APPLICATIONS OF NMR METABOLOMICS TO CELL MODELS OF DISEASE AND THERAPEUTIC MANUFACTURING

by

MAXWELL BACA COLONNA

(Under the Direction of Arthur S. Edison)

ABSTRACT

Small molecule metabolites are the last, often underappreciated level in the central dogma of molecular biology. Small molecules have become more appreciated for both their functional role and as a readout of the phenotype of a biological system. Metabolomics is the approach by which we identify and measure these small molecules. Nuclear Magnetic Resonance (NMR) is a common technique for performing metabolomics studies that possess several unique advantages for making such measurements. As the field of metabolomics has grown, so has the application of metabolomics in cell models of disease. Additionally, recent developments in cell-based therapeutics have also created a need for understanding and predicting function and product quality, which metabolomics of cell cultures is also well suited to provide. In this dissertation I will introduce the concepts and utility of NMR metabolomics to cell systems, then show how I have leveraged NMR metabolomics to gain a novel understanding of metabolic adaptations in cell models of cancer and rare genetic disease, as well as the measurement of culture media metabolites by NMR to develop a platform for predicting functional outputs of therapeutic cell products.

INDEX WORDS: NMR, metabolomics, cell culture, cancer cells, metabolites, cell manufacturing, cell therapy, metabolic disease, MCF-7, ALDH18A1, P5CS, CAR T cells, mesenchymal stromal cells, co-culture

APPLICATIONS OF NMR METABOLOMICS TO CELL MODELS OF DISEASE AND THERAPEUTIC MANUFACTURING

by

MAXWELL BACA COLONNA

B.S., The University of Texas at Austin, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

APPLICATIONS OF NMR METABOLOMICS TO CELL MODELS OF DISEASE AND THERAPEUTIC MANUFACTURING

by

MAXWELL BACA COLONNA

Major Professor:	Arthur S. Edison
Committee:	Shaying Zhao
	Kelley Moremen
	Ying Xu

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2022

DEDICATION

To my grandmother, Nancy Evans, who helped nurture my curiosity, showed me the value of travel, and always encouraged me to pursue my education. Thank you.

ACKNOWLEDGEMENTS

I have a lot of people to thank who helped me get to this point. First, my advisor, Art Edison, who took a chance on a musician with no research experience and has been patient, supportive, and encouraging throughout my time in his lab. You and Katherine have shown so much generosity to me over these last years, and given me a second family. Thank you to Shaying for welcoming me into her lab and for her and the rest of the lab teaching me so much in my first few years of grad school. Next, to all of the past and present members of the Edison group that have made the last 7 years a great environment to work, learn, play and teach. Thanks to Francesca for taking me under her wing at the beginning and showing me the ropes. Thank you to Jackie for her partnership and guidance. Thanks to Team Worm of Brie, Goncalo and Amanda for their comradery and support, and to many more years of friendship. Thanks to Nick for being a cool guy and a friend. Thanks to Bif, Tyler, Michael, Yue, Sicong, Fariba, Nicole, Omid, Zarif, Abby, Deanna, Conrad, Jojo, Ricardo and Nicci for all being great people and coworkers. Special thanks to Karen, Laura and Pam for keeping the lab running smoothly and their immeasurable help and support to make our lab an easy place to be. Thank you to all my wonderful friends (esp Chase, Josh and the Cheesers) and family at home who've been so supportive and encouraging of my journey. Thank you to my furbaby Naga who has helped me get through this last year and will be my best friend for a long time. Finally, a huge debt of gratitude to Margaret for being an incredibly supportive and loving partner to help me get through this process and over the finish line. And thanks to all of those not mentioned here who have encouraged my career in science, shown kindness, and impacted me.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Foreword

Sections 1.4 through 1.6 from Chapter 1 are reprinted with permission from Edison, A.S., Colonna, M.B., Gouveia, G.J., Holderman, N.R., Judge, M.T., Shen, X., and Zhang, S. NMR: Unique Strengths That Enhance Modern Metabolomics Research, Anal. Chem, 2021, 93, 1, 478–499, and is available at https://pubs.acs.org/doi/10.1021/acs.analchem.0c04414. Copyright 2021 American Chemical Society. The work was carried out with the goal of producing a review article, led by senior author Arthur S. Edison. My contribution to the work consisted of (i) writing the portion of the article titled "Metabolomics Applications in Mammalian Cell Cultures" and (ii) reviewing and editing the rest of the manuscript. Collaborators roles were as follows: Arthur S. Edison conceived and organized the content of the article, wrote the introduction and conclusion reviewed and edited the manuscript. Goncalo J. Gouveia wrote the portions of the article titled "Metabolite-Protein Interactions" and "Model Organisms and Metabolism." Nicole. R. Holderman wrote the portion of the article titled "Marine Environments and Carbon Cycling" and collected data for Figure 1. Michael T. Judge wrote the portion of the article titled "In-Vivo Metabolomics." Xunan Shen and Sicong Zhang both wrote the portion of the article titled "Chemoinformatics and Computational Modeling." All authors also reviewed and edited the manuscript. The work was supported by the following grants: NIH U2CES030167-03, NIH R01GM120151-04, NIH R01HD087306, NSF 1648035, NSF 1713746, the Gordon and Betty Moore Foundation, and the Georgia Research Alliance.

1.1 Introduction to Metabolomics

Metabolomics is a field that combines analytical chemistry, biology, and bioinformatics to identify and/or quantify small molecules present in a biological system. Small molecules are usually defined as compounds with a size of less than 1.5 kDa.[1] Metabolites are becoming more appreciated for both their active role in biology and their ability to reflect the dynamic state of an organism or system[2]. There are two primary methods for conducting metabolomics studies[1]. The first and most popular is mass spectrometry (MS), typically coupled to a form of chromatography, such as gas-chromatography (GC) or liquid-chromatography (LC). The other is nuclear magnetic resonance spectroscopy (NMR). While MS-based methods have the advantage of increased sensitivity, the large number of compounds detected, and obtaining elemental formulas of compounds. NMR is advantageous for its exceptional reproducibility, being sample non-destructive and inherently quantitative measurements. However these characteristics are not exhaustive, and unique advantages of NMR will be highlighted in the following sections of this introduction and have been extensively reviewed elsewhere[3,4]. Here I will briefly outline the key components of a general metabolomics study and their importance to biological data interpretation.

Metabolomics studies are composed of several key components, including study design, sample generation and collection, sample preparation, analytical measurement, data processing, analysis and interpretation. Crucial choices made at each of these steps will define the molecules to be measured, the comparability between samples within and between experiments, and the scope of conclusions able to be made. Study design includes the important choices of how samples will be grown/collected, how many samples, the treatments or perturbations given to each sample group, as well as the inclusion of relevant controls. These elements are all critical for obtaining

relevant information to make any statistically founded conclusions about the system being studied. Metabolomics studies are broadly categorized as being targeted or untargeted and is perhaps the first consideration in a metabolomics study[5]. Targeted studies are defined by *a priori* selection of a set or class of metabolites to be measured. This approach is usually used to test a hypothesis about a specific set of metabolites or metabolic pathway that can be interrogated directly. Alternatively, targeted studies can also be very broad, but use a predetermined library of compounds to measure (more common in MS-based approaches)[5,6]. Untargeted studies are defined by not assuming a specific class or set of compounds, or even knowing what compounds are to be measured in the study. All detectable compounds are measured and then identified *post hoc.* These studies are sometimes referred to as "profiling" or "fingerprinting." Their results can be used to generate hypotheses about the biological system to be subsequently tested directly[7]. Another key component of metabolomics study design is randomization, in the collection of samples, sample preparation and analytical measurements. Indeed, collecting or processing samples in batches can produce significant batch effects that can bias results and affect data interpretation.[8] Technical variability is introduced at each of these steps, and by introducing a randomized study design, can aid in distributing that variance across the data in a way that does not bias the results.[9]

Sample generation and collection are specific to the biological system being studied but will inform the subsequent steps of how samples are to be processed and metabolites extracted for the relevant analytical technique being employed. For example, a study of human urine samples will have much different sample collection and preparation steps than a study of tumor tissue[10]. Key considerations in the sample collection steps are minimizing perturbation of the biological system prior to collection and rapid quenching of run-on metabolism, typically achieved through rapid freezing and stable storage of the samples until they can be processed and analyzed[11]. Most

metabolomics studies offer a "snapshot" of the compounds present in a sample at the time they are collected and thus are highly influenced by the environmental conditions immediately around the time of sample collection. While no sample collection can be perfect in mitigating these confounding factors, achieving consistency between samples in a study is key to obtaining useful data and results.

Sample preparation is another critical process in metabolomics studies. The purpose of these steps is to take metabolites from the biological sample into a form that is amenable to analytical measurement. This typically involves the removal of components from the sample that will interfere with or disrupt data acquisition. Often these are macromolecules such as proteins and lipids.[12,13] In most sample matrices, these undesired components are removed through a form of chemical extraction, which selects for small molecules soluble in a chosen solvent while excluding the bulk mass of macromolecules. If the sample is solid or semisolid (i.e., tissue, cells, worms) this chemical extraction is usually accompanied by some physical disruption of the sample matrix, such as vortexing, bead-beating, grinding, or homogenization. The choice of extraction solvent will also determine the compounds that will be retained for analysis[14,15]. Indeed, it is common to perform multiple extractions with different solvents to maximize the recovery of small molecules of differing solubility. A fractionation during metabolite extraction can also be useful to reduce the chemical complexity of a single analytical sample, which may aid in quantitation and identification later in the analysis.[16,17] However, some sample matrices are amenable to little or no sample preparation depending on the methodology being used (see Section 1.5 for the example of culture media analysis by NMR). Regardless, sample handling and preparation is perhaps the most important step of determining the characteristics of the resulting data.[18]

Analytical measurements typically require the reconstitution of extracted and dried samples in a solvent. In the case of NMR, samples are reconstituted in a deuterium solvent containing a chemical shift reference standard. In the case of MS-based methods, samples are reconstituted in a solvent compatible with the chromatography being used in front of mass spectrometry. Specific analytical measurements are also study specific and will be discussed in more detail in the following section pertaining to NMR. Regardless of the methodology being employed, consistency of analytical conditions is paramount to obtaining reliable data.[18] This is especially relevant in large studies where data collection can last from many hours to many days. Deviations in conditions such as buffer lot, chromatography column, sample temperature, pH, etc. can result in observable effects that will influence measurements and interpretation of results.[9] In the case of pH shifts, often reconstitution solvents will also contain a buffer or be pH adjusted before analysis.[19]

Once the data are collected, appropriate pre-processing of raw spectra is necessary to filter out extraneous signals or artifacts from the analytical preparation/instrumentation, as well as adjusting spectral baselines, observing expected data from controls, and spectral alignment and normalization, making sure data from all samples are comparable. Typically, a final analysis will also include compound annotation to identify metabolites that have been measured. Ultimately, values of all spectral features or annotated metabolites are extracted from the processed data for individual samples and compiled into a table or matrix for statistical analysis. Analysis of the processed data is also very specific to the study goals but typically involves univariate and/or multivariate statistical analyses. Univariate tests such as a T-test or ANOVA consider values of one variable (spectral feature or annotated metabolite) at a time and compare them between experimental groups to determine if there are differences. Multivariate analyses, in contrast, look

at many or all variables from the dataset simultaneously. Multivariate methods are broadly categorized as supervised or unsupervised, which indicates whether the method is "told" which samples belong to which experimental groups. Unsupervised methods like principal component analysis use latent patterns in the dataset to reduce the dimensionality of the data, in which the individual samples are projected. Samples' relative distance to each other in these projected components indicates how similar or different their profiles are from each other. With supervised analyses such as partial-least squares discriminant analysis, group membership is known by the model, and thus identifies which combination of variables from the dataset distinguishes the groups from each other. While these are a small sample of basic methods for statistical analysis of metabolomics data, the subject has been well reviewed elsewhere.[20]

Through these general steps, metabolomics studies can provide both quantitative and qualitative information about biological based upon the measurement of metabolites and small molecules. However, it should be appreciated that the choices made at each step do highly influence the content of that information and how it is interpreted. Perhaps none is more important than the choice of technology used for a metabolomics study in determining the nature of that information. The nature of NMR data for metabolomics will be discussed in the following section.

1.2 NMR Metabolomics

As mentioned previously, NMR is one of the dominant technologies utilized for metabolomics studies. Here I will briefly summarize the basic principles and general applications of solution NMR to metabolomics. This is a simplified summary, and further details are excluded to keep subjects within the scope of this introduction. The goal is to provide a general background on the type of information that NMR provides, such that it can be appreciated how NMR provides

quantitative, structural information about molecules that can be leveraged for metabolomics studies.

NMR is an analytical method based on the quantum mechanical property of nuclei known as spin. Certain atomic nuclei are magnetically active: these nuclei can become polarized and will align under an applied external magnetic field. For NMR spectroscopy, these external magnetic fields are typically generated by superconducting magnets, many orders of magnitude greater than the earth's magnetic field (up to and above 20 Tesla). Nuclei that are aligned within an external magnetic field can be perturbed by radio-frequency pulses produced by transmitter coils around the sample. After being pushed out of alignment with the external magnetic field, nuclei will precess around the direction of the external field in a gyroscopic manner as they return to alignment with it. This precession of the nuclear magnetic fields induces a decaying, oscillating current in the detector coils. This electrical signal is received and recorded by the spectrometer in the form of a free induction decay (FID). Through the use of the Fourier transform, this signal as a function of time FID is converted to a spectrum of signal as a function of frequency. The central resonance frequency of a particular type of NMR nucleus (i.e., ^{1}H, ^{13}C etc.) in an external field is known as the Larmor frequency. The field strength of NMR magnets is typically referred to by the Larmor frequency of ^{1}H in those fields (i.e., 80 MHz, 600 MHz, 1.2 GHz, etc.). The specific resonance frequency as well as the pattern of signals (single peak, doublet peak, complex multiplet peaks, etc.) of nuclei within a molecule is determined by the chemical environment of the nucleus, or in other words, the structural connectivity of the molecule in which the nucleus is contained. Importantly, the signal intensity in an NMR spectrum is directly proportional to the number of those unique nuclei in the sample. Since spectral intensity is proportional to the number of resonating nuclei of that specific frequency, and frequency and signal patterns are unique to

specific molecular structures, compounds/metabolites have predictable patterns of spectral peaks whose intensity can be used to represent the relative concentration of that compound in the sample.

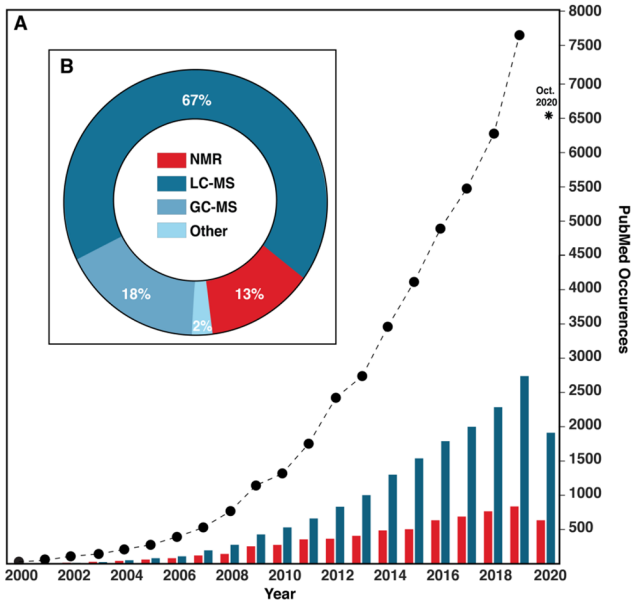


Figure 1.1: Overall trends. The bar chart in A shows PubMed search results for "metabolomics OR metabonomics" (black points connected by dashed lines), "metabolomics OR metabonomics AND mass spectrometry" (blue bars), and "metabolomics OR metabonomics AND NMR" (red bars). The inset pie chart B shows the current distribution of techniques used in studies deposited on the Metabolomics Workbench.[21] Both data were obtained Oct. 10, 2020 Reprinted with permission from Anal. Chem. 2021, 93, 1, 478–499. Copyright © 2020 American Chemical Society[3]

Despite the dominance of LC-MS as the primary metabolomics technology, NMR constitutes a significant portion of metabolomics studies that are published (Figure 1). ^{1}H (proton) NMR is the most commonly measured nucleus for metabolomics studies, due to the ubiquity of ^{1}H nuclei in organic molecules and having the highest sensitivity/detection by NMR due to its physical properties. One-dimensional ^{1}H NMR spectra are relatively quick to acquire and can detect tens to hundreds of compounds depending on the sample matrix. Thus, ^{1}H NMR is typically

used to profile all samples in a study. ^{1}H NMR methods can measure concentrations in the low micromolar range, with over 2 million to 1 dynamic range[3,22]. However, proton spectra of complex mixtures possess many convoluted peak patterns and signal overlaps, which can complicate both the quantitation and annotation of individual compounds. These splitting patterns are caused by scalar couplings, or the electromagnetic interactions between nuclei through chemical bonds. Multidimensional approaches can be used to separate otherwise convoluted signals. An example of this used in metabolomics is a homonuclear J-resolved (J-RES) experiment. This experiment generates two-dimensional spectra with the scalar couplings recorded in the second dimension, which eliminates the splitting patterns of typical ^{1}H spectra in a one-dimensional projection. Other two-dimensional homonuclear experiments such as Correlation Spectroscopy (COSY) and Total Correlation Spectroscopy (TOCSY) use those couplings to determine which proton signals are adjacent to each other within the same molecule. These methods are useful for both adding a second dimension of separation between signals. These and two-dimensional heteronuclear experiments are also commonly used for compound annotation and structural elucidation by NMR[4,23]. Heteronuclear Single Quantum Correlation (HSQC) and Heteronuclear Single Quantum Correlation-Total Correlation Spectroscopy (HSQC-TOCSY) experiments use scalar couplings between different NMR nuclei (most often ^{1}H and ^{13}C) to obtain further structural information about proton carbon pairs within molecules. In fact, HSQC spectra of complex mixtures can be used to match experimental data to spectral databases for compound annotation.[24,25] Multiple types of two-dimensional data, such as HSQC, TOCSY, and HSQC-TOCSY can be used to reinforce or confirm spectral matches, as they provide complementary structural information. The tradeoff of performing multidimensional experiments is typically paid in the form of experiment time. Heteronuclear experiments like HSQC, with samples containing a natural abundance of ^{13}C, can

take many hours to acquire, compared to minutes for a one-dimensional ^{1}H experiment. Additionally, careful considerations are necessary to be able to accurately quantitate two-dimensional spectra.[26] While methods have been developed to mitigate the time required for collecting multidimensional NMR spectra[27-29], they also require advanced expertise and/or processing procedures to perform. As a result, most studies only perform two-dimensional experiments on representative samples, such as internal pooled controls, to annotate the compounds that are present in experimental samples.

In summary, a typical NMR metabolomics study will use a combination of one-dimensional ^{1}H NMR for quantitation, in addition to two-dimensional experiments like HSQC to aid in annotating compounds measured in the proton spectrum. Once these data are collected, processed, and annotated, the data can be used for statistical analyses such as those described previously. With this general perspective on how NMR metabolomics is performed, I will now highlight some specific ways in which NMR metabolomics can be applied to gain unique insights into cell systems.

1.3 NMR Metabolomics of Mammalian Cell Systems

Cell culture is a mainstay of biomedical technology and research. From unraveling basic biology to manufacturing therapeutics, mammalian cell systems have been central to biomedicine. The ability to control and distinguish intracellular and extracellular components is something much more difficult in tissue or whole animals. In some contexts, knowledge of specific cell types under different conditions may provide the most relevant information. Genetic editing and other biochemical manipulations are also much easier to accomplish with cells, which is relevant for both basic research and therapeutic applications. It is therefore no surprise that metabolomics has been recognized as a valuable tool for cell culture applications.[30] The nature of clonal cell culture

results in lower biological variability between replicates, mitigating one of the major issues in metabolomics broadly.[31] While the vast bulk of human metabolomics studies are concerned with the analysis of biofluids and tissues, metabolomics studies of cells have continued to grow in number with the increasing numbers of metabolomics studies.[32] However, cell culture is often called into question for its relevance to *in vivo* biology, possibility for contamination or misidentification, and potential for genetic drift or phenotypic heterogeneity.[33] While some cell lines may have limited *in vivo* relevance, metabolomics data from cells is often complementary to that observed in whole tissues or animals.[32] Despite their limitations, the usefulness of cell models is in generating hypotheses and useful information where they could not be gained otherwise. Here I will highlight examples of the use and unique contributions of NMR in cell models of disease, particularly cancer, as well as in the context of mammalian cell manufacturing.

START OF PUBLISHED SECTION

1.4 Targeted Metabolism Studies

There are several unique features of NMR that allow for targeted analysis of specific metabolite classes or pathways of interest to cancer metabolism with the use of cell lines. This is a particularly powerful capability when advancing from common profiling or screening studies to understanding mechanisms or changes in metabolic flux that produce observed changes.

For example, coenzyme A species, redox metabolites such as NAD+ and NADH, as well as energy molecules like ATP are particularly useful for understanding cancer metabolism in vitro. However, these classes of molecules are difficult to measure with MS-based techniques due to their highly labile nature and structural similarity. These are challenging for NMR as well due to their low concentrations and aforementioned structural similarity which can produce high spectral overlap. Recently, Nagana Gowda et al. have optimized extraction and sample preparation techniques that

allow for preservation of these endogenous metabolites. A one-time addition of coenzyme standards to a reference sample was sufficient to identify and quantify signals from unique species using standard ^{1}H NMR across multiple samples.[34] Similarly, utilizing a combination of standard compound spiking and 2D correlation experiments, Nagana Gowda was also able to definitively identify and quantify redox coenzymes and adenosine phosphate species from extracted mammalian cells.[35] These methods highlight how, when combined with robust chemical extraction methods, the reproducibility and stability of metabolite chemical shifts in an NMR spectrum can enable comprehensive profiling of coenzymes and energy molecules with a single experiment. ^{13}C-labeled substrates are useful for tracking the flux of carbon sources and understanding the unique metabolism of different cancer types. While using SIL is not unique to NMR, the ability to directly and selectively detect molecules and the positions of atoms containing tracers such as ^{13}C is unparalleled. NMR provides atom-specific information, making it ideal for isotopomer analysis. Lane et al.[36] provide an excellent example of applying both NMR and LC-MS for investigating cancer energetics using both ^{13}C- and ^{15}N-labeled glucose, glutamine, glycerol, and octanoate to profile differential nutrient utilization between breast cancer cell lines of different histological subtypes. By using ^{1}H-detected 1D ^{13}C-HSQC (heteronuclear single quantum correlation) experiments, they compared NMR J-couplings at different nuclei in metabolites labeled by different isotopic substrates and were able to deduce the specific pathways used to metabolize these carbon sources. Combined with information from 2D ^{1}H- TOCSY (total correlation spectroscopy) experiments, they were able to quantify relative amounts of ribose contained in nucleotides generated by oxidative vs nonoxidative pentose phosphate pathways, among other insights.[36] Winnike et al. also used ^{13}C-labeled glucose and glutamine to determine the relative

flux of these nutrients in breast cancer cells using both directly detected ^{13}C 1D experiments as well as 2D ^{13}C-HSQC for metabolite annotation. ^{13}C detection provides greater chemical shift dispersion and less overlap than ^{1}H detected data.[37] Importantly, these studies and others have revealed that the typical classifications of breast cancer cells, such as proliferation rate or histological subtypes, do not necessarily predict metabolic pathway activity. To probe metabolism even more specifically at the enzyme level, Hattori et al.[38] utilized a variety of 1D and 2D NMR experiments on cell extracts labeled with either valine or its ketoacid, keto-isovalerate (KIV). To determine the directionality of the transamination reaction, which is catalyzed by branched-chain amino acid aminotransferase 1 (BCAT1), they either added ^{13}C-valine with natural abundance KIV or ^{13}C-KIV with natural abundance valine. Similar experiments were done with ^{15}N to follow the amino group. These data showed that leukemia cells preferentially transaminate branched-chain keto acids to their respective branched-chain amino acids, uncovering a novel behavior of leukemia cells shown to enhance their malignancy.[38] Judge et al. were able to use CIVM-NMR to reproduce this result using the same matched pairs of substrates in live cells utilizing a continuous 1D ^{13}C- HSQC experiment to detect the protons that were connected to labeled carbons, giving a simple and direct real-time display of KIV turnover (Figure 2).[39]

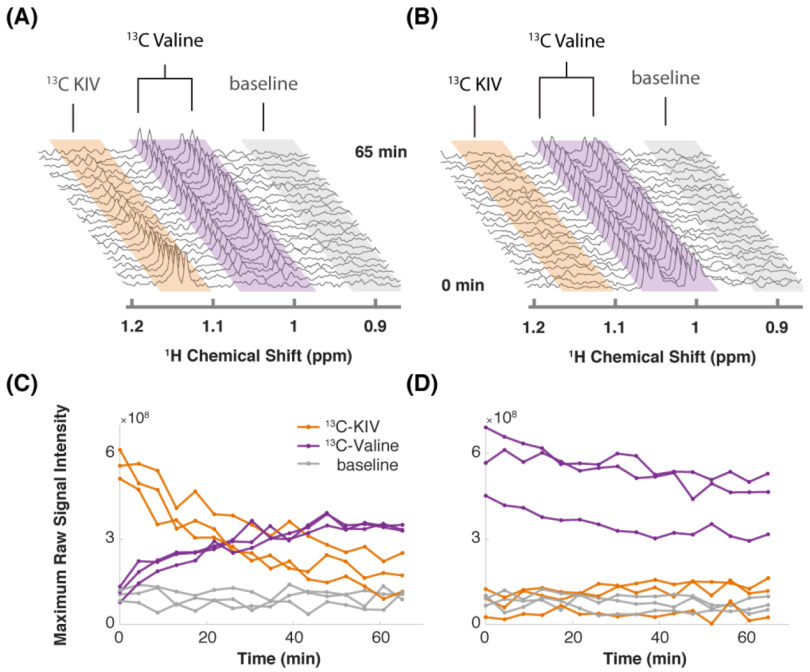


Figure 1.2 Targeted isotopic CIVM-NMR measurement of metabolic flux in human myeloid leukemia cells. (A) 13C-labeled keto-isovalerate (KIV) was converted to valine. (B) 13C-labeled valine was not converted to KIV, confirming unidirectional flux in ML cells. (C, D) Relative concentrations over time of 13C-labeled KIV (orange) and 13C-labeled valine (purple) compared to baseline noise (gray), obtained by taking the raw maximum spectral intensity within each region of the representative experiments in (A, B), respectively. Different lines show the data from 3 independent replicates of each experiment. Reprinted with permission from Judge, M. T.; Wu, Y.; Tayyari, F.; Hattori, A.; Glushka, J.; Ito, T.; Arnold, J.; Edison, A. S. Continuous in vivo Metabolism by NMR. Front. Mol. Biosci. 2019, 6, 26. doi: 10.3389/fmolb.2019.00026.

While ^{1}H and ^{13}C atoms are the most commonly used nuclei for profiling metabolites, there are other nuclei that can be leveraged for targeted analysis of metabolism in cancer cells by NMR. For instance, phosphocholine and phosphoethanolamine related molecules, which have been previously observed to be significant in cancer studies by NMR,[40,41] can be detected directly using ^{31}P NMR. Similar to ^{1}H, ^{31}P is an NMR active isotope and occurs at 100% natural abundance, which provides higher sensitivity than ^{13}C. There are fewer phosphorus resonances comprising a typical biological sample, resulting in a less crowded spectrum. Juranic and co-workers have developed useful NMR-based methods to characterize high-energy phosphometabolites like

ATP.[42] Not only do they take advantage of ^{31}P NMR, but they also label samples with added $H_2{}^{18}O$. The addition of ^{18}O is indirectly detectable through isotope effects that manifest on the ^{31}P nuclei, allowing for elucidation of valuable functional information in perfused tissues. Shah et al. demonstrated the utility of ^{31}P NMR in capturing the dynamics of different phosphoethanolamine species across cancerous and nonmalignant cell lines, revealing differential dependence of cancer cells on phospholipid synthesis when biosynthetic genes were knocked out.[43] Veronesi et al.[44] used direct ^{19}F detection of fluorine-labeled substrates and their enzymatic products to monitor the activity of a specific enzyme in living cells. Similar to ^{31}P, ^{19}F is an NMR active isotope that occurs at 100% natural abundance, with essentially no background resonances in most biological samples. This allows tracking the fate of ^{19}F tracer molecules without signals from endogenous compounds. With their system, Veronesi et al. were able to quantify changes in fatty acid amide hydrolase activity upon treatment with inhibitors.[44] As the first example of this type of quantitative kinetic data obtained in intact cells, ^{19}F has applications for both drug screening and targeted metabolism studies.

1.5 Media Analysis of Cell Cultures

The analysis of extracellular metabolites in culture media is important to the study of cancer cell metabolism in vitro. NMR is well suited for this due to the minimal sample preparation needed, which allows aliquots of culture media to be analyzed via NMR directly with the addition of a chemical shift reference in 5− 10% D_2O. Complementing the intracellular metabolome with changes in the extracellular environment provides greater context for interpreting the results of metabolomics studies. In addition, the use of small diameter, low volume sample tubes with a

high-sensitivity small volume probe enables sampling of media from the same culture repeatedly over time for time-course data.

Recently, Wojtowicz et al.[45] performed a time-course NMR analysis of media in the culture of breast cancer cells and compared the media profiles to NMR profiles of serum from breast cancer patients, to see if direct comparisons could reasonably be made between them when performing in vitro studies. Wojtowicz and colleagues collected media samples from breast cancer cell cultures 16 times over the course of 72 h, revealing dynamic, nonmonotonic changes in several detected metabolites. Lactate, alanine, glutamine, tyrosine, and glucose profiles in the culture media showed opposite trends of accumulation or depletion compared to changes in patient serum vs healthy controls, thus exhibiting some key limitations of direct comparisons of media with serum. However, the authors noted that for many metabolites, the interpretation of the results changed drastically depending on which time point was examined.[45] This again implicates the importance of collecting dynamic measurements at biologically relevant resolution to contextualize and properly interpret metabolomics data.

Mahar and colleagues recently showed the utility of combining isotopomer analysis with conditioned media analysis by NMR to quantify the Warburg effect, a metabolic signature of cancer cells where most glucose is converted to lactate instead of pyruvate.[46] Both normal and cancer cell lines were cultured in media containing [$^{2}H_7$]glucose, and the media were sampled over the course of 5 h. Due to the isotopic effect of deuterium incorporation, they were able to use ^{2}H decoupled ^{1}H NMR to resolve and track the accumulation of lactate isotopomers excreted into the media over the culture period, after correcting for changes in T_1 relaxation, another isotopic effect. They showed not only that the cancer cell line consumed almost six-times the amount of glucose compared to normal liver cells, but also nearly all of the [$^{2}H_7$]glucose consumed by the

cancer cells could be accounted for by glycolysis-produced lactate and monodeuterated water.[46] This study illustrates the power of measuring stable isotopes in conditioned media to derive definitive flux measurements through pathways enabled by the quantitative nature of NMR.

A unique twist on the analysis of conditioned media is to use it to simulate an in vivo environment for cancer cells. Luis et al. used conditioned media from cultured adipocytes to expose MCF-7 cells to an environment simulating an in vivo obesity phenotype. ^{1}H NMR analysis of the conditioned media before and after culturing MCF-7 cells showed an "inversion" of the Warburg effect as evidenced by increased glycolytic intermediates in the culture medium, along with increased proliferation and invasiveness characteristics of the cells.[47] The straightforward sampling and preparation of media for NMR analysis enabled the authors to conclude that the conditioned media metabolite profiles may be relevant to in vivo disease and have provided more evidence for how and why breast cancer prognoses are worse in obese women.

While the methods and approaches for interrogating mammalian (cancer) cell metabolism in vitro continue to become more advanced and comprehensive, the need for more physiologically relevant culture models amenable to these techniques is apparent. Ultimately, the application of these techniques to more complex models such as tissue-on-a-chip, coculture, or organoid systems will be able to provide metabolic discoveries with more relevance to disease and physiology. However, the direct application of classical untargeted approaches of intracellular and extracellular analysis can also be of use to profile the

1.6 Mammalian Cell-Based Manufacturing

Aside from basic biology research, another significant application of mammalian cell culture is in industrial production of biotherapeutics. While the industry of cell expressed biologics continues to grow and diversify, a new generation of cells-as-therapies is also on the horizon. Here

we will describe recent uses of NMR for mammalian cell-based biologics manufacturing in existing industrial processes and the future of NMR as a potential key technology for manufacturing of cell therapies.

The use of NMR, metabolomics, and other techniques for mammalian cell production of biologics manufacturing is not new,[48-50] but they have recently gained popularity for their utility in cell culture engineering and prediction of product quality. Recent work by Ali et al. and Zürcher et al. has highlighted the use of metabolomics data collected during cell culture bioprocesses to predict and improve the quality of products produced in cells.[51,52] However, there are some unique applications of NMR technology to improving cell-based bioprocesses. Brinson and Marino recently demonstrated the use of 2D NMR spectroscopy to provide a high-order structural fingerprint of cell-expressed protein products.[53] Since three-dimensional structure is critical to protein function, these NMR signatures can be used as metrics to assess product quality. This application is an example of the utility and versatility of structural information that is provided by NMR in cell manufacturing.

In addition, Blondeel et al. used time-course ^{1}H NMR metabolomics of culture media in cells expressing recombinant proteins in order to identify metabolites that were rapidly consumed, thus limiting cell density and production. By supplementing additional nutrients observed by NMR to be rapidly depleted, they were able to achieve a nearly 75% increase in cell density.[54] As mentioned elsewhere, the recurring measurement of metabolites in culture media is uniquely suited to NMR due to the simplicity of sample preparation and inherently quantitative nature of measurements.

As an evolution from cell expression-based production of biologics, cell therapies (such as CAR-T cells, mesenchymal stromal cells, and stem cell therapies) are poised to become a

significant sector of biopharmaceutical manufacturing.[55-57] Since cell therapies are even more complicated than the expression of biologics, cell products require increasingly comprehensive characterization and optimization for manufacture. Recent studies have already demonstrated the utility of metabolomics measurements to characterize and improve growth of cell therapy products.[58-61] ^{1}H NMR was recently used by Agostini et al. to characterize and predict the quality of platelet-derived media supplements for industrial cell manufacturing, leveraging the ease of sample preparation and throughput of media analysis by NMR.[62] Continuous flow NMR techniques for monitoring mammalian cell culture systems have existed for decades[63] but have not yet become a central technology in cell manufacturing. However, NMR could fill the need for a nonintrusive, high information content, online monitoring technique that can be leveraged to predict and improve product quality.

END OF PUBLISHED SECTION

1.7 Outline of my dissertation

This dissertation is split into two parts, both containing work of mine utilizing NMR metabolomics applied to mammalian cell cultures. The first part will outline my work applying untargeted NMR metabolomics to human cell models of disease to gain novel insights into possible disease mechanisms. The second part will highlight my work on developing a platform for the analysis of culture media in therapeutic cell manufacturing to predict cell product quality and function. I will conclude with comments on the future directions of these works.

1.8 References

(1) Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. *Chem Soc Rev* **2011**, *40*, 387-426.

(2) Rinschen, M. M.; Ivanisevic, J.; Giera, M.; Siuzdak, G. *Nat Rev Mol Cell Biol* **2019**, *20*, 353-367.

(3) Edison, A. S.; Colonna, M.; Gouveia, G. J.; Holderman, N. R.; Judge, M. T.; Shen, X.; Zhang, S. *Anal Chem* **2021**, *93*, 478-499.

(4) Wishart, D. S.; Cheng, L. L.; Copie, V.; Edison, A. S.; Eghbalnia, H. R.; Hoch, J. C.; Gouveia, G. J.; Pathmasiri, W.; Powers, R.; Schock, T. B.; Sumner, L. W.; Uchimiya, M. *Metabolites* **2022**, *12*.

(5) Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B. *Metabolomics* **2018**, *14*, 72.

(6) Ribbenstedt, A.; Ziarrusta, H.; Benskin, J. P. *PLoS One* **2018**, *13*, e0207082.

(7) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. *J Am Soc Mass Spectrom* **2016**, *27*, 1897-1905.

(8) Han, W.; Li, L. *Mass Spectrom Rev* **2022**, *41*, 421-442.

(9) Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D. *Bioanalysis* **2012**, *4*, 2249-2264.

(10) Beckonert, O.; Keun, H. C.; Ebbels, T. M.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Nat Protoc* **2007**, *2*, 2692-2703.

(11) Stevens, V. L.; Hoover, E.; Wang, Y.; Zanetti, K. A. *Metabolites* **2019**, *9*.

(12) Nagana Gowda, G. A.; Raftery, D. *Anal Chem* **2014**, *86*, 5433-5440.

(13) McHugh, C. E.; Flott, T. L.; Schooff, C. R.; Smiley, Z.; Puskarich, M. A.; Myers, D. D.; Younger, J. G.; Jones, A. E.; Stringer, K. A. *Metabolites* **2018**, *8*.

(14) Sauerschnig, C.; Doppler, M.; Bueschl, C.; Schuhmacher, R. *Metabolites* **2017**, *8*.

(15) Beltran, A.; Suarez, M.; Rodriguez, M. A.; Vinaixa, M.; Samino, S.; Arola, L.; Correig, X.; Yanes, O. *Anal Chem* **2012**, *84*, 5838-5844.

(16) Whiley, L.; Chekmeneva, E.; Berry, D. J.; Jimenez, B.; Yuen, A. H. Y.; Salam, A.; Hussain, H.; Witt, M.; Takats, Z.; Nicholson, J.; Lewis, M. R. *Anal Chem* **2019**, *91*, 8873-8882.

(17) Lopes, A. G.; Borges, R. M.; Kuhn, S.; Garrett, R.; Costa, F. D. N. *J Chromatogr A* **2022**, *1677*, 463211.

(18) Duarte, I. F.; Marques, J.; Ladeirinha, A. F.; Rocha, C.; Lamego, I.; Calheiros, R.; Silva, T. M.; Marques, M. P.; Melo, J. B.; Carreira, I. M.; Gil, A. M. *Anal Chem* **2009**, *81*, 5023-5032.

(19) Emwas, A. H.; Roy, R.; McKay, R. T.; Ryan, D.; Brennan, L.; Tenori, L.; Luchinat, C.; Gao, X.; Zeri, A. C.; Gowda, G. A.; Raftery, D.; Steinbeck, C.; Salek, R. M.; Wishart, D. S. *J Proteome Res* **2016**, *15*, 360-373.

(20) Blaise, B. J.; Correia, G. D. S.; Haggart, G. A.; Surowiec, I.; Sands, C.; Lewis, M. R.; Pearce, J. T. M.; Trygg, J.; Nicholson, J. K.; Holmes, E.; Ebbels, T. M. D. *Nat Protoc* **2021**, *16*, 4299-4326.

(21) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res* **2016**, *44*, D463-470.

(22) Godecke, T.; Napolitano, J. G.; Rodriguez-Brasco, M. F.; Chen, S. N.; Jaki, B. U.; Lankin, D. C.; Pauli, G. F. *Phytochem Anal* **2013**, *24*, 581-597.

(23) Nagana Gowda, G. A.; Raftery, D. *J Magn Reson* **2015**, *260*, 144-160.

(24) Wang, C.; Timari, I.; Zhang, B.; Li, D. W.; Leggett, A.; Amer, A. O.; Bruschweiler-Li, L.; Kopec, R. E.; Bruschweiler, R. *J Proteome Res* **2020**, *19*, 1674-1683.

(25) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A., et al. *Nucleic Acids Res* **2018**, *46*, D608-D617.

(26) Giraudeau, P. *Magn Reson Chem* **2014**, *52*, 259-272.

(27) Hyberts, S. G.; Milbradt, A. G.; Wagner, A. B.; Arthanari, H.; Wagner, G. *J Biomol NMR* **2012**, *52*, 315-327.

(28) Mori, S.; Abeygunawardana, C.; Johnson, M. O.; van Zijl, P. C. *J Magn Reson B* **1995**, *108*, 94-98.

(29) Hansen, A. L.; Kupce, E. R.; Li, D. W.; Bruschweiler-Li, L.; Wang, C.; Bruschweiler, R. *Anal Chem* **2021**, *93*, 6112-6119.

(30) Khoo, S. H.; Al-Rubeai, M. *Biotechnol Appl Biochem* **2007**, *47*, 71-84.

(31) Snijder, B.; Sacher, R.; Ramo, P.; Damm, E. M.; Liberali, P.; Pelkmans, L. *Nature* **2009**, *461*, 520-523.

(32) Cuperlovic-Culf, M.; Barnett, D. A.; Culf, A. S.; Chute, I. *Drug Discov Today* **2010**, *15*, 610-621.

(33) Kaur, G.; Dufour, J. M. *Spermatogenesis* **2012**, *2*, 1-5.

(34) Gowda, G. A. N. *Metabolites* **2018**, *8*.

(35) Nagana Gowda, G. A.; Abell, L.; Tian, R. *Anal Chem* **2019**, *91*, 2464-2471.

(36) Lane, A. N.; Tan, J.; Wang, Y.; Yan, J.; Higashi, R. M.; Fan, T. W. *Metab Eng* **2017**, *43*, 125-136.

(37) Winnike, J. H.; Stewart, D. A.; Pathmasiri, W. W.; McRitchie, S. L.; Sumner, S. J. *Int J Breast Cancer* **2018**, *2018*, 2063540.

(38) Hattori, A.; Tsunoda, M.; Konuma, T.; Kobayashi, M.; Nagy, T.; Glushka, J.; Tayyari, F.; McSkimming, D.; Kannan, N.; Tojo, A.; Edison, A. S.; Ito, T. *Nature* **2017**, *545*, 500-504.

(39) Judge, M. T.; Wu, Y.; Tayyari, F.; Hattori, A.; Glushka, J.; Ito, T.; Arnold, J.; Edison, A. S. *Front Mol Biosci* **2019**, *6*, 26.

(40) Verma, A.; Kumar, I.; Verma, N.; Aggarwal, P.; Ojha, R. *BBA Clin* **2016**, *5*, 170-178.

(41) Podo, F.; de Certaines, J. D. *Anticancer Res* **1996**, *16*, 1305-1315.

(42) Juranic, N.; Nemutlu, E.; Zhang, S.; Dzeja, P.; Terzic, A.; Macura, S. *J Biomol NMR* **2011**, *50*, 237-245.

(43) Shah, T.; Krishnamachary, B.; Wildes, F.; Wijnen, J. P.; Glunde, K.; Bhujwalla, Z. M. *NMR Biomed* **2018**, *31*, e3936.

(44) Veronesi, M.; Giacomina, F.; Romeo, E.; Castellani, B.; Ottonello, G.; Lambruschini, C.; Garau, G.; Scarpelli, R.; Bandiera, T.; Piomelli, D.; Dalvit, C. *Anal Biochem* **2016**, *495*, 52-59.

(45) Wojtowicz, W.; Wrobel, A.; Pyziak, K.; Tarkowski, R.; Balcerzak, A.; Bebenek, M.; Mlynarz, P. *Metabolites* **2020**, *10*.

(46) Mahar, R.; Donabedian, P. L.; Merritt, M. E. *Sci Rep* **2020**, *10*, 8885.

(47) Luis, C.; Duarte, F.; Faria, I.; Jarak, I.; Oliveira, P. F.; Alves, M. G.; Soares, R.; Fernandes, R. *Life Sci* **2019**, *223*, 38-46.

(48) Bradley, S. A.; Ouyang, A.; Purdie, J.; Smitka, T. A.; Wang, T.; Kaerner, A. *J Am Chem Soc* **2010**, *132*, 9531-9533.

(49) Chrysanthopoulos, P. K.; Goudar, C. T.; Klapa, M. I. *Metab Eng* **2010**, *12*, 212-222.

(50) Lewis, A. M.; Abu-Absi, N. R.; Borys, M. C.; Li, Z. J. *Biotechnol Bioeng* **2016**, *113*, 26-38.

(51) Zurcher, P.; Sokolov, M.; Bruhlmann, D.; Ducommun, R.; Stettler, M.; Souquet, J.; Jordan, M.; Broly, H.; Morbidelli, M.; Butte, A. *Biotechnol Prog* **2020**, *36*, e3012.

(52) Ali, A. S.; Chen, R.; Raju, R.; Kshirsagar, R.; Gilbert, A.; Zang, L.; Karger, B. L.; Ivanov, A. R. *Biotechnol J* **2020**, *15*, e1900565.

(53) Brinson, R. G.; Marino, J. P. *J Magn Reson* **2019**, *307*, 106581.

(54) Blondeel, E. J. M.; Ho, R.; Schulze, S.; Sokolenko, S.; Guillemette, S. R.; Slivac, I.; Durocher, Y.; Guillemette, J. G.; McConkey, B. J.; Chang, D.; Aucoin, M. G. *J Biotechnol* **2016**, *234*, 127-138.

(55) Feigal, E. G.; DeWitt, N. D.; Cantilena, C.; Peck, C.; Stroncek, D. *Nat Immunol* **2019**, *20*, 955-962.

(56) Roddie, C.; O'Reilly, M.; Dias Alves Pinto, J.; Vispute, K.; Lowdell, M. *Cytotherapy* **2019**, *21*, 327-340.

(57) Arjmand, B.; Sarvari, M.; Alavi-Moghadam, S.; Payab, M.; Goodarzi, P.; Gilany, K.; Mehrdad, N.; Larijani, B. *Front Endocrinol (Lausanne)* **2020**, *11*, 430.

(58) Doron, G.; Klontzas, M. E.; Mantalaris, A.; Guldberg, R. E.; Temenoff, J. S. *Biotechnol Bioeng* **2020**, *117*, 1761-1778.

(59) Wang, D.; Zhao, L.; Zheng, H.; Dong, M.; Pan, L.; Zhang, X.; Zhang, H.; Gao, H. *Mol Neurobiol* **2018**, *55*, 1112-1122.

(60) Lussey-Lepoutre, C.; Hollinshead, K. E.; Ludwig, C.; Menara, M.; Morin, A.; Castro-Vega, L. J.; Parker, S. J.; Janin, M.; Martinelli, C.; Ottolenghi, C.; Metallo, C.; Gimenez-Roqueplo, A. P.; Favier, J.; Tennant, D. A. *Nat Commun* **2015**, *6*, 8784.

(61) Wilkins, J.; Sakrikar, D.; Petterson, X. M.; Lanza, I. R.; Trushina, E. *Metabolomics* **2019**, *15*, 83.

(62) Agostini, F.; Ruzza, M.; Corpillo, D.; Biondi, L.; Acquadro, E.; Canepa, B.; Viale, A.; Battiston, M.; Serra, F.; Aime, S.; Mazzucato, M. *PLoS One* **2018**, *13*, e0203048.

(63) Gonzalez-Mendez, R.; Wemmer, D.; Hahn, G.; Wade-Jardetzky, N.; Jardetzky, O. *Biochim Biophys Acta* **1982**, *720*, 274-280.

PART 1

NMR METABOLOMICS IN CELL MODELS OF DISEASE

CHAPTER 2

A DIVERGED MCF-7 CELL LINE SHOWS ALTERNATIVE PI3K/AKT SIGNALING AND REWIRED METABOLISM ASSOCIATED WITH AN EMT-LIKE PHENOTYPE[1]

[1]Colonna, M.B., Edison, A.S., Zhao, S. A Diverged MCF-7 Cell Line Shows Alternative PI3K/Akt Signaling And Rewired Metabolism Associated With An EMT-Like Phenotype. To be submitted to Oncotarget, anticipated acceptance 2023

2.1 Abstract

Spontaneous changes in *in vitro* cancer cell phenotypes may lend insight into *in vivo* mechanisms of oncogenesis and metabolic changes that coincide with a change in cell phenotype are informative to understand the mechanisms and patterns that reflect cancer cell phenotypes [6]. We describe a divergent cell line from MCF-7 breast cancer cells, coined "Augusta", that spontaneously acquired phenotypic characteristics of Epithelial-mesenchymal-transition (EMT) after extended culture in a high-glucose media. However, untargeted metabolic and transcriptional profiling of these cells show changes inconsistent with EMT. Upon further analysis, the PI3K/Akt pathway showed highest enrichment of differentially expressed genes, particularly elements of PI3K**γ** signaling. Integrated analysis of metabolic pathways suggests increases in biosynthetic activity through glucose consumption, catabolism of branched-chain amino acids (BCAAs), and increased lipid turnover in Augusta cells alongside these changes in PI3K**γ** signaling and phenotype. These results illustrate the dramatic metabolic changes that can occur spontaneously in cancer cell systems and provides clues to potential mechanisms of cancer cell adaptation to changes in environment.

2.2 Introduction

Epithelial to Mesenchymal Transition (EMT) is a conserved process of cellular reprogramming that is associated with embryonic tissue development, but also in cancer progression[1,2]. Cancer cells from epithelial tissues sometimes undergo EMT on the way to becoming metastatic and invasive. It has been shown in models of EMT that metabolism is both a consequence and driver of cancer invasiveness.[3,4] Known metabolic adaptations in EMT include increased glucose consumption and glycolytic activity, decreased lipid biosynthesis, and evidence of mitochondrial dysfunction and oxidative stress[3,5,6]. There is also a well characterized gene expression program that is associated with EMT, characterized by regulation in several key effector genes such as ZEB, SNAIL, and TWIST.[7,8] Research has been put into understanding metabolic pathways involved in EMT in order to develop interventions for cancer progression and metastasis[9-11]. There is still a need for cell models to understand the phenomenon of invasive transitions, and particularly metabolic changes that coincide. We present here a variant of the MCF-7 cell line coined Augusta. Augusta cells emerged after extended culture in RPMI-1640 media, compared to typical DMEM media. Crucially, RPMI-1640 contains about double the amount of glucose compared to DMEM. Augusta cells possess a mesenchymal cell phenotype and increased cell proliferation: characteristics of invasive cancer cells having undergone EMT. We sought to assess if this cell line could be used as a spontaneous model of EMT *in vitro.* To achieve this, we performed untargeted ^{1}H NMR metabolomics and RNAseq on Augusta cells as well as "stock" MCF-7 cells acquired from ATCC in order to observe if Augusta cells displayed metabolic and gene expression changes consistent with EMT. Gene expression and metabolic analysis of Augusta cells showed changes partially inconsistent with a canonical EMT. Analysis of the most highly altered pathways highlighted the PI3K/Akt super-pathway, and in particular the PI3K**γ**

pathway. In addition, our integrated metabolic pathway analysis shows these changes in gene expression coincide with an increase in lipid metabolism, biosynthesis from glucose, and degradation of branched chain amino acids. These results are consistent with previous work identifying the PI3K**γ** pathway and extracellular acidification as a potentiator for cell invasiveness, along with identifying similar metabolic adaptations in invasive cell phenotypes. While these cells are not an obvious model for EMT, they provide an opportunity for further exploring the metabolic reprogramming among the multitude of cancer cell phenotypes.

2.3 Results

Augusta cells are phenotypically diverged from MCF-7 breast cancer cells possessing EMT-like characteristics

While MCF-7 cells retain many differentiated epithelial characteristics, including abundant cell-cell contacts, Augusta cells developed a marked change in cell morphology, corresponding to a more stromal or mesenchymal phenotype (Figure 2.1A,B). In addition to this change in cell morphology, Augusta cells to have a higher proliferation rate as shown by increased DNA content (Figure 2.1D), and a show a rapid acidification of the culture media (Figure 2.1C). When grown in co-culture with cancer associated fibroblasts, Augusta cells appear to show a more invasive phenotype in their growth and cell-cell interactions (Supplemental Figure 2.1). We sought to assess the metabolic and transcriptional reprogramming of Augusta cells to determine if they were consistent with known signatures of EMT in cancer cells.

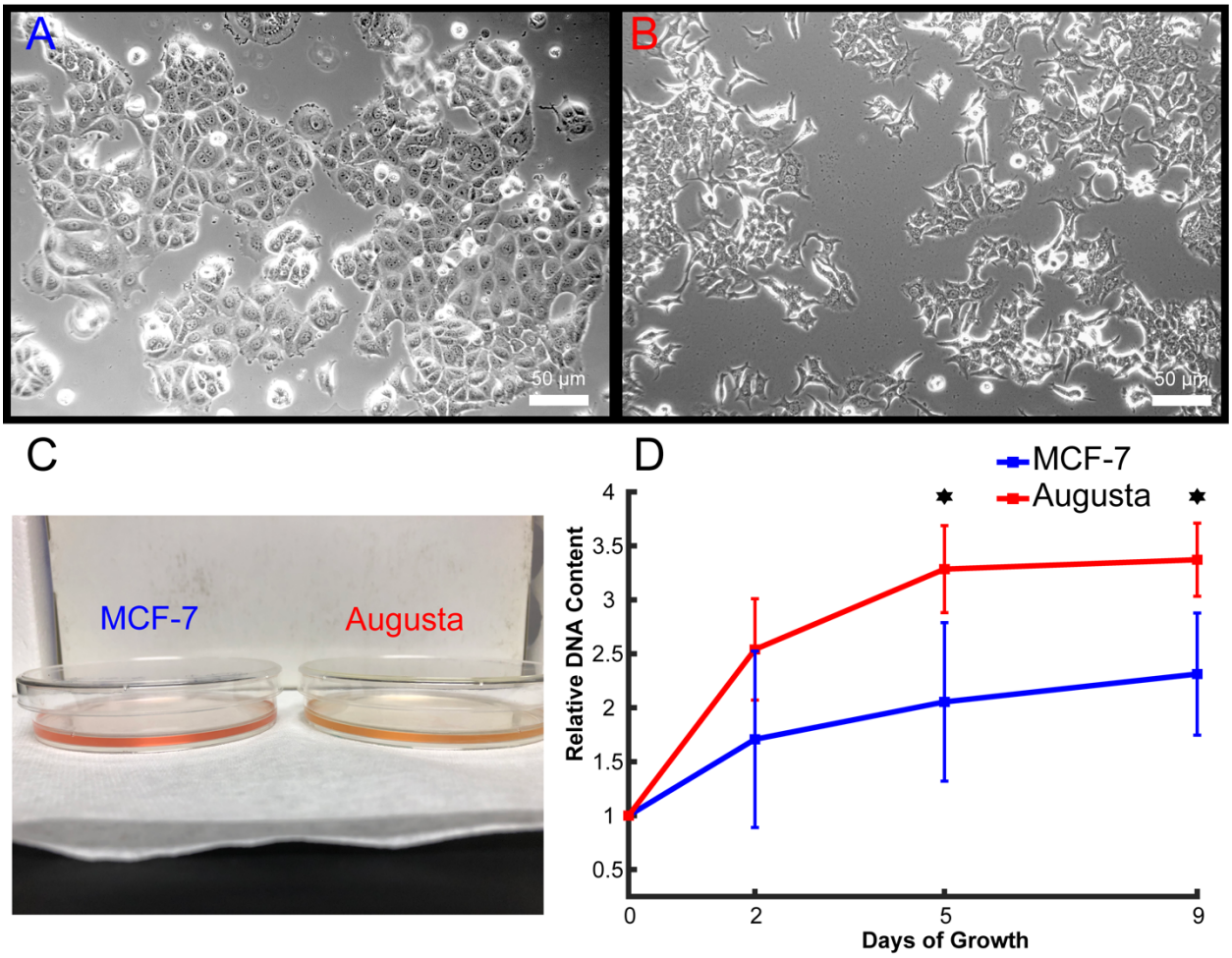


Figure 2.1: Augusta is a cell line diverged from the parental MCF-7 breast cancer cell line with altered morphology and proliferation characteristics. A and B) Phase contrast micrographs of MCF-7 and Augusta cells, respectively. Images taken at 400X total magnification. C) Photograph of 10cm culture plates after two days culture with equivalent number of cells. Media contains phenol red pH indicator. D) DNA fluorescence proliferation assay indicating relative DNA content of cell cultures over 9 days. N=4 for each cell line. Stars indicate p-value < 0.05

Table 2.1 Quantification of all significant annotated metabolites from ^{1}H NMR spectra of cell extracts (FDR < 0.05, MCF-7 N=10, Augusta N=11) and conditioned media (raw p-value < 0.05, MCF-7 N=6, Augusta N=6). Mean values are in relative units. Fold change values are in Augusta

relative to MCF-7. Shaded rows indicate increase in Augusta relative to parental MCF-7. [1]AXP represents adenosine phosphate species that could not be distinguished.

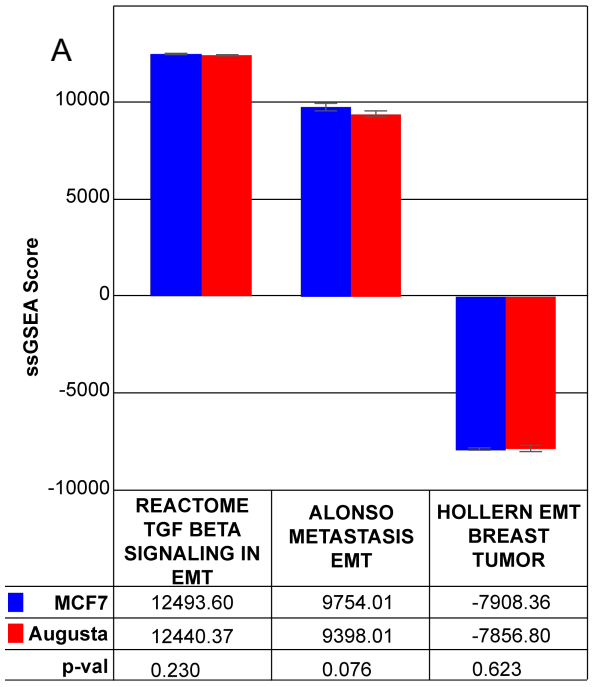
Metabolite	Confidence Score	Mean ± SD		log2(FC)	*p*-value	FDR
		MCF-7	Augusta			
CELL EXTRACTS						
myo-Inositol	4	0.192 ± 0.01	0.267 ± 0.01	0.476	4.68E-13	1.54E-11
Isoleucine	4	0.626 ± 0.04	0.461 ± 0.04	-0.442	4.14E-09	6.60E-08
Serine	4	0.451 ± 0.03	0.604 ± 0.04	0.421	6.00E-09	6.60E-08
Leucine	4	1.628 ± 0.11	1.208 ± 0.11	-0.431	3.61E-08	2.98E-07
Threonine	4	0.499 ± 0.02	0.422 ± 0.03	-0.242	5.07E-07	3.34E-06
Creatine phosphate	4	0.317 ± 0.08	0.519 ± 0.06	0.712	3.71E-06	2.04E-05
Lactate	4	2.442 ± 0.16	1.981 ± 0.18	-0.302	6.17E-06	2.91E-05
Putrescine	4	0.082 ± 0.01	0.099 ± 0.01	0.268	1.45E-05	5.99E-05
Glycerophosphocholine	3	0.938 ± 0.05	1.085 ± 0.09	0.211	2.08E-04	6.89E-04
Valine	4	0.394 ± 0.02	0.342 ± 0.03	-0.204	2.09E-04	6.89E-04
Glycine	3	0.869 ± 0.04	0.938 ± 0.04	0.109	5.69E-04	1.71E-03
Creatine	4	0.362 ± 0.09	0.499 ± 0.07	0.465	7.93E-04	2.18E-03
AXP[1]	2	0.166 ± 0.01	0.185 ± 0.01	0.154	1.79E-03	4.53E-03
Citrate	4	0.242 ± 0.04	0.289 ± 0.03	0.256	2.50E-03	5.88E-03
Phenylalanine	3	0.066 ± 0.01	0.051 ± 0.01	-0.380	3.66E-03	7.79E-03
Tyrosine	3	0.201 ± 0.03	0.160 ± 0.03	-0.334	3.78E-03	7.79E-03
Homoarginine	2	0.201 ± 0.04	0.259 ± 0.05	0.363	7.95E-03	1.54E-02
Acetate	3	0.743 ± 0.33	1.141 ± 0.30	0.619	8.90E-03	1.63E-02
CONDITIONED MEDIA						
Acetate	3	0.615 ± 0.01	5.119 ± 0.24	3.057	6.51E-09	1.69E-07
Pyruvate	4	0.195 ± 0.004	0.392 ± 0.02	1.009	3.47E-07	4.51E-06
Glucose	4	0.967 ± 0.04	0.270 ± 0.06	-1.839	3.41E-06	2.96E-05
Alanine	4	0.368 ± 0.004	0.427 ± 0.01	0.214	1.53E-03	8.46E-03
Proline	4	0.079 ± 0.004	0.107 ± 0.01	0.447	1.63E-03	8.46E-03
Lactate	4	1.533 ± 0.03	1.771 ± 0.06	0.208	3.56E-03	0.015
Ethanol	4	14.230 ± 0.40	15.901 ± 0.41	0.160	0.017	0.062
Glycine	3	0.331 ± 0.001	0.305 ± 0.01	-0.116	0.027	0.088
Glutamate	4	0.450 ± 0.005	0.481 ± 0.01	0.097	0.048	0.127
Glutamine	4	0.806 ± 0.01	0.882 ± 0.03	0.131	0.049	0.127

Metabolic pathway and transcriptional alterations of Augusta cells are not consistent with canonical EMT

EMT has well defined alterations in metabolism associated with cell state transition, and are in fact dependent on these metabolic transitions to maintain an EMT phenotype[4]. Increased consumption of glucose and glutamine, as well as decreases in de novo fatty acid synthesis and mitochondrial dysfunction are known signatures of cancer cells that have undergone EMT.[3] To

assess the consistency of Augusta cells' metabolism with EMT, we performed ^{1}H nuclear magnetic resonance (NMR) metabolomics of cell extracts and conditioned media, as well as RNAseq transcriptomics on the same cell cultures, analyzed separately as well as integrated.

NMR profiling of cell extracts quantified 33 annotated metabolites (Supplemental Table 2.1) A two-tailed t-test identified 18 metabolites that were significantly altered at FDR p-value < 0.05 (Table 2.1). Additional ^{1}H NMR analysis of the conditioned media at cell harvesting showed a significant depletion of glucose, but increased levels of acetate, pyruvate, alanine, proline, and lactate in Augusta cell cultures compared to MCF-7 (Table 2.1, Figure 2.2B).



	REACTOME TGF BETA SIGNALING IN EMT	ALONSO METASTASIS EMT	HOLLERN EMT BREAST TUMOR
MCF7	12493.60	9754.01	-7908.36
Augusta	12440.37	9398.01	-7856.80
p-val	0.230	0.076	0.623

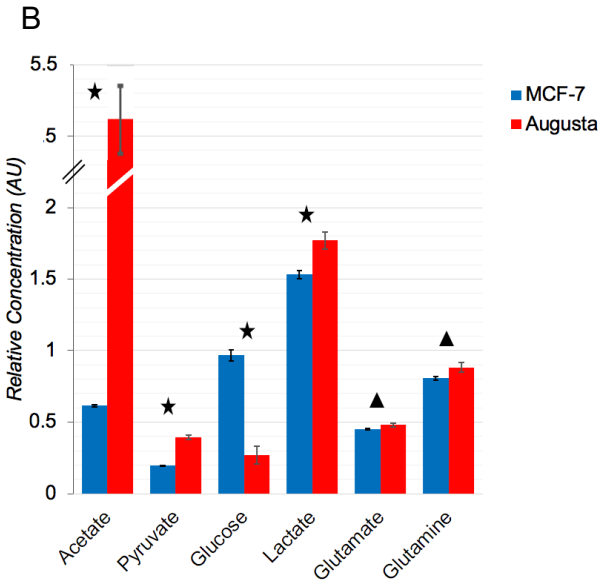


Figure 2.2 Metabolic and transcriptional patterns of Augusta cells inconsistent with EMT from MCF-7. A) ssGSEA scores for EMT related genesets. Error bars represent +/- standard error. N=3 for each cell line. B) Metabolite abundance in conditioned media as determined by NMR. N = 6 for each cell line. Stars indicate corrected p-value < 0.05. Triangle indicates raw p-value < 0.05. Error bars represent +/- standard error.

These results show increased consumption of glucose and production of lactate in the conditioned media, which are consistent with increased glycolysis. While increase in intracellular

citrate level could suggest changes in lipid metabolism, increased phospholipid precursors glycerophosphocholine and myo-inositol could be suggestive of increased lipid catabolism or synthesis. Additionally, no significant change in glutamine consumption was observed (Figure 2.2B). Finally, there was no significant change in redox metabolites such as glutathione and proline (which were quantified but not significant). Previous investigations into accumulation of ROS in mitochondria of Augusta cells were inconclusive (data not shown).

As an alternative multivariate approach to compare metabolite profiles, orthogonal partial least squares discriminant analysis (OPLS-DA) was able to both clearly separate the cell extracts and identified spectral features most important in modeling differences between the two cell lines (Supplemental Figure 2.2). Of note, features corresponding to branched-chain amino acids (BCAAs) and threonine were decreased Augusta cells, while increased intensity of features representing creatine phosphate, myo-inositol, serine and citrate were elevated in Augusta cells (Supplemental Figure 2.2) Overall, these metabolic results are only partly consistent with Augusta cells having an EMT-induced metabolic reprogramming.

To determine if Augusta cells had a transcriptional profile consistent with EMT, single-sample geneset enrichment analysis (ssGSEA) was performed for EMT-related genesets in the Molecular Signatures Database.[12,13] None of the EMT geneset scores examined appeared to be significantly altered between MCF-7 and Augusta cells (Figure 2.2A). While some individual genes of EMT were significantly altered, there is not overwhelming evidence of an EMT transcriptional profile based on these data. Differential gene expression analysis of all quantified genes resulted in identification of over 4600 genes significantly altered with a FDR adjusted p-value < 0.01. To identify pathways most affected, pathway enrichment analysis was performed using Paintomics 3.0.[14] Among all KEGG pathways, the PI3K-Akt pathway was the most enriched

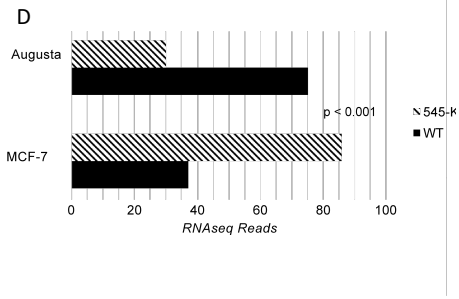
with differentially expressed genes (Figure 2.3A). This pathway contains 270 expressed genes quantified by RNAseq, 84 of which were differentially expressed with equal numbers being up and downregulated in Augusta (Figure 2.3A). ssGSEA analysis of all genes comprising the PI3K pathway in msigDB also shows a significant change in overall pathway expression in Augusta cells (Figure 2.3C).

A

KEGG Pathway	Total Genes	Sig. Genes	Combined *p*-value	BH FDR *p*-value
PI3K-Akt signaling pathway	270	84	6.19E-09	1.10E-06
Pathways in cancer	436	122	6.78E-09	1.10E-06
Axon guidance	160	57	2.79E-08	3.00E-06
Focal adhesion	172	56	3.47E-07	2.62E-05
Relaxin signaling pathway	107	40	4.87E-07	2.62E-05
Signaling pathways regulating pluripotency of stem cells	118	42	6.48E-07	2.99E-05
Proteoglycans in cancer	177	56	7.93E-07	3.20E-05
AGE-RAGE signaling pathway in diabetic complications	88	34	9.27E-07	3.33E-05
Thyroid hormone signaling pathway	104	38	2.36E-06	7.64E-05
Fluid shear stress and atherosclerosis	127	43	4.32E-06	1.27E-04

C

	MCF-7	Augusta	p-value
PI3K Combined	3232.99	4149.97	0.046
PI3K Upregulated	5226.11	3299.64	7.22E-04
PI3K Downregulated	1993.12	-850.33	4.59E-05



B

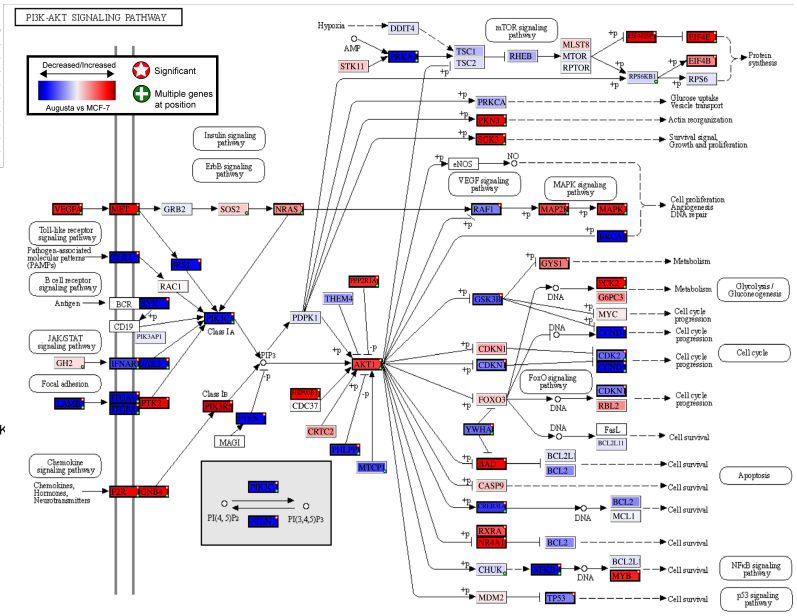


Figure 2.3: PI3K/Akt pathway, rather than EMT, is highly altered in Augusta cells. A) Top enriched KEGG pathways determined with Paintomics.[14] P-values are determined from Fischer-exact test . "Total" genes correspond to unique genes in dataset that are present in the indicated pathway. "Sig." genes are the subset of mapped features that were determined statistically previously (FDR <0.01). BH FDR values are p-values corrected for false discovery using Benjamini-Hochberg method. B) Pathway diagram for PI3K/Akt pathway produced by Paintomics. Rectangles indicate gene transcripts, those colored represent those quantified by RNAseq. Red rectangles represent transcripts increased in Augusta relative to MCF-7, blue decreased. Significantly differentially expressed genes (FDR < 0.01) are noted with red star icons. C) Mean ssGSEA scores for PI3k/Akt pathway genes according to KEGG geneset. P-values determined from two-tailed T-test D) Quantification of RNAseq read counts containing residue 545 of PIK3CA gene. P-value calculated by chi-squared test.

Augusta cells exhibit downregulated expression of mutant PIK3CA transcript and upregulate components of G-protein and PI3Kγ signaling

PIK3CA is a component of the PI3K complex and one of the most commonly mutated genes in breast cancers. Of these, a known oncogenic mutation occurs at residue 545. Sequencing reads mapped to this residue were examined and counted, showing a decrease in reads containing the mutant residue in Augusta cells (Figure 2.3D). The expression of this mutant transcript of PIK3CA is known to affect glutamine metabolism in colorectal cancer cells.[15]

ssGSEA was performed again for all PI3K-related genesets in the Molecular Signatures Database in attempt to narrow down the specific context of PI3K signaling that is altered in Augusta. The Reactome geneset "G beta gamma signaling through PI3K-gamma signaling" provided the lowest p-value when comparing enrichment scores between Augusta and MCF-7 (Figure 2.4A).[16] Among the 44 expressed genes in this pathway, 18 were differentially expressed, and 16 of those were increased in Augusta cells. These significantly increased genes corresponded to signaling kinases PAK1 and AKT1/2, an activator protein of PAK1, G-protein beta and gamma subunits, phospholipase C beta, effector proteins of CDC42, a regulatory subunit of PI3K-gamma, and inhibitor protein of BTK (Figure 2.4B).

A

Geneset Name	Mean ssGSEA score		Δ	p-val
	MCF-7	Augusta		
REACTOME G BETA GAMMA SIGNALLING THROUGH PI3KGAMMA	7166.28	7528.65	362.37	0.005
REACTOME PI3K EVENTS IN ERBB2 SIGNALING	8906.40	8509.86	-396.53	0.010
REACTOME NEGATIVE REGULATION OF THE PI3K AKT NETWORK	6452.00	6697.74	245.74	0.010
GSE19772 HCMV INFL VS HCMV INF MONOCYTES AND PI3K INHIBITION	-5348.59	-5143.92	204.67	0.021
GSE19772 CTRL VS HCMV INF MONOCYTES AND PI3K INHIBITION	1819.50	1603.07	-216.44	0.030
REACTOME ERYTHROPOIETIN ACTIVATES PHOSPHOINOSITIDE 3 KINASE PI3K	5136.27	4522.00	-614.27	0.031
REACTOME CD28 DEPENDENT PI3K AKT SIGNALING	8407.08	8155.96	-251.11	0.070
REACTOME PI3K EVENTS IN ERBB4 SIGNALING	3339.43	2571.38	-768.05	0.086

B

Gene	log2(FC)	log(FC) SE	p-value	FDR p-value	Description
PAK1	0.849	0.078	2.03E-27	1.14E-25	Signalling kinase
PIK3R5	2.159	0.266	4.90E-16	1.17E-14	PI3Kgamma subunit
ARHGEF6	1.981	0.246	7.55E-16	1.77E-14	Activator of PAK1
IBTK	0.718	0.109	4.74E-11	6.84E-10	Inhibitor of BTK
AKT1	0.346	0.055	2.16E-10	2.91E-09	Signalling kinase
GNG4	0.878	0.139	2.33E-10	3.13E-09	Ggamma subunit
GNB4	1.050	0.168	3.98E-10	5.18E-09	Gbeta subunit
GNB5	0.429	0.084	3.48E-07	2.91E-06	Gbeta subunit
RHOA	-0.288	0.057	5.03E-07	4.09E-06	GTPase
GNB1	0.230	0.050	4.42E-06	3.07E-05	Gbeta subunit
GNG7	0.669	0.166	5.80E-05	0.0003	Ggamma subunit
CDC42SE1	0.258	0.064	6.18E-05	0.0003	effector of cdc42
PLCB1	0.346	0.087	6.36E-05	0.0003	Phospholipase C beta
GNG12	-0.267	0.072	0.0002	0.0011	Ggamma subunit
AKT2	0.286	0.082	0.0005	0.0021	Signalling kinase
GNB1L	0.629	0.180	0.0005	0.0022	Gbeta subunit
CDC42BPB	0.190	0.059	0.0012	0.0048	effector of cdc42
GNG13	0.755	0.244	0.0020	0.0075	Ggamma subunit

C

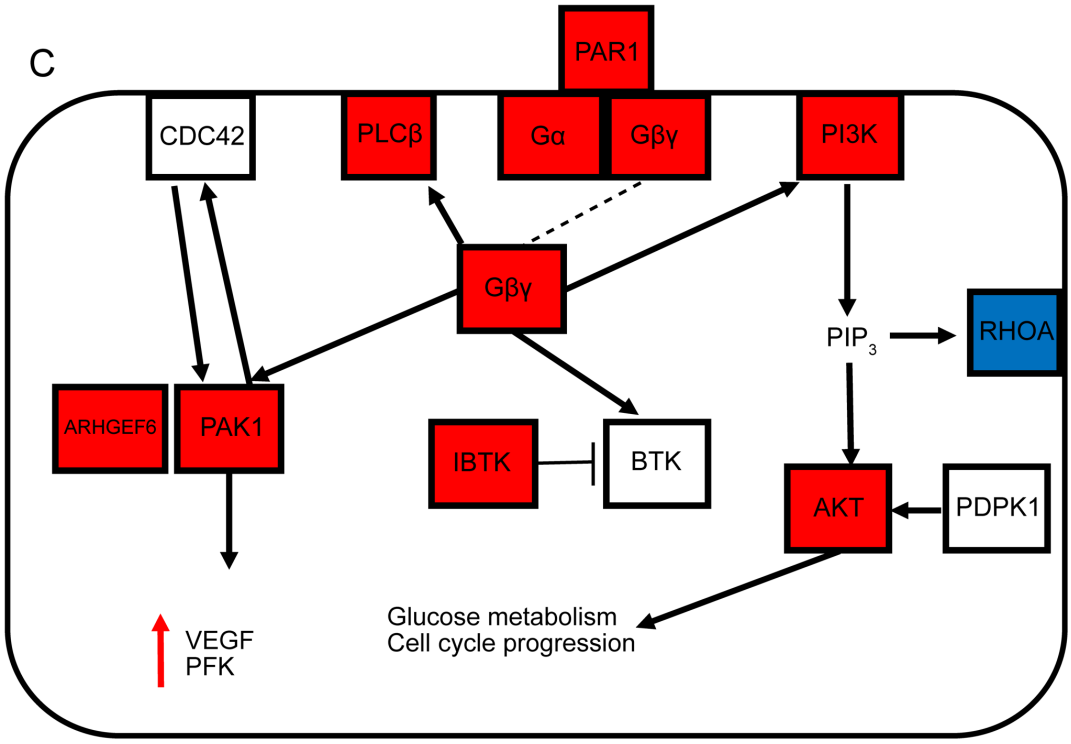


Figure 2.4 Augusta cells show increased expression of G beta-gamma signaling components that contribute to cell invasiveness and increased metabolic activity. A) Single-sample geneset enrichment results for top significant genesets involved in PI3K signaling N=3. B) Differentially expressed genes from G beta gamma signaling pathway. Fold change is Augusta relative to MCF-

7. Shaded rows indicate genes upregulated in Augusta. C) Schematic showing expression of G beta gamma signalling network, adapted from Reactome pathway diagram. Rectangles indicate gene products quantified by RNAseq, color indicates up or down regulation in Augusta, uncolored are not differentially expressed. Black arrows indicate activation reactions/interactions.

This pattern of upregulated components of G beta-gamma signaling points towards increased activation of PAK1, CDC42 and AKT. PAK1 and CDC42 signaling in particular have been implicated in cell migration and invasion, cytoskeletal reorganization and cell cycling.[17-19] PAK1 regulates expression of several targets including VEGF and PFKM, which are both upregulated in Augusta, providing more evidence of increased PAK1 signaling via G-protein activation.[20,21] Components of PI3K**γ** have also seen recent attention for its role in increasing cancer cell invasiveness and proliferation.[22-24] Together, these gene expression data provide evidence of cell morphological and proliferation changes driven by G protein beta-gamma activity ultimately leading to activation via PI3K**γ**.

Table 2.2 Top enriched KEGG Metabolism pathways determined with Paintomics[14]. P-values are determined from Fischer-exact test . "Mapped" genes and metabolites correspond to unique genes or metabolites in dataset that are present in the indicated pathway. "Sig." genes and metabolites are the subset of mapped species that were determined statistically significant within each -omic dataset previously. Q-values corrected for false discovery rate among KEGG Metabolism pathway enrichment tests using Benjamini-Hochberg method.

KEGG Metabolism Pathway	Total Mapped Genes	Sig. Genes	Total Mapped Metabolites	Sig. Metabolites	Combined *p*-value	Combined *q*-value
Valine, leucine and isoleucine degradation	46	19	3	3	0.0003	0.0149
Purine metabolism	156	48	3	2	0.0003	0.0149
Steroid biosynthesis	17	10			0.0006	0.0177
Fatty acid metabolism	46	17			0.0017	0.0376
Pyrimidine metabolism	95	30	2	0	0.0033	0.0591
Terpenoid backbone biosynthesis	21	9			0.0042	0.0601
Pentose and glucuronate interconversions	25	10			0.0047	0.0601
Glycolysis / Gluconeogenesis	56	19	2	2	0.0055	0.0612
Fatty acid degradation	38	13			0.0084	0.0797
Synthesis and degradation of ketone bodies	9	5			0.0090	0.0797
Glycerolipid metabolism	49	15			0.0138	0.1022

Metabolic alterations in Augusta cells suggests an increase in biosynthetic metabolism and lipid turnover

To better characterize the metabolic alterations occurring in Augusta cells, metabolic pathway enrichment was performed utilizing both transcriptomic and metabolomic data. Among KEGG metabolism pathways, branched-chain amino acid degradation is the most enriched (Table 2.2). Genes involved in the early, rate limiting steps of BCAA degradation (BCAT2, BCKDH), were significantly upregulated in Augusta cells, corresponding with the decrease in abundance of intracellular BCAAs themselves (Figure 2.5). Additionally, expression of genes comprising the LAT1 transporter (SLC7A2, SLC7A5), are also significantly upregulated in Augusta cells (Figure 2.5, Supplementary Data) suggesting an increased catabolism of BCAAs for lipid biosynthesis or TCA cycle cataplerosis.

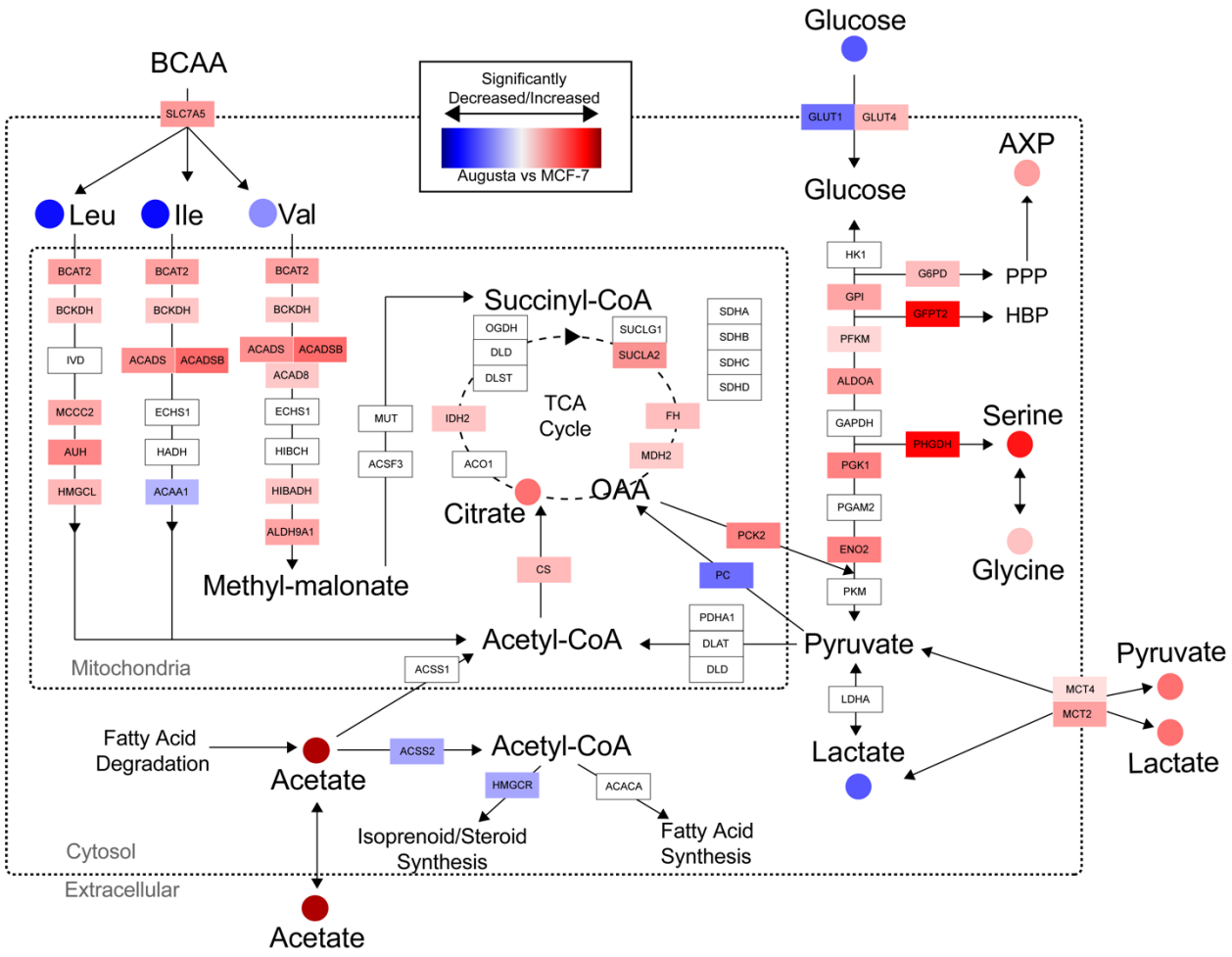


Figure 2.5 Model of metabolic alterations in Augusta cells compared to MCF-7 coinciding with differential PI3K**γ** signaling. Rectangles indicate gene transcripts quantified by RNAseq, circles represent metabolites quantified by NMR, other important metabolites/pathways labeled with text, arrows indicate biochemical reactions/pathways. Colored nodes indicate features statistically significant, and shade indicates relative log2 fold change. AXP= adenosine mono/di/tri-phosphates PPP = Pentose Phosphate Pathway, HBP = Hexoasmine biosynthetic pathway, OAA = Oxaloacetate

Another highly enriched metabolic pathway with multiple quantified metabolites is glycolysis/gluconeogenesis. The glucose transporter GLUT4 and approximately half of the genes involved in transformation of glucose to pyruvate are significantly upregulated (FDR p-value < 0.01) in Augusta, suggesting an increase in overall glucose uptake and utilization by Augusta cells consistent with profiles of conditioned media (Table 2.2, Figure 2.5). This is consistent with

enrichment of pyrimidine and purine metabolic pathways (Table 2.2), coupled with the significant increase of adenosine phosphate in Augusta cells (Table 2.1). Furthermore, the genes corresponding to branch points into the pentose phosphate, hexosamine, and serine biosynthetic pathways (G6PD, GFPT and PHGDH respectively) are significantly upregulated in Augusta cells (Figure 2.5), providing more evidence of increased glucose flux into biosynthesis of these metabolites.

The remaining enriched metabolic pathways are broadly related to lipid metabolism (Table 2.2). Most genes comprising the terpenoid and steroid biosynthetic pathways are significantly downregulated. This includes HMGCR, which encodes the enzyme responsible for the committed step in terpenoid and steroid biosynthesis from acetyl-CoA. (Figure 2.5, Supplementary Figures). Fatty acid metabolism/degradation, ketone body metabolism and glycerolipid metabolism are highly branched and complex in their patterns of gene expression, yet their enrichment in significantly altered genes indicate an increased level of phospholipid turnover and/or biosynthesis (Supplementary Figures). Significantly increased abundance of phospholipid precursors myo-inositol and glycerophosphocholine in cell extracts are also indicative of phospholipid biosynthesis/turnover occurring within Augusta cells (Table 2.1).

Together, these enrichment and expression patterns in the above metabolic pathways point to an increased glucose flux for biosynthesis, increased BCAA catabolism and increased lipid metabolism (but less steroid biosynthesis) in Augusta cells compared to the parental MCF-7 (Figure 2.5). These patterns are congruent with the differences of cell phenotype such as increased proliferation rate, which requires nucleotides and amino acids for replication, and altered cell membrane morphology reflecting changes in lipid composition. The implications of these will be discussed further in the following section.

2.4 Discussion

As described previously, metabolic reprogramming is a hallmark of EMT that has been characterized across several studies including increased glycolysis, oxidative stress, increased glutamine consumption, and lipid catabolism. Here we discuss some of the major metabolic pathways found to be altered between MCF-7 and Augusta cells and their implications in cancer invasive transition.

Metabolic alterations in Augusta cells not consistent with canonical EMT metabolic reprogramming, but consistent with other observations of invasive transition processes

Highly proliferative cells, and cancer cell particularly, exhibit high rates of aerobic glycolysis, or the Warburg effect. Current understanding points to several ways that the Warburg effect benefits cancer cells, including acidification of the tumor microenvironment for immunosuppression, and providing biosynthetic precursors.[25] The combination of our metabolomics and gene expression data support increase in glycolysis and glucose flux towards biosynthesis of amino acids, nucleotides, and lipids in Augusta cells (Figure 2.5). The true fate and utilization of glucose in our cell system could be strengthened with use of isotopically labeled glucose.

BCAAs have previously been implicated in potentiating cancer cell proliferation and tumor growth by stimulating mTOR signaling.[26,27] While these previous studies have shown that an *accumulation* of intracellular BCAAs increase in cancer cell proliferation, our results indicate a different role in Augusta cells. Our pathway level analysis suggests that BCAAs may be catabolized to substrates for TCA cycle anapleroisis and/or increased lipid biosynthesis. The latter scenario was recently demonstrated in pancreatic cancer cells, where lipid synthesis from BCAA catabolism was required for cell growth.[28] Another recent study by Shafei *et al*. demonstrated that

inhibition of BCAA metabolism decreased cell proliferation and migration in breast cancer cells.[29] Future analysis of non-polar metabolites of cells could provide more evidence of BCAA catabolism intermediates and lipid synthesis not detected by the methods used here. Consistent with this, the BCAA transporter (LAT1) is upregulated in Augusta cells (Figure 2.5); however, quantification of BCAA uptake were not assessed directly, and there was no significant change in abundance of BCAAs measured in conditioned media (Table 2.1). Future work should include time-course analysis of extracellular media during culture to establish relative rates of BCAA uptake between cell lines. Isotopic labeling experiments would be needed to conclusively determine the absolute consumption of BCAAs or other nutrients. As it stands, this study adds to evidence supporting the role of BCAAs as important carbon sources for increased proliferation in cancer cells.

Changes in lipid metabolism observed in Augusta cells match previously described changes in membrane lipid composition during EMT, characterized by a decrease in cholesterol content and increase in phospholipids.[30] Along with this altered membrane lipid composition, EMT/invasive transition is associated with increased expression of extracellular matrix proteases and receptors such as MMP15, MMP17, and PAR1 which are all upregulated in Augusta cells (Figure 2.4B, Supplemental Data).[31-33] EMT signatures quantified by ssGSEA did not show any consistent differences between Augusta and MCF-7 cells, suggesting that Augusta may be in a pre-EMT transition, or perhaps our cell culture system lacks the signals (i.e. TGF-β**)** to induce a strong canonical EMT transcriptional signature (Figure 2.2A).

Alternative PI3K signaling is potentially responsible for maintaining growth signaling through invasiveness transition

The PI3K pathway has been extensively studied in the context of cancer and disease.[34,35] PIK3CA is part of the Class IA PI3 kinases, activated primarily by receptor tyrosine kinases and Ras-related GTPases. Specific oncogenic mutations to this gene have been shown to have specific effects on reprogramming metabolism and malignancy.[15,36,37] Our RNAseq analysis showed a significant decrease in overall PIK3CA expression, while examination of the mapped reads shows a significant decrease in reads specifically containing the oncogenic E545K mutation (Figure 2.3D). Expression of this specific mutation has been shown to increase glutamine consumption in colorectal carcinoma cells, but our data did not provide evidence of significant alterations in this pathway.[15] PI3K**γ**, on the other hand, is of Class IB, which is activated primarily by G-protein coupled receptors through interaction with the beta-gamma subunit. Though typically active in immune cells, PI3K**γ** components have been identified to regulate invasiveness and growth in cancer cells. [17-19,38] The specific component found to be upregulated in Augusta cells is the regulatory subunit also known as p101, which was previously identified as having a potential oncogenic role in cancer cells.[22] The corresponding Class IB catalytic subunit, PIK3CG, was not quantified in our RNAseq data. Future work will verify it's presence by protein blotting and/or qPCR of these specific components to investigate further how activity of this pathway is differentially regulated in these cells.

Extracellular acidosis as an ontology for invasive transition of cell phenotype

One of the initial observations of the divergence of Augusta cells was a rapid and severe acidification of the culture media indicated by the phenol red in RPMI 1640 medium (Figure 2.1C). This acidification is confirmed by the significant increases in lactate and acetate measured in conditioned media (Table 2.2). Several recent studies have investigated the effect of an acidic extracellular environment on cancer metabolism and phenotype, including EMT.[39-41] Several

reported observations, including increased glycolytic capacity, pentose phosphate activity, altered lipid metabolism, and upregulation of some EMT markers such as TWIST1 and IDH2 were consistent here, suggesting that acidosis may contribute to the EMT-like transformation of Augusta cells. It is worth re-addressing that Augusta cells diverged after many passages in RPMI-1640 media. RMPI 1640 base formula contains twice the concentration of glucose compared to DMEM media typically used for MCF-7 cells. This higher glucose content could have allowed for increased glycolytic capacity and acidification of culture media, which may have contributed to this invasive transition through the mechanisms proposed in these recent studies.

In summary, we performed the multi-omic profiling of a diverged cell line to describe the metabolic and signaling pathways associated with changes in cell morphology and proliferation resembling an EMT phenotype. Increased glucose consumption, BCAA catabolism, and phospholipid metabolism are concomitant with a pre-EMT invasive cell phenotype stimulated by increased G beta-gamma signaling and alternative PI3K activation. These results add to the growing body of evidence implicating these metabolic and signaling pathways as connected hallmarks of invasive transition in cancer cells.

2.5 Acknowledgements

We would like to thank Dr. Huidong Shi for donating the Augusta cells. Special thanks to Yanfang Sun for performing initial cell culture and co-culture experiments that inspired this work, and contributing the confocal microscopy images in Supplemental Figure 2.1. RNA sequencing was performed at the Georgia Genomics and Bioinformatics Core. All NMR spectroscopy was performed in the NMR facility at the Complex Carbohydrate Research Center, with the assistance of Fariba Tayyari and John Gluska.

Funding

Work presented here was funded by NIH R01CA182093 and the Georgia Research Alliance.

2.6 Materials and Methods

Cell culture

MCF-7 cells were purchased from ATCC, Augusta cells were a gift from Huidong Shi at Augusta University. Cell lines were maintained in RPMI 1640 media (Corning Cellgro) supplemented with 10% fetal bovine serum (Alphabioregen) and 1% penecilin/streptomycin (Life Technologies). Cells used for all experiments were between passage number 10 and 20, and passaged at least twice after recovery of frozen stocks. Passages were performed at 1:5 or 1:6 split ratios. During culture of cells for metabolomics analysis, media were replaced every two days, which included a wash step of phosphate buffered saline (PBS) prior to addition of fresh media.

Cell proliferation assay

Freshly passaged cells were resuspended in growth medium. Serial dilutions ranging from 50 to 50,000 cells in 100 ml final volumes were prepared in 96-well clear-bottom microplates (Corning Costar). Four replicate samples were prepared in wells for each serial dilution of each cell line, along with control samples for media background fluorescence measurement, and incubated at 37°C at 5% CO_2 . Every 2 days, growth medium was aspirated and replaced with fresh growth medium. Microplate cultures were harvested by placing directly into -80°C freezer. Plates were harvested at 2, 4, 7, and 11 days after initial seeding and kept at -80°C until all plates were harvested. Microplates were then thawed at room temperature, and 100 mL of SYBR Green (Life Technologies) lysis buffer (at a concentration of .33 uL SYBR Green per mL) was thoroughly mixed into each well. Plates samples were incubated in darkness for one hour. Fluorescence was then recorded on Cytation5 plate reader (BioTek Instruments) with an excitation wavelength of 485 nm, and emission wavelength of 530 nm. Average media control values from each plate were

subtracted from cell sample measurements to remove background. Relative signal from first timepoint for each cell line was plotted against time for Figure 1C.

Metabolomics sample preparation

MCF-7 and Augusta cells were grown simultaneously in 100mm plates to ~80% confluency, with media being replaced every two days during culture. Media reagents used throughout each culture period were from same source and lot. At the end of the culture period, media was removed and flash frozen in liquid nitrogen. Cell monolayer was washed three times with PBS (-/-) . Approximately $1/8^{th}$ of cells were removed with a clean cell scraper, collected in PBS, and resuspended in RNA lysis buffer (Qiagen) for RNA extraction. 1 mL of ice cold extraction solvent (80:20 methanol:water, HPLC grade) was added to remaining cells on the plate. Remaining cells were scraped, collected in a 1.5 mL microfuge tube, and flash frozen in liquid nitrogen. For an extraction control, another tube was filled with the same volume of extraction solvent, flash frozen and stored with experimental samples. Samples were stored at -80°C until extraction. For extraction, cells were thawed on ice and then vortexed at maximum speed for 2 minutes to lyse cells and extract polar metabolites. Cell lysates were then centrifuged at 14,000 rpm (Eppendorf 5417C) for 15 min at 4°C. Supernatants were transferred to new microfuge tubes. Solvent was evaporated using vaccum centrifuge instrument (Labconco) until dry (~6 hrs). Dried extracts were stored at -80°C until just prior to NMR analysis. Dried extracts were resuspended in 200 uL of 100 mM phosphate buffer in D_2O with 1/3 mM deuterated sodium trimethylsilylpropanesulfonate (DSS-D6) (Cambridge Isotope Laboratories) adjusted to pH of 7.4. 20 uL from each reconstituted experimental sample was taken and combined to form an internal pooled control sample. 180 uL of samples were transferred into 3mm NMR tubes (Bruker Biospin).

Conditioned media was thawed on ice for 1 hour, clarified by centrifugation at max speed for 15 minutes and aliquoting supernatant to new tube. An equal amount from experimental sample combined to create internal pool. 100 mM DSS-D6 in D_2O was added to aliquot of clarified medium for a final concentration of 10% in a total volume of 200 uL, 180 uL of which was transferred to 3mm SampleJet NMR tubes (Bruker Biospin).

NMR data acquisition

All data were collected on an 800 MHz Bruker Avance III HD spectrometer with a 5mm TCI cryoprobe using TopSpin software (Bruker Biospin). NOESYPR1D spectra were collected on each sample in automation using ICON-NMR (Bruker Biospin). In addition, two-dimensional ^{13}C-^{1}H HSQC and ^{1}H-TOCSY spectra were collected for the pooled samples.

Metabolite Annotation

Two dimensional spectra were processed in NMRpipe.[42] Processed HSQC and TOCSY spectra were submitted to the COLMARm web server for database matching of HSQC peaks. [43] HSQC matches were manually reviewed with TOCSY and proton spectra to confirm the match. Annotations were assigned a confidence score based upon the levels of spectral data supporting the match as previously described.[44] For annotations with confidence score of 3 or higher, a single spectral bin was selected from 1D data to quantitate that compound.

NMR Data Processing

All one-dimensional spectra were manually phased and baseline corrected using TopSpin 3.5 (Bruker Biospin). Processed spectra were imported into MATLAB (Mathworks, Inc.). Using an in-house MATLAB toolbox (available at https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA), spectra were referenced to DSS, aligned, and normalized using the PQN algorithm.[45] Using an interactive script,

boundaries of all spectral features were manually selected and automatically integrated to produce a matrix of feature intensities. Cell extract feature intensities were batch corrected between independent experiments using the ComBat method implemented in Metaboanalyst web-server.[46,47] These batch corrected data were used to perform subsequent statistical analyses.

Metabolomics Statistical Analysis

The batch corrected intensities for all spectral features were quantile normalized and pareto scaled before input to create an orthogonalized partial least squares discriminant analysis (OPLS-DA) model in Metaboanalyst[46]. Features with high variable importance to projection (VIP) scores were matched to annotated metabolites. After assignment of all identifiable metabolites to a single spectral feature, univariate analysis was performed in MATLAB. False discovery rate correction was applied to all p-values using the Benjamini-Hochberg method.[48]

RNAseq sample preparation

Cells from six cultures of the two cell lines described were combined pairwise to produce 3 unique biological samples for each cell line. Total RNA was extracted from cells using AllPrep DNA/RNA Mini kit according to manufacturer's instructions (Qiagen). RNA concentration and quality was estimated with nanoDrop instrument (Thermo) (280/260 ratio ~2) and by agarose gel electrophoresis. Approximately 1 ug of total RNA was submitted to the Georgia Genomics and Bioinformatics Core for library preparation and Illumina sequencing of 75bp paired end reads.

Transcriptomics analysis

Sequencing reads were mapped to human genome assembly hg38 with HISAT2.[49,50] Uniquely mapped paired reads were input to Stringtie with the –e option for gene level quantification using the refSeq human genome annotation.[51] To filter the set of genes used for analysis to consistently expressed genes, genes were excluded that did not have an FPKM value

greater than zero for all samples AND greater than 1 for at least one sample. From this set of genes (13036), gene level read counts were used as input for DEseq2 using the GenePattern web-server.[52,53]

Pathway Analysis

Log_2 fold changes of all expressed genes (13036) and annotated metabolites (31) between Augusta and MCF-7 cells were input to Paintomics 3.0. "relevant features" used for the analysis were Metabolites that were significant with FDR p-value < 0.05 (Table 1), and all genes with FDR p-value < 0.01 determined by DESeq2.

Data Availability

Raw 1D and 2D NMR spectra, along with all associated scripts and MATLAB workflows available on request. Manually selected spectral features used for OPLS-DA, and full table of all annotated metabolite values and unknown features available in Supplementary Data. Raw RNAseq data available on request. RNAseq mapping and processing scripts available on request. Full DESeq2 results, including fold change values and significant genes used for pathway analysis available in Supplementary Data.

2.7 References

(1) Thiery, J. P.; Acloque, H.; Huang, R. Y.; Nieto, M. A. *Cell* **2009**, *139*, 871-890.

(2) Kalluri, R.; Weinberg, R. A. *J Clin Invest* **2009**, *119*, 1420-1428.

(3) Sciacovelli, M.; Frezza, C. *The FEBS Journal* **2017**, *284*, 3132-3144.

(4) Zhou, L.; Luo, M.; Cheng, L.-j.; Li, R.-n.; Liu, B.; Linghu, H. *Pathology - Res Pract* **2019**, *215*, 152681.

(5) Yang, J. H.; Kim, N. H.; Yun, J. S.; Cho, E. S.; Cha, Y. H.; Cho, S. B.; Lee, S.-H.; Cha, S. Y.; Kim, S.-Y.; Choi, J.; Nguyen, T.-T. M.; Park, S.; Kim, H. S.; Yook, J. I. *Life Sci Alliance* **2020**, *3*, e202000683.

(6) Hua, W.; Kostidis, S.; Mayboroda, O.; Giera, M.; Hornsveld, M.; Ten Dijke, P. *Metabolites* **2021**, *11*.

(7) Peinado, H.; Olmeda, D.; Cano, A. *Nat Rev Cancer* **2007**, *7*, 415-428.

(8) Lamouille, S.; Xu, J.; Derynck, R. *Nat Rev Mol Cell Biol* **2014**, *15*, 178-196.

(9) Ogrodzinski, M. P.; Teoh, S. T.; Lunt, S. Y. *Cancer Research* **2020**, canres.1666.2020.

(10) Halldorsson, S.; Rohatgi, N.; Magnusdottir, M.; Choudhary, K. S.; Gudjonsson, T.; Knutsen, E.; Barkovskaya, A.; Hilmarsdottir, B.; Perander, M.; Maelandsmo, G. M.; Gudmundsson, S.; Rolfsson, O. *Cancer Lett* **2017**, *396*, 117-129.

(11) Sun, X.; Wang, M.; Wang, M.; Yao, L.; Li, X.; Dong, H.; Li, M.; Li, X.; Liu, X.; Xu, Y. *Front Cell Dev Biol* **2020**, *8*, 655.

(12) Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J. P.; Tamayo, P. *Cell Systems* **2015**, *1*, 417-425.

(13) Barbie, D. A.; Tamayo, P.; Boehm, J. S.; Kim, S.; Moody, S. E.; Dunn, I. F.; Schinzel, A. C.; Sandy, P.; Meylan, E.; Scholl, C.; Fröhling, S.; Chan, E. M.; Sos, M. L.; Michel, K.; Mermel, C.; Silver, S. J.; Weir, B. A.; Reiling, J. H.; Sheng, Q.; Gupta, P. B., et al. *Nature* **2009**, *462*, 108-112.

(14) Hernández-de-Diego, R.; Tarazona, S.; Martínez-Mira, C.; Balzano-Nogueira, L.; Furió-Tarí, P.; Pappas, G. J.; Conesa, A. *Nucleic acids research* **2018**, *46*, W503-W509.

(15) Hao, Y.; Samuels, Y.; Li, Q.; Krokowski, D.; Guan, B.-J.; Wang, C.; Jin, Z.; Dong, B.; Cao, B.; Feng, X.; Xiang, M.; Xu, C.; Fink, S.; Meropol, N. J.; Xu, Y.; Conlon, R. A.; Markowitz, S.; Kinzler, K. W.; Velculescu, V. E.; Brunengraber, H., et al. *Nature Communications* **2016**, *7*, 11971.

(16) Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; Loney, F.; May, B.; Milacic, M.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Weiser, J.; Wu, G., et al. *Nucleic Acids Research* **2019**, *48*, D498-D503.

(17) Li, Z.; Hannigan, M.; Mo, Z.; Liu, B.; Lu, W.; Wu, Y.; Smrcka, A. V.; Wu, G.; Li, L.; Liu, M.; Huang, C.-K.; Wu, D. *Cell* **2003**, *114*, 215-227.

(18) Liu, F.; Li, X.; Wang, C.; Cai, X.; Du, Z.; Xu, H.; Li, F. *Int J Cancer* **2009**, *125*, 2511-2519.

(19) Balasenthil, S.; Sahin, A. A.; Barnes, C. J.; Wang, R.-A.; Pestell, R. G.; Vadlamudi, R. K.; Kumar, R. *J Biol Chem* **2004**, *279*, 1422-1428.

(20) Bagheri-Yarmand, R.; Vadlamudi, R. K.; Wang, R.-A.; Mendelsohn, J.; Kumar, R. *J Biol Chem* **2000**, *275*, 39451-39457.

(21) Singh, R. R.; Song, C.; Yang, Z.; Kumar, R. *J Biol Chem* **2005**, *280*, 18130-18137.

(22) Brazzatti, J. A.; Klingler-Hoffmann, M.; Haylock-Jacobs, S.; Harata-Lee, Y.; Niu, M.; Higgins, M. D.; Kochetkova, M.; Hoffmann, P.; McColl, S. R. *Oncogene* **2012**, *31*, 2350-2361.

(23) Turvey, M. E.; Klingler-Hoffmann, M.; Hoffmann, P.; McColl, S. R. *Immunol Cell Biol* **2015**, *93*, 735-743.

(24) Xie, Y.; Abel, P. W.; Kirui, J. K.; Deng, C.; Sharma, P.; Wolff, D. W.; Toews, M. L.; Tu, Y. *Biochem Pharmacol* **2013**, *85*, 1454-1462.

(25) Chen, Y. J.; Huang, X.; Mahieu, N. G.; Cho, K.; Schaefer, J.; Patti, G. J. *Biochemistry-us* **2014**, *53*, 4755-4757.

(26) Hattori, A.; Tsunoda, M.; Konuma, T.; Kobayashi, M.; Nagy, T.; Glushka, J.; Tayyari, F.; McSkimming, D.; Kannan, N.; Tojo, A.; Edison, A. S.; Ito, T. *Nature* **2017**, *545*, 500-504.

(27) Zhang, L.; Han, J. *Biochemical and Biophysical Research Communications* **2017**, *486*, 224-231.

(28) Lee, J.; Cho, Y.-R.; Kim, J.; Kim, J.; Nam, H.; Kim, S.; Son, J. *Exp Mol Medicine* **2019**, *51*, 146.

(29) Shafei, M. A.; Flemban, A.; Daly, C.; Kendrick, P.; White, P.; Dean, S.; Qualtrough, D.; Conway, M. E. *Breast Cancer-tokyo* **2020**, 1-16.

(30) Sampaio, J. L.; Gerl, M. J.; Klose, C.; Ejsing, C. S.; Beug, H.; Simons, K.; Shevchenko, A. *Proceedings of the National Academy of Sciences* **2011**, *108*, 1903-1907.

(31) Rizki, A.; Weaver, V. M.; Lee, S.-Y.; Rozenberg, G. I.; Chin, K.; Myers, C. A.; Bascom, J. L.; Mott, J. D.; Semeiks, J. R.; Grate, L. R.; Mian, S. I.; Borowsky, A. D.; Jensen, R. A.; Idowu, M. O.; Chen, F.; Chen, D. J.; Petersen, O. W.; Gray, J. W.; Bissell, M. J. *Cancer research* **2008**, *68*, 1378-1387.

(32) Tan, B.; Jaulin, A.; Bund, C.; Outilaft, H.; Wendling, C.; Chenard, M.-P.; Alpy, F.; Cicek, A. E.; Namer, I. J.; Tomasetto, C.; Dali-Youcef, N. *Cancers* **2020**, *12*, 2357.

(33) Boire, A.; Covic, L.; Agarwal, A.; Jacques, S.; Sherifi, S.; Kuliopulos, A. *Cell* **2005**, *120*, 303-313.

(34) Martini, M.; Santis, M.; Braccini, L.; Gulluni, F.; Hirsch, E. *Ann Med* **2014**, *46*, 372-383.

(35) Fruman, D. A.; Chiu, H.; Hopkins, B. D.; Bagrodia, S.; Cantley, L. C.; Abraham, R. T. *Cell* **2017**, *170*, 605-635.

(36) Ilic, N.; Birsoy, K.; Aguirre, A. J.; Kory, N.; Pacold, M. E.; Singh, S.; Moody, S. E.; DeAngelo, J. D.; Spardy, N. A.; Freinkman, E.; Weir, B. A.; Tsherniak, A.; Cowley, G. S.; Root, D. E.; Asara, J. M.; Vazquez, F.; Widlund, H. R.; Sabatini, D. M.; Hahn, W. C. *Proceedings of the National Academy of Sciences* **2017**, *114*, E3434-E3443.

(37) Samuels, Y.; Diaz, L. A.; Schmidt-Kittler, O.; Cummins, J. M.; DeLong, L.; Cheong, I.; Rago, C.; Huso, D. L.; Lengauer, C.; Kinzler, K. W.; Vogelstein, B.; Velculescu, V. E. *Cancer Cell* **2005**, *7*, 561-573.

(38) Andrews, S.; Stephens, L. R.; Hawkins, P. T. *Sci STKE* **2007**, *2007*, cm2.

(39) Gao, J.; Guo, Z.; Cheng, J.; Sun, B.; Yang, J.; Li, H.; Wu, S.; Dong, F.; Yan, X. *Scientific Reports* **2020**, *10*, 21967.

(40) Urbanelli, L.; Buratta, S.; Logozzi, M.; Mitro, N.; Sagini, K.; Raimo, R. D.; Caruso, D.; Fais, S.; Emiliani, C. *J Enzym Inhib Med Ch* **2020**, *35*, 963-973.

(41) Riemann, A.; Rauschner, M.; Gießelmann, M.; Reime, S.; Haupt, V.; Thews, O. *Neoplasia* **2019**, *21*, 450-458.

(42) Delaglio, F.; Grzesiek, S.; Vuister, G.; Zhu, G.; Pfeifer, J.; Bax, A. *Journal of Biomolecular NMR* **1995**, *6*, 277-293.

(43) Wang, C.; Timari, I.; Zhang, B.; Li, D. W.; Leggett, A.; Amer, A. O.; Bruschweiler-Li, L.; Kopec, R. E.; Bruschweiler, R. *J Proteome Res* **2020**, *19*, 1674-1683.

(44) Walejko, J. M.; Chelliah, A.; Keller-Wood, M.; Gregg, A.; Edison, A. S. *Metabolites* **2018**, *8*.

(45) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Analytical chemistry* **2006**, *78*, 4281-4290.

(46) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Res* **2015**, *43*, W251-257.

(47) Johnson, W. E.; Li, C.; Rabinovic, A. *Biostatistics* **2007**, *8*, 118-127.

(48) Benjamini, Y.; Hochberg, Y. *J Royal Statistical Soc Ser B Methodol* **1995**, *57*, 289-300.

(49) Pertea, M.; Kim, D.; Pertea, G. M.; Leek, J. T.; Salzberg, S. L. *Nat Protoc* **2016**, *11*, 1650-1667.

(50) Kim, D.; Langmead, B.; Salzberg, S. L. *Nat Methods* **2015**, *12*, 357-360.

(51) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res* **2007**, *35*, D61-65.

(52) Love, M. I.; Huber, W.; Anders, S. *Genome Biol* **2014**, *15*, 550.

(53) Reich, M.; Liefeld, T.; Gould, J.; Lerner, J.; Tamayo, P.; Mesirov, J. P. *Nat Genet* **2006**, *38*, 500-501.

CHAPTER 3

FUNCTIONAL ASSESSMENT OF HOMOZYGOUS ALDH18A1 VARIANTS REVEALS ALTERATIONS IN AMINO ACID AND ANTIOXIDANT METABOLISM[1]

[1]Colonna, M. B.; Moss, T.; Mokashi, S.; Srikanth, S.; Jones, J. R.; Foley, J. R.; Skinner, C.; Lichty, A.; Kocur, A.; Wood, T.; Stewart, T. M.; Casero, R. A.; Flanagan-Steet, H.; Edison, A. S.; Lyons, M. J.; Steet, R. *Hum Mol Genet* **2022**. Reprinted here with permission from publisher.

Foreword

Chapter 3 is reprinted with permission from Maxwell B Colonna, Tonya Moss, Sneha Mokashi, Sujata Srikanth, Julie R Jones, Jackson R Foley, Cindy Skinner, Angie Lichty, Anthony Kocur, Tim Wood, Tracy Murray Stewart, Robert A Casero Jr., Heather Flanagan-Steet, Arthur S Edison, Michael J Lyons, Richard Steet, Functional assessment of homozygous *ALDH18A1* variants reveals alterations in amino acid and antioxidant metabolism, *Human Molecular Genetics*, 2022;, ddac226, available at https://doi.org/10.1093/hmg/ddac226. The motivation and my inclusion on this work was driven by Richard Steet. My role on this project was to (i) design metabolomics study for both HEK and patient fibroblast cell cultures, (ii) prepare and perform metabolomics experiments on cell samples that were generated, (iii) analyze NMR data, (iv), interpret the results, (v) write relevant sections of the manuscript pertaining to the metabolomics experiments and data interpretation (vi), review and edit the manuscript. All other work were performed by co-authors, including clinical evaluation of patients, exosome sequencing, CRISPR-Cas genome editing of HEK cells, transfection, electrophoresis and western blotting of cell extracts, RNAseq analysis, culture of patient fibroblast cells and polyamine analysis. Research reported in this manuscript was funded by Greenwood Genetic Center, the National Science Foundation (NSF 1946970 and 1648035 to A.S.E. and M.C.); and the University of Pennsylvania Orphan Disease Center Million Dollar Bike Ride (MDBR-20-135-SRS) and the Chan Zuckerberg Initiative (to R.A.C./T.M.S.); and the National Cancer Institute (R01 CA235863 to R.A.C.).

Abstract

Mono- and biallelic variants in *ALDH18A1* cause a spectrum of human disorders associated with cutaneous and neurological findings that overlap with both cutis laxa and spastic paraplegia. *ALDH18A1* encodes the bifunctional enzyme pyrroline-5-carboxylate synthetase (P5CS) that plays a role in the *de novo* biosynthesis of proline and ornithine. Here we characterize a previously unreported homozygous *ALDH18A1* variant (p.Thr331Pro) in four affected probands from two unrelated families, and demonstrate broad-based alterations in amino acid and antioxidant metabolism. These four patients exhibit variable developmental delay, neurological deficits, and loose skin. Functional characterization of the p.Thr331Pro variant demonstrated a lack of any impact on the steady-state level of the P5CS monomer or mitochondrial localization of the enzyme, but reduced incorporation of the monomer into P5CS oligomers. Using an unlabeled NMR-based metabolomics approach in patient fibroblasts and *ALDH18A1*-null human embryonic kidney cells expressing the variant P5CS, we identified reduced abundance of glutamate and several metabolites derived from glutamate, including proline and glutathione. Biosynthesis of the polyamine putrescine, derived from ornithine, was also decreased in patient fibroblasts, highlighting the functional consequence on another metabolic pathway involved in antioxidant responses in the cell. RNA sequencing of patient fibroblasts revealed transcript abundance changes in several metabolic and ECM-related genes, adding further insight into pathogenic processes associated with impaired P5CS function. Together these findings shed new light on amino acid and antioxidant pathways associated with *ALDH18A1*-related disorders, and underscore the value of metabolomic and transcriptomic profiling to discover new pathways that impact disease pathogenesis.

Introduction

ALDH18A1 encodes delta-1-pyrroline-5-carboxylate synthase (P5CS), a bifunctional mitochondrial enzyme involved in the *de novo* biosynthesis of several amino acids and metabolites, including proline and ornithine [1,2]. This enzyme catalyzes the conversion of glutamate to pyrroline-5-carboxylate, which is subsequently used by either ornithine aminotransferase (OAT) to make ornithine [3], or pyrroline-5-carboxylate reductase (PYCR1) to make proline (**Figure 1A**). The metabolism of glutamate by P5CS connects this enzyme with both the urea cycle and TCA cycle as well as the biosynthesis of several different amino acids and polyamines. In addition, several of the downstream reaction products of P5CS, including the polyamines and proline, are involved in cellular antioxidant responses [4-8]. The P5CS substrate, glutamate, is also utilized by the cell for glutathione biosynthesis, further highlighting the importance of this metabolic pathway in establishing redox and antioxidant capacity.

ALDH18A1 was first associated with human disease following the identification of two patients with a neurocutaneous disorder who were shown to bear biallelic mutations in the gene [9]. *ALDH18A1* has also been implicated in several human cancers including breast cancer and melanoma [10-12]. The clinical complexity of *ALDH18A1*-related disorders expanded with the identification of variants in this gene associated with complicated forms of inherited spastic paraplegia [13,14]. Based on the analysis of this broadening clinical spectrum and the impact of known *ALDH18A1* mutations, *ALDH18A1*-related disorders in humans are thought to encompass at least two distinct syndromes - hereditary spastic paraplegia 9 (SPG9A and B) and cutis laxa 3 (ADCL3 and ARCL3A) [15]. Dominant and recessive mutations have now been identified for both disorders. The cutaneous phenotypes associated with cutis laxa may arise in part from impaired proline biosynthesis, which may limit the production of collagen and elastin in the skin and other

tissues. The mechanistic basis of the neurological manifestations is not fully understood. These impairments may relate to the altered availability of key metabolites for both the urea cycle and TCA cycle (reviewed in [15]). The connection to the urea cycle is reinforced by similarities in motor neuron degeneration evident in patients with arginase and ornithine transporter defects [14].

Monitoring the fate of isotopically-labeled glutamate in patient cells has been useful in establishing the pathogenicity of *ALDH18A1* variants [16,17]. This functional approach, however, does not provide insight regarding the complete landscape of metabolic changes that stem from impaired P5CS function. This limits our understanding of the factors and metabolites that contribute to disease pathogenesis. Here we describe a previously unreported missense homozygous variant of uncertain significance (p.Thr331Pro) in *ALDH18A1* present in four affected probands from two unrelated families. These patients show many clinical features consistent with SPG9B, including global developmental delay and hypotonia, but share other symptoms with ARCL3A such as loose, hyperelastic skin. Studies were performed in patient cells to functionally resolve this variant of uncertain significance, and explore the metabolic and transcriptomic profiles that accompany impaired P5CS function. NMR-based metabolomic studies in two different cell systems and transcriptomic profiling on patient fibroblasts uncovered the involvement of numerous metabolites and pathways not identified in prior studies, including extensive mobilization of glutamate-derived antioxidant pathways, and abnormal abundance of extracellular matrix (ECM) proteins and ECM-modifying enzymes. The implications of these findings with regard to antioxidant responses, metabolic adaptations and cutaneous disease pathogenesis in these patients are discussed.

Results

Clinical summary of patients

The pedigrees of the two families are shown in **Figure 3.1B**. The domain organization of the P5CS enzyme and the position of the homozygous variant found in all four affected patients is depicted in **Figure 3.1C**. The clinical and diagnostic findings in the four patients from two unrelated families is described below, and the clinical features of these patients compared to those with ARCL3A and SPG9B are summarized in **Table 3.1**.

Table 3.1 Summary of Clinical Findings in the Patients Compared to ARCL3A and SPG9B.

Clinical Features	**Family 1**	**Family 2**					
	P1	**P2**	**P3**	**P4**	**Total**	**ARCL3A**	**SPG9B**
Age of onset	6 mo	7 mo	7 mo	2 mo	2-7 mo	0-6 mo	~5 yo
Cutis laxa	-	-	-	-	0/4	+++	-
Skin hyperelasticity	-	+	+	+	3/4	-	-
Global developmental delay	+	+	+	+	4/4	+++	+++
Hypotonia	+	+	+	+	4/4	+++	+
Joint hypermobility	-	+	+	+	3/4	+++	+
Short stature	-	+	+	+	3/4	+++	++
Microcephaly	+	+	+	-	3/4	+++	++
Visible veins	+	+	+	+	4/4	++	-

Cataracts	-	+	-	-	1/4	++	+
Hypertonia/spasticity	+	-	-	-	1/4	+	+++
Corpus callosum hypogenesis	+	+	?	?	2/2	+++	+

Family 1

Patient 1 (P1) is a 10-year-old female who initially presented for a genetics evaluation at 16 months old for developmental delay. She was first noted to have developmental delay at 6 months old. She sat at 16 months, started taking steps with assistance at 5 years, and walking at 7 years. She started saying mama when she was 2 years old. She now has a few signs but no other words. She has a history of pyloric stenosis, gastroesophageal reflux, constipation, gastric dysmotility, obstructive sleep apnea, and bruxism. Her neurologic findings include a generalized resting tremor, hypotonia, periods of hyperventilation, head titubation, and spastic diplegia. She wears splints on her arms to prevent self-injurious behavior. At her most recent physical exam, she had a weight of 28.1 kg (Z = -0.90), height of 128.6 cm (Z = -1.47), and head circumference of 48.2 cm (Z = -3.24). She was noted to have a broad face, deep-set eyes, synophrys, high nasal bridge, mild 2-3 toe syndactyly, central hypotonia, and lower limb spasticity. Her skin exam was normal without loose skin. A brain MRI revealed prominent ventricles and subarachnoid spaces with thin corpus callosum and incomplete myelination. She has had normal urine organic acids, plasma amino acids, and carnitine levels. Chromosomal microarray analysis (NCBI 36/hg18) did not identify any copy number variants but did reveal a 20-megabase block of homozygosity on chromosome 10q22.3q24.1. She has had normal methylation analysis for Angelman syndrome, sequencing and Multiplex Ligation-dependent Probe Amplification (MLPA) analysis of the

MECP2 gene, myotonic dystrophy testing, high-resolution chromosome analysis, and sequencing of the *TCF4*, *UBE3A*, *SLC9A6*, *CDKL5*, and *FOXG1* genes. A syndromic autism panel revealed a variant in *SHANK3* which was maternally inherited and not felt to be clinically significant. Whole exome sequencing revealed a variant in *CACNA1G* which was maternally inherited and not felt to be clinically relevant. The (NM_002860.3:c.991G>T; p.Thr331Pro) variant in *ALDH18A1* that was initially reported as a variant of uncertain significance but has subsequently been re-classified as likely pathogenic. This amino acid alteration is located in the C-lobe of the glutamate-5-kinase domain of the enzyme with no other known disease-causing variants in close proximity. With the new functional evidence provided below, the *ALDH18A1* alteration has been reclassified as likely pathogenic based on ACMG criteria. The evidence used for this reclassification includes 1) it is present at a very low frequency in the public SNP databases (PM2) 2) parental testing indicated all four parents are heterozygous carriers of this alteration which confirms homozygosity in the probands (PM3), and 3) functional analyses using patient fibroblasts and *ALDH18A1* knockout HEK293 cells expressing the p.Thr331Pro variant enzyme demonstrated alterations in metabolites consistent with a pathogenic effect for this variant (PS3).

Family 2

Patient 2 (P2) is a 6-year-old female who initially presented for a genetics evaluation at 10 months old due to developmental delay. She has a history of poor weight gain, gastroesophageal reflux, cataract, and hypotonia. She babbles but does not have any words. She can pull herself to a stand but does not walk. At her most recent physical exam, she had a weight of 12.1 kg (Z = -5.18), height of 96.2 cm (Z = -4.31), and head circumference of 45 cm (Z = -4.59). She has full cheeks, epicanthal folds, supraorbital fullness, deep-set eyes, full lips, open mouth, small jaw, mild

tooth decay, prominent antihelices, decreased muscle mass, hypotonia, joint hypermobility, mild skin hyperelasticity, and visible veins on her face and trunk. Brain MRI at 7 months revealed thin corpus callosum and delayed myelination. She has had normal conventional karyotyping, whole genome chromosomal microarray analysis, plasma amino acids, acylcarnitine profile, carnitine levels, urine organic acids, and myotonic dystrophy testing. Whole exome sequencing revealed a heterozygous variant in *GJC2* but no evidence of an alteration of the other allele on deletion/duplication testing. She also was found to have a heterozygous secondary finding in *PALB2* associated with increased cancer risk. Exome sequencing also revealed a homozygous NM_002860.3: c.991A>C (p.Thr331Pro) variant in *ALDH18A1* that was initially reported as a variant of uncertain significance but has subsequently been re-classified as likely pathogenic.

Patient 3 (P3) is a 4-year-old male who initially presented for a genetics evaluation at 16 months old for developmental delay and poor growth. He has a history of hypotonia, gastroesophageal reflux, and feeding difficulties. He is able to prop sit but is nonambulatory and nonverbal. At his most recent physical exam, he had a weight of 9.25 kg (Z = -7.06), height of 85.3 cm (Z = -4.62), and head circumference of 46.1 cm (Z = -3.48). He has full cheeks, epicanthal folds, supraorbital fullness, unfurled superior helices, full lips, open mouth, small jaw, tooth decay, decreased muscle mass, hypotonia, joint hypermobility, mild skin hyperelasticity, and veins visible on his face and trunk (see **Supplemental File 3.1**). He had normal plasma amino acids, acylcarnitine profile, carnitine levels, and urine organic acids. Targeted gene testing revealed homozygosity for the p.Thr331Pro variant in *ALDH18A1* previously identified in his older sister.

Patient 4 (P4) is a 16-month-old female who initially presented for a genetics evaluation at 4 months old due to the family history of two siblings with homozygous *ALDH18A1* variants. She has a history of hypotonia, poor weight gain, constipation, gastroesophageal reflux, and

eczema. She can prop sit but is not walking. She has no words. At her most recent physical exam, she had a weight of 7.1 kg (Z = -2.80), height of 69.7 cm (Z = -3.29), and head circumference of 44.2 cm (Z = -1.43). She has full cheeks, epicanthal folds, supraorbital fullness, deep-set eyes, open mouth, protruding tongue, hypotonia, joint hypermobility, mild skin hyperelasticity, and visible veins on her face and trunk. Her only genetic testing has been targeted testing which revealed homozygosity for the p.Thr331Pro variant in *ALDH18A1*.

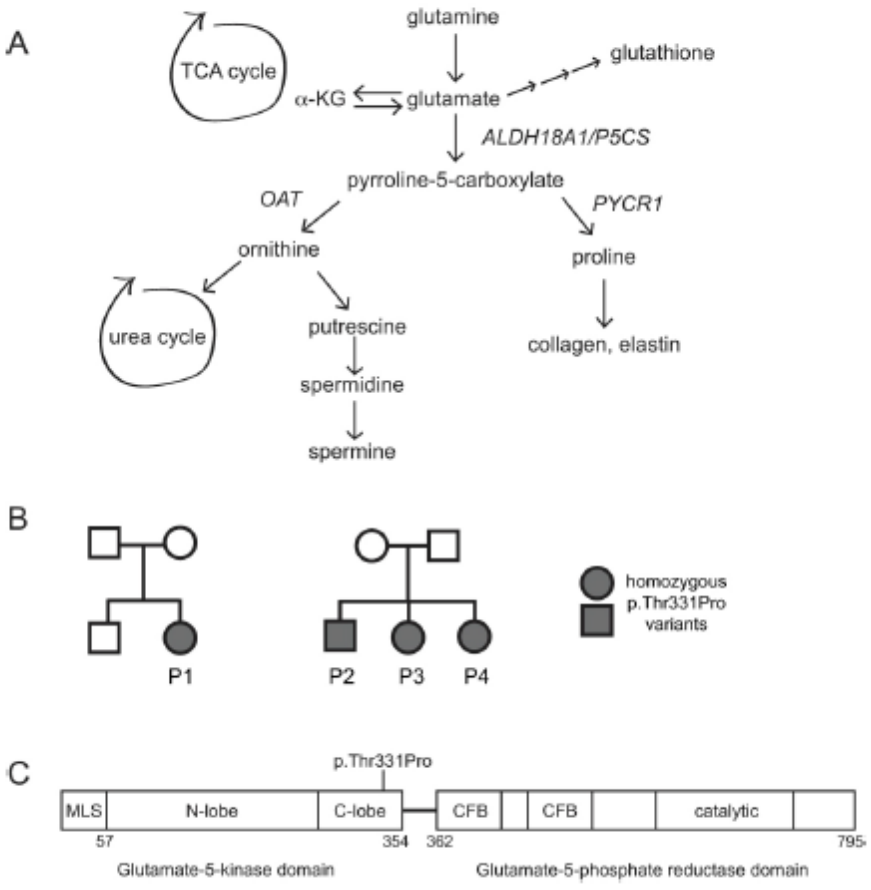


Figure 3.1 Overview of metabolic pathways connected to P5CS and the domain organization of the enzyme. (A) Schematic of the different metabolic pathways that are connected to the products of the P5CS-mediated reaction, and utilize the P5CS substrate, glutamate; a-KG, a-ketoglutarate; *OAT*, ornithine transaminase; *PYCR1*, pyrroline-5-carboxylate reductase (B) Pedigrees of the two families and four affected individuals included in this study; (C) Domain organization of the P5CS enzyme and location of the homozygous p.Thr331Pro variant found in all four patients; MLS, mitochondrial localization signal; CFB, co-factor binding domain.

Functional characterization of p.Thr331Pro variant in HEK293 and patient-derived cells

To explore the functional impact of this variant, we generated an *ALDH18A1*-knockout HEK293 cell line using CRISPR-Cas9 gene editing, and used this cell line to re-express either the WT or p.Thr331Pro variant-bearing P5CS enzyme. As shown in **Figure 3.2A**, the *ALDH18A1*-KO cells show undetectable levels of the P5CS protein. Upon expression of either the WT or variant enzyme, no major differences in the abundance or electrophoretic mobility of the P5CS enzyme were noted when transfection efficiencies were equivalent, indicating that the variant protein is not inherently unstable or subject to proteolysis. This same set of cells were stained with an anti-P5CS antibody to examine whether the mitochondrial localization of the enzyme was altered (**Figure 3.2B**). The P5CS enzyme localizes to structures consistent with mitochondria in WT HEK293 but is undetectable in the *ALDH18A1* KO cells. The WT and p.Thr331Pro P5CS were shown to localize to the mitochondria when expressed in the KO HEK293 cells, as determined by co-staining with a TOM20 antibody. We did not note any substantial differences in the localization of the p.Thr331Pro P5CS in these immunostains. Skin fibroblasts from P2 were obtained and used to investigate the level and localization of endogenous P5CS. We showed that the steady-state level of the P5CS monomer in the patient cells is comparable to control fibroblasts (**Figure 3.2C**). Prior studies on *ALDH18A1* variants have demonstrated effects on the ability of the enzyme to form homo-oligomers as the basis for compromised function [16,18]. This possibility was tested in the primary fibroblasts by resolving WT and patient cell lysates on a non-denaturing native gel. A reproducible decrease in a reactive band at 170kDa, presumably the dimeric form of the enzyme, was detected in the patient fibroblasts, suggesting that the p.Thr331Pro variant may

alter its ability to incorporate into P5CS oligomers (**Figure 3.2D**). Quantification of the ratio of P5CS dimer/monomer, with the latter abundance determined following parallel resolution on a denaturing SDS-PAGE gel, showed a significant 40% reduction in this ratio in patient cells, supporting reduced formation of oligomers as the primary mechanism for impaired P5CS function.

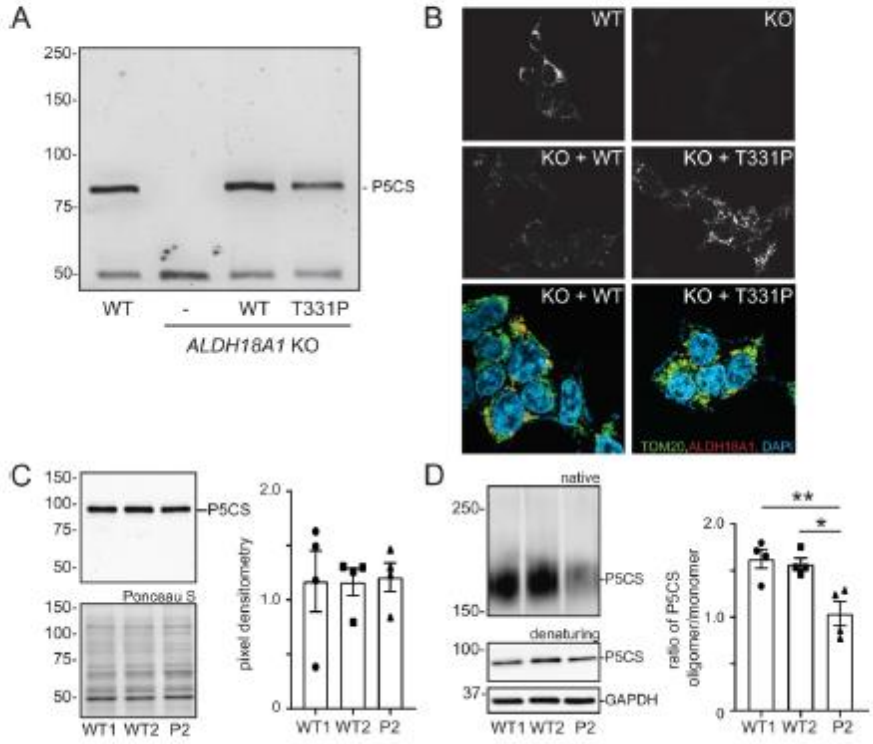


Figure 3.2 Functional analysis of the p.Thr331Pro variant in HEK293 cells and patient fibroblasts demonstrate normal steady-state levels and mitochondrial localization but impaired incorporation of the P5CS monomer into homo-oligomeric structures. (A) Representative Western blot of the P5CS enzyme in WT, *ALDH18A1*-null cells, and *ALDH18A1*-null cells transfected with either WT or p.Thr331Pro *ALDH18A1* DNA; (B) Immunostaining of P5CS and TOM20 in the KO HEK293 cells transfected with WT and p.Thr331Pro *ALDH18A1* DNA; overall transfection efficiencies ranged from 50-60% across the different runs and were roughly equivalent with WT or p.Thr331Pro *ALDH18A1* DNA in individual experiments; (C) Representative Western blot of the P5CS monomer in WT and patient fibroblasts and quantification of relative abundance from four independent experiments; (D) Native gel electrophoresis and Western blot analysis of P5CS in WTnd patient fibroblast lysates. The ratio of P5CS oligomer to monomer was quantified from four independent experiments, and average values plotted. A one-way ANOVA was performed to determine statistical significance. P value smaller than 0.05 is considered statistically significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns = not significant.

NMR metabolomics reveals broad-based adaptations in amino acid metabolism and antioxidant biosynthesis

We next sought to understand global metabolic alterations that arise from impaired P5CS function. Using both the HEK293 cell system and WT and patient fibroblasts, ^{1}H NMR-based metabolomics were performed on cells grown under standard culture conditions. Features in the resulting spectra were annotated and quantified, culminating in 26 quantified and annotated metabolites. We observed broad and significant alterations in the abundance of several different metabolites when comparing the WT and *ALDH18A1* KO HEK293 cells, highlighting the adaptations these cells undergo in order to maintain viability and essential components of P5CS-related metabolic pathways. To what extent such differences also arise due to karyotype differences in the WT and KO cells, as opposed to the underlying metabolic defect, is not known. Nonetheless, we were able to compare the relative abundance of several key metabolites in the *ALDH18A1* KO cells expressing either the WT or p.Thr331Pro P5CS enzyme. These findings show that while the profiles of both enzyme expressing cells are more similar to each other than the KO, the p.Thr331Pro enzyme restored metabolite levels less effectively in comparison to the WT enzyme (**Figure 3.3; Supplemental File 3.2**). This is evident for metabolites such as proline, glycine, succinate, UDP-GlcNAc, and glutamine. There were several examples where the opposite trend was noted, including glutamate and aspartate. In total, there were 6 metabolites that showed significant changes in abundance between p.Thr331Pro and WT enzyme expressing cells, including proline, glycine, succinate, NAD+, phosphorylcholine, and creatine (**Figure 3.3**).

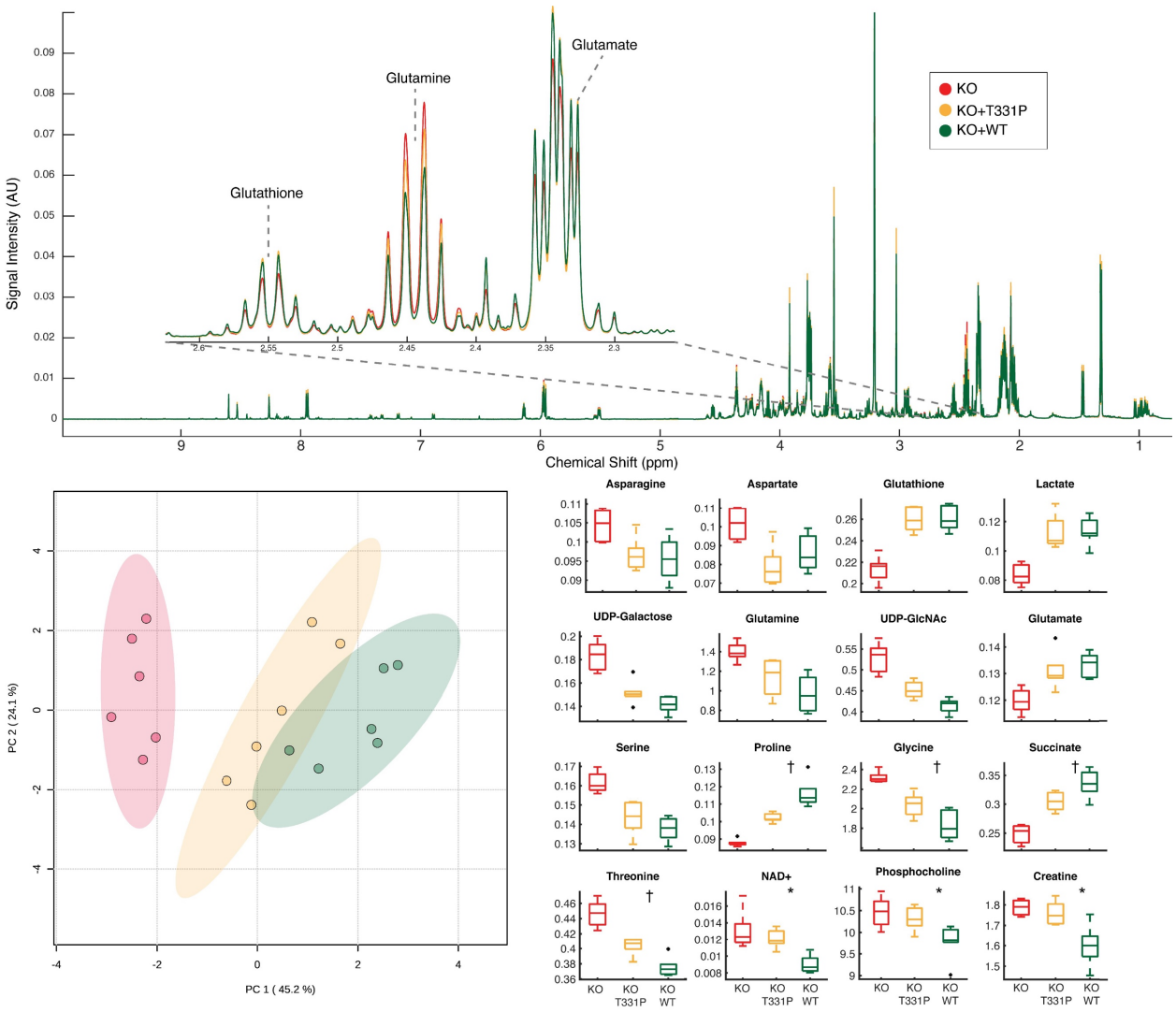


Figure 3.3 Expression of p.Thr331Pro variant enzyme partially restores metabolic profile of ALDH18A1 knockout compared to expression of WT enzyme. (A) Average ^{1}H NMR spectra of HEK 293 cell extracts. KO - CRISPR knockout of ALDH18A1, KO + p.Thr331Pro (labeled T331P in the figure) - knockout cells exogenously expressing T331P variant of ALDH18A1 enzyme. KO + WT - knockout cells exogenously expressing wild type ALDH18A1 enzyme; overall transfection efficiencies ranged from 50-60% across the different runs and were largely equivalent with WT or p.Thr331Pro *ALDH18A1* DNA; (B) Principal component analysis scores plot of KO, KO+T331P, and KO+WT cells. Integrated spectral features used for analysis; (C) Box and whisker plots of metabolites annotated from NMR spectra found significantly different by one-way ANOVA (FDR adjusted p-value < 0.05). All comparisons significant between KO and KO+T331P, KO and KO+WT unless otherwise noted. † indicates additional significant comparison between KO+T331P and KO+WT. * indicates significant comparisons between KO and KO+WT, and KO+T331P and KO+WT only.

A parallel analysis was undertaken using WT and patient fibroblasts. Reduction in the abundance of both proline and glutathione were observed in the patient cells (**Figure 3.4A**). The decrease in proline levels was consistent with the analysis in HEK293 cell system but diminished glutathione levels has not been previously shown in *ALDH18A1* patient cells. Another striking observation was the overall reduction in the level of several amino acids in the patient cells, likely indicating the need for these cells to consume other amino acids for cataplerosis and/or conversion to glutamate in order to replenish the TCA cycle and glutamate-derived metabolites. Pathway enrichment analysis of metabolites significantly altered between WT and p.Thr331Pro enzyme-expressing cells in both the HEK293 system and primary fibroblasts showed consistency in the pathways being impacted. These include arginine and proline metabolism, glycine and serine metabolism, and glutamate metabolism (**Figure 3.4B**). Metabolite changes that were consistent between the two cell systems suggest that these trends stem directly from impaired P5CS function. For metabolites where the changes varied between the two cell systems, we believe the patient fibroblasts may more accurately reflect the biochemistry of the underlying disease, as these are untransfected primary cells with residual P5CS function that have not undergone clonal proliferation or selection.

A

Metabolite	HEK293		Primary Fibroblast	
	T331P vs WT	*p*-val	P2 vs WT	*p*-val
Proline	-0.627	0.019	-0.165	0.224
Glutathione	-0.025	0.488	-0.199	0.038
Arginine	-0.013	0.880	0.195	0.026
Lysine	0.008	0.938	0.083	0.093
myo Inositol	0.012	0.820	0.544	0.030
Serine	0.024	0.727	-0.047	0.726
Lactic acid	0.026	0.794	0.099	0.294
Glutamate	0.030	0.622	-0.127	0.013
Phosphorylcholine	0.073	0.018	-0.164	0.186
Threonine	0.103	9.89E-03	-0.412	7.17E-06
Alanine	0.105	0.374	-0.108	0.491
Valine	0.114	0.477	-0.102	8.94E-03
Creatine	0.137	0.006	-0.084	0.611
Leucine	0.179	0.352	-0.106	0.101
Phenylalanine	0.179	0.282	-0.197	4.77E-04
Tyrosine	0.191	0.387	-0.135	0.078
Isoleucine	0.191	0.380	-0.165	2.28E-04
Glycine	0.202	0.046	-0.378	0.012
Glutamine	0.209	0.170	-0.540	0.202

B

HEK293	FDR	Primary Fibroblast	FDR
Arginine and Proline Metabolism	0.000174	Glutathione Metabolism	0.0495
Carnitine Synthesis	0.00811	Glycine and Serine Metabolism	0.0495
Glutamate Metabolism	0.0532	Glutamate Metabolism	0.173
Ketone Body Metabolism	0.0532	Alanine Metabolism	0.173
Glycine and Serine Metabolism	0.063	Arginine and Proline Metabolism	0.173

Figure 3.4 Patient fibroblasts show unique metabolite changes but consistent alterations of proline and glutamyl metabolism. (A) Table of fold change and *p*-values for all annotated metabolites common to both HEK 293 and primary fibroblast cell extracts. Red indicates relative increase in T331P variant enzyme expressing cells, blue a decrease. Bold metabolite names indicate those with consistent trend between cell systems. Green values indicate significant change (*p*-value < 0.05 by two-tailed T-test) for that comparison. These significant metabolites were used (but not exclusively) as input for metabolite set over representation analysis in (B); (B) Metabolite Set Enrichment Analysis results of significant metabolites for each cell system. Top 5 significant metabolite sets from SMPDB shown for each cell system, along with FDR *p*-values. List of all annotated metabolites with raw *p*-value < 0.05 between p.Thr331Pro and WT enzyme expressing cells was used for over representation analysis in each cell system.

Analysis of polyamine levels demonstrates a reduction in putrescine in patient fibroblasts

The apparent increased utilization of glutathione in the patient cells suggests a possible response to oxidative stress caused by mitochondrial dysfunction or metabolic toxicity. To look at other antioxidant molecules derived from glutamate metabolism, polyamines including putrescine, spermidine and spermine were analyzed in WT and patient fibroblasts (**Figure**

3.5). None of these molecules were detected or identified in the NMR-based metabolomics but can be quantified using high-performance liquid chromatography [19]. *De novo* production of these polyamines is derived almost exclusively from ornithine in cells. These results showed a robust reduction in putrescine levels (and total polyamine levels), with spermidine and spermine not significantly impacted. It is possible the patient fibroblasts prioritize the synthesis of spermidine and spermine as part of a broader antioxidant response, leading to overutilization of their precursor putrescine.

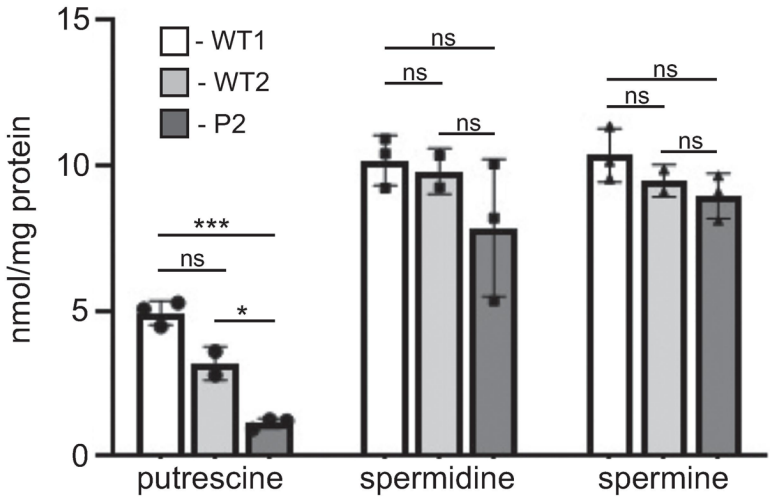
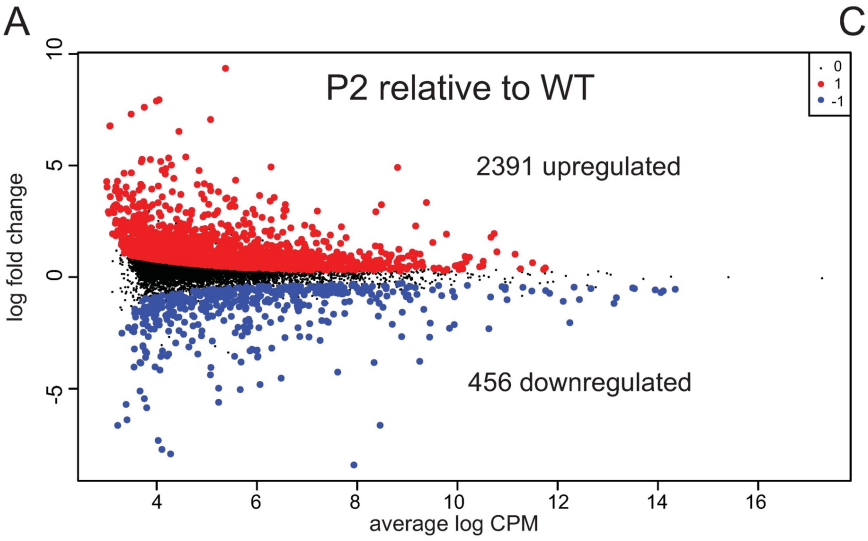


Figure 3.5 Reduced levels of the polyamine putrescine in patient fibroblasts. The abundance of putrescine, spermidine and spermine in the two WT and patient fibroblast lines was determined using mass spectrometry in three independent experiments. A one-way ANOVA was performed to determine statistical significance; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns = not significant.

Altered transcript abundance of several ECM and metabolic genes in ALDH18A1 patient fibroblasts

RNA sequencing was performed on WT and patient fibroblasts to examine alterations in gene expression in response to impaired P5CS activity. The volcano plot in **Figure 3.6A** shows upregulation of 2391 genes and downregulation of 456 genes in the patient fibroblasts (full dataset is shown in **Supplemental File 3.3**). Among the dysregulated genes in the patient fibroblasts,

several genes involved in glutathione metabolism, including GPX1, GSTT2B and GSS, were noted. Transcript abundance of the glutamate transporter SLC7A11 was decreased by 1.7-log fold in the patient cells. If this reduction corresponds to lower amounts of the SLC7A11 transporter, these cells may have a limited capacity to resupply glutamate for the necessary metabolic reactions, including proline and glutathione biosynthesis. In addition to differential expression of several ECM-related genes, numerous genes involved in cell signaling were also altered in patient cells. This is highlighted by the Reactome and Panther pathway analysis which shows overrepresentation of several genes in the MAPK, SEMA3 and RUNX2 pathways, as well as genes involved in cholesterol metabolism (**Figure 3.6B** and **Supplemental File 3.3**). Among the most significantly altered transcripts in the patient cells were the metalloproteinase MMP3 (up 4.9-log fold) and the matrix protein, COMP (cartilage oligomeric matrix protein; down 6.6-log fold) (**Figure 3.6C**). Elevation in the transcript and protein abundance of several MMPs has been observed in cutis laxa patient fibroblasts [20,21]. Western blot analysis was performed on WT and patient fibroblast lysates revealing that levels for both proteins correspond to their relative transcript abundance (**Figure 3.6D** and **E**). The reduction in COMP levels in the patient cells is accompanied by altered transcript abundance of other ECM genes, as well as several ECM-related growth factors (TGFB3R) and enzyme inhibitors (TIMP3) (**Figure 3.6B**). Together, these findings point to altered transcript abundance in response to impaired P5CS function, and alterations in ECM-associated proteins that appear to be consistent across human disorders with cutis laxa.



C

Gene name	log fold change	FDR
MMP3	4.91	3.04E-176
COMP	-6.64	1.67E-172
PI16	-8.42	5.17E-156
SVIL	4.93	1.36E-100
KRT19	-4.81	4.46E-76
CLDN11	3.25	6.96E-76
AKAP12	2.96	4.44E-75
F2R	3.57	4.74E-73
LRRN4CL	3.40	2.10E-71
RNVU1-28	-4.26	1.43E-68
COLEC12	3.01	1.10E-67
SULF1	-3.77	2.98E-66
EFEMP1	-3.83	1.54E-65
COL6A1	1.95	8.91E-65
CCBE1	3.25	3.21E-63
WNT16	4.34	7.13E-63
COL6A2	1.79	4.35E-62
BGN	-2.06	1.53E-60
OLFML2A	7.06	2.58E-58
MIR604	9.36	5.62E-58
TGFBR3	3.00	5.62E-58
TIMP3	-2.31	7.46E-52
MALL	-5.04	1.26E-49
AHNAK2	2.26	2.49E-49
PRUNE2	3.47	3.60E-49

B

Reactome pathways	PANTHER pathways
Olfactory Signaling Pathway (R-HSA-381753)	Integrin signalling pathway (P00034)
Immune System (R-HSA-168256)	CCKR signaling map (P06959)
G alpha (s) signalling events (R-HSA-418555)	Angiogenesis (P00005)
Signaling by Receptor Tyrosine Kinases (R-HSA-9006934)	Huntington disease (P00029)
Membrane Trafficking (R-HSA-199991)	Ras Pathway (P04393)
Metabolism of proteins (R-HSA-392499)	Gonadotropin-releasing hormone receptor pathway (P06664)
Neutrophil degranulation (R-HSA-6798695)	EGF receptor signaling pathway (P00018)
Extracellular matrix organization (R-HSA-1474244)	VEGF signaling pathway (P00056)
Metabolism (R-HSA-1430728)	PDGF signaling pathway (P00047)
Innate Immune System (R-HSA-168249)	Cytoskeletal regulation by Rho GTPase (P00016)

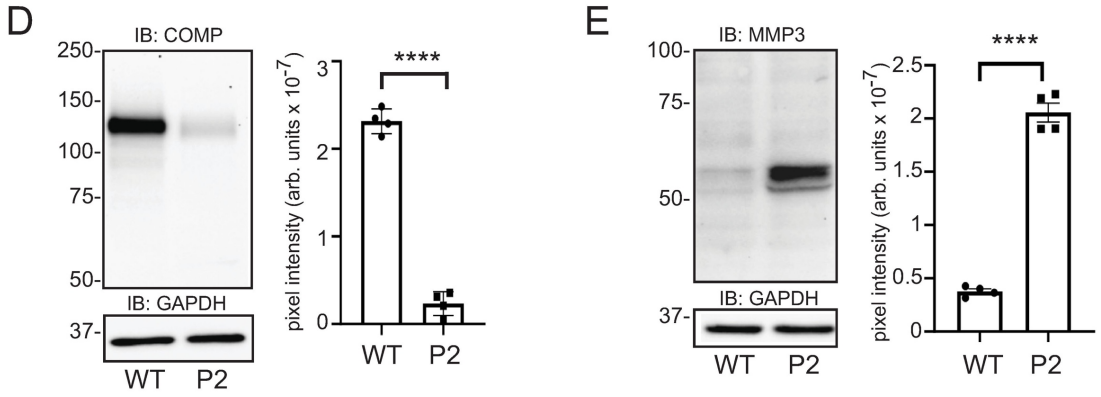


Figure 3.6 RNA sequencing of WT and patient fibroblasts reveals altered transcript abundance of multiple genes. (A) Volcano plot of the differentially expressed genes (2391 upregulated and 456 downregulated compared to WT control fibroblasts); data represents analysis of three independent biological replicates for each cell line; (B) Overrepresented genes/pathways from the Reactome and Panther GO analyses; (C) Table of the most significant transcript abundance changes in patient cells compared to WT controls; (D) Western blot analysis of COMP protein and quantification of relative levels (n=4; statistical significance determined using a Student t-test, **** $P < 0.0001$); (E) Western blot analysis of MMP3 protein and quantification of relative levels (n=4; statistical significance determined using a Student t-test, **** $P < 0.0001$). GAPDH is shown as a loading control.

Discussion

We describe a previously unreported homozygous missense variant in the *ALDH18A1* gene that is present in four affected patients from two unrelated families. The amino acid change resides in the glutamate-5-kinase domain of the P5CS enzyme, in a region where few other variants have been identified. The clinical phenotype of these patients appears to be intermediate between ARCL3 and SPG9B with regard to severity and specificity. They share many neurological features associated with SPG9B, including hypotonia, but also had loose skin characteristic of ARCL3A. Interestingly, P1 has hypertonia and spasticity consistent with SPG9B but the other patients do not. P1 also did not have any evidence of loose skin whereas all three of the affected patients in Family 2 have this cutaneous phenotype. The intermediate nature of the clinical manifestations highlights the complexity of *ALDH18A1*-related disease, and the competition between different metabolic pathways for common substrates. The current functional evidence strongly supports pathogenicity of the p.Thr331Pro variant, despite no reduction in the steady-state level of the monomeric P5CS enzyme nor impaired mitochondrial localization. A defect in homo-oligomer formation was observed using native gel electrophoresis, suggesting the p.Thr331Pro variant may affect complex formation like other reported variants [16,18]. The pathogenicity of this variant and evidence for impaired P5CS function is most clearly supported by the metabolomics data showing reduction in products downstream of P5CS including proline and the ornithine-derived polyamine, putrescine. The new insights into the pathogenic mechanisms and the global alterations in metabolism associated with *ALDH18A1*-related disorders are summarized in **Figure 3.7** and discussed below.

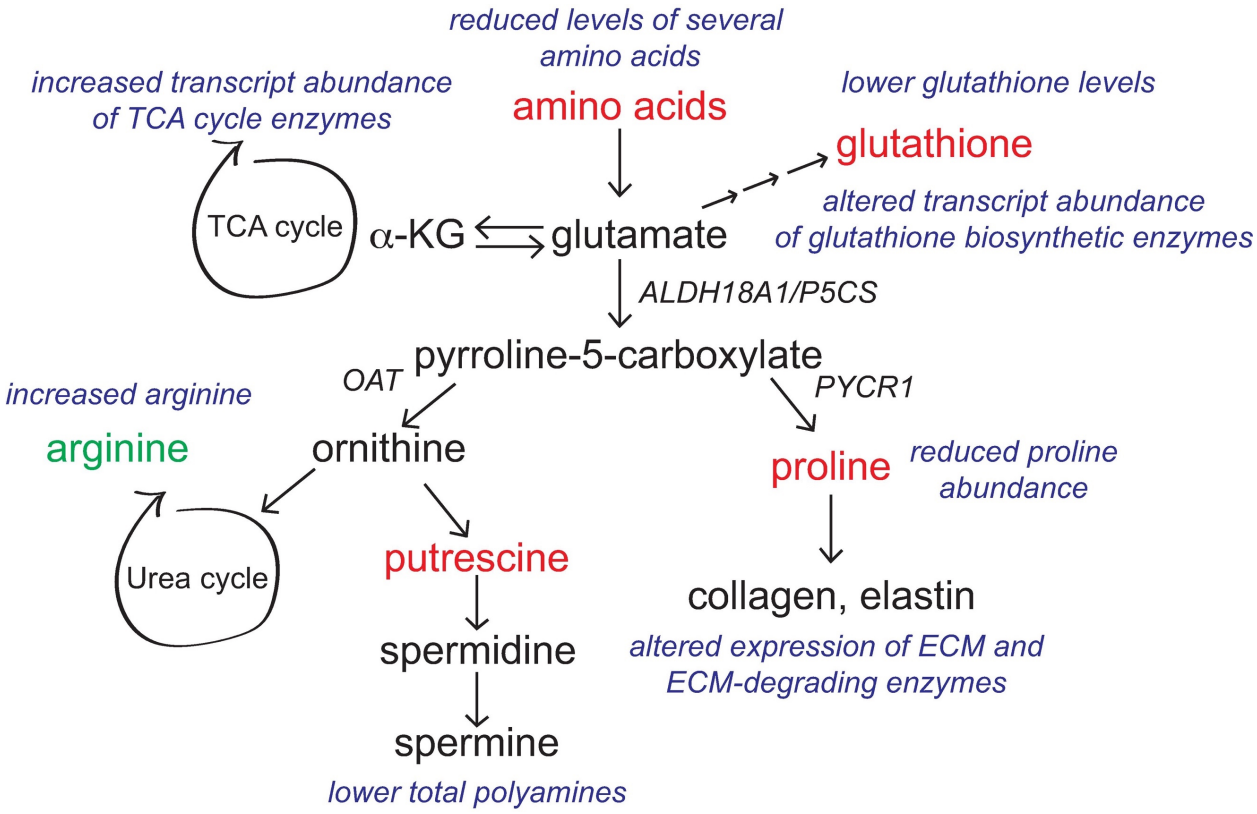


Figure 3.7 Summary of metabolic and transcriptomic findings in the context of P5CS-mediated metabolism. Metabolomic findings point to the possible involvement of glutathione and polyamines in the context of *ALDH18A1*-related disorders. As both are relevant to antioxidant responses, their reduction (e.g. total polyamines) and /or overutilization (e.g. glutathione) may indicate the presence of oxidative stress. The reduced levels of several amino acids, and increased transcript abundance of TCA cycle genes, in the patient cells may indicate the need to replenish glutamate pools depleted by the increased metabolic demand for this amino acid. In addition to the lower levels of proline (which can directly impact the production of ECM proteins), altered transcript and protein abundance of the matrix metalloproteinase, MMP3, and the matrix component, COMP, were noted in patient fibroblasts, highlighting broader alterations in ECM homeostasis associated with impaired P5CS function. Lastly, the increase in arginine levels indicates that this amino acid may not be efficiently converted to ornithine by arginase, thus exacerbating the reduction in ornithine-derived polyamines.

The NMR-based metabolomics in two different cell systems uncovered the involvement of several pathways and metabolites that had not be appreciated in other studies. Most striking, and consistent between the two cell systems, was the reduction in glutathione levels in cells with impaired P5CS function. Glutathione is synthesized in a multistep pathway beginning with the

conversion of glutamate to γ-glutamyl-cysteine by the enzyme, glutamate-cysteine ligase. The reduction in glutathione levels may be interpreted as an overutilization of this pathway, further stressed by the need for glutamate in the biosynthesis of both ornithine and proline. Levels of several amino acids were significantly reduced in the patient fibroblasts, possibly reflecting their conversion to glutamate in response to increased need for this key substrate. It will be of interest to explore whether the depletion of these amino acids might trigger other consequences such as mTOR inhibition or an imbalance in proteostasis. We are currently exploring whether the depletion of these amino acids, such as the nutrient sensor leucine, may result in the upregulation of autophagic pathways [22-24]. The inhibition of mTOR-related pathways and increased autophagy is intriguing as it could also increase mitophagy and the degradation of mitochondria or other organelles [25,26]. Reduction in the level of other amino acids such as arginine could indirectly impact the urea cycle leading to additional insults to sensitive tissues such as neurons. The upregulation of MMP3 and other matrix-degrading enzymes may be part of the same response to amino acid deprivation and increased need for glutamate to support key metabolic processes.

In the HEK293 expression system, significant changes in metabolites such as lactate, succinate, NAD+, and creatine suggest broad differences in the energy and/or redox status of the cells based on which enzyme is expressed. Similarly, transcriptome analysis of patient fibroblasts shows an enrichment of TCA cycle genes that are upregulated, and significant downregulation of several biosynthetic off-ramp transcripts such as PCK and GOT1. These observations imply that the metabolic adaptations required to compensate for P5CS dysfunction are energy intensive, and require increased TCA cycle activity to maintain energy requirements within the cell. Alternatively, increased expression of TCA cycle genes could be generating intermediates and cofactors to be used for glutamate/glutathione biosynthesis, such as 2-oxoglutarate and NADPH.

In addition, phospholipid precursors myo-inositol and phosphorylcholine were observed to be significantly altered in patient cells and HEK cells, respectively. These, in addition to the enrichment of differentially expressed genes in several lipid pathways, suggest broad changes in lipid metabolism. These again could be indicative of metabolic adaptations to balance the energy needs of the cell, or reflective of altered membrane lipid composition. We believe some caution is warranted in the interpretation of trends from the metabolomics results from the HEK293 cells in light of the large differences in some metabolites in the WT vs. KO cells.

The reduction and/or overutilization of glutathione may indicate the presence of oxidative stress. Preliminary experiments performed to identify oxidative stress in the patient cells were inconclusive, although it is possible that the antioxidant responses, including the production of glutathione, is sufficient to prevent detectable oxidative stress in fibroblasts. Whether such antioxidant responses are inadequate in other cell types such as neurons remains to be determined. We speculate that the increased need for glutamate in the brain as a precursor for neurotransmitters such as GABA may put further stress on P5CS-related metabolism, creating mitochondrial stress and production of reactive oxygen species that impact survival and function of neurons and other sensitive cell types. Aside from the requirement for proline in the biosynthesis of many ECM proteins, this amino acid is also an antioxidant [4]. Thus, under conditions where oxidative stress is abundant, competition for proline may create an imbalance that can contribute to pathogenesis.

Polyamines are critical cellular components required for a multitude of functions, including the regulation of receptor ion channels in the CNS [27,28]. As such, their concentrations are strictly controlled to maintain homeostasis [29]. The altered polyamine levels detected in the *ALDH18A1*-variant patient fibroblasts may therefore implicate a role for polyamines in the patient phenotype.

The recent identification of a growing number of patients with gene variants affecting polyamine metabolic enzymes has created a family of polyamine-related disorders. Although yet to be fully characterized, these rare, neurodevelopmental syndromes, including Snyder-Robinson syndrome [30,31] and Bachmann-Bupp syndrome [32], share certain components of the clinical phenotype, including developmental delay and hypotonia [33,34]. As each of these syndromes, including the *ALDH18A1* variant, causes a different perturbance in the individual polyamine pools, it will be important to decipher the contributions of each to the associated phenotypes. Additionally, the combined total amount of polyamines may be important, particularly regarding antioxidant capacity, as polyamines can protect against oxidative damage [35]. To our knowledge, this is the first report of a genetic variant causing an overall reduction in polyamine concentration.

The transcriptome analysis of patient fibroblasts revealed altered abundance of numerous ECM-related transcripts. Of note, transcript abundance of the cartilage oligomeric matrix protein (COMP) was reduced nearly 7-log fold in the patient cells, while the matrix-degrading enzyme, MMP-3, had a 6-log fold elevation in transcript abundance. Analysis of enriched pathways among the differentially expressed genes also uncovered several metabolic pathways, including cholesterol metabolism, that had not been previously noted in the context of *ALDH18A1*-related disease. We confirmed that altered transcript abundance of both COMP and MMP3 correlate to the same changes in protein level by Western blot. It is possible that impaired P5CS function and *de novo* proline biosynthesis cause a shift in the types of ECM proteins made by the patient fibroblasts. Collectively, the combination of transcriptomics and metabolomics in patient cells has uncovered multiple new pathways and processes that are sensitive to impaired P5CS, setting the stage for a deeper investigation of how these altered pathways relate to the tissue pathogenesis and phenotypic specificity in patients with *ALDH18A1* variants.

Materials and Methods

Exome sequencing

DNA libraries were prepared from genomic DNA isolated from the proband fibroblasts, using the Agilent SureSelect^{XT} Clinical Research Exome v2 capture kit (Agilent Technologies, Santa Clara, CA). Briefly, DNA was fragmented using the Covaris ME220 system (Covaris, Woburn, MA), and fragments of 150-200 bp were selected using AMPure XP beads (Beckman Coulter, Brea, CA). Fragments were subsequently end-repaired, adenylated at the 3' end, ligated to sequencing adaptors, and then PCR-amplified using the SureSelect^{XT} Library Preparation kit (Agilent Technologies) with DNA being purified using AMPure XP beads after each of these steps. 750 ng of each DNA library was used for hybrdization and capture with the SureSelect^{XT} Clinical Research Exome v2 probes (Agilent Technologies). Captured fragments were amplified by PCR and purified. The quality of the enriched libraries was evaluated using a D1000 Tape on the TapeStation 4200 (Agilent Technologies). Libraries were quantified on a Victor Nivo Fluorometer (PerkinElmer, Waltham, MA) using a Quant-IT Broad Range kit (Life Technologies, Carlsbad, CA), and separate libraries were pooled and sequenced using an Illumina NovaSeq 6000^{TM} Sequencing System at 157x coverage (Illumina Inc., San Diego, CA) per the manufacturer's protocol.

The Agilent SureSelect^{XT} Clinical Research Exome V2 kit was used to target known disease-associated exonic regions of the genome (coding sequences and splice junctions of known protein-coding genes associated with disease, as well as an exomic backbone) using genomic DNA isolated from peripheral blood samples. The targeted regions were sequenced using the Illumina NovaSeq^{TM} 6000 System with 150 bp paired-end reads. Using Illumina DRAGEN Bio-IT Platform$^{®}$ software, the DNA sequence was aligned and compared to the human genome build 19

(hg19/NCBI build 37). The emedgene® software was used to filter and analyze sequence variants identified in the patient and compare them to the sequences of affected and unaffected family members. Sanger sequencing was subsequently performed to confirm the variant of interest.

Antibodies

The rabbit anti-P5CS polyclonal antibody was purchased from Novus Biologicals (cat# NBP1-83324; Centennial, CO, USA), the mouse anti-TOM20 monoclonal antibody was obtained from Santa Cruz (cat# sc-17764; Dallas, TX, USA) the goat anti-COMP polyclonal antibody was from R&Systems (cat# AF3134; Minneapolis, MN, USA), the mouse anti-MMP3 monoclonal antibody was obtained from R&D Systems (cat# MAB513; Minneapolis, MN, USA), and the HRP-conjugated, rabbit anti-GAPDH monoclonal antibody was purchased from Cell Signaling Technology (cat# 3683; Danvers, MA, USA).

Generation of ALDH18A1-KO HEK293T cells and cell culture

ALDH18A1-knockout HEK293 cells were generated by CRISPR-Cas9 editing using a parental HEK293T cell line (Canopy Biosciences, Saint Louis, MO). The clonal knockout line was sequenced and shown to bear a 1bp insertion in the first allele and a 1bp deletion in the second allele. Both the parental and KO HEK293 cell lines were shown to have abnormal numbers of certain chromosomes, and a translocation of chromosomes 4 and 10, following conventional karyotyping.

Allele 1: 1bp deletion of exon 2

CGCAGCATGTTGAGTCAAGTTTACCGCTG -
GGGTTCCAGCCCTTCAACCAACATCTTCTG

CGCAGCATGTTGAGTCAAGTTTACCGCTGTGGGTTCCAGCCCTTCAACCAACATCTT
CTG

Allele 2: 1bp insertion of exon 2

CGCAGCATGTTGAGTCAAGTTTACCGCTGGTGGGTTCCAGCCCTTCAACCAACATCT

TCT

CGCAGCATGTTGAGTCAAGTTTACCGCTG -

TGGGTTCCAGCCCTTCAACCAACATCTTCT

Patient fibroblasts were obtained from P2 following consent. All cell lines were maintained in DMEM with 10% fetal bovine serum and penicillin/streptomycin in a humidified incubator with 5% CO_2. Monolayers were subcultured every 2-3 days using trypsin/EDTA.

Transfection, Gel Electrophoresis and Western blotting

Transfections in the HEK293 cell system were performed using Lipofectamine Plus. Transfection efficiency was determined for each experiment by immunostaining for the P5CS enzyme and calculating the percentage of cells with detectable P5CS enzyme. For denaturing Western blot analysis, 20 µg of the cell lysates prepared in RIPA buffer were separated on a 10% SDS-PAGE gel. For the native gel analysis, cell lysates were prepared using buffer containing 1% NP-40, passed five times through a 21-gauge needle syringe and separated on a 6% PAGE gel using running buffer without SDS. Resolved protein was transferred to 0.45µm pore nitrocellulose at 110 volts for 2 h at 4°C. The membrane was rinsed and blocked membranes were blocked with 5% milk/TBST for 1 hour at room temperature. Blots were incubated overnight with antibodies 1:1000 at 4°C, washed and then incubated with HRP conjugated anti-rabbit (1:2000) for 1 hour before washing and developing with SuperSignal West Pico PLUS ECL reagent (ThermoFisher #34577). All images were captured on the Bio-Rad ChemiDoc MP Imaging System (Bio-Rad #12003154). Analysis was done with Image Lab Software (Bio-Rad #1709690,

ver5.2.1). For the native gel, NativeMark unstained protein standards (Fisher; cat# LC0725) were used.

NMR-based metabolomics studies

Each plate of adherent HEK293 cells was washed with PBS, scraped, and pellets flash-frozen in liquid nitrogen. After transport on dry-ice and just prior to extraction, ice-cold 80% methanol/water extraction solvent was added to pellets. Primary fibroblast culture plates were flash-frozen and transported on dry ice. Due to lower cell density, fibroblasts were scraped from three 10 cm culture plates in ice-cold extraction solvent and combined to create each experimental replicate prior to extraction. Aqueous metabolites were extracted by vortexing/lysing cell pellets in the extraction solvent, pelleting cell debris by centrifugation, and collecting the supernatant. 10% of the supernatant was taken from each sample to form an internal pooled sample. The solvent was then evaporated to produce dried extracts using a CentriVap Benchtop Vacuum Concentrator (Labconco, Kansas City, MO, USA). Extracts were reconstituted in a deuterium oxide phosphate buffer (pH 7.4) and kept at 4°C before data acquisition.

^{1}H NMR spectra were acquired on all samples using noesypr1d pulse sequence on a 600 MHz Bruker Avance III HD spectrometer using a 5mm TCI cryoprobe. Additional ^{1}H-^{13}C heteronuclear single quantum correlation (1H-13C HSQC) and ^{1}H-^{1}H total correlation spectroscopy (^{1}H-^{1}H TOCSY) spectra were acquired on the internal pooled sample and used for metabolite annotation with COLMARm [36]. One dimensional 1H NMR spectra were referenced, solvent/water regions removed, and normalized with PQN algorithm [37] using an in-house MATLAB toolbox (github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA). Relative quantification of all spectral features was performed using a semi-automatic workflow within the toolbox.

Statistical analysis of these features including PCA, fold change calculations, and ANOVA was performed using Metaboanalyst [38]. One-way ANOVA was performed using these features to identify significant features, using Tukey post-hoc test to determine significant comparisons. Non-overlapped features were chosen to represent relative abundance of annotated metabolites, and assigned a confidence score as previously described [39]. Pathway enrichment analysis was also performed in Metaboanalyst using hypergeometric test for overrepresentation, mapping to pathways described in SMPD [40]. For pairwise comparisons of p.Thr331Pro vs WT HEK293 and WT vs P2 primary fibroblast cells, a Student's T-test was performed. Metabolites with a raw *p*-value < 0.05 were included in the list of compounds for pathway enrichment. For integrated transcriptome and metabolome pathway analysis of primary fibroblast cells, fold changes for all annotated metabolites and transcripts between WT and P2 cells, along with lists of metabolites and transcripts with p-value < 0.05 were submitted to Paintomics 3.0 to establish pathway enrichment using Fisher exact test [41].

Polyamine measurements

Intracellular polyamine concentrations were measure by the reverse-phase HPLC methods of Kabra et al. using 1,7-diaminoheptane as an internal standard [19].

RNA sequencing

To prepare libraries for RNA sequencing, triplicate fibroblast cultures at 80-90% confluence in 10 cm culture dishes were collected directly in 1.5 mL Trizol. Total RNA was extracted using the Direct-Zol miniprep kit RNA extraction kit with a DNA digestion step according to the manufacturer's instructions (Zymo Research, Irvine, CA). RNA was eluted with 30 μL water. We depleted ribosomal RNA using the Universal Plus Total RNA Seq kit with Human AnyDeplete (Tecan, Männedorf, Switzerland) and prepared bar-

coded cDNA libraries for sequencing on an SP flow cell on the NovaSeq 6000 platform (Illumina, San Diego, CA) similar to previously described work [42,43]. We performed the initial steps of raw read processing and normalization as previously described except that we used the human reference genome (GRCh38) for alignment [42,43]. We used the edgeR package for differential expression analysis and performed Gene Ontology analysis by statistical overrepresentation tests using PantherDB [44-46]. The code for all the analyses is available on Github and the data is deposited in GEO (https://github.com/snehamokashi/ADH_RNAseq and GEO accession number: GSE202424).

Acknowledgements

We are grateful to the patients and families for their participation in this study. We thank Bonne Lethco and Dr. Barb Dupont at the GGC for assistance with the conventional karyotyping of the WT parental and *ALDH18A1* KO cells. This work was supported by the Greenwood Genetic Center, the National Science Foundation (NSF 1946970 and 1648035; to ASE and MC), and the University of Pennsylvania Orphan Disease Center Million Dollar Bike Ride (MDBR-20-135-SRS) and the Chan Zuckerberg Initiative (to RAC/TMS), and the National Cancer Institute (R01 CA235863 to RAC).

Conflict of Interest: The Greenwood Genetic Center receives revenue from diagnostic testing performed in the GGC Molecular Diagnostic Laboratory.

Ethics statement: Informed consent was obtained from the families of the affected individuals involved in this study (Self Regional Healthcare; IRB Number: Pro00085001). This consent included permission to use patient photos in published manuscripts.

References

1. Hu, C.A., Khalil, S., Zhaorigetu, S., Liu, Z., Tyler, M., Wan, G., and Valle, D. (2008). Human Delta1-pyrroline-5-carboxylate synthase: function and regulation. Amino Acids *35*, 665-672. 10.1007/s00726-008-0075-0.

2. Hu, C.A., Delauney, A.J., and Verma, D.P. (1992). A bifunctional enzyme (delta 1-pyrroline-5-carboxylate synthetase) catalyzes the first two steps in proline biosynthesis in plants. Proc Natl Acad Sci U S A *89*, 9354-9358. 10.1073/pnas.89.19.9354.

3. Ginguay, A., Cynober, L., Curis, E., and Nicolis, I. (2017). Ornithine Aminotransferase, an Important Glutamate-Metabolizing Enzyme at the Crossroads of Multiple Metabolic Pathways. Biology (Basel) *6*. 10.3390/biology6010018.

4. Krishnan, N., Dickman, M.B., and Becker, D.F. (2008). Proline modulates the intracellular redox environment and protects mammalian cells against oxidative stress. Free Radic Biol Med *44*, 671-681. 10.1016/j.freeradbiomed.2007.10.054.

5. Yang, Z., Zhao, X., Shang, W., Liu, Y., Ji, J.F., Liu, J.P., and Tong, C. (2021). Pyrroline-5-carboxylate synthase senses cellular stress and modulates metabolism by regulating mitochondrial respiration. Cell Death Differ *28*, 303-319. 10.1038/s41418-020-0601-5.

6. Ha, H.C., Sirisoma, N.S., Kuppusamy, P., Zweier, J.L., Woster, P.M., and Casero, R.A., Jr. (1998). The natural polyamine spermine functions directly as a free radical scavenger. Proc Natl Acad Sci U S A *95*, 11140-11145. 10.1073/pnas.95.19.11140.

7. Ha, H.C., Yager, J.D., Woster, P.A., and Casero, R.A., Jr. (1998). Structural specificity of polyamines and polyamine analogues in the protection of DNA from strand breaks induced by reactive oxygen species. Biochem Biophys Res Commun *244*, 298-303. 10.1006/bbrc.1998.8258.

8. Rider, J.E., Hacker, A., Mackintosh, C.A., Pegg, A.E., Woster, P.M., and Casero, R.A., Jr. (2007). Spermine and spermidine mediate protection against oxidative damage caused by hydrogen peroxide. Amino Acids *33*, 231-240. 10.1007/s00726-007-0513-4.

9. Kamoun, P., Aral, B., and Saudubray, J.M. (1998). [A new inherited metabolic disease: delta1-pyrroline 5-carboxylate synthetase deficiency]. Bull Acad Natl Med *182*, 131-137; discussion 138-139.

10. Craze, M.L., Cheung, H., Jewa, N., Coimbra, N.D.M., Soria, D., El-Ansari, R., Aleskandarany, M.A., Wai Cheng, K., Diez-Rodriguez, M., Nolan, C.C., et al. (2018). MYC regulation of glutamine-proline regulatory axis is key in luminal B breast cancer. Br J Cancer *118*, 258-265. 10.1038/bjc.2017.387.

11. Liu, W., Le, A., Hancock, C., Lane, A.N., Dang, C.V., Fan, T.W., and Phang, J.M. (2012). Reprogramming of proline and glutamine metabolism contributes to the proliferative and metabolic responses regulated by oncogenic transcription factor c-MYC. Proc Natl Acad Sci U S A *109*, 8983-8988. 10.1073/pnas.1203244109.

12. Phang, J.M., Liu, W., Hancock, C.N., and Fischer, J.W. (2015). Proline metabolism and cancer: emerging links to glutamine and collagen. Curr Opin Clin Nutr Metab Care *18*, 71-77. 10.1097/MCO.0000000000000121.

13. Coutelier, M., Goizet, C., Durr, A., Habarou, F., Morais, S., Dionne-Laporte, A., Tao, F., Konop, J., Stoll, M., Charles, P., et al. (2015). Alteration of ornithine metabolism leads to dominant and recessive hereditary spastic paraplegia. Brain *138*, 2191-2205. 10.1093/brain/awv143.

14. Panza, E., Martinelli, D., Magini, P., Dionisi Vici, C., and Seri, M. (2019). Hereditary Spastic Paraplegia Is a Common Phenotypic Finding in ARG1 Deficiency, P5CS

Deficiency and HHH Syndrome: Three Inborn Errors of Metabolism Caused by Alteration of an Interconnected Pathway of Glutamate and Urea Cycle Metabolism. Front Neurol *10*, 131. 10.3389/fneur.2019.00131.

15. Marco-Marin, C., Escamilla-Honrubia, J.M., Llacer, J.L., Seri, M., Panza, E., and Rubio, V. (2020). Delta(1) -Pyrroline-5-carboxylate synthetase deficiency: An emergent multifaceted urea cycle-related disorder. J Inherit Metab Dis *43*, 657-670. 10.1002/jimd.12220.
16. Fischer-Zirnsak, B., Escande-Beillard, N., Ganesh, J., Tan, Y.X., Al Bughaili, M., Lin, A.E., Sahai, I., Bahena, P., Reichert, S.L., Loh, A., et al. (2015). Recurrent De Novo Mutations Affecting Residue Arg138 of Pyrroline-5-Carboxylate Synthase Cause a Progeroid Form of Autosomal-Dominant Cutis Laxa. Am J Hum Genet *97*, 483-492. 10.1016/j.ajhg.2015.08.001.
17. Bicknell, L.S., Pitt, J., Aftimos, S., Ramadas, R., Maw, M.A., and Robertson, S.P. (2008). A missense mutation in ALDH18A1, encoding Delta1-pyrroline-5-carboxylate synthase (P5CS), causes an autosomal recessive neurocutaneous syndrome. Eur J Hum Genet *16*, 1176-1186. 10.1038/ejhg.2008.91.
18. Panza, E., Escamilla-Honrubia, J.M., Marco-Marin, C., Gougeard, N., De Michele, G., Morra, V.B., Liguori, R., Salviati, L., Donati, M.A., Cusano, R., et al. (2016). ALDH18A1 gene mutations cause dominant spastic paraplegia SPG9: loss of function effect and plausibility of a dominant negative mechanism. Brain *139*, e3. 10.1093/brain/awv247.
19. Kabra, P.M., Lee, H.K., Lubich, W.P., and Marton, L.J. (1986). Solid-phase extraction and determination of dansyl derivatives of unconjugated and acetylated polyamines by

reversed-phase liquid chromatography: improved separation systems for polyamines in cerebrospinal fluid, urine and tissue. J Chromatogr *380*, 19-32. 10.1016/s0378-4347(00)83621-x.

20. Hatamochi, A., Kuroda, K., Shinkai, H., Kohma, H., Oishi, Y., and Inoue, S. (1998). Regulation of matrix metalloproteinase (MMP) expression in cutis laxa fibroblasts: upregulation of MMP-1, MMP-3 and MMP-9 genes but not of the MMP-2 gene. Br J Dermatol *138*, 757-762. 10.1046/j.1365-2133.1998.02210.x.
21. Hatamochi, A., Mori, K., Arakawa, M., Ueki, H., and Kondo, M. (1996). Collagenase gene expression in cutis laxa fibroblasts is upregulated by transcriptional activation of the promoter gene through a 12-0-tetradecanoyl-phorbol-13-acetate (TPA)-responsive element. J Invest Dermatol *106*, 631-636. 10.1111/1523-1747.ep12345435.
22. Tsien, C., Davuluri, G., Singh, D., Allawy, A., Ten Have, G.A., Thapaliya, S., Schulze, J.M., Barnes, D., McCullough, A.J., Engelen, M.P., et al. (2015). Metabolic and molecular responses to leucine-enriched branched chain amino acid supplementation in the skeletal muscle of alcoholic cirrhosis. Hepatology *61*, 2018-2029. 10.1002/hep.27717.
23. Wyant, G.A., Abu-Remaileh, M., Wolfson, R.L., Chen, W.W., Freinkman, E., Danai, L.V., Vander Heiden, M.G., and Sabatini, D.M. (2017). mTORC1 Activator SLC38A9 Is Required to Efflux Essential Amino Acids from Lysosomes and Use Protein as a Nutrient. Cell *171*, 642-654 e612. 10.1016/j.cell.2017.09.046.
24. Zheng, L., Zhang, W., Zhou, Y., Li, F., Wei, H., and Peng, J. (2016). Recent Advances in Understanding Amino Acid Sensing Mechanisms that Regulate mTORC1. Int J Mol Sci *17*. 10.3390/ijms17101636.

25. Markaki, M., Tsagkari, D., and Tavernarakis, N. (2021). Mitophagy mechanisms in neuronal physiology and pathology during ageing. Biophys Rev *13*, 955-965. 10.1007/s12551-021-00894-7.

26. Sukhorukov, V., Voronkov, D., Baranich, T., Mudzhiri, N., Magnaeva, A., and Illarioshkin, S. (2021). Impaired Mitophagy in Neurons and Glial Cells during Aging and Age-Related Disorders. Int J Mol Sci *22*. 10.3390/ijms221910251.

27. Cheriyan, J., Balsara, R.D., Hansen, K.B., and Castellino, F.J. (2016). Pharmacology of triheteromeric N-Methyl-D-Aspartate Receptors. Neurosci Lett *617*, 240-246. 10.1016/j.neulet.2016.02.032.

28. Hirose, T., Saiki, R., Yoshizawa, Y., Imamura, M., Higashi, K., Ishii, I., Toida, T., Williams, K., Kashiwagi, K., and Igarashi, K. (2015). Spermidine and Ca(2+), but not Na(+), can permeate NMDA receptors consisting of GluN1 and GluN2A or GluN2B in the presence of Mg(2+). Biochem Biophys Res Commun *463*, 1190-1195. 10.1016/j.bbrc.2015.06.081.

29. Pegg, A.E. (2016). Functions of Polyamines in Mammals. J Biol Chem *291*, 14904-14912. 10.1074/jbc.R116.731661.

30. Arena, J.F., Schwartz, C., Ouzts, L., Stevenson, R., Miller, M., Garza, J., Nance, M., and Lubs, H. (1996). X-linked mental retardation with thin habitus, osteoporosis, and kyphoscoliosis: linkage to Xp21.3-p22.12. Am J Med Genet *64*, 50-58. 10.1002/(SICI)1096-8628(19960712)64:1<50::AID-AJMG7>3.0.CO;2-V.

31. Cason, A.L., Ikeguchi, Y., Skinner, C., Wood, T.C., Holden, K.R., Lubs, H.A., Martinez, F., Simensen, R.J., Stevenson, R.E., Pegg, A.E., and Schwartz, C.E. (2003). X-linked

spermine synthase gene (SMS) defect: the first polyamine deficiency syndrome. Eur J Hum Genet *11*, 937-944. 10.1038/sj.ejhg.5201072.

32. Bupp, C.P., Schultz, C.R., Uhl, K.L., Rajasekaran, S., and Bachmann, A.S. (2018). Novel de novo pathogenic variant in the ODC1 gene in a girl with developmental delay, alopecia, and dysmorphic features. Am J Med Genet A *176*, 2548-2553. 10.1002/ajmg.a.40523.

33. Peron, A., Spaccini, L., Norris, J., Bova, S.M., Selicorni, A., Weber, G., Wood, T., Schwartz, C.E., and Mastrangelo, M. (2013). Snyder-Robinson syndrome: a novel nonsense mutation in spermine synthase and expansion of the phenotype. Am J Med Genet A *161A*, 2316-2320. 10.1002/ajmg.a.36116.

34. VanSickle, E.A., Michael, J., Bachmann, A.S., Rajasekaran, S., Prokop, J.W., Kuzniecky, R., Hofstede, F.C., Steindl, K., Rauch, A., Lipson, M.H., and Bupp, C.P. (2021). Expanding the phenotype: Four new cases and hope for treatment in Bachmann-Bupp syndrome. Am J Med Genet A *185*, 3485-3493. 10.1002/ajmg.a.62473.

35. Murray Stewart, T., Dunston, T.T., Woster, P.M., and Casero, R.A., Jr. (2018). Polyamine catabolism and oxidative damage. J Biol Chem *293*, 18736-18745. 10.1074/jbc.TM118.003337.

36. Bingol, K., Li, D.W., Zhang, B., and Bruschweiler, R. (2016). Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. Anal Chem *88*, 12411-12418. 10.1021/acs.analchem.6b03724.

37. Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem *78*, 4281-4290. 10.1021/ac051632c.

38. Chong, J., Wishart, D.S., and Xia, J. (2019). Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Curr Protoc Bioinformatics *68*, e86. 10.1002/cpbi.86.

39. Walejko, J.M., Chelliah, A., Keller-Wood, M., Gregg, A., and Edison, A.S. (2018). Global Metabolomics of the Placenta Reveals Distinct Metabolic Profiles between Maternal and Fetal Placental Tissues Following Delivery in Non-Labored Women. Metabolites *8*. 10.3390/metabo8010010.

40. Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C., et al. (2010). SMPDB: The Small Molecule Pathway Database. Nucleic Acids Res *38*, D480-487. 10.1093/nar/gkp1002.

41. Hernandez-de-Diego, R., Tarazona, S., Martinez-Mira, C., Balzano-Nogueira, L., Furio-Tari, P., Pappas, G.J., Jr., and Conesa, A. (2018). PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res *46*, W503-W509. 10.1093/nar/gky466.

42. Johnstun, J.A., Shankar, V., Mokashi, S.S., Sunkara, L.T., Ihearahu, U.E., Lyman, R.L., Mackay, T.F.C., and Anholt, R.R.H. (2021). Functional Diversification, Redundancy, and Epistasis among Paralogs of the Drosophila melanogaster Obp50a-d Gene Cluster. Mol Biol Evol *38*, 2030-2044. 10.1093/molbev/msab004.

43. Mokashi, S.S., Shankar, V., MacPherson, R.A., Hannah, R.C., Mackay, T.F.C., and Anholt, R.R.H. (2021). Developmental Alcohol Exposure in Drosophila: Effects on Adult Phenotypes and Gene Expression in the Brain. Front Psychiatry *12*, 699033. 10.3389/fpsyt.2021.699033.

44. Chen, Y., Lun, A.T., and Smyth, G.K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res *5*, 1438. 10.12688/f1000research.8987.2.

45. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res *40*, 4288-4297. 10.1093/nar/gks042.

46. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140. 10.1093/bioinformatics/btp616.

PART 2

NMR METABOLOMICS FOR IMPROVEMENT OF CELL THERAPY MANUFACTURING

CHAPTER 4

PREDICTING T-CELL QUALITY DURING MANUFACTURING THROUGH AN ARTIFICIAL INTELLIGENCE-BASED INTEGRATIVE MULTIOMICS ANALYTICAL PLATFORM[1]

[1]Odeh-Couvertier VY*, Dwarshuis NJ*, Colonna MB*, Levine BL, Edison AS, Kotanchek T, Roy K, Torres-Garcia W. Predicting T-cell quality during manufacturing through an artificial intelligence-based integrative multiomics analytical platform. Bioeng Transl Med. 2022 Jan 4;7(2):e10282. doi: 10.1002/btm2.10282. PMID: 35600660; PMCID: PMC9115702. Reprinted here with permission from the publisher (* authors contributed equally to this work).

Foreword

Chapter 3 is reprinted from Odeh-Couvertier VY*, Dwarshuis NJ*, Colonna MB*, Levine BL, Edison AS, Kotanchek T, Roy K, Torres-Garcia W. Predicting T-cell quality during manufacturing through an artificial intelligence-based integrative multiomics analytical platform. Bioeng Transl Med. 2022 Jan 4;7(2):e10282. doi: 10.1002/btm2.10282. (* authors contributed equally to this work). This work is reproduced under the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/legalcode). My contributions to this work were (i) developing study design for metabolomics, (ii) developing protocol for sample collection and storage, (ii), sample preparation and NMR analysis, (iii) quantitative NMR data analysis, (iv) annotation of media spectra, (v) biological interpretation of modeling results, (vi) creating spectral simulations of culture media as proof of concept, (vii) writing appropriate sections pertaining to results, and (viii) reviewing and editing manuscript. All other experimental and analysis work was performed by Nathan J. Dwarshuis and Valerie Odeh-Couvertier, respectively. Categories and levels of other author contributions are as follows (included as part of publication): Valerie Odeh-Couvertier: Conceptualization (equal); data curation (equal); formal analysis (lead); investigation (lead); methodology (lead); software (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). Nathan J. Dwarshuis: Conceptualization (equal); data curation (lead); formal analysis (equal); investigation (lead); methodology (lead); resources (equal); validation (lead); writing – original draft (equal); writing – review and editing (equal). Maxwell B. Colonna: Conceptualization (equal); data curation (lead); formal analysis (equal); investigation (lead); methodology (lead); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). Bruce L. Levine: Conceptualization (equal); supervision (equal); writing – original draft

(equal); writing – review and editing (equal). Arthur S. Edison: Conceptualization (equal); funding acquisition (equal); investigation (equal); resources (lead); writing – original draft (equal); writing – review and editing (equal). Theresa Kotanchek: Conceptualization (equal); formal analysis (lead); methodology (lead); software (equal); supervision (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). Krishnendu Roy: Conceptualization (lead); funding acquisition (lead); project administration (equal); resources (lead); supervision (equal); writing – original draft (equal); writing – review and editing (equal). Wandaliz Torres-Garcia: Conceptualization (equal); formal analysis (lead); methodology (equal); project administration (lead); resources (equal); software (lead); supervision (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead).

Abstract

Large-scale, reproducible manufacturing of therapeutic cells with consistently high quality is vital for translation to clinically effective and widely accessible cell therapies. However, the biological and logistical complexity of manufacturing a living product, including challenges associated with their inherent variability and uncertainties of process parameters, currently make it difficult to achieve predictable cell-product quality. Using a degradable microscaffold-based T-cell process, we developed an artificial intelligence (AI)-driven experimental-computational platform to identify a set of critical process parameters and critical quality attributes from heterogeneous, high-dimensional, time-dependent multiomics data, measurable during early stages of manufacturing and predictive of end-of-manufacturing product quality. Sequential, design-of-experiment-based studies, coupled with an agnostic machine-learning framework, were used to extract feature combinations from early in-culture media assessment that were highly predictive of the end-product CD4/CD8 ratio and total live $CD4^+$ and $CD8^+$ naïve and central memory T cells ($CD63L^+CCR7^+$). Our results demonstrate a broadly applicable platform tool to predict end-product quality and composition from early time point in-process measurements during therapeutic cell manufacturing.

Introduction

T-cell-based immunotherapies have received great interest from clinicians and industry due to their potential to treat, and often functionally cure some hematological cancers and their potential applicability in many other diseases.[1, 2] Since 2017, four genetically modified autologous Chimeric Antigen Receptor (CAR) T-cell therapies (*Yescarta*™, *Kymriah*™, *Tecartus*™, and *Breyanzi*®) have received approval from the U.S. Food and Drug Administration to treat certain B-cell malignancies. Despite these successes, T-cell-based immunotherapies are constrained by time-intensive, high cost, complex manufacturing processes that are time-intensive, expensive, and difficult to scale[3, 4] and lack methods and tools to predict the end-product quality during manufacturing. Quality assessment is performed only at the end of manufacturing which takes many days. Identification of early putative critical quality attributes (CQAs) and the associated critical process parameters (CPPs) that can be measured nondestructively during culture and can predict end-product attributes early in the manufacturing timeline could be transformative for the cell therapy field.

Translating laboratory-scale T-cell expansion experiments into a large-scale manufacturing process is hindered by the incomplete understanding of cell properties and how they are affected by process variables, lack of detailed characterization, and high variability of materials during manufacturing.[5] These challenges of manufacturing a "living product" are further magnified since current chemistry, manufacturing, and control, analytics, regulations, and product specifications are designed for conventional chemical and biopharmaceutical manufacturing systems.[6] This underscores the need to develop innovative tools, methods, and standards to ensure appropriate quality controls, and new strategies involving quality by design and good manufacturing practices for cell-based therapies.[7-9] The intricate manufacturing process for T cells and other cell therapies

must be deeply assessed and appropriately controlled to ensure scalability, predictability, and a high-quality manufacturing process at the most reasonable cost. A key step for reaching this goal is to identify putative CQAs and CPPs early in the manufacturing process that can predict the quality of the manufactured cell-therapy product. We hypothesized that rigorous characterization of process parameters along with longitudinal measurements of cell-secreted cytokine, chemokine, and metabolites from the culture media early during manufacturing will allow us to develop an artificial intelligence (AI)-based mathematical-computational framework for the identification of multivariate parameters that are predictive of the end-of-manufacturing product phenotypes.

Characterization studies of approved autologous anti-CD19 CAR-T cell therapies have recently revealed initial sets of candidate quality attributes, that is, percent transduction, vector copy number, and interferon-γ production for axicabtagene ciloleucel (Yescarta™),[10] while CAR expression and release of interferon-γ are a few of those identified for tisagenlecleucel (Kymriah™).[11] Many of these attributes are calculated as endpoint responses and thus a deeper understanding of the cell growth process impacted by starting conditions and performance during their manufacturing is essential. Hence, CQAs that enable early monitoring through real-time process measurements such as multiomics cell characterization can overcome current challenges in assessing product consistency. Yet, the computational complexity of dealing with the heterogeneity and multivariate nature of multiomics measurements to characterize T-cell quality, that is, high-definition phenotyping of naïve and memory subsets, remains a challenge.

Generally, T cells with a lower differentiation state such as naïve and stem cell or central memory cells have been shown to provide superior anti-tumor potency, presumably due to their higher potential to replicate, migrate, and engraft, leading to a long-term, durable response.[12-15] Likewise, CD4 T cells are similarly important to anti-tumor potency due to their cytokine release properties

and ability to resist exhaustion.[16, 17] Our group has developed a novel degradable microscaffold (DMS)-based method using porous microcarriers functionalized with anti-CD3 and anti-CD28 mAbs for use in T-cell expansion cultures. We showed that compared to commercially available microbeads (Miltenyi), DMSs generated a higher number of migratory naïve (T_N) and central memory (T_{CM}) ($CCR7^+CD62L^+$) T cells and $CD4^+$ T cells across multiple donors.[18] We used this manufacturing process as an exemplar to develop an experimental-computational AI-based tool to predict product quality from early process measurements. This two-phase approach consists of (1) the optimization of process parameters through experimental designs, and (2) the extraction of early predictive signatures of T-cell quality by multiomics integration using regression models. This agnostic computational approach provides a platform to discover early predictive CQAs and CPPs to ensure consistent product quality that can be widely applicable for other cellular therapies.

Results

Overall multiomics study design

T cells were expanded ex vivo for 14 days and 100 µl of supernatant media samples were collected at days 4, 6, 8, 11, and 14 to measure cytokine profiles and perform nuclear magnetic resonance (NMR) analysis. Endpoint responses on DMS-based T-cell extracts were measured for different combinations of DMS parameters: IL2 concentration, DMS concentration, and functionalized antibody percent. Two experimental regions were determined using a design-of-experiments (DOE) methodology to maximize the yields of $CD62L^+CCR7^+$ cells (i.e., naïve and central memory T cells, $T_N + T_{CM}$) as a function of these process parameters. The first DOE resulted in a randomized 18-run I-optimal custom design where each DMS parameter was evaluated at three levels. To further optimize this DOE in terms of total live $CD4^+$ $T_N + T_{CM}$ cells, a sequential adaptive design-of-experiment (ADOE) was designed with 12 additional samples

(**Figure 4.1b**). All 30 runs from both experiments (DOE, ADOE) were molecularly characterized to model total live $T_N + T_{CM}$ (a) $CD4^+$, (b) $CD8^+$, and (c) their ratio (Supporting Figure 4.S1). The extraction of early predictive CPPs and CQAs for the expansion of $T_N + T_{CM}$ cells during ex vivo culture was performed in two phases: (1) optimization of process parameters and (2) integration of multiomics for predictive modeling (**Figure 4.1**).

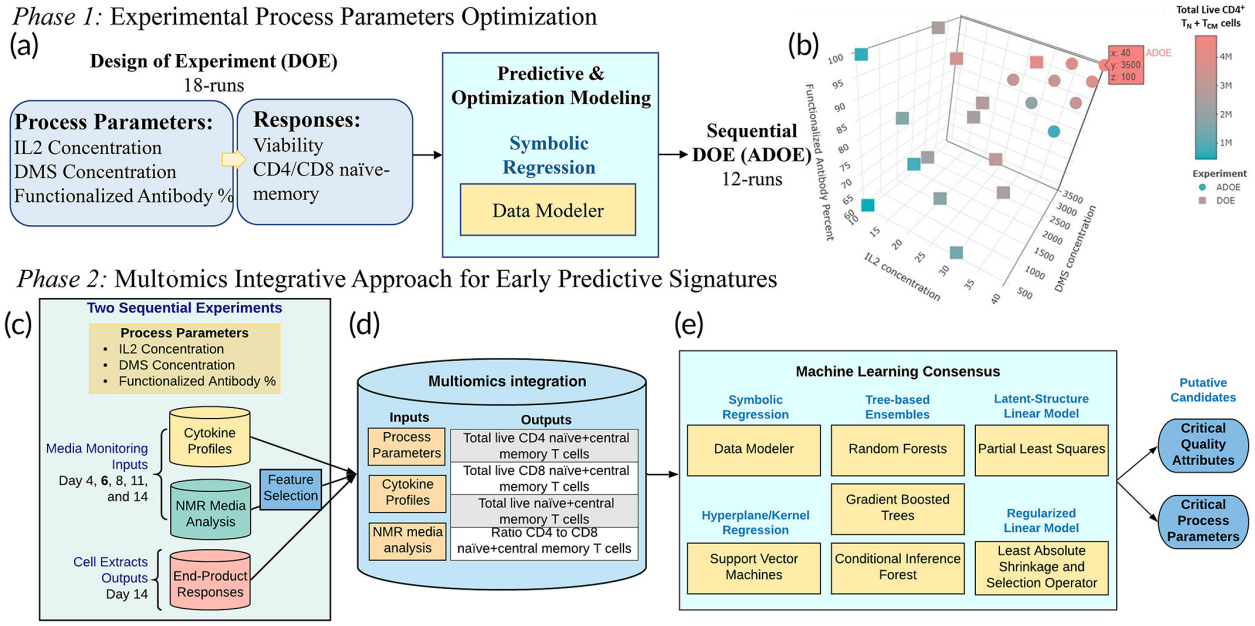


Figure 4.1 Two-phase approach to extract early predictive critical process parameters (CPPs) and critical quality attributes (CQAs) for $CD4^+/CD8^+$ $T_N + T_{CM}$ cells. (a) Design-of-experiment (DOE) modeling and optimization of process parameters. (b) Experimental region studied and optimized for total live $CD4^+$ $T_N + T_{CM}$ cells. (c) Total live $CD4^+$ $T_N + T_{CM}$ cells across the overall study design (two experiments varying process parameters). (d) Integrative multiomics approach through (e) a machine learning consensus analysis to identify early predictive CPPs and CQAs putative candidates for both total live $CD4^+$ and $CD8^+$ $T_N + T_{CM}$ cells

Optimization of $T_N + T_{CM}$ cells as a function of process parameters

Using symbolic regression (Data Modeler software from Evolved Analytics LLC), we examined the interactive effects of the DMS parameters on yield to simultaneously predict and optimize both $CD4^+$ and $CD8^+$ $T_N + T_{CM}$. A model ensemble predicted 4.2 $\times 10^6$ $CD4^+$ $T_N + T_{CM}$ cells at an optimum setting of 30 U/µl IL2, 2500 carriers/µl, and 100% functionalized mAbs (Supporting Figure **S2**). This result was consistent with the observed

maximum value of 4.0×10^6, highlighting that $CD4^+$ $T_N + T_{CM}$ yield was maximized at high levels of DMS parameters (**Figure 4.1b**). In contrast, the predicted optimum yield for $CD8^+$ $T_N + T_{CM}$ was 1.9×10^7 cells at a setting of 30 U/µl IL2, 600 carriers/µl, and 100% functionalized mAbs (data not shown). Although this combination was not experimentally tested, the closest measured record (30 U/µl IL2, 500 carriers/µl, 100% functionalized mAbs) achieved the predicted maximum yield. Hence, the $CD8^+$ $T_N + T_{CM}$ yield was maximized at high IL2 concentration and functionalized mAbs percentage but low DMS concentration.

The DOE analysis highlighted the potential for further optimization of total live $CD4^+$ $T_N + T_{CM}$ cells, as well as the potential to optimize the $CD4^+$ to $CD8^+$ $T_N + T_{CM}$ cells ratio, at DMS levels greater than those originally evaluated (DOE). Therefore, to test and validate, a second adaptive design of experiment (ADOE) was designed to maximize the total live $CD4^+$ $T_N + T_{CM}$ cells. We expanded the parameter range, assessing IL2 concentration >30 U/µl and DMS concentration >2500 carriers/µl (Figure **1b**). $CD4^+$ $T_N + T_{CM}$ and its ratio to $CD8^+$ T_N+ T_{CM}, 4.7×10^6 cell and 0.49 respectively, were maximized when IL2 concentration (40 U/µl) and DMS concentration (3500 carriers/µl) were maximized (**Figure 4.1b**; Supporting Table 4.S1; Figure **S2**). Utilizing the ADOE data set, new response ensembles were generated enabling more robust prediction over the expanded parameter space (↑IL2 and ↑DMS concentrations).

Multiomic integrative analysis for early monitoring of T-cell manufacturing

Due to the heterogeneity of the multivariate data collected and knowing that no single model structure is perfect for all applications, we implemented an agnostic modeling approach to better understand these $T_N + T_{CM}$ responses. To achieve this, a consensus analysis using seven machine learning (ML) techniques, random forest (RF), gradient boosted machine (GBM), conditional inference forest (CIF), least absolute shrinkage and selection operator (LASSO),

partial least-squares regression (PLSR), support vector machine (SVM), and data modeler's symbolic regression (SR), was implemented to molecularly characterize $T_N + T_{CM}$ cells and to extract predictive features of quality early on their expansion process (**Figure 4.1d,e**).

SR models achieved the highest predictive performance ($R^2 > 93\%$) when using multiomics predictors for all endpoint responses (**Table 4.1**). SR achieved $R^2 > 98\%$, while GBM tree-based ensembles showed leave-one-out cross-validated R^2 (LOO-R^2) >95% for $CD4^+$ and $CD4^+/CD8^+$ $T_N + T_{CM}$ responses. Similarly, LASSO, PLSR, and SVM methods showed consistent high LOO-R^2, 92.9%, 99.7%, and 90.5%, respectively, to predict the $CD4^+/CD8^+$ $T_N + T_{CM}$. Yet, about 10% reduction in LOO-R^2, 72.5%–81.7%, was observed for $CD4^+$ $T_N + T_{CM}$ with these three methods. Lastly, SR and PLSR achieved $R^2 > 90\%$ while other ML methods exhibited exceedingly variable LOO-R^2 (0.3%, RF-51.5%, LASSO) for $CD8^+$ $T_N + T_{CM}$ cells. The top-performing technique, SR, showed that the median aggregated predictions for total live $CD4^+$ and $CD8^+$ $T_N + T_{CM}$ cells increases when IL2 concentration, IL15, and IL2R increase, while IL17a decreases in conjunction with other interactive features. These patterns combined with low values of DMS concentration and GM_CSF uniquely characterized maximum $CD8^+$ $T_N + T_{CM}$. Meanwhile, higher glycine but lower IL13 in combination with others showed maximum $CD4^+$ $T_N + T_{CM}$ predictions (**Figure 4.2a**).

Table 4.1 LOO-R^2 prediction performance results for all machine learning (ML) models when evaluating process parameters, and features from cytokine and nuclear magnetic resonance (NMR) media analysis at day 6 or day 4 *Notes*: ML models' prediction performance is measured as the leave-one-out cross-validated R^2 (LOO-R^2) while SR prediction performance is measured as R^2 of the ensemble prediction where the ensemble is composed of diverse models with complexity constrained. Predictors evaluated: (PP) Process parameters, (N) NMR, (S) Cytokines measured at day 4 or 6. Maximum R^2 within each ML method are shown in bold.

LOO-R^2	ML						
Response/predictors	**SR**	**RF**	**GBM**	**CIF**	**LASSO**	**PLSR**	**SVM**
Ratio of CD4 to CD8 $T_N + T_{CM}$ cells							
PP + N4	**99%**	86.8%	96.3%	**84.5%**	88.6%	92.5%	88.5%
PP + N6	**99%**	73.6%	95.9%	70.1%	81.0%	95.8%	79.7%
PP + S6	**99%**	**87.1%**	**99.9%**	83.4%	87.2%	97.9%	86.8%
PP + S6 + N6	**99%**	85.5%	95.3%	83.4%	**92.9%**	**99.7%**	**90.5%**
Total live $CD4^+$ $T_N + T_{CM}$ cells							
PP + N4	97%	67.0%	93.6%	69.3%	34.3%	90.1%	75.5%
PP + N6	96%	45.9%	92.6%	51.2%	42.8%	**92.1%**	**79.4%**
PP + S6	**98%**	**71.4%**	**99.9%**	**75.0%**	**74.9%**	80.0%	75.5%
PP + S6 + N6	**98%**	68.2%	95.6%	74.4%	72.5%	81.7%	77.0%
Total live $CD8^+$ $T_N + T_{CM}$ cells							
PP + N4	93%	4.7%	**44.4%**	9.2%	1.2%	65.1%	9.1%
PP + N6	86%	2.0%	29.9%	**15.8%**	28.5%	63.3%	30.6%
PP + S6	**93%**	**7.8%**	28.0%	15.1%	**76.2%**	**98.4%**	**49.8%**
PP + S6 + N6	**93%**	0.3%	32.7%	9.8%	51.5%	96.4%	37.8%

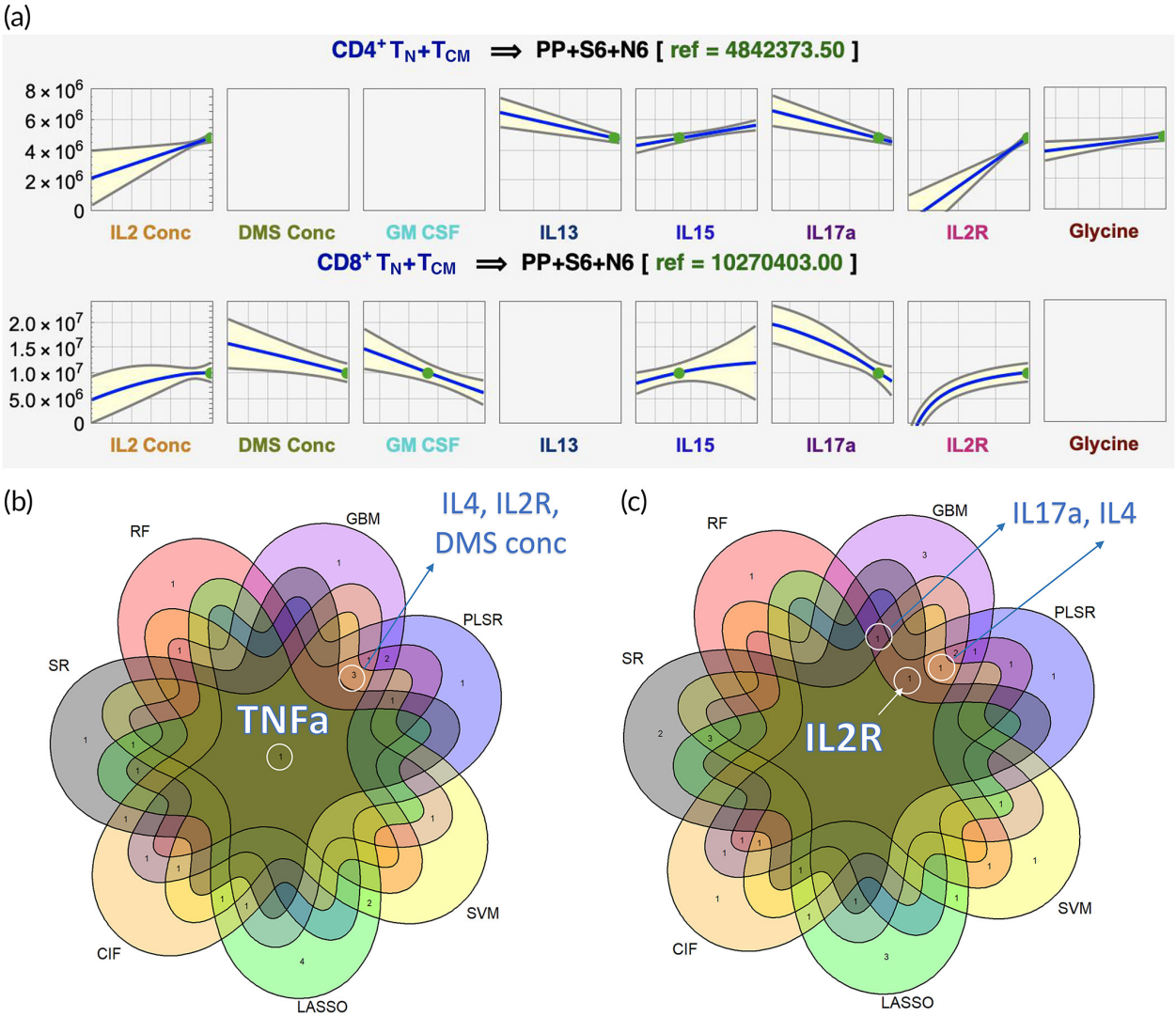


Figure 4.2 Multiomics culturing media prediction profiles of highly predictive features for early monitoring of T-cell manufacturing. (a) Prediction model profiles from day 6 cultured media monitoring where total live $CD4^+$ $T_N + T_{CM}$ is maximized. (b) Machine learning (ML) models consensus for ratio $CD4^+$ to $CD8^+$ $T_N + T_{CM}$ cells, and (c) ML models consensus for total live $CD4^+$ $T_N + T_{CM}$ cells. Feature names are shown for consensus with 5 or more ML models at the highest-ranking standing (see the Materials and Methods section)

Selecting CPPs and CQAs candidates consistently for T-cell memory across different models is desired. Here, TNFα was found in consensus across all seven ML methods for predicting $CD4^+/CD8^+$ $T_N + T_{CM}$ when considering features with the highest importance scores across models (**Figure 4.2b**; Materials and Methods section). Other features, IL2R, IL4, IL17a, and DMS

concentration, were commonly selected in ≥5 ML methods (**Figure 4.2b,c**). Moreover, IL13 and IL15 were found predictive in combination with these using SR (Table 4.S2).

This integrative analysis of cytokine and NMR media analysis monitored at early stages of the T-cell process provided highly predictive feature combinations of end-product quality particularly for total live $T_N + T_{CM}$ $CD4^+$ cells and $CD4^+/CD8^+$ ratio as shown in **Figure 4.3a,b**. However, when translating a real-time monitoring strategy to a large-scale manufacturing process, measuring both cytokine and NMR features from media can be difficult and expensive. To be cost-efficient and translatable, we demonstrated that either cytokine profiles or NMR media analysis alone is sufficient to find predictive features without compromising prediction performance.

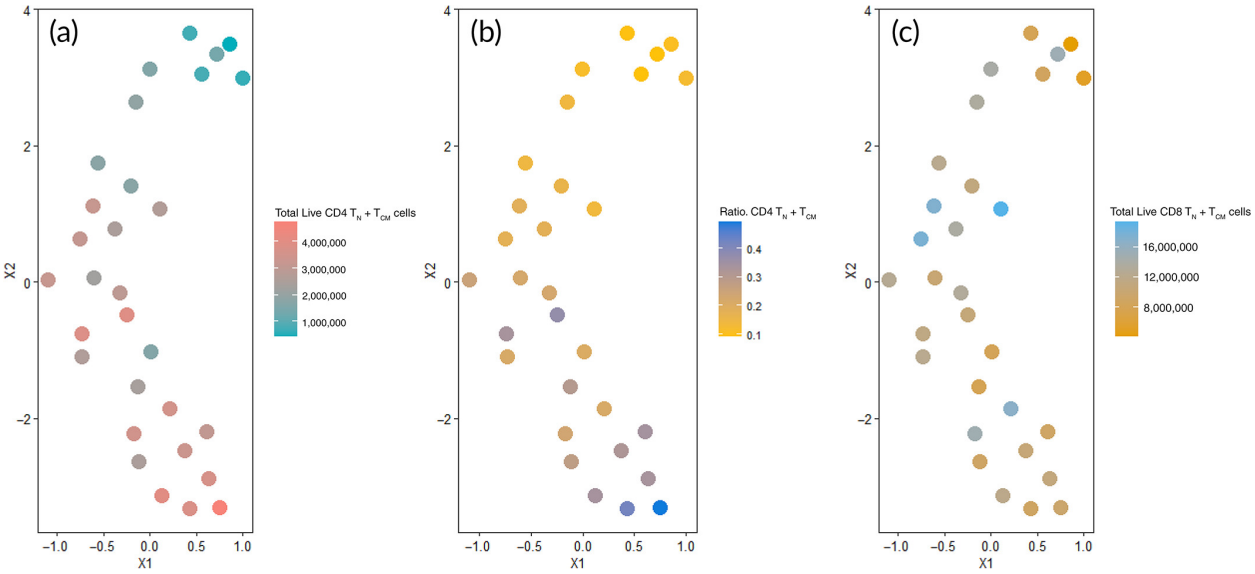


Figure 4.3 Uniform manifold approximation and projection (UMAP) clustering in 2D (X1, X2) of T-cell samples from early predictive from nuclear magnetic resonance (NMR) and cytokine media features at day 6 of T-cell culturing (formate, lactate, histidine, ethanol, dimethylamine, branch chain amino acids (BCAAs), glucose, glutamine, TNFα, IL2R, IL4, IL17a, IL13, IL15, and GM-CSF): for (a) ratio $CD4^+$ to $CD8^+$ $T_N + T_{CM}$, (b) total live $CD4^+$ $T_N + T_{CM}$ cells, and (c) total live $CD8^+$ $T_N + T_{CM}$ cells

Single-omics media analysis for early prediction

ML models using solely media cytokine profiles at day 6 reached similar or higher R^2 than those of the multiomics models ($CD4^+$ $T_N + T_{CM}$: 71.4%–99.9%; $CD4^+/CD8^+$: 83.4%–99.7%).

However, CD8$^+$ $T_N + T_{CM}$ still had variable LOO-R^2, 7.8%–93%. Overall, higher cytokine media profiles showed higher CD4$^+$ $T_N + T_{CM}$ and consequently its ratio with CD8$^+$ (**Figure 4.4a**). This behavior was evident, even beyond day 6, for TNFα, IL2R, IL17a, and IL4 which were frequently selected as predictive features across models (**Figures 4.4b,c** and **S3**g–i). A more complex behavior was detected for CD8$^+$ $T_N + T_{CM}$ which cannot be explained by cytokine secretion alone (**Figure 4.4d**).

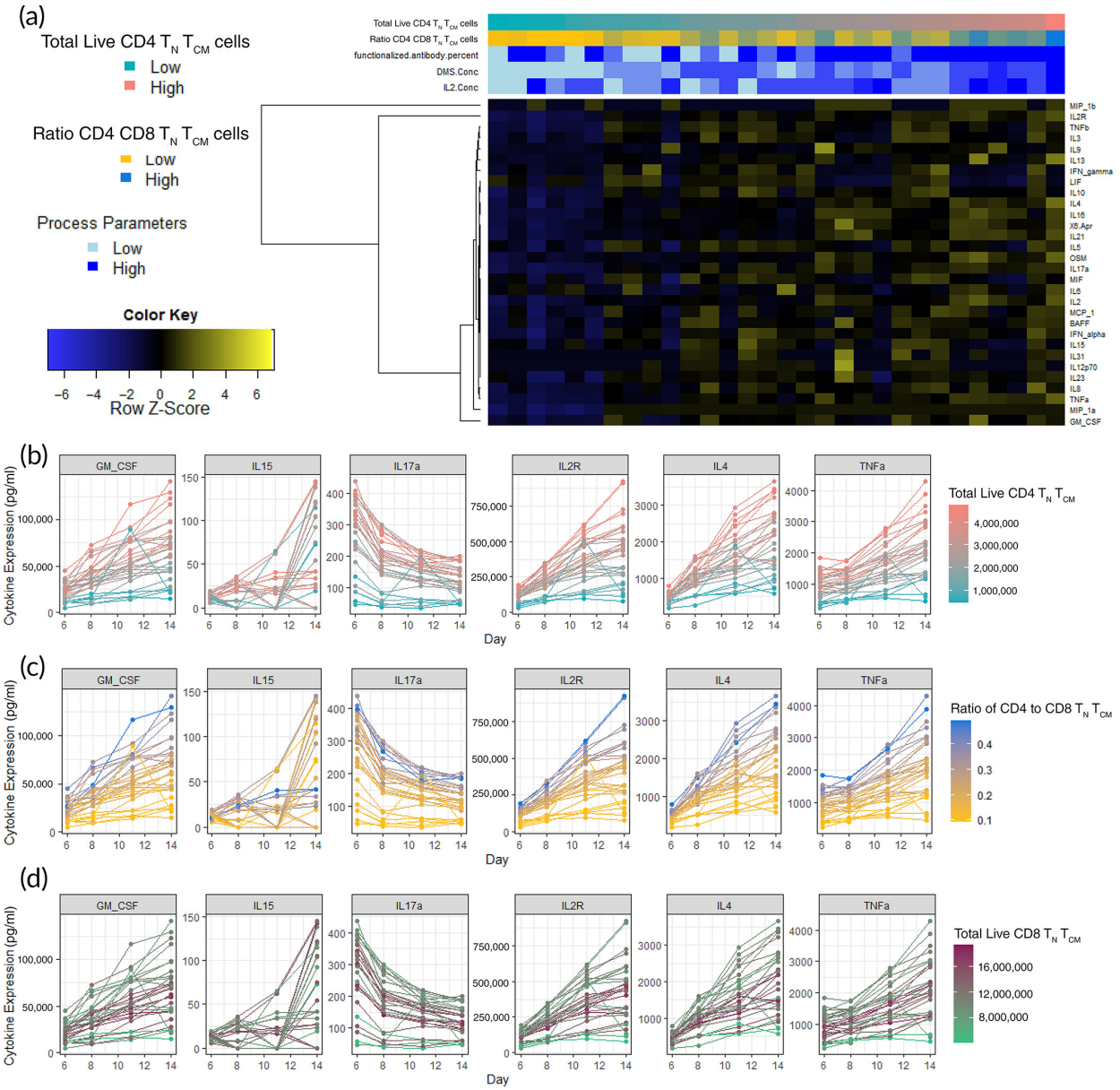


Figure 4.4 General characteristics of cytokine media profiles. (a) Heatmap for cytokine profiles from media samples on day 6. Expression in picograms/milliliter across time points for relevant

cytokine features for (b) ratio CD4$^+$ to CD8$^+$ $T_N + T_{CM}$ cells, (c) total live CD4$^+$ $T_N + T_{CM}$ cells, and (d) total live CD8$^+$ $T_N + T_{CM}$ cells

Models using only NMR media intensities on day 6 revealed an R^2 decrease of 8.8% and 11.1%, on average, compared with the multiomics and cytokine models, respectively. Yet, SR, GBM, and PLSR reached high LOO-R^2 (92.1%–99%), specifically for CD4$^+$/CD8$^+$ and CD4$^+$ $T_N + T_{CM}$. Although good prediction was achieved with NMR media analysis on day 6, we obtain slightly better predictions with NMR media analysis on day 4 (**Table 4.1**). From these models, formate, lactate, DMS concentration were highly ranked to predict both, ratio CD4$^+$/CD8$^+$ and CD4$^+$ $T_N + T_{CM}$ (Figure **S3**a–f). Some variable combinations also contained histidine, ethanol, dimethylamine, branch chain amino acids (BCAAs), glucose, and glutamine (Table **S3**). Lower intensity values for BCAAs, dimethylamine, glucose, and glutamine displayed higher CD4$^+$ $T_N + T_{CM}$ cells across the different media monitoring times (Figure **S5**a). Inversely, higher intensities of formate and lactate showed higher CD4$^+$ $T_N + T_{CM}$ and its ratio with CD8$^+$ consistently across time (**Figure 4.5a,b**).

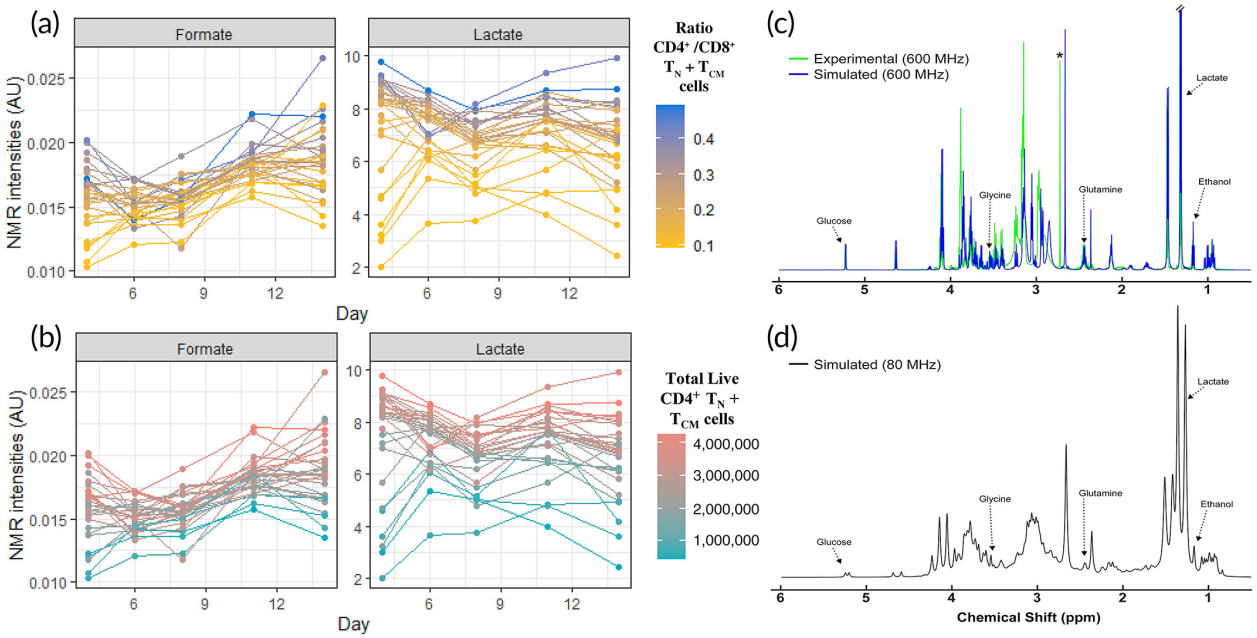


Figure 4.5 Top-performing features nuclear magnetic resonance (NMR) media analysis. NMR intensities in arbitrary units (AU) across time points for (a) ratio $CD4^+/CD8^+$ $T_N + T_{CM}$ cells, and (b) total live $CD4^+$ $T_N + T_{CM}$ cells. (c) Simulation of ^1H NMR spectrum shows the potential to detect multiple predictive features at lower magnetic fields. Overlay of a pooled experimental spectrum of T-cell culture medium (green) and GISSMO[19, 20] simulated spectrum (blue), composed of 19 compounds that reasonably approximate the experimental spectrum acquired at 600 MHz. Asterisk indicates an unknown feature of high intensity that was simulated with 2,3-dimethylamine (blue feature to right). Annotated features in the spectrum correspond to those identified as being highly predictive of output responses across computational methods. (d) GISSMO[19, 20] simulated spectrum at 80 MHz, corresponding to a field strength of commercially available benchtop NMR systems

The initial screening of a few samples from a different experimental batch shows much lower values of $T_N + T_{CM}$ responses but maintains a similar NMR and cytokine media patterns as the DOE and ADOE experiments (lower value intensities/secretion, lower $T_N + T_{CM}$ response) in terms of the total live $T_N + T_{CM}$ cells for $CD4^+$ and $CD8^+$. However, the decay in total live $T_N + T_{CM}$ cells for $CD8^+$ is much rapid than $CD4^+$ which makes the ratio behave in a more complex behavior (Figures **S7** and **S8**).

Discussion

CPP's understanding is critical to new product development and, especially in cell therapy development, it can have life-saving implications. The challenges for effective modeling grow with the increasing complexity of processes due to high dimensionality, and the potential for process interactions and nonlinear relationships. Another critical challenge is the limited amount of available data, mostly small DOE data sets. SR has the necessary capabilities to resolve the issues of process effects modeling and has been applied across multiple industries.[21] SR discovers mathematical expressions that fit a given sample and differs from conventional regression techniques in that a model structure is not defined a priori.[22] Hence, a key advantage of this methodology is that transparent, human-interpretable models can be generated from small and large data sets with no prior assumptions.[23, 24]

Since the model search process lets the data determine the model, diverse and competitive (e.g., accuracy and complexity) model structures are typically discovered. An ensemble of diverse models can be formed where its constituent models will tend to agree when constrained by observed data yet diverge in new regions. Collecting data in these regions helps to ensure that the target system is accurately modeled, and its optimum is accurately located.[23, 24] Exploiting these features allows adaptive data collection and interactive modeling. Consequently, this adaptive-DOE approach is useful in a variety of scenarios, including maximizing model validity for model-based decision making, optimizing processing parameters to maximize target yields, and developing emulators for online optimization and human understanding.[23, 24]

An in-depth characterization of potential DMS-based T-cell CQAs includes a list of cytokine and NMR features from media samples that are crucial in many aspects of T-cell fate decisions and effector functions of immune cells. Cytokine features were observed to slightly improve prediction and dominated the ranking of important features and variable combinations when modeling together with NMR media analysis and process parameters (**Figure 4.2a,b**). Predictive cytokine features such as TNFα, IL2R, IL4, IL17a, IL13, and IL15 were biologically assessed in terms of their known functions and activities associated with T cells. T helper cells secrete more cytokines than T cytotoxic cells, as per their main functions, and activated T cells secrete more cytokines than resting T cells. It is possible that some cytokines simply reflect the $CD4^+/CD8^+$ ratio and the activation degree by proxy proliferation. However, the exact ratio of expected cytokine abundance is less clear and depends on the subtypes present, and thus examination of each relevant cytokine is needed.

IL2R is secreted by activated T cells and binds to IL2, acting as a sink to dampen its effect on T cells.[25] Since IL2R was much greater than IL2 in solution, this might reduce the overall effect

of IL2, which could be further investigated by blocking IL2R with an antibody. In T cells, TNF can increase IL2R, proliferation, and cytokine production.[25] It may also induce apoptosis depending on concentration and alter the $CD4^+$ to $CD8^+$ ratio.[26] Given that TNF has both a soluble and membrane-bound form, this may either increase or decrease $CD4^+$ ratio and/or memory T cells depending on the ratio of the membrane to soluble TNF.[27] Since only soluble TNF was measured, membrane TNF is needed to understand its impact on both $CD4^+$ ratio and memory T cells. Furthermore, IL13 is known to be critical for Th2 response and therefore could be secreted if there are significant Th2 T cells already present in the starting population.[28] This cytokine has limited signaling in T cells and is thought to be more of an effector than a differentiation cytokine.[29] This feature might be emerging as relevant due to an initially large number of Th2 cells or because Th2 cells were preferentially expanded; indeed, IL4 is the conical cytokine that induces Th2 cell differentiation and was observed to be an important variable (Figure **2b,c**). The role of these cytokines could be investigated by quantifying the Th1/2/17 subsets both in the starting population and longitudinally. Similar to IL13, IL17 is an effector cytokine produced by Th17 cells[30] thus may reflect the number of Th17 subset of T cells. GM-CSF has been linked with activated T cells, specifically Th17 cells, but it is not clear if this cytokine is inducing differential expansion of $CD8^+$ T cells or if it is simply a covariate with another cytokine inducing this expansion.[31] Finally, IL15 has been shown to be essential for memory signaling and effective in skewing CAR-T cells toward the Tscm phenotype when using membrane-bound IL15Ra and IL15R.[32] Its high predictive behavior goes with its ability to induce large numbers of memory T cells by functioning in an autocrine/paracrine manner and could be explored by blocking either the cytokine or its receptor.

Similarly, literature suggests that many of the predictive metabolites found here are consistent with metabolic activity associated with T-cell activation and differentiation, yet it is not

clear how the various combinations of metabolites relate with each other in a heterogeneous cell population and should be explored. Formate and lactate were found to be highly predictive and observed to positively correlate with higher values of total live $CD4^+$ $T_N + T_{CM}$ cells (**Figures 4.5a,b** and **S6**). Formate is a byproduct of the one-carbon cycle implicated in promoting T-cell activation.[33] Importantly, this cycle occurs between the cytosol and mitochondria of cells and formate excreted.[34] Mitochondrial biogenesis and function have been shown necessary for memory cell persistence.[35, 36] Therefore, increased formate in media could be an indicator of one-carbon metabolism and mitochondrial activity in the culture.

In addition to formate, lactate was found as a putative CQA of $T_N + T_{CM}$. Lactate is the end-product of aerobic glycolysis, characteristic of highly proliferating cells and activated T cells.[37, 38] Glucose import and glycolytic genes are immediately upregulated in response to T-cell stimulation and thus the generation of lactate. At earlier time points, this abundance suggests a more robust induction of glycolysis and higher overall T-cell proliferation. Interestingly, our models indicate that higher lactate predicts higher $CD4^+$, both in total and in proportion to $CD8^+$, seemingly contrary to previous studies showing that $CD8^+$ T cells rely more on glycolysis for proliferation following activation.[39] It may be that glycolytic cells dominate in the culture at the early time points used for prediction, and higher lactate reflects more cells.

Ethanol patterns are difficult to interpret since its production in mammalian cells is still poorly understood.[40] Fresh media analysis indicates ethanol presence in the media used, possibly utilized as a carrier solvent for certain formula components. However, this does not explain the high variability and trend of ethanol abundance across time (Figure **S5**). As a volatile chemical, variation could be introduced by sample handling throughout the analysis process. Nonetheless, it is also possible that ethanol excreted into media over time, impacting processes regulating redox

and reactive oxygen species which have previously been shown to be crucial in T-cell signaling and differentiation.[41]

Metabolites that consistently decreased over time are consistent with the primary carbon source (glucose) and essential amino acids (BCAA, histidine) that must be continually consumed by proliferating cells. Moreover, the inclusion of glutamine in our predictive models also suggests the importance of other carbon sources for certain T-cell subpopulations. Glutamine can be used for oxidative energy metabolism in T cells without the need for glycolysis.[39] Overall, these results are consistent with existing literature that show different T-cell subtypes require different relative levels of glycolytic and oxidative energy metabolism to sustain the biosynthetic and signaling needs of their respective phenotypes.[42, 43] It is worth noting that the trends of metabolite abundance here are potentially confounded by the partial replacement of media that occurred periodically during expansion (see the Materials and Methods section), thus likely diluting some metabolic byproducts (i.e., formate and lactate) and elevating depleted precursors (i.e., glucose and amino acids). More definitive conclusions of metabolic activity across the expanding cell population can be addressed by a closed system, ideally with online process sensors and controls for formate, lactate, along with ethanol and glucose.

We demonstrated the ability to identify predictive markers using high-magnetic field NMR spectrometers. However, these are expensive, require a significant amount of resources to house and maintain, and would be the unlikely option for routine monitoring in industrial cell-manufacturing. Another common method, liquid chromatography (LC) coupled to mass spectrometry, has the advantage of a relatively smaller footprint and less upfront cost but it has other drawbacks such as destruction of the sample and difficulty with components in culture media that damage LC columns without extraction. Nevertheless, methods like continuous closed-loop

sampling are being developed to address this and might be readily available in the future.[44] Recently, permanent magnet-based NMR spectrometers (benchtop-size) have become available at a lower cost. Many of these are readily configured for flow-through reaction monitoring, which can be leveraged in a closed-cell manufacturing process. To explore the feasibility of such system, we utilized a spectral simulation to evaluate if putative CQAs identified here could theoretically be observed and quantified at a magnetic field strength of 80 MHz (common commercial benchtop systems). First, the experimental data acquired at 600 MHz was approximated by creating a simulated mixture of identified metabolites (**Figure 4.5c**) and then simulated at 80 MHz (**Figure 4.5d**). While the spectral resolution is significantly reduced compared to a spectrum at high-field, there are still numerous features that can be attributed to unique metabolites, including those identified as highly predictive (**Figure 4.5c,d**). Although this is promising, there will be challenges to acquiring high-quality data in a closed bioreactor system, that is, cells/DMS-particles present in suspension, final media formulation dictated by the amount of spectral complexity/overlap, and accurate quantitation of features with high overlap from other signals. However, a dedicated benchtop NMR coupled to a bioreactor could provide a simple system for real-time monitoring of CQAs.

Conclusions

Henceforth, this two-phase approach enabled in-depth characterization and identification of potential CQAs and CPPs for T cells. More sampling is needed to explore aspects like donor-to-donor variability or orthogonal behaviors from failed expansions when available it can be incorporated into this workflow which will be enriched due to its data-driven iterative design that fine-tunes model parameters as more data fit back into it, providing a powerful framework to optimize a complex experimental space during the cell-manufacturing process, and to facilitate the

identification of CPPs and early predictive CQAs from multiomics, which can be used broadly in the cell therapy and regenerative medicine field to accurately predict end-of-manufacturing quality at early stages.

The workflow and methods developed here could eventually allow manufacturers to identify deviations and problems with a manufacturing batch early during the culture and potentially implement corrective in-process controls. This could provide a more thorough understanding of the process parameters and their influence on end-product quality, and allow manufacturers to reduce batch failures, and thus improve cost, reduce risk, and increase access to cell-based therapies.

Materials and methods

Microcarrier fabrication

DMSs were fabricated as previously described.[18] To vary the surface concentration of the antibodies, the anti-CD3/anti-CD28 mAb mixture was further combined with a biotinylated isotype control to reduce the overall fraction of targeted mAbs. All mAbs were low endotoxin azide-free (Biolegend custom, LEAF specification). The surface concentration of the antibodies was quantified as previously described using a bicinchoninic acid assay kit (Thermo Fisher 23227).[18] See Supplementary Methods.

T-cell culture including sample collection

Cryopreserved primary human T cells were obtained as sorted CD3 subpopulations (Astarte Biotech). T cells were activated by adding DMSs (amount specified by the DOE) at day 0 of culture immediately after thaw. DMSs were not added or removed during the culture and had antibodies that were conjugated in proportions specified by the DOE. Initial cell density was 2.0×10^6 cells/ml in a 96-well plate with 300 µl volume. Media was serum-free TexMACS

(Miltentyi Biotech 170-076-307) supplemented with recombinant human IL2 in concentrations specified by the DOE (Peprotech 200-02). Cell cultures were expanded for 14 days as counted from the time of initial seeding and activation. Cell counts and viability were assessed using acridine orange/propidium iodide (AO/PI) and a Countess Automated Cell Counter (Thermo Fisher). Media was added to cultures every 2 days to 3 days in a 3:1 ratio (new volume:old volume) or based on a 300 mg/dl glucose threshold. The ADOE was done using the same feeding schedule as the initial DOE to maintain consistency for validation. Media glucose was measured using a ChemGlass glucometer to confirm cell growth and activation.

Flow cytometry

At the end of culture, at least 1e5 T cells from each run were washed with PBS once, resuspended in PBS, and stained with Zombie UV (Biolegend, 423107) for 30 min at room temperature in the dark at a 1:1000 dilution. Cells were spun and resuspended in FACS buffer (1X PBS, 2% bovine serum albumin, 5 mM EDTA) and were stained with antibodies according to **Table S1** for 60 min in the dark at 4°C. Cells were then resuspended in fresh FACS buffer, after which they were run on a BD LSR ortessa. All stained was performed in a 96 well v-bottom plate. See **Supplementary Methods**.

Cytokine measurements

Cytokines were measured using a custom ProcartaPlex Luminex kit (Thermo Fisher). The assay was performed using media samples taken at various time points throughout the T-cell culture according to the manufacturer's instructions with modifications to half the reagent requirements. Data available at Supporting **Dataset S1**. See **Supplementary Methods**.

NMR metabolomics sample preparation

Fifty microliter of media was collected from each culture at each time point (before media exchange, if applicable), flash-frozen in liquid nitrogen, and stored at −80°C. Samples were shipped to CCRC on dry ice for NMR analysis. Run order of samples was randomized. Samples were prepared in two batches for each rack of NMR samples to be run. For each rack, samples were pulled and sorted on dry ice, then thawed at 4°C for 1 h. Samples were then centrifuged at 2990 × *g* at 4°C for 20 min to pellet any cells or debris that may have been collected with the media. 5 µl of 100/3 mM DSS-D6 in deuterium oxide (Cambridge Isotope Laboratories) were added to 1.7 mm NMR tubes (Bruker BioSpin), followed by 45 µl of media from each sample that was added and mixed, for a final volume of 50 µl in each tube. Samples were prepared on ice and in predetermined, randomized order. The remaining volume from each sample in the rack (~4 µl) was combined to create an internal pool. This material was used for internal controls within each rack as well as metabolite annotation.

NMR data collection and processing

NMR spectra were collected on a Bruker Avance III HD spectrometer at 600 MHz using a 5-mm TXI cryogenic probe and TopSpin software (Bruker BioSpin). One-dimensional spectra were collected on all samples using the noesypr1d pulse sequence under automation using ICON NMR software. Two-dimensional (2D) HSQC and TOCSY spectra were collected on internal pooled control samples for metabolite annotation. One-dimensional spectra were manually phased and baseline corrected in TopSpin. 2D spectra were processed in NMRpipe.[45] One dimensional spectra were referenced, water/end regions removed, and normalized with the PQN algorithm[46] using an in-house MATLAB (The MathWorks, Inc.) toolbox (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA).

NMR feature selection

To reduce the total number of spectral features from approximately 250 peaks and enrich for those that would be most useful for statistical modeling, a variance-based feature selection was performed within MATLAB. For each digitized point on the spectrum, the variance was calculated across all experimental samples and plotted. Clearly resolved features corresponding to peaks in the variance spectrum were manually binned and integrated to obtain quantitative feature intensities across all samples (Figure **S4**). In addition to highly variable features, several other clearly resolved and easily identifiable features were selected (glucose, BCAA region, etc.). Some features were later discovered to belong to the same metabolite but were included in further analysis. Data are available at **Dataset S1**.

Metabolite annotation

2D spectra collected on pooled samples were uploaded to COLMARm web server,[47] where HSQC peaks were automatically matched to database peaks. HSQC matches were manually reviewed with additional 2D and proton spectra to confirm the match. Annotations were assigned a confidence score based upon the levels of spectral data supporting the match as previously described.[48] Annotated metabolites were matched to previously selected features used for statistical analysis. Several low abundance features selected for analysis did not have database matches and were not annotated.

Low-field spectrum simulation

Using the list of annotated metabolites obtained above, an approximation of a representative experimental spectrum was generated using the GISSMO mixture simulation tool.[19, 20] With the simulated mixture of compounds, generated at 600 MHz to match the experimental data, a new simulation was generated at 80 MHz to match the field strength of commercially available benchtop NMR spectrometers. The GISSMO tool allows visualization of

signals contributed from each individual compound as well as the mixture, which allows annotation of features in the mixture belonging to specific compounds.

ML modeling

Seven ML techniques were implemented to predict T_N and T_{CM} responses related to the memory phenotype of the cultured T cells under different process parameters conditions. The ML methods executed were RF, GBM, CIF, LASSO, PLSR, SVM, and SR. Primarily, SR models were used to optimize process parameter values based on $T_N + T_{CM}$ phenotype and to extract early predictive variable combinations from the multiomics experiments. SR was done using Evolved Analytics' Data Modeler software (Evolved Analytics LLC). While nonparametric tree-based ensembles were done through the *randomForest*, *gbm*, and *cforest* regression functions in R, for RF, gradient boosted trees, and CIF models, respectively. Prediction performance was evaluated using LOO-R^2 and permutation-based variable importance scores assessing % increase of mean squared errors, relative influence based on the increase of prediction error, coefficient values for RF, GBM, and CID, respectively. Partial least squares regression was executed using the *plsr* function from the *pls* package in R while LASSO regression was performed using the *cv.glmnet* R package, both using leave-one-out cross-validation. Finally, the *kernlab* R package was used to construct the SVM regression models. Parameter tuning was done for all models in a grid search manner using the *train* function from the *caret* R package using LOO-R^2 as the optimization criteria. Prediction performance was measured for all models using the final model with LOO-R^2 tuned parameters. More details at **Table S2**. See Supplementary Methods.

ML consensus analysis

Consensus analysis of the relevant variables extracted from each ML model was done to identify consistent predictive features of quality at the early stages of manufacturing. Using

importance scores, key predictive variables were selected if their importance scores were within the 80th percentile ranking for the following ML methods: RF, GBM, CIF, LASSO, PLSR, SVM while for SR variables present in >30% of the top-performing SR models from Data Modeler ($R^2 \geq 90\%$, complexity ≤ 100) were chosen to investigate consensus. Only variables with those high percentile scoring values were evaluated in terms of their logical relation (intersection across ML models). See Supplementary Methods.

Acknowledgments

The material is based upon work supported by the National Science Foundation under Grant No. EEC-1648035. The work and views presented are those of the authors and do not reflect the views of the National Science Foundation. The research work from Nathan J. Dwarshuis and Krishnendu Roy was also partially supported by funds from The Billie and Bernie Marcus Foundation, The Georgia Research Alliance, and the Georgia Tech Foundation through their support of the Marcus Center for Therapeutic Cell Characterization and Manufacturing (MC3M) at Georgia Tech. Nathan J. Dwarshuis would like to thank Melissa Kemp for access to the Bioplex 200 machine and to Levi Wood/Laura Weinstock for the optimized Luminex protocol. Maxwell B. Colonna would like to thank Hesam Dashti for assistance with getting additional GISSMO compound entries and simulation frequencies uploaded to enable the mixture simulation.

Conflict of interest

Bruce L. Levine declares financial interest intellectual property and patents in the field of cell and gene therapy (University of Pennsylvania Alliance with Novartis, licensing, and royalty fees). Bruce L. Levine is a consultant for Novartis, Terumo, and Lilly Asia Ventures and he is part of the Scientific Advisory Board for Avectas, Brammer Bio/TF Viral Vector Services, Immuneel, Incysus, Ori Biotech, and Vycellix. Moreover, Bruce L. Levine is the co-founder and equity holder

Tmunity Therapeutics and all of his conflict of interest is managed in accordance with University of Pennsylvania policy and oversight. Theresa Kotanchek is the Chief Executive Officer of Evolved Analytics, LLC. The remaining authors declare no competing interests. Krishnendu Roy declares consulting, intellectual property, and patents in cell and gene therapy. Krishnendu Roy is a consultant to Terumo, Merck, LEK consulting, Mubadala Ventures, Anzu Partners, Decibio, and Clearview Healthcare Partners. Krishnendu Roy also serves on the advisory board of the MIT-Singapore Cell therapy Partnership.

References

1. Fesnak AD, June CH, Levine BL. Engineered T cells: the promise and challenges of cancer immunotherapy. *Nat Rev Cancer*. 2016;16(9):566-581. doi: 10.1038/nrc.2016.97
2. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science (New York, NY)*. 2015;348(6230):62-68. doi: 10.1126/science.aaa4967
3. Dwarshuis NJ, Parratt K, Santiago-Miranda A, Roy K. Cells as advanced therapeutics: state-of-the-art, challenges, and opportunities in large scale biomanufacturing of high-quality cells for adoptive immunotherapies. *Adv Drug Deliv Rev*. 2017;114:222-239. doi: 10.1016/j.addr.2017.06.005
4. Roddie C, O'Reilly M, Pinto JDA, Vispute K, Lowdell M. Manufacturing chimeric antigen receptor T cells: issues and challenges. *Cytotherapy*. 2019;21:327-340. doi: 10.1016/j.jcyt.2018.11.009
5. Roh K-H, Nerem RM, Roy K. Biomanufacturing of therapeutic cells: state of the art, current challenges, and future perspectives. *Annu Rev Chem Biomol Eng*. 2016;7(1):455-478. doi: 10.1146/annurev-chembioeng-080615-033559
6. Carmen J, Burger SR, McCaman M, Rowley JA. Developing assays to address identity, potency, purity and safety: cell characterization in cell therapy process development. *Regen Med*. 2012;7(1):85-100. doi: 10.2217/rme.11.105
7. Simon CG, Lin-Gibson S, Elliott JT, Sarkar S, Plant AL. Strategies for achieving measurement assurance for cell therapy products. *Stem Cells Transl Med*. 2016;5(6):705-708. doi: 10.5966/sctm.2015-0269

8. Campbell A, Brieva T, Raviv L, et al. Concise review: process development considerations for cell therapy. *Stem Cells Transl Med*. 2015;4(10):1155-1163. doi: 10.5966/sctm.2014-0294
9. Lipsitz YY, Timmins NE, Zandstra PW. Quality cell therapy manufacturing by design. *Nat Biotechnol*. 2016;34(4):393-400. doi: 10.1038/nbt.3525
10. Better M, Chiruvolu V, Sabatino M. Overcoming challenges for engineered autologous T cell therapies. *Cell Gene Ther Insights*. 2018;4(4):173-186.
11. Tyagarajan S, Spencer T, Smith J. Optimizing CAR-T cell manufacturing processes during pivotal clinical trials. *Mol Ther Methods Clin Dev*. 2020;16:136-144. doi: 10.1016/j.omtm.2019.11.018
12. Xu Y, Zhang M, Ramos CA, et al. Closely related T-memory stem cells correlate with in vivo expansion of CAR.CD19-T cells and are preserved by IL-7 and IL-15. *Blood*. 2014;123(24):3750-3759. doi: 10.1182/blood-2014-01-552174
13. Gattinoni L, Klebanoff CA, Restifo NP. Paths to stemness: building the ultimate antitumour T cell. *Nat Rev Cancer*. 2012;12(10):671-684. doi: 10.1038/nrc3322
14. Fraietta JA, Lacey SF, Orlando EJ, et al. Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia. *Nat Med*. 2018;24(5):563-571. doi: 10.1038/s41591-018-0010-1
15. Gattinoni L, Lugli E, Ji Y, et al. A human memory T cell subset with stem cell–like properties. *Nat Med*. 2011;17(10):1290-1297. doi: 10.1038/nm.2446
16. Wang D, Aguilar B, Starr R, et al. Glioblastoma-targeted CD4+ CAR T cells mediate superior antitumor activity. *JCI Insight*. 2018;3(10):e99048. doi: 10.1172/jci.insight.99048.

17. Yang Y, Kohler ME, Chien CD, et al. TCR engagement negatively affects CD8 but not CD4 CAR T cell expansion and leukemic clearance. *Sci Transl Med*. 2017;9(417):eaag1209. doi: 10.1126/scitranslmed.aag1209
18. Dwarshuis NJ, Song HW, Patel A, Kotanchek T, Roy K. Functionalized microcarriers improve T cell manufacturing by facilitating migratory memory T cell production and increasing CD4/CD8 ratio. *bioRxiv*. 2019;646760. doi: 10.1101/646760
19. Dashti H, Westler WM, Tonelli M, Wedell JR, Markley JL, Eghbalnia HR. Spin system modeling of nuclear magnetic resonance spectra for applications in metabolomics and small molecule screening. *Anal Chem*. 2017;89(22):12201-12208. doi: 10.1021/acs.analchem.7b02884
20. Dashti H, Wedell JR, Westler WM, et al. Applications of parametrized NMR spin systems of small molecules. *Anal Chem*. 2018;90(18):10646-10649. doi: 10.1021/acs.analchem.8b02660
21. Kordon AK, Lue C-T. Symbolic regression modeling of blown film process effects. *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*. Vol 1. IEEE; 2004:561-568. doi: 10.1109/CEC.2004.1330907
22. Koza JR. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*. 1994;4(2):87. 10.1007/bf00175355
23. Kotanchek M, Smits G, Vladislavleva E. Pursuing the Pareto paradigm: tournaments, algorithm variations and ordinal optimization. In: Riolo R, Soule T, Worzel B, eds. *Genetic Programming Theory and Practice IV. Genetic and Evolutionary Computation*. Springer; 2007:167-185. doi: 10.1007/978-0-387-49650-4_1

24. Kotanchek M, Smits G, Vladislavleva E. Exploiting trustable models via Pareto GP for targeted data collection. *Genetic Programming Theory and Practice VI. Genetic and Evolutionary Computation*. Springer; 2009:1-18. doi: 10.1007/978-0-387-87623-8_10

25. Witkowska AM. On the role of sIL-2R measurements in rheumatoid arthritis and cancers. *Mediators Inflamm*. 2005;2005(3):121-130. doi: 10.1155/MI.2005.121

26. Vudattu NK, Holler E, Ewing P, et al. Reverse signalling of membrane-integrated tumour necrosis factor differentially regulates alloresponses of CD4+ and CD8+ T cells against human microvascular endothelial cells. *Immunology*. 2005;115(4):536-543. doi: 10.1111/j.1365-2567.2005.02190.x

27. Mehta AK, Gracias DT, Croft M. TNF activity and T cells. *Cytokine*. 2018;101:14-18. doi: 10.1016/j.cyto.2016.08.003

28. Wong FS. Stimulating IL-13 receptors on T cells: a new pathway for tolerance induction in diabetes? *Diabetes*. 2011;60(6):1657-1659. doi: 10.2337/db11-0353]

29. Junttila IS. Tuning the cytokine responses: an update on interleukin (IL)-4 and IL-13 receptor complexes. *Front Immunol*. 2018;9:888. doi: 10.3389/fimmu.2018.00888.

30. Amatya N, Garg AV, Gaffen SL. IL-17 signaling: the Yin and the Yang. *Trends Immunol*. 2017;38(5):310-322. doi: 10.1016/j.it.2017.01.006

31. Becher B, Tugues S, Greter M. GM-CSF: from growth factor to central mediator of tissue inflammation. *Immunity*. 2016;45(5):963-973. doi: 10.1016/j.immuni.2016.10.026

32. Hurton LV, Singh H, Najjar AM, et al. Tethered IL-15 augments antitumor activity and promotes a stem-cell memory subset in tumor-specific T cells. *Proc Natl Acad Sci USA*. 2016;113(48):E7788-E7797. doi: 10.1073/pnas.1610544113

33. Ron-Harel N, Santos D, Ghergurovich JM, et al. Mitochondrial biogenesis and proteome remodeling promotes one carbon metabolism for T cell activation. *Cell Metab*. 2016;24(1):104-117. doi: 10.1016/j.cmet.2016.06.007

34. Pietzke M, Meiser J, Vazquez A. Formate metabolism in health and disease. *Mol Metab*. 2020;33:23-37. doi: 10.1016/j.molmet.2019.05.012

35. van der Windt GJW, Everts B, Chang C-H, et al. Mitochondrial respiratory capacity is a critical regulator of CD8+ T cell memory development. *Immunity*. 2012;36(1):68-78. doi: 10.1016/j.immuni.2011.12.007

36. Vardhana SA, Hwee MA, Berisa M, et al. Impaired mitochondrial oxidative phosphorylation limits the self-renewal of T cells exposed to persistent antigen. *Nat Immunol*. 2020;13:1-12. doi: 10.1038/s41590-020-0725-2

37. Lunt SY, Vander Heiden MG. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annu Rev Cell Dev Biol*. 2011;27(1):441-464. doi: 10.1146/annurev-cellbio-092910-154237

38. Chang C-H, Curtis JD, Maggi LB, et al. Posttranscriptional control of T cell effector function by aerobic glycolysis. *Cell*. 2013;153(6):1239-1251. doi: 10.1016/j.cell.2013.05.016

39. Cao Y, Rathmell JC, Macintyre AN. Metabolic reprogramming towards aerobic glycolysis correlates with greater proliferative ability and resistance to metabolic inhibition in CD8 versus CD4 T cells. *PLoS One*. 2014;9(8):e104104. doi: 10.1371/journal.pone.0104104

40. Antoshechkin AG. On intracellular formation of ethanol and its possible role in energy metabolism. *Alcohol Alcohol*. 2001;36(6):608. doi: 10.1093/alcalc/36.6.608

41. Sena LA, Li S, Jairaman A, et al. Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. *Immunity*. 2013;38(2):225-236. doi: 10.1016/j.immuni.2012.10.020

42. Almeida L, Lochner M, Berod L, Sparwasser T. Metabolic pathways in T cell activation and lineage differentiation. *Semin Immunol*. 2016;28(5):514-524. doi: 10.1016/j.smim.2016.10.009

43. Wang R, Green DR. Metabolic checkpoints in activated T cells. *Nat Immunol*. 2012;13(10):907-915. doi: 10.1038/ni.2386

44. Chilmonczyk MA, Kottke PA, Stevens HY, Guldberg RE, Fedorov AG. Dynamic mass spectrometry probe for electrospray ionization mass spectrometry monitoring of bioreactors for therapeutic cell manufacturing. *Biotechnol Bioeng*. 2019;116(1):121-131. doi: 10.1002/bit.26832

45. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*. 1995;6(3):277-293. doi: 10.1007/BF00197809

46. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem*. 2006;78(13):4281-4290. doi: 10.1021/ac051632c

47. Bingol K, Li D-W, Zhang B, Brüschweiler R. Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. *Anal Chem*. 2016;88(24):12411-12418. doi: 10.1021/acs.analchem.6b03724

48. Walejko JM, Chelliah A, Keller-Wood M, Gregg A, Edison AS. Global metabolomics of the placenta reveals distinct metabolic profiles between maternal and fetal placental tissues following delivery in non-labored women. *Metabolites*. 2018;8(1). doi: 10.3390/metabo8010010

49. Sud M, Fahy E, Cotter D, et al. Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res*. 2016;44(D1):D463-D470. doi: 10.1093/nar/gkv1042

50. Holmes E, Cloarec O, Nicholson JK. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: application to HgCl2 toxicity. *J Proteome Res*. 2006;5(6):1313-1320. doi: 10.1021/pr050399w

CHAPTER 5

DISCUSSION AND FUTURE DIRECTIONS

I have shown my work in leveraging NMR metabolomics in cell models of disease and as a functional readout of manufactured cells. In this final chapter I will discuss the potential avenues for advancing metabolomics in cell models and cell manufacturing, including current work nearing completion.

Metabolomics in advanced cell systems

While most untargeted human metabolomics studies have been aimed towards finding biomarkers that can indicate disease, the use of cells has the ability to directly interrogate mechanisms of disease and metabolism.[1,2] However it is often difficult to accurately recreate, or even be aware of, all the relevant conditions to include in a cell model. While all cell models are inherently simplistic and imperfect representations, their advantages in being able to get at molecular mechanisms and generating hypotheses are still of huge value. With these considerations, there are steps that can be taken to improve the accuracy and relevance of metabolomics studies in cell systems.

The use of primary cells is probably the simplest way to get a little bit closer to physiological relevance in cell models. As illustrated in Chapter 3, there can be significant differences between metabolic adaptations in genome edited cell lines compared to patient derived cells. Cell lines have a theoretical advantage of being reproducible and comparable across experiments/labs, however the scientific community has largely been dissuaded from that illusion. In the end, primary cells probably generate more relevant results.

Many cells that are cultured in 2D are also able to be cultured in 3D, typically by embedding cells into a gel-like matrix. Altered cell-cell contacts and interactions with extracellular proteins/scaffolds in 3D cultures result in vastly different gene expression and metabolic profiles.[3] While 3D cell culture has been around for some time, the lack of standard and reproducible 3D matrix has also hampered its ability to produce consistent results. This problem is complicated further when considering biologically accurate matrices would also be highly dependent upon the cell type and tissue being simulated. Indeed, specific interactions with components of the microenvironment around cells determines the cell phenotype.[4] There are of course challenges to applying traditional metabolomics methods to these culture systems, such as attempting to recover cells from an extracellular matrix, or extracting metabolites from embedded cells, introduce additional sources of variance and bias through more sample handling and matrix effects

Alternatively, some cell types can spontaneously form organoids, which can more closely reflect tissue characteristics. These organoids are 3D multicellular colonies, and through their cell-cell interactions, polarization, and super structure develop properties closer to real tissue. One tradeoff here is organoid systems can blur the lines between intracellular and extracellular metabolites if they develop cysts where metabolites are being exchanged in a closed environment.

To add another layer of complexity, interactions between different cell types are also key to understanding most disease mechanisms. Thus, co-culture systems are also important applications of cell culture that are pose some difficulty to applying traditional metabolomics studies, as contributions from different cell types are difficult to distinguish. Particularly relevant to the work described here, are interactions between cancer cells and stromal cells [5] and between cancer cell and immune cell interactions. In fact, other co-authored works I have contributed to (in progress) have attempted to examine the extracellular environment of glioma organoid and CAR-

T cells. These analyses have proved difficult due to not being able to easily account for which cell type is causing observed changes in the media, in addition to how they are affecting each other. There have been methods proposed to be able to disentangle metabolite contributions of different cell types when both analyzed separately and in co-culture when part of the same study, and should be further explored for these types of interaction studies.[6] Being able to separate cells from a co-culture after an exposure period would be ideal for being able to more accurately understand how cell-cell interactions affect internal metabolism of individual cell types. A reproducible, 3D printed, perfusable organoid chip could be a solution for being able to more closely simulate a tissue microenvironment with multiple cell types.[7,8]

Another frontier in cell-based metabolomics is in both temporal and spatial resolution of metabolism, which is also hugely important to understanding and treating disease. Currently, there are methods being developed to address this and gain one or both of those dimensions of information.[9-11]

Future work developing better cell models for cancer and disease metabolism should strive to match as closely as possible the metabolic and transcription signatures of whole tissues, providing evidence for more accurate approximations. Ultimately, integrating these more relevant cell systems with biochemical genetics afforded by cell culture will provide the most informative knowledge that can be gained from these metabolomics studies. When combined with the unique capabilities of NMR as outlined in Chapter 1, there is immense potential for the use of NMR based metabolism studies in cell models as technology advances.

Future Directions Improving Cell Therapies

We've begun a new study to further demonstrate the use of NMR media analysis of cell cultures to predict cell product function and create hypotheses of mechanism of action, addressing

two vital needs in the cell therapy space. Here I will describe this ongoing work and summarize the results thus far.

Mesenchymal stromal cells are a promising cell therapy that has several potential uses, including modulating inflammation and immune response in tissues. So far it's applications have been limited by inconsistent results in trials, unknown mechanisms of action and no critical quality attributes that can be reliable indicators of product potency. Contrastingly to CAR T therapies described in Chapter 3, MSC therapies are primarily being developed as allogeneic therapies, or "off-the-shelf" banks of manufactured donor cells that can be given to many patients. Some of the inconsistency in effect is due to heterogeneity in donor cells that used for therapy, different donors can have different potency in different patients. Characterizing both this variation in different donor cells as well early markers predictive of potency irrespective of donor are crucial for improving these therapies.

To approach these goals, we expanded 10 donor MSC lines identically in a cell manufacturing facility at Georgia Tech University. Media was collected from each day of culture for analysis by NMR. At the endpoint of culture, the cells were harvested, and assays performed to assess function. The collected culture media from all cell lines and replicates were analyzed by NMR as described in Chapter 3 (Figure 5.1).

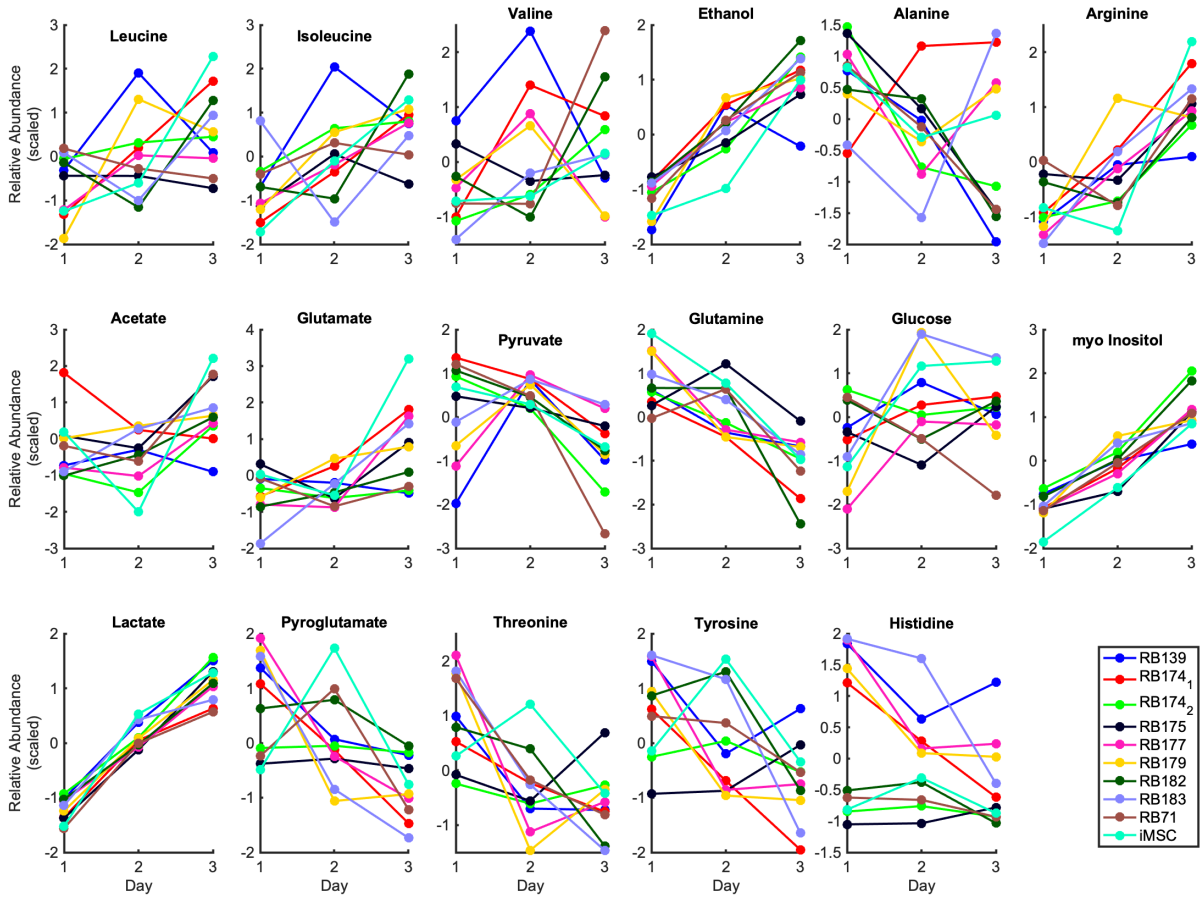


Figure 5.3 Metabolite abundance trajectories for 10 MSC donor lines over three days of continuous culture. Y axis is zero scaled

Interestingly, the trajectories of some NMR features/metabolites in the culture media do not appear to keep a linear trajectory, or indeed a monotonic trajectory. This suggests a higher time resolution is required to capture the true dynamics of certain metabolites in these early culture periods.

Features were taken from the media NMR spectra and used as input to predict the functional values of each donor cell line (Figure 5.2). To focus on the early predictive features, we looked specifically at media from the first three days of culture

A

LOOR2	ML						
	SR	PLSR	SVM	GBM	LASSO	RF	DT
NMR DAY 2 Change	97%	60%	71%	84%	54%	71%	47%
NMR DAY 3 Change	99%	93%	97%	89%	97%	73%	90%

B

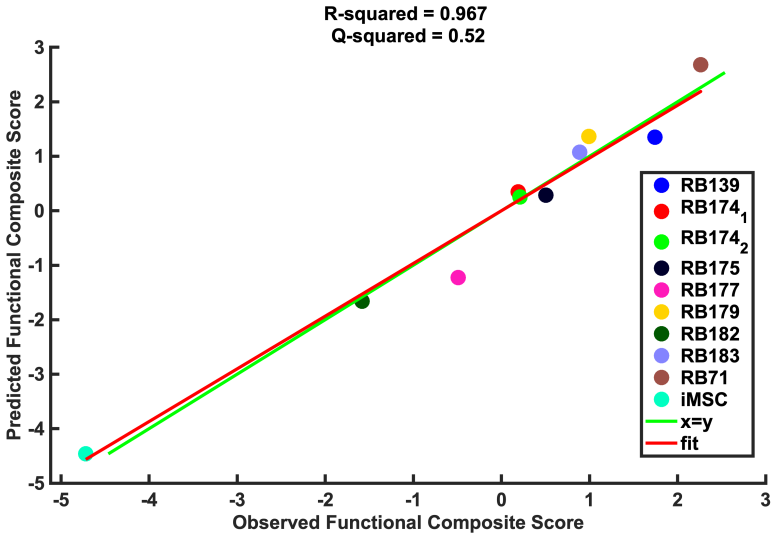


Figure 5.4 Machine Learning (ML) results modeling MSC function with NMR features. A) Table of leave-one-out-R2 performance values for the variable sets and machine learning methods indicated. B) Response plot of PLSR model showing model predicted functional composite scores against actual measured scores. SR = symbolic regress, PLSR = partial least squares regression, SVM = support vector machine, GBM = gradient boosted machine, LASSO = , RF = random forest, DT = decision tree.

I have been able to generate good models to predict the function of different MSC donor lines with changes in NMR feature intensity over two or three days of culture. These models can each provide information on which features are most important for their predictive ability. We are currently assessing these features to determine which pathways may be indicated as being most affected in good vs bad performing cell products.

As mentioned earlier in this dissertation, I believe the work exhibited here shows the potential for NMR as a method for continuous monitoring of culture media for cell manufacturing. Benchtop NMR, which is a more affordable and portable option to the high field instruments used

to conduct the work in this dissertation, could still provide useful information on-line for monitoring product quality (Figure 5.3).

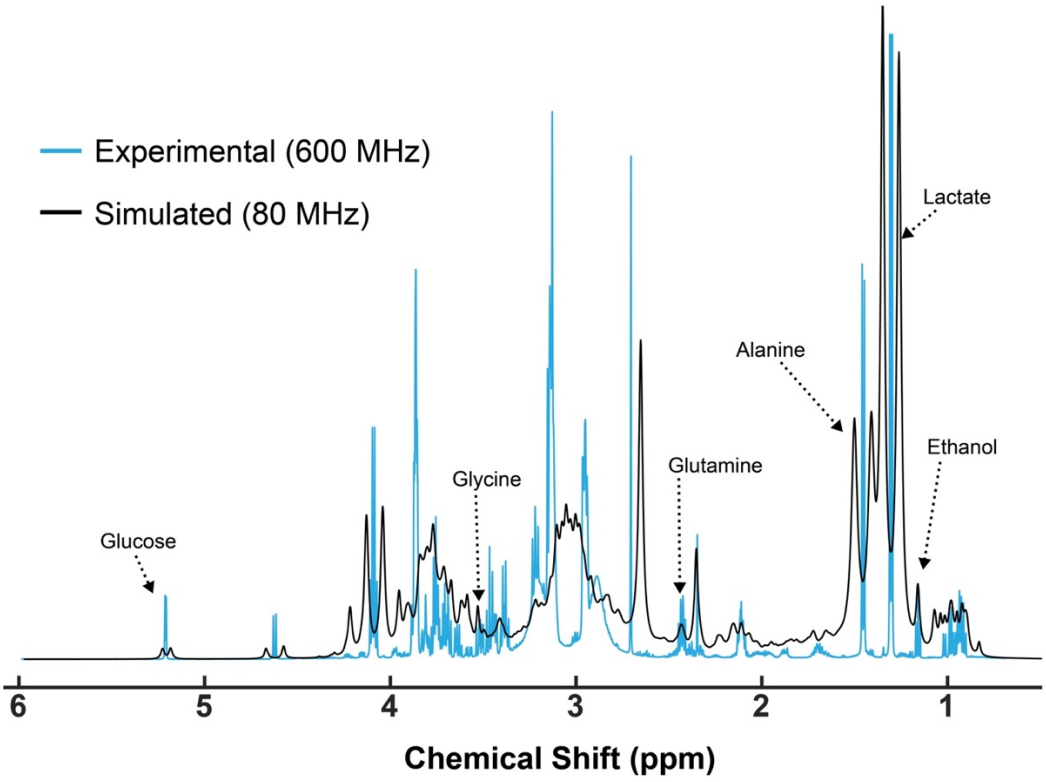


Figure 5.5 Comparison of experimental NMR spectrum at 600 MHz and simulated spectrum of the same mixture at 80 MHz. Peaks corresponding to annotated metabolites are labeled with arrows.

Concluding remarks

Realtime monitoring is something only feasible in cell culture which gives important information. Incorporation of other techniques such as isotopic labeling and use of more relevant culture systems could be leveraged to get detailed dynamics of metabolism under conditions of cell manufacturing, disease, or drug treatment. Understanding these dynamics will undoubtedly be invaluable for generating new knowledge of human health. I can envision a future of precision

medicine and cell manufacturing where patient cells are collected, grown and tested in physiologically relevant systems *ex vivo*, and monitored to understand the nature of their metabolism. As the basic knowledgebase continues to grow (including the examples in this dissertation), we will work towards identifying the metabolic patterns of different disease states will be essential to then identifying specific interventions and improving treatment outcomes.

References

(1) Downs, D. M.; Bazurto, J. V.; Gupta, A.; Fonseca, L. L.; Voit, E. O. *AIMS Microbiol* **2018**, *4*, 289-303.

(2) Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. *Nat Rev Mol Cell Biol* **2016**, *17*, 451-459.

(3) Wang, T.; Kwon, S. H.; Peng, X.; Urdy, S.; Lu, Z.; Schmitz, R. J.; Dalton, S.; Mostov, K. E.; Zhao, S. *iScience* **2020**, *23*, 101629.

(4) Krause, S.; Maffini, M. V.; Soto, A. M.; Sonnenschein, C. *Bmc Cancer* **2010**, *10*, 263.

(5) Martinez-Outschoorn, U. E.; Balliet, R. M.; Rivadeneira, D. B.; Chiavarina, B.; Pavlides, S.; Wang, C.; Whitaker-Menezes, D.; Daumer, K. M.; Lin, Z.; Witkiewicz, A. K.; Flomenberg, N.; Howell, A.; Pestell, R. G.; Knudsen, E. S.; Sotgia, F.; Lisanti, M. P. *Cell Cycle* **2010**, *9*, 3256-3276.

(6) Azzollini, A.; Boggia, L.; Boccard, J.; Sgorbini, B.; Lecoultre, N.; Allard, P. M.; Rubiolo, P.; Rudaz, S.; Gindro, K.; Bicchi, C.; Wolfender, J. L. *Front Microbiol* **2018**, *9*, 72.

(7) Tang, M.; Xie, Q.; Gimple, R. C.; Zhong, Z.; Tam, T.; Tian, J.; Kidwell, R. L.; Wu, Q.; Prager, B. C.; Qiu, Z.; Yu, A.; Zhu, Z.; Mesci, P.; Jing, H.; Schimelman, J.; Wang, P.; Lee, D.; Lorenzini, M. H.; Dixit, D.; Zhao, L., et al. *Cell Res* **2020**, *30*, 833-853.

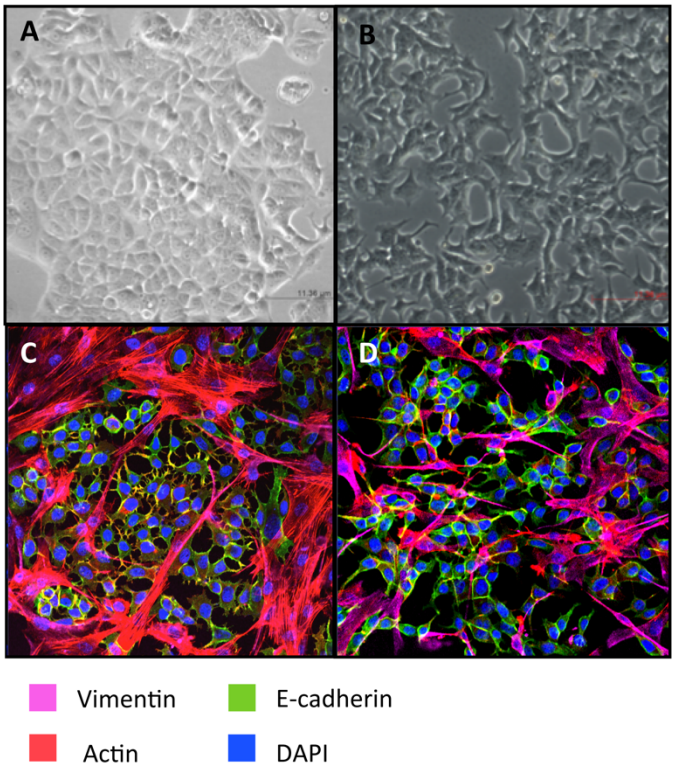
(8) Dornhof, J.; Kieninger, J.; Muralidharan, H.; Maurer, J.; Urban, G. A.; Weltin, A. *Lab Chip* **2022**, *22*, 225-239.

(9) Judge, M. T.; Wu, Y.; Tayyari, F.; Hattori, A.; Glushka, J.; Ito, T.; Arnold, J.; Edison, A. S. *Front Mol Biosci* **2019**, *6*, 26.

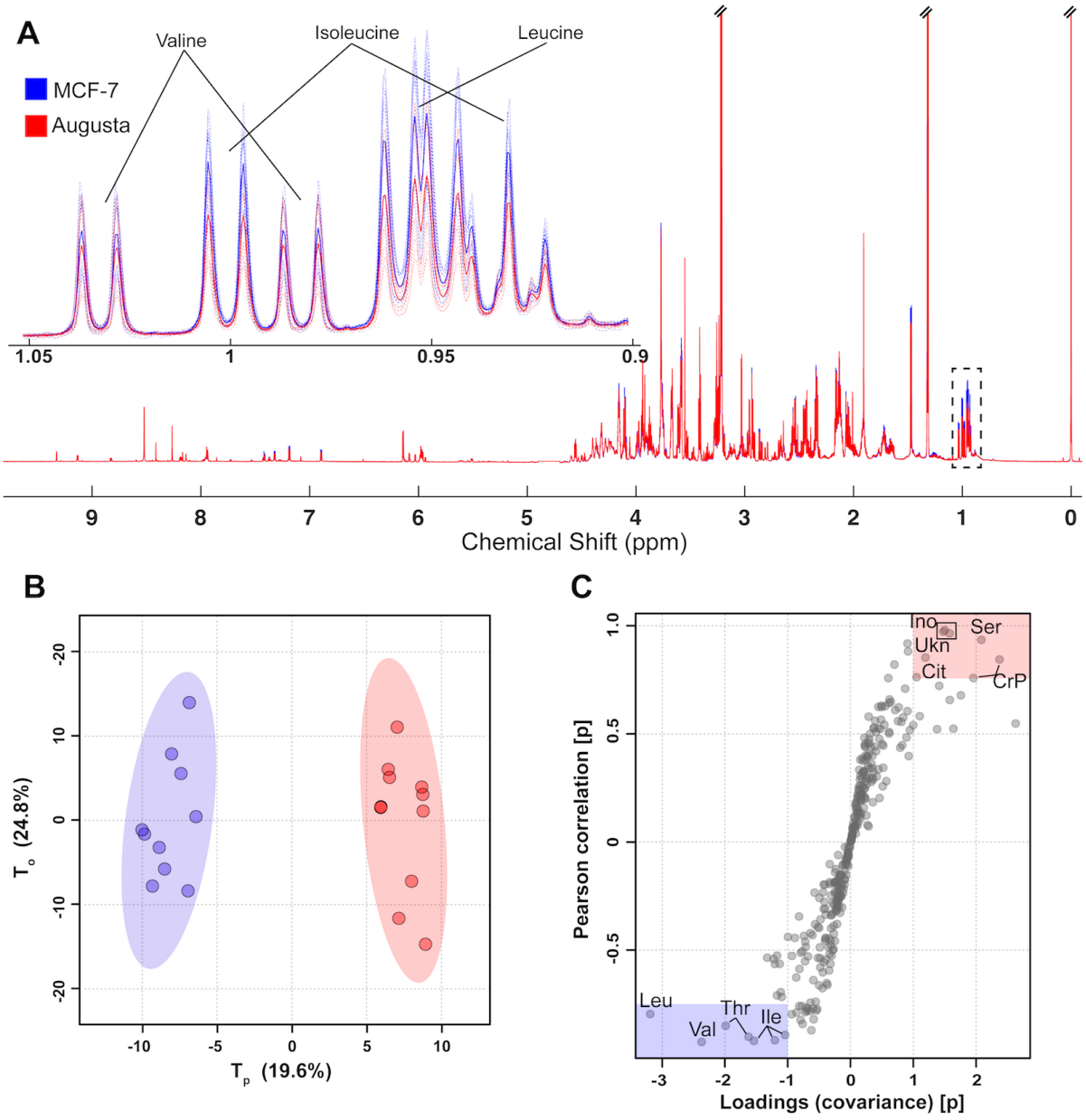
(10) Knitsch, R.; AlWahsh, M.; Raschke, H.; Lambert, J.; Hergenroder, R. *Anal Chem* **2021**, *93*, 13485-13494.

(11) Zang, Q.; Sun, C.; Chu, X.; Li, L.; Gan, W.; Zhao, Z.; Song, Y.; He, J.; Zhang, R.; Abliz, Z. *Anal Chim Acta* **2021**, *1155*, 338342.

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

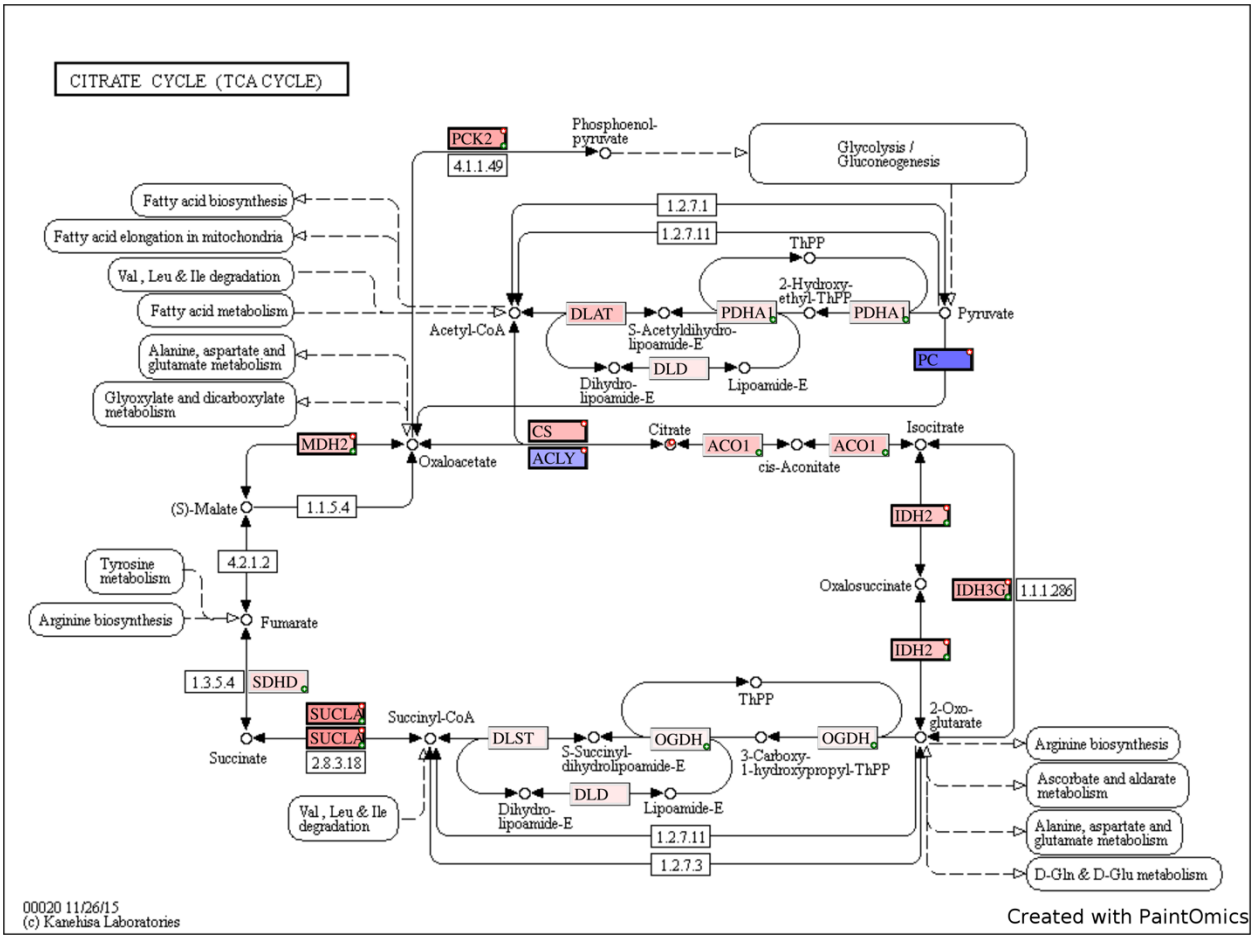


Supplemental Figure 2.6: Augusta cells show more invasive phenotype when cocultured with cancer associated fibroblasts. A and B) Phase contrast micrographs of MCF-7 and Augusta cells respectively, shown at 400X total magnification. C and D) Confocal fluorescence micrographs of MCF-7 and Augusta (respectively) cells co-cultured with cancer associated fibroblasts.

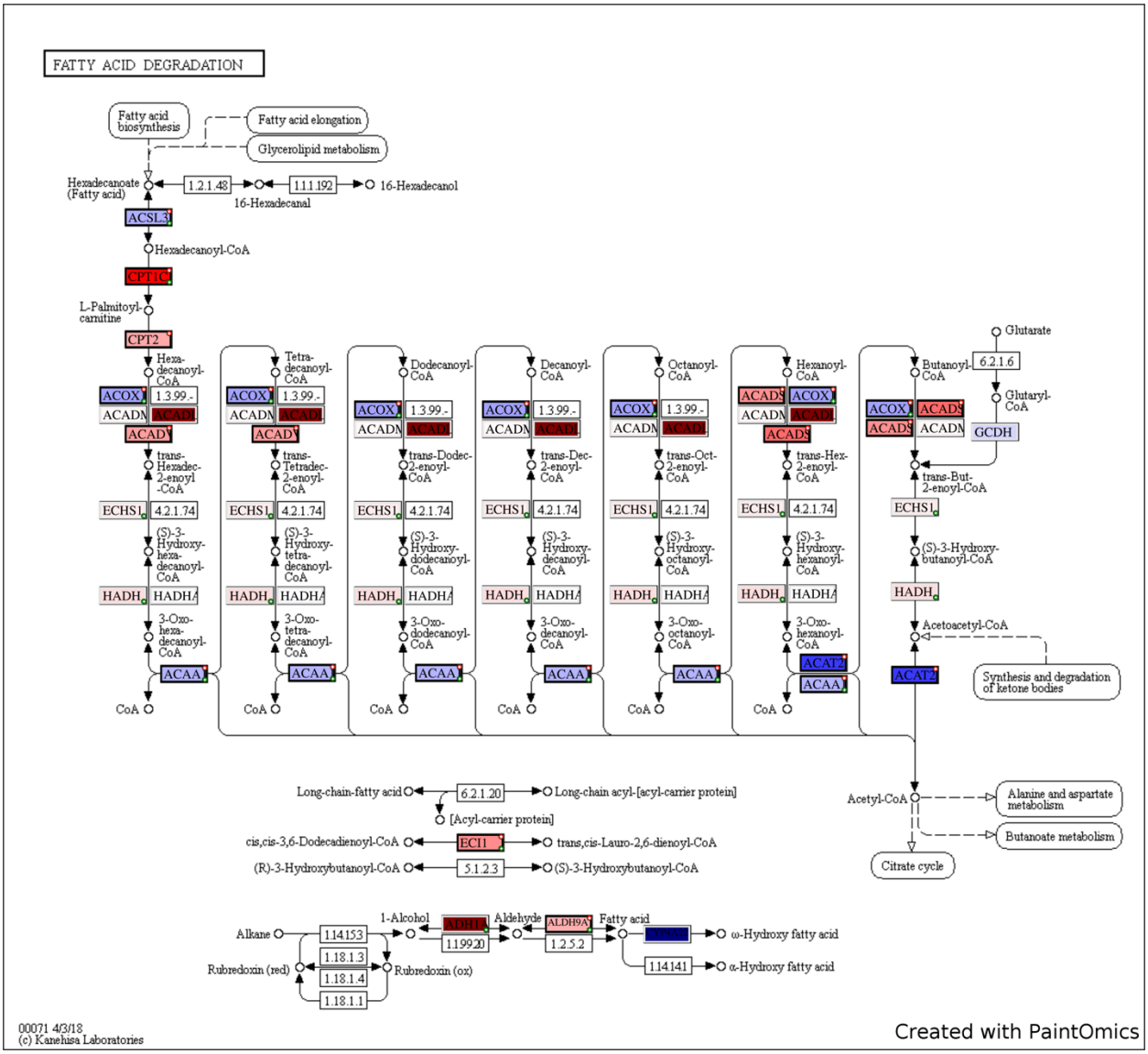


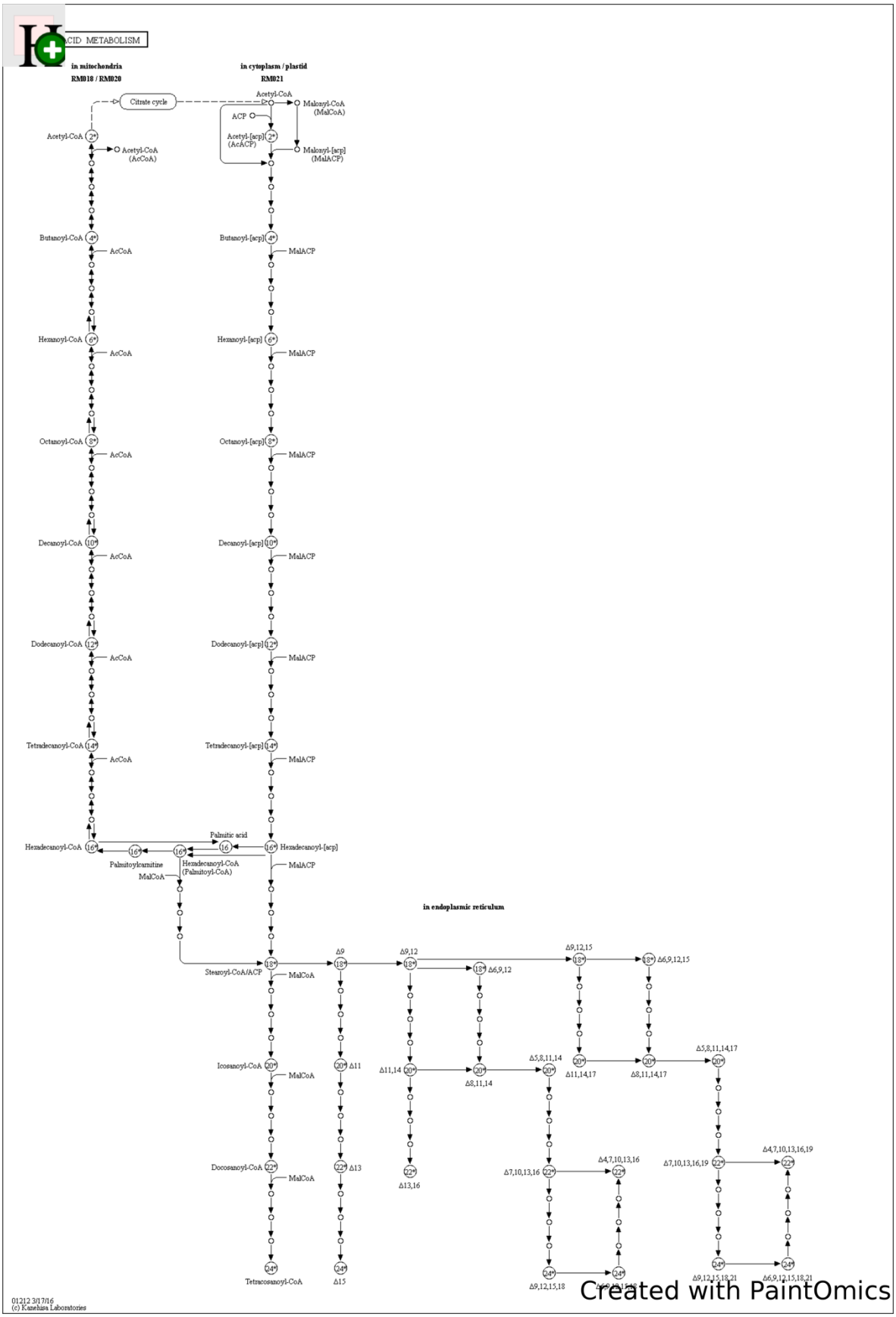
Supplemental Figure 2.2: Untargeted ^{1}H NMR metabolomics shows metabolic reprogramming largely inconsistent with EMT in Augusta cells. A) Overlay of average ^{1}H NMR spectra for each cell line. Inset shows zoom of region corresponding to signals from BCAAs. Dashed lines are spectra from individual samples across two independent experiments. B) Scores plot of OPLS-DA model generated using 340 manually selected features. Each dot represents an independent culture replicate. $R^2Y = .679$ $Q^2 = .544$. MCF-7 N=10, Augusta N=11. C) Variable importance

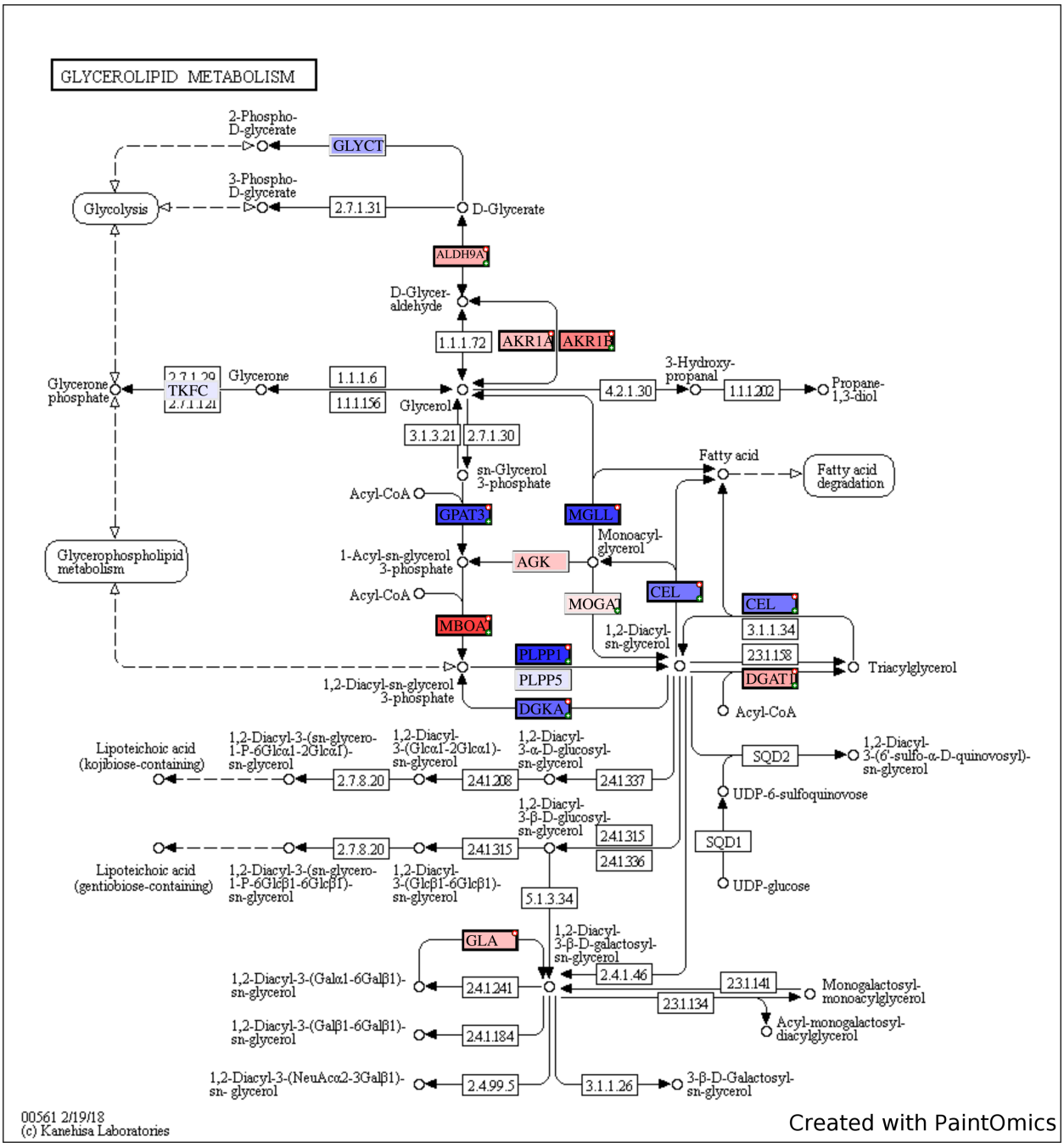
projection plot indicating contribution of each feature to classification in OPLS-DA model (loadings) and correlation of feature value with T_p component score. Highlighted areas indicate features with loadings > |1| and correlation > |.75|. Leu = leucine, Val = valine, Thr = threonine, Ile = isoleucine, Ino = myo-inositol, Ukn = unknown, Cit = citrate, Ser = serine, CrP = creatine phosphate

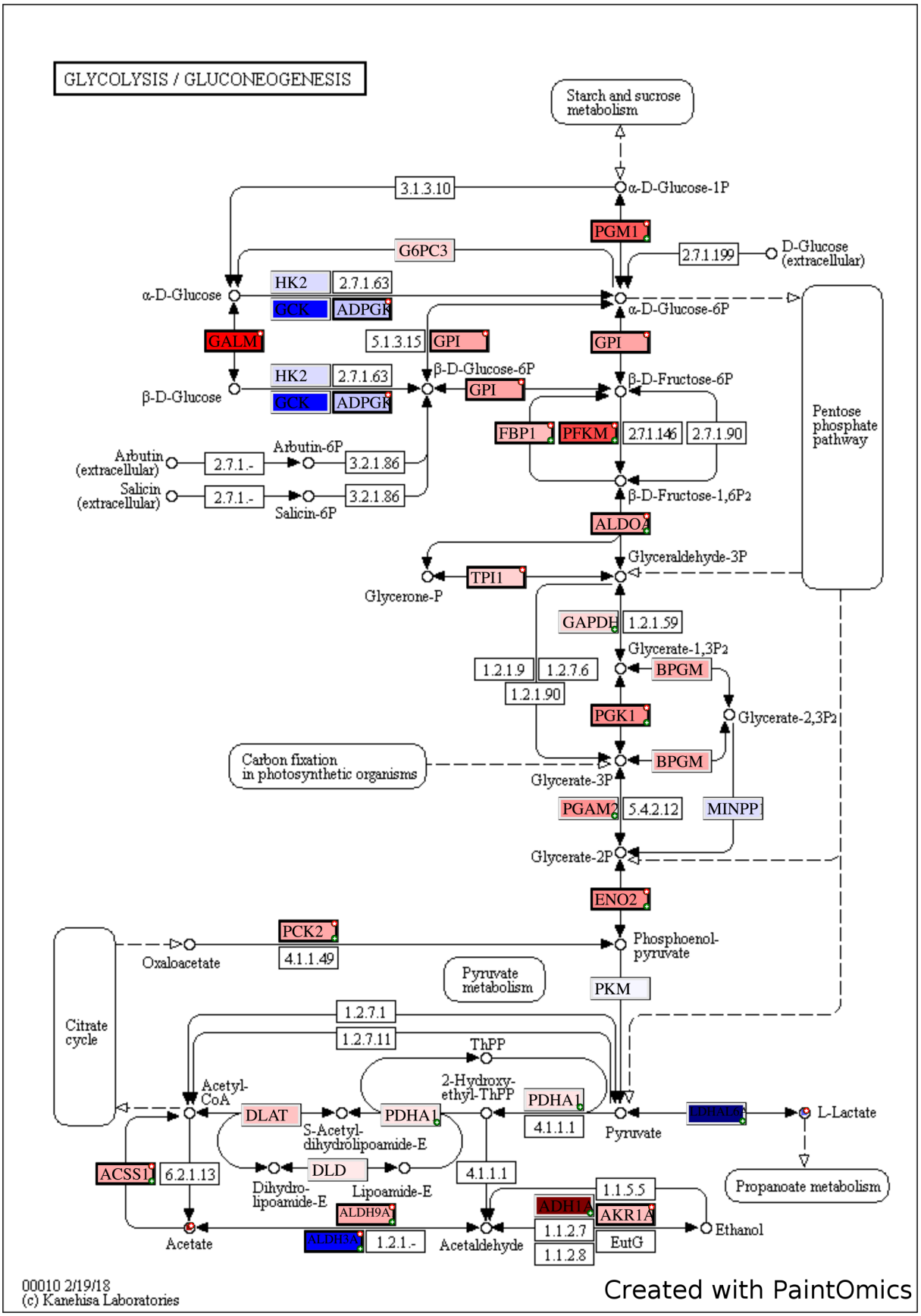


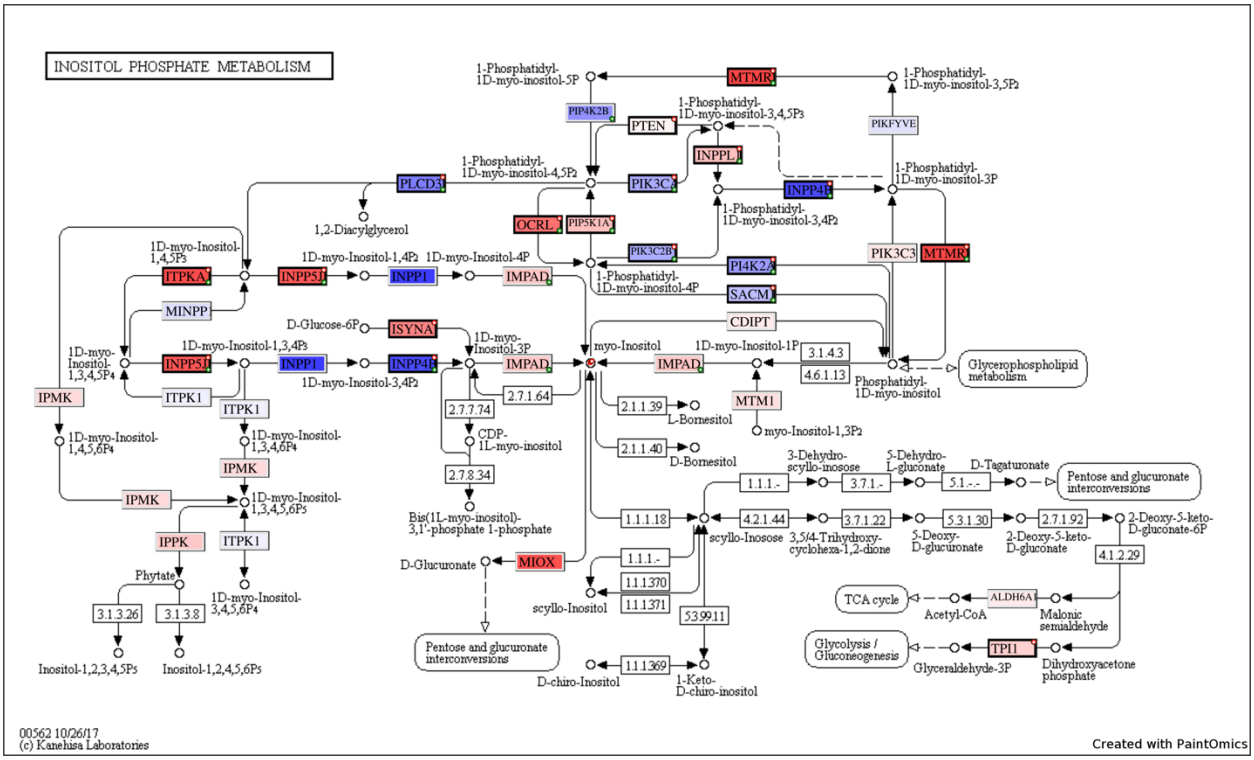
Supplemental Figures 2.3-15: KEGG maps of most highly enriched metabolic pathways with metabolomics and transcriptomics data represented.

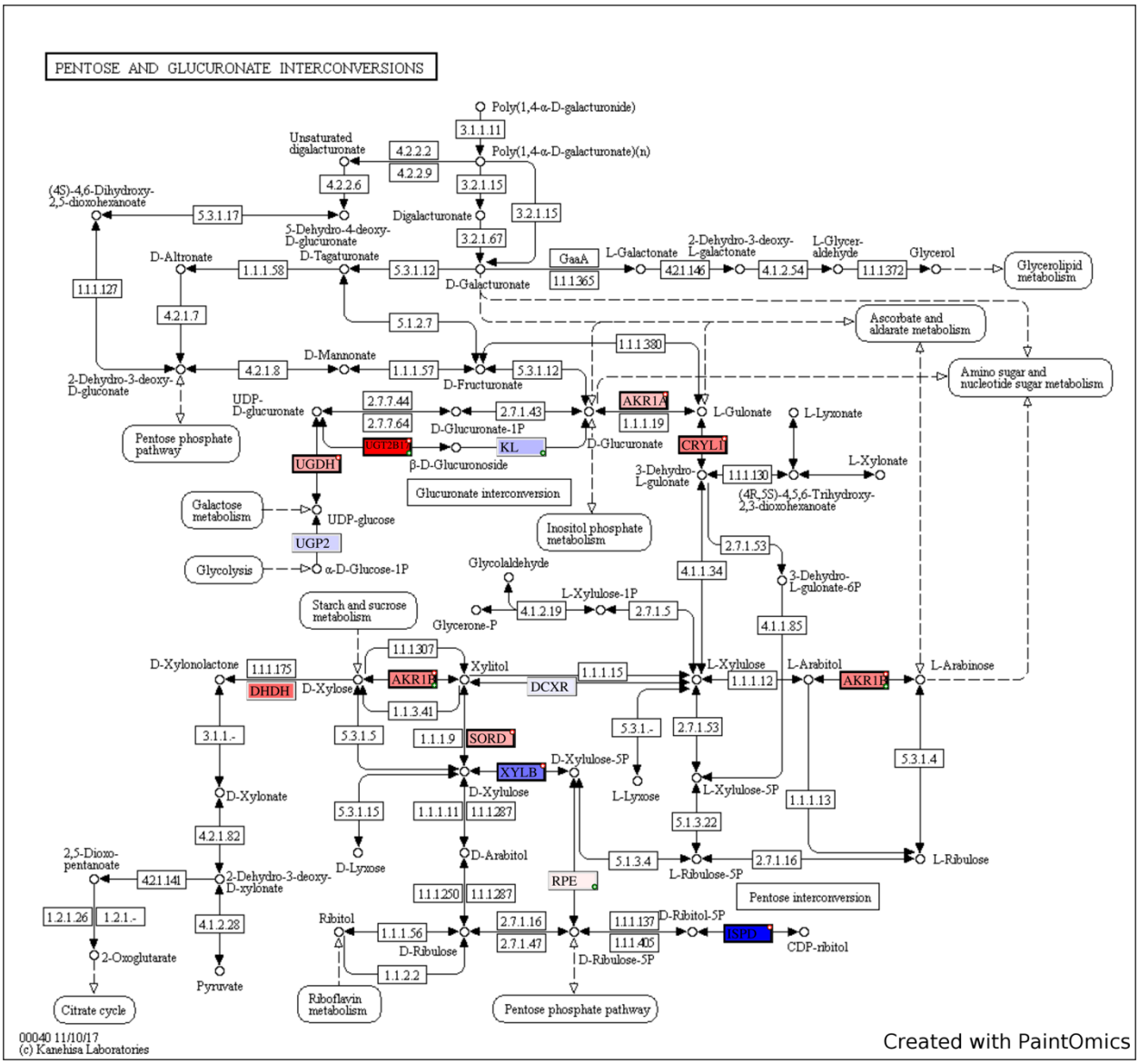

FATTY ACID DEGRADATION
Fatty acid biosynthesis
Fatty acid elongation
Glycerolipid metabolism
Hexadecanoate (Fatty acid)
1.2.1.48
16-Hexadecanal
1.1.1.192
16-Hexadecanol
ACSL3
Hexadecanoyl-CoA
CPT1C
L-Palmitoyl-carnitine
CPT2
Hexa-decanoyl-CoA
Tetra-decanoyl-CoA
Dodecanoyl-CoA
Decanoyl-CoA
Octanoyl-CoA
Hexanoyl-CoA
Butanoyl-CoA
Glutarate
6.2.1.6
Glutaryl-CoA
GCDH
ACOX
ACADN
ACADL
ACADM
ACADS
1.3.99.-
trans-Hexadec-2-enoyl-CoA
trans-Tetradec-2-enoyl-CoA
trans-Dodec-2-enoyl-CoA
trans-Dec-2-enoyl-CoA
trans-Oct-2-enoyl-CoA
trans-Hex-2-enoyl-CoA
trans-But-2-enoyl-CoA
ECHS1
4.2.1.74
(S)-3-Hydroxy-hexa-decanoyl-CoA
(S)-3-Hydroxy-tetra-decanoyl-CoA
(S)-3-Hydroxy-dodecanoyl-CoA
(S)-3-Hydroxy-decanoyl-CoA
(S)-3-Hydroxy-octanoyl-CoA
(S)-3-Hydroxy-hexanoyl-CoA
(S)-3-Hydroxy-butanoyl-CoA
HADH
3-Oxo-hexa-decanoyl-CoA
3-Oxo-tetra-decanoyl-CoA
3-Oxo-dodecanoyl-CoA
3-Oxo-decanoyl-CoA
3-Oxo-octanoyl-CoA
3-Oxo-hexanoyl-CoA
Acetoacetyl-CoA
ACAA
ACAT2
CoA
Synthesis and degradation of ketone bodies
Long-chain-fatty acid
6.2.1.20
Long-chain acyl-[acyl-carrier protein]
[Acyl-carrier protein]
cis,cis-3,6-Dodecadienoyl-CoA
ECI1
trans,cis-Lauro-2,6-dienoyl-CoA
(R)-3-Hydroxybutanoyl-CoA
5.1.2.3
(S)-3-Hydroxybutanoyl-CoA
Acetyl-CoA
Alanine and aspartate metabolism
Butanoate metabolism
Citrate cycle
Alkane
1.14.15.3
1-Alcohol
ADH1A
1.1.99.20
Aldehyde
ALDH9A1
1.2.5.2
Fatty acid
CYP4A11
1.14.14.1
ω-Hydroxy fatty acid
α-Hydroxy fatty acid
Rubredoxin (red)
1.18.1.3
1.18.1.4
1.18.1.1
Rubredoxin (ox)
00071 4/3/18
(c) Kanehisa Laboratories
Created with PaintOmics

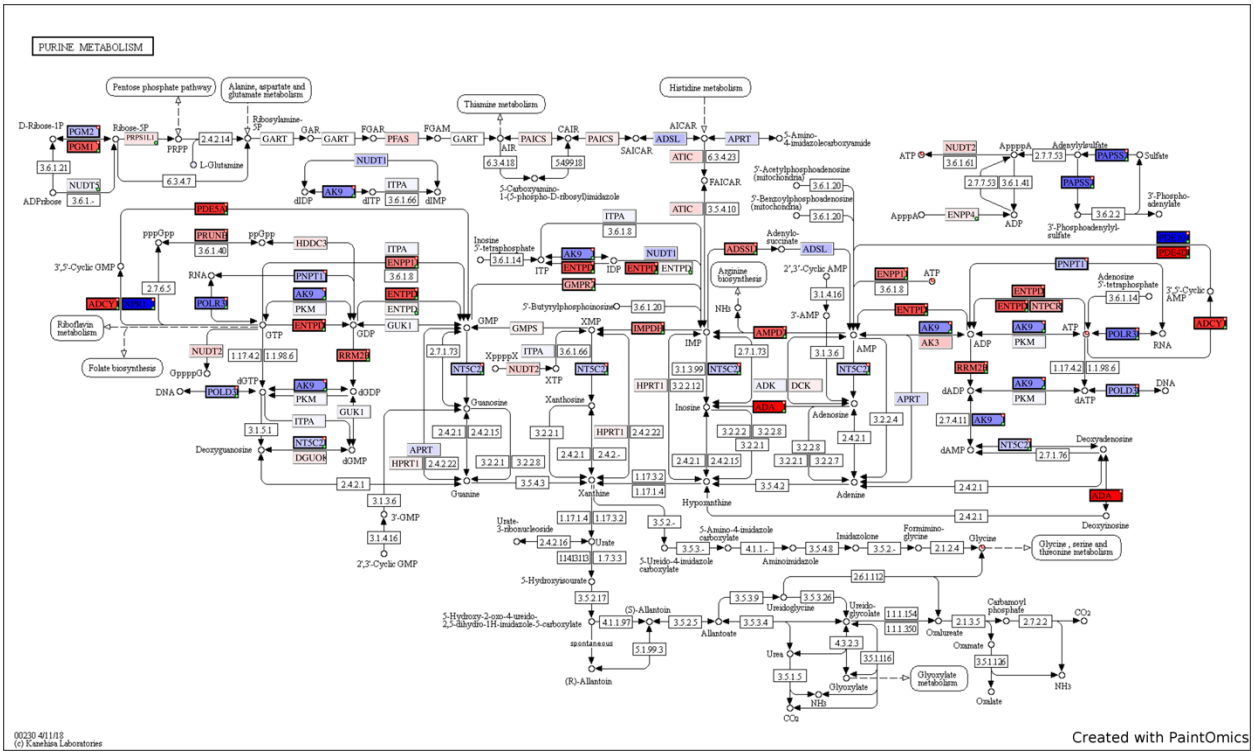

ACID METABOLISM
in mitochondria
RM018 / RM020
in cytoplasm / plastid
RM021
Citrate cycle
Acetyl-CoA
Malonyl-CoA (MalCoA)
ACP
Acetyl-[acp] (AcACP)
Malonyl-[acp] (MalACP)
Acetyl-CoA (AcCoA)
Butanoyl-CoA
Hexanoyl-CoA
Octanoyl-CoA
Decanoyl-CoA
Dodecanoyl-CoA
Tetradecanoyl-CoA
Hexadecanoyl-CoA
Palmitic acid
Palmitoylcarnitine
Hexadecanoyl-CoA (Palmitoyl-CoA)
Hexadecanoyl-[acp]
MalCoA
MalACP
AcCoA
in endoplasmic reticulum
Stearoyl-CoA/ACP
Icosanoyl-CoA
Docosanoyl-CoA
Tetracosanoyl-CoA
Δ9
Δ11
Δ13
Δ15
Δ9,12
Δ6,9,12
Δ11,14
Δ8,11,14
Δ13,16
Δ9,12,15
Δ6,9,12,15
Δ11,14,17
Δ8,11,14,17
Δ5,8,11,14
Δ7,10,13,16
Δ4,7,10,13,16
Δ5,8,11,14,17
Δ7,10,13,16,19
Δ4,7,10,13,16,19
Δ9,12,15,18
Δ9,12,15,18,21
Δ6,9,12,15,18,21
01212 3/17/16
(c) Kanehisa Laboratories
Created with PaintOmics

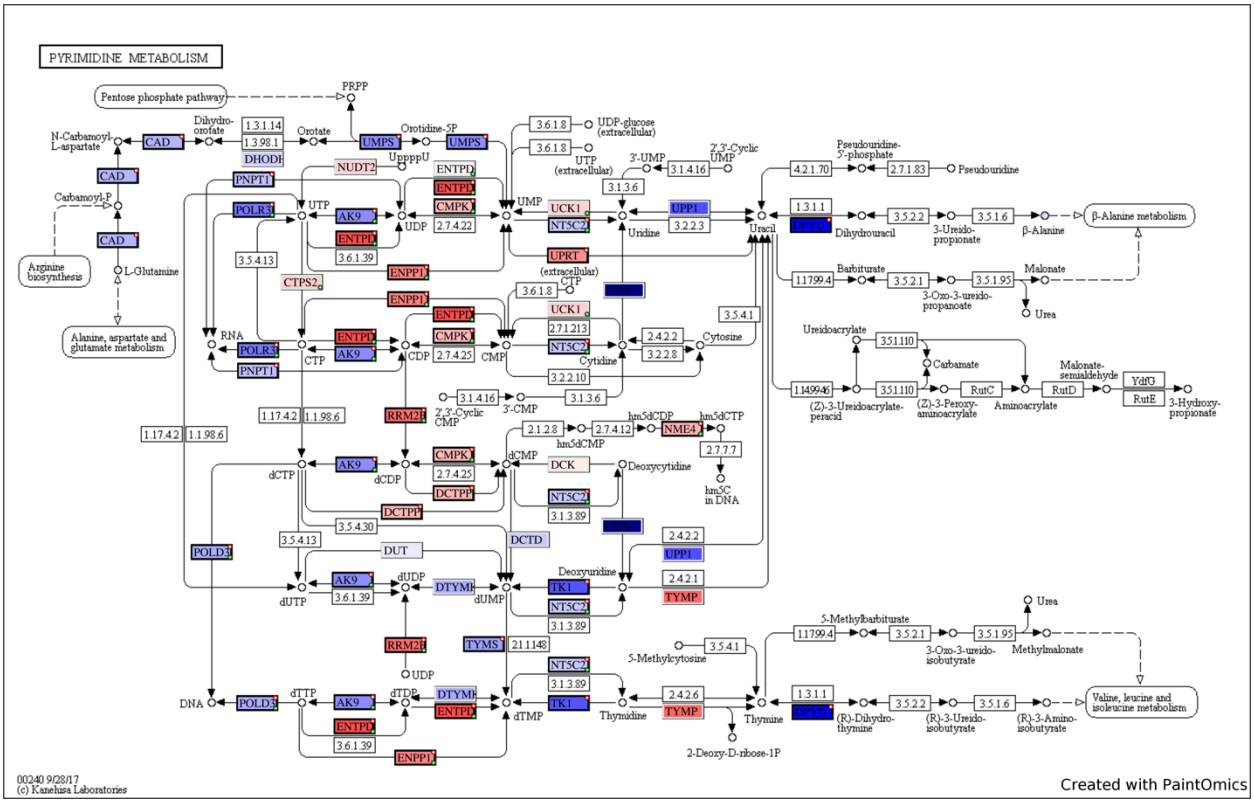

GLYCEROLIPID METABOLISM
2-Phospho-D-glycerate
GLYCT
3-Phospho-D-glycerate
Glycolysis
2.7.1.31
D-Glycerate
ALDH9A1
D-Glycer-aldehyde
1.1.1.72
AKR1A1
AKR1B1
Glycerone phosphate
TKFC
Glycerone
1.1.1.6
1.1.1.156
Glycerol
4.2.1.30
3-Hydroxy-propanal
1.1.1.202
Propane-1,3-diol
3.1.3.21
2.7.1.30
sn-Glycerol 3-phosphate
Fatty acid
Fatty acid degradation
Acyl-CoA
GPAT3
MGLL
Monoacyl-glycerol
Glycerophospholipid metabolism
1-Acyl-sn-glycerol 3-phosphate
AGK
MOGAT
CEL
CEL
Acyl-CoA
MBOA
1,2-Diacyl-sn-glycerol
3.1.1.34
23.1.158
PLPP1
PLPP5
DGAT1
Triacylglycerol
1,2-Diacyl-sn-glycerol 3-phosphate
DGKA
Acyl-CoA
Lipoteichoic acid (kojibiose-containing)
1,2-Diacyl-3-(sn-glycero-1-P-6Glcα1-2Glcα1)-sn-glycerol
1,2-Diacyl-3-(Glcα1-2Glcα1)-sn-glycerol
1,2-Diacyl-3-α-D-glucosyl-sn-glycerol
2.7.8.20
2.4.1.208
2.4.1.337
SQD2
1,2-Diacyl-3-(6'-sulfo-α-D-quinovosyl)-sn-glycerol
UDP-6-sulfoquinovose
1,2-Diacyl-3-β-D-glucosyl-sn-glycerol
2.4.1.315
2.4.1.336
SQD1
2.7.8.20
2.4.1.315
Lipoteichoic acid (gentiobiose-containing)
1,2-Diacyl-3-(sn-glycero-1-P-6Glcβ1-6Glcβ1)-sn-glycerol
1,2-Diacyl-3-(Glcβ1-6Glcβ1)-sn-glycerol
UDP-glucose
5.1.3.34
GLA
1,2-Diacyl-3-β-D-galactosyl-sn-glycerol
2.4.1.46
1,2-Diacyl-3-(Galα1-6Galβ1)-sn-glycerol
2.4.1.241
23.1.141
Monogalactosyl-monoacylglycerol
23.1.134
Acyl-monogalactosyl-diacylglycerol
1,2-Diacyl-3-(Galβ1-6Galβ1)-sn-glycerol
2.4.1.184
1,2-Diacyl-3-(NeuAcα2-3Galβ1)-sn- glycerol
2.4.99.5
3.1.1.26
3-β-D-Galactosyl-sn-glycerol
00561 2/19/18
(c) Kanehisa Laboratories
Created with PaintOmics

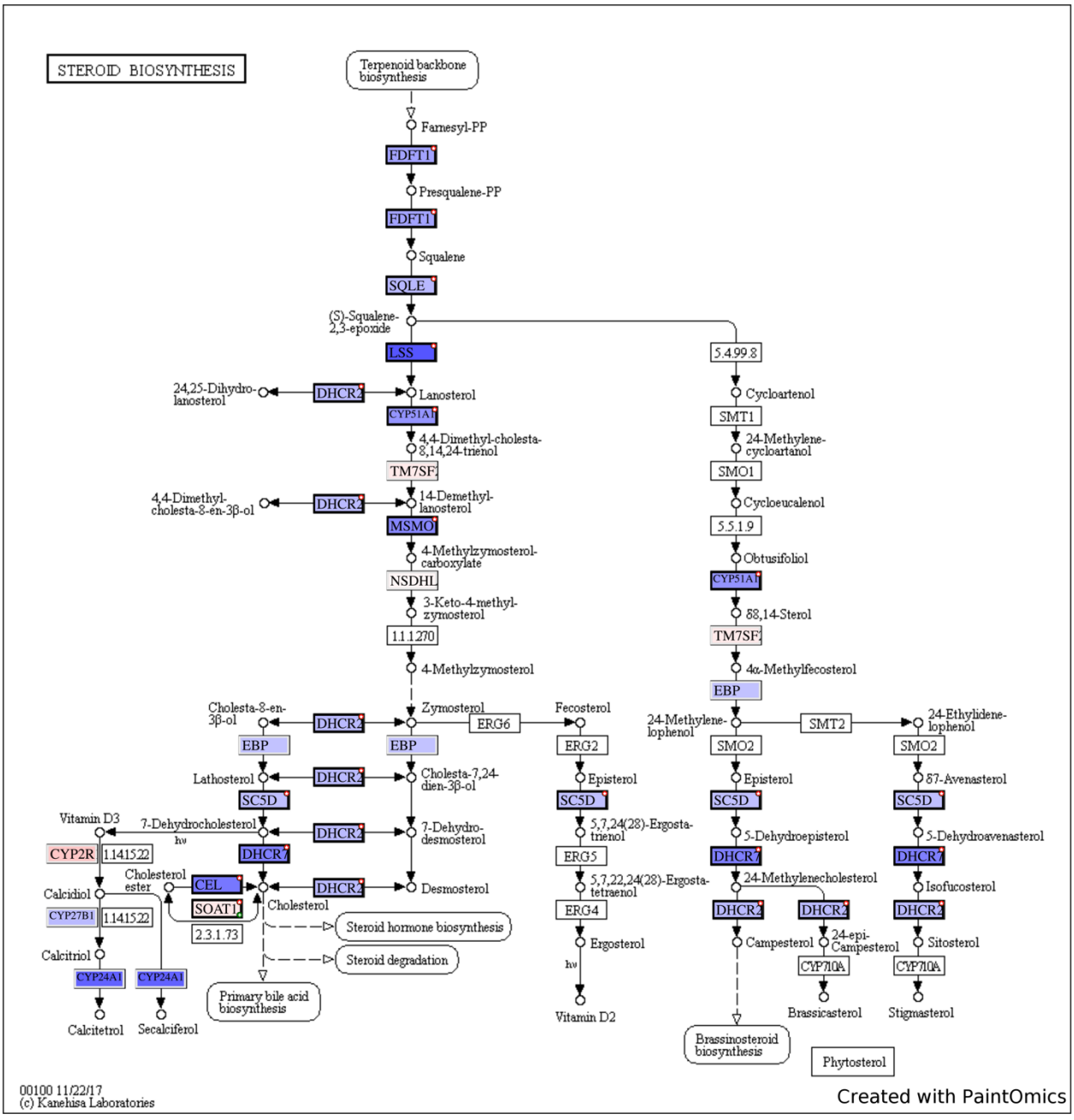

GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
α-D-Glucose-1P
3.1.3.10
PGM1
G6PC3
2.7.1.199
D-Glucose (extracellular)
α-D-Glucose
HK2
2.7.1.63
GCK
ADPGK
α-D-Glucose-6P
GALM
5.1.3.15
GPI
β-D-Glucose-6P
β-D-Glucose
β-D-Fructose-6P
Pentose phosphate pathway
FBP1
PFKM
2.7.1.146
2.7.1.90
Arbutin (extracellular)
2.7.1.-
Arbutin-6P
3.2.1.86
Salicin (extracellular)
Salicin-6P
β-D-Fructose-1,6P2
ALDOA
Glyceraldehyde-3P
TPI1
Glycerone-P
GAPDH
1.2.1.59
Glycerate-1,3P2
1.2.1.9
1.2.7.6
1.2.1.90
BPGM
PGK1
Glycerate-2,3P2
Carbon fixation in photosynthetic organisms
Glycerate-3P
PGAM2
5.4.2.12
MINPP1
Glycerate-2P
ENO2
PCK2
4.1.1.49
Oxaloacetate
Phosphoenol-pyruvate
Pyruvate metabolism
PKM
Citrate cycle
1.2.7.1
1.2.7.11
ThPP
Acetyl-CoA
2-Hydroxy-ethyl-ThPP
DLAT
PDHA1
PDHA1
4.1.1.1
Pyruvate
LDHAL6
L-Lactate
S-Acetyl-dihydrolipoamide-E
ACSS1
6.2.1.13
DLD
4.1.1.1
Propanoate metabolism
Dihydro-lipoamide-E
Lipoamide-E
1.1.5.5
ADH1A
AKR1A
ALDH9A
1.2.1.-
Acetaldehyde
1.1.2.7
1.1.2.8
EutG
Ethanol
Acetate
00010 2/19/18
(c) Kanehisa Laboratories
Created with PaintOmics

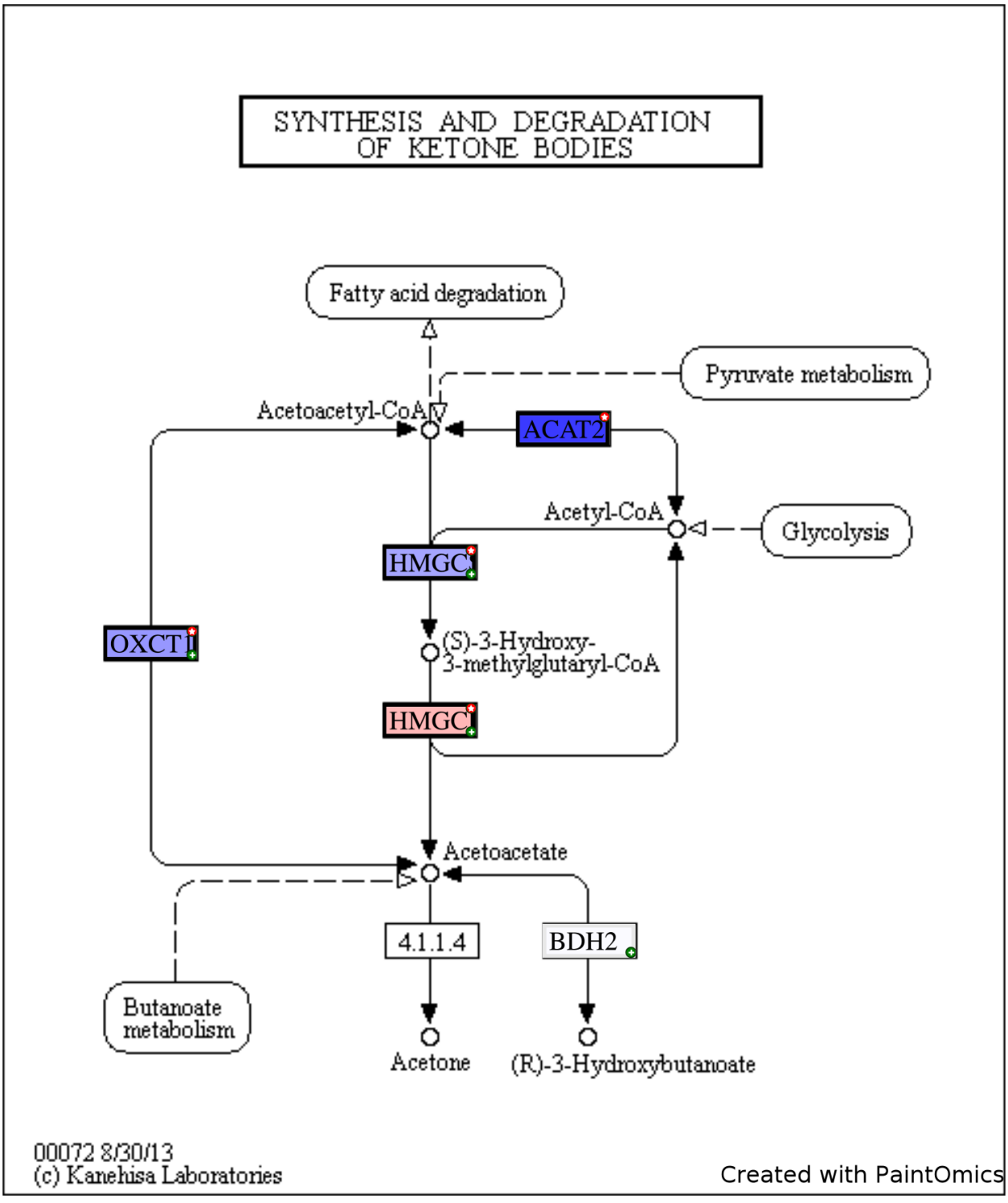

INOSITOL PHOSPHATE METABOLISM
1-Phosphatidyl-1D-myo-inositol-5P
MTMR
1-Phosphatidyl-1D-myo-inositol-3,5P2
PIP4K2B
1-Phosphatidyl-1D-myo-inositol-3,4,5P3
PTEN
PIKFYVE
INPPL1
1-Phosphatidyl-1D-myo-inositol-4,5P2
PLCD3
PIK3CA
INPP4A
1-Phosphatidyl-1D-myo-inositol-3P
1,2-Diacylglycerol
OCRL
PIP5K1A
1-Phosphatidyl-1D-myo-inositol-3,4P2
1D-myo-Inositol-1,4,5P3
1D-myo-Inositol-1,4P2
1D-myo-Inositol-4P
PIK3C2B
PIK3C3
MTMR
ITPKA
INPP5J
INPP1
IMPAD
PI4K2A
1-Phosphatidyl-1D-myo-inositol-4P
SACM
MINPP
CDIPT
D-Glucose-6P
ISYNA
1D-myo-Inositol-3P
myo-Inositol
1D-myo-Inositol-1P
3.1.4.3
4.6.1.13
Glycerophospholipid metabolism
Phosphatidyl-1D-myo-inositol
1D-myo-Inositol-1,3,4P3
1D-myo-Inositol-1,3,4,5P4
INPP5J
INPP1
INPP4B
IMPAD
IMPAD
IPMK
ITPK1
ITPK1
1D-myo-Inositol-3,4P2
2.7.7.74
2.7.1.64
2.1.1.39
L-Bornesitol
MTM1
myo-Inositol-1,3P2
1D-myo-Inositol-1,4,5,6P4
1D-myo-Inositol-1,3,4,6P4
CDP-1L-myo-inositol
2.1.1.40
D-Bornesitol
IPMK
2.7.8.34
1.1.1.-
3-Dehydro-scyllo-inosose
3.7.1.-
5-Dehydro-L-gluconate
5.1.-.-
D-Tagaturonate
Pentose and glucuronate interconversions
IPMK
1D-myo-Inositol-1,3,4,5,6P5
Bis(1L-myo-inositol)-3,1'-phosphate 1-phosphate
1.1.1.18
4.2.1.44
3.7.1.22
5.3.1.30
2.7.1.92
2-Deoxy-5-keto-D-gluconate-6P
scyllo-Inosose
3,5/4-Trihydroxy-cyclohexa-1,2-dione
5-Deoxy-D-glucuronate
2-Deoxy-5-keto-D-gluconate
4.1.2.29
IPPK
ITPK1
D-Glucuronate
MIOX
1.1.1.-
1.1.1.370
1.1.1.371
scyllo-Inosol
Phytate
1D-myo-Inositol-3,4,5,6P4
TCA cycle
ALDH6A1
Acetyl-CoA
Malonic semialdehyde
3.1.3.26
3.1.3.8
53.99.11
Pentose and glucuronate interconversions
Glycolysis / Gluconeogenesis
TPI1
Glyceraldehyde-3P
Dihydroxyacetone phosphate
Inositol-1,2,3,4,5P5
Inositol-1,2,4,5,6P5
D-chiro-Inositol
1.1.1.369
1-Keto-D-chiro-inositol
00562 10/26/17
(c) Kanehisa Laboratories
Created with PaintOmics

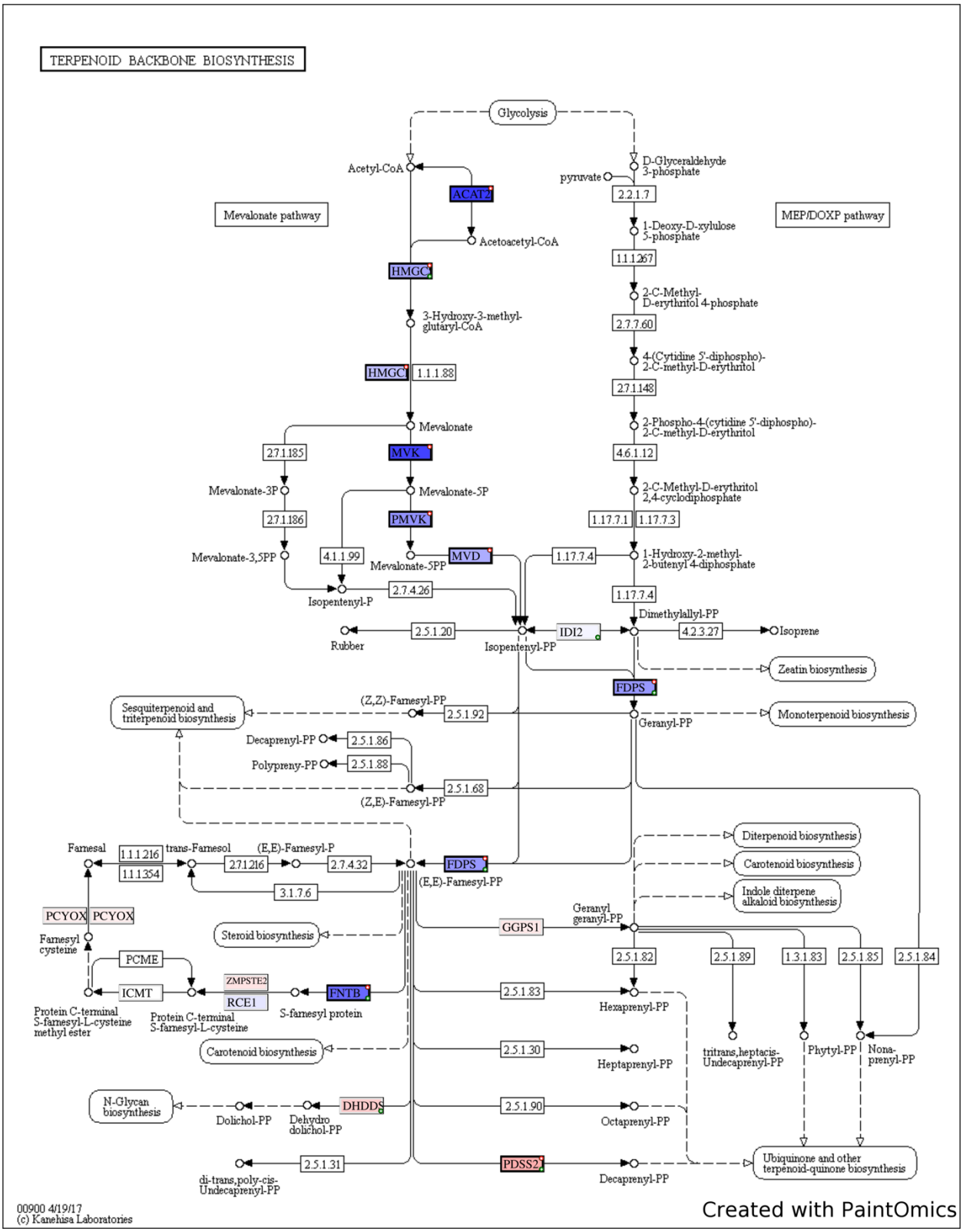

PENTOSE AND GLUCURONATE INTERCONVERSIONS
Poly(1,4-α-D-galacturonide)
3.1.1.11
Unsaturated digalacturonate
4.2.2.2
4.2.2.9
Poly(1,4-α-D-galacturonate)(n)
4.2.2.6
3.2.1.15
(4S)-4,6-Dihydroxy-2,3-dioxohexanoate
5.3.1.17
5-Dehydro-4-deoxy-D-glucuronate
Digalacturonate
3.2.1.15
3.2.1.67
D-Altronate
1.1.1.58
D-Tagaturonate
5.3.1.12
D-Galacturonate
GaaA
1.1.1.365
L-Galactonate
4.2.1.146
2-Dehydro-3-deoxy-L-galactonate
4.1.2.54
L-Glycer-aldehyde
1.1.1.372
Glycerol
Glycerolipid metabolism
1.1.1.127
4.2.1.7
5.1.2.7
Ascorbate and aldarate metabolism
1.1.1.380
D-Mannonate
4.2.1.8
1.1.1.57
5.3.1.12
D-Fructuronate
Amino sugar and nucleotide sugar metabolism
2-Dehydro-3-deoxy-D-gluconate
UDP-D-glucuronate
2.7.7.44
2.7.7.64
2.7.1.43
D-Glucuronate-1P
AKR1A
1.1.1.19
L-Gulonate
L-Lyxonate
Pentose phosphate pathway
UGT2B7
KL
β-D-Glucuronoside
D-Glucuronate
CRYL1
UGDH
Glucuronate interconversion
3-Dehydro-L-gulonate
1.1.1.130
L-Xylonate
(4R,5S)-4,5,6-Trihydroxy-2,3-dioxohexanoate
Galactose metabolism
UDP-glucose
Inositol phosphate metabolism
2.7.1.53
UGP2
Glycolysis
α-D-Glucose-1P
Glycolaldehyde
4.1.1.34
3-Dehydro-L-gulonate-6P
L-Xylulose-1P
Starch and sucrose metabolism
4.1.2.19
2.7.1.5
Glycerone-P
4.1.1.85
1.1.1.307
D-Xylonolactone
1.1.1.175
DHDH
D-Xylose
AKR1B
Xylitol
DCXR
1.1.1.15
L-Xylulose
1.1.1.12
L-Arabitol
AKR1B
L-Arabinose
1.1.3.41
3.1.1.-
5.3.1.5
1.1.1.9
SORD
5.3.1.-
2.7.1.53
XYLB
D-Xylulose-5P
5.3.1.4
D-Xylonate
D-Xylulose
L-Lyxose
L-Xylulose-5P
1.1.1.13
5.3.1.15
1.1.1.11
1.1.1.287
5.1.3.22
4.2.1.82
2,5-Dioxo-pentanoate
D-Lyxose
D-Arabitol
5.1.3.4
2.7.1.16
L-Ribulose
4.2.1.141
2-Dehydro-3-deoxy-D-xylonate
1.1.1.250
1.1.1.287
RPE
L-Ribulose-5P
Pentose interconversion
1.2.1.26
1.2.1.-
4.1.2.28
Ribitol
1.1.1.56
2.7.1.16
2.7.1.47
1.1.1.137
1.1.1.405
D-Ribitol-5P
ISPD
CDP-ribitol
D-Ribulose
D-Ribulose-5P
2-Oxoglutarate
1.1.2.2
Pyruvate
Citrate cycle
Riboflavin metabolism
Pentose phosphate pathway
00040 11/10/17
(c) Kanehisa Laboratories
Created with PaintOmics

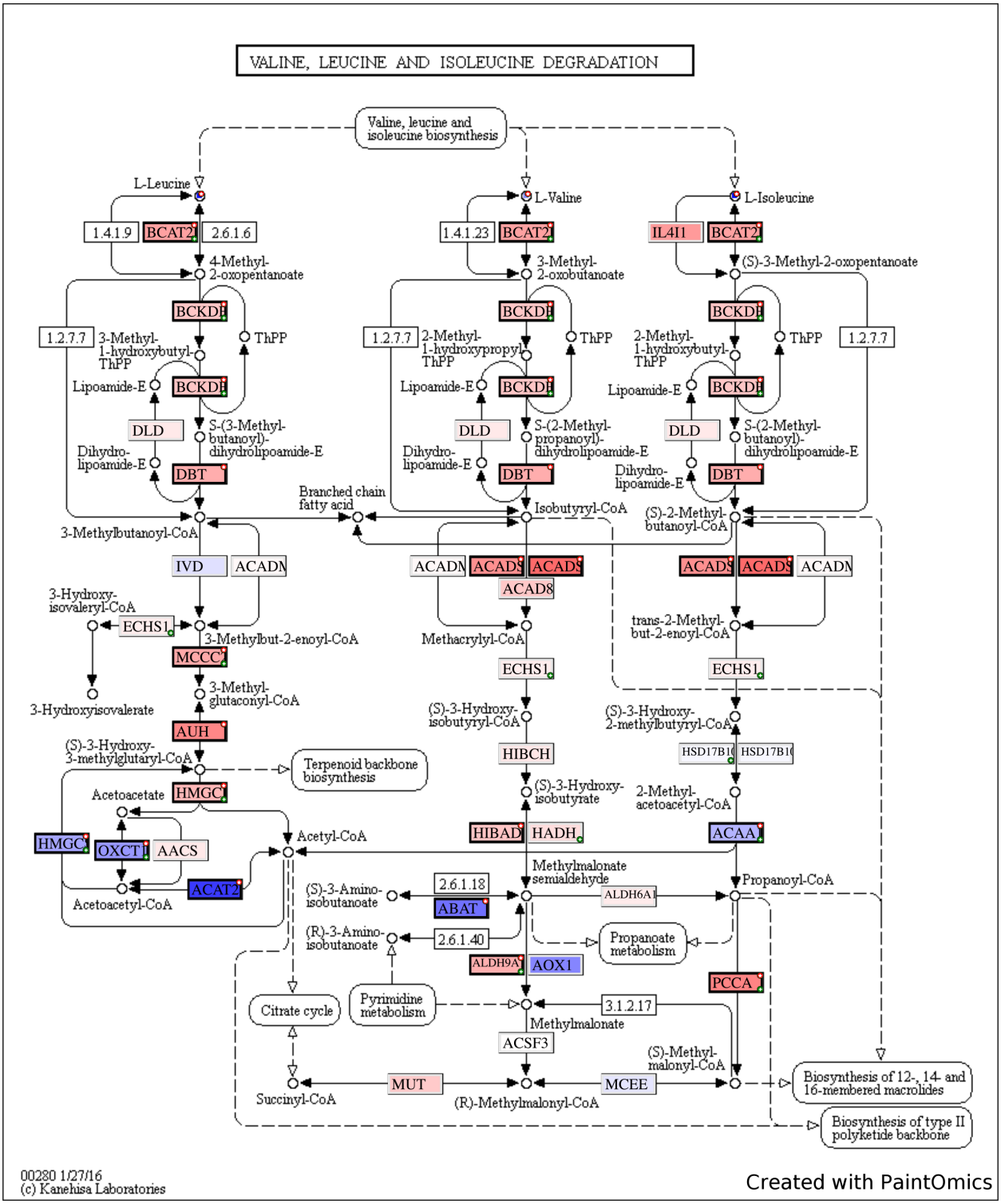

PURINE METABOLISM
Pentose phosphate pathway
Alanine, aspartate and glutamate metabolism
Thiamine metabolism
Histidine metabolism
Arginine biosynthesis
Riboflavin metabolism
Folate biosynthesis
Glycine, serine and threonine metabolism
Glyoxylate metabolism
β-Alanine metabolism
Valine, leucine and isoleucine metabolism
00230 4/11/13
(c) Kanehisa Laboratories
Created with PaintOmics

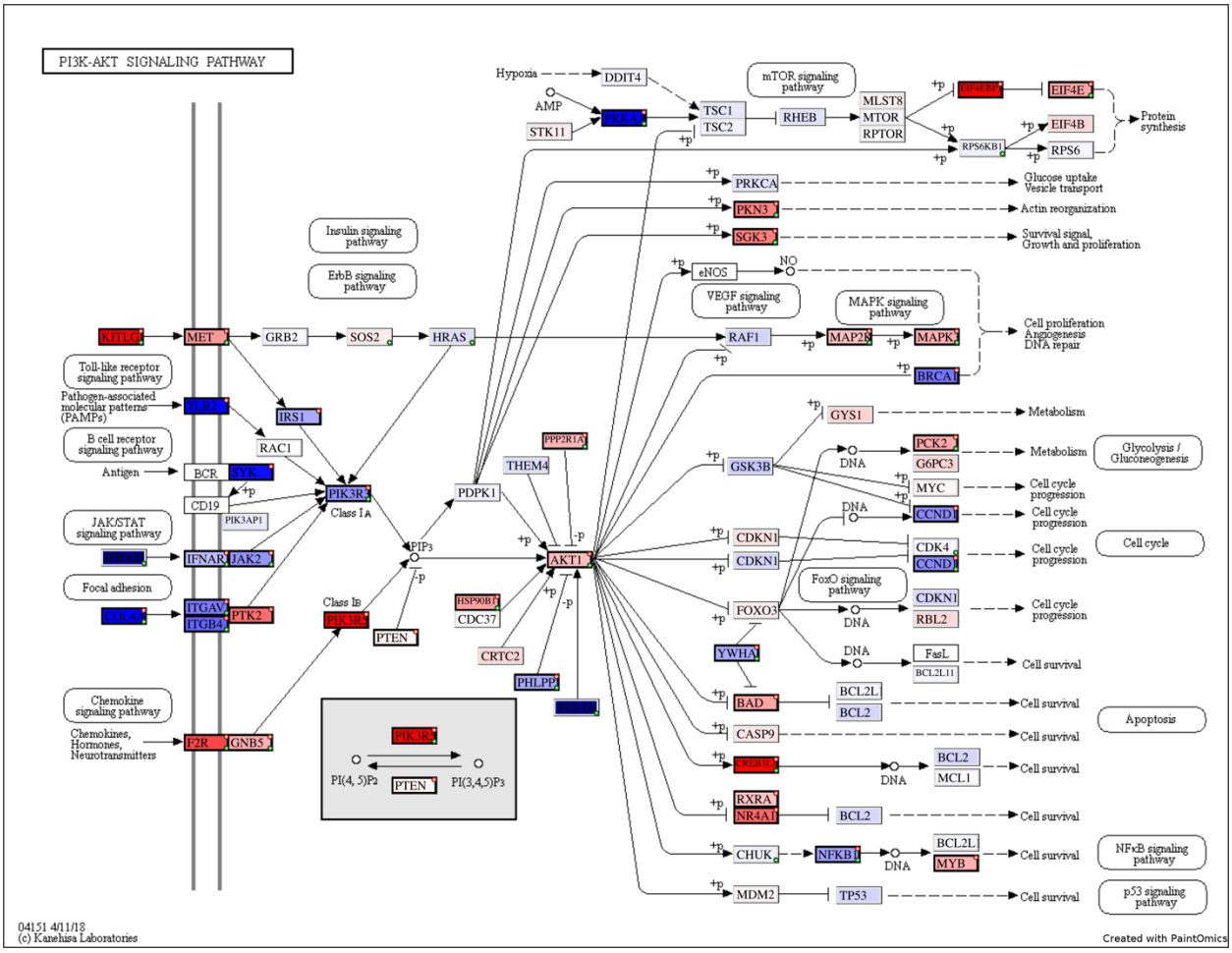

PYRIMIDINE METABOLISM
Pentose phosphate pathway
Arginine biosynthesis
Alanine, aspartate and glutamate metabolism
β-Alanine metabolism
Valine, leucine and isoleucine metabolism
00240 9/28/17
(c) Kanehisa Laboratories
Created with PaintOmics


STEROID BIOSYNTHESIS
Terpenoid backbone biosynthesis
Farnesyl-PP
FDFT1
Presqualene-PP
FDFT1
Squalene
SQLE
(S)-Squalene-2,3-epoxide
LSS
5.4.99.8
24,25-Dihydro-lanosterol
DHCR24
Lanosterol
CYP51A1
Cycloartenol
SMT1
4,4-Dimethyl-cholesta-8,14,24-trienol
TM7SF2
24-Methylene-cycloartanol
SMO1
4,4-Dimethyl-cholesta-8-en-3β-ol
DHCR24
14-Demethyl-lanosterol
MSMO1
Cycloeucalenol
5.5.1.9
4-Methylzymosterol-carboxylate
NSDHL
Obtusifoliol
CYP51A1
3-Keto-4-methyl-zymosterol
1.1.1.270
88,14-Sterol
TM7SF2
4-Methylzymosterol
4α-Methylfecosterol
EBP
Cholesta-8-en-3β-ol
DHCR24
Zymosterol
ERG6
Fecosterol
24-Methylene-lophenol
SMT2
24-Ethylidene-lophenol
EBP
EBP
ERG2
SMO2
SMO2
Lathosterol
DHCR24
Cholesta-7,24-dien-3β-ol
Episterol
Episterol
87-Avenasterol
SC5D
SC5D
SC5D
SC5D
Vitamin D3
7-Dehydrocholesterol
hν
DHCR24
7-Dehydro-desmosterol
5,7,24(28)-Ergosta-trienol
5-Dehydroepisterol
5-Dehydroavenasterol
CYP2R1
1.14.15.22
DHCR7
ERG5
DHCR7
DHCR7
Cholesterol ester
CEL
DHCR24
Desmosterol
5,7,22,24(28)-Ergosta-tetraenol
24-Methylenecholesterol
Isofucosterol
Calcidiol
SOAT1
Cholesterol
ERG4
DHCR24
DHCR24
DHCR24
CYP27B1
1.14.15.22
2.3.1.73
Steroid hormone biosynthesis
Ergosterol
Campesterol
24-epi-Campesterol
Sitosterol
Calcitriol
Steroid degradation
hν
CYP710A1
CYP710A1
CYP24A1
CYP24A1
Primary bile acid biosynthesis
Vitamin D2
Brassicasterol
Stigmasterol
Calcitetrol
Secalciferol
Brassinosteroid biosynthesis
Phytosterol
00100 11/22/17
(c) Kanehisa Laboratories
Created with PaintOmics


SYNTHESIS AND DEGRADATION OF KETONE BODIES
Fatty acid degradation
Pyruvate metabolism
Acetoacetyl-CoA
ACAT2
Acetyl-CoA
Glycolysis
HMGCS
OXCT1
(S)-3-Hydroxy-3-methylglutaryl-CoA
HMGCL
Acetoacetate
4.1.1.4
BDH2
Butanoate metabolism
Acetone
(R)-3-Hydroxybutanoate
00072 8/30/13
(c) Kanehisa Laboratories
Created with PaintOmics


TERPENOID BACKBONE BIOSYNTHESIS
Glycolysis
Acetyl-CoA
ACAT2
Acetoacetyl-CoA
Mevalonate pathway
HMGCS
3-Hydroxy-3-methyl-glutaryl-CoA
HMGCL
1.1.1.88
Mevalonate
2.7.1.185
MVK
Mevalonate-3P
Mevalonate-5P
2.7.1.186
PMVK
Mevalonate-3,5PP
4.1.1.99
Mevalonate-5PP
MVD
2.7.4.26
Isopentenyl-P
2.5.1.20
Rubber
Isopentenyl-PP
IDI2
pyruvate
D-Glyceraldehyde 3-phosphate
2.2.1.7
MEP/DOXP pathway
1-Deoxy-D-xylulose 5-phosphate
1.1.1.267
2-C-Methyl-D-erythritol 4-phosphate
2.7.7.60
4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol
2.7.1.148
2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol
4.6.1.12
2-C-Methyl-D-erythritol 2,4-cyclodiphosphate
1.17.7.1
1.17.7.3
1-Hydroxy-2-methyl-2-butenyl 4-diphosphate
1.17.7.4
1.17.7.4
Dimethylallyl-PP
4.2.3.27
Isoprene
Zeatin biosynthesis
FDPS
Geranyl-PP
Monoterpenoid biosynthesis
Sesquiterpenoid and triterpenoid biosynthesis
(Z,Z)-Farnesyl-PP
2.5.1.92
Decaprenyl-PP
2.5.1.86
Polypreny-PP
2.5.1.88
2.5.1.68
(Z,E)-Farnesyl-PP
Farnesal
1.1.1.216
1.1.1.354
trans-Farnesol
(E,E)-Farnesyl-P
2.7.1.216
2.7.4.32
FDPS
(E,E)-Farnesyl-PP
3.1.7.6
Diterpenoid biosynthesis
Carotenoid biosynthesis
Indole diterpene alkaloid biosynthesis
PCYOX
PCYOX
Farnesyl cysteine
Steroid biosynthesis
GGPS1
Geranyl geranyl-PP
PCME
ICMT
ZMPSTE2
RCE1
FNTB
Protein C-terminal S-farnesyl-L-cysteine methyl ester
Protein C-terminal S-farnesyl-L-cysteine
S-farnesyl protein
2.5.1.82
2.5.1.89
1.3.1.83
2.5.1.85
2.5.1.84
2.5.1.83
Hexaprenyl-PP
tritrans,heptacis-Undecaprenyl-PP
Phytyl-PP
Nona-prenyl-PP
Carotenoid biosynthesis
2.5.1.30
Heptaprenyl-PP
N-Glycan biosynthesis
Dolichol-PP
Dehydro dolichol-PP
DHDDS
2.5.1.90
Octaprenyl-PP
di-trans,poly-cis-Undecaprenyl-PP
2.5.1.31
PDSS2
Decaprenyl-PP
Ubiquinone and other terpenoid-quinone biosynthesis
00900 4/19/17
(c) Kanehisa Laboratories
Created with PaintOmics


VALINE, LEUCINE AND ISOLEUCINE DEGRADATION
Valine, leucine and isoleucine biosynthesis
L-Leucine
L-Valine
L-Isoleucine
1.4.1.9
BCAT2
2.6.1.6
1.4.1.23
IL4I1
4-Methyl-2-oxopentanoate
3-Methyl-2-oxobutanoate
(S)-3-Methyl-2-oxopentanoate
BCKDH
1.2.7.7
3-Methyl-1-hydroxybutyl-ThPP
2-Methyl-1-hydroxypropyl-ThPP
2-Methyl-1-hydroxybutyl-ThPP
ThPP
Lipoamide-E
DLD
S-(3-Methyl-butanoyl)-dihydrolipoamide-E
S-(2-Methyl-propanoyl)-dihydrolipoamide-E
S-(2-Methyl-butanoyl)-dihydrolipoamide-E
Dihydro-lipoamide-E
DBT
Branched chain fatty acid
Isobutyryl-CoA
(S)-2-Methyl-butanoyl-CoA
3-Methylbutanoyl-CoA
IVD
ACADM
ACADS
ACAD8
3-Hydroxy-isovaleryl-CoA
ECHS1
3-Methylbut-2-enoyl-CoA
Methacrylyl-CoA
trans-2-Methyl-but-2-enoyl-CoA
MCCC
3-Methyl-glutaconyl-CoA
3-Hydroxyisovalerate
(S)-3-Hydroxy-isobutyryl-CoA
(S)-3-Hydroxy-2-methylbutyryl-CoA
AUH
(S)-3-Hydroxy-3-methylglutaryl-CoA
Terpenoid backbone biosynthesis
HIBCH
HSD17B10
(S)-3-Hydroxy-isobutyrate
2-Methyl-acetoacetyl-CoA
Acetoacetate
HMGCS
HIBADH
HADH
ACAA
Acetyl-CoA
HMGCL
OXCT1
AACS
Methylmalonate semialdehyde
Propanoyl-CoA
ACAT2
Acetoacetyl-CoA
(S)-3-Amino-isobutanoate
2.6.1.18
ABAT
ALDH6A1
(R)-3-Amino-isobutanoate
2.6.1.40
Propanoate metabolism
ALDH9A1
AOX1
PCCA
Citrate cycle
Pyrimidine metabolism
3.1.2.17
Methylmalonate
ACSF3
(S)-Methyl-malonyl-CoA
Succinyl-CoA
MUT
MCEE
(R)-Methylmalonyl-CoA
Biosynthesis of 12-, 14- and 16-membered macrolides
Biosynthesis of type II polyketide backbone
00280 1/27/16
(c) Kanehisa Laboratories
Created with PaintOmics



Supplemental Figure 2.16: KEGG Map of PI3K/AKT signalling pathway with gene expression data represented. Note: Not all quantified and/or significant genes within pathway are shown in image.

Supplemental Table 2.1: All quantified intracellular metabolites measured by ^{1}H NMR. Values are in arbitrary units. Bold line indicates FDR p-value threshold of 0.05.

Metabolite	Confidence Score	Mean ± SD		log2(FC)	*p*-value	BH FDR *p*-value
		MCF-7	Augusta			
myo-Inositol	4	0.192 ± 0.01	0.267 ± 0.01	0.476	4.68E-13	1.54E-11
Isoleucine	4	0.626 ± 0.04	0.461 ± 0.04	-0.442	4.14E-09	6.60E-08
Serine	4	0.451 ± 0.03	0.604 ± 0.04	0.421	6.00E-09	6.60E-08
Leucine	4	1.628 ± 0.11	1.208 ± 0.11	-0.431	3.61E-08	2.98E-07
Threonine	4	0.499 ± 0.02	0.422 ± 0.03	-0.242	5.07E-07	3.34E-06
Creatine phosphate	4	0.317 ± 0.08	0.519 ± 0.06	0.712	3.71E-06	2.04E-05
Lactate	4	2.442 ± 0.16	1.981 ± 0.18	-0.302	6.17E-06	2.91E-05
Putrescine	4	0.082 ± 0.01	0.099 ± 0.01	0.268	1.45E-05	5.99E-05
Glycerophosphocholine	3	0.938 ± 0.05	1.085 ± 0.09	0.211	2.08E-04	6.89E-04
Valine	4	0.394 ± 0.02	0.342 ± 0.03	-0.204	2.09E-04	6.89E-04
Glycine	3	0.869 ± 0.04	0.938 ± 0.04	0.109	5.69E-04	1.71E-03
Creatine	4	0.362 ± 0.09	0.499 ± 0.07	0.465	7.93E-04	2.18E-03
AXP	2	0.166 ± 0.01	0.185 ± 0.01	0.154	1.79E-03	4.53E-03
Citrate	4	0.242 ± 0.04	0.289 ± 0.03	0.256	2.50E-03	5.88E-03
Phenylalanine	3	0.066 ± 0.01	0.051 ± 0.01	-0.380	3.66E-03	7.79E-03
Tyrosine	3	0.201 ± 0.03	0.160 ± 0.03	-0.334	3.78E-03	7.79E-03
Homoarginine	2	0.201 ± 0.04	0.259 ± 0.05	0.363	7.95E-03	1.54E-02
Acetate	3	0.743 ± 0.33	1.141 ± 0.30	0.619	8.90E-03	1.63E-02
Proline	4	0.049 ± 0.01	0.056 ± 0.01	0.174	5.10E-02	8.86E-02
Aspartate	4	0.147 ± 0.03	0.162 ± 0.02	0.143	0.167	0.262
Lysine	4	0.153 ± 0.01	0.159 ± 0.01	0.051	0.163	0.262
O-Phosphocholine	4	2.174 ± 0.26	2.005 ± 0.34	-0.117	0.219	0.328
Asparagine	4	0.289 ± 0.03	0.272 ± 0.04	-0.088	0.267	0.383
Alanine	4	1.591 ± 0.12	1.511 ± 0.23	-0.074	0.343	0.472
NAD+	3	0.087 ± 0.01	0.093 ± 0.01	0.087	0.373	0.492
Methionine	3	0.410 ± 0.03	0.400 ± 0.03	-0.037	0.460	0.584
Arginine	4	0.193 ± 0.01	0.188 ± 0.02	-0.037	0.554	0.631
Glutamine	4	0.285 ± 0.02	0.276 ± 0.04	-0.046	0.542	0.631
Hydroxyproline	2	0.150 ± 0.02	0.155 ± 0.02	0.042	0.543	0.631
Taurine	4	0.752 ± 0.07	0.763 ± 0.07	0.020	0.727	0.800
Glutamate	4	0.077 ± 0.01	0.079 ± 0.01	0.021	0.772	0.821
Glutathione	4	0.366 ± 0.04	0.370 ± 0.03	0.016	0.801	0.826
UDP-GlcNAc	3	0.036 ± 0.01	0.037 ± 0.01	0.046	0.855	0.855

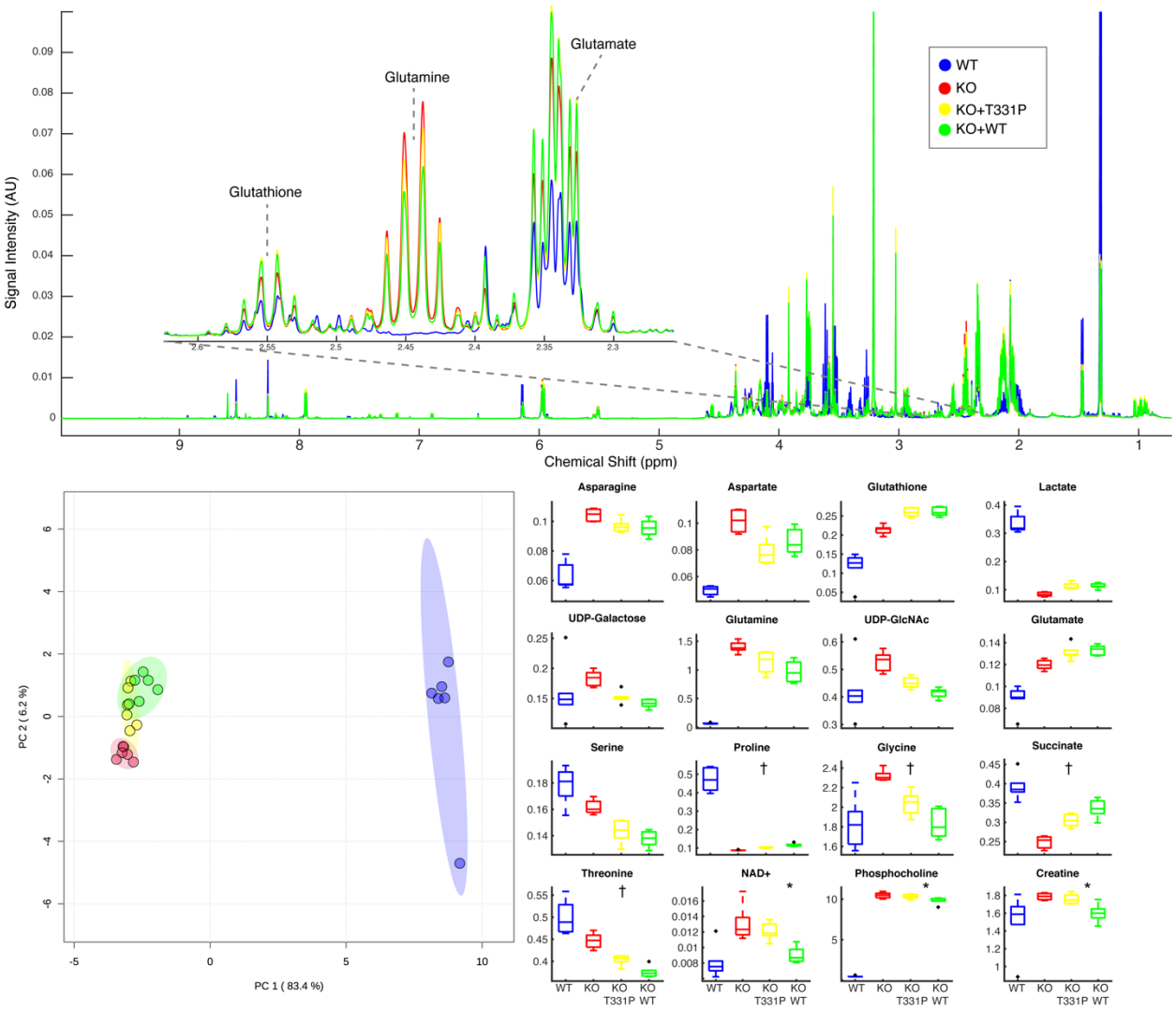
Supplemental Table 2.2: All quantified conditioned media metabolites measured by ^{1}H NMR. Values are in arbitrary units. [1]KIC = ketoisovaleric acid, [2]KMV = ketomethylvaleric acid.

Metabolite	Confidence Score	Mean ± SE		log2(FC)	*p*-value	FDR
		MCF-7	Augusta			
Acetate	3	0.615 ± 0.01	5.119 ± 0.240	3.057	6.51E-09	1.69E-07
Pyruvate	4	0.195 ± 0.004	0.392 ± 0.016	1.009	3.47E-07	4.51E-06
Glucose	4	0.967 ± 0.04	0.270 ± 0.061	-1.839	3.41E-06	2.96E-05
Alanine	4	0.368 ± 0.004	0.427 ± 0.014	0.214	1.53E-03	8.46E-03
Proline	4	0.079 ± 0.004	0.107 ± 0.005	0.447	1.63E-03	8.46E-03
Lactate	4	1.533 ± 0.03	1.771 ± 0.058	0.208	3.56E-03	0.015
Ethanol	4	14.230 ± 0.396	15.901 ± 0.407	0.160	0.017	0.062
Glycine	3	0.331 ± 0.001	0.305 ± 0.011	-0.116	0.027	0.088
Glutamate	4	0.450 ± 0.005	0.481 ± 0.014	0.097	0.048	0.127
Glutamine	4	0.806 ± 0.014	0.882 ± 0.033	0.131	0.049	0.127
Isoleucine	4	0.648 ± 0.005	0.603 ± 0.027	-0.102	0.107	0.252
myo-Inositol	4	0.157 ± 0.009	0.179 ± 0.017	0.193	0.242	0.485
Sucrose	3	0.068 ± 0.004	0.081 ± 0.010	0.252	0.233	0.485
Arginine	4	0.926 ± 0.012	0.958 ± 0.028	0.049	0.288	0.499
KIC[1]	4	0.072 ± 0.005	0.082 ± 0.008	0.195	0.277	0.499
Histidine	4	0.077 ± 0.005	0.085 ± 0.007	0.146	0.334	0.523
Serine	4	0.236 ± 0.009	0.221 ± 0.012	-0.095	0.342	0.523
KMV[2]	4	0.138 ± 0.001	0.141 ± 0.004	0.029	0.450	0.650
Tyrosine	4	0.152 ± 0.004	0.155 ± 0.003	0.030	0.501	0.685
Phenylalanine	4	0.165 ± 0.004	0.162 ± 0.003	-0.026	0.579	0.753
Leucine	4	0.711 ± 0.008	0.707 ± 0.033	-0.006	0.921	0.953
Lysine	4	0.169 ± 0.002	0.169 ± 0.004	0.007	0.850	0.953
Methionine	4	0.046 ± 0.003	0.045 ± 0.004	-0.042	0.776	0.953
Styachyose-Raffinose	3	0.018 ± 0.003	0.018 ± 0.009	-0.042	0.953	0.953
Threonine	4	0.158 ± 0.013	0.154 ± 0.021	-0.036	0.870	0.953
Valine	4	0.594 ± 0.005	0.595 ± 0.021	0.003	0.945	0.953

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

Supplemental File 1 – Photos of the soft loose skin noted in P3. This patient and his siblings (P2 and P4) have loose and supple skin (but not classic cutis laxa phenotype). P1, who bears the same genotype, has no evidence of a cutaneous phenotype.

Supplemental File 2- NMR spectra, PCA plot and quantification of selected metabolites in the WT, KO, KO + WT and KO + p.Thr331Pro HEK293 cell system



Average ^{1}H NMR spectra of HEK 293 cell extracts. WT – WT HEK293 cells, KO - CRISPR knockout of *ALDH18A1*, KO + p.Thr331Pro (labeled T331P in the figure) - knockout cells exogenously expressing T331P variant of P5CS enzyme. KO + WT - knockout cells

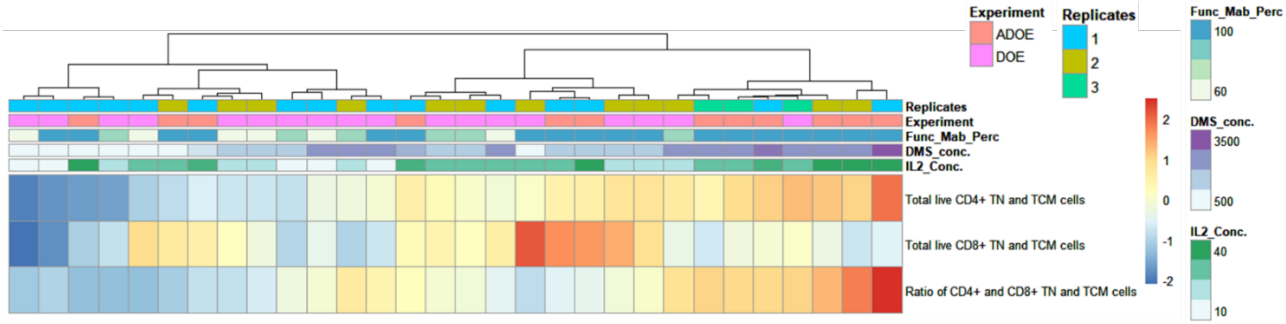
exogenously expressing wild type P5CS enzyme; (B) Principal component analysis scores plot of WT, KO, KO+T331P, and KO+WT cells. Integrated spectral features used for analysis; (C) Box and whisker plots of metabolites annotated from NMR spectra found significantly different by one-way ANOVA (FDR adjusted p-value < 0.05). All comparisons significant between KO and KO+T331P, KO and KO+WT unless otherwise noted. † indicates additional significant comparison between KO+T331P and KO+WT. * indicates significant comparisons between KO and KO+WT, and KO+T331P and KO+WT only.

SUPPLEMENTAL MATERIAL FOR CHAPTER 4

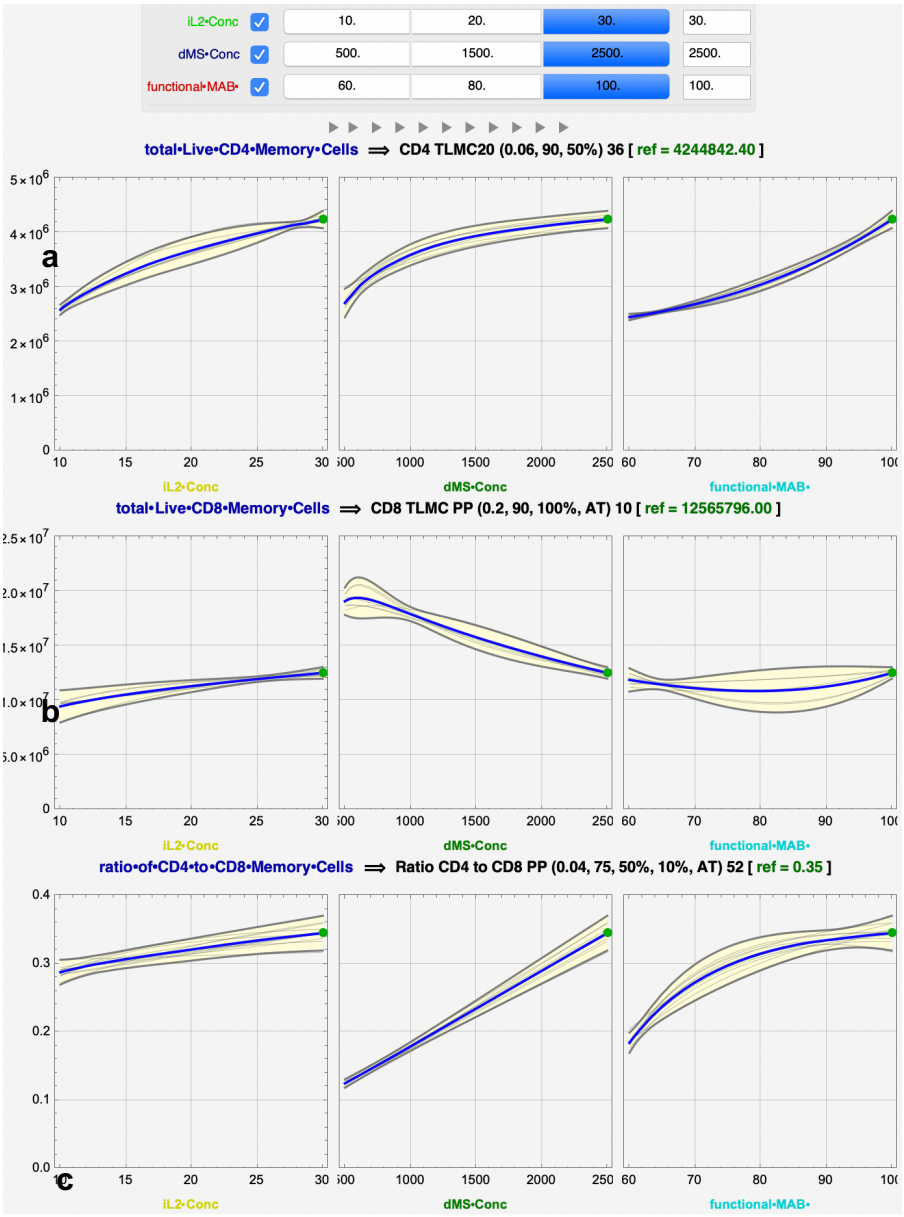
Supplemental information

Dataset S1 (separate file: btm_SuppDatasetS1.xlsx). Process parameters, Cytokine, NMR metabolomics, end-product responses (i.e., T_N+T_{CM} cells), other cell morphology details can be found for both experiments performed (DOE, ADOE). Column names are self-explanatory, and their categories followed as Experiments information, Process Parameters, Media cytokine secretion at day 6, 8, 11, and 14, Media NMR analysis at day 4, 6, 8, 11, and 14, and other info.

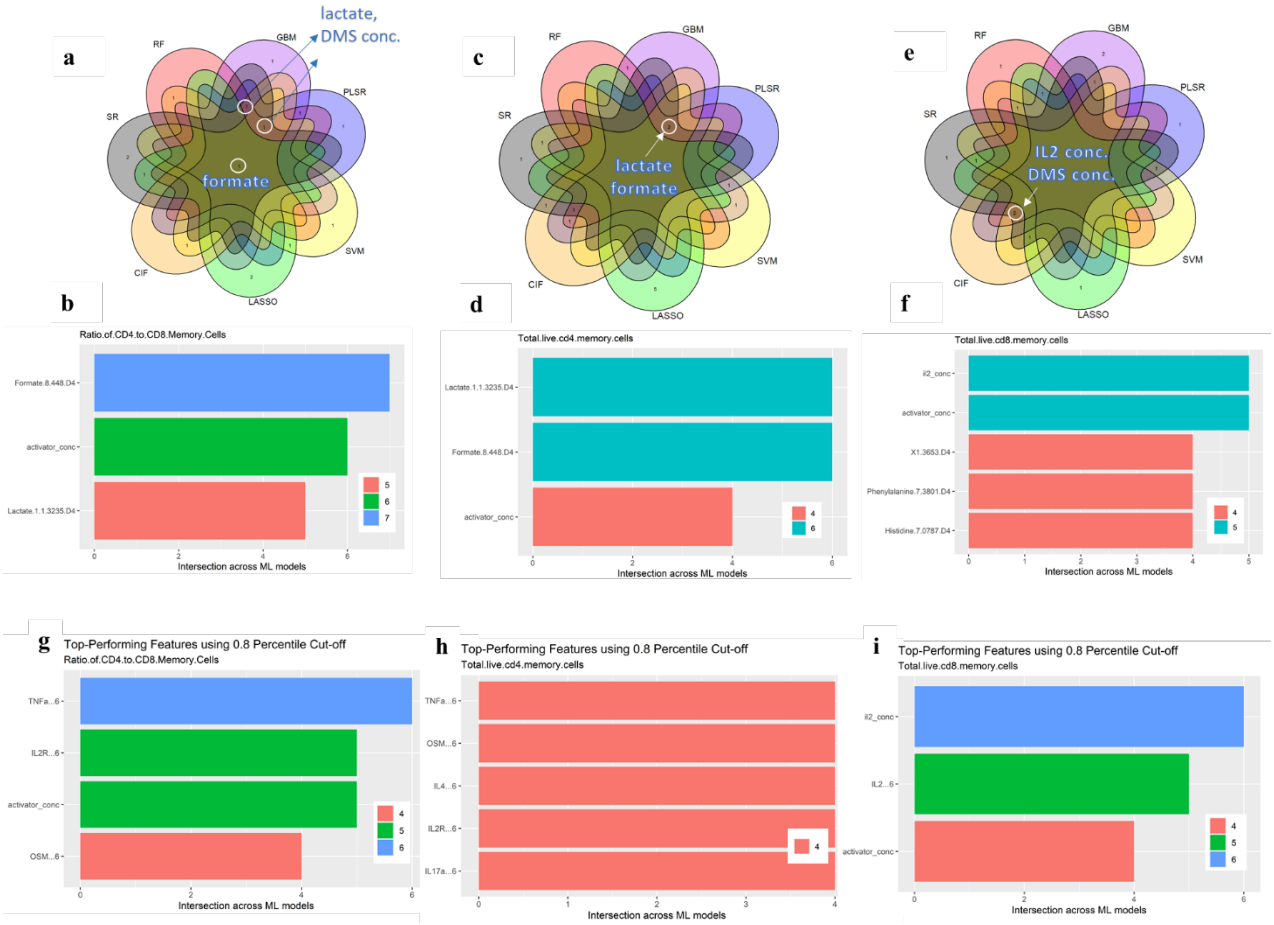
Supplementary Figure



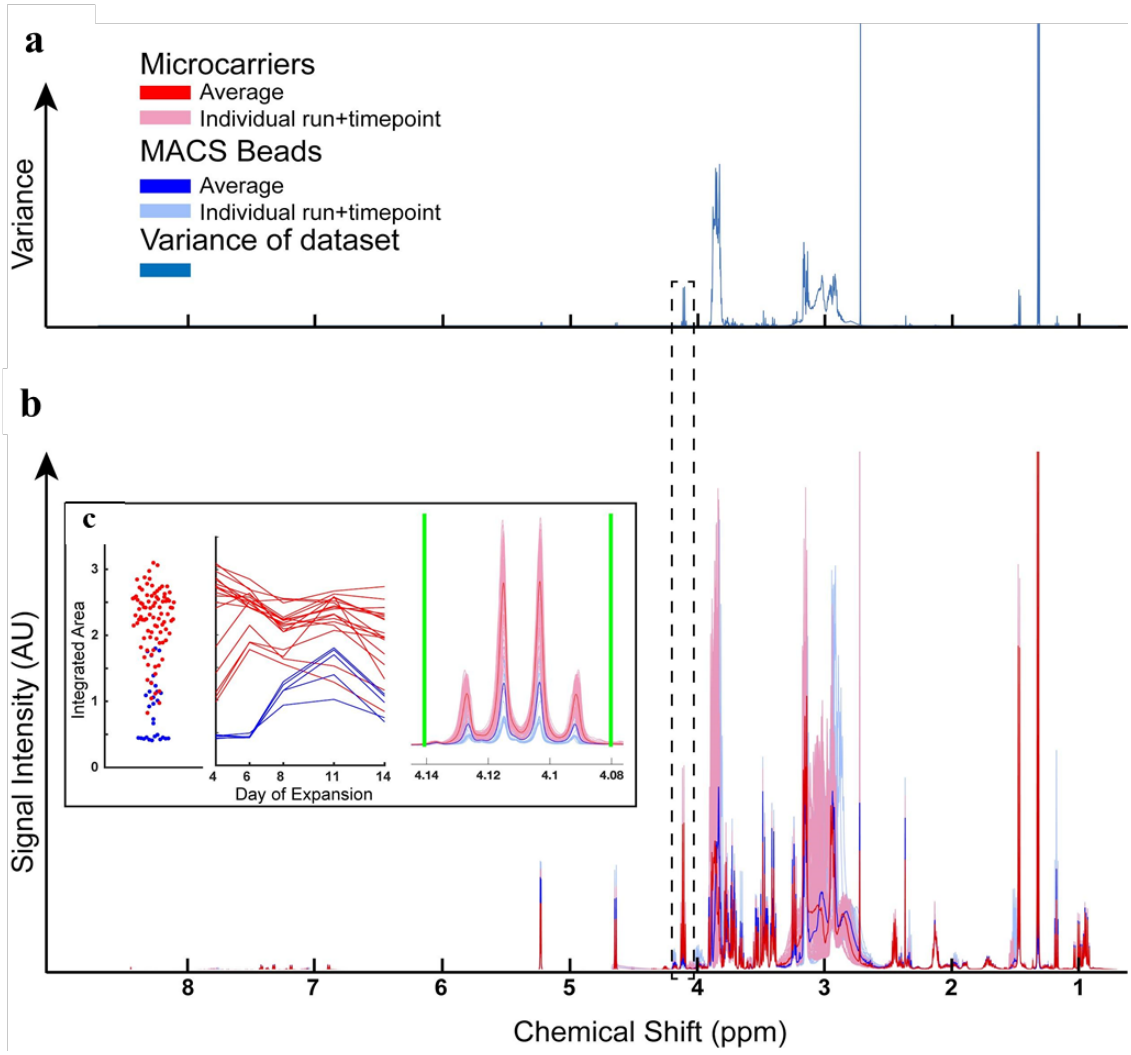
Supp.Fig.S1. Heatmap display of hierarchical clustering for all three T_N+T_{CM} endpoint responses using Ward.D agglomeration and Euclidean distance. Replicates represent the number of samples for that particular process parameter combination.



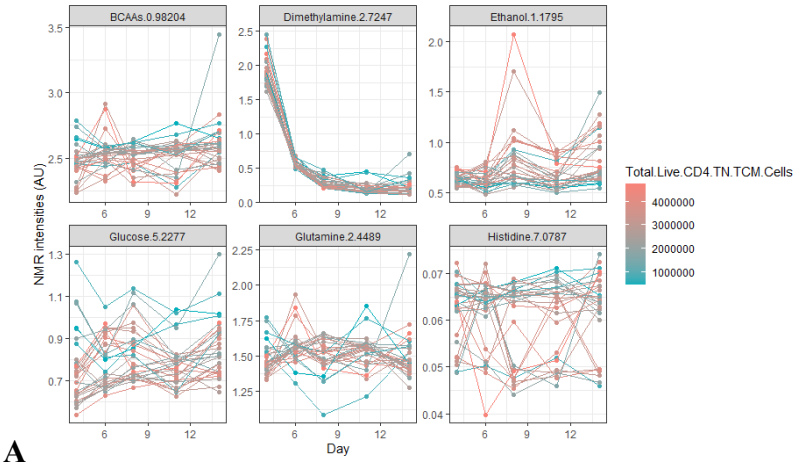
Supp.Fig.S2. Symbolic regression ensemble plots as given by DataModeler optimizing for Total live CD4$^+$ T_N+T_{CM} cells. Predicted response profiles of a) Total live CD4$^+$ T_N+T_{CM} cells, b) Total live CD8$^+$ T_N+T_{CM} cells and c) Ratio of CD4$^+$ to CD8$^+$ T_N+T_{CM} cells at the predicted optimum for Total live CD4$^+$ T_N+T_{CM} cells.



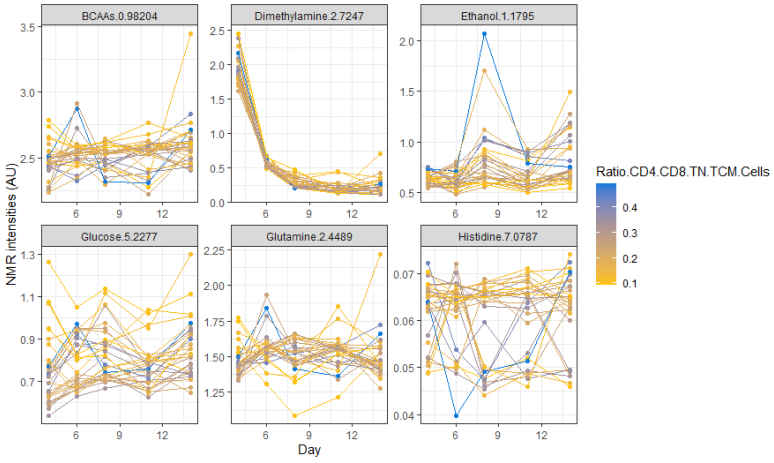
Supp.Fig.S3. Overall feature consensus analysis of top-performing features in single-omics (a-f) NMR models at day 4 and (g-i) Cytokine models at day 6 for **a,b,g**) ratio of total live $CD4^+$ to $CD8^+$ T_N+T_{CM} cells, **c,d,h**) total live $CD4^+$ T_N+T_{CM} cells, and **e,f,i**) total live $CD8^+$ T_N+T_{CM} cells



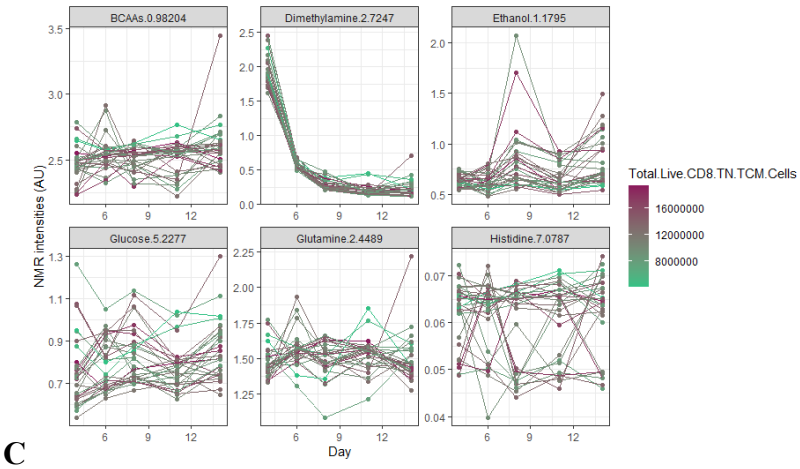
Supp.Fig.S4. Variance based feature selection of NMR features for computational modeling. a) Plot of variance across ^{1}H NMR spectrum for all experimental samples. X-axis matched to below. b) Plot of all experimental spectra for both microcarrier and MACS bead process runs. Averages shown in bold lines. c) Integration of feature highlighted in dashed box. Far right plot shows overlay of all experimental spectra and averages for both groups. Vertical green lines correspond to boundaries for feature integration. Center plot shows the trajectory of integrated values for individual runs (represented as continuous lines) over the expansion period indicated on X-axis. Far left-plot shows a distribution of integrated values for all samples over all timepoints.



A

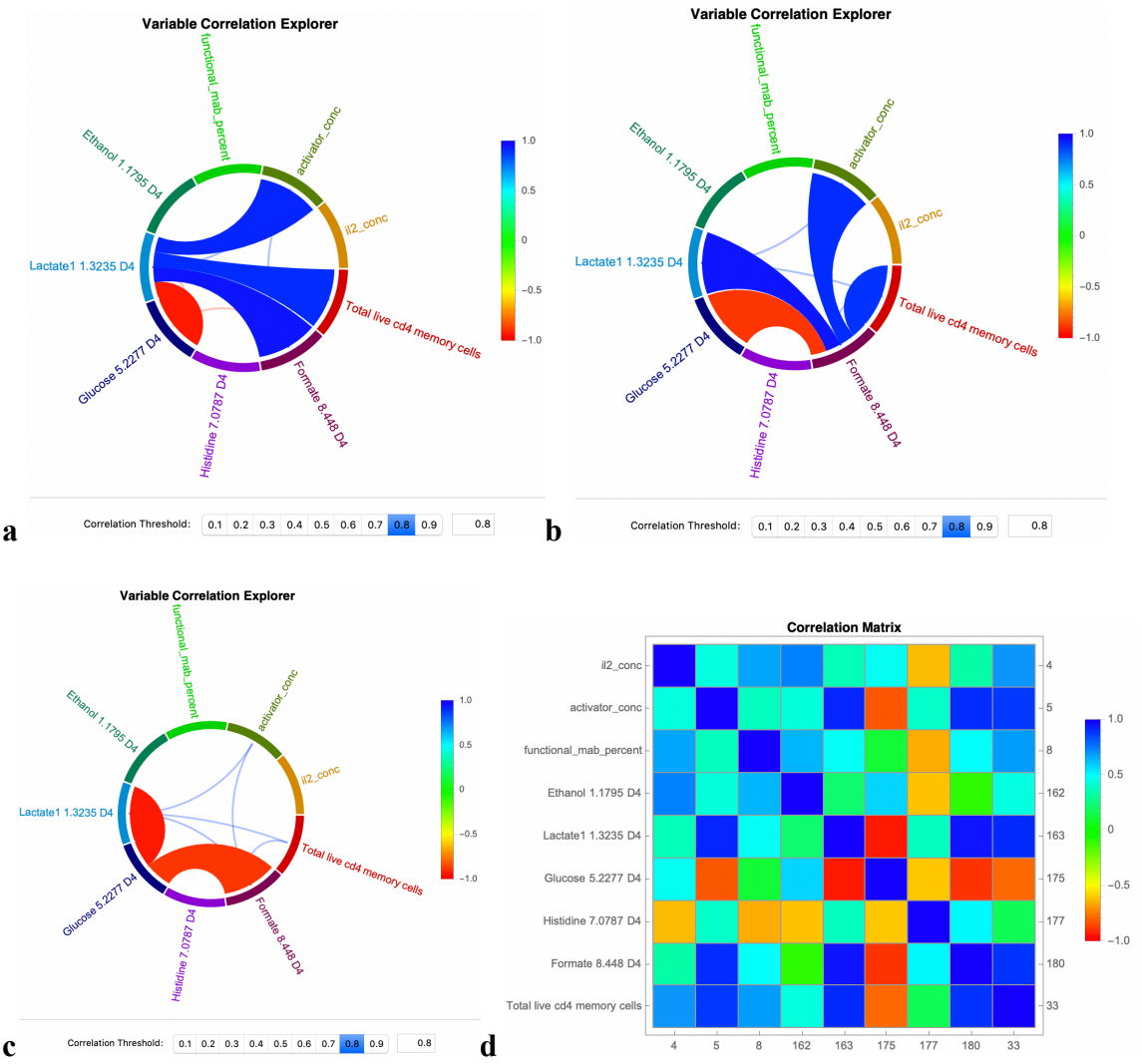


B

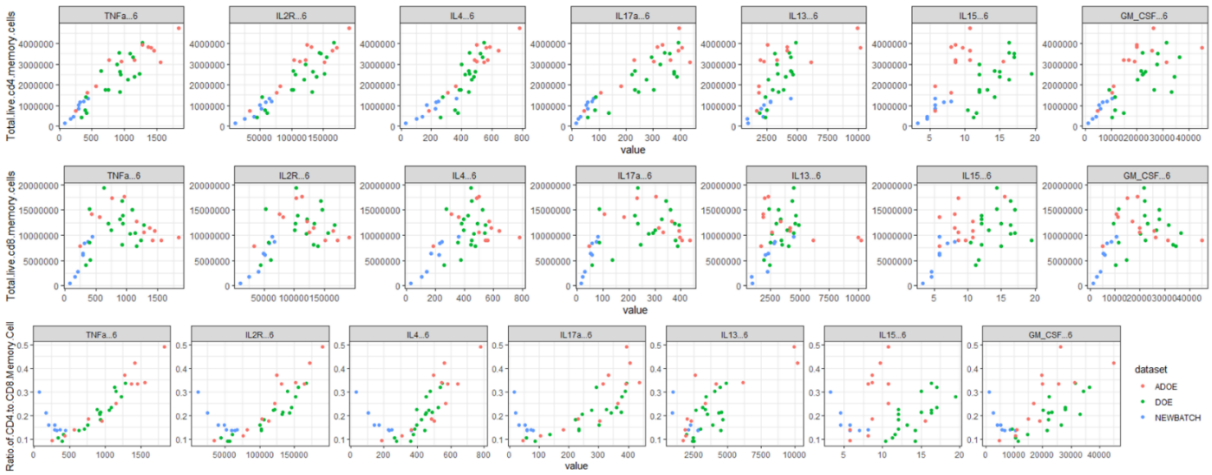


C

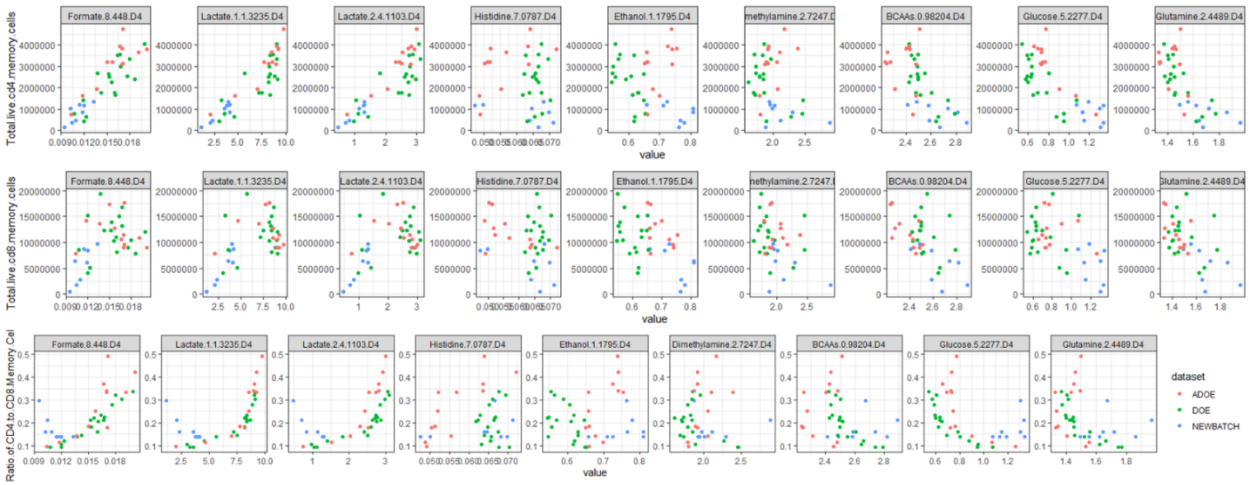
Supp.Fig.S5. Media NMR intensities across monitoring times for a) total live $CD4^+$ T_N+T_{CM} cells, b) ratio $CD4^+/CD8^+$ T_N+T_{CM}, and c) total live $CD8^+$ T_N+T_{CM} cells.



Supp.Fig.S6: NMR Feature Correlation for SR-DataModeler models for NMR media analysis at day 4: **a)** lactate is strongly positively correlated with formate, DMS Conc and Total Live CD4^{+} and negatively correlated with glucose; **b)** formate is strongly positively correlated with lactate, DMS Conc and Total Live CD4^{+} and negatively correlated with glucose; **c)** glucose is negatively correlated with lactate and formate; **d)** NMR correlation Matrix.



Supp.Fig.S7: Intensity values for predictive cytokine features from media analysis on T cell culturing at day 6 for T_N + T_{CM} responses.



Supp.Fig.S8: Intensity values for predictive NMR features from media analysis on T cell culturing at day 4 for T_N + T_{CM} responses.

Supplementary Tables

Supp.Table.S1. Summary Statistics for Day 14 Total Live (CD4+, CD8+) T_N and T_{CM} cells and Ratios for DOE/ADOE

Response	**Experiment**	**Minimum**	**Median**	**Mean**	**Maximum**
Total live CD4+ T_N and T_{CM} cells	DOE: 18-runs	4.3×10^5	2.5×10^6	2.3×10^6	4.0×10^6
	ADOE: 12-runs	7.4×10^5	3.2×10^6	3.1×10^6	4.7×10^6
Total live CD8+ T_N and T_{CM} cells	DOE: 18-runs	4.1×10^6	1.2×10^7	1.1×10^7	1.9×10^7
	ADOE: 12-runs	7.8×10^6	1.1×10^6	1.2×10^6	1.8×10^7
Ratio live CD4+/CD8+ T_N and T_{CM} cells	DOE: 18-runs	0.09	0.21	0.20	0.34
	ADOE: 12-runs	0.09	0.29	0.27	0.49

Supp.Table.S2. Variable combinations across responses from top-performing SR DataModeler

Response	Predictors	Top Symbolic Regression DataModeler Combinations
Ratio CD4+/CD8+ T_N+T_{CM} cells	**PP+N4**	DMS Conc + Functional Mabs %+Histidine+Formate+Lactate
		IL2 Conc+DMS Conc + Functional Mabs %+Histidine+Formate+Ethanol
		IL2 Conc+DMS Conc + Functional Mabs %+Histidine+Lactate
	PP+N6	DMS Conc + Functional Mabs%+UK 1.3653+ Dimethylamine+ Glycine+UK 7.5387
		DMS Conc+Functional Mabs%+ Lactate1+Histidine +UK 7.5387
	PP+S6	DMS Conc + Functional Mabs %+GMCSF+IL2R+MIF
		DMS Conc +GMCSF+IL2R+ IL5+MIF
		DMS Conc + Functional Mabs %+GMCSF+TNFa
	PP+S6+N6	GMCSF+ IL3+TNFa+ Tyrosine+Formate
		IL3+TNFa+ Tyrosine+Formate
		IL3+TNFa+Formate+UK 7.5387
Total Live CD4+ T_N+T_{CM} cells	**PP+N4**	IL2 Conc+DMS Conc+Functional Mabs%+Ethanol+Lactate
		IL2 Conc+DMS Conc+Functional Mabs%+Ethanol+Formate
		IL2 Conc+DMS Conc+Functional Mabs%+Ethanol+Lactate+Formate

		IL2 Conc+DMS Conc+Functional Mabs%+Ethanol+Formate+Histidine
		IL2 Conc+Functional Mabs%+Ethanol+Dimethylamine+Lactate
	PP+N6	IL2 Conc + DMS Conc + Functional Mabs% + Lactate + Phenylalanine
		IL2 Conc + DMS Conc + Functional Mabs%+UK 4.1784
	PP+S6	IL2 Conc + IL13 + IL15 + IL17a + IL2R
		IL2 Conc + IFN alpha+ IL13 + IL15 + IL2R
		IL2 Conc + IL13 + IL15 + IL2R
	PP+S6+N6	IL2 Conc+IFN Alpha+ IL13+IL15+Histidine
		IL2 Conc+ IL13+IL15+IL17a+IL2R+Glycine
		IL2 Conc+IL13+IL15+IL2R+MIF+Glycine
Total Live CD8+ T_N+T_{CM} cells	**PP +N4**	IL2 Conc + DMS Conc + Lactate + Ethanol + Histidine + BCAAs
		IL2 Conc + DMS Conc + Functional Mabs% + Lactate + Ethanol + Histidine
		IL2 Conc + DMS Conc + Formate + Ethanol + Glucose + BCAAs
		IL2 Conc + DMS Conc + Formate + Glucose + BCAAs
		IL2 Conc + DMS Conc + Ethanol + Lactate+ Glutamine
	PP+N6	IL2 Conc + DMS Conc + Ethanol + Pyruvate + UK 7.5387
		IL2 Conc + DMS Conc + Ethanol +Tyrosine

		IL2 Conc + DMS Conc + Ethanol +Lactate
	PP+S6	IL2 Conc+ DMS Conc+ IL15+IL17a+TNFa
	PP+S6+N6	IL2 Conc +DMS Conc + GM CSF+IL15+IL17a+UK 1.5208
		IL2 Conc +DMS Conc + GM CSF+IL15+IL17a+IL2R
		IL2 Conc +DMS Conc +IL15+IL17a+TNFa

Note: UK means unknown or unidentified.

Supp.Table.S3. Variables present in >30% of the top-performing Symbolic Regression models from DataModeler (R2≥ 90%, Complexity ≤ 100) for the different end-product responses.

Input	**Ratio CD4+/CD8+ T_N+T_{CM} cells**	**Total Live CD4+ T_N+T_{CM} cells**	**Total Live CD8+ T_N+T_{CM} cells**
PP+N4	DMS Conc Functional Mabs % Lactate Formate Histidine Ethanol	IL2 Conc DMS Conc Functional Mab % Lactate Formate Ethanol	IL2 Conc DMS Conc Lactate Formate Histidine BCAAs Ethanol
PP+N6	Functional Mabs% DMS conc UK 75387 Dimethylamine Glycine UK 13653 Lactate Histidine	IL2 Conc Functional Mab % DMS Conc Lactate Phenylalanine UK41784	IL2 Conc DMS Conc Ethanol UK 75387 Tyrosine Pyruvate Lactate
PP+S6	DMS conc GMCSF IL2R MIF	IL Conc IL2R IL13 IL15	IL2 Conc DMS Conc IL15 IL17a

	IL5 Functional Mabs % TNFa	IL17a IFN alpha MIF	TNFa
PP+S6+N6	TNFa IL3 Formate Tyrosine GMCSF UK75387	IL2 Conc IL2R IL13 IL15 Glycine Histidine IL17a MIF IFN alpha	IL2 Conc IL15 DMS Conc IL17a GM CSF UK 15208 IL2R TNFa

Supplementary Materials and Methods

Overall multi-omics study design and development. The first DOE resulted in a randomized 18-run I-optimal custom design where each DMS parameter was evaluated at three levels: IL2 concentration (10, 20, and 30 U/□L), DMS concentration (500, 1500, 2500 carrier/□L), and functionalized antibody percent (60%, 80%, 100%). These 18 runs consisted of 14 unique parameter combinations where 4 of them were replicated twice to assess prediction error. Process parameters for the ADOE were evaluated at multiple levels: IL2 concentration (30, 35, and 40 U/□L), DMS concentration (500, 1000, 1500, 2000, 2500, 3000, 3500 carrier/□L), and functionalized antibody percent (100%) as depicted in Fig.1*B*. To further optimize the initial region explored (DOE) in terms of total live $CD4^+$ T_N+T_{CM} cells, a sequential adaptive design-of-experiment (ADOE) was designed with 10 unique parameter combinations, two of these replicated twice for a total of 12 additional samples (Fig.1*B*). The fusion of cytokine and NMR profiles from media to model these responses included 30 cytokines from a custom Thermo Fisher ProcartaPlex Luminex kit and 20 NMR features. These 20 spectral features from NMR media analysis were selected out of approximately 250 peaks through the implementation of a variance-based feature selection approach and some manual inspection steps.

Microcarrier fabrication. Degradable microscaffolds were fabricated as previously described[1]. Briefly, gelatin microcarriers (CuS, GE Healthcare DG-2001-OO) were suspended at 20 mg/mL in 1X phosphate-buffered saline (PBS). Sulfo-NHS-biotin (SNB) (Thermo Fisher 21217 or Apex Bio A8001) was dissolved at 10 µM in ultrapure water and 7.5 µL SNB/mL PBS was added to carrier suspension and allowed to react for 60 min. After washing the carriers three times in PBS, 40 µg/mL streptavidin (Jackson Immunoresearch 016-000-114) was added and allowed to react for 60 min. Biotinylated mAbs against human CD3 and CD28 were combined in a 1:1 mass ratio

and added to the carriers at 2 μg mAbs/mg carriers. To vary the surface concentration of the antibodies, the anti-CD3/anti-CD28 mAb mixture was further combined with a biotinylated isotype control to reduce the overall fraction of targeted mAbs. mAbs were allowed to bind to the carriers for 60 min. All mAbs were low endotoxin azide-free (Biolegend custom, LEAF specification). Fully functionalized DMSs were washed in sterile PBS and washed once again in the cell culture media to be used for the T cell expansion. The surface concentration of the antibodies was quantified as previously described using a bicinchoninic acid assay (BCA) kit (Thermo Fisher 23227)[1].

Flow cytometry. At the end of culture, at least 1e5 T cells from each run were washed with PBS once, resuspended in PBS, and stained with Zombie UV (Biolegend, 423107) for 30 minutes at room temperature in the dark at a 1:1000 dilution. Cells were spun and resuspended in FACS buffer (1X PBS, 2% bovine serum albumin, 5 mM EDTA) and were stained with antibodies according to **Supp.MM.Table.1** for 60 minutes in the dark at 4C.

Supp.MM.Table.1. Flow cytometry antibodies

Antigen	**Fluorophore**	**Vendor**	**Cat Number**
CD3	APC-Fire	Biolegend	34839
CD4	PerCP-Cy5.5	BD	561438
CCR7	AF647	BD	561438
CD62L	PE	BD	341012

Cytokine measurements. Cytokines were measured using a custom ProcartaPlex Luminex kit (Thermo Fisher). The assay was performed using media samples taken at various time points throughout the T cell culture according to the manufacturer's instructions with modifications to

half the reagent requirements. Briefly, an 8-point standard curve was created with all included standards. 25 □L magnetic beads were added to all required wells and washed three times. 25 □L of each standard or sample was added to the wells and the plate was sealed and spun at 850 rpm for 120 minutes followed by three washes. 12.5 □L detection antibody was added followed by sealing the plate and spinning for 60 minutes at 850 rpm and three washes. 25 □L streptavidin PE was added followed by the same spin and wash steps. 120 □L of reading buffer was added to the plate, the plate was analyzed on a BioPlex 200 (BioRad). Any samples that were majority over-range (denoted as "OOR >" in the output spreadsheet) were deemed too concentrated at run at 1/10th their original concentration to put them within range. All samples were run without technical replicates. Luminex data was preprocessed using R for inclusion in the analysis pipeline as follows. Any cytokine level that was over-range ("OOR >" in output) was set to the maximum value of the standard curve for that cytokine. Any value that was under-range ("OOR <" in output spreadsheet) was set to zero. All values that were extrapolated from the standard curve were left unchanged. Data available at **Supp.Dataset.1**.

NMR unknown identification. Several low abundance features selected for analysis did not have database matches and were not annotated. Statistical total correlation spectroscopy[2] suggested that some of these unknown features belonged to the same molecules (not shown). Additional multidimensional NMR experiments will be required to determine their identity.

Symbolic regression. Symbolic regression (SR) was done using Evolved Analytics' Data Modeler software (Evolved Analytics LLC, Midland, MI). Data Modeler utilizes genetic programming to evolve symbolic regression models (both linear and non-linear) rewarding simplicity and accuracy. Using the selection criteria of highest accuracy (R^2>90% or noise-power) and lowest complexity, the top-performing models were identified. Driving variables,

variable combinations, and model dimensionality tables were generated. The top-performing variable combinations were used to generate model ensembles. In this analysis, Data Modeler's *SymbolicRegression* function was used to develop explicit algebraic (linear and nonlinear) models. The fittest models were analyzed to identify the dominant variables using the *VariablePresence* function, the dominant variable combinations using the *VariableCombinations* function, and the model dimensionality (number of unique variables) using the *ModelDimensionality* function. *CreateModelEnsemble* was used to define trustable model ensembles using selected variable combinations and these were summarized (model expressions, model phenotype, model tree plot, ensemble quality, model quality, variable presence map, ANOVA tables, model prediction plot, exportable model forms) using the *ModelSummaryTable* function. Ensemble prediction and residual performance were respectively assessed via the *EnsemblePredictionPlot* and *EnsembleResidualPlot* subroutines. Model maxima (*ModelMaximum* function) and model minima (*ModelMinimum* function) were calculated and displayed using the *ResponsePlotExplorer* function. Trade-off performance of multiple responses was explored using the *MultiTargetResponseExplorer* and *ResponseComparisonExplorer* with additional insights derived from the *ResponseContourPlotExplorer*. Graphics and tables were generated by Data Modeler. These model ensembles were used to identify predicted response values, potential optima in the responses, and regions of parameter values where the predictions diverge the most.

Other ML Methods. Non-parametric tree-based ensembles were done through the *randomForest, gbm,* and *cforest* regression functions in R, for random forest, gradient boosted trees, and conditional inference forest models, respectively. Both random forest and conditional inference forest construct multiple decision trees in parallel, by randomly choosing a subset of

features at each decision tree split, in the training stage. Random forest individual decision trees are split using the Gini Index, while conditional inference forest uses a statistical significance test procedure to select the variables at each split, reducing correlation bias. In contrast, gradient boosted trees construct regression trees in series through an iterative procedure that adapts over the training set. This model learns from the mistakes of previous regression trees in an iterative fashion to correct errors from its precursors' trees (i.e., minimize mean squared errors). Prediction performance was evaluated using leave-one-out cross-validation (LOO)-R^2 and permutation-based variable importance scores assessing % increase of mean squared errors (MSE), relative influence based on the increase of prediction error, coefficient values for RF, GBM, and CID, respectively. Partial least squares regression was executed using the *plsr* function from the *pls* package in R while LASSO regression was performed using the *cv.glmnet* R package, both using leave-one-out cross-validation. Finally, the *kernlab* R package was used to construct the Support Vector Machine regression models.

Parameter tuning was done for all models in a grid search manner using the *train* function from the *caret* R package using LOO-R^2 as the optimization criteria. Specifically, the number of features randomly sampled as candidates at each split (mtry) and the number of trees to grow (ntree) were tuned parameters for random forest and conditional inference forest. In particular, minimum sum of weights in a node to be considered for splitting and the minimum sum of weights in a terminal node were manually tuned for building the CIF models. Moreover, GBM parameters such as the number of trees to grow, maximum depth of each tree, learning rate, and the minimal number of observations at the terminal node, were tuned for optimum LOO-R^2 performance as well. For PLSR, the optimal number of components to be used in the model was assessed based on the standard error of the cross-validation residuals using the function

selectNcomp from the *pls* package. Moreover, LASSO regression was performed using the *cv.glmnet* package with *alpha* = 1. The best lambda for each response was chosen using the minimum error criteria. Lastly, a fixed linear kernel (i.e., svmLinear) was used to build the SVM regression models evaluating the cost parameter value with best LOO-R^2. Prediction performance was measured for all models using the final model with LOO-R^2 tuned parameters. **Supp.MM.Table.2** shows the parameter values evaluated per model at the final stages of results reporting. Machine learning implementation codes used in this work are available at GitHub (https://github.com/wandaliz/CMaT_TCell_MachineLearning/). DataModeler information can be requested at http://www.evolved-analytics.com/.

Supp.MM.Table.2. ML parameter values evaluated and tuned

ML Model	**Tuned Parameter Values**
RF	ntree=c(500,1000,1500,2000,2500) mtry=all possibilities
GBM	interaction.depth=c(1:4) n.trees = (1:20)*10 shrinkage=c(0.1,0.01, 0.02) n.minobsinnode=c(2:6) bag.fraction=0.5
CIF	mtry=all possibilities ntree*=100 minsplit* = 6 minbucket* = 3

LASSO	alpha=1 lambda=seq(0.001,0.05,by = 0.001)
PLSR	ncomp = 1:15
SVM	svmLinear cost=seq(0.05,2,.05)
	*other values besides the ones shown were optimized manually

Machine Learning Consensus Analysis. All regression methods were executed, and the high-performing models were used to perform a consensus analysis of the important variables to extract potential critical quality attributes and critical process parameters predictive of T-cell potency, safety, and consistency at the early stages of the manufacturing process. Consensus analysis of the relevant variables extracted from each machine learning model was done to identify consistent predictive features of quality at the early stages of manufacturing. First importance scores for all features were measured across all ML models using *varImp* with *caret* R package except for scores for SVM which *rminer* R package was used. These importance scores were percent increase in mean squared error (MSE), relative importance through average increase in prediction error when a given predictor is permuted, permuted coefficients values, absolute coefficient values, weighted sum of absolute coefficients values, and relative importance from sensitivity analysis determined for RF, GBM, CIF, LASSO, PLSR, and SVM, respectively. Using these scores, key predictive variables were selected if their importance scores were within the 80^{th} percentile ranking for the following ML methods: RF, GBM, CIF, LASSO, PLSR, SVM while for SR variables present in >30% of the top-performing SR models from Data Modeler ($R2 \geq 90\%$, Complexity ≤ 100) were chosen to investigate consensus except for NMR

media models at day 4 which were considered a combination of the top-performing results of models excluding lactate ppms, and include those variables which were in > 40% of the best performing models. Only variables with those high percentile scoring values were evaluated in terms of their logical relation (intersection across ML models) and depicted using a Venn diagram from the *venn* R package.

Supplementary References

1. Dwarshuis NJ, Song HW, Patel A, Kotanchek T, Roy K. Functionalized microcarriers improve T cell manufacturing by facilitating migratory memory T cell production and increasing CD4/CD8 ratio. *bioRxiv*. Published online 2019:646760.

2. Holmes E, Cloarec O, Nicholson JK. Probing Latent Biomarker Signatures and in Vivo Pathway Activity in Experimental Disease States via Statistical Total Correlation Spectroscopy (STOCSY) of Biofluids: Application to HgCl2 Toxicity. *J Proteome Res*. 2006;5(6):1313-1320. doi:10.1021/pr050399w