5 ppm mass tolerance and a pooled-library of metabolite standards to confirm metabolite identity. Metabolite levels were normalized by BCA protein quantification for each condition.

Immunoblotting

125K cells were seeding in a 6 well plate for 24hrs before collection. Each condition was washed twice in cold PBS and lysed in a buffer containing 10 mM Tris-Cl pH 7.5, 150 NaCl, 1 mM EDTA, 1% Triton X-100, 2% SDS, 0.1% CHAPS, protease inhibitors (Roche), and PhosphoStop PHOSSTOP (Roche 04906837001). Lysates were sonicated, centrifuged at 20,000g, and total protein quantified using BCA Protein Assay Kit (Thermo Fisher). Supernatants were run on 10-20% Tri-Glycine gel with PageRuler ladder and analyzed via immunoblotting. Blots were developed using the Chemiluminescent detection and film exposure.

Oxygen consumption measurements

The basal oxygen consumption of intact FTC133 sorted population cells was measured using an XF96 Extracellular Flux Analyzer (Seahorse Bioscience). 40,000 cells were plated 24hrs before the assay using the previously mentioned DMEM/F12 media condition. Basal oxygen consumption measurements were normalized by protein levels after the assay was completed.

RNA Whole Exome Sequencing

For bulk polar RNAseq, 100K low and high heteroplasmy cells were seeding triplicate in 6 well plates and grown for 48hrs under indicated treatments. Cells were washed 2x with cold PBS and RNA was harvested using RNAeasy mini kit (Qiagen) according to manufacturer's protocol. 2.5ug of RNA from each well was collected and shipped to Genewiz. The fastq processing was completed through Genewiz's standard human whole exome sequencing protocol. STAR and RESM, described in the software

and algorithm section, were used to determine the normalized transcript per million values. All reads were aligned to the hg38 human genome assembly.

Software and Algorithms

Name	Company	Link
Skyline Daily	MacCoss Lab	https://skyline.ms/project/home/software/Skyline/begin.view
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR
RESM	Li and Dwey, 2011	https://github.com/deweylab/RSEM
GSEA	Subramanian et al., 2005	https://www.gsea-msigdb.org/gsea/index.jsp
Prism 7	GraphPad	https://www.graphpad.com/scientific-software/prism/
Seurat	Butler et al. 2018	https://satijalab.org/seurat/
Monocle 3	Cao J, et al 2019	https://cole-trapnell-lab.github.io/monocle3/
R 3.6.3	R Core Team 2020	https://www.R-project.org/
RStudio 1.2.5042		https://rstudio.com/products/rstudio/download/

Microsoft Excel 16.16.26	Microsoft	https://www.microsoft.com/en-us/microsoft-365/excel
RNA velocity	La Manno, et al. 2018	https://github.com/velocyto-team/velocyto.R
Gene Ontology	Huang, Mi H, et al. 2019	http://geneontology.org/docs/downloads/

Chapter 3: Future directions and Perspectives

Improving the co-essentiality networks

In chapter 1, we found that *C12orf49* functions to regulated the SREBP1 pathway, while *TMEM41a* is associated with saturated lipid metabolism. There remain many poorly characterized genes such as *VKORC1L1* for fatty acid utilization and *TMEM189* in plasmalogen synthesis. Additional validation of these genes would thereby strengthen the confidence of the already present networks. In this section, I would like to propose additional data sources that could create new networks or refine edges that motivate experimental validation.

One of the simplest alterations that requires the fewest number of changes in the existing pipeline is to create different gene subsets. In this instance we largely chose metabolic genes because of the lab's skillsets and the amount of well-validated pathways that can be easily grouped into a gene network. Perhaps the major advantage is the metabolic gene filtration greatly reduces the number of false positives and simplifies the manual curation of each network through literature research. There, however, is no technical reason to limit the analysis to just metabolic genes. There is a growing number of studies that are conducting CRISPR-Cas9 genetic screens using specific gene sets (Williams et al., 2020). One growing field is the identification of various mitochondrial transporters. One can simultaneously create two gene lists, the first being any genes that are predicted to be transmembrane (putative transporters) and any genes that are found to be in the mitochondria (MitoCarta 2.0). Reutilization of our pipeline, while editing the gene lists, can prove to generate many new hypotheses.

Another idea, that is specific to the work in the Birsoy lab, is to overlay the vast number of genetics screens that we have done onto our published networks. Over the years, the Birsoy Lab has completed many types of genetic screens, like knockout and activating, on many cell lines, and under many conditions. Normally, we only consider the top scoring genes as hits, however combining the screens with the networks may allow us to go deeper down the genetic screen list or reduce noise in our networks. One example is to identify genes that are associated with fatty acid metabolism. We have already completed LOF and activation genetic screens using palmitate and arachidonate treatment. These screens have revealed *CHP1* and *VKORC1L1* as hits which support the idea that the fatty acid network is valid (Zhu et al., 2019). The logical extension can be done using other treatments for each network of interest.

Along the same lines of combing more information is to use additional publicly available datasets. Large-scale RNAseq data, proteomics, conservation, and literature databases can be useful in these efforts to either create more edges between nodes or increase confidence in existing interactions (Chen et al., 2018; Meyers et al., 2017; Pellegrini et al., 1999; Sowa et al., 2009). An example of this is the STRING database, which creates edges between genes that are associated with each other through literature keyword associations. In summary, I believe further improvements on these networks will allow researchers to more quickly narrow down hypotheses that have more support and thus have a higher likelihood of being valid.

Generalizing the FTC133 heteroplasmic population isolation method

The scRNAseq analysis combined with the FACS-based sorting method has proven to be an effective strategy for isolating low and high heteroplasmy FTC133 cells. There are many benefits to this method as we described above, but a major limitation that we seek to overcome is that the current data suggest that this method is limited to the FTC133 cell line. The next step is to look at additional approaches to either determine whether our results are generalizable in other settings or if what we find is an example of a general trend.

The approach that we are currently taking to determine whether our findings are generalizable is to see if our markers or differentially expressed genes are also shared in other systems. A brute force method is to take any of our antibody markers and profile other cells that have the same ND4 mutation or other mtDNA LOF mutations. The idea behind this approach is to determine whether these markers are representative of mtDNA LOF mutations across different cell lines and cancer types. Ideally, we will see similar clustering in the FACS analysis, but the lack of a cluster defined by CD82 and CD9 does not mean that there will not be a change in heteroplasmy after the sort. The simplest approach is to choose cell lines with the same ND4 mutation, with a heteroplasmy around 50%, and across different cancer types. A slightly more complicated approach that may prove to be favorable is to choose cell lines that are around 50% heteroplasmy and transcriptomically similar to FTC133. For example, we can develop a FTC133 gene signature by performing differential gene expression compared to other cell lines. Then, we can rank all cell lines by their FTC133

enrichment score. Similarly, we can develop a gene signature based on our cluster 4 marker genes and perform a similar analysis. The limitation is that the signature may only be present in a subset of cells and is thus drowned out in the mix population RNAseq profiling. Another consideration is the bulk heteroplasmy levels in general may trend alongside the enrichment of cluster 4 markers, which would still present a correlative association between these genes and heteroplasmic mutations. Alternatively, we can do differentially expression analysis between tumor samples with LOF heteroplasmic mutations compared to samples without and identify cell surface proteins that may be enriched.

The other possibility is that the FTC133 scRNAseq results are representative of a general trend. Cluster 4 marker genes are comprised of 26% cell surface proteins, where approximately 7.4% of all human genes encode for cell surface proteins (Bausch-Fluck et al., 2015). This difference suggests that there is a potential enrichment of cell surface proteins in cells with high heteroplasmic LOF mutations. The question that comes to mind is whether there is a general trend that connects LOF heteroplasmic mutations with cell surface proteins. In order to more formally make this connection, I would create a matrix of cancer samples from the TCGA by all known cell surface proteins for each cancer type. Hierarchical clustering may potentially group samples by cell surface proteins. Then, we can determine whether there is a cluster that is defined by high or low heteroplasmic LOF mutations. In the event that there are many clusters, we could consider all clusters with heteroplasmic LOF mutation levels that are likely above the threshold effect. This strategy may potentially result in cell surface proteins

enriched in the low and high heteroplasmic clusters that would then give a pool of antibodies that can be used for the brute forced method mentioned above. If true, these finding could open up a new mitochondrial question asking why LOF mutations in the mtDNA lead to increased cell surface protein expression. This question may provide insights into uncovering the role of heteroplasmic LOF mutations.

Bibliography

Aryaman, J., Johnston, I.G., and Jones, N.S. (2019). Mitochondrial Heterogeneity. Front. Genet. *9*.

Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell *174*, 1293-1308.e36.

Bacman, S.R., Williams, S.L., Pinto, M., Peralta, S., and Moraes, C.T. (2013). Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. Nat. Med. *19*, 1111–1113.

Baksh, S.C., and Finley, L.W.S. (2020). Metabolic Coordination of Cell Fate by α-Ketoglutarate-Dependent Dioxygenases. Trends Cell Biol. *0*.

Baksh, S.C., Todorova, P.K., Gur-Cohen, S., Hurwitz, B., Ge, Y., Novak, J.S.S., Tierney, M.T., dela Cruz-Racelis, J., Fuchs, E., and Finley, L.W.S. (2020). Extracellular serine controls epidermal stem cell fate and tumour initiation. Nat. Cell Biol. *22*, 779–790.

Basu, S., Duren, W., Evans, C.R., Burant, C.F., Michailidis, G., and Karnovsky, A. (2017). Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. Bioinforma. Oxf. Engl. *33*, 1545–1553.

Baughman, J.M., Perocchi, F., Girgis, H.S., Plovanich, M., Belcher-Timme, C.A., Sancak, Y., Bao, X.R., Strittmatter, L., Goldberger, O., Bogorad, R.L., et al. (2011). Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. Nature *476*, 341–345.

Bausch-Fluck, D., Hofmann, A., Bock, T., Frei, A.P., Cerciello, F., Jacobs, A., Moest, H., Omasits, U., Gundry, R.L., Yoon, C., et al. (2015). A Mass Spectrometric-Derived Cell Surface Protein Atlas. PLOS ONE *10*, e0121314.

Bayraktar, E.C., La, K., Karpman, K., Unlu, G., Ozerdem, C., Ritter, D.J., Alwaseem, H., Molina, H., Hoffmann, H.-H., Millner, A., et al. (2020). Metabolic coessentiality mapping identifies C12orf49 as a regulator of SREBP processing and cholesterol metabolism. Nat. Metab. *2*, 487–498.

Beadnell, T.C., Scheid, A.D., Vivian, C.J., and Welch, D.R. (2018). Roles of the mitochondrial genetics in cancer metastasis: Not to be ignored any longer. Cancer Metastasis Rev. *37*, 615–632.

Bertomeu, T., Coulombe-Huntington, J., Chatr-aryamontri, A., Bourdages, K.G., Coyaud, E., Raught, B., Xia, Y., and Tyers, M. (2018). A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural Features and Contextual Diversity of the Human Cell-Essential Proteome. Mol. Cell. Biol. *38*.

Bhatti, J.S., Bhatti, G.K., and Reddy, P.H. (2017). Mitochondrial dysfunction and oxidative stress in metabolic disorders — A step towards mitochondria based therapeutic strategies. Biochim. Biophys. Acta BBA - Mol. Basis Dis. *1863*, 1066–1077.

Birsoy, K., Possemato, R., Lorbeer, F.K., Bayraktar, E.C., Thiru, P., Yucel, B., Wang, T., Chen, W.W., Clish, C.B., and Sabatini, D.M. (2014). Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. Nature *508*, 108–112.

Birsoy, K., Wang, T., Chen, W.W., Freinkman, E., Abu-Remaileh, M., and Sabatini, D.M. (2015). An Essential Role of the Mitochondrial Electron Transport Chain in Cell Proliferation Is to Enable Aspartate Synthesis. Cell *162*, 540–551.

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.

Brandon, M., Baldi, P., and Wallace, D.C. (2006). Mitochondrial mutations in cancer. Oncogene *25*, 4647–4662.

Brown, M.S., and Goldstein, J.L. (1997). The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. Cell *89*, 331–340.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., Natarajan, K.N., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.

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

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Carey, B.W., Finley, L.W.S., Cross, J.R., Allis, C.D., and Thompson, C.B. (2015). Intracellular α-ketoglutarate maintains the pluripotency of embryonic stem cells. Nature *518*, 413–416.

Cavalli, L.R., Varella-Garcia, M., and Liang, B.C. (1997). Diminished tumorigenic phenotype after depletion of mitochondrial DNA. Cell Growth Differ. Mol. Biol. J. Am. Assoc. Cancer Res. *8*, 1189–1198.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chandel, N.S., Maltepe, E., Goldwasser, E., Mathieu, C.E., Simon, M.C., and Schumacker, P.T. (1998). Mitochondrial reactive oxygen species trigger hypoxia-induced transcription. Proc. Natl. Acad. Sci. U. S. A. *95*, 11715–11720.

Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., Kasapi, M., Ferguson, M.L., et al. (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell *173*, 386-399.e12.

Chinnery, P.F., Taylor, G.A., Howell, N., Brown, D.T., Parsons, T.J., and Turnbull, D.M. (2001). Point Mutations of the mtDNA Control Region in Normal and Neurodegenerative Human Brains. Am. J. Hum. Genet. *68*, 529–532.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Corver, W.E., Demmers, J., Oosting, J., Sahraeian, S., Boot, A., Ruano, D., Wezel, T. van, and Morreau, H. (2018). ROS-induced near-homozygous genomes in thyroid cancer. Endocr. Relat. Cancer *25*, 83–97.

Edwards, P.A., Tabor, D., Kast, H.R., and Venkateswaran, A. (2000). Regulation of gene expression by SREBP and SCAP. Biochim. Biophys. Acta *1529*, 103–113.

Ericsson, J., Jackson, S.M., and Edwards, P.A. (1996). Synergistic binding of sterol regulatory element-binding protein and NF-Y to the farnesyl diphosphate synthase promoter is critical for sterol-regulated expression of the gene. J. Biol. Chem. *271*, 24359–24364.

Gallaher, J.A., Brown, J.S., Anderson, A.R.A., Brown, J.S., and Anderson, A.R.A. (2019). The impact of proliferation-migration tradeoffs on phenotypic evolution in cancer. Sci. Rep. *9*, 2425.

Gallego-García, A., Monera-Girona, A.J., Pajares-Martínez, E., Bastida-Martínez, E., Pérez-Castaño, R., Iniesta, A.A., Fontes, M., Padmanabhan, S., and Elías-Arnanz, M. (2019). A bacterial light response reveals an orphan desaturase for human plasmalogen synthesis. Science *366*, 128–132.

Gammage, P.A., Moraes, C.T., and Minczuk, M. (2018). Mitochondrial Genome Engineering: The Revolution May Not Be CRISPR-Ized. Trends Genet. *34*, 101–110.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Garcia-Bermudez, J., Baudrier, L., La, K., Zhu, X.G., Fidelin, J., Sviderskiy, V.O., Papagiannakopoulos, T., Molina, H., Snuderl, M., Lewis, C.A., et al. (2018). Aspartate is a limiting metabolite for cancer cell proliferation under hypoxia and in tumours. Nat. Cell Biol. *20*, 775–781.

Garcia-Bermudez, J., Baudrier, L., Bayraktar, E.C., Shen, Y., La, K., Guarecuco, R., Yucel, B., Fiore, D., Tavora, B., Freinkman, E., et al. (2019). Squalene accumulation in cholesterol auxotrophic lymphomas prevents oxidative cell death. Nature *567*, 118–122.

Gaude, E., Schmidt, C., Gammage, P.A., Dugourd, A., Blacker, T., Chew, S.P., Saez-Rodriguez, J., O'Neill, J.S., Szabadkai, G., Minczuk, M., et al. (2018). NADH Shuttling Couples Cytosolic Reductive Carboxylation of Glutamine with Glycolysis in Cells with Mitochondrial Dysfunction. Mol. Cell *69*, 581-593.e7.

Gorelick, A.N., Kim, M., Chatila, W.K., La, K., Hakimi, A.A., Taylor, B.S., Gammage, P.A., and Reznik, E. (2020). Respiratory complex and tissue lineage drive mutational patterns in the tumor mitochondrial genome. BioRxiv 2020.08.18.256362.

Gorman, G.S., Chinnery, P.F., DiMauro, S., Hirano, M., Koga, Y., McFarland, R., Suomalainen, A., Thorburn, D.R., Zeviani, M., and Turnbull, D.M. (2016). Mitochondrial diseases. Nat. Rev. Dis. Primer *2*, 16080.

Grandhi, S., Bosworth, C., Maddox, W., Sensiba, C., Akhavanfard, S., Ni, Y., and LaFramboise, T. (2017). Heteroplasmic shifts in tumor mitochondrial genomes reveal tissue-specific signals of relaxed and positive selection. Hum. Mol. Genet. *26*, 2912–2922.

Guan, G., Dai, P.H., Osborne, T.F., Kim, J.B., and Shechter, I. (1997). Multiple sequence elements are involved in the transcriptional regulation of the human squalene synthase gene. J. Biol. Chem. *272*, 10295–10302.

Guerra, F., Guaragnella, N., Arbini, A.A., Bucci, C., Giannattasio, S., and Moro, L. (2017). Mitochondrial Dysfunction: A Novel Potential Driver of Epithelial-to-Mesenchymal Transition in Cancer. Front. Oncol. *7*.

Han, Y.H., Kim, S.H., Kim, S.Z., and Park, W.H. (2008). Antimycin A as a mitochondrial electron transport inhibitor prevents the growth of human lung cancer A549 cells. Oncol. Rep. *20*, 689–693.

Harris, M. (1980). Pyruvate blocks expression of sensitivity to antimycin A and chloramphenicol. Somatic Cell Genet. *6*, 699–708.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

Hood, L., and Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. Genome Med. *5*, 79.

Horton, J.D., Goldstein, J.L., and Brown, M.S. (2002). SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. J. Clin. Invest. *109*, 1125–1131.

Hua, X., Nohturfft, A., Goldstein, J.L., and Brown, M.S. (1996). Sterol resistance in CHO cells traced to point mutation in SREBP cleavage-activating protein. Cell *87*, 415–426.

Ishikawa, K., Takenaga, K., Akimoto, M., Koshikawa, N., Yamaguchi, A., Imanishi, H., Nakada, K., Honma, Y., and Hayashi, J.-I. (2008). ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis. Science *320*, 661–664.

Jackson, C.B., Turnbull, D.M., Minczuk, M., and Gammage, P.A. (2020). Therapeutic Manipulation of mtDNA Heteroplasmy: A Shifting Perspective. Trends Mol. Med. *26*, 698–709.

Jain, I.H., Zazzeron, L., Goli, R., Alexa, K., Schatzman-Bone, S., Dhillon, H., Goldberger, O., Peng, J., Shalem, O., Sanjana, N.E., et al. (2016). Hypoxia as a therapy for mitochondrial disease. Science *352*, 54–61.

Kalina, T., Fišer, K., Pérez-Andrés, M., Kuzílková, D., Cuenca, M., Bartol, S.J.W., Blanco, E., Engel, P., and van Zelm, M.C. (2019). CD Maps—Dynamic Profiling of CD1–CD100 Surface Expression on Human Leukocyte and Lymphocyte Subsets. Front. Immunol. *10*.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428.

Kenny, T.C., Hart, P., Ragazzi, M., Sersinghe, M., Chipuk, J., Sagar, M. a. K., Eliceiri, K.W., LaFramboise, T., Grandhi, S., Santos, J., et al. (2017). Selected mitochondrial DNA landscapes activate the SIRT3 axis of the UPR mt to promote metastasis. Oncogene *36*, 4393–4404.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kim, E., Dede, M., Lenoir, W.F., Wang, G., Srinivasan, S., Colic, M., and Hart, T. (2019). A network of human functional gene interactions from knockout fitness screens in cancer cells. Life Sci. Alliance *2*.

Kim, H., Shim, J.E., Shin, J., and Lee, I. (2015). EcoliNet: a database of cofunctional gene network for Escherichia coli. Database J. Biol. Databases Curation *2015*.

King, M.P., and Attardi, G. (1989). Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. Science *246*, 500–503.

Kleinfelter, L.M., Jangra, R.K., Jae, L.T., Herbert, A.S., Mittler, E., Stiles, K.M., Wirchnianski, A.S., Kielian, M., Brummelkamp, T.R., Dye, J.M., et al. (2015). Haploid Genetic Screen Reveals a Profound and Direct Dependence on Cholesterol for Hantavirus Membrane Fusion. MBio *6*, e00801.

Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. *25*, 1860–1872.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst. Biol. *5*, 21.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial–mesenchymal transition. Nat. Rev. Mol. Cell Biol. *15*, 178–196.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Martínez-Reyes, I., Chandel, N.S., and Chandel, N.S. (2020). Mitochondrial TCA cycle metabolites control physiology and disease. Nat. Commun. *11*, 102.

Matsuda, M., Korn, B.S., Hammer, R.E., Moon, Y.A., Komuro, R., Horton, J.D., Goldstein, J.L., Brown, M.S., and Shimomura, I. (2001). SREBP cleavage-activating protein (SCAP) is required for increased lipid synthesis in liver induced by cholesterol deprivation and insulin elevation. Genes Dev. *15*, 1206–1216.

Mazumder, R., and Hastie, T. (2012). Exact Covariance Thresholding into Connected Components for Large-Scale Graphical Lasso. J. Mach. Learn. Res. JMLR *13*, 781–794.

McBride, H.M., Neuspiel, M., and Wasiak, S. (2006). Mitochondria: more than just a powerhouse. Curr. Biol. CB *16*, R551-560.

McDavid, A., Finak, G., Gottardo, R., Finak, G., and Gottardo, R. (2016). The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. Nat. Biotechnol. *34*, 591–593.

McDonald, E.R., de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell *170*, 577-592.e10.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Mok, B.Y., de Moraes, M.H., Zeng, J., Bosch, D.E., Kotrys, A.V., Raguram, A., Hsu, F., Radey, M.C., Peterson, S.B., Mootha, V.K., et al. (2020). A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature *583*, 631–637.

Morais, R., Zinkewich-Péotti, K., Parent, M., Wang, H., Babai, F., and Zollinger, M. (1994). Tumor-forming ability in athymic nude mice of human cell lines devoid of mitochondrial DNA. Cancer Res. *54*, 3889–3896.

Morris, J.P., Yashinskie, J.J., Koche, R., Chandwani, R., Tian, S., Chen, C.-C., Baslan, T., Marinkovic, Z.S., Sánchez-Rivera, F.J., Leach, S.D., et al. (2019). α-Ketoglutarate links p53 to cell fate during tumour suppression. Nature *573*, 595–599.

Münch, C., Harper, J.W., and Harper, J.W. (2016). Mitochondrial unfolded protein response controls matrix pre-RNA processing and translation. Nature *534*, 710–713.

Osuna-Ramos, J.F., Reyes-Ruiz, J.M., and Del Ángel, R.M. (2018). The Role of Host Cholesterol During Flavivirus Infection. Front. Cell. Infect. Microbiol. *8*, 388.

Pan, J., Meyers, R.M., Michel, B.C., Mashtalir, N., Sizemore, A.E., Wells, J.N., Cassel, S.H., Vazquez, F., Weir, B.A., Hahn, W.C., et al. (2018). Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. Cell Syst. *6*, 555-568.e7.

Pavlova, N.N., and Thompson, C.B. (2016). The Emerging Hallmarks of Cancer Metabolism. Cell Metab. *23*, 27–47.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. *96*, 4285–4288.

Picard, M., Zhang, J., Hancock, S., Derbeneva, O., Golhar, R., Golik, P., O'Hearn, S., Levy, S., Potluri, P., Lvova, M., et al. (2014). Progressive increase in mtDNA 3243A>G heteroplasmy causes abrupt transcriptional reprogramming. Proc. Natl. Acad. Sci. U. S. A. *111*, E4033-4042.

Polyak, K., Li, Y., Zhu, H., Lengauer, C., Willson, J.K., Markowitz, S.D., Trush, M.A., Kinzler, K.W., and Vogelstein, B. (1998). Somatic mutations of the mitochondrial genome in human colorectal tumours. Nat. Genet. *20*, 291–293.

Pombo, J.P., and Sanyal, S. (2018). Perturbation of Intracellular Cholesterol and Fatty Acid Homeostasis During Flavivirus Infections. Front. Immunol. *9*, 1276.

Possemato, R., Marks, K.M., Shaul, Y.D., Pacold, M.E., Kim, D., Birsoy, K., Sethumadhavan, S., Woo, H.-K., Jang, H.G., Jha, A.K., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature *476*, 346–350.

Potter, M., Newport, E., and Morten, K.J. (2016). The Warburg effect: 80 years on. Biochem. Soc. Trans. *44*, 1499–1505.

Quirós, P.M., Prado, M.A., Zamboni, N., D'Amico, D., Williams, R.W., Finley, D., Gygi, S.P., and Auwerx, J. (2017). Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. J. Cell Biol. *216*, 2027–2045.

Saiselet, M., Floor, S., Tarabichi, M., Dom, G., Hébrant, A., van Staveren, W.C.G., and Maenhaut, C. (2012). Thyroid cancer cell lines: an overview. Front. Endocrinol. *3*.

Sakai, J., Duncan, E.A., Rawson, R.B., Hua, X., Brown, M.S., and Goldstein, J.L. (1996). Sterol-regulated release of SREBP-2 from cell membranes requires two sequential cleavages, one within a transmembrane segment. Cell *85*, 1037–1046.

Sakakura, Y., Shimano, H., Sone, H., Takahashi, A., Inoue, N., Toyoshima, H., Suzuki, S., Yamada, N., and Inoue, K. (2001). Sterol regulatory element-binding proteins induce an entire pathway of cholesterol synthesis. Biochem. Biophys. Res. Commun. *286*, 176–183.

Schumacher, M.M., Elsabrouty, R., Seemann, J., Jo, Y., and DeBose-Boyd, R.A. (2015). The prenyltransferase UBIAD1 is the target of geranylgeraniol in degradation of HMG CoA reductase. ELife *4*, e05560.

Serin, E.A.R., Nijveen, H., Hilhorst, H.W.M., and Ligterink, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. Front. Plant Sci. *7*.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science *343*, 84–87.

Shimada, K., Muhlich, J.L., and Mitchison, T.J. (2019). A tool for browsing the Cancer Dependency Map reveals functional connections between genes and helps predict the efficacy and selectivity of candidate cancer drugs. BioRxiv 2019.12.13.874776.

Shimano, H., Sato, R., and Sato, R. (2017). SREBP-regulated lipid metabolism: convergent physiology — divergent pathophysiology. Nat. Rev. Endocrinol. *13*, 710–730.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389–403.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tan, A.S., Baty, J.W., Dong, L.-F., Bezawork-Geleta, A., Endaya, B., Goodwin, J., Bajzikova, M., Kovarova, J., Peterka, M., Yan, B., et al. (2015). Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. Cell Metab. *21*, 81–94.

Tasdogan, A., McFadden, D.G., and Mishra, P. (2020). Mitochondrial DNA Haplotypes as Genetic Modifiers of Cancer. Trends Cancer *0*.

Taylor, R.W., and Turnbull, D.M. (2005). MITOCHONDRIAL DNA MUTATIONS IN HUMAN DISEASE. Nat. Rev. Genet. *6*, 389–402.

Vallett, S.M., Sanchez, H.B., Rosenfeld, J.M., and Osborne, T.F. (1996). A direct role for sterol regulatory element binding protein in activation of 3-hydroxy-3-methylglutaryl coenzyme A reductase gene. J. Biol. Chem. *271*, 12247–12253.

Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029–1033.

Varmus, H. (2002). Genomic empowerment: the importance of public databases. Nat. Genet. *32*, 3–3.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Vergara, J.R., and Estévez, P.A. (2014). A Review of Feature Selection Methods Based on Mutual Information. Neural Comput. Appl. *24*, 175–186.

Vivian, C.J., Brinker, A.E., Graw, S., Koestler, D.C., Legendre, C., Gooden, G.C., Salhia, B., and Welch, D.R. (2017). Mitochondrial Genomic Backgrounds Affect Nuclear DNA Methylation and Gene Expression. Cancer Res. *77*, 6202–6214.

Vyas, S., Zaganjor, E., and Haigis, M.C. (2016). Mitochondria and Cancer. Cell *166*, 555–566.

Wainberg, M., Kamber, R.A., Balsubramani, A., Meyers, R.M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., et al. (2019). A genome-wide almanac of co-essential modules assigns function to uncharacterized genes. BioRxiv 827071.

Wallace, D.C. (2012). Mitochondria and cancer. Nat. Rev. Cancer *12*, 685–698.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science *343*, 80–84.

Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell *168*, 890-903.e15.

Wang, X., Sato, R., Brown, M.S., Hua, X., and Goldstein, J.L. (1994). SREBP-1, a membrane-bound transcription factor released by sterol-regulated proteolysis. Cell *77*, 53–62.

Warburg, O. (1956). On the origin of cancer cells. Science *123*, 309–314.

Warburg, O., Wind, F., and Negelein, E. (1927). THE METABOLISM OF TUMORS IN THE BODY. J. Gen. Physiol. *8*, 519–530.

Weber, R.A., Yen, F.S., Nicholson, S.P.V., Alwaseem, H., Bayraktar, E.C., Alam, M., Timson, R.C., La, K., Abu-Remaileh, M., Molina, H., et al. (2020). Maintaining Iron Homeostasis Is the Key Role of Lysosomal Acidity for Cell Proliferation. Mol. Cell *77*, 645-655.e7.

Weinberg, S.E., and Chandel, N.S. (2015). Targeting mitochondria metabolism for cancer therapy. Nat. Chem. Biol. *11*, 9–15.

West, A.P., and Shadel, G.S. (2017). Mitochondrial DNA in innate immune responses and inflammatory pathology. Nat. Rev. Immunol. *17*, 363–375.

White, S.L., Collins, V.R., Wolfe, R., Cleary, M.A., Shanske, S., DiMauro, S., Dahl, H.H., and Thorburn, D.R. (1999). Genetic counseling and prenatal diagnosis for the mitochondrial DNA mutations at nucleotide 8993. Am. J. Hum. Genet. *65*, 474–482.

Williams, R.T., Guarecuco, R., Gates, L.A., Barrows, D., Passarelli, M.C., Carey, B., Baudrier, L., Jeewajee, S., La, K., Prizer, B., et al. (2020). ZBTB1 Regulates Asparagine Synthesis and Leukemia Cell Response to L-Asparaginase. Cell Metab. *31*, 852-861.e6.

Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics *6*, 227.

Yang, J., Goldstein, J.L., Hammer, R.E., Moon, Y.A., Brown, M.S., and Horton, J.D. (2001). Decreased lipid synthesis in livers of mice with disrupted Site-1 protease gene. Proc. Natl. Acad. Sci. U. S. A. *98*, 13607–13612.

Yuan, Y., Wang, W., Li, H., Yu, Y., Tao, J., Huang, S., and Zeng, Z. (2015). Nonsense and missense mutation of mitochondrial ND6 gene promotes cell migration and invasion in human lung adenocarcinoma. BMC Cancer *15*, 346.

Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J., Weinstein, J.N., et al. (2020). Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat. Genet. *52*, 342–352.

Zhu, X.G., Nicholson Puthenveedu, S., Shen, Y., La, K., Ozlu, C., Wang, T., Klompstra, D., Gultekin, Y., Chi, J., Fidelin, J., et al. (2019). CHP1 Regulates Compartmentalized Glycerolipid Synthesis by Activating GPAT4. Mol. Cell *74*, 45-58.e7.