# HOW TO PLAN FOR THE UNKNOWN: A GUIDE TO *CAENORHABDITIS ELEGANS* AS A MODEL ORGANISM IN METABOLOMICS

by

AMANDA OLIVIA SHAVER

(Under the Direction of ARTHUR S. EDISON)

## ABSTRACT

Untargeted metabolomics studies measure tens of thousands of features in a single biological sample. However, most features detected are unknown compounds. This creates a great need for a reliable approach to identify unknown compounds. A factor contributing to the large number of unknown compounds is that metabolomics studies usually apply genetics and pathway mapping *after* analytical measurements are collected. The problem with this approach is that unknown spectral features are challenging to resolve outside the context of a pathway. Here, we put genetic strain selection *before* data collection, thus established and hypothesized pathways can put unknown spectral features into context and help narrow possibilities during compound identification. Here, the model organism *Caenorhabditis elegans* is used to develop, test, and validate a pipeline to identify unknown metabolites. First, culturing and assaying large mixed-stage *C. elegans* populations in large-scale culture plates yield enough animals to collect phenotypic and population data, along with analytical chemical data. This method standardizes culturing conditions crucial for reproducible data. Second, three disparate study groups of *C. elegans* strains are compared (*i.e.*, genetically distinct natural strains; primary and secondary metabolism mutants) to showcase how an augmented design coupled with meta-analysis effectively handles known obstacles in metabolomics experiments to compare data in long-term studies. Technical obstacles encompassing non-linear batch variation, limited overlap in technology coverage, instability of spectral features, and challenging statistical analysis caused by heteroscedasticity are overcome using our approach. This project demonstrates the importance of using pipeline validation and proper study design, yielding reliable data for downstream unknown compound identification and metabolic pathway interpretation.

INDEX WORDS: *C. elegans*, untargeted metabolomics, meta-analysis, experimental design

# HOW TO PLAN FOR THE UNKNOWN: A GUIDE TO *CAENORHABDITIS ELEGANS* AS A MODEL ORGANISM IN METABOLOMICS

by

AMANDA OLIVIA SHAVER

B.S., University of Kansas, 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

HOW TO PLAN FOR THE UNKNOWN: A GUIDE TO *CAENORHABDITIS ELEGANS* AS A MODEL ORGANISM IN METABOLOMICS

by

AMANDA OLIVIA SHAVER

| | |
|---|---|
| Major Professor: | Arthur S. Edison |
| Committee: | Erik Andersen |
| | Casey Bergman |
| | Franklin E. Leach III |
| | Patricia Moore |
| | John Wares |

Electronic Version Approved:

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

# ACKNOWLEDGEMENTS

I would like first to acknowledge my advisor, Dr. Arthur Edison, who believed in me as a scientist and welcomed me into his lab, and provided me with the best community and support imaginable. I could not have asked for a more kind, imaginative, and patient mentor during my graduate training. May I become a scientist who is just as humble, creative, innovative, and kind as you are. To my committee who has supported me and championed my success even during the darkest hours of my training. I will always be grateful for your support and encouragement. I want to thank “Team Worm,” Goncalo Gouveia and Brianna Garcia, for being the best colleagues, companions, and friends. I do not want to imagine what graduate school would have been like without you, and I hope no amount of space or time will break our Team Worm bond. To honorary Team Worm member, Max Colonna, thank you for your friendship, feedback, and endless Slack messages filled with lots of laughs. A special thank you to Pamela Kirby, who provided me with ample technical support and was my right-hand woman in the madness of generating samples over the years for this project. To my partner, Joel Campbell, for listening to all I have to say about worms at any time. Thank you to my family for pretending to be excited about my research, but always truly caring about my happiness and success as a scientist. Stephanie Halmo and Cecile Riker, for your friendship both in and outside the dance studio. Many, many others whom I can’t name here.

## TABLE OF CONTENTS

APPENDICES

# LIST OF TABLES

Page

# LIST OF FIGURES

Page

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BA | Bland Altman |
| BMRB | Biological Magnetic Resonance Data Bank |
| BRM | Biological reference material |
| CeNDR | *Caenorhabditis elegans* Natural Diversity Resource |
| CGC | *Caenorhabditis* Genetics Center |
| CM | Central metabolism |
| COLMARm | Complex Mixture Analysis by NMR |
| CT | Controlled temperature |
| CV | Coefficient of variation |
| DDA | Data-dependent acquisition |
| EXT | Optical extinction |
| FE | Fixed effects |
| HILIC | Hydrophilic interaction liquid chromatography |
| HMDB | Human Metabolome Database |
| IBAT | Iterative batch averaging method |
| IPA | Isopropanol |
| LC-MS | Liquid chromatography-mass spectrometry |
| LPFC | Large particle flow cytometer |
| LSCP | Large scale culture plate |
| NGMA | Nematode growth media agarose |
| NMR | Nuclear magnetic resonance |

| | |
|---|---|
| NS | Natural strains |
| PCA | Principal component analysis |
| QA | Quality assurance |
| QC | Quality control |
| RP | Reverse phase |
| RT | Room temperature |
| SED | Standard Euclidean Distance |
| SMD | Standardized mean difference |
| TIC | Total ion count |
| TOF | Time of flight |
| UGT | UDP-glycosyltransferases |

# CHAPTER 1

# INTRODUCTION AND LITERATURE REVIEW

## Introduction to *Caenorhabditis elegans*

Victor Nigon and Ellsworth Dougherty were the first to study the small, free-living nematode *Caenorhabditis elegans* in the 1940s[1]. However, it was not until the early seventies that the worm gained scientific traction when introduced by Sydney Brenner as a multicellular genetic model[2]. *C. elegans* has few somatic cells, which made it possible to reconstruct the cell lineage and map the wiring to the nervous system, one of the initial traits that made this worm an invaluable resource[3]. Since then, *C. elegans* has revolutionized cellular and molecular biology, developmental biology, neurobiology, and genetics[4-7]. *C. elegans* grows well under laboratory conditions, either on agar plates or in liquid culture, feeding on *E. coli* bacteria[8]. The worm has a quick life cycle taking approximately three days to grow from embryo to egg-laying adult in the laboratory-adapted Bristol strain, N2 (**Figure 1.1**)[9]. *C. elegans* primarily exists as a self-fertilizing hermaphrodite, producing males at a low frequency (<1% in N2), which allows for ease of generation as populations are driven to homozygosity (*i.e.*, hermaphrodites cannot mate with other hermaphrodites; thus, strains are nearly isogenic)[8]. This trait allows for the generation of new recombinant strains via simple genetic crosses. A single hermaphrodite can produce up to 300 offspring, and if mated with males, hermaphrodites can produce up to 1000 offspring[8]. Importantly, nematodes can be cryopreserved for long-term storage which allows strains to be preserved without worrying about mutation accumulation over time[3].

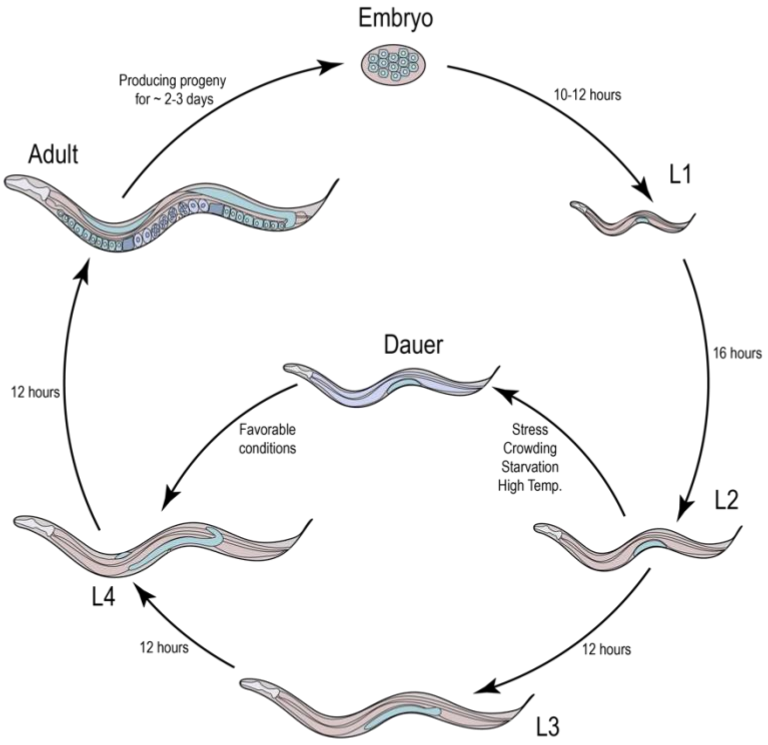


**Figure 1.1** ***Caenorhabditis elegans*** **life cycle.** Worms hatch from an embryo where they start to eat and increase in size through four larval stages (*i.e.*, L1 – L4). Newly hatched L1 worms are approximately 0.25 millimeters (mm) long and continue to grow, shedding their cuticle after a period of inactivity (*i.e.*, lethargus)[10], and continue into the next larval stage with a new cuticle. Approximately 12 hours after the L4 molt, adult hermaphrodites (1 mm) begin producing progeny. Hermaphrodites can lay eggs for 2-3 days, laying about 300 embryos. Males are not shown in this diagram but can be distinguished starting at the L4 stage. After reproduction, hermaphrodites can live for several weeks before dying of senescence. If animals are under stressful conditions at the L2 stage, animals can molt in an alternative L3 stage (*i.e.*, dauer), a stage of arrested development where the animal can survive for months. All times displayed for development are based on the laboratory-derived Bristol strain, N2, at 20°C[9].

## *C. elegans* as a Model Organism in Metabolomics

Just as *C. elegans* has proved to be a powerful model organism studied widely in biology, *C. elegans* also has some key advantages to be a key model organism in metabolomics. *C. elegans* has similar nutritional requirements as humans, including the same essential amino acids and vitamins, homologous metabolic pathways, and canonical metabolic regulatory pathways such as insulin and target of rapamycin (TOR) signaling[9, 11, 12]. The nematode has a well-annotated genome in addition to a variety of genome-wide technologies that are available and enable the genome-scale characterization of metabolic phenotypes, for instance, in response to dietary changes[4, 13, 14]. These include genome wide RNAi libraries and deletion mutants grown and maintained by the *Caenorhabditis* Genetics Center (CGC)[15, 16]. In addition, large-scale protein-protein and protein-DNA interaction mapping efforts have identified molecular connections that can be integrated with phenotypic data[7, 17, 18]. These tools have helped gain new insights into metabolic gene regulatory networks, which showcase clear advantages in using *C. elegans* for system-level studies of metabolism.

In addition to its most well-known and recognized contributions to biomedicine – the discovery of genes involved in apoptosis, RNAi, and the first use of green fluorescent protein to determine cell-specific gene expression were accomplished first in *C. elegans*, adding to the compelling reasons to use *C. elegans* as a platform in unknown compound identification[6, 19, 20]. From a practical perspective, *C. elegans* can be easily cultured in plates or liquid with a diet of a simple bacteria, *E. coli*[8]. From a resource perspective, there are many reasons to use *C. elegans*. First, the CGC maintains and distributes ~ 10,000 genetic strains (and growing) of *C. elegans* and related nematodes, providing many mutant strains for primary and secondary metabolism. Second, the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) contains hundreds of genetically diverse natural strains[21]. To date, greater than 100,000 natural isotypes represent more than 85% of the total genetic diversity of the species. For example, 120 of these strains capture as much genetic diversity as some human sub-populations (*e.g.*, Europeans)[21]. Therefore genome-wide association studies (GWAS) can readily be performed to identify genes that underlie quantitative trait differences[21]. From a scientific perspective, a growing body of work is demonstrating the importance of *C. elegans* as a chemical model organism[22-24]. Studies in chemical ecology, metabolomics, and pheromone signaling have collectively showcased how *C. elegans* is an indispensable tool and model organism to these fields[25-28]. These studies laid the foundation for metabolomics research to leverage the extensive genetic and phenotypic data in *C. elegans*. Just as model organisms were important for completing the human genome (*i.e.*, used in DNA sequencing technologies and data analysis techniques), they also play an essential role in metabolomics technique development[14, 27, 29].

As the metabolome can vary greatly among genotypes within a species as well as across environments, using a model organism such as *C. elegans* assists in identifying the differences detected in the metabolome due to genetic and environmental effects. We believe that systematic comparisons between genes and metabolites for many organisms will enhance our understanding of evolution and biology, and aid in the identification of unknown metabolites[30-33].

Thus, when approaching an untargeted metabolomics study to identify unknown metabolites where lots of variabilities are at play, using a model organism as the biological subject can be a great aid. Much of what we know about biological processes in every scientific field has been learned through basic research in model organisms. Specifically, *C. elegans*, one of the most extensively studied animals with ample genetic resources, is a perfect model for untargeted metabolomics studies[4, 15, 21, 34].

## Introduction to Metabolomics

Metabolomics is the study of metabolites, a collection of small molecules (<1500 Da) in a biological system[35]. Metabolites are deeply connected to the biological system interacting with the genome, epigenome, transcriptome, and proteome, representing the most downstream stage in the dynamic biochemical organization that is the central dogma of molecular biology (**Figure 1.2**)[36-40]. Small molecules that make up the metabolome are quite functionally and chemically diverse including, (i) endogenous molecules biosynthesized in primary metabolism (*e.g.*, directly involved in normal growth, development, and reproduction), (ii) secondary metabolite signaling molecules (*e.g.*, not essential for growth, but usually with important ecological functions), (iii) exposome (*i.e.*, external molecules, xenobiotics), and (iv) the microbiome (*i.e.*, molecules from the microbial community)[41]. These small molecules interact and control the functions of DNA, RNA, and proteins through chemical modifications and metabolite-macromolecule interactions[39]. Therefore, metabolites are not only the downstream products of cellular regulatory processes but interactors and direct modulators of biological processes and phenotypes to genetic or environmental changes at a given moment in time[39, 42, 43].

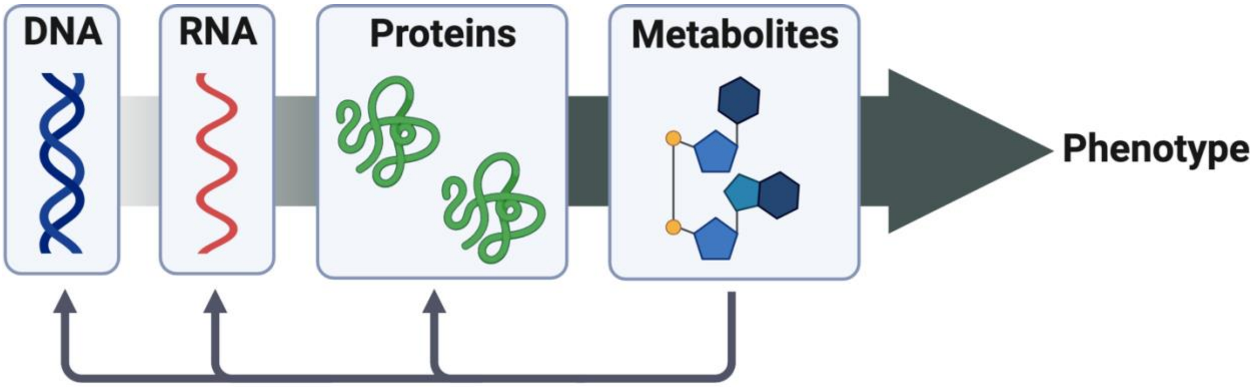


**Figure 1.2 Metabolites interact and connect to all levels of the central dogma.** The central dogma flows from DNA to RNA via transcription, to proteins via translation, and then to metabolites. Unlike DNA, RNA, and proteins, metabolites provide feedback to each central dogma stage. Metabolites are the closest measurable biological output to an organism's phenotype. Created with Biorender.com.

For decades, identifying metabolites as active participants in biological processes has been of interest, from the seminal finding of lactose-dependent regulation of gene expression in bacteria from the *lac* operon to the discovery of the key cellular roles of nutrient and energy sensor, mTOR kinase[12, 44]. As analytical technologies have evolved, the involvement of metabolites in biological processes has become more understood and proliferated[39, 45-47]. More recent work has showcased the vast number of small molecule interactions with macromolecules; for example, endogenous metabolites were found to aid in regulating cellular identity and activity where taurine, an aminosulfonic acid, enhances oligodendrocyte differentiation from stem cells[48]. There has also been compelling evidence of metabolites chemically modifying macromolecules such as lysine acetylation[49]. Altogether, metabolomics has evolved into a field that can unravel complex biological processes and impact many scientific disciplines.

However, with great potential comes significant challenges. The complexity of the metabolome (*e.g.*, ever-changing chemical interactions, chemicals present at vast ranges of concentration) and unique challenges in metabolomics experiments compared to other omics fields cannot be overlooked[50, 51]. One inherent challenge is handling the chemical complexity in metabolomics studies (**Figure 1.3**)[52]. For example, out of the omics fields, metabolomics is often compared to proteomics and has undoubtedly leveraged the proteomics experience. However, metabolomics has different challenges, the biggest being the compound identification process[53-55]. Peptides and proteins are primarily linear polymers that can be sequenced. Proteins are commonly identified by collecting MS/MS spectra that serve as an input to databases (*e.g.*, MASCOT, COMET)[56, 57]. Even unidentified spectra can be processed with *de novo* sequencing resulting in *de novo* peptide sequences assembled to complete sequences (*e.g.*, for antibody screening) or mapped to a protein database (*e.g.*, PepExplorer or MS BLAST)[58, 59].

In contrast, metabolites often lack a common building block. While they do use common elements (*i.e.*, C, H, O, N, S, P), there are on the order of 2 million possible chemical structures (dependent on the number of atoms) a metabolite can take (**Figure 1.3**), making the annotation process challenging [42, 52]. Since the number of building blocks of metabolomics is so vast compared to other omics fields, identifying metabolites becomes the central dilemma of metabolomics (**Figure 1.3**)[52].

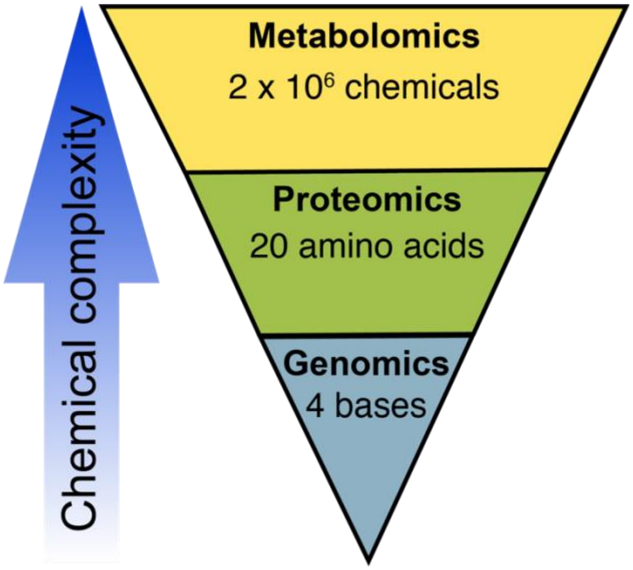


**Figure 1.3 The central dilemma in metabolomics, unknown metabolites.** Chemical complexity increases as the building blocks of the omics field increases. The foundation of metabolomics is so vast that the chemical complexity causes the central dilemma of metabolomics. Figure modified from Wishart 2011[52].

## Targeted vs. Untargeted Metabolomics

The field of metabolomics is primarily broken into two approaches, targeted or untargeted. Targeted metabolomics detects and quantifies sets of metabolites that are targeted and defined *before* the start of an experiment[42]. Untargeted metabolomics allows for the collection of data *without* preexisting knowledge to identify biologically significant known and unknown compounds[42] (**Figure 1.4**).

Broadly, targeted metabolomics studies aim to identify and quantify a limited number of known metabolites through a *validation* process[5, 35, 60-62]. Conversely, untargeted metabolomics studies acquire data for as many chemical species as possible, annotate metabolites, and review known and unknown feature changes, generating new hypotheses and *discovering* new small molecules[63-65]. Data can be used for relative quantification across sample groups and provide a hypothesis that is often further studied with targeted approaches[42].

When identifying metabolites, classes of compounds are defined as “known knowns,” “known unknowns,” or “unknown unknowns” (**Figure 1.5**)[66]. “Known knowns” are metabolites whose identity can be confirmed by analytical platforms, are routinely identified, and exist in reference databases. “Known unknowns” are compounds that are unknown to the investigator at the time of investigation but are cited in chemical literature or chemical databases. “Unknown unknowns" are truly unknown compounds that require extensive sample history and characterization.

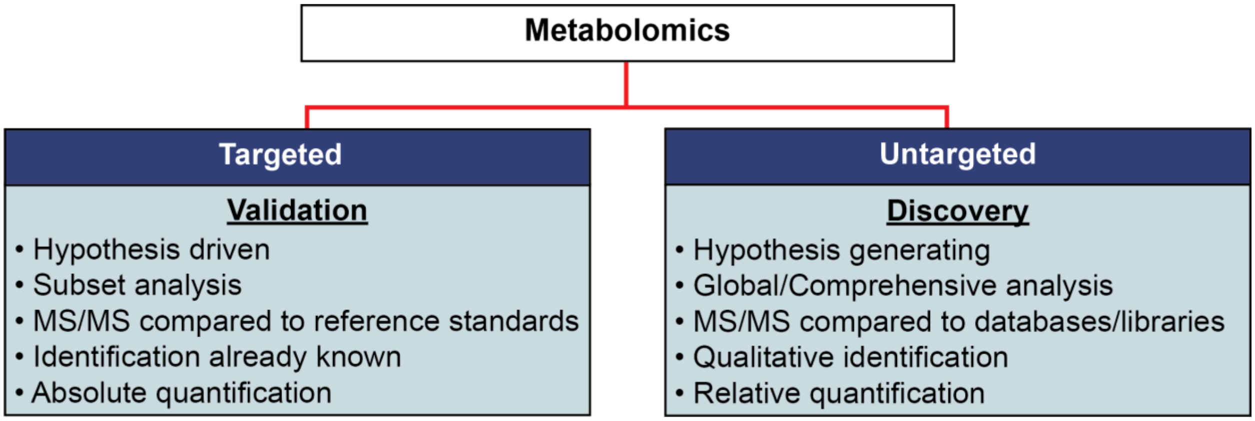
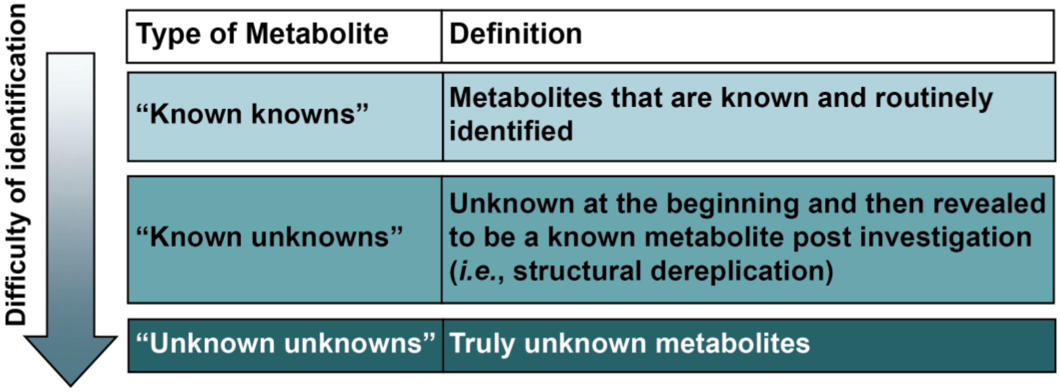

Metabolomics
Targeted
Validation
• Hypothesis driven
• Subset analysis
• MS/MS compared to reference standards
• Identification already known
• Absolute quantification
Untargeted
Discovery
• Hypothesis generating
• Global/Comprehensive analysis
• MS/MS compared to databases/libraries
• Qualitative identification
• Relative quantification

**Figure 1.4 Targeted vs. untargeted metabolomics.** Targeted metabolomics is validation-based, and measures defined groups of metabolites for absolute quantification. Untargeted metabolomics is discovery-based and allows for collection of data without preexisting knowledge. Modified from Schrimpe-Rutledge, Codreanu, Sherrod and McLean [42].


Difficulty of identification

| Type of Metabolite | Definition |
|---|---|
| "Known knowns" | Metabolites that are known and routinely identified |
| "Known unknowns" | Unknown at the beginning and then revealed to be a known metabolite post investigation (*i.e.*, structural dereplication) |
| "Unknown unknowns" | Truly unknown metabolites |

**Figure 1.5 Types of known and unknown metabolites.** Definitions from Little, Cleven, and Brown (2011)[66].

## Untargeted Metabolomics

A major advantage of untargeted metabolomics is the collection of data without preexisting knowledge. The idea that untargeted metabolomics analysis will result in an extensive list of identified small molecules that can be mapped to known networks and pathways is assumed, yet confidently assigning identifications often cannot be made due to the fundamental challenges of the metabolomics identification process[32, 42]. For example, chemical features can be assigned to many tentative or preliminary structures, or a current candidate match may not be available in metabolome databases. Since metabolites lack a specific genetic template, metabolomics databases are currently incomplete, unlike in other omics fields[42, 67]. In-silico metabolite databases can provide guidance, but confirmation of compound identification must be made through methods such as validation of covalent bonds within a molecule to obtain the complete molecular structure in nuclear magnetic resonance (NMR) spectroscopy, standard compound spike-in as the "true" validation of metabolite identify in liquid chromatography-mass spectrometry (LC-MS) or NMR[68], retention times (time taken for the solute to pass through a chromatography column) and tandem mass spectrometry (MS/MS) fragmentation data with a reference standard (**Table 1.1**)[69].

Despite technological advances in NMR and LC-MS, the two analytical instruments most applied to metabolomics studies, metabolite identification remains the overwhelming bottleneck in untargeted metabolomics experiments. Only a tiny fraction, less than 2-10%, of the detected compounds in untargeted metabolomics studies can be reliably annotated, leaving most chemical species unknown[51, 70]. This unknown chemical space continues to be the focus of intense scientific research, developing new methods and hardware for improved chemical separation and structural identification, along with new databases, libraries, and algorithms that can predict chemical molecular properties[51, 71]. While the contribution of new technologies and databases is undoubtedly needed in untargeted metabolomics studies, the field also needs to integrate more comprehensive experimental design approaches that can create a streamlined compound identification process.

## Experimental Design in Large-Scale Untargeted Metabolomics Studies

Without careful planning and implementation of a proper experimental design, large metabolomics datasets can be acquired that do not answer the experiment's biological objectives or, even worse, produce data that leads to false conclusions. Thus, an experimental design that can handle variation (*e.g.*, sample preparation, instrument) appropriately and ensure data collected addresses the study hypothesis and will be usable for downstream experiments should be implemented (**Figure 1.6**). Since untargeted studies provide relative comparisons between samples (*i.e.*, metabolite concentrations are generally not reported) compared to targeted studies that provide quantitative data related to metabolite concentrations, it is even more critical to ensure significant detail and quality assurance occurs in and across all analytical strategies.

Generally, metabolomics experiments have four main stages where variation is introduced, sample development, sample collection, sample preparation, and data acquisition. Thus, when planning out a large-scale untargeted metabolomics study, the following should always be considered: (i) reproducibility of pre-analytical sample collection, (ii) analytical sample preparation across sites and/or over time, (iii) requirement for multiple analytical experiments and instrument maintenance, (iv) randomization strategy for sample preparation and analysis, (v) appropriate quality control (QC) samples, and (vi) post-analytical data processing.

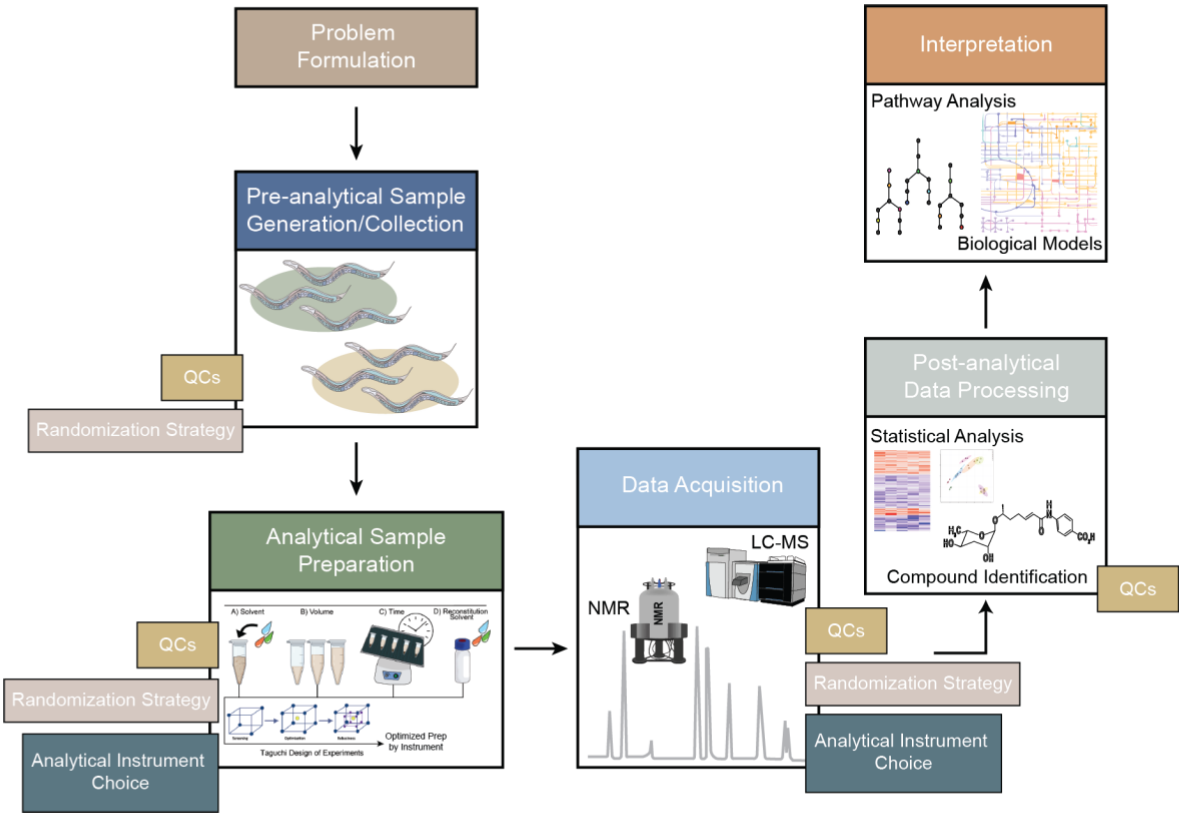


**Figure 1.6 General Metabolomics Experimental Design and Workflow.** At the simplest level, metabolomics studies include a hypothesis (targeted studies) or problem (untargeted studies), pre-analytical sample generation and collection, analytical sample preparation, data acquisition, post-analytical data processing, and interpretation. As variation is a large problem in metabolomics studies proper quality controls (QCs), analytical instrument choice, and randomization strategies should be applied throughout the workflow to ensure introduced variation is not greater than the biological variation being measured.

***(i) Reproducibility of pre-analytical sample collection***

In many untargeted metabolomics studies, samples are collected at multiple sites and/or across long periods. Since large-scale studies can include hundreds of samples collected over time, it is feasible that several researchers will handle and process samples with different equipment and consumables throughout the study. Therefore, a standard operating procedure (SOP) should be designed and used to train all researchers prior to sample collection. Since the metabolome of a sample is a snapshot of the metabolic interactions at a period in time, any differences in sample handling, collection, processing, storage, and transportation can impact the metabolic profile. This is especially true for blood serum and plasma that contain concentrations of enzymes that provide the capability for metabolism to operate post sample collection[72, 73]. Without appropriately quenching metabolism, the metabolic profile of the sample analyzed will differ from the sample's metabolic profile at the time of collection[74]. Therefore, integrating an SOP is essential to minimize intra- or inter-researcher and inter-site variation that could impact the sample. It has been shown that the most significant variation at this stage comes from inter-individual processing/handling associated with the sample[75, 76].

***(ii) Analytical sample preparation across sites and/or over time***

In large untargeted metabolomics studies, analytical sample preparation cannot be done in one setting; therefore, the study should be divided into small analytical experiments to achieve an appropriate analysis. It is good practice to identify the equipment (*e.g.*, centrifuge, homogenizer) or process (*e.g.*, user handling capacity) that is the limiting factor of the study and use that information to guide the structure for sample preparation appropriately.

Additionally, while untargeted metabolomics studies seek full metabolome coverage it is not currently possible from a technical point of view. While a study may be “untargeted”, chosen homogenization and extraction protocols, along with downstream instrument choice, will impact which metabolites are successfully measured and are associated with the biological question [77-79]. Variation in solvent, extraction volume, and reconstitution solvent all affect the metabolic sample output. Taguchi Design of Experiments (DoE) approaches allow for the optimization and should be used when planning and conducting experiments to obtain high quality data to best answer the biological questions at hand (see Appendix C).

***(iii) Requirement for multiple analytical experiments and instrument maintenance***

NMR and LC-MS are complementary analytical platforms used for metabolomics, each with advantages and disadvantages in terms of sensitivity, information content differences, and variation over time (**Table 1.1**)[80, 81]. These platforms can be used individually but can be combined to yield better metabolome coverage and enable more accurate metabolite annotation[82]. The integration of multiple analytical platforms provides deeper understanding of these studies[83, 84]. However, it is essential to understand how each instrument works and the inherent requirements, limits, maintenance, and challenges for both.

NMR spectroscopy exploits quantum mechanical interactions to provide an atomic-level detail of small molecule interactions occurring in a sample. NMR can rigorously quantify abundant compounds in various sample types (*e.g.*, biofluids, tissues, cell extracts, whole organisms) in a non-destructive manner without the need for elaborate sample preparation[80]. In NMR, samples are introduced into the spectrometer in sealed glass tubes, so no sample-instrument interaction results in instrument drift like in LC-MS (see below), however chemical changes can occur due to instrument settings, high salt, or pH issues[85]. NMR offers advantages for compounds that are difficult to ionize or require derivatization for MS. NMR allows for the identification of compounds with identical masses and remains the mainstay for determining structures of unknown compounds[45].

In contrast to NMR, LC-MS has the sensitivity and ability to detect tens of thousands of features in a sample and because of this has become the most widely used technology in metabolomics studies[86, 87]. Fundamentally, MS measures the mass-to-charge ratio (*m/z*) of molecules present in a sample, which can be converted into a neutral mass measurement representing the molecular weight of molecules in a sample, aiding in identification of unique molecules in complex sample types. However, since biological samples are inherently complex and contain high matrix components (*i.e.*, endogenous and exogenous factors that can interfere with the ionization process of the analyte) and metabolites, matrix components and metabolites physically interact with chromatography and MS platforms, leading to degradation of the analytical performance and sensitivity (*e.g.*, a buildup of matrix and metabolites on columns can cause chromatographic performance changes and introduce retention-time variability within data). Another LC-MS challenge is MS 'drift' (*i.e.*, measured variable response, *m/z* and retention time), which can become significant as an experiment runs or over the course of multiple injections – thus integrating multiple injections of the same sample (*e.g.*, QC sample) when running an experiment can aid in identifying drift and subsequent fixes. While there are more examples than showcased here, routine LC-MS instrument maintenance is critical. The variability introduced into the data during sample collection can become significantly equal to or greater than the biological variability in the dataset.

Taken together, the integration of NMR and LC-MS data can be taken for structure elucidation or for better metabolome coverage. Specifically, each analytical approach brings unique attributes – NMR brings structural connectivity and accurate relative concentrations of higher-concentration metabolites, and LC-MS brings molecular weights and elemental formulas to molecules (**Table 1.1**). Additionally, applying an appropriate number of samples or injections ensures that the within-experiment variability introduced is small compared with the biological variability in the sample being studied.

| **Characterization** | **LC-MS** | **NMR** |
|---|---|---|
| Sample | | |
| Intervention | Destructive | Non-destructive |
| Preparation | Extensive | Simple |
| Reproducibility | More variable | Very reproducible |
| Sensitivity | Higher | Lower |
| Speed of Analysis (per sample) | 5-30 min | 20 - 30 min |
| Chromatic Separation | Medium-resolution separation | No separation |
| Structural Information | Low | High |
| Chemophysical Information | More information (time separation) | Less information |
| **Main Advantages** | Soft ionization | Minimal smaple preparation |
| | Large mass range | Non-destructive |
| | Wide dynamic range | Suitable for comounds which are difficult to ionize or require derivatization |
| **Main Distadvantages** | Slow analysis time | Poor sensitivity and dynamic range |
| | Requires ionizable metabolites | Some chemical classes not detected |
| | Inter-and intra-variability | |

**Table 1.1 Two complimentary technologies: NMR vs. LC-MS.** A general comparison of the main characterization differences and advantages between LC-MS and NMR.

### *(iv) Randomization strategy for sample preparation and analysis*

During sample collection, preparation, and analysis, randomization is imperative to ensure sound biological conclusions are reached. The randomization design that will yield the most robust dataset to answer the short and long-term hypotheses should be used for metabolomics studies collected over months or years. One option is to keep experimental units (*e.g.*, genotypes) collected in the same sample collection block (*e.g.*, start/harvest date) together in the same batch for data acquisition, known as conditional randomization (CR). Alternatively, samples can be randomized to new batches for data acquisition without regard to the original sample collection block, known as re-randomization (RR). Simulations of natural populations have shown that CR can best detect genotypic differences when looking for genotype effects. When experimental errors of blocks/batches are confounded, they are jointly estimated as a single effect leaving more degrees of freedom for estimating the residual error. Thus, all genotypes are measured simultaneously in the same ‘environment,’ reducing the genotype comparison variance. In contrast, RR results separate the estimate block/batch effects, effectively constructing different environments for each genotype and making the accurate estimation of these effects necessary for comparing genotypes.

***(v) Appropriate quality control (QC) samples***

Quality control (QC) samples should represent the qualitative and quantitative composition of the subject samples being analyzed in the study; sometimes, this can be the average composition of all samples studied. In analytical experiments, QC samples should be analyzed intermittently, ideally in the beginning, middle, and end of the sample layout and the composition of each QC sample should be equal in theory. Realistically, variation will be introduced during an analytical process (*e.g.*, injection volume, ion-transmission efficiency, needle change), and QC samples will aid in identifying when and where this variation is most prevalent.

Different QC sample types can benefit an untargeted metabolomics study including a reference strain/sample type, pooled reference strains/samples, pooled test strains/samples, and a long-term reference material (RM) (*i.e.*, either an in-house iterative batch averaging method (IBAT)[88] or commercially produced RM). A reference strain or sample type that can be compared to all test strains should always be included. All reference strain samples extracted and included in each instrument run should also be pooled and used. Additionally, a pool of all extracted test strains/samples in each instrument run should be included. Lastly, as we show in Gouveia *et al.* 2021, an IBAT sample is a stable reference material that can be produced over time in any context where multiple small batches of starting material are produced, aliquoted, and then pooled to generate an RM [88]. Throughout all steps of the process, the importance of collecting meta-data can aid in the assessment of sample bias, outliers and can relate data to confounding factors that can be incorporated into statistical analysis.

***(vi) Post-analytical data processing***

There are many facets to post-analytical processing, including peak picking/selection, deconvolution, spectral alignment to identify features and their corresponding abundances (*e.g.*, peak height or peak area), normalization [89-91], scaling [92], transformations (*e.g.*, log or rank transformations) [93], and statistical analysis (*e.g.*, principal component analysis [PCA], analysis of variation [ANOVA], partial least squares – discriminate analysis [PLS-DA], orthogonal projections to latent structures [OPLSDA], or meta-analysis) [50]. It is after these processing steps that the identified statistically significant spectral features can be used in the compound identification process and to determine the structure of the significant molecules, and ultimately make meaningful biological conclusions and connect to other datasets (*e.g.*, RNA Sequencing, GWAS, pathway mapping). Choices made at each step of data processing comes with it benefits and associated pitfalls [50].

## Study Motivation and Summary

In untargeted metabolomics studies we can measure up to tens of thousands of features in a single biological sample, with sensitivity, resolution, and reproducibility by both LC-MS and NMR. However, the major challenge is that most of the features we can now detect are unknown compounds[80]. A factor contributing to the large number of unknown compounds is that metabolomics studies usually apply genetics and pathway mapping *after* collecting analytical measurements[71]. The problem with this approach is that unknown spectral features are challenging to resolve outside of the context of a pathway. Here, we put genetic strain selection *before* data collection, thus established and hypothesized pathways can be used to put unknown spectral features into context and help narrow down possibilities during compound identification. Using this approach, we can leverage publicly available genetic and metabolic pathway data. Known metabolic pathways provide substantial chemical constraints on unknown compound structures which can narrow the options to identify unknowns. However, it is worth mentioning that known pathways do not imply all metabolites on a pathway, as known pathways can still be incomplete and not indicate all possible metabolic interactions in their current state. Various metabolomics studies have demonstrated that "known" pathways are incomplete, and that unknown chemical species are indicative of much more extensive pathways and networks than currently documented[71].

Our project is largely driven by the premise that genetics is a critical but untapped tool that will enable the development of a pipeline aiding in unknown metabolite identification. Genetics provides defined genes, their functions, and involvement in known pathways that enable prediction of gene involvement in metabolic networks[94]. Here we can approach compound identification differently by combining the elements of traditional metabolomics workflows (*i.e.*, LC-MS/MS and NMR) with genetic tools to computationally link molecular formulas with NMR spectra and genetically defined and targeted pathways. This dissertation will cover components of using genetic tools to aid in the path to compound identification.

To demonstrate the overall approach, we use the model organism *C. elegans* because (i) it has extensive genetic and pathway data freely available (*e.g.*, WormBase[13], WormFLux[94], WormCat), (ii) genetically diverse and mutant strains of interest can be obtained by the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR)[21] and Caenorhabditis Genetics

Center (CGC), respectively, (iii) it shares much of its central metabolic pathways with humans, and (iv) can be grown with relative ease in many environmental conditions in laboratory conditions. Our argument for using *C. elegans* as a model organism to address our biological problem of interest is addressed in Chapter 2.

The purpose of this dissertation work is to use the genetic model organism *C. elegans* to (i) develop, test, and validate a pipeline for the identification of unknown metabolites and (ii) improve our understanding of conserved metabolic pathways and identify associated metabolites. In Chapter 2, I lay out our published method for culturing and assaying large-scale mixed-stage *C. elegans* populations in large scale culture plates (LSCP), which generates sufficient numbers of animals to collect phenotypic and population data, along with analytical chemical data (*i.e.*, LC-MS/MS and NMR). This method standardizes culturing conditions crucial for reproducible data[95].

Chapter 3 builds on the data collected in Chapter 2 by showing how an augmented design and meta-analytic approach were used to overcome the well-known technical obstacles to large-scale metabolomics studies encompassing high levels of non-linear batch variation, limited overlap in technology coverage, instability of spectral features, and challenging statistical analysis caused by widespread heteroscedasticity. Two meta-analysis approaches were used on strains collected in Chapter 2 to identify spectral features that differ between (i) the reference strain and each strain and (ii) across all strains in a study group. With this design, we identified putative compounds to prioritize for downstream compound identification.

Taken together, the work presented in this dissertation lays out an approach to better leverage data across metabolomics experiments to have a more efficient process to compound identification, the biggest bottleneck in untargeted metabolomics. Lastly, Chapter 4 reviews how the methods and approaches here (*i.e.*, batch effects, rank transformation, meta-analysis post-processing issues) can be a great asset to ongoing and future metabolomics studies. Additionally, current conclusions, work in progress, and future directions regarding this project will be reviewed.

## References

1. Nigon, V.M. & Felix, M.A. History of research on C. elegans and other free-living nematodes as model organisms. *WormBook* **2017**, 1-84 (2017).
2. Brenner, S. The genetics of behaviour. *Br Med Bull* **29**, 269-271 (1973).
3. Brenner, S. The genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94 (1974).
4. Consortium, C.e.S. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
5. Denzel, M.S., Lapierre, L.R. & Mack, H.I.D. Emerging topics in C. elegans aging research: Transcriptional regulation, stress response and epigenetics. *Mech Ageing Dev* **177**, 4-21 (2019).
6. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811 (1998).
7. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-543 (2004).
8. Stiernagle, T. Maintenance of C. elegans. *WormBook*, 1-11 (2006).
9. Corsi, A.K., Wightman, B. & Chalfie, M. A Transparent Window into Biology: A Primer on Caenorhabditis elegans. *Genetics* **200**, 387-407 (2015).
10. Raizen, D.M. et al. Lethargus is a Caenorhabditis elegans sleep-like state. *Nature* **451**, 569-572 (2008).
11. Watson, E. & Walhout, A.J. Caenorhabditis elegans metabolic gene regulatory networks govern the cellular economy. *Trends Endocrinol Metab* **25**, 502-508 (2014).

12. Zoncu, R., Efeyan, A. & Sabatini, D.M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nat Rev Mol Cell Biol* **12**, 21-35 (2011).

13. Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. & Spieth, J. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res* **29**, 82-86 (2001).

14. Wilson, R.K. How the worm was won. The C. elegans genome sequencing project. *Trends Genet* **15**, 51-58 (1999).

15. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-237 (2003).

16. Rual, J.F. et al. Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. *Genome Res* **14**, 2162-2168 (2004).

17. Deplancke, B. et al. A gene-centered C. elegans protein-DNA interaction network. *Cell* **125**, 1193-1205 (2006).

18. Reece-Hoyes, J.S. et al. Extensive rewiring and complex evolutionary dynamics in a C. elegans multiparameter transcription factor network. *Mol Cell* **51**, 116-127 (2013).

19. Peden, E., Killian, D.J. & Xue, D. Cell death specification in C. elegans. *Cell Cycle* **7**, 2479-2484 (2008).

20. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802-805 (1994).

21. Cook, D.E., Zdraljevic, S., Roberts, J.P. & Andersen, E.C. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res* **45**, D650-D657 (2017).

22. Weinhouse, C., Truong, L., Meyer, J.N. & Allard, P. Caenorhabditis elegans as an emerging model system in environmental epigenetics. *Environ Mol Mutagen* **59**, 560-575 (2018).

23. Hulme, S.E. & Whitesides, G.M. Chemistry and the worm: Caenorhabditis elegans as a platform for integrating chemical and biological research. *Angew Chem Int Ed Engl* **50**, 4774-4807 (2011).

24. Jones, A.K., Buckingham, S.D. & Sattelle, D.B. Chemistry-to-gene screens in Caenorhabditis elegans. *Nat Rev Drug Discov* **4**, 321-330 (2005).

25. Golden, J.W. & Riddle, D.L. A pheromone influences larval development in the nematode Caenorhabditis elegans. *Science* **218**, 578-580 (1982).

26. Jeong, P.Y. et al. Chemical structure and biological activity of the Caenorhabditis elegans dauer-inducing pheromone. *Nature* **433**, 541-545 (2005).

27. Edison, A.S. et al. Metabolomics and Natural-Products Strategies to Study Chemical Ecology in Nematodes. *Integr Comp Biol* **55**, 478-485 (2015).

28. Zdraljevic, S. et al. Natural variation in C. elegans arsenic toxicity is explained by differences in branched chain amino acid metabolism. *Elife* **8** (2019).

29. Tennessen, J.M., Barry, W.E., Cox, J. & Thummel, C.S. Methods for studying metabolism in Drosophila. *Methods* **68**, 105-115 (2014).

30. Wan, Q.L. et al. Metabolomic signature associated with reproduction-regulated aging in Caenorhabditis elegans. *Aging (Albany NY)* **9**, 447-474 (2017).

31. Cho, K., Mahieu, N.G., Johnson, S.L. & Patti, G.J. After the feature presentation: technologies bridging untargeted metabolomics and biology. *Curr Opin Biotechnol* **28**, 143-148 (2014).

32. Karnovsky, A. & Li, S. Pathway Analysis for Targeted and Untargeted Metabolomics. *Methods Mol Biol* **2104**, 387-400 (2020).

33. Wu, S. et al. Mapping the Arabidopsis Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. *Mol Plant* **11**, 118-134 (2018).

34. Edison, A.S. et al. The Time Is Right to Focus on Model Organism Metabolomes. *Metabolites* **6** (2016).

35. Liu, X. & Locasale, J.W. Metabolomics: A Primer. *Trends Biochem Sci* **42**, 274-284 (2017).

36. Guijas, C., Montenegro-Burke, J.R., Warth, B., Spilker, M.E. & Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat Biotechnol* **36**, 316-320 (2018).

37. Metallo, C.M. & Vander Heiden, M.G. Understanding metabolic regulation and its influence on cell physiology. *Mol Cell* **49**, 388-398 (2013).

38. Rabinowitz, J.D. & Silhavy, T.J. Systems biology: metabolite turns master regulator. *Nature* **500**, 283-284 (2013).

39. Rinschen, M.M., Ivanisevic, J., Giera, M. & Siuzdak, G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol* **20**, 353-367 (2019).

40. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

41. Weckwerth, W. & Morgenthal, K. Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* **10**, 1551-1558 (2005).

42. Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **27**, 1897-1905 (2016).

43. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **48**, 155-171 (2002).

44. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356 (1961).

45. Edison, A.S. et al. NMR: Unique Strengths That Enhance Modern Metabolomics Research. *Anal Chem* **93**, 478-499 (2021).

46. Cui, L., Lu, H. & Lee, Y.H. Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases. *Mass Spectrom Rev* **37**, 772-792 (2018).

47. Putri, S.P., Yamamoto, S., Tsugawa, H. & Fukusaki, E. Current metabolomics: technological advances. *J Biosci Bioeng* **116**, 9-16 (2013).

48. Beyer, B.A. et al. Metabolomics-based discovery of a metabolite that enhances oligodendrocyte maturation. *Nat Chem Biol* **14**, 22-28 (2018).

49. Choudhary, C., Weinert, B.T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat Rev Mol Cell Biol* **15**, 536-550 (2014).

50. Gertsman, I. & Barshop, B.A. Promises and pitfalls of untargeted metabolomics. *J Inherit Metab Dis* **41**, 355-366 (2018).

51. Monge, M.E., Dodds, J.N., Baker, E.S., Edison, A.S. & Fernandez, F.M. Challenges in Identifying the Dark Molecules of Life. *Annu Rev Anal Chem (Palo Alto Calif)* **12**, 177-199 (2019).

52. Wishart, D.S. Advances in metabolite identification. *Bioanalysis* **3**, 1769-1782 (2011).

53. Kim, T. et al. A hierarchical approach to removal of unwanted variation for large-scale metabolomics data. *Nat Commun* **12**, 4992 (2021).

54. Blazenovic, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8** (2018).

55. Misra, B.B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *Eur J Mass Spectrom (Chichester)* **26**, 165-174 (2020).

56. Eng, J.K. et al. A deeper look into Comet--implementation and features. *J Am Soc Mass Spectrom* **26**, 1865-1874 (2015).

57. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22-24 (2013).

58. Muth, T., Hartkopf, F., Vaudel, M. & Renard, B.Y. A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *Proteomics* **18**, e1700150 (2018).

59. Leprevost, F.V. et al. PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Mol Cell Proteomics* **13**, 2480-2489 (2014).

60. Dall, K.B. & Faergeman, N.J. Metabolic regulation of lifespan from a C. elegans perspective. *Genes Nutr* **14**, 25 (2019).

61. Jin, K. et al. Genetic and metabolomic architecture of variation in diet restriction-mediated lifespan extension in Drosophila. *PLoS Genet* **16**, e1008835 (2020).

62. Shen, P., Yue, Y. & Park, Y. A living model for obesity and aging research: Caenorhabditis elegans. *Crit Rev Food Sci Nutr* **58**, 741-754 (2018).

63. Menni, C., Zierer, J., Valdes, A.M. & Spector, T.D. Mixing omics: combining genetics and metabolomics to study rheumatic diseases. *Nat Rev Rheumatol* **13**, 174-181 (2017).

64. Lewis, G.D., Asnani, A. & Gerszten, R.E. Application of metabolomics to cardiovascular biomarker and pathway discovery. *J Am Coll Cardiol* **52**, 117-123 (2008).

65. Gebauer, J. et al. A Genome-Scale Database and Reconstruction of Caenorhabditis elegans Metabolism. *Cell Syst* **2**, 312-322 (2016).

66. Little, J.L., Cleven, C.D. & Brown, S.D. Identification of "known unknowns" utilizing accurate mass data and chemical abstracts service databases. *J Am Soc Mass Spectrom* **22**, 348-359 (2011).

67. Matsuda, F. Rethinking Mass Spectrometry-Based Small Molecule Identification Strategies in Metabolomics. *Mass Spectrom (Tokyo)* **3**, S0038 (2014).

68. Dona, A.C. et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput Struct Biotechnol J* **14**, 135-153 (2016).

69. Witting, M. & Bocker, S. Current status of retention time prediction in metabolite identification. *J Sep Sci* **43**, 1746-1754 (2020).

70. Metabolomics: Dark matter. *Nature* **455**, 698 (2008).

71. da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* **112**, 12549-12550 (2015).

72. Teahan, O. et al. Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Anal Chem* **78**, 4307-4318 (2006).

73. Dunn, W.B., Wilson, I.D., Nicholls, A.W. & Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **4**, 2249-2264 (2012).

74. Winder, C.L. et al. Global metabolic profiling of Escherichia coli cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. *Anal Chem* **80**, 2939-2948 (2008).

75. Barton, R.H., Nicholson, J.K., Elliott, P. & Holmes, E. High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. *Int J Epidemiol* **37 Suppl 1**, i31-40 (2008).

76. Dunn, W.B. et al. A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int J Epidemiol* **37 Suppl 1**, i23-30 (2008).

77. Geier, F.M., Want, E.J., Leroi, A.M. & Bundy, J.G. Cross-platform comparison of Caenorhabditis elegans tissue extraction strategies for comprehensive metabolome coverage. *Anal Chem* **83**, 3730-3736 (2011).

78. Mushtaq, M.Y., Choi, Y.H., Verpoorte, R. & Wilson, E.G. Extraction for metabolomics: access to the metabolome. *Phytochem Anal* **25**, 291-306 (2014).

79. Maria John, K.M., Harnly, J. & Luthria, D. Influence of direct and sequential extraction methodology on metabolic profiling. *J Chromatogr B Analyt Technol Biomed Life Sci* **1073**, 34-42 (2018).

80. Markley, J.L. et al. The future of NMR-based metabolomics. *Curr Opin Biotechnol* **43**, 34-40 (2017).

81. Bingol, K. & Bruschweiler, R. Two elephants in the room: new hybrid nuclear magnetic resonance and mass spectrometry approaches for metabolomics. *Curr Opin Clin Nutr Metab Care* **18**, 471-477 (2015).

82. Gonzalez-Dominguez, A. et al. An Overview on the Importance of Combining Complementary Analytical Platforms in Metabolomic Research. *Curr Top Med Chem* **17**, 3289-3295 (2017).

83. Dunn, W.B. et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **6**, 1060-1083 (2011).

84. Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R. & Griffin, J.L. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **40**, 387-426 (2011).

85. Wishart, D.S. Quantitative metabolomics using NMR. *Trends in Analytical Chemistry* **27** (2008).

86. Gika, H., Virgiliou, C., Theodoridis, G., Plumb, R.S. & Wilson, I.D. Untargeted LC/MS-based metabolic phenotyping (metabonomics/metabolomics): The state of the art. *J Chromatogr B Analyt Technol Biomed Life Sci* **1117**, 136-147 (2019).

87. Kodra, D. et al. Is Current Practice Adhering to Guidelines Proposed for Metabolite Identification in LC-MS Untargeted Metabolomics? A Meta-Analysis of the Literature. *J Proteome Res* (2021).

88. Gouveia, G.J. et al. Long-Term Metabolomics Reference Material. *Anal Chem* **93**, 9193-9199 (2021).

89. Li, B. et al. NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* **45**, W162-W170 (2017).

90. Ejigu, B.A. et al. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *OMICS* **17**, 473-485 (2013).

91. Rosen Vollmar, A.K. et al. Normalizing Untargeted Periconceptional Urinary Metabolomics Data: A Comparison of Approaches. *Metabolites* **9** (2019).

92. Tugizimana, F., Steenkamp, P.A., Piater, L.A. & Dubery, I.A. A Conversation on Data Mining Strategies in LC-MS Untargeted Metabolomics: Pre-Processing and Pre-Treatment Steps. *Metabolites* **6** (2016).

93. Di Guida, R. et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **12**, 93 (2016).

94. Yilmaz, L.S. & Walhout, A.J. A Caenorhabditis elegans Genome-Scale Metabolic Network Model. *Cell Syst* **2**, 297-311 (2016).

95. Shaver, A.O., Gouveia, G.J., Kirby, P.S., Andersen, E.C. & Edison, A.S. Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations. *J Vis Exp* (2021).

# CHAPTER 2

# METHOD TO CULTURE AND ASSAY LARGE-SCALE MIXED-STAGE *CAENORHABDITIS ELEGANS* POPULATIONS FOR OMICS STUDIES[1]

[1] Shaver, A. O., Gouveia, G. J. , Kirby, P. S., Andersen E. C., and Edison, A. S. Culture and Assay of Large-Scale Mixed-Stage *Caenorhabditis elegans* Populations. *Journal of Visualized Experiments* (171), e61453, doi:10.3791/61453 (2021).
Reprinted here with permission from the publisher.

Chapter 2 is adapted from Shaver, A. O., Gouveia, G. J., Kirby, P. S., Andersen, E. C., Edison, A. S. Culture and Assay of Large-Scale Mixed-Stage *Caenorhabditis elegans* Populations. *J. Vis. Exp.* (171), e61453, doi:10.3791/61453 (2021), available at https://www.jove.com/t/61453/culture-assay-large-scale-mixed-stage-caenorhabditis-elegans. My role on this project was to (i) generate a reproducible growth protocol to achieve large mixed-stage populations of *C. elegans* for analytical measurements, (ii) collect and analyze phenotypic data for each sample, (iii) generate all figures for the *JoVE* manuscript, (iv) write the *JoVE* manuscript, and (v) catalog and distribute samples for downstream analytical analysis. Collaborator's roles were as follows: Goncalo J. Gouveia generated the stable *E. coli* food, through an iterative batch averaging method (IBAT), to feed the worms. The IBAT process was published after the *JoVE* manuscript and is available at https://pubs.acs.org/doi/full/10.1021/acs.analchem.1c01294. Pamela S. Kirby sterilized equipment, prepared the nematode growth media agarose, and spread bacterial lawns for downstream sample growth. Erik C. Andersen provided *C. elegans* strains from the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) along useful feedback and troubleshooting ideas. Arthur S. Edison provided useful intellectual contributions, feedback, and financial support. Research reported in this manuscript was supported by the National Institutes of Health Award Number U2CES030167.

## Abstract

*Caenorhabditis elegans* (*C. elegans*) has been and remains a valuable model organism to study developmental biology, aging, neurobiology, and genetics. The large body of work on *C. elegans* makes it an ideal candidate to integrate into large-population, whole-animal studies to dissect the complex biological components and their relationships in a given organism. In order to use *C. elegans* in collaborative omics research, a method is needed to generate large populations of animals where a single sample can be split and assayed across diverse platforms for comparative analyses.

Here, a method to culture and collect an abundant mixed-stage *C. elegans* population on a large-scale culture plate (LSCP) and subsequent phenotypic data is presented. This pipeline yields sufficient numbers of animals to collect phenotypic and population data, along with any data needed for omics experiments (*i.e.*, genomics, transcriptomics, proteomics, and metabolomics). In addition, the LSCP method requires minimal manipulation to the animals themselves, less user prep time, provides tight environmental control, and ensures that handling of each sample is consistent throughout the study for overall reproducibility. Lastly, methods to document population size and population distribution of *C. elegans* life stages in a given LSCP are presented.

## Introduction

*C. elegans* is a small free-living nematode that is found throughout the world in a variety of natural habitats [1]. Its relative ease of growth, fast generation time, reproduction system, and transparent body make it a powerful model organism that has been widely studied in developmental biology, aging, neurobiology, and genetics [2, 3]. The copious work on *C. elegans* makes it a prime candidate to use in omics studies to comprehensively link phenotypes with complex biological components and their relationships in a given organism.

In order to use *C. elegans* in collaborative omics research, a method is needed to generate large mixed-stage populations of animals where a single sample can be split and used across diverse platforms and instruments for comparative analyses. Creating a pipeline to generate such a sample requires keen awareness of diet, environment, stress, population structure, and sample handling and collection. Therefore, it is crucial to have standard and reproducible culturing conditions integrated into large-scale pipelines. In *C. elegans* research, two traditional methods are used to culture worms - agar petri dishes and liquid culture [4].

Historically when large quantities of *C. elegans* are needed, they are grown in liquid culture [4]. The steps involved in generating a large population of worms in liquid culture require multiple handling steps that often include bleach synchronization to rupture gravid adult cuticles, releasing embryos to achieve the desired population size. However, when bleach synchronization is used, population growth is dependent on starting census size and thus effects subsequent growth and population numbers. In addition, *C. elegans* strains vary in their cuticle sensitivity, exposure time, and stress response to bleach synchronization making it difficult to assay a large number of strains at a time [5-9].

Additionally, worm growth in liquid culture requires a couple of transfer steps as it is often recommended to grow just one generation of worms before harvesting because overcrowding can easily occur if grown for multiple generations and lead to dauer formation despite the presence of food [10]. Dauer formation occurs through small signaling molecules such as ascarosides, often referred to as “dauer pheromones” [10-13] that are released into liquid media and affect the growth of the population. Furthermore, growing large worm populations in liquid culture leads to excess bacteria accumulation in the culture, creating difficulties when a clean sample is needed for downstream phenotypic assays. Lastly, when a liquid culture becomes contaminated, it is more difficult to maintain as fungal spores or bacterial cells are easily dispersed throughout the media [14].

The other traditional method of growing *C. elegans* is on agar petri dishes. Commercially available petri dishes allow one to easily grow multiple generations of mixed-stage worms without the rapid effects of overcrowding and high dauer formation as seen in liquid cultures. However, a disadvantage to worm growth on traditional agar petri dishes are that the largest commercially available petri dish does not yield large worm populations for an omics study without adding in a bleach synchronization step. In summary, culturing mixed-stage populations of *C. elegans* on agar petri dishes is more suitable for collecting omics data, but we required a method to generate very large population sizes without liquid culturing.

Here, we present a method to culture and collect large mixed-stage *C. elegans* populations on a LSCP. Collecting samples through this pipeline yields enough sample to gather phenotypic and population data, along with any data needed for omics experiments (*i.e.*, genomics, transcriptomics, proteomics, and metabolomics). In addition, the LSCP method requires minimal manipulation of the animals, less user prep time, provides tight environmental control, and ensures that handling of each sample is consistent throughout the study for overall reproducibility.

## Protocol

### 1 Sterilize LSCP and Equipment

NOTE: A variety of vessels can be used as a LSCP. In this protocol, a standard glass baking dish was used. The LSPCs in use had outer dimensions of 35.56 x 20.32 cm, inner dimensions of 27.94 x 17.78 cm, and approximately 4.45 cm deep and came with a fitted lid. Ensure the LSCPs are dishwasher and autoclave safe. Ensure the LSCP lids are dishwasher safe.

NOTE: Throughout each step of this protocol ensure working space is cleaned with 70% ethanol and 10% bleach. If available, treat used areas with UV light for 30 minutes and turn on a HEPA air-filter 30 minutes prior to starting each step.

1.1 Prep glass LSCPs by handwashing, followed by dishwashing, and subsequent autoclaving to ensure glassware is free of contaminants prior to starting the experiment. Store autoclaved LSCPs in a clean dry location until in use.

1.2 Prep LSCP lids by handwashing followed by dishwashing. Store LSCP lids in a clean bin until needed.

1.3 On the day Nematode Growth Media Agarose (NGMA) is prepped, wipe LSCP lids with 10% bleach solution twice, followed by 70% ethanol. Once wiped down with 10% bleach and 70% ethanol, keep LSCP lids in a clean bin in the laminar flow hood where the NGMA will be prepped.

**2 Prepare Nematode Growth Media Agarose (NGMA)**

NOTE: The preparation steps for the NGMA as described here will yield enough material for 2.5 LSCPs. The protocol can be tailored to the needed LSCP batch size in a given experiment.

2.1 Prepare NGMA by combining the following reagents into an autoclaved 2 L Erlenmeyer flask with stir bar on a stir plate: 2.5 g peptone, 3 g NaCl, 7 g agarose, 10 g agar, and 975 mL sterile water [15]. Ensure that the total volume equals 1 L. Tape a foil cap to the flask.

2.2. Autoclave on liquid cycle at 121°C and 21 p.s.i. for 45 minutes.

2.4 Turn on water bath set to 50°C.

2.5 Bring the autoclaved NGMA to the water bath to cool to 50°C.

2.6 Bring 2 L Erlenmeyer flask of NGMA into the hood or cleaned space and set on a stir plate. Use a thermometer to track NGMA temperature.

2.7 After the NGMA has reached 50°C, add the following in the order listed with a sterile disposable pipette inside the hood or cleaned space: 25 mL of 1 M $KH_2PO_4$ (K phosphate buffer), 1 mL of cholesterol (5mg/mL in ethanol), 1 mL of 1 M $CaCl_2$, 1 mL of 1 M $MgSO_4$, 1 mL of nystatin (10mg/mL), and 1 mL of streptomycin (100 mg/mL) [15].

2.8 Pour 400 mL of NGMA into a sterile glass LSCP, approximately 1.3 cm deep, allow the LSCP to solidify on flat surface in the hood and place the autoclaved foil lid back on the LSCP.

2.9 Once agar is set, remove the foil and place a clean air-tight lid onto LSCP and move to 4°C for storage.

2.10 NGMA in LSCPs should be used within 5 days and can be stored at 4°C until in use.

### 3 Generate *E. coli* food for NGMA on LSCP

NOTE: In order to generate a stable food source, batches of HT115 (DE3) *E. coli* have been generated using a small batch averaging concept consistent with the central limit theorem [16]. Details described in this protocol start once *E. coli* has been generated and subsequently frozen for use.

### 4 Bacterial Lawn on NGMA

4.1 Bring NGMA LSCPs out of 4°C to room temperature (RT) for several hours before spreading the bacterial lawn to allow entire dish to reach RT.

4.2 Pull out needed *E. coli* bacterial stock(s) from -80°C to thaw [17].

NOTE: In this protocol, *E. coli* bacterial stocks were grown in a bioreactor. At the end of the culture growth, the culture was diluted 1:50, and the measured $OD_{600}$ was 0.4. Thus, the culture had an effective $OD_{600}$ of 20. Bacteria were pelleted, weighed, and resuspended in K-medium at a concentration of 0.5 g/mL (wet weight), transferred into 2 mL aliquots, and frozen [18].

4.3 Dilute *E. coli* bacterial stock(s) with 2 mL sterile K-medium to achieve 0.5 g *E. coli* in 4 mL per NGMA LSCP.

4.4 Carefully pipette 4 mL of *E. coli* in the middle of the NGMA LSCP.

NOTE: The amount of bacteria used here has been optimized for a LSCP with a dimension of 35.56  x 20.32 cm and to yield a large population of mixed-stage worms. Bacterial volume and concentration can be adjusted to fit experimental needs.

4.5 Use a sterile spreader to spread the bacteria into a rectangle leaving approximately 3.8 cm of room around the edges of the NGMA *E. coli* free.

4.6 Leave the NGMA LSCP with *E. coli* in the hood with the fan on for 1 hour to ensure the *E. coli* suspension fully dries.

4.7 Once the bacterial lawn is dry, push lid on tightly and store at 4°C until used.

**5 Chunk Worms to Reduce Stress and Age Variability Across Samples**

NOTE: The age and health of a given worm can influence fecundity and subsequent population growth time. Ensure worms are maintained in healthy conditions with minimal stress prior to being used in this pipeline. It is assumed stock samples have been created, frozen, and kept at -80°C to reduce genetic drift over time.

NOTE: Chunking is an optimal method to transfer worms from a homozygous strain [4]. If a strain is heterozygous or needs to be maintained by picking and mating, chunking is not advisable. Chunking frequency may need to be optimized depending on worm genotypes used, temperature chosen for growth, and downstream steps.

5.1 Streak worms from a frozen worm stock to a newly seeded 6 cm plate[4]. This plate will serve as the “master chunk” plate.

5.2 After the master chunk plate is full of healthy gravid adults (approximately 3 days) with plenty of *E. coli* lawn still present, follow standard *C. elegans* chunking guidelines as described in WormBook to produce four total chunk plates[4].

NOTE: All chunk plates should be stored in a controlled temperature (CT) room at 20°C unless otherwise specified for growth.

NOTE: If users of this protocol do not have access to a CT room as described here it is recommended to use either (A) a small incubator where the temperature can be controlled or (B) a designated room where environmental conditions can be controlled as much as possible. If neither of these alternate options are available, note that variation in sample growth may be greater.

5.3 Once many gravid adults are observed in the 4th chunk plate, move on to Step 6.

## 6 Spot Bleaching Gravid Adults onto LSCP

NOTE: This bleaching technique is used to eradicate most contaminants and dissolve the cuticle of the hermaphrodites releasing embryos from the adult worm. The bleach solution will soak into the NGMA prior to the embryos hatching.

6.1 Bring LSCP produced in Steps 1-4 out to RT for several hours prior to spot bleaching worms.

6.2 Prepare a 7:2:1 ratio of $ddH_20$:bleach:5 M NaOH. Make this alkaline hypochlorite solution fresh just before use.

NOTE: Use the same stock of bleach and NaOH throughout the duration of a given experiment to avoid bleach batch affects. Bleach used in this protocol was 5-10% Sodium hypochlorite.

6.3 Light a Bunsen burner and flame a worm pick before proceeding. Scoop fresh *E. coli* onto a sterile pick from the edge of the bacterial lawn on the LSCP.

6.4 Pick a single gravid adult from the 4th chunk plate for spot bleaching.

6.5 Pipette 5 μl of the alkaline hypochlorite solution into one corner of the LSCP away from the *E. coli* lawn.

6.6 Place the picked gravid adult into the 5 μl alkaline hypochlorite solution. Tap the nematode to help disrupt the cuticle and release eggs.

6.7 Repeat steps 6.4 – 6.6 four times for a total of five gravid adults placed evenly around the *E. coli* lawn. Pick all five gravid adults from the same 4th chunk plate to ensure nearly genetically isogenic individuals are added to a given sample.

NOTE: Five gravid adults are used to seed each LSCP for the following reasons: (a) A simple, fast, and efficient way to seed many *C. elegans* strains onto LSCPs at one time was needed and (b) to reduce the age differences amongst the gravid adults picked that could lead to growth heterogeneity.

NOTE: Depending on the needs of a given experiment, the number of starting gravid adults on a LSCP can be changed. Altering the number of starting gravid adults on the LSCP will change growth rate and thus time to harvest.

6.8 Place lid back onto the LSCP.

6.9 Repeat steps for all LSCPs.

**7 Worm Growth in Controlled Temperature (CT) Room**

NOTE: The CT room was set to 20°C as most *C. elegans* stocks are maintained at this temperature. Temperature can be adjusted to fit the needs of an experiment.

7.1 Following spot bleaching, place the lid tightly onto the LSCP and place in the CT room set to 20°C with constant airflow and a 12L:12D photoperiod (12 hours light and 12 hours darkness).

7.2 Note time and position sample was placed in CT room.

NOTE: Position within the room should always be documented to record any environmental differences samples could potentially encounter while growing. Once the sample is in the CT room, it should remain in its assigned spot undisturbed. Do not open the lid of the LSCP in the CT Room to decrease the chance of contamination.

7.3 Take LSCP to a microscope, outside of the CT room, to observe population growth and density.

NOTE: Each *C. elegans* strain and sample will vary in its growth, so monitor samples closely.

NOTE: While it is recommended to not disturb the growth of the LSCP while in the CT Room, LSCPs were transported out of the CT Room and lids were opened every 2 days to monitor sample growth. Taking the sealed lids off the LSCPs every 2 days also allows for $O_2$ to flow into the LSCP.

NOTE: While checking population growth on the LSCP, ensure condensation is not accumulating. Significant condensation accumulation was not observed in the worm growth process reported here, however if condensation accumulation occurs take a clean delicate task wiper and remove the condensation to avoid pooling on the agar.

7.4 Once the LSCP has become full of a large population of worms, it is ready to be harvested. The following criteria are used to decide if a LSCP is ready to be collected (7.5 – 7.8):

7.5 The LSCP is (1) full of gravid adult worms, 7.6 (2) reached a large population size (*i.e.*, worms cover the entire surface of the agar),

7.7 (3) does not have many eggs on the surface of the agar (*i.e.*, the maximum number of worms should have hatched), and

7.8 (4) has minimal to no *E. coli* left indicating the worms would starve and generate dauer larvae if left on the plate for an additional two days. If the LSCP meets all four of the criteria described in 7.5-7.8, the sample is ready to be harvested.

NOTE: Although most LSCPs are ready to harvest between 10 to 20 days, depending on strain and sample **(see Figure 2.4)**, check each LSCP frequently upon establishing this protocol to determine normal harvest times.

7.9 Clean gloves and area with 70% ethanol between handling LSCPs to avoid cross contamination between strains.

**8 Harvesting the LSCP Sample**

8.1 Turn on and allow centrifuge to cool to 4°C prior to harvesting samples.

8.2 Prepare three 50 mL conical tubes with M9 solution per LSCP to be harvested.

8.3 Label one 15 mL conical tube per LSCP.

NOTE: All centrifugation steps take place in the 15 mL conical tube, as worms pellet well in these tubes.

8.4 Pour 50 mL of M9 solution (from one 50 mL conical tube in Step 8.3) onto the LSCP surface and swirl around to ensure that M9 covers the entire NGMA surface.

8.5 While M9 sits on the LSCP surface, prime a sterile serological pipette with M9.

NOTE: By priming the sterile serological pipette with M9 this ensures that less worms stick to the inside of the plastic pipette, preventing sample loss.

8.6 Tilt the LSCP so M9 and the worm population gather in one corner of the LSCP.

NOTE: The mixture of M9 solution and worms from the LSCP will be referred to as the “worm suspension” in downstream steps.

8.7 Using a primed serological pipette with an automatic pipettor, pipette worm suspension and place into the original 50 mL conical tube. Once 50 mL of worm suspension is collected, place conical tube on a rocker to disrupt bacteria clumps and debris.

8.8 Repeat steps 8.4 - 8.7 collecting 150 mL of worm suspension per LSCP.

8.9 Transfer 15 mL of worm suspension, from one of the three 50 mL conical tubes, by pouring to the labelled 15 mL conical tube set aside in Step 8.4. Centrifuge the 15 mL conical tube at 884 x G for 1 minute at 4°C. The majority of worms will pellet at the bottom of the tube.

8.10 Aspirate off supernatant ensuring to not disturb the worm pellet.

8.11 Continue adding approximately 13 mL of worm suspension to the same 15 mL conical tube repeating steps 8.10 and 8.11 until all 150 mL of worm suspension are consumed. Invert the tube and disturb the pellet between centrifugations to wash and aspirate off as much bacteria and debris as possible.

NOTE: At this step, the contents from all three 50 mL conical tubes are condensed in a single 15 mL tube.

8.12 Add 10 mL of clean M9 to the 15 mL conical tube and agitate the worm pellet by inverting. Centrifuge the 15 mL conical tube at 884 x G for 1 minute at 4°C. Aspirate off supernatant ensuring to not disturb the worm pellet. Repeat twice.

NOTE: If there is a great amount of debris or bacteria in sample, repeat step 8.12 until the sample is clean.

8.13 Once the sample is clean, add $ddH_20$ to the worm pellet for a total of 10 mL of $ddH_20$ and worms. Agitate the worm pellet by inverting. Move quickly into step 9.1, as worms must remain in $ddH_20$ for five minutes or less to avoid osmotic stress.

NOTE: Suspending worm pellets in $ddH_20$ is the preferred solvent for downstream omics steps. Worms can be suspended in other solvents or buffers if they are compatible with a given experimental workflow.

NOTE: The mixture of $ddH_20$ and worms from step 8.14 are referred to as the "worm sample" in subsequent steps.

**9 Estimate Population Size**

NOTE: Move through Steps 9.1 – 9.7 quickly.

9.1 Prior to pipetting the worm sample, prime pipette tip being used with M9 to avoid worms sticking to the inside of the plastic pipette preventing sample loss and reducing count variation.

9.2 Take a 100 μL aliquot of worm sample and dilute it into 900 μL of M9. Mix well and make a serial dilution (1:10, 1:100, 1:1000). Repeat this step twice to achieve a total of three sets of aliquot replicates.

NOTE: Pipetting worms can cause high variability in sample population counts. Ensure that the worm sample is homogenous prior to pipetting the desired aliquot.

9.3 Set the 15 mL conical tube on a rocker to continue moving culture while aliquots are counted.

9.4 Ensure the worm sample is well mixed and homogenous. Pipette 5 μL from the 1:10 worm sample, dispense it onto a microscopy slide, and count the number of worms. If this number is <~50, then also count the 1:100 and 1:1000 dilutions. If it is more than 50, move to the next serial dilution.

NOTE: If too many worms cannot be accurately counted, use the next serial dilution for counting instead.

9.5 Count each aliquot replicate of each dilution three times. At the end of counting, for most cultures, nine total counts will be documented (*i.e.*, three total counts for each aliquot replicate).

9.6 Average the dilution counts to determine the estimated population size of the worm sample. These dilution counts will determine the volume of worm sample needed to create desired aliquot size for omics steps.

NOTE: In this experiment, aliquots of approximately 200,000 mixed-stage worms were generated. In addition, one aliquot of approximately 50,000 mixed-stage worms was set aside for sorting in a large particle flow cytometer (described in Step 10).

NOTE: As aliquots are being created, ensure the worm sample is homogenous with a mixed-stage population of worms to ensure that each aliquot has approximately the same number of worms.

9.7 Once the worm sample has been split into appropriate aliquots, flash freeze in liquid nitrogen and store the sample at -80°C.

NOTE: Do not freeze the aliquot intended for large particle flow cytometry.

**10 (Optional) Prepping Sample for Large Particle Flow Cytometry**

NOTE: Steps 10, 11, and 12 are the authors' preferred method to record sample growth (*i.e.*, population size and population distribution of *C. elegans* life stages) and determine success of a culture. Users of this protocol can substitute optional Steps 10, 11, and 12 with their own metrics of growth success. Steps 10, 11, and 12 are described here for two reasons; First, so users who have equipment used in Steps 10, 11, and 12 can replicate these steps and secondly, to show validation of this growth method.

NOTE: Step 9 above provides a good estimation of total number of worms to determine aliquot sizes, and step 10 is a more quantitative metric to estimate the number and population distribution of worms in a given sample.

10.1 Bring the aliquot of approximately 50,000 mixed-stage worms (set aside in Step 9.6) up to 10 mL total volume in M9 solution.

10.2 Make a solution composed of 1 mg/mL of *E. coli* and a 1:50 dilution of 0.5 μM red fluorescent microspheres [18].

10.3 Add 200 μL of this solution to the 10 mL of mixed-stage worms in M9 and incubate while rocking for 20 minutes.

10.4 After 20 minutes, centrifuge the 15 mL conical tube at 884 x G for 1 minute at 4°C.

10.5 Aspirate off supernatant ensuring to not disturb the worm pellet.

10.6 Wash worm pellet twice with M9 solution to eliminate excess bacteria and red fluorescent microspheres.

10.7 Add 5 mL of M9 to the worm pellet and ensure pellet looks clean. If pellet is clean, add 5 mL of M9 with 50 mM sodium azide to both straighten and kill the worms for accurate counting and sizing [19].

10.8 Document time and date when sodium azide is added to sample.

10.9 Set sample aside on rocker until needed for large particle flow cytometry.

NOTE: Sodium azide is known to affect nematode physiology (*i.e* body length, metabolism, and thermotolerance). Therefore, it is critical to note the time worms are exposed to sodium azide as many of these physiological affects happen within a matter of minutes [20]. Due to the known physiological effects of sodium azide on worms, this treatment will affect downstream image quality and should be considered.

## 11 (Optional) Documenting Population Distribution and Prepping 384-well Plate for Imaging

NOTE: Step 11 uses a large particle flow cytometer (LPFC). Basic knowledge of a LPFC is assumed in this protocol. Other methods can be substituted to document the growth and population distribution of samples.

NOTE: Steps documented here are for users who plan to use a LPFC in their pipeline [21].

11.1 Clean and prime the LPFC and allow laser(s) to warm for 1 hour prior to sorting samples.

11.2 After the laser has warmed, open the "Histogram" profile and scale to a Time of Flight (TOF) of 2050.

11.3 Add a bar region to the "Histogram" spanning a TOF range of 100. The first bar region covers a TOF of 50-150.

11.4 Continue to create twenty bar regions each spanning a TOF range of 100. These bar regions will span the entire TOF range from 50 – 2050. See **Supplementary Table 2.1** for the exact gated regions to use across the TOF distribution.

11.5 Save this Histogram set up as an "experiment" to use in future LPFC runs.

11.6 Select calibrated 384-well plate or calibrate instrument to a 384-well plate to dispense objects into.

11.7 Once in the calibrated 384-well plate template, set template to dispense 20 gated objects into four wells (four technical replicates of each gated region) for each of the 20 bar regions created during Steps 11.3-4. See **Supplementary Table 2.2** for an example layout of how to dispense worms into the 384-well plate.

11.8 Transfer sample from Step 10.9 into a 50 mL conical tube and add additional M9 solution to achieve approximately 40 mL total volume.

11.9 Start sorting the sample while simultaneously dispensing objects from the sample into the calibrated 384-well plate.

NOTE: Ensure the flow rate of the LPFC is operating between 15-20 objects per second and specify no doubles to be sorted.

11.10. Once entire sample has been sorted and the maximum number of gated regions have been dispensed into the 384-well plate, take the sample off LPFC and clean instrument.

NOTE: When larger TOF regions are reached, it may become challenging to continue to fill the 384-well plate due to low event counts in that TOF region. Fill as many of the gated regions as possible to get the best idea of where *C. elegans* life-stages fall within the LPFC distribution prior to running out of sample.

11.11 Place a sealed film on top of the 384-well plate until imaged.

NOTE: Image plate as quickly as possible after sorting since samples are treated with sodium azide [20].

NOTE: Red fluorescent microspheres can be seen in the LPFC data files to help identify which objects are alive worms, dead worms, dauers, or junk [22].

**12 (Optional) Imaging 384-well Plate**

NOTE: Step 12 uses a plate-reading micro confocal microscope. Basic knowledge of a micro confocal microscope is assumed in this protocol. Other methods can be substituted to document the growth and population distribution of samples.

12.1 Using a plate-reading micro confocal microscope with a 20X lens, take images of each well in the 384-well plate with the following settings (12.2-12.6):

12.2 (A) Open the "Objective and Camera" tab and set to "10x Plan ApoLambda" mode.

12. 3 (B) Open the "Camera Binning" tab and set to "2".

12.4 (C) Open the "Sites to Visit on Plate" tab and set to "4" sites per well and "overlap sites 10%" to later stitch together images.

12.5 (D) Open "Wavelength" tab and set to "Brightfield 1".

12.6 (E) Open "Illumination" tab and set to "Transmitted light, bright sample".

12.7 Place the 384-well plate in the microscope and set the "Z Stack" to "Calculate Offset" and find the proper focal plane for the samples in the 384-well plate.

12.8 Run the 384-well plate on the micro confocal microscope collecting four images per well.

12.9 Montage the four images together to create one image per well.

## Representative Results

**Growth of *C. elegans* using the LSCP method yields an average of approximately 2.4 million mixed-stage worms per sample over 12.2 days.** Growth of *C. elegans* using the LSCP method enables users to generate large mixed-stage populations of *C. elegans* with little handling and manipulation of the animals, which is ideal for large-scale omics studies **(Figure 2.1)**. Once a LSCP has become full of adult worms, reached a large population size, and has minimal bacteria left, users can harvest and estimate the population size. This point can also serve as a quality control by evaluating whether the population is sufficient to use in an omics pipeline (**Figure 2.2**). Population dynamics are dependent on the strain itself, behavior of the strain (*i.e.*, burrowing strains tended to have lower worm recovery), and growth success (*i.e.*, contamination). The LSCP method was tested on 15 strains of *C. elegans* containing a mixture of *Caenorhabditis* Genetics Center (CGC) mutants and *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) wild strains [23], strain genotypes are described in **Supplementary Table 2.3**.

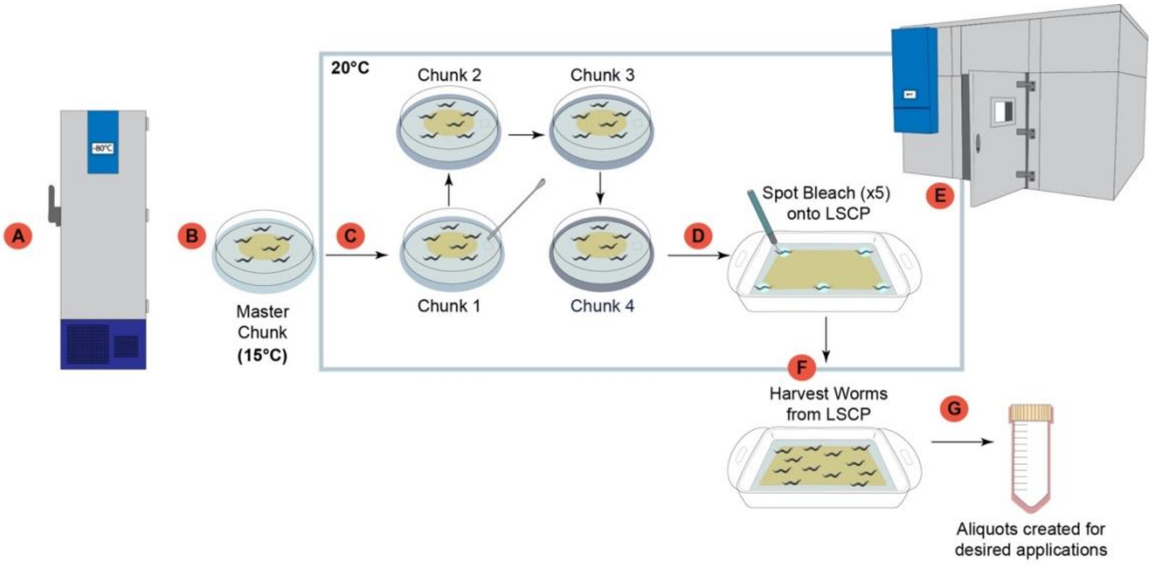


**Figure 2.1: Overview of the LSCP worm growth pipeline. (A)** Once received in the lab, all strains are prepared and frozen for long-term storage at -80°C[2]. **(B)** A "master chunk" plate is prepared from a frozen worm stock and stored at 15°C to be used for *no longer than* one month. **(C)** Each sample goes through four successive chunking steps to reduce generational stress prior to growing on the LSCP. **(D)** Five individual gravid adults are picked from the "chunk 4" 6 cm plate in Step (D) and spot bleached on five given areas of the LSCP. **(E)** The LSCP is placed in a Controlled Temperature room to grow at 20°C until it has become full of adult worms, reached a large population size, and has minimal bacteria left. **(F)** The worm population is harvested and collected for downstream steps. **(G)** Aliquots are created from LSCP and are flash frozen for downstream desired applications.

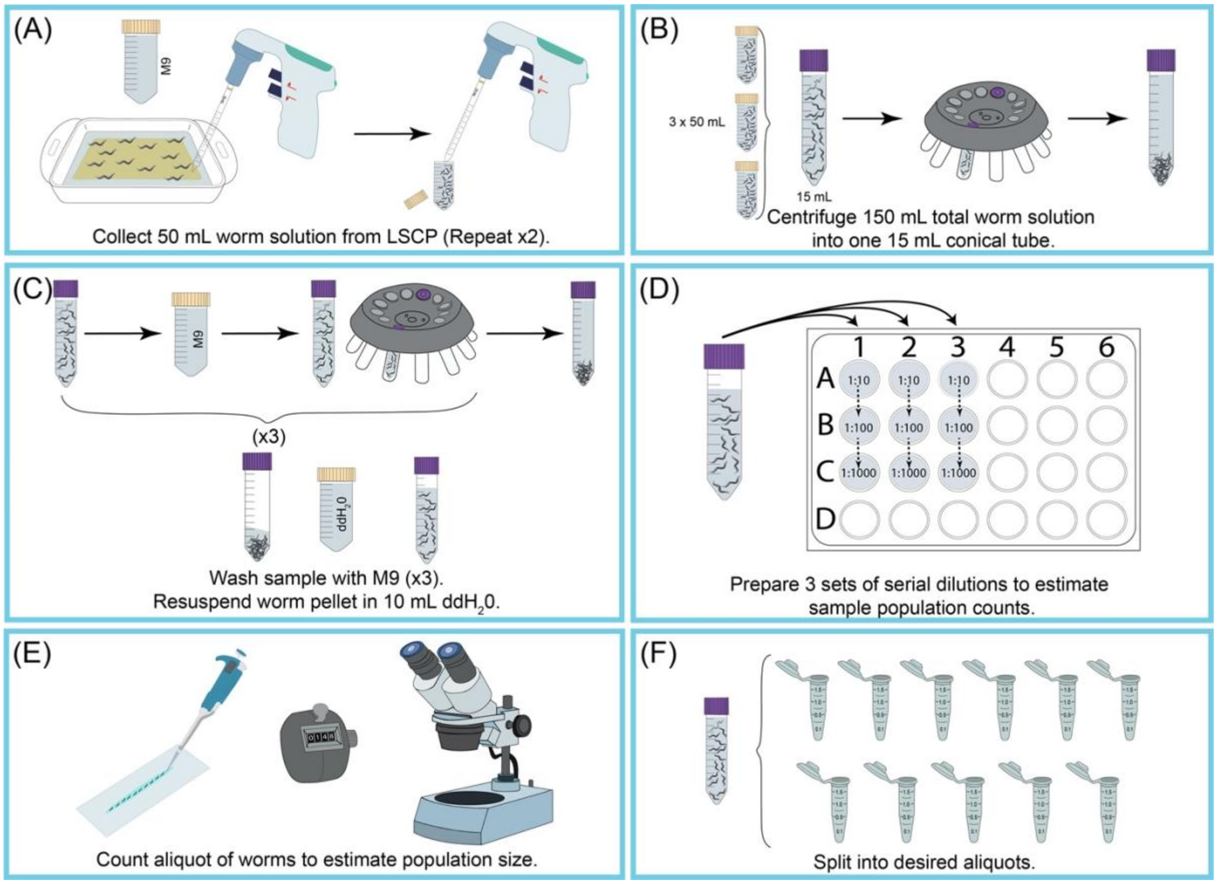

(A)
M9
Collect 50 mL worm solution from LSCP (Repeat x2).
(B)
3 x 50 mL
15 mL
Centrifuge 150 mL total worm solution into one 15 mL conical tube.
(C)
M9
(x3)
ddH$_2$0
Wash sample with M9 (x3). Resuspend worm pellet in 10 mL ddH$_2$0.
(D)
1 2 3 4 5 6
A B C D
1:10
1:100
1:1000
Prepare 3 sets of serial dilutions to estimate sample population counts.
(E)
0148
Count aliquot of worms to estimate population size.
(F)
Split into desired aliquots.

**Figure 2.2: Overview of LSCP harvesting and estimating population size. (A)** Use 50 mL of M9 to wash worms off the NGMA surface and pipette worm suspension into a 50 mL conical tube. Repeat twice. **(B)** Pour 15 mL of worm suspension into a new 15 mL conical tube. Pellet worms by centrifuging. Aspirate off M9 + debris without disturbing worm pellet. Repeat until all 150 mL of worm suspension are collected. **(C)** Wash and centrifuge the worm pellet three times with M9 to eliminate remaining debris. Once the sample is clean, resuspend worm pellet in 10 mL of ddH$_2$0. **(D)** Create a serial dilution of sample to estimate worm population size. Choose the dilution that allows you to count worms most accurately. The dilution used may change depending on the population size of the LSCP. Once a dilution is chosen, ensure you count worms from all three

aliquot replicates of that dilution. **(E)** Aliquot the sample onto a clean slide and count worms present under a dissecting microscope. **(F)** Split sample into appropriate-sized aliquots.

The LSCP method yielded population sizes from approximately 94,500 to 9,290,000. The mean population size within the reference strain, PD1074, and across strains was approximately 2.4 million worms **(Figure 2.3)**. No significant differences were found in estimated population sizes between *C. elegans* strains over the course of an average of 12.2 LSCP growth days **(Figure 2.4)**. PD1074 LSCPs took between 10 – 14 days to grow to a full mixed-stage population. The mean growth time across PD1074 was 10 days. The slowest growing strain grew for a maximum of 20 days, and the fastest growing strain grew for a minimum of 10 days (**Figure 2.4**).

Therefore, using this LSCP method, users can easily integrate new strains of interest into a study with little knowledge of developmental timing and background expertise. Note that strains and phenotypes that have to be maintained by picking, have fecundity defects, are heterozygous, or have growth defects may not work well in this pipeline.

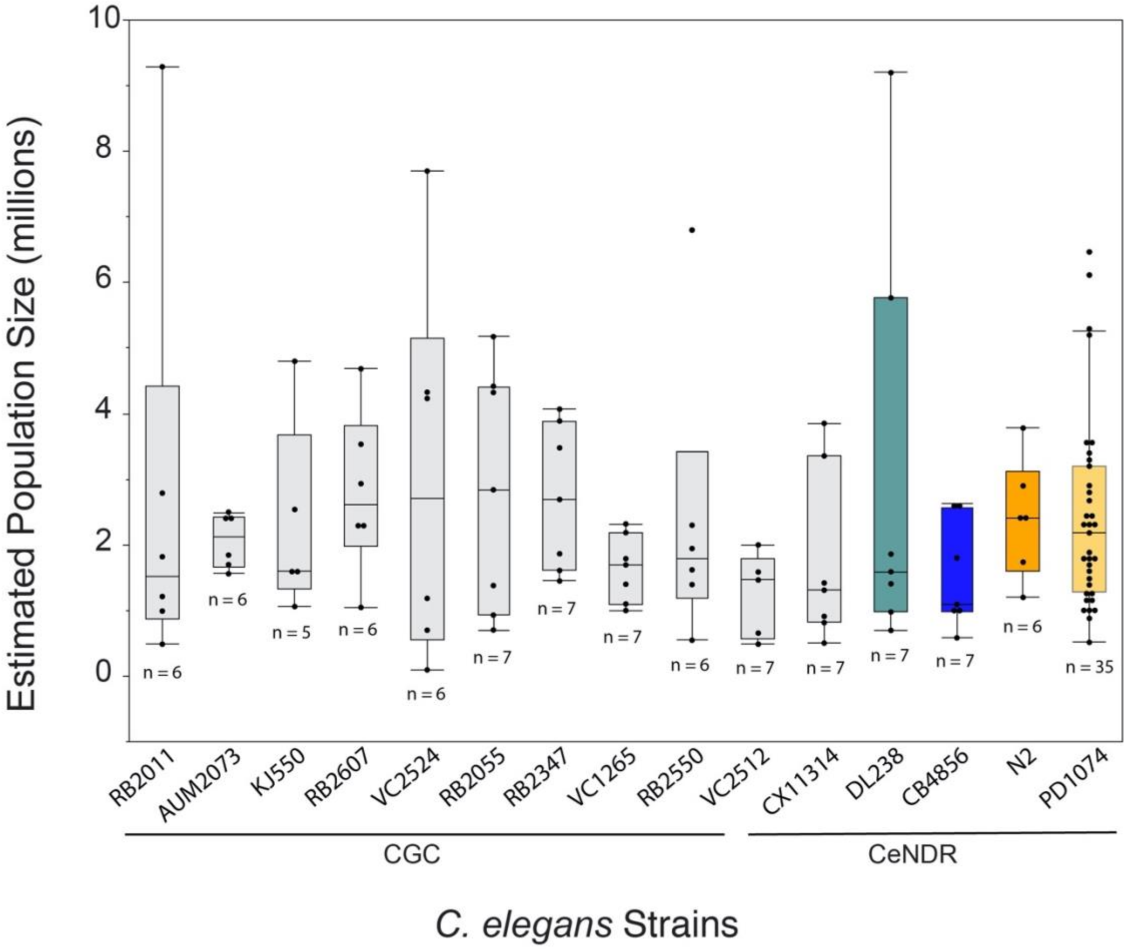


**Figure 2.3: LSCP method generates an average population of 2.4 million mixed-stage worms.** The LSCP yields population sizes in the smallest population growths at around 94,500 and at the biggest population growths at around 9,290,000. The mean population size across all strains was 2.4 million worms. Bars underneath *C. elegans* strain names indicate whether a strain is a CGC mutant or CeNDR natural isolate. LSCP sample size is displayed for each strain. Comparisons for all pairs using Tukey's HSD Test were performed. No significant differences were observed between estimated population sizes across *C. elegans* strains (F(14,108) = 0.7, *p* = 0.77). Colored bars indicate standard color displays for respective *C. elegans* strain representation.

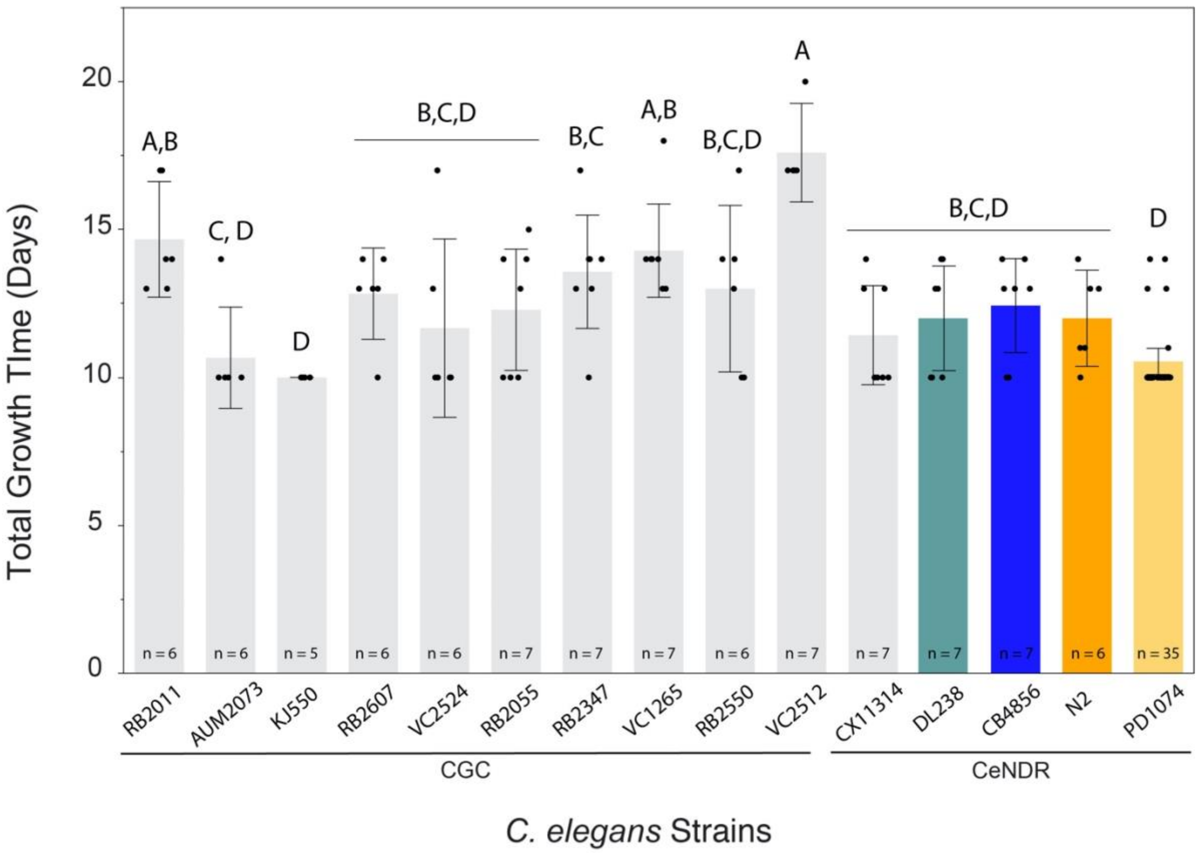


**Figure 2.4: LSCP method generates large mixed-stage populations of worms in 10 – 20 days.** *C. elegans* LSCP grew until the sample was full of adult worms, reached a large population size, and had minimal bacterial left. LSCPs took between 10 – 20 days to grow to a full mixed-stage population, depending on the strain. The mean growth time across the strains was 12.2 days. LSCP sample size is displayed for each strain. Each error bar was constructed using 1 standard deviation from the mean. Levels not connected by same letter are significantly different. Comparisons for all pairs using Tukey's HSD Test. A significant difference was found in the amount of growth time on LSCP needed across *C. elegans* strains (F(14,108) = 8.8, $p$ < 0.0001*). Colored bars indicate standard color displays for respective *C. elegans* strain representation.

**Large particle flow cytometry and imaging of samples allows users to document population distribution.** A wide variety of platforms can be used to measure successful population growth. For reproducible omics measurements, it is important to grow consistent cultures. The metrics of culture reproducibility are number of worms and a consistent size distribution for a given strain. We show that the sample distribution for the reference strain, PD1074 – a variant of the original N2 Bristol strain, using the LPFC [18, 21] and micro confocal microscope as proxies for growth success. As worms were measured from the L1 stage through gravid adult on the LPFC distribution **(Figure 2.5),** subsequent imaging **(Figure 2.6)**, and the variation in the population distribution across samples **(Figure 2.7),** we can see that this pipeline is generating a mixed-stage population of *C. elegans*.

To take a closer look at the population distribution of our mixed-stage samples, we looked at the distribution of 35 PD1074 LSCPs by looking at the percent of worms that fall within each region across the entire Time of Flight (TOF) (*i.e.*, body length) distribution **(Figure 2.7A and B)**.

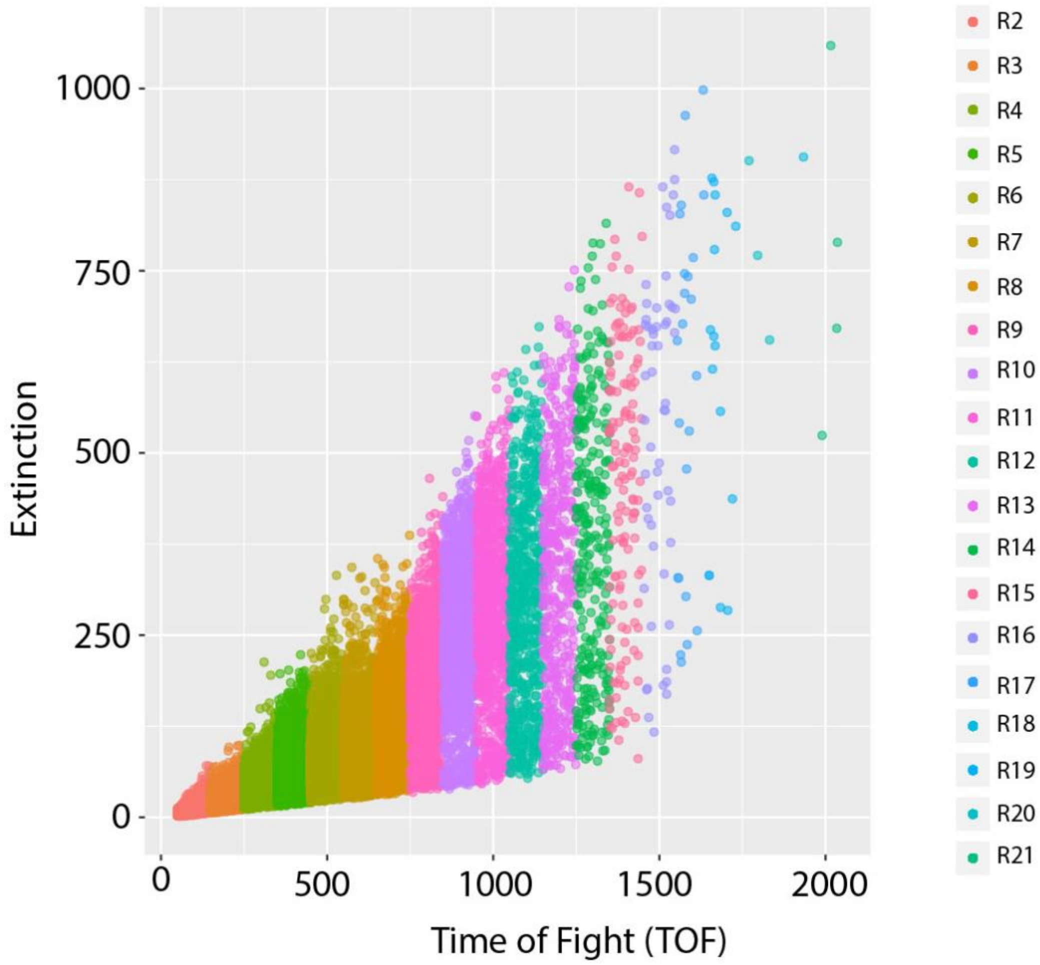


**Figure 2.5: Mixed population and growth measurement of the wild-type reference strain, PD1074.** A representative LPFC distribution of one LSCP growth of the wild-type reference strain, a variant of the original N2 Bristol strain, (PD1074) documents the size distribution and event counts of a mixed-stage population. The x-axis displays the length (Time of Flight, TOF) of the worms sorted. The y-axis displays the optical density (optical extinction, EXT) of the worms sorted. Each data point is a worm that was documented in the sample. Each TOF region that was used for image analysis is displayed in a different color. Twenty TOF regions were created (R2 –

R21) ranging from a TOF of 50 to 2050. Details on each TOF region can be found in Supplementary Table 2.1.

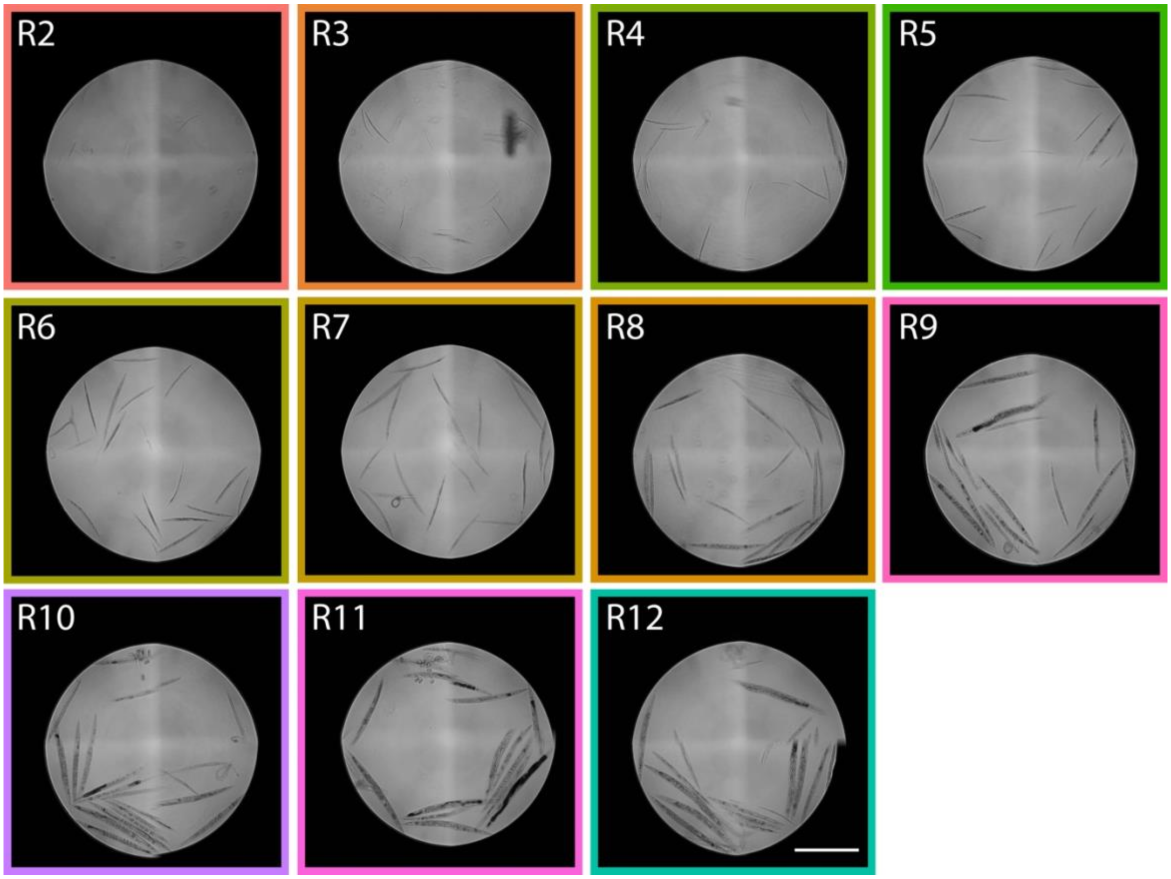


**Figure 2.6: Images of worms sorted from TOF regions ranging from R2 – R12 show the PD1074 LPFC distribution.** In region R2, L1 worms can be identified and in region R9 predominately gravid adults are identified spanning the two developmental larval extremes giving us approximate regions within the flow cytometer distribution of where stages are expected in the distribution. Scale bar represents 1 mm. Representative images were taken from the LPFC distribution displayed in Figure 2.5, and the colored boxes correspond to regions from Figure 2.5.

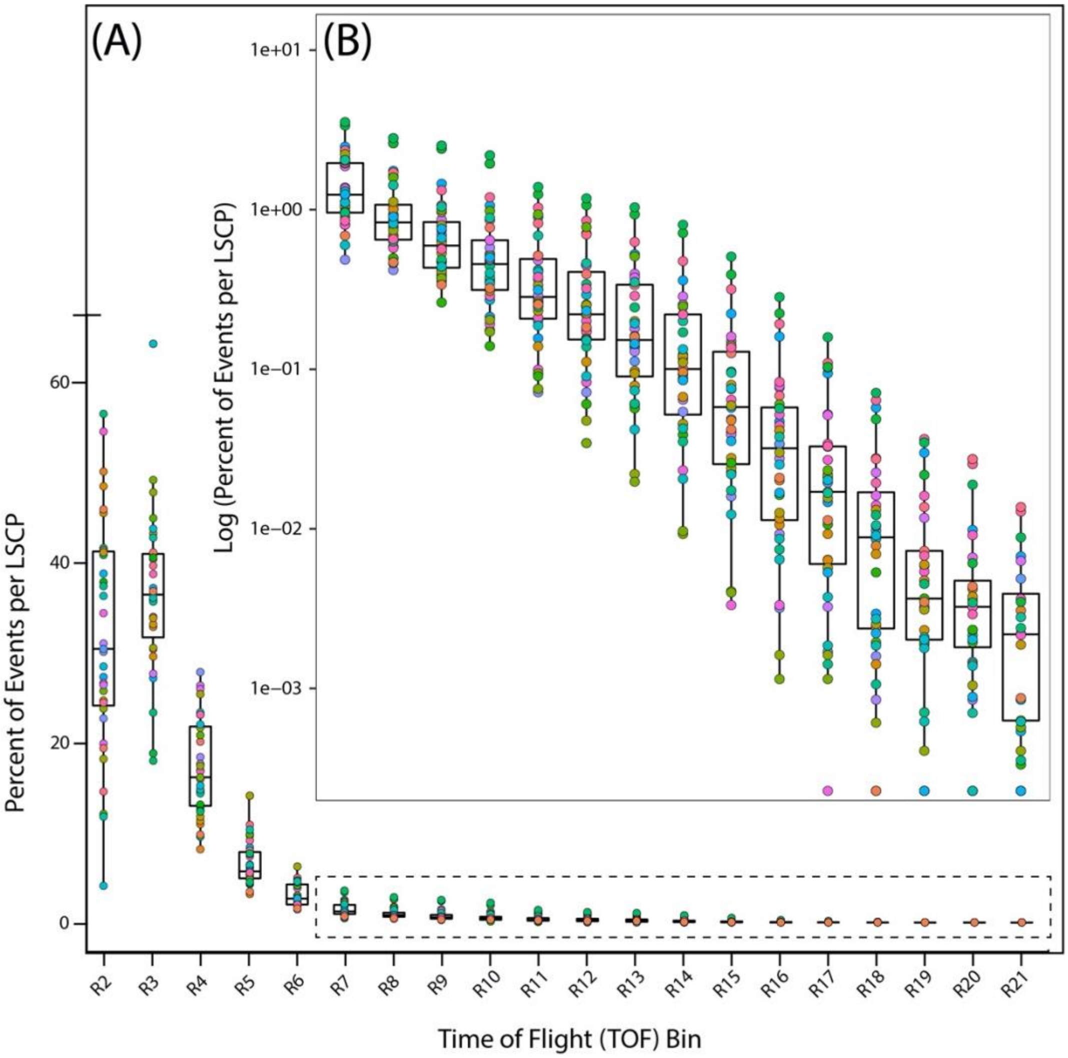


**Figure 2.7: Population distribution across time of flight (TOF) regions in the wild-type reference strain, PD1074.** Distribution of worms across the entire TOF region showing in which regions worms are found. Each PD1074 LSCP is represented as an individual color. **(A)** The x-axis shows the twenty TOF regions (R2 – R21) observed and counted for the LSCP displaying the entire size distribution. The y-axis shows the percent of worms from a given LSCP that had a body size that fell into a given TOF region. **(B)** As a smaller fraction of the worm population falls between the R7-R21 regions, the log of the percent of worms that fell within each region was taken

to display the population distribution. The x-axis displays the R7-R21 TOF regions. The y-axis displays the log of the percent of worms from a given LSCP that had a body size that fell into a given TOF region.

## Discussion

Contamination by mold, fungi, or other bacterial sources can occur at any step in the LSCP method, so handle samples with care. By growing the LSCP in a controlled setting, the user can more easily track the growth of the sample and document potential contamination. If the surface of the LSCP becomes contaminated, either cut out the contamination when possible and let the sample continue to grow or discard the sample if the contamination is not possible to control. It is imperative to address contamination quickly to reduce unwanted growth and to ensure it is not outcompeting worms for resources.

This method is meant for those who want to grow large-scale mixed-population cultures of *C. elegans*. Although it may be possible to grow synchronized populations of worms on the LSCP as done on commercially available petri dishes and in liquid culture, the authors have not tested this option. Additionally, if users wish to grow more than approximately 2.4 million worms on average in a given sample, a different method is recommended. Growth success is dependent on the strain being processed in the pipeline. The authors were able to successfully grow populations of approximately 2.4 million worms in at least five biological replicates of 15 *C. elegans* strains, indicating that the method is robust.

This method allows the user to harvest large populations of worms with all life cycle stages present. With current methods available, collecting large-scale samples of *C. elegans* requires bleach synchronization to obtain the number of worms desired for downstream work. Given this approach, one can now grow as many worms as previously possible in fermenters or large-scale liquid cultures without the difficulties associated with bleach synchronizing and multiple handling steps. Our protocol allows one to target strains of interest efficiently, use minimal handling time in growing the sample itself, and isolate stages of worms or the population as needed in downstream pipelines.

## Future Applications

The authors are using samples grown in the method described here to identify unknown metabolites in various strains of *C. elegans* via Liquid Chromatography – Mass Spectrometry, NMR spectroscopy, and RNA sequencing. The authors plan to continue to use this method for growth of samples in this pipeline with a variety of *C. elegans* strains as new strains of interest can be easily processed using this pipeline.

## Acknowledgements

We thank members of the Edison Lab for helpful discussions and feedback on this manuscript; in particular, B.M. Garcia. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440), and CeNDR, which is funded by NSF Living Collections CSBR 1930382. This work was supported by a grant from the NIH (U2CES030167).

## References

1. Felix, M.A. & Braendle, C. The natural history of Caenorhabditis elegans. *Curr Biol* **20**, R965-969 (2010).
2. Corsi, A.K., Wightman, B. & Chalfie, M. A Transparent Window into Biology: A Primer on Caenorhabditis elegans. *Genetics* **200**, 387-407 (2015).
3. Brenner, S. The genetics of behaviour. *Br Med Bull* **29**, 269-271 (1973).

4. Stiernagle, T. Maintenance of C. elegans. *WormBook*, 1-11 (2006).

5. Xiong, H., Pears, C. & Woollard, A. An enhanced C. elegans based platform for toxicity assessment. *Sci Rep* **7**, 9839 (2017).

6. Loer, C.M. et al. Cuticle integrity and biogenic amine synthesis in Caenorhabditis elegans require the cofactor tetrahydrobiopterin (BH4). *Genetics* **200**, 237-253 (2015).

7. Li, Y. & Paik, Y.K. A potential role for fatty acid biosynthesis genes during molting and cuticle formation in Caenorhabditis elegans. *BMB Rep* **44**, 285-290 (2011).

8. Meli, V.S., Osuna, B., Ruvkun, G. & Frand, A.R. MLT-10 defines a family of DUF644 and proline-rich repeat proteins involved in the molting cycle of Caenorhabditis elegans. *Mol Biol Cell* **21**, 1648-1661 (2010).

9. Fritz, J.A. & Behm, C.A. CUTI-1: A novel tetraspan protein involved in C. elegans CUTicle formation and epithelial integrity. *PLoS One* **4**, e5117 (2009).

10. Golden, J.W. & Riddle, D.L. A pheromone influences larval development in the nematode Caenorhabditis elegans. *Science* **218**, 578-580 (1982).

11. Jeong, P.Y. et al. Chemical structure and biological activity of the Caenorhabditis elegans dauer-inducing pheromone. *Nature* **433**, 541-545 (2005).

12. Srinivasan, J. et al. A blend of small molecules regulates both mating and development in Caenorhabditis elegans. *Nature* **454**, 1115-1118 (2008).

13. Kaplan, F. et al. Ascaroside expression in Caenorhabditis elegans is strongly dependent on diet and developmental stage. *PLoS One* **6**, e17804 (2011).

14. Celen, I., Doh, J.H. & Sabanayagam, C.R. Effects of liquid cultivation on gene expression and phenotype of C. elegans. *BMC Genomics* **19**, 562 (2018).

15. Andersen, E.C., Bloom, J.S., Gerke, J.P. & Kruglyak, L. A variant in the neuropeptide receptor npr-1 is a major determinant of Caenorhabditis elegans growth and physiology. *PLoS Genet* **10**, e1004156 (2014).

16. Rosenblatt, M. A Central Limit Theorem and a Strong Mixing Condition. *Proc Natl Acad Sci U S A* **42**, 43-47 (1956).

17. Boyd, W.A., Smith, M.V. & Freedman, J.H. Caenorhabditis elegans as a model in developmental toxicology. *Methods Mol Biol* **889**, 15-24 (2012).

18. Nika, L., Gibson, T., Konkus, R. & Karp, X. Fluorescent Beads Are a Versatile Tool for Staging Caenorhabditis elegans in Different Life Histories. *G3 (Bethesda)* **6**, 1923-1933 (2016).

19. Denver, D.R. et al. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nat Genet* **37**, 544-548 (2005).

20. Massie, M.R., Lapoczka, E.M., Boggs, K.D., Stine, K.E. & White, G.E. Exposure to the metabolic inhibitor sodium azide induces stress protein expression and thermotolerance in the nematode Caenorhabditis elegans. *Cell Stress Chaperones* **8**, 1-7 (2003).

21. Pulak, R. Techniques for analysis, sorting, and dispensing of C. elegans on the COPAS flow-sorting system. *Methods Mol Biol* **351**, 275-286 (2006).

22. Lee, D. et al. Selection and gene flow shape niche-associated variation in pheromone response. *Nat Ecol Evol* **3**, 1455-1463 (2019).

23. Cook, D.E., Zdraljevic, S., Roberts, J.P. & Andersen, E.C. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res* **45**, D650-D657 (2017).

# CHAPTER 3

# AN ANCHORED EXPERIMENTAL DESIGN AND META-ANALYSIS APPROACH TO ADDRESS BATCH EFFECTS IN LARGE-SCLAE METABOLOMICS [1]

[1] Shaver, A.O., Garcia, B.M., Gouveia, G.J, Morse, A.M., Liu, Z., Asef, C.K., Borges, R.M., Leach III, F.E., Andersen, E.C., Amster, I.J., Fernández, F.M., Edison, A.S., and McIntyre, L.M. An anchored experimental design and meta-analysis approach to address batch effects in large-scale metabolomics. Submitted to Nature Communications (April 2022). DOI: https://doi.org/10.1101/2022.03.25.485859

Chapter 3 is adapted from Shaver, A.O., Garcia, B.M., Gouveia, G.J, Morse, A.M., Liu, Z., Asef, C.K., Borges, R.M., Leach III, F.E., Andersen, E.C., Amster, I.J., Fernández, F.M., Edison, A.S., and McIntyre, L.M. An anchored experimental design and meta-analysis approach to address batch effects in large-scale metabolomics. Submitted to Nature Communications in April 2022. A pre-print is available on bioRxiv at DOI: https://doi.org/10.1101/2022.03.25.485859. My roles on this project were to (i) generate *C. elegans* samples, (ii) configure the study design, (iii) prepare, extract, and acquire the NMR data, (iv) aid in NMR data processing, (v) aid in quality control assessments of LC-MS and NMR data, (vi) analyzed the meta-analysis data, (vii) created 12 figures, and (viii) wrote the manuscript. I am sharing co-authorship on this manuscript with Brianna M. Garcia, where her roles were to: (i) lyophilize all LC-MS samples, (ii) aid in NMR data processing, (iii) perform LC-MS data processing, (iv) aid in quality control assessments of LC-MS and NMR data, (v) analyzed the meta-analysis data, (vi) created 3 figures, and (vii) wrote the manuscript. Collaborator's roles were as follows: Goncalo J. Gouveia generated the stable *E. coli* food, through an iterative batch averaging method (IBAT), to feed the worms. He also generated the *C. elegans* IBAT samples. The IBAT process was published and is available at https://pubs.acs.org/doi/full/10.1021/acs.analchem.1c01294. Goncalo aided in NMR analysis and generated Figure 3.5. Alison M. Morse generated workflows to perform analysis in HiPerGator and aided in running quality control and meta-analysis workflows. Zihao Liu performed preliminary comparisons between the linear models analysis versus meta-analysis. Carter K. Asef prepared, extracted, and aquired the LC-MS data. Ricardo M. Borges provided advice and input on the LC-MS data processing methods and manuscript edits. Franklin E. Leach III provided useful feedback, troubleshooting ideas, and manuscript edits. Erik C. Andersen provided *C. elegans*

strains from the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) along useful feedback, troubleshooting ideas, and manuscript edits. I. Jonathan Amster provided useful feedback and manuscript edits. Facundo M. Fernández provided useful intellectual contributions and manuscript feedback. Arthur S. Edison provided useful intellectual contributions, feedback, and financial support. Lauren M. McIntyre helped generate workflows to perform analysis in HiPerGator, aided in running quality control and meta-analysis workflows, provided intellectual contributions and support, and manuscript edits. Research reported in this manuscript was supported by the National Institutes of Health Award Number U2CES030167.

## Abstract

Large-scale untargeted metabolomics studies suffer from batch effects and instrument variability issues, making comparisons within and across studies difficult. To identify stable features, we used a set of anchors: a reference strain during sample collection, blocks during data collection, and appropriate controls. Anchor samples allowed for the identification of stable features. Five datasets from three studies of *Caenorhabditis elegans* were collected by nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography-mass spectrometry (LC-MS). By exploiting variation among anchor samples, we found 34% and 14% of features to be significant in LC-MS and NMR, respectively. Comparing differences between mutants and natural strains (NS) and independently assaying NS for variation in the same features, we found that 20-50% of features varied in a mutant and among a set of genetically diverse NS. Features that varied in response to a mutation and showed variation in NS are excellent targets for compound identification.

## Introduction

Untargeted metabolomics studies compare the variation in small molecules caused by genetic perturbations, treatments, and environmental differences[1]. Metabolomics is a powerful tool in biomarker discovery and holds great promise for precision medicine[2-4]. Targeted metabolomics is common in studies exploring human health questions that range from aging[5, 6] to complex diseases[7-12]. An advantage of untargeted metabolomics for these questions is the ability to reach beyond sets of well studied compounds to explore differences in an unbiased way[13]. Despite the attractiveness of an unbiased survey, untargeted metabolomics has well known challenges. In particular, the collection of highly variable biological material in a reproducible manner across batches remains an unsolved analytical challenge. These effects make the identification of differential compounds and comparisons of their abundances across datasets challenging. Chemical annotation of compounds requires considerable time and labor[14]; therefore, it is essential to prioritize spectral features based on their likely biological significance. Given this bottleneck, it is essential to find novel ways to prioritize spectral features and overcome intractable challenges such as matrix effects, instrument drift, and batch variation[15-18].

Batch effects across experiments are a problem in untargeted metabolomics and a barrier to adopting these methods[19]. Although normalization strategies are improving[15, 16]; non-linear effects[20], sample variation, the inability to separate environmental variance, and analytical artifacts[17] still pose ongoing challenges. While different approaches to sample-based and data-based normalization have been described, such as total protein content, total ion count (TIC), and pooled QCs[12, 21], reproducibility and heteroscedasticity (variability across variables) issues remain problematic[22-26].

Our goal, and that of many studies, is to compare groups across large numbers of samples[27-30]. As sample size increases, challenges associated with variation must be accounted for appropriately. In metabolomics studies, variation in pre-analytical sample collection (growth), analytical sample preparation (extraction), and data collection (instrument)[31] can be confounded (**Figure 3.1**). Identification of shared spectral features using biological reference materials (BRM) is a successful strategy[31, 32] that has proven essential in large-scale studies[32-34]. Implementation of BRM controls for instrument variation can estimate and normalize extraction variation[16, 18, 31]. However, the challenge of inter-sample variation remains. Metabolites may only be present in some samples or some batches. This inter-sample variation in metabolites present within a study may be due to genetic variation among individuals or variation in environmental conditions. In both liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR) spectroscopy, ambiguity in whether features are generated by genetic or environmental factors coupled with batch effects and challenges in peak picking algorithms present obstacles to apply untargeted metabolomics to broader studies[17, 35].

Although tools to handle extraction and instrumentation variation exist, they can still be limited in their utility in large studies for samples with complex matrices[33, 35, 36]. Here, we use the model system *Caenorhabditis elegans* to demonstrate how anchor samples, defined as replicates of the reference strain that augment the experimental design, can be combined with known control strategies to disentangle heteroscedasticity in order to focus on genetic variation in spectral features without compound identification[37-39].

*C. elegans* is a model organism ideally suited to study conserved small molecules in metabolism[40-42]. The worm's short life cycle, self-fertilization of homozygous hermaphroditic individuals, ease of cultivation, and ability to propagate large numbers of animals[43] are ideal for large-scale studies[42, 44-46]. These traits allow one to (i) develop, test, and validate approaches to identify stable spectral features, (ii) demonstrate the feasibility of large-scale biochemical pathway analyses with genetic mutants, and (iii) focus on spectral features likely to reveal essential components of metabolic pathways by comparing features that vary in large and small genetic perturbations.

We designed three *C. elegans* studies to link metabolic variation across mutants and wild strains. The first and second studies comprised central metabolism (CM) mutants and UDP-glycosyltransferases (UGT) mutants as examples of primary and secondary metabolism, respectively. CM mutants have been used in studies showing that diagnostic changes can be associated with human disease[47, 48]. UGTs are an evolutionarily diverse class of Phase 2 enzymes involved in detoxification[49, 50]. Although UGTs are vital to internal detoxification across species, the functions of UGTs have not been well described due to complex regulatory factors[49-52]. The third study comprises genetically diverse natural strains (NS) from a broad geographic base, used to describe natural variation in the metabolome of *C.* elegans[53]. We included N2, a widely used laboratory-adapted strain, and three strains selected from natural habitats worldwide[54].

Collectively, CM and UGT mutants, and NS, allow us to (i) identify spectral features that vary between genetic perturbations, (ii) compare the same spectral features across all three studies without compound identification, and (iii) plan future experiments that can be directly compared to these studies. The anchor design proposed here is straightforward to execute in model systems. An anchor design is an experimental design that includes a reference strain alongside every test strain in an experiment during growth and collection, augmenting the design. This technique is common in agricultural studies as it was first proposed to account for the rampant variation among field plots and the need to compare large numbers of genotypes across heterogeneous environments[38, 39]. The inclusion of anchor samples during data acquisition prioritizes the focus on stable features across a wide range of environmental conditions. Significant spectral features that vary across studies are of interest and more likely to elucidate reactions in biochemical pathways.

## Results

Here, we provide a method to identify spectral features and a straightforward meta-analytic approach that includes three studies dispersed across six acquisition batches. This demonstration is comprised of 104 independent samples to produce five analytical datasets (3 LC-MS and 2 NMR, two complementary technologies commonly used in untargeted metabolomics). This method (**Figures 3.1 and S3.1**) showcases how the use of anchor samples, strict quality control and assurance (QA/QC), and peak picking strategies identifies stable spectral features that reflect the genotype rather than experimental or environmental variation.

**Anchor samples are pooled to identify stable spectral features**

Stable spectral features are identified across three diverse studies of *C. elegans* strains. Our first study comprised of CM mutants (n = 5) identifies spectral features involved in central metabolism. The second study, UGT mutants (n = 5), identifies spectral features affected by Phase 2 enzymes involved in the detoxification system. The third study, NS (n = 4), assesses natural genetic variation. Collectively, these studies represent the common genetic paradigms of general interest to metabolomics and genetic researchers (See **Table S3.1** for full strain details).

Due to stringent QA/QC methods, filtering, and peak picking, datasets were reduced to only include stable spectral features detected in the anchor strain. Thus, the number of stable spectral features detected across LC-MS and NMR fractions are as follows: 3953 in reverse phase (RP) LC-MS positive, 377 in RP LC-MS negative, 199 in hydrophilic interaction liquid chromatography (HILIC) LC-MS positive, 585 in NMR polar, and 487 in NMR non-polar. An instrument failure occurred during the collection of the HILIC negative data preventing the use of that dataset (see Methods).

LC-MS spectral feature identification is challenged by inconsistencies in the presence of chromatographic peaks across replicates, retention time drift complicating peak alignment, non-linear batch effects, and algorithmic limitations in estimating peak abundances in complex spectra[26, 55-58]. Including multiple anchor samples and pooled anchor samples in each batch can mitigate these issues. We collected samples from genotype PD1074 in every block during the large-scale culture plate (LSCP) growth process[43]. These samples anchor the three studies and enable inter-study comparisons[37, 39].

PD1074 LSCPs are genetically identical, leading to the expectation that the spectral features present in each biological replicate are a result of the strain's genetic composition as they are stable across an extensive range of growth conditions (samples collected over six months) (**Figure S3.2**). Individual PD1074 anchor samples are extracted separately and pooled to average the growth and extraction variance (repeat extractions, n=20 for NMR, n =18 for LC-MS). During data acquisition, we included an anchor pool in each batch as one of several controls (**Figure 3.1C**). Comparing the anchor pools for each batch (n=12) across all batches (n=6) enables the identification of spectral features present across instrument runs over months. Further, the selection of features present in all anchor samples ensures stable features across a range of environmental conditions (**Figure 3.1**). Iterative batch average method (IBAT) controls in the NMR study combined with the anchor PD1074 samples and the anchor pools enabled us to estimate the relative contribution of growth (~60%), extraction (~40%), and instrument variance (negligible)[31]. By including anchor samples during peak picking (see Methods), we found that 97% of the selected features are present in the independently collected IBAT controls in the NMR experiment, validating this approach for feature selection.

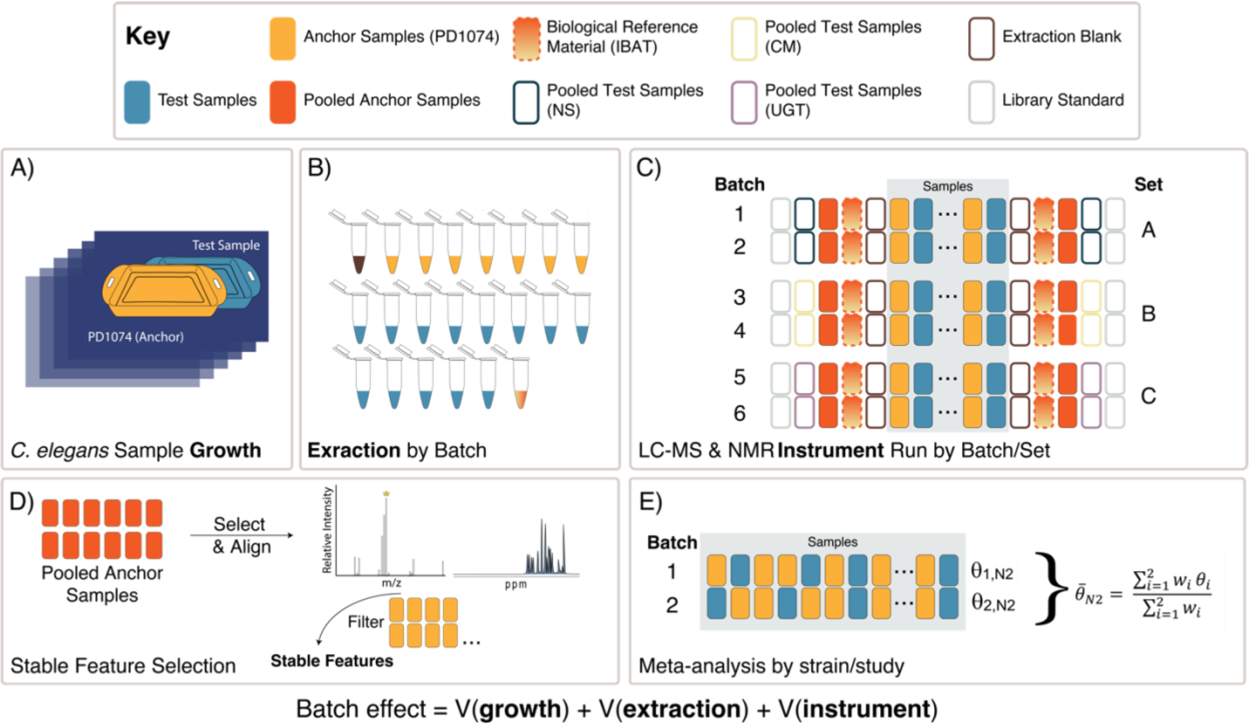


**Figure 3.1. Experimental design overview. (A)** Each *C. elegans* LSCP is grown and harvested with at least one PD1074 anchor control (sample growth variation captured) (**Figure S3.2**). **(B)** Multiple independent PD1074 anchor samples and test strains (NS, CM mutants, or UGTs), IBAT references, and blanks are included in each batch for LC-MS or NMR (batch preparation variation captured). **(C)** Each test strain is collected in two sequential batches. A total of six batches in three sets with instrument controls, library standards, replicate injections of the pooled test, and PD1074 anchor samples (instrumentation variation captured) are in each run. **(D)** In LC-MS, stable PD1074 spectral features were first identified from PD1074 pools and retained if present above the level of the blank in 100% of the individual PD1074 spectra. In NMR, semi-automated peak-picking and binning were performed to extract peak heights and identify stable peaks present in PD1074 samples. (**E**) Data analysis was performed using meta-analysis models to identify spectral features of interest.

### Meta-analysis identifies differences in spectral features between test and reference strains without the need for complex normalization

For each spectral feature, the difference between the PD1074 anchor LSCP (n=6-10) and each test strain (n=2-6) was estimated for each batch. We identified statistically significant spectral features by comparing a single test strain to the control and performing a meta-analysis across the two batches[59]. By comparing test samples to the PD1074 anchor sample within the batch, we did not need to estimate and remove batch variance [37]. Although it is common in metabolomics experiments to use human plasma, urine, or commercially available reference materials, these samples are made in batches of finite material, creating referencing issues in long-term studies when batches of standards are exhausted[31]. Instead, we used the inter-sample variation within batches to compare the test and control samples and estimated a relative effect size in each batch (meta-analysis) (**Figure 3.1E**). We compared this new approach with a linear models analysis[60] (**Figure S3.3**) and demonstrated that final inferences are very similar, as predicted in larger studies that have compared individual analyses and meta-analytic approaches[60]. An advantage of the meta-analysis is the ability to apply this technique generally, even when there are complex patterns of variance such as those present in cohort studies or due to technical variation (*e.g.*, after an instrument interruption). We then leverage the effect sizes calculated from the anchored design to compare strains to each other even when data acquisition occurred independently.

We see a similar pattern across platforms for the percentage of significant features identified across the three studies, with the highest percentage found in the RP LC-MS (-) dataset (**Figure 3.2A**). The highest percentage of significant spectral features was 58% in the CM mutant study. In the individual strains, the CM mutant, VC1265 (*pyk-1*) had the largest overall effect across platforms and fractions, followed by RB2347 (*idh-2*). AUM2073 (*unc-119*) and KJ550 (*aco-1*) had the smallest overall effects (**Figures 3.2B, 3.3, and S3.4**). For the UGT mutants, VC2512 (*ugt-60*) had the largest overall effect, followed by RB2607 (*ugt-49*). RB2011 (*ugt-62*) had the smallest overall effect (**Figures 3.2C and S3.5**). These patterns demonstrate the variation in single knockouts of different genes with similar functions. In the NS, the most genetically divergent strains from PD1074 (CB4856 and DL238) had the largest overall effect in both platforms, and N2 had a small set of differences, as expected, since N2 is a variant of PD1074 (**Figures 3.2D and S3.6**).

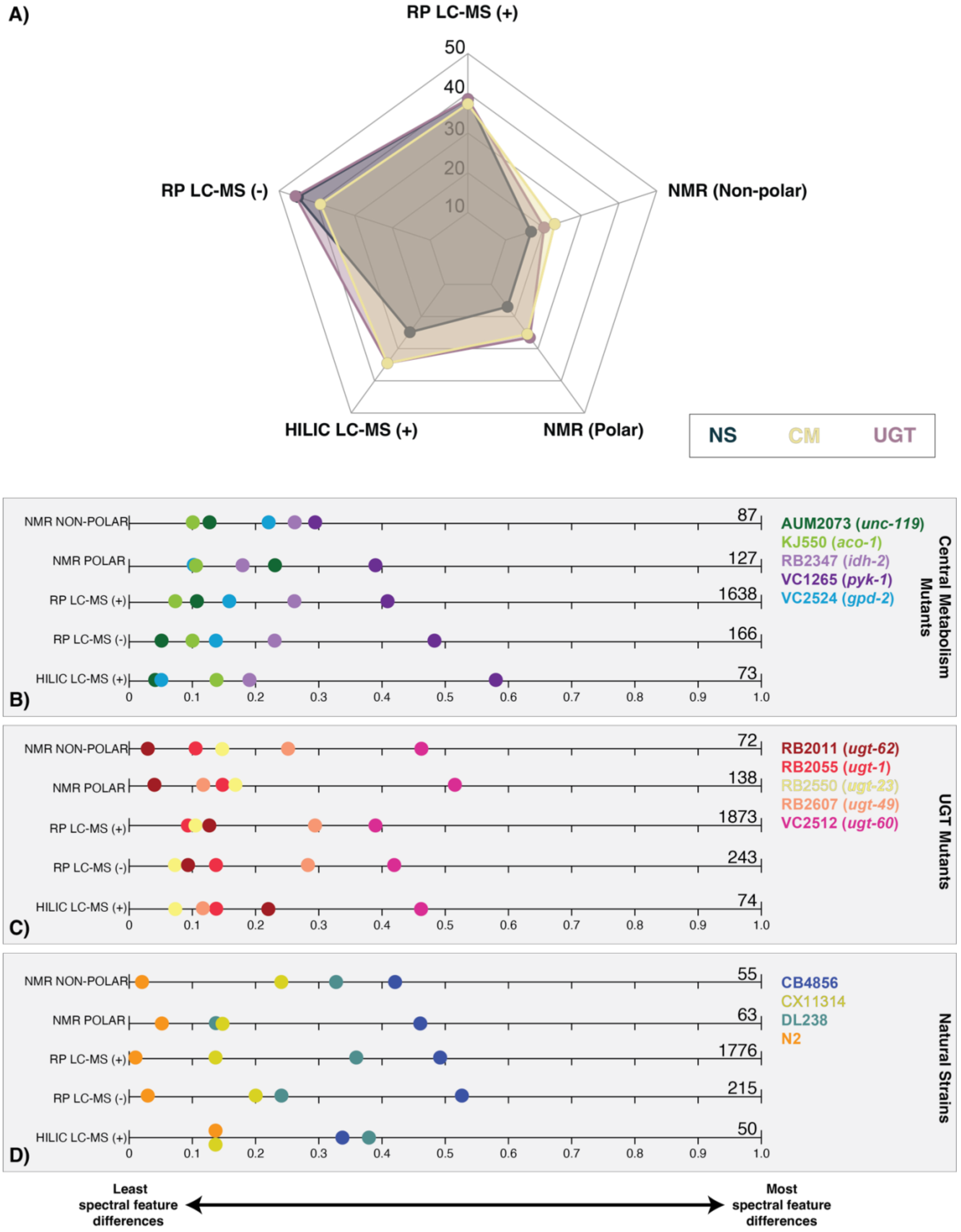

A)
RP LC-MS (+)
50
40
30
20
10
NMR (Non-polar)
NMR (Polar)
HILIC LC-MS (+)
RP LC-MS (-)
NS
CM
UGT
B)
NMR NON-POLAR
NMR POLAR
RP LC-MS (+)
RP LC-MS (-)
HILIC LC-MS (+)
87
127
1638
166
73
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AUM2073 (unc-119)
KJ550 (aco-1)
RB2347 (idh-2)
VC1265 (pyk-1)
VC2524 (gpd-2)
Central Metabolism Mutants
C)
NMR NON-POLAR
NMR POLAR
RP LC-MS (+)
RP LC-MS (-)
HILIC LC-MS (+)
72
138
1873
243
74
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
RB2011 (ugt-62)
RB2055 (ugt-1)
RB2550 (ugt-23)
RB2607 (ugt-49)
VC2512 (ugt-60)
UGT Mutants
D)
NMR NON-POLAR
NMR POLAR
RP LC-MS (+)
RP LC-MS (-)
HILIC LC-MS (+)
55
63
1776
215
50
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
CB4856
CX11314
DL238
N2
Natural Strains
Least spectral feature differences
Most spectral feature differences

**Figure 3.2. Summary of significant spectral features found in each analytical platform and across the three studies identified via the meta-strain model. (A)** Percent of significant features found in at least one strain in each of the five technologies. Significant spectral features identified in at least one strain by study are displayed for **(B)** central metabolism mutants, **(C)** UGT mutants, and **(D)** natural strains. Strains at zero have no significant spectral feature differences from the anchor strain, PD1074. Strains at one have the most significant spectral feature differences from PD1074. Significant feature totals are summarized at the end of the plot and detailed in **Table 3.1**, which lists the total spectral features identified via the meta-strain model.

| **Study Group** | **Strain** | **Number of significant spectral features** | | | | |
|---|---|---|---|---|---|---|
| | | **NMR** | | **LC-MS** | | |
| | | **Non-polar** | **Polar** | **RP +** | **RP -** | **HILIC +** |
| Central metabolism mutants (CM) | AUM2073 | 11 | 29 | 175 | 8 | 3 |
| | KJ550 | 9 | 14 | 110 | 17 | 10 |
| | RB2347 | 23 | 22 | 421 | 38 | 14 |
| | VC1265 | 25 | 49 | 671 | 79 | 42 |
| | VC2524 | 19 | 13 | 261 | 24 | 4 |
| | Total CM sig. features by platform | 87 | 127 | 1638 | 166 | 73 |
| UGT mutants (UGT) | RB2011 | 2 | 6 | 237 | 21 | 16 |
| | RB2055 | 8 | 20 | 161 | 34 | 10 |
| | RB2550 | 11 | 23 | 200 | 18 | 5 |
| | RB2607 | 18 | 17 | 539 | 69 | 9 |
| | VC2512 | 33 | 72 | 736 | 101 | 34 |
| | Total UGT sig. features by platform | 72 | 138 | 1873 | 243 | 74 |
| Natural Strains (NS) | N2 | 1 | 3 | 22 | 6 | 7 |
| | DL238 | 18 | 15 | 631 | 52 | 19 |
| | CX11314 | 13 | 16 | 254 | 44 | 7 |
| | CB4856 | 23 | 29 | 869 | 113 | 17 |
| | Total NS sig. features by platform | 55 | 63 | 1776 | 215 | 50 |
| Total sig. features by platform | | 146 | 228 | 2541 | 281 | 115 |

**Table 3.1. Summary of significant spectral features found in all three studies across NMR and LC-MS in the meta-strain model.** The total number of significant spectral features in the meta-strain model ($p < 0.05$) for a given strain and each analytical platform are listed.

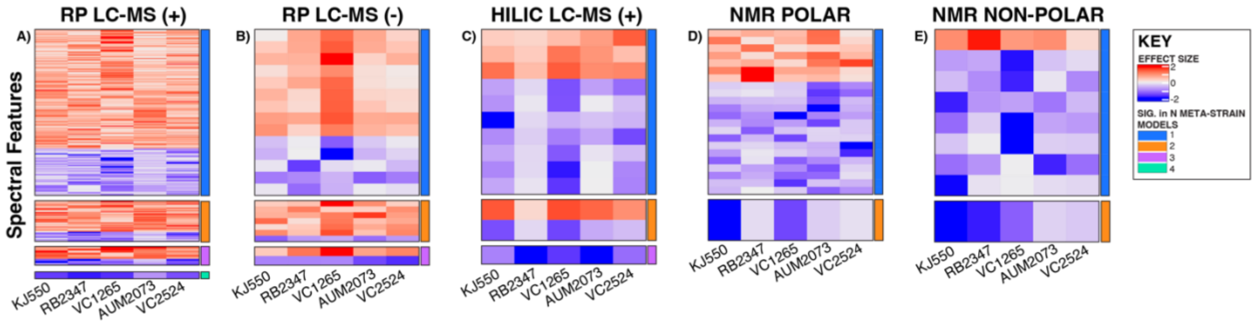


**Figure 3.3. Heatmaps of significant spectral features identified via the meta-strain model in the CM mutant study. (A)** RP LC-MS positive mode (**B)** RP LC-MS negative mode (**C)** HILIC LC-MS positive mode (**D)** NMR polar (**E)** NMR non-polar. For each heatmap, the first five columns are strains, and each row represents a spectral feature with an effect size that is consistently higher or lower relative to PD1074 in that study. The effect sizes range from (2 to -2). Positive effect sizes (*i.e.*, the strain had a higher peak at that given metabolic feature than the anchor PD1074) are displayed in red. Negative effect sizes (*i.e.*, the anchor PD1074 had a higher peak at that given metabolic feature than the test strain) are displayed in blue. The right-hand column indicates the number of strains in which a given spectral feature is statistically significant. See **Figure S3.4** for significant spectral features identified via both the meta-strain and meta-study models.

### Analysis of features across studies

The percentage of significant features in each of the mutant studies (CM and UGT) that overlapped in at least one NS (**Figure 3.4**) were features of interest for follow-up compound identification. CM mutant strains AUM2073 (*unc-119*) and RB2347 (*idh-2*) share 75% and 68% of their significant features with a NS, respectively. UGT mutants, RB2607 (*ugt-49*) and RB2055 (*ugt-1*) share 67% and 62% of their significant features with a NS, respectively. RB2011 (*ugt-62*) had the most overlap with the NS sharing 67% of its significant features in RP LC-MS (+) and 44% in HILIC LC-MS (+). See **Figures S3.7 and S3.8** for significant feature overlap across study comparisons.

We focused on compounds affected in any CM mutants and used those to identify which UGTs and NS had genetic variation in those same compounds for the NMR polar data. Using COLMAR[61], we identified three putative compounds significant in strains from all three studies. Of the 35 putative compounds showing evidence for metabolic variation in the NMR data 13 were annotated (**see Table S3.2**).

Nine putative compounds show metabolic variation in response to the *pyk-1* mutation (**Figure 3.5**). The mutation in *pyk-1* affects a large portion of the metabolome. The gene *pyk-1*, is involved in one of the last enzymes of glycolysis, encoding for pyruvate kinase and responsible for glycolytic ATP production. The depletion of lactic acid production is consistent with the mutation in *pyk-1*[62] in the strain VC1265. We saw the depletion of lactic acid in the DL238 (NS), and an increase in VC2512 (*ugt-60*) (**Figure 3.5**). As expected, none of the 13 compounds identified in the NMR polar dataset were significant in N2 (**Figure 3.5**). Interestingly, these compounds were also insignificant in CX11314, RB2055 (*ugt-1*), RB2607 (*ugt-49*), and RB2011 (*ugt-62*). Using the meta-study model, we identified significantly different features within the NS that were otherwise not significant in a single strain (meta-strain model) (**Figure S3.9**).

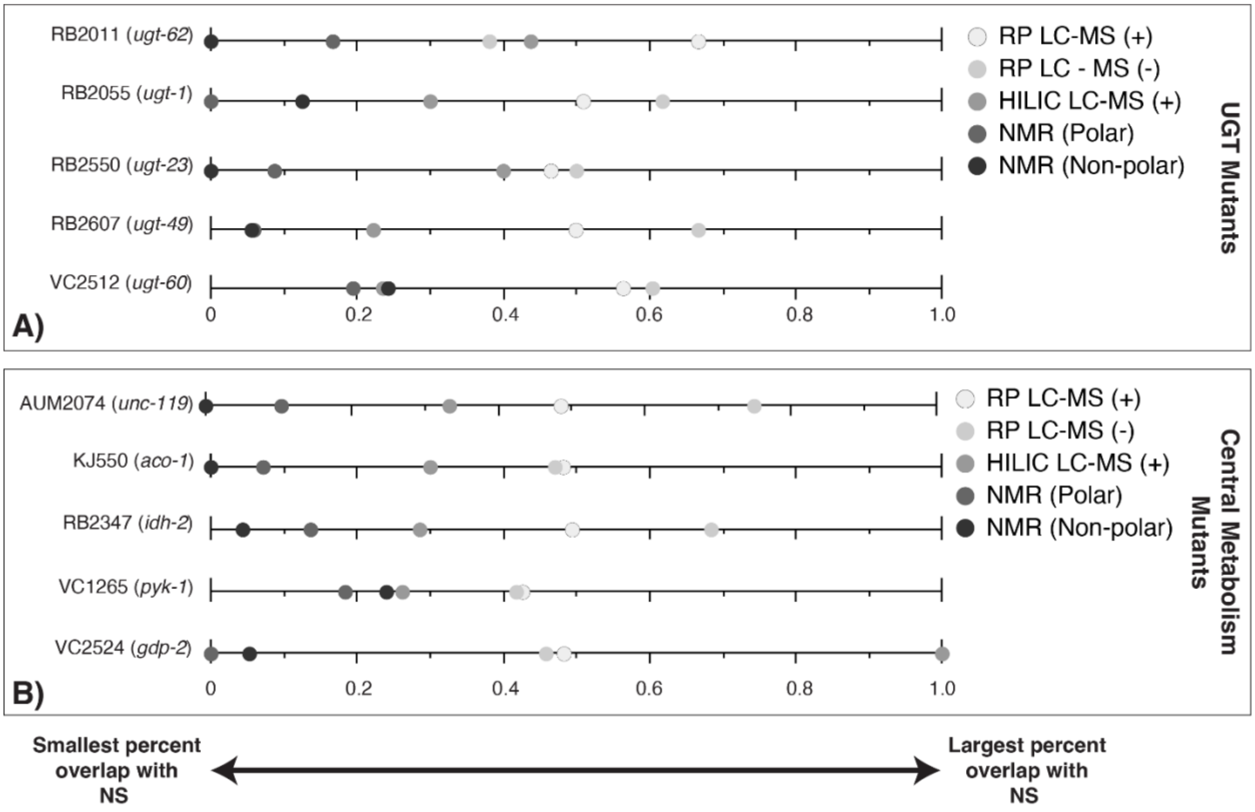


**Figure 3.4. Percent of significant features for each of the mutant studies (CM and UGT) that are also significant in at least one NS by analytical platform. (A)** UGT mutants **(B)** CM mutants. Data points at zero indicate the analytical platform detected no significant spectral features shared between the mutant strain and a natural strain. Data points at one indicate all significant spectral features for the mutant strain are shared with a natural strain for that analytical platform.

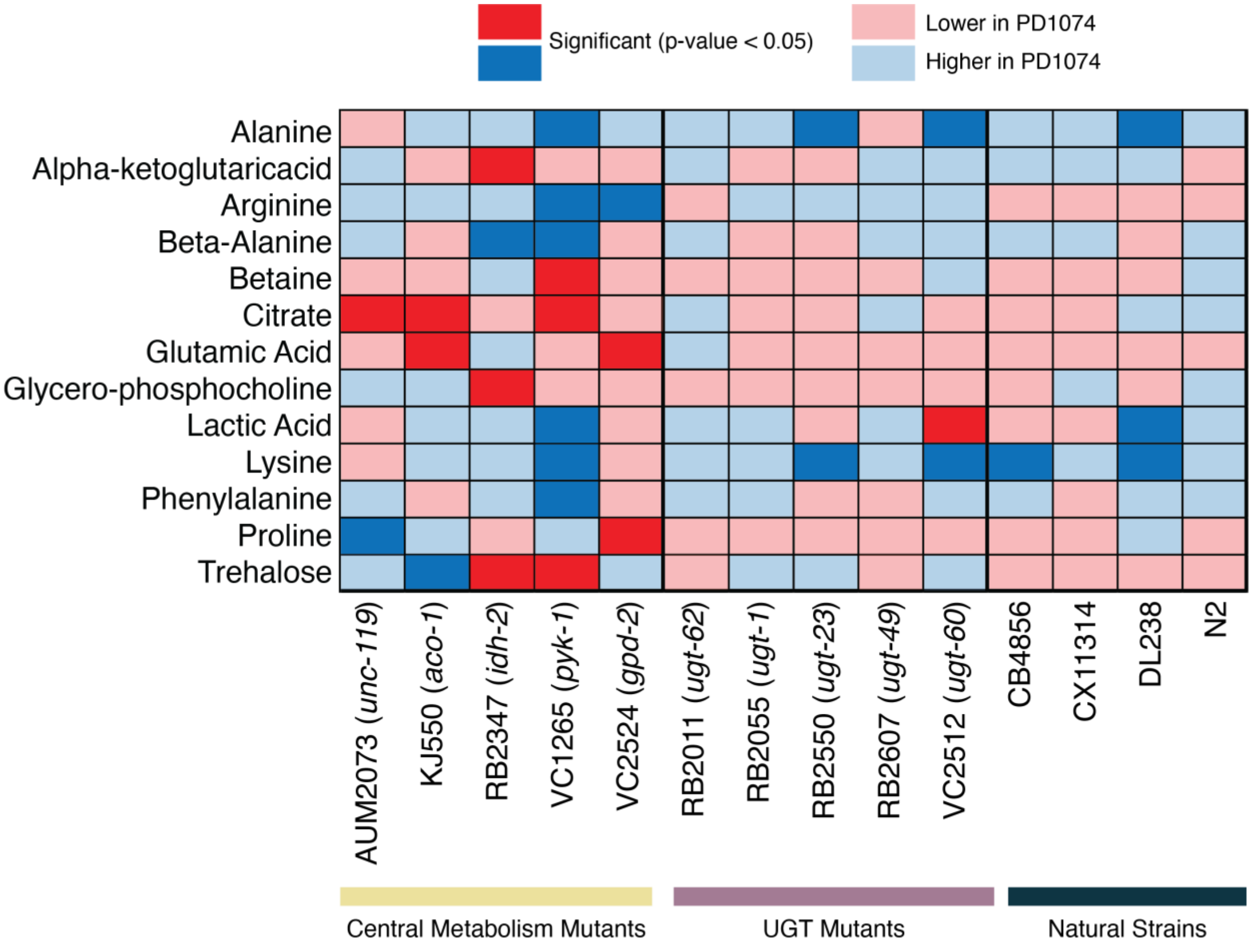


**Figure 3.5. Heatmap of metabolites identified by NMR.** Significant NMR spectral features in the central metabolism mutants are compared across UGT mutants and natural strains. Muted boxes indicate that the metabolite is not significant for that strain. Deep blue boxes indicate the metabolite is significant and more abundant in the anchor strain, PD1074. Deep red boxes indicate the metabolite is significant and more abundant in the strain listed. For compounds with more than one significant feature, the highest effect size feature is used for this figure. The significant compound list provides metabolites to pursue in subsequent experiments. See **Table S3.2** for compound annotation list and details.

## Discussion

The anchor design enables the same spectral features to be analyzed across current and future experiments. We elected to use PD1074 as our anchor as it is a trackable variant of the laboratory-adapted strain, N2. A decision to focus on stable features in the anchor strain instead of unique features in test strains enables the use of the anchor strain in downstream studies. The implementation of this method is not limited to a single genotype; multiple strains of interest can be used as anchors if they are included as individual samples and QC pools throughout the study. A focus on spectral features present despite environmental variation eliminates the need to annotate compounds prior to across study comparisons. The barrier is not a limit in the number of spectral features that can be detected, but the number that can be identified.

In non-model organism experiments, implementation of a BRM[31, 34] with multiple extractions in each batch and a pooled control enables stable feature identification. We recently developed techniques that reduce the burden of BRM[31] development. However, in the LC-MS data, we demonstrate that in the absence of a BRM, pooled anchor samples can be used to connect the experiment over time and identify stable features even with variable data acquisition conditions. For example, these scenarios can include periods of instrument maintenance which can affect data quality.

A similar approach has been suggested by normalizing to a QC or BRM material included in each batch[10, 15, 16]. Perhaps of most value to experimental success is the inclusion of several alternate strategies for stable feature identification. Even though a high abundance of features due to contaminants from sample preparation are present in the LC-MS dataset, the anchor design and use of individual PD1074 samples along with their pools enable the identification of stable spectral features. We include multiple individual samples from a control group and a pool in each batch to filter features that vary due to environmental variation during sample collection. Additionally, this provides the added benefit of enabling a straightforward calculation of the relative effect size for each test strain by batch that can be used later to combine batches or compare test strains.

Our analysis focuses on a single anchor genotype and allows us to compare the effect of all common stable spectral features across experiments. The same result can be achieved by identifying a control group. Studies where data are acquired over time can be compared because they are connected by the anchor samples. Anchors enable spectral features to be prioritized for future studies so that database matching and, ultimately, compound identification efforts are focused on the most likely biologically important spectral features. This aspect is important as model systems and genetic studies[7, 9, 63] increase in size and complexity[31, 36, 64, 65].

Effect sizes calculated in the meta-strain model are comparable to those calculated by a linear model analysis, demonstrating the successful implementation of meta-analysis when sample sizes are small (**Figure S3.3**). Additionally, we demonstrate how a list of significant spectral features can be used to focus NMR compound identification efforts (13 annotated compounds) (**Figure 3.5**). A similar approach can be used for LC-MS where features are annotated using accurate mass, elemental formula, MS/MS database matching, and *in silico* predictions of spectral features. Compound identification approaches for LC-MS are challenging and oftentimes require orthogonal data for confident annotations. This approach allows future MS/MS experiments to target spectral features of interest. The existence of multiple anchor control samples can be used to collect these specific features rather than relying on the data-dependent acquisition (DDA) or iterative DDA approaches. *In silico* prediction methods for NMR and MS/MS have improved accuracy; however, they remain computationally expensive and intractable, especially as molecular weight increases.$^{1}$H and $^{13}$C 1D NMR and MS/MS fragmentation *in silico* predictions can be prioritized for target features identified with this approach[66]. However, ambiguity is expected to remain for large molecular weight formulas, although the set of possible compounds can be significantly reduced[67].

Mapping metabolites in pathways is complicated because many metabolites are involved in multiple pathways and/or have yet to be described. The genetic mutation approach used to annotate gene function in pathways has had limited success in untargeted metabolomics because of the inability to track features across experiments, the exclusion of relevant mutants needed for a comprehensive study, and the necessity of subsequent rescue experiments to discern pathway-gene relationships. Using a meta-study approach, we identify significant features across a study that are not differentially expressed in meta-strain comparisons (**Figure S3.8**). This allows for the identification of significant spectral features when small sample sizes prohibit statistical inferences in the meta-strain model or when batch effects are large contributing to heteroscedasticity. Similarly, untargeted studies of collections of genotypes[68] using an anchor genotype, in this case PD1074, can leverage data across experiments and increase the utility of untargeted metabolomics for genetic studies and may increase the efficiency of the compound identification process[14].

## Online Methods

### *C. elegans* Strain Selection

This study used 15 *Caenorhabditis elegans* strains obtained from the *Caenorhabditis* Genetics Center (CGC) and *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) [41]. Fourteen *C. elegans* strains were used as ‘comparison strains,’ and one strain, PD1074, was used as the ‘reference strain’ (**Table S3.1**). These strains were selected to cover the diversity of interests in the metabolomics community, to encompass samples with mutations in primary and secondary metabolism, along with natural strains.

### *C. elegans* Sample Growth and Preparation

Large populations of nematodes were generated for every biological replicate with minimal variability[43]. The stable *Escherichia coli* (*E. coli*) IBAT (iterative batch average method) BRM and food source used throughout this experiment was described previously[31]. Briefly, a large-scale culture plate (LSCP) was used for each biological sample to generate a large mixed-stage population of worms (four to seven LSCP replicates per test strain). For each LSCP, worms were collected, population size estimated, and subsequently divided into at least 12 identical aliquots of 200,000 worms in ddH$_2$O and flash-frozen in liquid nitrogen and stored at -80°C[43]. As a quality control sample, *C. elegans* IBAT BRM was generated and saved in 200,000 worm aliquots[31].

### Study Design

Each *C. elegans* strain was reared with at least one PD1074 LCSP (*i.e.*, anchor sample). Sample collection for all three studies lasted more than six months. Each culture initiation and harvest included at least one LSCP of PD1074. To ensure handling was consistent, no more than five LSCPs were handled at a given time. There are 29 independent PD1074 LCSPs collected and 104 independent test strain sample LSCPs. The PD1074 represents an augmented (*i.e.*, anchor) design[37, 39], where one PD1074 replicate ('check') was matched with each test strain replicate ('new treatments') where the test strains have fewer replicates than PD1074s, augmenting the standard design.

### Iterative Batch Average Method (IBAT) in PD1074

An IBAT control[31], made up of pools of PD1074, was generated to assess batch variance across the six batches in this study. Briefly, aliquots of PD1074 were pooled together to generate a BRM that (i) minimizes the variance between batches of PD1074 BRM, (ii) can be used throughout large-scale experiments, and (iii) can be used to determine the magnitude of variation at multiple points in a metabolomics experiment. See Gouveia *et al.*, 2021 for more details on the IBAT process[31].

### Lyophilization

Frozen aliquots of 200,000 *C. elegans* worms were retrieved from -80°C and lyophilized in a VirTis® BenchTop™ "K" Series Freeze Dryer (*SP Industries, Inc.)*. After lyophilization, each aliquot was weighed and stored at -80°C until homogenization.

### Batching and Quality Control Across Analytical Platforms

Up to 24 extractions could be performed simultaneously based on centrifuge capacity limitations. Six extraction batches were needed to accommodate all the strains. Extraction batches were designed in sets of two consecutive batches so that each test strain has all replicates measured in close proximity. The test strains belong to three studies, and the studies are used to create three sets of two batches each, for a total of six batches. The three sets were collected back-to-back in NMR but are separated in time by some months in the LC-MS, although the column and instrument are the same for all three sets. There was a needle failure between batches 5 and 6 in the HILIC LC-MS run. The NS were collected in batches 1 and 2, most of the CM mutants in batches 3 and 4 (exception, AUM2073 and VC2524 were collected in batches 5 and 6), and the UGT mutants were mostly in batches 5 and 6 with (exception, RB2011 was collected in batch 1). Each extraction batch includes half of the replicates for each test sample type (balanced across two consecutive batches), a set of PD1074 anchor LCSPs, the IBAT control, and an extraction blank. Extraction blanks were processed with test strain and PD1074 aliquots to control for homogenization and extraction steps to account for non-biologically related LC-MS or NMR features that arise from sample preparation. Test LSCPs were unique to a batch, but aliquots from the same PD1074 LSCPs may be included more than once. Multiple aliquots of the same LSCP enables QC of feature selection and alignment were included as these differ only by technical variance (*i.e.*, instrument and extraction) (**Figure 3.1**).

### NMR Sample Homogenization and Extraction

Frozen lyophilized *C. elegans* aliquots were retrieved from -80°C. 200 μL of 1 mm zirconia beads (BioSpec Products) were added to each sample and homogenized at 420 rcf for 90 seconds in a FastPrep-96 homogenizer and subsequently placed on dry ice for 90 seconds to avoid overheating; this step was repeated twice for a total of three rounds.

Using the homogenized samples, 1 mL of 100% IPA chilled to -20°C was added to the lyophilized/homogenized sample powder and Zirconia beads in two increments of 500 μL. After each addition of 500 μL, samples were vortexed for 30 seconds – 1 min., and left at room temperature (RT) for 15 - 20 minutes. After RT incubation, samples were stored overnight (~12 hours) at -20°C. Samples were centrifuged for 30 minutes at 4°C (20,800 rcf). The supernatant was transferred to a new tube to analyze non-polar molecules. 1 mL of pre-chilled 80:20 $CH_3OH:H_2O$ (4°C) was added to the remaining worm pellet to analyze polar molecules. The polar fraction was allowed to shake at 4°C for 30 minutes. Samples were centrifuged at 20,800 rcf for 30 minutes at 4°C. The supernatant was transferred to a new tube to analyze non-polar molecules. Both polar and non-polar samples were placed in a Labconco Centrivap at RT and monitored until completely dry. Once dry, polar samples were reconstituted in $D_2O$ (99%, Cambridge Isotope Laboratories, Inc.) in a 100 mM sodium phosphate buffered solution with 0.11 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS-D6; 98%; Cambridge Isotope Laboratories, Inc.). Once dry, non-polar samples were reconstituted in $CDCl_3$ (99.96%; Cambridge Isotope Laboratories, Inc.). Samples were vortexed until fully soluble, and 45 μL of each sample were transferred into 1.7 mm NMR tubes (Bruker SampleJet) for acquisition.

**NMR Acquisition**

To collect the polar fraction, one-dimensional (1D) $^1$H NMR spectra were acquired with a noesypr1d pulse sequence on a NEO 800 MHz Bruker NMR spectrometer equipped with a 1.7mm TCI cryoprobe and a Bruker SampleJet autosampler cooled to 6°C. During acquisition, 32,768 complex data points were collected using 128 scans with two additional dummy scans. The spectral width was set to 15 ppm.

To collect the non-polar fraction, one-dimensional (1D) $^1$H NMR spectra were acquired with a zg pulse sequence (zg30). During acquisition, 65,536 complex data points were collected using 64 scans with four additional dummy scans. The spectral width was set to 20.2 ppm.

In addition, immediately after each 1D acquisition, a 2D J-resolved spectrum is collected using the Bruker pulse program jresgpprqf. For both the polar and non-polar fractions, 8,192 and 40 points were collected using eight scans, four dummy scans, and spectral widths of 16 and 0.09 ppm, respectively. See metabolomics workbench Study IDs (NMR polar: ST002095; NMR non-polar: ST002096) for additional acquisition parameters and data.

For metabolite identification the web server COLMARm was used. As inputs three two-dimensional experiments 1H-1H TOCSY (dipsi2gppphzspr), 1H-13C HSQC (hsqcetgpsisp2.2) and 1H-13C HSQC-TOCSY (hsqcdietgpsisp.2) collected on separate pooled PD1074 polar samples were used. The HSQC experiment was collected using 6250 and 720 points in the indirect and direct dimensions, 32 scans and 16 dummy scans and a spectral width of 13 ppm for the proton and 165 ppm for the carbon dimensions. The HSQC-TOCSY experiment parameters were identical to HSQC except for 32 dummy scans and a 90 ms mixing time. The TOCSY experiment was collected with 7272 points and 800 points in the indirect and direct dimensions, 32 scans and 16 dummy scans, a spectral width of 11.367 ppm in both dimensions and a mixing time of 90 ms. Peak picking and spectral match against hydrophilic metabolite databases (*i.e.*, HMDB and BMRB) was carried out by COLMARm using 0.04 and 0.3 ppm chemical shift cutoffs for $^1$H and $^{13}$C respectively and a matching ratio cutoff of 0.6. See metabolomics workbench Study IDs (NMR polar: ST002095; NMR non-polar: ST002096) for all the acquisition parameters and data.

**NMR Data Processing**

Following data acquisition, the data were processed using NMRPipe [69]. Fourier transform, an exponential line broadening of 1.5 Hz and manual phase correction were carried out. Using the tools from (MATLAB, The MathWorks, R2019a[70]), the spectra were referenced at 7.24 ppm using the $CDCl_3$ resonance, and the polar extracts are referenced at 0.00 ppm using DSS. Solvent regions were removed followed by baseline correction using a statistical smoothing function[71]. Alignment was performed using CCOW[72] and PAFFT[73] algorithms. Manual curation of semi-automated peak-picking was carried out by peak picking that used a binning algorithm[74] to extract peak heights. This was done separately for blanks and samples. Individual spectral features were removed if detected in the solvent and process blanks using the BFF function in SECIMtools[75].

Two-dimensional NMR experiments were also processed using NMRPipe. Spectra were Fourier transformed, a $90^O$ shifted sine window function and automatic zero filled applied, manually phased and referenced to DSS.

Stable spectral features were compared between individual PD1074 samples, PD1074 pools, and IBAT controls.

## LC-MS Sample Homogenization and Extraction

Using glass and zirconium oxide beads, the aliquots were homogenized for three minutes in a Qiagen Tissuelyser 2. Homogenized worms were extracted with 1.5 mL of isopropanol (IPA) at -20°C overnight (approximately 12 hours), then pelleted and the supernatant transferred to separate 2 mL centrifuge tubes. Supernatants were then dried to completion in a Labconco Centrivap and stored at -80°C for non-polar LC-MS analysis. The pellet was extracted a second time using 80:20 methanol:water ($CH_3OH:H_2O$) (v:v) for 20 minutes at RT while shaking at 1500 rpm. Samples were again pelleted to separate proteins, and the supernatant was transferred to separate 2 mL centrifuge tubes, dried down to completion, and stored at -80°C for polar LC-MS analysis.

## LC-MS Acquisition and Processing

Each instrument run for a single batch included the following controls with replicate injections at the beginning and end of the batch: instrument control, extraction blanks, pooled test sample aliquots, and pooled PD1074 anchor aliquots. In the middle of the batch, individual test samples and PD1074 controls were injected in a randomized order to reduce systematic bias.

Non-polar extracts were reconstituted in 75 µL of IPA containing isotopically labeled lipid standards and analyzed by LC-MS using a ThermoFisher Scientific Accucore C30 150 x 2.1mm, 2.6 µm column paired with a Thermo Fisher Orbitrap ID-X in positive and negative polarity. Polar (80:20 $CH_3OH:H_2O$) extracts were reconstituted in 75 µL of 80:20 $CH_3OH:H_2O$ containing isotopically labeled arginine, hypoxanthine, hippuric acid, and methionine (Cambridge Isotope Laboratories, Inc.) and analyzed by LC-MS using a Waters BEH Amide 150 x 2.1 mm, 1.7 µm column paired with a Thermo Fisher Orbitrap ID-X in positive and negative polarity. LC-MS/MS data for each mode of analysis was collected using three rounds of iterative DDA (Thermo Scientific AcquireX) performed on pooled test samples.

Data for each sample was collected in full MS1 with a resolution of 240,000 FWHM (full-width half-maximum) and MS/MS spectra of pooled samples were collected at a resolution of 30,000 FWHM using a 0.8da isolation window and stepped HCD collision energies of 15, 30, and 45. See supplemental information for detailed LC-MS parameter settings. Thermo .raw files were converted to centroid mode and .mzML format using Proteowizard's MSconvertGUI tool[76]. Raw files are deposited at metabolomics workbench Study ID ST002092. Pre-processing steps, input parameters, and set values used for LC-MS data are listed in **Table S3.4**.

### Selection of stable LC-MS spectral features

A plasticizer contamination event precluded us from quantitatively assessing the performance of an IBAT control in the LC-MS experiments and was eliminated from the LC-MS data. Instead, we used the 12 PD1074 anchor pools encompassing the individual PD1074 anchor samples in a two-step procedure. We use the PD1074 anchor samples here because these anchor pools are present by design in each of the three sets and six batches for the initial peak picking and alignment. First, these samples were averaged over extraction variance but differ based on instrumentation variation across batches. In the second step, we retained the subset of peaks only present in 100% of the individual PD1074 samples to focus on stable peaks across growth conditions. Here, we focus on peaks present across multiple individual samples of the same genotype, PD1074, a variant of the laboratory-adapted strain N2, but any strain of *C. elegans* could serve this purpose.

The 12 pooled PD1074 samples were used to estimate optimal parameters, and then these parameters were applied to all samples using a memory-efficient algorithm SLAW (https://github.com/zamboni-lab/SLAW).[77] Only spectral features above the blank threshold of 100 for all 12 anchor pools were retained for further analysis. SLAW offers the following peak picking algorithms: XCMS centWave[78, 79], OpenMS FeatureFinderMetabo[80, 81], and MZmine ADAP[82, 83] For this study, ADAP was selected as the peak picking algorithm[84]. The SLAW algorithm is predicated upon the assumption that the experimental design includes identical QC samples across an experiment (*e.g.*, BRM) in intervals during data collection. This inclusion is typical in large-scale studies[31, 85, 86], but the selection of stable spectral features across extraction variance is not standard. While the benefits of including QC samples are known and recently have been implemented in peak picking and alignment optimization workflows that traditionally have not scaled to large data[87], the inclusion of anchored samples during sample generation, analytical measurement, and data processing is novel.

Spectral peaks were filtered further using the individual PD1074 anchor samples. We took a conservative approach requiring 100% of the PD1074 samples to have each spectral feature present above the blank. This focuses the experiment and our attention on spectral features that are likely to be present in a subsequent independently prepared MS2 experiment in the compound identification process, and not spectral features present sporadically due to variation in growth or extraction.

### Quality Control Assessments for LC-MS and NMR Data

Stable spectral features were rank transformed (*i.e.*, raw data is replaced by ranks where the lowest rank has the smallest peak height, and the highest rank has the largest peak height for a given spectral feature). QC assessments included Standard Euclidean Distance (SED), principal component analysis (PCA), coefficient of variation (CV), Bland Altman (BA), and sample density distributions to identify potential feature artifacts and/or atypical samples[75]. See **Table S3.3** for QC parameters and thresholds used to identify stable mass features[75]. PCA is used to visualize distortions due to batch or genotype. BA plots on pools and anchor samples within a batch were used to visualize alignment variation, and BA plots on replicate aliquots of the same anchor samples were used to verify the success of the alignment across batches. Per feature CV is examined to identify any wildly aberrant features and was used to help refine the quantification of the solvent front.

Sample outliers were identified based on the SED plots. Chromatograms of samples whose distance to other samples did not cross the 95% percentile for the distribution of pairwise distances were manually examined for chromatography failure. The PD1074 LSCP sample "aos54" failed the QC assessment for NMR. The PD1074 LSCP samples "aos53" and "aos41" failed the QC assessment for RP LC-MS datasets. Test strain "aos49" in batch 5 is removed from all datasets, and test strain "aos25" in batch 1 was removed from the HILIC LC-MS positive dataset. Samples were removed from further consideration in their respective datasets.

**Meta-analysis on LC-MS and NMR Data**

All replicates of a particular genotype were contained within two sequential batches: however, different test strains within the same study span multiple sets. We used meta-analysis for each feature to compare the test genotype to the control, where each batch is treated as an 'experiment' using a fixed effects (FE) model using standardized mean difference (SMD)[75], referred to as "meta-strain" model throughout. Positive effect sizes indicate that the test strain has a higher peak than PD1074 for a given chemical feature. Negative effect sizes indicate PD1074 has a higher peak than the test strain for a given chemical feature. Each strain was tested against PD1074 to see if that feature is differentially expressed between PD1074 and the test strain. We also used meta-analysis to compare test genotypes to each other, referred to as the "meta-study" model throughout. For example, for the five UGT mutants (*i.e.*, RB2011, RB2550, RB2055, RB2607, and VC2512) we tested whether a feature is differentially expressed between the test strain and anchor in all five test genotypes. See supplemental information for more details.

### NMR $^1$H 1D spectra Annotation

Significant features obtained from the "meta-strain" analysis of the CM mutants were selected for identification. The 2D experiments HSQC, HSQC-TOCSY, and TOCSY were collected from a pooled PD1074 sample. These data served as inputs to the public webserver COLMARm[88] (Complex Mixture Analysis by NMR), an application that allows us to simultaneously and interactively compare multiple 2D spectra data to HMDB[89], BMRB[90], and NMRShiftDB[91] publicly available databases. Only the significant features were annotated. Annotation confidence scores per compound are detailed in **Table S3.2** according to the previously reported levels as described elsewhere[92].

Further annotation details can be found in the COLMAR outputs submitted to Metabolomics Workbench. **Figure 3.5** illustrates the annotated compounds. Only the feature with the highest effect size was selected for compounds with more than one significant feature. After a list of compounds was identified, WormFlux[44] was used to explore the effects of the CM mutants on the *C. elegans* metabolic network.

### Code Availability

The python code for QA/QC is available through GitHub (https://github.com/secimTools/SECIMTools) and can be run via a Galaxy install (https://docs.galaxyproject.org/en/master/) or from a command line interface. The meta-analysis (meta_analysis.py) and rank transformation (add_group_rank.py) python code are available on the SECIMtools GitHub page. The Matlab functions used as well as instructions and version control are available at https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA.

## Data Availability

These data are available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, **https://www.metabolomicsworkbench.org** where it has been assigned Study ID (LC-MS: ST002092; NMR polar: ST002095; NMR non-polar: ST002096). The data can be accessed directly via its Project DOI: http://dx.doi.org/10.21228/M82978. This work is supported by NIH grant U2C-DK119886. Methods and protocols used in this study are available on protocols.io (NMR: dx.doi.org/10.17504/protocols.io.b2rbqd2n), LCMS: (dx.doi.org/10.17504/protocols.io.bahjib4n).

Acknowledgements

Research reported in this manuscript was supported by the National Institutes of Health Award Number U2CES030167 (A.S.E., E.C.A., F.E.L., F.M.F., I.J.A., K.M.M., L.M.M.). F.M.F. also acknowledged support from NIG grant 1R01CA218664-01. The authors would like to thank Pamela Kirby at the Edison Lab for assistance with material generation and storage logistics. Thank you to John Glushka for NMR support. Thank you to Oleksandr (Alex) Moskalenko for application and bioinformatics support, and development of SECIMTools 3.0. HiPerGator, the University of Florida High Performance Computing Platform and it's support staff in particular Erik Deumens for co-ordination. Kelsey Sinclair and Carter Johnson for help with documentation for the SECIMTools 2.0 updates. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). Some strains were provided by CeNDR, which is funded by NSF Living Collections CSBR 1930382. Lastly, we thank WormBase.

# References

1. Liu, X. & Locasale, J.W. Metabolomics: A Primer. *Trends Biochem Sci* **42**, 274-284 (2017).
2. Burgess, D.J. The TOPMed genomic resource for human health. *Nat Rev Genet* **22**, 200 (2021).
3. Peng, B., Li, H. & Peng, X.X. Functional metabolomics: from biomarker discovery to metabolome reprogramming. *Protein Cell* **6**, 628-637 (2015).
4. Schmidt, J.C. et al. Metabolomics as a Truly Translational Tool for Precision Medicine. *Int J Toxicol* **40**, 413-426 (2021).
5. Hastings, J. et al. Multi-Omics and Genome-Scale Modeling Reveal a Metabolic Shift During C. elegans Aging. *Front Mol Biosci* **6**, 2 (2019).
6. Jones, D.P., Park, Y. & Ziegler, T.R. Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu Rev Nutr* **32**, 183-202 (2012).
7. Menni, C., Zierer, J., Valdes, A.M. & Spector, T.D. Mixing omics: combining genetics and metabolomics to study rheumatic diseases. *Nat Rev Rheumatol* **13**, 174-181 (2017).
8. van der Sijde, M.R., Ng, A. & Fu, J. Systems genetics: From GWAS to disease pathways. *Biochim Biophys Acta* **1842**, 1903-1909 (2014).
9. Lewis, G.D., Asnani, A. & Gerszten, R.E. Application of metabolomics to cardiovascular biomarker and pathway discovery. *J Am Coll Cardiol* **52**, 117-123 (2008).
10. Barupal, D.K. et al. Generation and quality control of lipidomics data for the alzheimer's disease neuroimaging initiative cohort. *Sci Data* **5**, 180263 (2018).

11. Rahman, M.L. et al. Plasma lipidomics profile in pregnancy and gestational diabetes risk: a prospective study in a multiracial/ethnic cohort. *BMJ Open Diabetes Res Care* **9** (2021).

12. Sindelar, M. et al. Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity. *Cell Rep Med* **2**, 100369 (2021).

13. Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal Chem* **88**, 524-545 (2016).

14. Blazenovic, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8** (2018).

15. Fan, S. et al. Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal Chem* **91**, 3590-3596 (2019).

16. Kim, T. et al. A hierarchical approach to removal of unwanted variation for large-scale metabolomics data. *Nat Commun* **12**, 4992 (2021).

17. Misra, B.B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *Eur J Mass Spectrom (Chichester)* **26**, 165-174 (2020).

18. Sherman, E. et al. Reference samples guide variable selection for correlation of wine sensory and volatile profiling data. *Food Chem* **267**, 344-354 (2018).

19. De Livera, A.M. et al. Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem* **87**, 3606-3615 (2015).

20. Huaxu Yu, T.H. Comprehensive assessment of the diminished statistical power caused by nonlinear electrospray ionization responses in mass spectrometry-based metabolomics. *Analytica Chimica Acta* **1200**, 9 (2022).

21. Wulff, J.E.M., M.W. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *Advances in Bioscience and Biotechnology* **9**, 339-351 (2018).

22. Sumner, L.W. et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211-221 (2007).

23. Fiehn, O. et al. Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J* **53**, 691-704 (2008).

24. Spicer, R.A., Salek, R. & Steinbeck, C. Compliance with minimum information guidelines in public metabolomics repositories. *Sci Data* **4**, 170137 (2017).

25. Broadhurst, D. et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **14**, 72 (2018).

26. Dunn, W.B., Wilson, I.D., Nicholls, A.W. & Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **4**, 2249-2264 (2012).

27. Fang, C. & Luo, J. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J* **97**, 91-100 (2019).

28. Molon, M. et al. Effects of Temperature on Lifespan of Drosophila melanogaster from Different Genetic Backgrounds: Links between Metabolic Rate and Longevity. *Insects* **11** (2020).

29. Smirnoff, N. Ascorbic acid metabolism and functions: A comparison of plants and mammals. *Free Radic Biol Med* **122**, 116-129 (2018).

30. Helf, M.J., Fox, B.W., Artyukhin, A.B., Zhang, Y.K. & Schroeder, F.C. Comparative metabolomics with Metaboseek reveals functions of a conserved fat metabolism pathway in C. elegans. *Nat Commun* **13**, 782 (2022).

31. Gouveia, G.J. et al. Long-Term Metabolomics Reference Material. *Anal Chem* **93**, 9193-9199 (2021).

32. Liu, K.H. et al. Reference Standardization for Quantification and Harmonization of Large-Scale Metabolomics. *Anal Chem* **92**, 8836-8844 (2020).

33. Beisken, S., Eiden, M. & Salek, R.M. Getting the right answers: understanding metabolomics challenges. *Expert Rev Mol Diagn* **15**, 97-109 (2015).

34. Wasito, H. et al. Yeast-based reference materials for quantitative metabolomics. *Anal Bioanal Chem* (2021).

35. Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **27**, 1897-1905 (2016).

36. Chamberlain, C.A., Rubio, V.Y. & Garrett, T.J. Impact of matrix effects and ionization efficiency in non-quantitative untargeted metabolomics. *Metabolomics* **15**, 135 (2019).

37. Federer, W.T., Reynolds, M., and Crossa J. Combining Results from Augmented Designs over Sites. *AGRONOMY JOURNAL* **93**, 389-395 (2001).

38. Federer, W.T. & Zelen, M. Analysis of multifactor classifications with unequal numbers of observations. *Biometrics* **22**, 525-552 (1966).

39. Federer, W.T.a.S., C. S. The Use of Covariance to Control Gradients in Experiments. *Biometrics* **10**, 282-290 (1954).

40. Consortium, C.e.S. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

41. Cook, D.E., Zdraljevic, S., Roberts, J.P. & Andersen, E.C. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res* **45**, D650-D657 (2017).

42. Edison, A.S. et al. The Time Is Right to Focus on Model Organism Metabolomes. *Metabolites* **6** (2016).

43. Shaver, A.O., Gouveia, G.J., Kirby, P.S., Andersen, E.C. & Edison, A.S. Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations. *J Vis Exp* (2021).

44. Yilmaz, L.S. & Walhout, A.J. A Caenorhabditis elegans Genome-Scale Metabolic Network Model. *Cell Syst* **2**, 297-311 (2016).

45. Girard, L.R. et al. WormBook: the online review of Caenorhabditis elegans biology. *Nucleic Acids Res* **35**, D472-475 (2007).

46. Hodgkin, J. What does a worm want with 20,000 genes? *Genome Biol* **2**, COMMENT2008 (2001).

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

48. Marquez, J. et al. Rescue of TCA Cycle Dysfunction for Cancer Therapy. *J Clin Med* **8** (2019).

49. Meech, R. et al. The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms. *Physiol Rev* **99**, 1153-1222 (2019).

50. Yang, N., Sun, R., Liao, X., Aa, J. & Wang, G. UDP-glucuronosyltransferases (UGTs) and their related metabolic cross-talk with internal homeostasis: A systematic review of UGT isoforms for precision medicine. *Pharmacol Res* **121**, 169-183 (2017).

51. Hasegawa, K., Miwa, S., Tsutsumiuchi, K. & Miwa, J. Allyl isothiocyanate that induces GST and UGT expression confers oxidative stress resistance on C. elegans, as demonstrated by nematode biosensor. *PLoS One* **5**, e9267 (2010).

52. Stupp, G.S. et al. Chemical detoxification of small molecules by Caenorhabditis elegans. *ACS Chem Biol* **8**, 309-313 (2013).

53. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**, 862-872 (2006).

54. Zhang, G., Mostad, J.D. & Andersen, E.C. Natural variation in fecundity is correlated with species-wide levels of divergence in Caenorhabditis elegans. *G3 (Bethesda)* **11** (2021).

55. Liu, Q. et al. Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Sci Rep* **10**, 13856 (2020).

56. Brunius, C., Shi, L. & Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **12**, 173 (2016).

57. Lange, E., Tautenhahn, R., Neumann, S. & Gropl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **9**, 375 (2008).

58. Smith, R., Ventura, D. & Prince, J.T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief Bioinform* **16**, 104-117 (2015).

59. Hedges, L.V. & Olkin, I. Statistical methods for meta-analysis. (Academic Press, Orlando; 1985).

60. Liu, Z., Vol. Master of Science 61 (University of Florida, Gainesville, FL; 2021).

61. Zhang, F., Robinette, S.L., Bruschweiler-Li, L. & Bruschweiler, R. Web server suite for complex mixture analysis by covariance NMR. *Magn Reson Chem* **47 Suppl 1**, S118-122 (2009).

62. Luz, A.L. et al. From the Cover: Arsenite Uncouples Mitochondrial Respiration and Induces a Warburg-like Effect in Caenorhabditis elegans. *Toxicol Sci* **152**, 349-362 (2016).

63. Gebauer, J. et al. A Genome-Scale Database and Reconstruction of Caenorhabditis elegans Metabolism. *Cell Syst* **2**, 312-322 (2016).

64. Annesley, T.M. Ion suppression in mass spectrometry. *Clin Chem* **49**, 1041-1044 (2003).

65. Wehrens, R. et al. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12**, 88 (2016).

66. Borges, R.M. et al. Quantum Chemistry Calculations for Metabolomics. *Chem Rev* **121**, 5633-5670 (2021).

67. Das, S., Edison, A.S. & Merz, K.M., Jr. Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal Chem* **92**, 10412-10419 (2020).

68. Noble, L.M., Rockman, M.V. & Teotonio, H. Gene-level quantitative trait mapping in Caenorhabditis elegans. *G3 (Bethesda)* **11** (2021).

69. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-293 (1995).

70. , Edn. R2019a (The MathWorks, Inc., Natick, Massachusetts, United States; 2019).

71. Xi, Y. & Rocke, D.M. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics* **9**, 324 (2008).

72. Tomasi G., v.d.B.F., Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **18**, 231-241 (2004).

73. Wong, J.W., Durante, C. & Cartwright, H.M. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal Chem* **77**, 5655-5661 (2005).

74. S.A.A. Sousa, A.M., Márcia Miguel Castro Ferreira Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 93-102 (2013).

75. Kirpich, A.S. et al. SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* **19**, 151 (2018).

76. Chambers, M.C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918-920 (2012).

77. Delabriere, A., Warmer, P., Brennsteiner, V. & Zamboni, N. SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Analytical Chemistry* **93**, 15024-15032 (2021).

78. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).

79. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **78**, 779-787 (2006).

80. Kenar, E. et al. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol Cell Proteomics* **13**, 348-359 (2014).

81. Röst, H.L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **13**, 741-748 (2016).

82. Myers, O.D., Sumner, S.J., Li, S., Barnes, S. & Du, X.A.-O. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.

83. Pluskal, T., Castillo S Fau - Villar-Briones, A., Villar-Briones A Fau - Oresic, M. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.

84. Myers, O.D., Sumner, S.J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem* **89**, 8696-8703 (2017).

85. Han, W. & Li, L. Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom Rev* (2020).

86. Peng, J., Chen, Y.T., Chen, C.L. & Li, L. Development of a universal metabolome-standard method for long-term LC-MS metabolome profiling and its application for bladder cancer urine-metabolite-biomarker discovery. *Anal Chem* **86**, 6540-6547 (2014).

87. Delabriere, A., Warmer, P., Brennsteiner, V. & Zamboni, N. SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Anal Chem* **93**, 15024-15032 (2021).

88. Bingol, K., Li, D.W., Zhang, B. & Bruschweiler, R. 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 (2016).

89. Wishart, D.S. et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **50**, D622-D631 (2022).

90. Ulrich, E.L. et al. BioMagResBank. *Nucleic Acids Res* **36**, D402-408 (2008).

91. Kuhn, S. & Schlorer, N.E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2--a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn Reson Chem* **53**, 582-589 (2015).

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

# CHAPTER 4

# DISCUSSION AND FUTURE DIRECTIONS

Untargeted metabolomics is a powerful approach to understand how metabolites shape complex phenotypes and their interactions. To take full advantage of the rich data generated in metabolomics studies, a robust experimental design and workflow must be implemented to collect data that addresses the study hypothesis and makes biologically meaningful conclusions. This dissertation summarizes suggested methodological approaches to implement in untargeted metabolomics experiments.

## Benefits and Limitations to the Presented Methods

The LSCP growth method, implementation of an anchor design, and identification of stable spectral features enables future experiments to adopt these methods and compare results to past and future studies. Chapter 2 showed that the LSCP growth method allows users to grow and harvest large mixed-stage populations of *C. elegans* for metabolomics studies without the challenges associated with multiple bleaching steps, liquid culture handling, or synchronization of multiple strains with different developmental timings. LSCPs allowed us to collect strains of interest and generate large *C. elegans* populations in each sample for LC-MS and NMR data collection at multiple institutions.

The biggest limitation in the LSCP method is the inability to identify unique metabolites or metabolic shifts for a given developmental stage, as we did not focus on culturing a single *C. elegans* life cycle stage. It may be possible to grow synchronized populations of worms on the

LSCPs as done on commercially available petri dishes and in liquid culture. However, if other researchers are interested in the LSCP method to synchronize strains to a single life cycle stage, a few factors should be kept in mind: (i) sample handling time will increase and (ii) growth rate will vary depending on the strain's mutation(s), behavior, and growth conditions.

In Chapter 3, the ability to track stable features in the anchor strain PD1074 provides an opportunity to compare *C. elegans* LC-MS and/or NMR spectroscopy data across experiments, laboratories, and project goals of any study that implements a common anchor strain. The anchor design and use of anchor samples allowed us to apply rigorous QA/QC methods that aided in the identification of stable features that reflect the genotype of the anchor strain instead of experimental or environmental variation, an issue commonly faced in metabolomics experiments. While metabolomics experiments regularly detect tens of thousands of spectral features, most of these detected features are unstable or are not biologically meaningful[1]. Our methods greatly reduce the number of spectral features analyzed in a dataset, which will allow us to focus on a narrowed list of stable and biologically meaningful features during compound identification. Additionally, these methods enable a comparison of data across studies. The benefits listed here will be of great use to the untargeted metabolomics community.

While there are many benefits to the approaches presented in Chapter 3, there are limitations to consider. Mapping metabolites in genetic pathways is complicated because many are involved in multiple pathways and/or pathways that have yet to be described[2, 3]. Here, we attempted to identify how central metabolism (CM) and UDP-glucronosyltransferases (UGT) mutations affected the identified spectral features and/or compounds. While we did track stable anchor features across mutant strains, our approach is limited as not all relevant mutants needed were included, and subsequent rescue experiments to discern gene pathway relationships were not performed. To pinpoint how genetic mutations affect the metabolome, the implementation of stable features for the test strains, the inclusion of all relevant mutants in a pathway, and rescue experiments would need to be performed.

Additionally, we decided to focus on stable features in the anchor strain instead of unique features in the test strains to use the anchor strain in downstream studies that include large mutant or natural strain panels. Thus, if a feature is not identified and stable in the anchor strain, it was omitted from the performed analyses. All test strains have biologically unique features and metabolites that are not present in the anchor strain; this approach loses that information. Notably, while we did not focus on unique test strain features here, a user could adjust the peak picking strategy and thresholds to focus on those unique features.

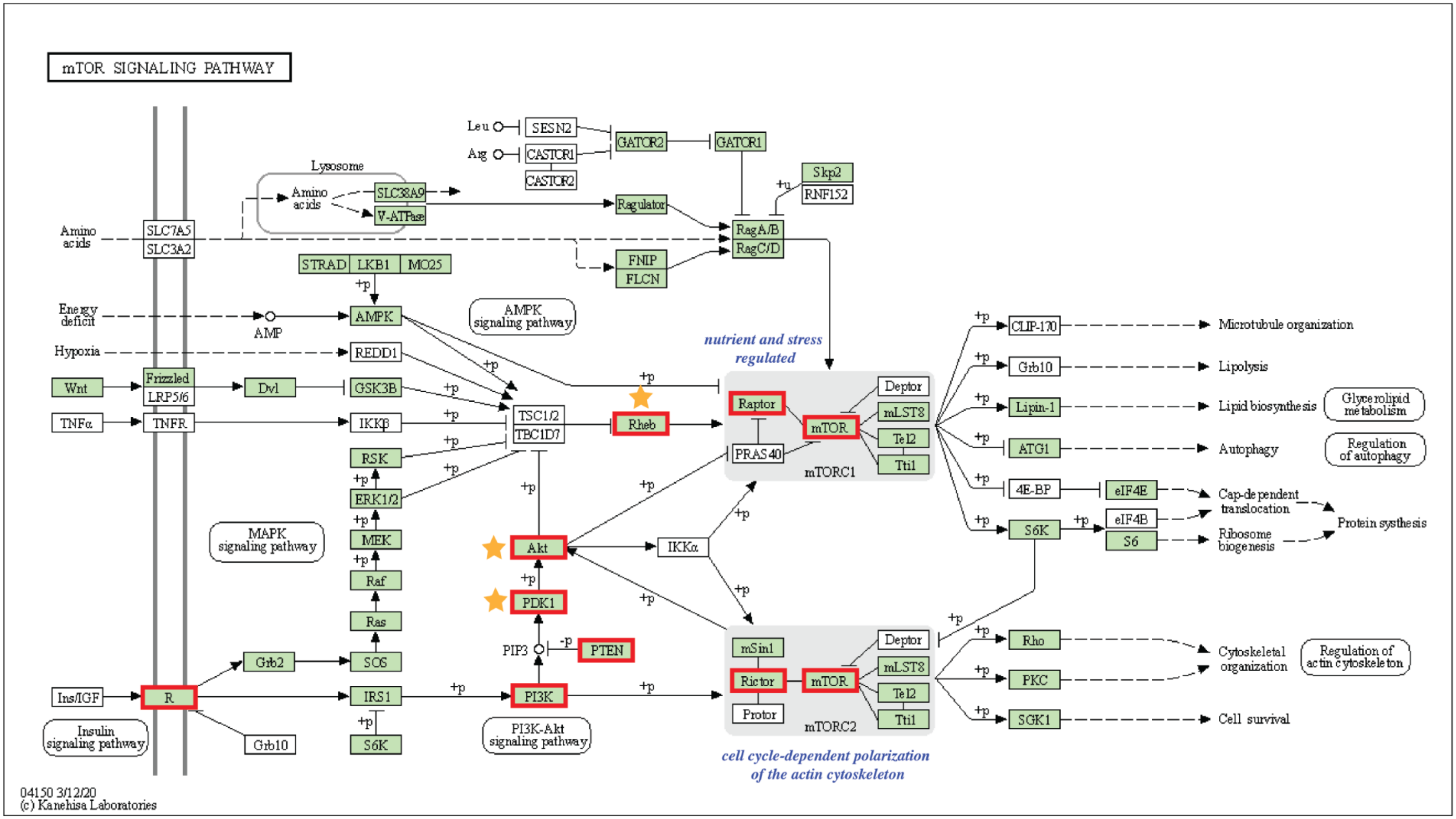
## Future Studies: Implementation of Methods in the TOR Signaling Pathway

Over the past two years, we have collected data on the target of rapamycin (TOR or mTOR) pathway. TOR is a conserved serine/threonine kinase that regulates cell growth and metabolism in response to environmental cues[4]. Here, we have implemented the experimental methods described in Chapters 2 and 3 to focus on *C. elegans* strains with mutations in (1) multiple genes and (2) the same genes in the TOR pathway (**Figure 4.1**).

By focusing on one pathway and collecting data on multiple genes within the TOR pathway, we will be able to see the metabolite shifts and changes that are up or downstream from mutations. For example, since PDK is upstream of AKT (**Figure 4.1**) we hypothesize that the knockout of *pdk-1* would affect AKT and its subsequent metabolic contribution. By including a test strain with a mutation for *pdk-1* and *akt-1*, we can directly test our hypothesis by collecting the metabolic profile of two strains directly next to one another in the TOR pathway.

The value of focusing on different strains with a mutation at the *same* gene allows us to build in an additional control, as we hypothesize that strains with a mutation in the same gene display the same metabolomic differences. For example, we have collected data on GR1318 (*pdk-1*) and PJ1134 (*pdk-1*), which are both strains with mutations for the gene *pdk-1* and dominant suppressors of the daf-c phenotype of *age-1* (*i.e.*, animals become increasingly sluggish/immobilized as they age). Therefore, we hypothesize that GR1318 and PJ11314 will have similar metabolic profiles. Once data from the TOR samples are processed, we plan to compare that study to those presented here. We also plan to continue using described methods to analyze new strains of interest.

**Figure 4.1 mTOR Signaling Pathway.** The mammalian (mechanistic) target of rapamycin (mTOR) is a highly conserved serine/threonine protein kinase, which exists in two complexes termed mTOR complex 1 (mTORC1) and 2 (mTORC2). mTORC1 contains mTOR, Raptor, PRAS40, Deptor, mLST8, Tel2 and Tti1. mTORC1 is activated by the presence of growth factors, amino acids, energy status, stress, and oxygen levels to regulate several biological processes, including lipid metabolism, autophagy, protein synthesis and ribosome biogenesis. On the other hand, mTORC2, which consists of mTOR, mSin1, Rictor, Protor, Deptor, mLST8, Tel2 and Tti1, responds to growth factors and controls cytoskeletal organization, metabolism, and survival. Red boxes indicate biological replicates for strains carrying a mutation at that point in the pathway have been collected. Collection of additional strains in the mTOR pathway is in progress. Stars indicate we have multiple strains carrying a mutation for that gene in the mTOR pathway. mTOR signaling pathway reproduced with permission from KEGG Pathway entry: cel04150 (https://www.kegg.jp/entry/pathway+cel04150).


mTOR SIGNALING PATHWAY
Leu
SESN2
Arg
CASTOR1
CASTOR2
GATOR2
GATOR1
Skp2
+u
RNF152
Lysosome
Amino acids
SLC38A9
V-ATPase
Regulator
Amino acids
SLC7A5
SLC3A2
RagA/B
RagC/D
FNIP
FLCN
STRAD
LKB1
MO25
+p
Energy deficit
AMP
AMPK
AMPK signaling pathway
Hypoxia
REDD1
nutrient and stress regulated
Wnt
Frizzled
LRP5/6
Dvl
GSK3B
+p
TNFα
TNFR
IKKβ
TSC1/2
TBC1D7
Rheb
RSK
ERK1/2
MEK
Raf
Ras
SOS
Grb2
MAPK signaling pathway
Akt
PDK1
PIP3
-p
PTEN
IKKα
Ins/IGF
R
IRS1
PI3K
S6K
Grb10
Insulin signaling pathway
PI3K-Akt signaling pathway
Raptor
mTOR
Deptor
mLST8
Tel2
Tti1
PRAS40
mTORC1
mSin1
Rictor
mTOR
Deptor
mLST8
Tel2
Tti1
Protor
mTORC2
cell cycle-dependent polarization of the actin cytoskeleton
CLIP-170
Microtubule organization
Grb10
Lipolysis
Lipin-1
Lipid biosynthesis
Glycerolipid metabolism
ATG1
Autophagy
Regulation of autophagy
4E-BP
eIF4E
Cap-dependent translocation
eIF4B
S6K
S6
Ribosome biogenesis
Protein systhesis
Rho
PKC
SGK1
Cytoskeletal organization
Regulation of actin cytoskeleton
Cell survival
04150 3/12/20
(c) Kanehisa Laboratories

## Future Studies: Application of Methods to the Fraction Library and Compound Identification Processes

Consortium collaborators have created a fraction library on the anchor strain PD1074. Spectral features and overlapping compounds identified through processes described in Chapter 3 (**Figure 3.5** and **Table S3.2**) can be further explored, annotated, and validated through a fraction library process. As the interactions between metabolites and macromolecules are dynamic, these functional interactions are difficult to capture and are largely ignored in metabolomics research[5]. However, NMR is an avenue to study the function of metabolites through the creation and analysis of fraction libraries[5]. Briefly, chromatographic fractionation allows for the collection of fractions at distinct time intervals and concentrates complex biological samples (*e.g.*, from *C. elegans* samples)[5, 6]. Concentrated fractions can bridge the sensitivity gap between NMR and LC-MS, and, importantly, reduce spectral overlap, improving the process of metabolite annotation.

## Future Studies: Implementation of MALDI-MSI to Identify Stage Specific Chemical Features

Finally, matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) is an ionization technique that uses a laser energy absorbing matrix to create ions with minimal fragmentation. MALDI MS has low volume requirements and can quickly produce information rich MS datasets using a 384 well-plate format. As mentioned above, a limitation to the presented methods is the inability to identify *C. elegans* stage-specific chemical features. However, using samples sorted into wells by TOF from the LPFC (Chapter 2), we hypothesize that MALDI-MS can be used to identify size specific metabolites detected by changes in the *C. elegans* cuticle. *C. elegans* shed their cuticles at each life cycle stage (**see Figure 1.1**), thus the MALDI-MS technique can be used to differentiate chemicals between life cycle stages. MALDI imaging MS (MALDI-IMS) is a technique that records mass spectra as a function of position across a biological tissue sample, yielding images of chemical distribution. Methods have been developed to use MALDI-MSI to record chemically relevant data on single intact nematodes[7]. We hypothesize that the stage-specific features identified can then be spatial localization and mapped by MALDI-IMS.

We collected preliminary MALDI-MS (+) data on a PD1074 sample sorted by TOF to see if we could identify separation by *C. elegans* body size (**Figure 4.2**). Preliminary data shows that small regions and large regions do separate well. Additional studies should be performed to determine the stage of worms that are encompassed in each of those TOF regions to see if stages are indeed being separated. Once stage of worms is confirmed by microscopy (Chapter 2), we can assign chemical specific features to a life-cycle stage. This would aid studies presented here and future studies that use a LSCP growth method as we would be able to still take advantage of using a mixed-stage population while also knowing the relative abundance of each stage in a given sample and better understand the chemical dynamic occurring in a population of nematodes.

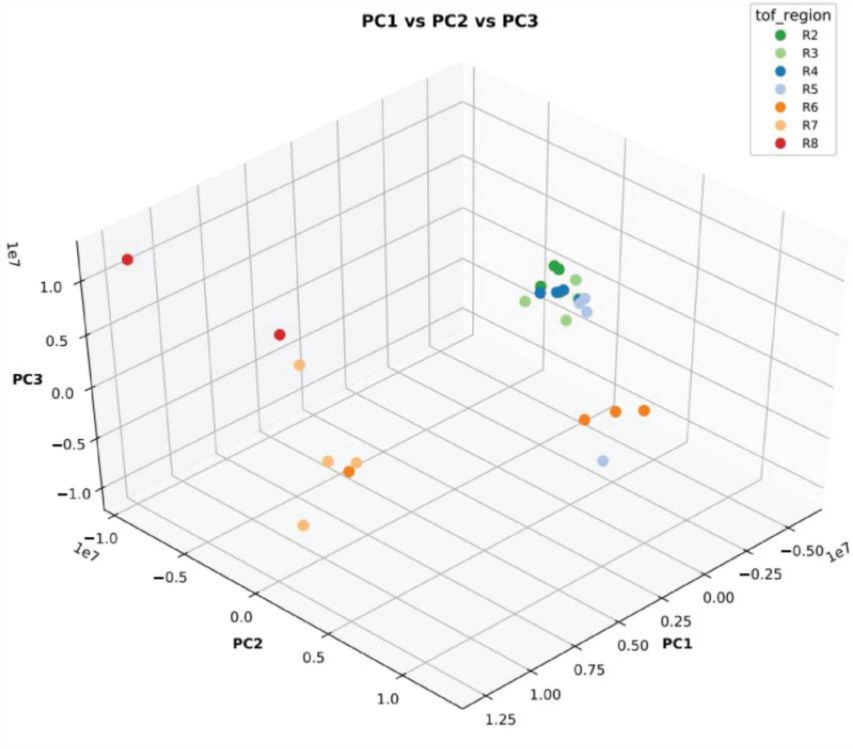


**Figure 4.2 PCA of MALDI-MS chemical features from a PD1074 sample that has been sorted by TOF.** TOF regions are displayed by color. Each region spans a TOF range of 100. Regions here encompass a total TOF range of 50 – 750. Data points of the same color represent technical replicates. Each data point encompasses 20 nematodes. See LPFC methods in Chapter 2 for TOF region details.

In conclusion, better tools and methods continue to improve the process of compound annotation. Work presented in this thesis showcases a unique approach not commonly used or adapted in untargeted metabolomics studies but can be integrated across studies on most organisms.

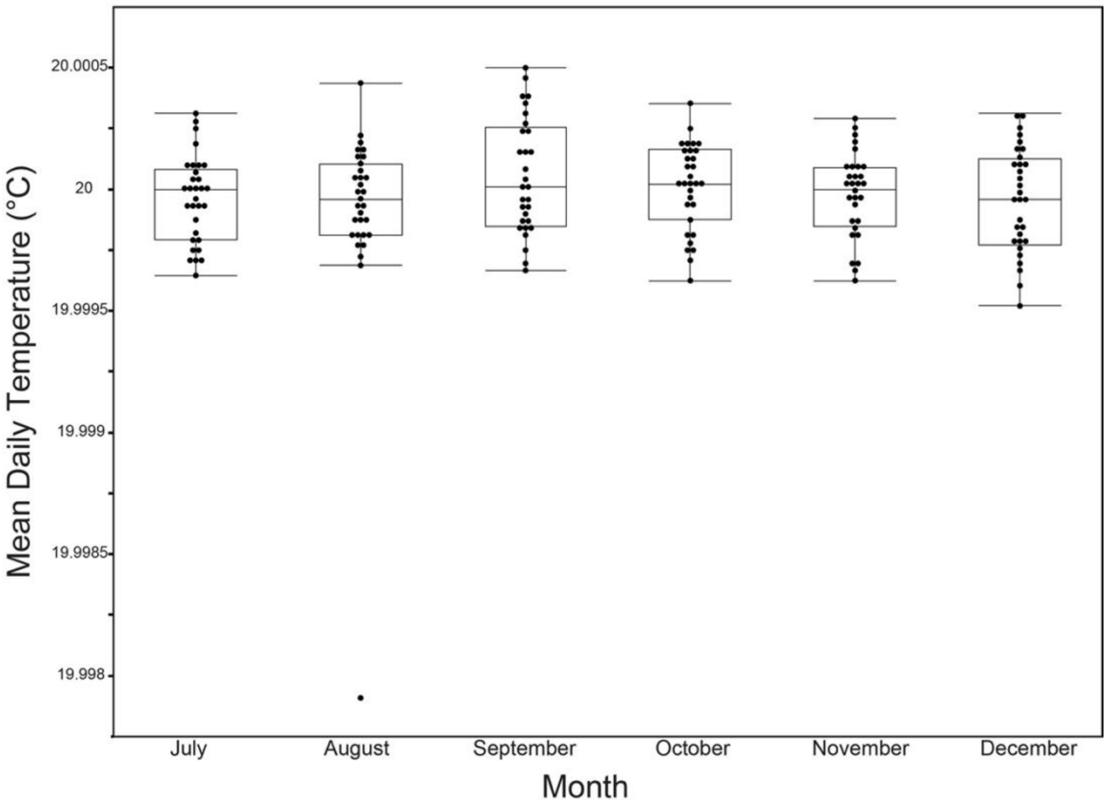
## References

1. Mahieu, N.G., Huang, X., Chen, Y.J. & Patti, G.J. Credentialing features: a platform to benchmark and optimize untargeted metabolomic methods. *Anal Chem* **86**, 9583-9589 (2014).
2. Chan, E.K., Rowe, H.C., Hansen, B.G. & Kliebenstein, D.J. The complex genetic architecture of the metabolome. *PLoS Genet* **6**, e1001198 (2010).
3. Matsuda, F. et al. Dissection of genotype-phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant J* **70**, 624-636 (2012).
4. Wullschleger, S., Loewith, R. & Hall, M.N. TOR signaling in growth and metabolism. *Cell* **124**, 471-484 (2006).
5. Edison, A.S. et al. NMR: Unique Strengths That Enhance Modern Metabolomics Research. *Anal Chem* **93**, 478-499 (2021).

6. Weller, M.G. A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques. *Sensors (Basel)* **12**, 9181-9209 (2012).

7. Menger, R.F., Clendinen, C.S., Searcy, L.A., Edison, A.S., and Yost, R.A. MALDI Mass Spectrometric Imaging of the Nematode Caenorhabditis elegans. *Current Metabolomics* **3**, 130-137 (2015).

# SUPPLEMENTAL MATERIAL CHAPTER 2



**Supplementary Figure 2.1: Mean daily temperature (°C) of growth conditions under which the LSCP was grown and handled.** Reported temperatures of the Controlled Temperature (CT) room were documented and collected throughout the six-month span of sample growth and collection. The average daily temperature is reported here. No significant differences were observed between the temperature in which the LCSP grew during the duration of the project (F(5,24) = 2.59, *p* = 0.0524). The entire temperature difference spanned no greater than 0.003°C throughout the six-month duration of sample growth and generation.

| Gated Region | TOF Distribution (.2us) |
|---|---|
| R2 | 50 - 150 |
| R3 | 150 -250 |
| R4 | 250 - 350 |
| R5 | 350 - 450 |
| R6 | 450 - 550 |
| R7 | 550 - 650 |
| R8 | 650- 750 |
| R9 | 750 - 850 |
| R10 | 850 - 950 |
| R11 | 950 - 1050 |
| R12 | 1050- 1150 |
| R13 | 1150 - 1250 |
| R14 | 1250 - 1350 |
| R15 | 1350- 1450 |
| R16 | 1450 - 1550 |
| R17 | 1550 - 1650 |
| R18 | 1650 - 1750 |
| R19 | 1750 - 1850 |
| R20 | 1850 - 1950 |
| R21 | 1950 - 2050 |

**Supplementary Table 2.1: TOF gated regions used to sort worms into 384-well plates for imaging.** Binned regions were created to span a TOF of 100 across the entire TOF distribution from 50 – 2050. Gated regions can be changed and optimized to suit your needs. Each TOF region that was used for image analysis is displayed in a different color.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | R2 | R2 | R2 | R2 | R3 | R3 | R3 | R3 | R4 | R4 | R4 | R4 | R5 | R5 | R5 | R5 | R6 | R6 | R6 | R6 | R7 | R7 | R7 | R7 |
| B | R8 | R8 | R8 | R8 | R9 | R9 | R9 | R9 | R10 | R10 | R10 | R10 | R11 | R11 | R11 | R11 | R12 | R12 | R12 | R12 | R13 | R13 | R13 | R13 |
| C | R14 | R14 | R14 | R14 | R15 | R15 | R15 | R15 | R16 | R16 | R16 | R16 | R17 | R17 | R17 | R17 | R18 | R18 | R18 | R18 | R19 | R19 | R19 | R19 |
| D | R20 | R20 | R20 | R20 | R21 | R21 | R21 | R21 | | | | | | | | | | | | | | | | |
| E | | | | | | | | | | | | | | | | | | | | | | | | |
| F | | | | | | | | | | | | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | | | | | | | | | | | | |
| H | | | | | | | | | | | | | | | | | | | | | | | | |
| I | | | | | | | | | | | | | | | | | | | | | | | | |
| J | | | | | | | | | | | | | | | | | | | | | | | | |
| K | | | | | | | | | | | | | | | | | | | | | | | | |
| L | | | | | | | | | | | | | | | | | | | | | | | | |
| M | | | | | | | | | | | | | | | | | | | | | | | | |
| N | | | | | | | | | | | | | | | | | | | | | | | | |
| O | | | | | | | | | | | | | | | | | | | | | | | | |
| P | | | | | | | | | | | | | | | | | | | | | | | | |

**Supplementary Table 2.2: 384-well plate template of TOF regions and replicate layout.** Every sample was sorted into a 384-well plate for imaging. Four replicates were created for each region selected for sorting. Gated regions can be changed and optimized to suit your needs. See Supplementary Table 1 for specific gated regions created and used in this protocol. Each TOF region that was used for image analysis is displayed in a different color.

| **Strain** | **Geneotype** | **Source** | **About** |
|---|---|---|---|
| N2 | WT | CeNDR | WT |
| RB2011 | *ugt-62*<br>(ok2663) | CGC | Homozygous. Gene knockout *ugt-62* |
| AUM2073 | *unc-119* | CGC | GSK-3 promotes S-phase entry and progress in germline stem clls to maintain tissue output |
| KJ550 | *aco-1* | CGC | *aco-1* encodes an aconitase that is homologous to mammalian iron regulatory protein-1 (IRP1); *aco-1* activity is required for normal brood sizes and, under iron stress conditions, for normal lifespan and L4-to-adult growth rates |
| DL238 | WT | CeNDR | WT |
| RB2607 | *ugt-49* | CGC | Homozygous. Gene knockout *ugt-49* |
| VC2524 | *gpd-2* | CGC | *gpd-2* encodes one of four glyceraldehyde-3-phosphate dehydrogenases (GAPDHs) |
| CX11314 | WT | CeNDR | WT |
| RB2055 | *ugt-1*<br>(ok2718) | CGC | Homozygous. Gene knockout *ugt-1* |
| RB2347 | *idh-2* | CGC | *idh-2* encodes a predicted mitochondrial isocitrate dehydrogenase; by homology, IDH-2 is predicted to catalyze the formation of alpha-ketoglutarate from isocitrate as part of the citric acid cycle |
| VC1265 | *pyk-1* | CGC | *pyk-1* encodes for one of two pyruvate kinases. Essential for embryonic devleopment |
| RB2550 | *ugt-23*<br>(ok3541) | CGC | Homozygous. Gene knockout *ugt-23* |
| CB4856 | WT | CeNDR | WT |
| VC2512 | *ugt-60* | CGC | Homozygous. Gene knockout *ugt-60* |
| PD1074 | WT | CeNDR | WT |

**Supplementary Table 2.3: *C. elegans* strains used in this protocol contain a mixture of CGC and CeNDR strains.** The strain, genotype, strain source, and details are described in this table.

# SUPPLEMENTAL MATERIAL CHAPTER 3

Supplemental Information

## Reverse Phase (RP) chromatography method

Non-polar extracts were separated using a Vanquish liquid chromatograph (ThermoFisher Scientific), fitted with a ThermoFisher Scientific Accucore™ C30 UPLC RP column (2.1 x 150 mm, 2.6 µm particle size). The compounds were eluted with the following gradient: 60:40 acetonitrile:water (ACN:$H_2O$) with 10 mM ammonium formate and 0.1% formic acid (mobile phase A) and 90:10 isopropanol:acetonitrile with 10 mM ammonium formate and 0.1% formic acid (mobile phase B) using the following gradient program: 0.0 min 20% B; 1.0 min 60% B; 5.0 min 70% B; 5.5 min 85% B; 8.0 min 90% B; 8.2-10.5 min 100% B; 10.7-12.0 min 20% B. A curve 5 value was set for 0.0 minutes, and a curve 6 for the remainder of the gradient. The flow rate was set at 0.400 mL $min^{-1}$. The column temperature was set to 50°C, and the injection volume was 2 µL. The following internal standards were spiked in for RP analysis: 15:0-18:1(d7) PC, 15:0-18:1(d7) PE, 15:0-, 18:1(d7) PS, 15:0-18:1(d7) PG, 15:0-18:1(d7) PI, 18:1(d7) LPC, 18:1(d7) LPE, 18:1(d7) Chol Ester, 15:0-18:1(d7) DG, 15:0-18:1(d7)-15:0 TG, 18:1(d9) SM, and Cholesterol (d7).

## Hydrophilic Interaction Liquid Chromatography (HILIC) method

Polar extracts were separated using a Vanquish liquid chromatograph (ThermoFisher Scientific), fitted with a Waters Acquity UPLC BEH Amide column (2.1 x 150 mm, 1.7 µm particle size). The compounds were eluted with the following gradient: 80:20 water:acetonitrile ($H_2O$:ACN) with 10 mM ammonium formate and 0.1% formic acid (mobile phase A) and 100% ACN with 0.1% formic acid (mobile phase B) using the following gradient program: 0.0-0.5 min 95% B; 8.0-9.4 min 40% B; 9.5-11.0 min 95% B. A curve 5 value was set for 0.0 minutes, a curve 6 at 0.5 min, curve 7 at 8.0 min, and a curve 6 for the remainder of the gradient. The flow rate was set at 0.400 mL min$^{-1}$. The column temperature was set to 40°C, and the injection volume was 2 µL.

### Mass spectrometer settings and methods

An Orbitrap ID-X Tribrid mass spectrometer (ThermoFisher Scientific) equipped with a HESI ion source was used for all mass spectrometry data collection. For HILIC analysis the mass spectrometer was run in full MS mode at a resolution of 240,000 (FWHM at *m/z* 200) for the duration of the chromatographic gradient. A normalized automatic gain control (AGC) target of 100% was set with a maximum injection time of 100 ms. A tune file with the following source conditions was used for positive and negative mode: spray voltage (+) 3500, spray volage (-) 2500, vaporizer temperature: 275 ºC, sheath gas: 40, aux gas: 8, sweep gas: 1, and S-Lens RF level: 60%. Scan range covered *m/z* 70-1050. Calibration was conducted using ThermoFisher Pierce™ Negative Ion Calibration Solution and Pierce™ LTQ Velos ESI Positive Ion Calibration Solution prior to collecting negative and positive mode data, respectively.

Identical parameters were used for reverse-phase analysis with the following exceptions: spray voltage (+) 3500, spray volage (-) 2800, maximum injection time of 200ms, vaporizer temperature: 425 ºC, sheath gas: 60, aux gas: 18, sweep gas: 4, and a scan range of *m/z* 150-2000.

### Growth of *C. elegans* using the LSCP method yields on average 2.4 million mixed-stage worms per sample

Growth of *C. elegans* using the LSCP method generates large mixed-stage populations of *C. elegans* with minor handling and manipulation of the animals, ideal for metabolomics experiments. Population dynamics depend on the strain’s behavior (*i.e.*, burrowing strains tend to have lower worm recovery) and growth success (*i.e.*, contamination). The LSCP method yields population sizes from approximately 94,500 to 9,290,000. The mean population size within the anchor strain, PD1074, and across strains is approximately 2.4 million worms (**Figure S3.2**). PD1074 LSCPs take between 10 – 14 days to grow to a full mixed-stage population. The mean growth time for PD1074 is ten days. The slowest growing strain grows for a maximum of 20 days, and the fastest growing strain for a minimum of 10 days (**Figure S3.2**). Strain-dependent variation in growth rate is to be expected because of the wide range of strains and traits included in this study. Notably, growth time is not indicative of the final population size (**Figure S3.2**). Additional information on *C. elegans* growth and population dynamics has been presented previously[1].

### Variation between PD1074 and N2

The anchor strain PD1074, obtained from CeNDR, is a variant of the traditionally used laboratory-adapted N2 Bristol strain. PD1074 is used in this study as it is a variant of N2 with a trackable evolutionary history[2]. Here, we include N2, obtained from CeNDR, as one of our natural strains as an additional way to validate our methods and processes. Phenotypically, compared to PD1074, N2 is not significantly different in the amount of time needed to cultivate a population or the resulting population size (**Figure S3.2**). Metabolically, N2 is most similar to PD1074 showing 16 total feature differences in the RP LC-MS (+) data, six feature differences in the RP LC-MS (-), and seven differences in the HILIC LC-MS (+) data (**Table 3.1**), indicating little difference between N2 and PD1074. In the non-polar NMR data, N2 has three significant feature differences and seven significant feature differences in the polar NMR data (**Table 3.1**), having many fewer spectral features differentiating from PD1074. Both the phenotypic and metabolic data showcase that these strains are very similar. In contrast, in the RP LC-MS modes, N2 has up to a 50-fold difference from the other strains (**Figure S3.5**).

### Model comparison for batch effect corrections

Meta-analysis is a statistical analysis that combines summary statistics instead of an analysis of individual samples[3, 4]. A meta-analysis can be used to account for the batch effects in untargeted metabolomics. In a traditional meta-analysis, an effect size is calculated for each study and then combined and weighed by the individual study sample sizes [5, 6]. As meta-analysis is a promising approach to address the complicated variance structure in a straightforward way, and has been shown to be equivalent to more complex linear model approaches on individual data on larger sample sizes[3], we demonstrate that meta-analysis, even with relatively small sample sizes per group (n=6 for test samples), is very similar to a mixed effects model with the variance modeled appropriately[7]. This study design makes it possible to apply both approaches and compare inferences directly. In other more complex situations modeling the individual-level data can be very challenging. The meta-analysis model was formally compared to the linear model in several formulations, and in each case, the meta-strain model has similar statistical inferences as the linear model, consistent with the literature on larger sample sizes[4]. To mirror the comparison in Lin & Zeng (2010)[3], a fixed-effect model or random effect model can be used to infer a true biological effect from the effects estimated from individual batches. The fixed-effect model assumes that the underlying effect sizes in all batches are identical and treats the inverse of the variance of each batch effect size as the weight to account for all variabilities in each batch.

Contrary to the fixed-effect model, a random-effect model assumes that the underlying effect size in all batches is similar but not identical. As there are just a few batches in this study, a fixed-effect model is applied. We use the fixed-effect model effect size estimate:

$$\bar{\theta}_w = \frac{\sum_{i=1}^{k} w_i \, \theta_i}{\sum_{i=1}^{k} w_i}$$

where $w_i$ is a weight calculated as the inverse of variance for the effect size in batch $i$, and $\bar{\theta}_w$ is the effect size of interest inferred from the individual effect sizes in batches. For the ANOVA comparison the effect size is calculated as:

$$\bar{\theta} = \frac{lsmean_{test} - \ lsmean_{PD1074}}{sd}$$

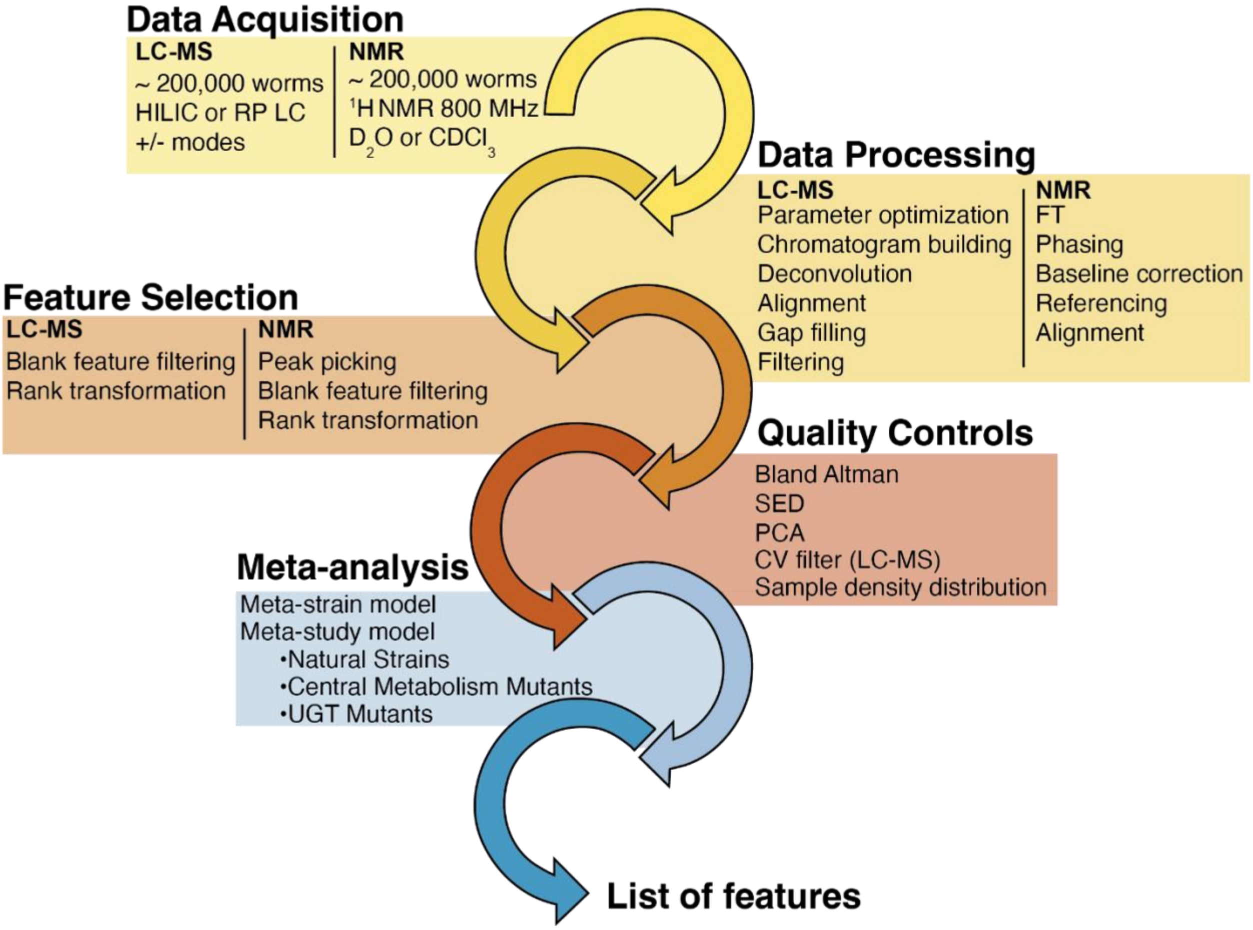
$$sd = \ \sqrt{n} * se$$

To illustrate the comparability between these approaches, we compare the linear model by batch $l$, where strain $i$ is the independent variable and ion signal for each spectral feature $m$, and test replicate *j* is the response variable*:*

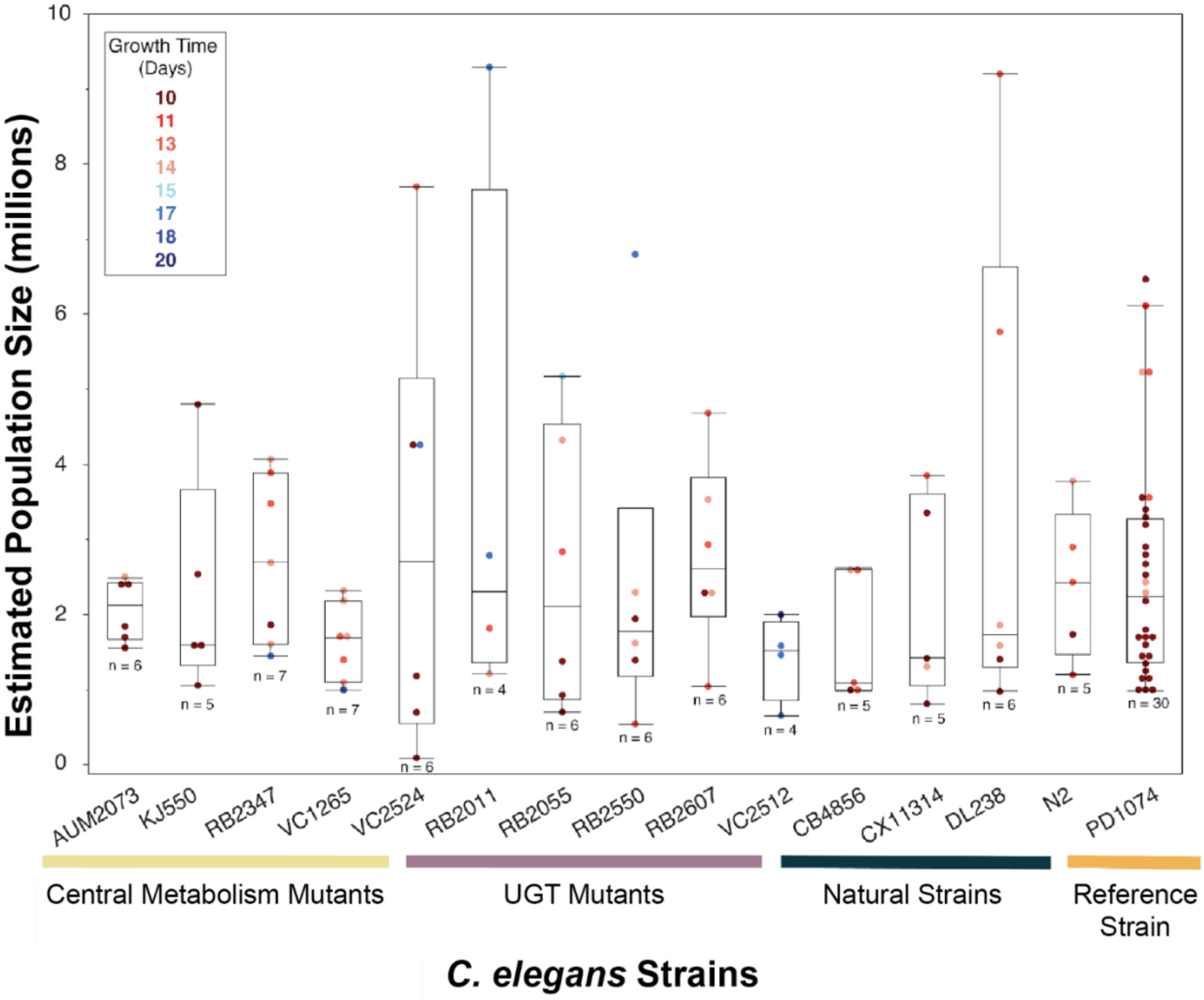
$$Y_{mlij} = \mu + batch_l + \ strain_i + e_{lij}$$

The example depicted is typical, with the number of features detected in the linear model consistently slightly higher than the meta-analysis[4]. This is consistent with the slight benefit in the degrees of freedom estimates from the combined model on individual data. The effect size estimates are consistent. Batch effects across experiments are an enormous problem in metabolomics experiments, and the inability to adequately address this in a mixed model analyses is a well-known problem[4, 7]. Meta-analysis presents a straightforward way of comparing data across experiments.

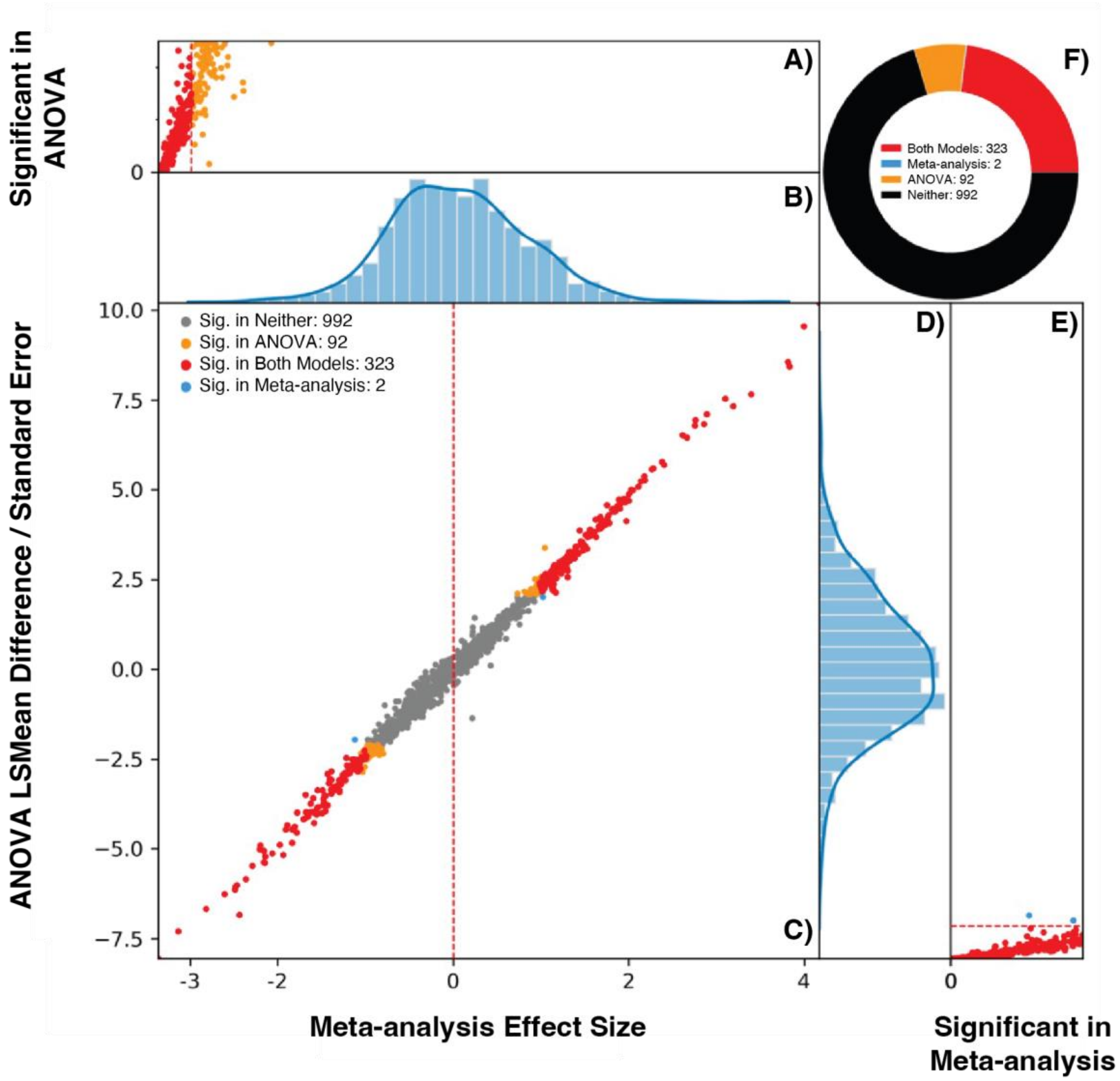
Supplemental Figures



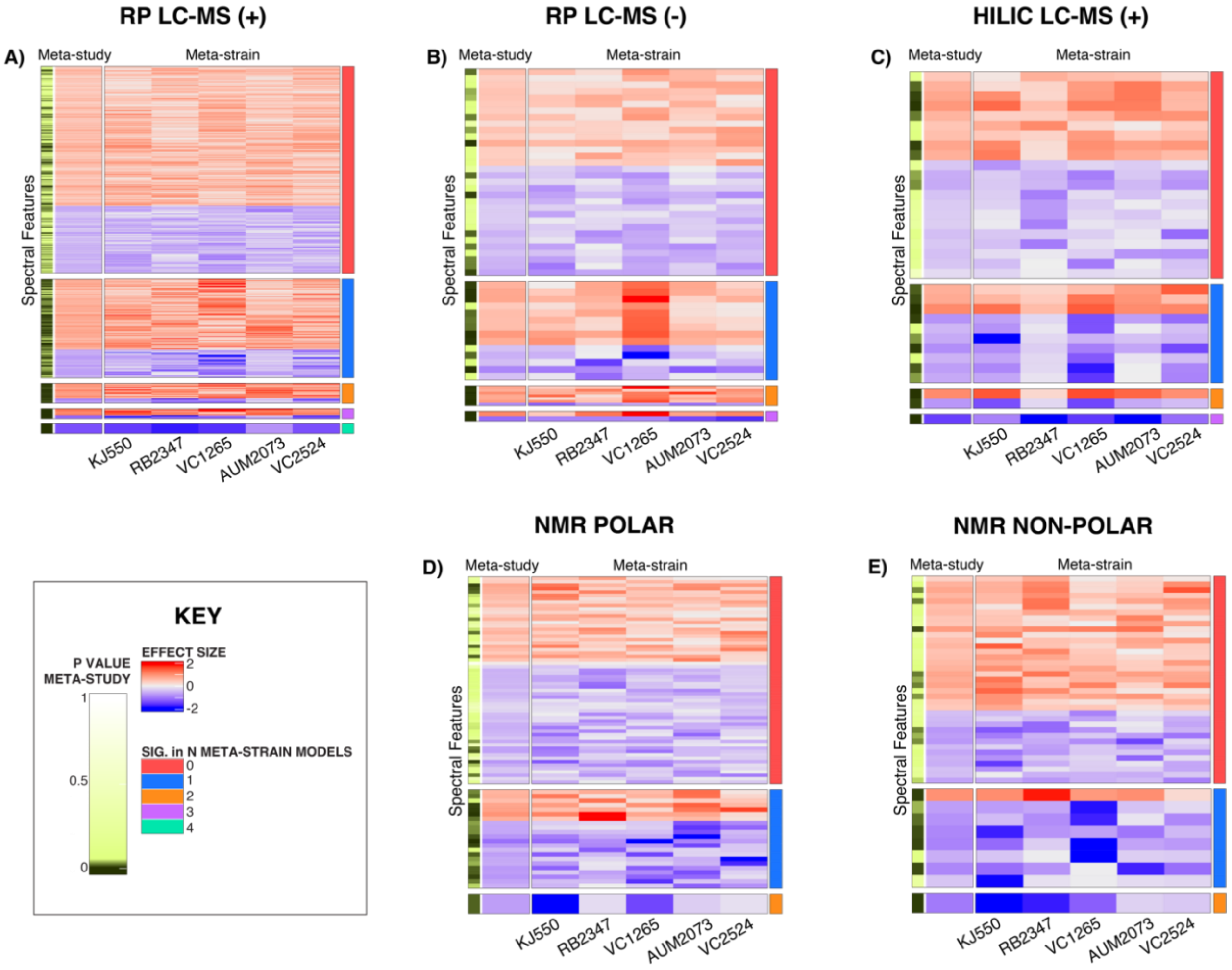
**Supplemental Figure 3.1. Unified workflow for spectral feature selection across platforms.** Steps and software listed to obtain spectral features. LC-MS-specific steps are listed on the left side of each process. NMR-specific steps are denoted on the right side of each process. When neither NMR nor LC-MS is indicated, the same steps are performed on data from both analytical platforms (see Methods for detailed documentation).



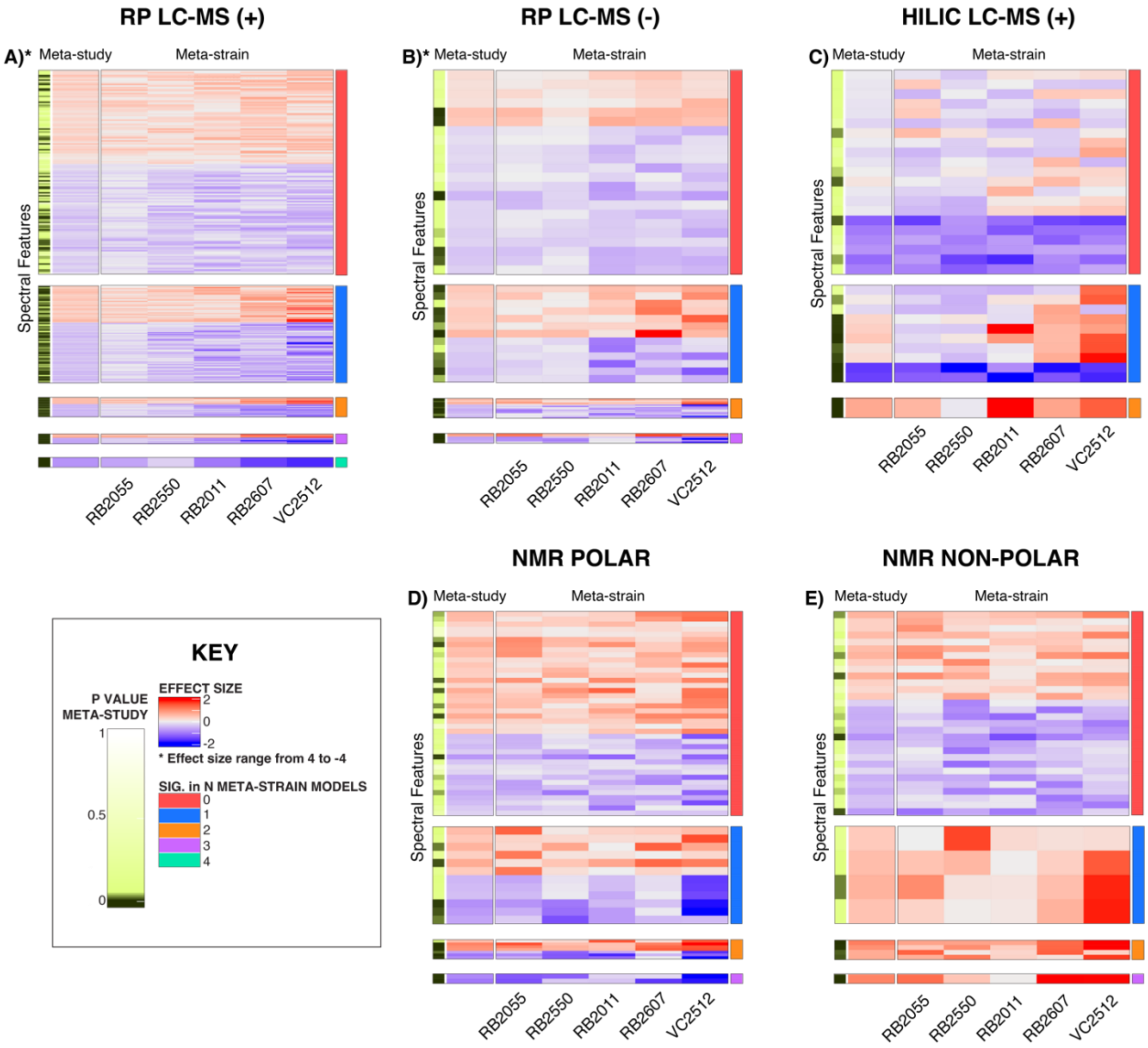
**Supplemental Figure 3.2**. **LSCP method generates, on average, a population of 2.4 million mixed-stage *C. elegans* populations.** The LSCP yields population sizes in the smallest population growths at around 94,500 worms and the biggest population growths at around 9,290,000 worms. The mean population size across all strains was 2.4 million worms. Bars underneath *C. elegans* strain names indicate each strains study. Comparisons of population size for all pairs using Tukey's HSD test were performed. No significant differences are observed between estimated population sizes across *C. elegans* strains. Colored data points indicate the growth time (days) to generate a given LSCP sample.



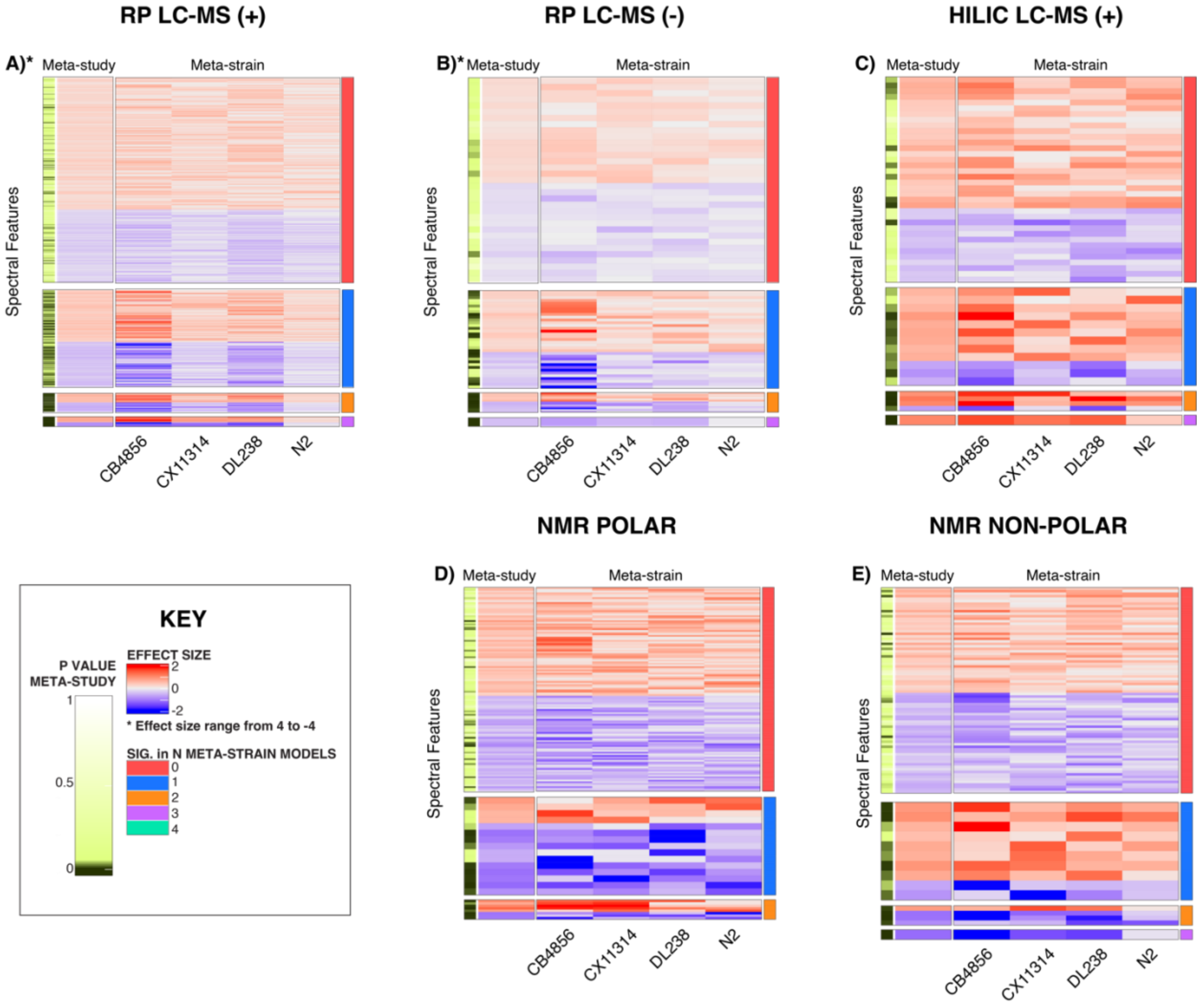
**Supplemental Figure 3.3: Comparing the inferences from a linear model and meta-analysis (meta-strain) using the test genotype, VC1265.** There are six panels in the plot. **(A)** Displays the p-values where the y-axis is from 0-1 and represents the significant results (p-value < 0.05) in the linear model. **(B)** displays the effect size distribution for effect sizes in the x-axis for the scatter plot. **(C)**  is a scatter plot, where the x-axis is the effect sizes calculated by the meta-anlaysis model and the y-axis is the *lsmean* difference calculation. Each point is one spectral feature. **(D)** displays the effect size distribution for effect sizes in the y-axis for the scatter plot. Point colors represent significance of the test of the null hypothesis where the mean peak height for VC1265 is not different from the mean peak height of PD1074 with a nominal threshold p-value < 0.05. Red points are significant in both models. Orange points are significant in the linear model. Blue points are significant results in the meta-analysis model. Grey points indicates that spectral feature is not significant in either model. Red lines in **(A)** and **(E)** are significance thresholds for the nominal p-value < 0.05 on the other test. **(F)** summarizes the results.



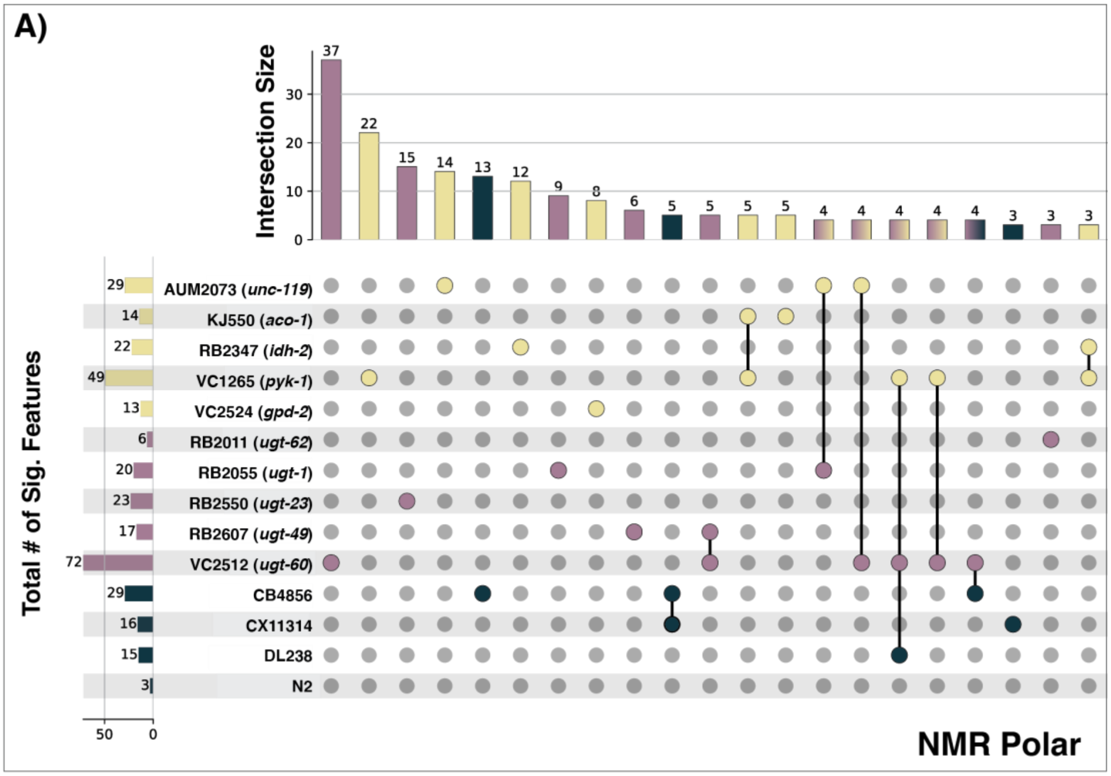
**Supplemental Figure 3.4. Heatmaps of significant spectral features identified via the meta-strain and meta-study models in the CM mutants study. (A)** RP LC-MS positive mode **(B)** RP LC-MS negative mode **(C)** HILIC LC-MS positive mode **(D)** NMR polar and **(E)** NMR non-polar. The first two columns on the left-hand side pertain to the meta-study model results. The yellow and black bar highlights the significant features found in the meta-study model, followed by the meta-study heatmap. The following five columns within the heatmap compare the spectral features for a given strain in the meta-strain model. For each heatmap, each row represents a spectral feature with an effect size that is consistently higher or lower relative to PD1074 in that study (meta-strain) or across strains (meta-study). The effect sizes range from (2 to -2). Positive effect sizes (*i.e.*, the strain had a higher peak at that given metabolic feature than the anchor PD1074) are displayed in red. Negative effect sizes (*i.e.*, the anchor PD1074 had a higher peak at that given metabolic feature than the test strain) are displayed in blue. The right-hand column indicates the number of strains in which a given spectral feature is statistically significant in the meta-strain model.

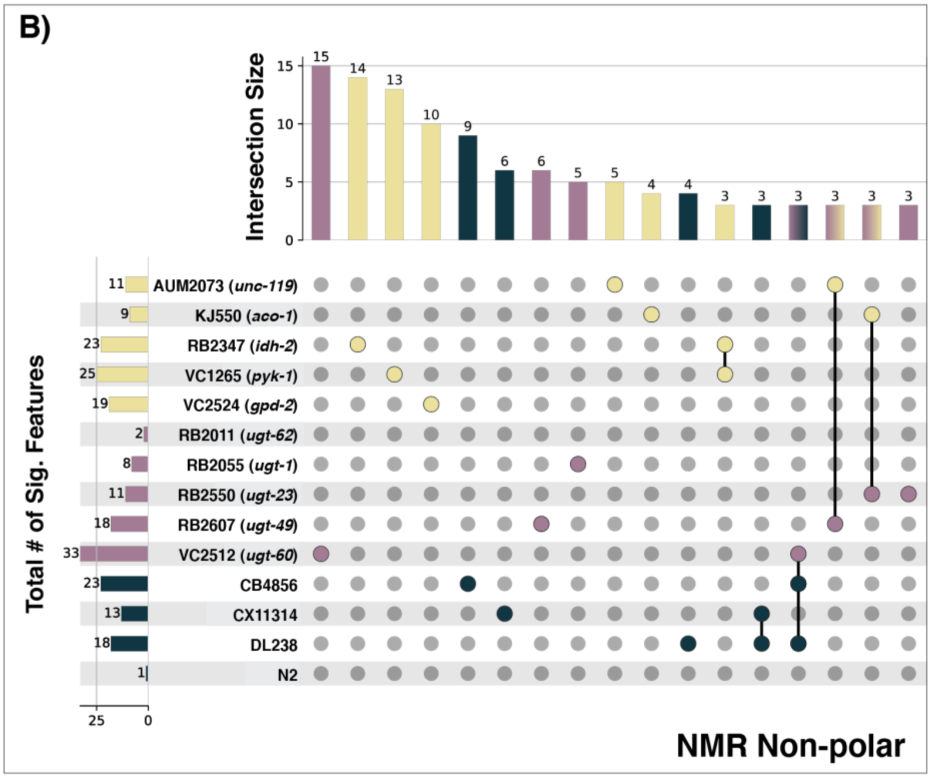


**Supplemental Figure 3.5. Heatmaps of significant spectral features identified via the meta-strain and meta-study models in the UGT study. (A)** RP LC-MS positive mode **(B)** RP LC-MS negative mode **(C)** HILIC LC-MS positive mode **(D)** NMR polar and **(E)** NMR non-polar. The first two columns on the left-hand side pertain to the meta-study model results. The yellow and black bar highlights the significant features found in the meta-study model, followed by the meta-study heatmap. The following five columns within the heatmap compare the spectral features for a given strain in the meta-strain model. For each heatmap, each row represents a spectral feature with an effect size that is consistently higher or lower relative to PD1074 in that study (meta-strain) or across strains (meta-study). The effect sizes range from (2 to -2). Modes denoted with an asterisk have effect sizes that range from (4 to -4). Positive effect sizes (*i.e.*, the strain had a higher peak at that given metabolic feature than the anchor PD1074) are displayed in red. Negative effect sizes (*i.e.*, the anchor PD1074 had a higher peak at that given metabolic feature than the test strain) are displayed in blue. The right-hand column indicates the number of strains in which a given spectral feature is statistically significant in the meta-strain model.

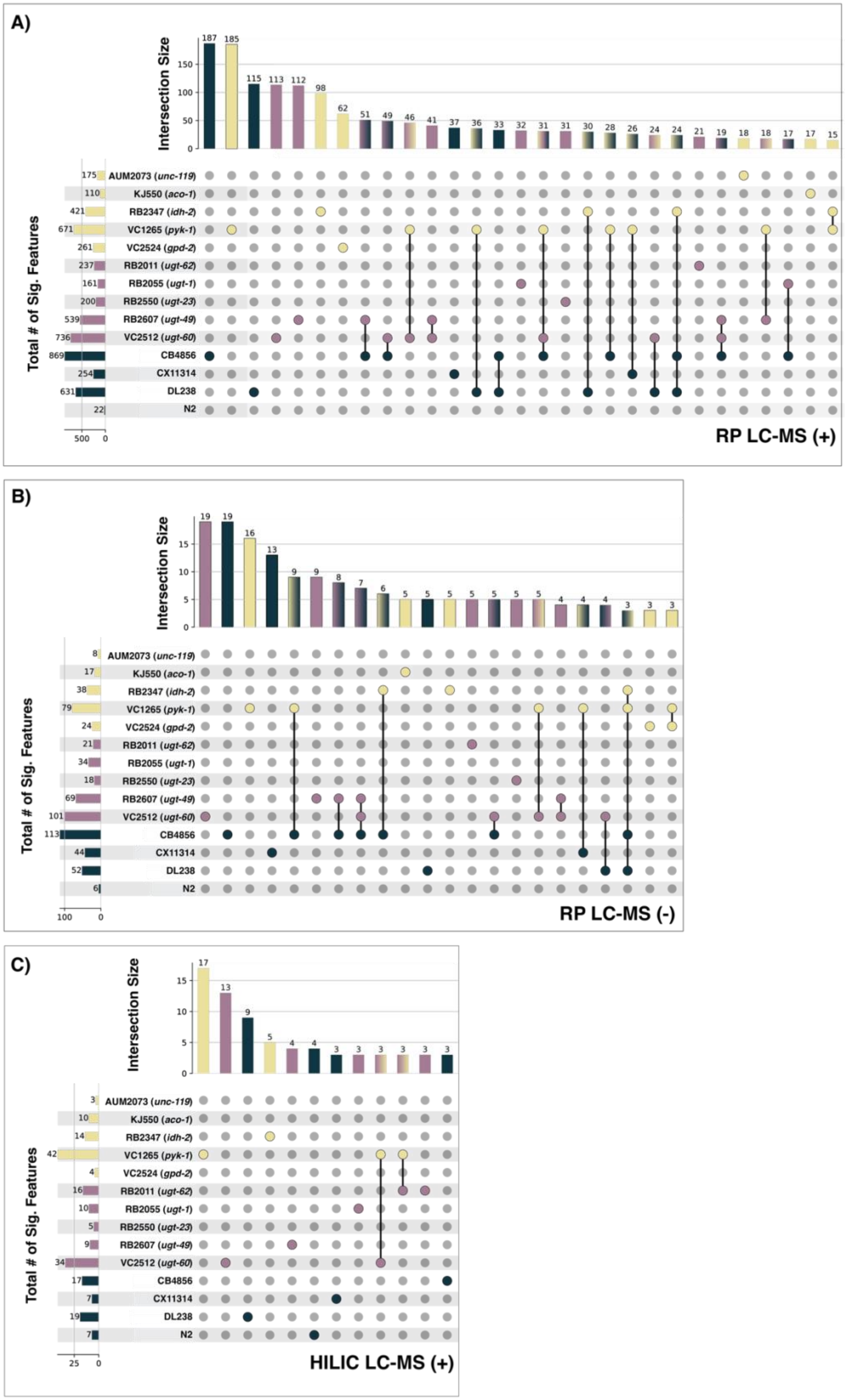


**Supplemental Figure 3.6. Heatmaps of significant spectral features identified via the meta-strain and meta-study models in the NS study. (A)** RP LC-MS positive mode **(B)** RP LC-MS negative mode **(C)** HILIC LC-MS positive mode **(D)** NMR polar and **(E)** NMR non-polar. The first two columns on the left-hand side pertain to the meta-study model results. The yellow and black bar highlights the significant features found in the meta-study model, followed by the meta-study heatmap. The following five columns within the heatmap compare the spectral features for a given strain in the meta-strain model. For each heatmap, each row represents a spectral feature with an effect size that is consistently higher or lower relative to PD1074 in that study (meta-strain) or across strains (meta-study). The effect sizes range from (2 to -2). Modes denoted with an asterisk have effect sizes that range from (4 to -4). Positive effect sizes (*i.e.*, the strain had a higher peak at that given metabolic feature than the anchor PD1074) are displayed in red. Negative effect sizes (*i.e.*, the anchor PD1074 had a higher peak at that given metabolic feature than the test strain) are displayed in blue. The right-hand column indicates the number of strains in which a given spectral feature is statistically significant in the meta- strain model.


A)
Intersection Size
0
10
20
30
37
22
15
14
13
12
9
8
6
5
5
5
5
4
4
4
4
4
3
3
3
Total # of Sig. Features
29 AUM2073 (unc-119)
14 KJ550 (aco-1)
22 RB2347 (idh-2)
49 VC1265 (pyk-1)
13 VC2524 (gpd-2)
6 RB2011 (ugt-62)
20 RB2055 (ugt-1)
23 RB2550 (ugt-23)
17 RB2607 (ugt-49)
72 VC2512 (ugt-60)
29 CB4856
16 CX11314
15 DL238
3 N2
50
0
NMR Polar

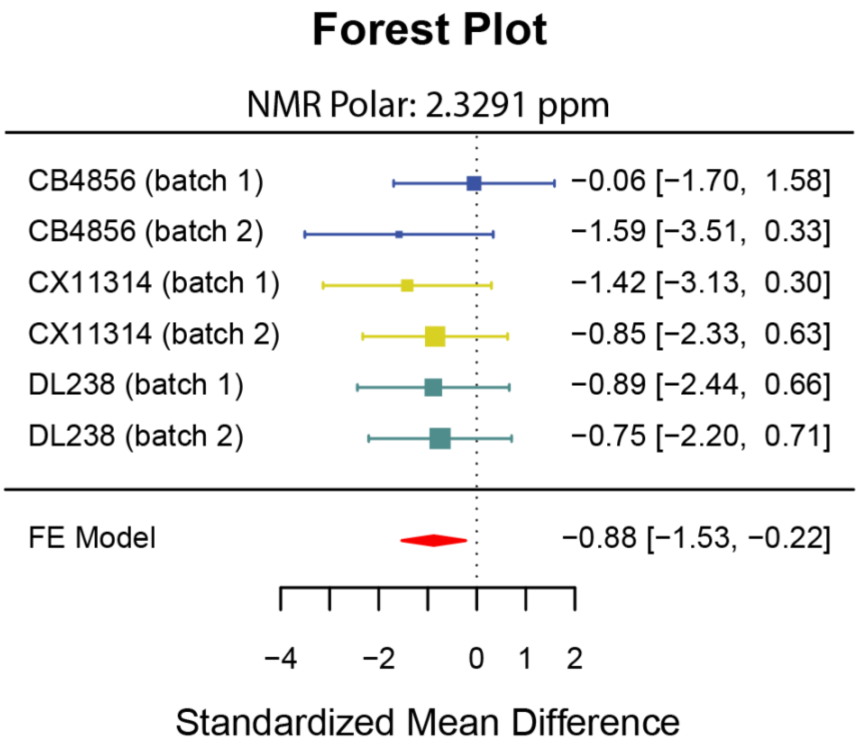

B)
Intersection Size
0
5
10
15
15
14
13
10
9
6
6
5
5
4
4
3
3
3
3
3
3
Total # of Sig. Features
11 AUM2073 (unc-119)
9 KJ550 (aco-1)
23 RB2347 (idh-2)
25 VC1265 (pyk-1)
19 VC2524 (gpd-2)
2 RB2011 (ugt-62)
8 RB2055 (ugt-1)
11 RB2550 (ugt-23)
18 RB2607 (ugt-49)
33 VC2512 (ugt-60)
23 CB4856
13 CX11314
18 DL238
1 N2
25
0
NMR Non-polar

**Supplemental Figure 3.7. Significant features found via the meta-strain model across the three study groups in NMR.** (**A)** NMR polar (**B)** NMR non-polar. Fourteen strains in the CM mutants (yellow), UGT mutants (purple), and NS (green) are displayed. Horizontal bar plots sum the total number of significant features for a strain. Vertical bar plots sum significant feature interactions within and across strains. Significant feature connections below three are not displayed.



A)
Intersection Size
0
50
100
150
187
185
115
113
112
98
62
51
49
46
41
37
36
33
32
31
31
30
28
26
24
24
21
19
18
18
17
17
15
Total # of Sig. Features
175 AUM2073 (unc-119)
110 KJ550 (aco-1)
421 RB2347 (idh-2)
671 VC1265 (pyk-1)
261 VC2524 (gpd-2)
237 RB2011 (ugt-62)
161 RB2055 (ugt-1)
200 RB2550 (ugt-23)
539 RB2607 (ugt-49)
736 VC2512 (ugt-60)
869 CB4856
254 CX11314
631 DL238
22 N2
500
0
RP LC-MS (+)
B)
Intersection Size
0
5
10
15
19
19
16
13
9
9
8
7
6
5
5
5
5
5
5
5
4
4
4
3
3
3
Total # of Sig. Features
8 AUM2073 (unc-119)
17 KJ550 (aco-1)
38 RB2347 (idh-2)
79 VC1265 (pyk-1)
24 VC2524 (gpd-2)
21 RB2011 (ugt-62)
34 RB2055 (ugt-1)
18 RB2550 (ugt-23)
69 RB2607 (ugt-49)
101 VC2512 (ugt-60)
113 CB4856
44 CX11314
52 DL238
6 N2
100
0
RP LC-MS (-)
C)
Intersection Size
0
5
10
15
17
13
9
5
4
4
3
3
3
3
3
3
Total # of Sig. Features
3 AUM2073 (unc-119)
10 KJ550 (aco-1)
14 RB2347 (idh-2)
42 VC1265 (pyk-1)
4 VC2524 (gpd-2)
16 RB2011 (ugt-62)
10 RB2055 (ugt-1)
5 RB2550 (ugt-23)
9 RB2607 (ugt-49)
34 VC2512 (ugt-60)
17 CB4856
7 CX11314
19 DL238
7 N2
25
0
HILIC LC-MS (+)

**Supplemental Figure 3.8. Significant features found via the meta-strain model across the three study groups in LC-MS. (A)** RP LC-MS positive mode (**B)** RP LC-MS negative mode (**C)** HILIC LC-MS positive mode. Fourteen strains in the CM mutants (yellow), UGT mutants (purple), and NS (green) are displayed. Horizontal bar plots sum the total number of significant features for a strain. Vertical bar plots sum significant feature interactions within and across strains. Significant feature connections below three are not displayed for all modes, except for the RP LC-MS positive mode, where connections below 15 are not displayed.



**Supplemental Figure 3.9. Forest plot comparing the meta-strain and meta-study models in the NS for the NMR polar ppm 2.3291.** The meta-strain model for each NS in each batch is displayed, where the area of each square is proportional to the study's weight in the meta-analysis with confidence intervals (CI) represented by whiskers. The dashed vertical line represents the overall measure of effect. The right-hand column is the measure of effect (odds ratio) for each study. The meta-study model (FE model) is represented by the red diamond, where the lateral points indicate CI for the estimate.

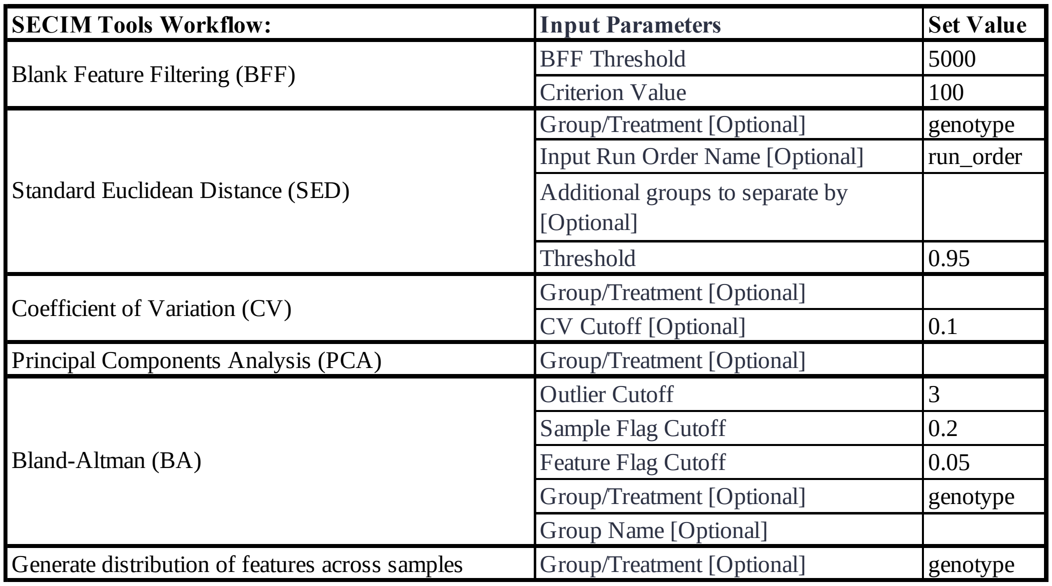
## Supplemental Tables

| Strain | Genotype | Source | About | Study Group |
|---|---|---|---|---|
| PD1074 | wild type | CeNDR | wild type | Anchor Strain |
| AUM2073 | *unc-119; vizSi34 II; unc-119(ed3) III.* | CGC | GSK-3 promotes S-phase entry and progress in germline cells to maintain tissue output | Central Metabolism Mutants |
| KJ550 | *aco-1 (jh131) X.* | | Exhibits aconitate hydratase activity. Is involved in tricarboxylic acid metabolic process. Localizes to cytosol. Is expressed in seam cell. Is an ortholog of human ACO1 (aconitase 1). | |
| RB2128 | *idh-1; F59B8.2(ok2832) IV.* | | *idh-1 encodes a predicted cytosclic isocitrate dehydrogenase; by homology, IDH-1 is the formation of alpha-ketoglutarate from isocitrate, Is predicted to have isocitrate dehydrogenase (NADP+) activity and metal ion binding activity. Is expressed in tail. Is an ortholog of human IDH1 (isocitrate dehydrogenase (NADP(+)) 1).* | |
| RB2347 | *idh-2(ok3183) X.* | | Is predicted to have isocitrate dehydrogenase (NADP+) activity and metal ion binding activity. Human ortholog(s) of this gene are implicated in D-2-hydroxyglutaric aciduria 2. Is an ortholog of human IDH2 (isocitrate dehydrogenase (NADP(+)) 2). | |
| VC1265 | *pyk-1; F25H5.3(ok1754) I.* | | *pyk-1 encodes for one of two pyruvate kinases. essential for embryonic development, Is predicted to have kinase activity; metal ion binding activity; and pyruvate kinase activity. Human ortholog(s) of this gene are implicated in pyruvate kinase deficiency of red cells. Is an ortholog of human PKLR (pyruvate kinase L/R) and PKM (pyruvate kinase M1/2).* | |
| VC2524 | *gpd-2 (ok3243) X.* | | Is predicted to have NAD binding activity; NADP binding activity; and glyceraldehyde-3-phosphate dehydrogenase (NAD+) (phosphorylating) activity. Is expressed in several structures, including AB; Psub1; and head. Is an ortholog of human GAPDH (glyceraldehyde-3-phosphate dehydrogenase). | |
| RB2011 | *ugt-62 (ok2663)* | CGC | homozygous. gene knockout ugt-62, Is predicted to have glucuronosyltransferase activity. Human ortholog(s) of this gene are implicated in Crigler-Najjar syndrome and Gilbert syndrome. Is an ortholog of several human genes including UGT1A6 (UDP glucuronosyltransferase family 1 member A6); UGT1A8 (UDP glucuronosyltransferase family 1 member A8); and UGT1A9 (UDP glucuronosyltransferase family 1 member A9). | UGT Mutants |
| RB2055 | *ugt-1 (ok2718) V.* | | homozygous. gene knockout ugt-1, Is predicted to have UDP-glycosyltransferase activity. Human ortholog(s) of this gene are implicated in Crigler-Najjar syndrome and Gilbert syndrome. Is an ortholog of several human genes including UGT1A4 (UDP glucuronosyltransferase family 1 member A4); UGT1A8 (UDP glucuronosyltransferase family 1 member A8); and UGT1A9 (UDP glucuronosyltransferase family 1 member A9). | |
| RB2550 | *ugt-23 (ok3541) X.* | | homozygous. gene knockout ugt-23, Is predicted to have glucuronosyltransferase activity. Is involved in gastrulation. Human ortholog(s) of this gene are implicated in Crigler-Najjar syndrome and Gilbert syndrome. Is an ortholog of human UGT3A1 (UDP glycosyltransferase family 3 member A1) and UGT3A2 (UDP glycosyltransferase family 3 member A2). | |
| RB2607 | *ugt-49(ok3633) V.* | | homozygous. gene knockout ugt-49, Is predicted to have glucuronosyltransferase activity. Human ortholog(s) of this gene are implicated in Crigler-Najjar syndrome and Gilbert syndrome. Is an ortholog of several human genes including UGT2A3 (UDP glucuronosyltransferase family 2 member A3); UGT2B10 (UDP glucuronosyltransferase family 2 member B10); and UGT2B11 (UDP glucuronosyltransferase family 2 member B11). | |
| VC2512 | *ugt-60(ok3248) III/hT2 [bli-4(e937) let-?(q782) qIs48] (I;III).* | | homozygous. gene knockout ugt-60, Is predicted to have glucuronosyltransferase activity. Human ortholog(s) of this gene are implicated in Crigler-Najjar syndrome and Gilbert syndrome. Is an ortholog of human UGT2B7 (UDP glucuronosyltransferase family 2 member B7). | |
| N2 | wild type | CeNDR | wild type | Natural Strains |
| DL238 | | | | |
| CX11314 | | | | |
| CB4856 | | | | |

**Supplemental Table 3.1.** ***C. elegans*** **genotypes used in study.**

| Feature ID | Putative Annotation | Strain | Confidence score |
|---|---|---|---|
| ppm_1_4761 | Alanine | VC1265 | 4 |
| ppm_1_4896 | | | |
| ppm_2_5027 | Alpha-ketoglutaricacid | RB2347 | 3 |
| ppm_2_237 | Aminoadipic Acid* | RB2347 | 2 |
| ppm_3_2462 | Arginine | VC1265 | 4 |
| ppm_3_2547 | | VC2524 | |
| ppm_1_6361 | Arginine* | AUM2073 | 4 |
| ppm_1_6525 | | | |
| ppm_3_1839 | Beta-Alanine | RB2347 | 4 |
| ppm_3_2672 | Betaine | VC1265 | 4 |
| ppm_0_86209 | CH3-Lipoprotein | AUM2073 | 1 |
| ppm_2_5584 | Citrate (carbon shift outside threshold criteria) | AUM2073 | 2 |
| ppm_2_5788 | | KJ550 | |
| ppm_2_5866 | | AUM2073 | |
| ppm_2_0689 | Glutamic Acid | VC2524 | 4 |
| ppm_2_0778 | | | |
| ppm_2_1094 | | KJ550 | |
| ppm_2_1186 | | | |
| ppm_2_3474 | Glutamic Acid* | RB2347 | 4 |
| ppm_2_3585 | | | |
| ppm_2_4498 | Glutamine-Unknown* | AUM2073 | 4 |
| ppm_2_4587 | | | |
| ppm_2_4673 | | | |
| ppm_2_4774 | | | |
| ppm_2_1285 | Glutamine/Glutamic Acid* | RB2347 | 3 |
| ppm_3_5782 | Glycerol* | AUM2073 | 3 |
| ppm_3_9436 | Glycero-phosphocholine | RB2347 | 3 |
| ppm_1_3238 | Lactic Acid | VC1265 | 4 |
| ppm_1_3396 | | | |
| ppm_1_6686 | Leucine* | AUM2073 | 4 |
| ppm_1_755 | Lysine | KJ550 | 4 |
| ppm_3_0189 | | VC1265 | |
| ppm_3_029 | | | |
| ppm_3_0393 | | | |
| ppm_1_9208 | Lysine-Acetic Acid-Arginine* | AUM2073 | 3 |
| ppm_1_6832 | Lysine-Arginine* | KJ550 | 4 |
| ppm_1_7156 | | VC1265 | |
| ppm_1_7348 | | | |
| ppm_7_3231 | Phenylalanine | VC1265 | 4 |
| ppm_7_338 | | | |
| ppm_7_4387 | | RB2347 | |
| ppm_3_1231 | Phenylalanine* | RB2347 | 4 |
| ppm_7_3876 | | | |
| ppm_2_007 | Proline (low level) | AUM2073 | 3 |
| ppm_2_0306 | | VC2524 | |
| ppm_3_4544 | Trehalose | RB2347 | 4 |
| ppm_3_4704 | | VC1265 | |
| ppm_5_188 | | RB2347 | |
| ppm_5_2069 | | KJ550 | |
| ppm_3_6408 | Trehalose-Glycerol* | VC1265 | 4 |
| ppm_3_6575 | | RB2347 | |
| ppm_3_8518 | Trehalose* | KJ550 | 4 |
| ppm_3_8699 | | | |
| ppm_1_591 | Unknown 1 | AUM2073 | n/a |
| ppm_1_5981 | | | |
| ppm_1_6058 | Unknown 2 | KJ550 | n/a |
| ppm_3_2244 | Unknown 3 | VC1265 | n/a |
| ppm_8_1863 | Unknown 4 | AUM2073 | n/a |
| ppm_0_94007 | Unknown 5* | AUM2073 | n/a |
| ppm_1_9545 | Unknown 6* | KJ550 | n/a |
| ppm_2_2768 | Unknown 7* | RB2347 | n/a |
| ppm_2_2952 | Unknown 8* | RB2347 | n/a |
| ppm_3_7919 | Unknown 9* | RB2347 | n/a |

**Supplemental Table 3.2.** List of significant features and respective annotation. FeatureID indicates the chemical shift of each feature. Putative annotation indicates compound name as obtained from COLMAR. Strain indicates the corresponding mutant for each feature deemed significantly different (p-value < 0.005). Confidence score defined as 1 to 5, with 5 being the highest. The scale is defined as follows: (1) putatively characterized compound classes or annotated compounds, (2) matched to literature and/or 1D spectra of a reference standard, (3) matched to HSQC, (4) matched to HSQC and validated by HSQC–TOCSY and TOCSY (COLMARm), and (5) validated by spiking the authentic compound into the sample$^{8}$. Unknown has no matches in COLMAR database. Low level indicates features were low intensity in 2D spectra. Abbreviations: asterisk – feature indicated was found to be overlapped within the 2D spectra, n/a – not applicable.



| SECIM Tools Workflow: | Input Parameters | Set Value |
|---|---|---|
| Blank Feature Filtering (BFF) | BFF Threshold | 5000 |
| | Criterion Value | 100 |
| Standard Euclidean Distance (SED) | Group/Treatment [Optional] | genotype |
| | Input Run Order Name [Optional] | run_order |
| | Additional groups to separate by [Optional] | |
| | Threshold | 0.95 |
| Coefficient of Variation (CV) | Group/Treatment [Optional] | |
| | CV Cutoff [Optional] | 0.1 |
| Principal Components Analysis (PCA) | Group/Treatment [Optional] | |
| Bland-Altman (BA) | Outlier Cutoff | 3 |
| | Sample Flag Cutoff | 0.2 |
| | Feature Flag Cutoff | 0.05 |
| | Group/Treatment [Optional] | genotype |
| | Group Name [Optional] | |
| Generate distribution of features across samples | Group/Treatment [Optional] | genotype |

**Supplemental Table 3.3.** SECIM Tools workflow, input parameters, and set values used for QC steps on analytical data.

| | RP (+) | HILIC (+) | RP (-) |
|---|---|---|---|
| | | | |
| fold blank | 1 | 1 | 1 |
| frac qc | 1 | 1 | 1 |
| | | | |
| alpha* | 0.1 | 0.1 | 0.1 |
| dmz* | 0.005 | 0.005 | 0.005 |
| drt* | 0.03 | 0.03 | 0.03 |
| extracted quantity | height | height | height |
| num references | 150 | 150 | 150 |
| ppm | 5 | 5 | 5 |
| | | | |
| adducts positive | [M+H]+, [M+2H]2+, [M+Na]+, [M+K]+, [M+NH4]+, [M+2Na-H]+, [2M+H]+, [2M+2H]2+, [2M+H+Na]2+, [2M+Na]+, [2M+2Na-H]+, [M+2H-NH3]2+, [M+H-H2O]+, [M+H-H2O]+, [M+2H-H2O]2+, [M+3H]3+, [M+CH3COONa+H]+, [M+CH3COONa+Na]+, [M+CH3COONa+NH4]+ | [M+H]+, [M+2H]2+, [M+Na]+, [M+K]+, [M+NH4]+, [M+2Na-H]+, [2M+H]+, [2M+2H]2+, [2M+H+Na]2+, [2M+Na]+, [2M+2Na-H]+, [M+2H-NH3]2+, [M+H-H2O]+, [M+H-H2O]+, [M+2H-H2O]2+, [M+3H]3+, [M+CH3COONa+H]+, [M+CH3COONa+Na]+, [M+CH3COONa+NH4]+ | [M-H]-, [M-2H]2-, [M-2H+Na]-, [M-H+Cl]2-, [M-2H+K]-, [M+Cl]-, [2M-H]-, [2M-2H+Na]-, [2M-H+Cl]2-, [2M-2H+K]-, [2M+Cl]-, [M-H-H2O]- |
| dmz | 0.005 | 0.005 | 0.005 |
| main adducts positive | [M+H]+, [M+2H]2+, [M+Na]+, [M+NH4]+ | [M+H]+, [M+2H]2+, [M+Na]+, [M+NH4]+ | [M-H]-, [M+Cl]-, [2M-H]- |
| max charge | 3 | 3 | 3 |
| max isotopes | 4 | 4 | 4 |
| min filter | 2 | 2 | 2 |
| num files | 100 | 100 | 100 |
| polarity | positive | positive | negative |
| ppm | 5 | 5 | 5 |
| | | | |
| files used | 12 | 12 | 12 |
| need optimization | TRUE | TRUE | TRUE |
| noise threshold | 1000 | 1000 | 1000 |
| num iterations | 5 | 5 | 5 |
| number of points | 30 | 30 | 50 |
| | | | |
| ms1 | gap-filled data matrix | gap-filled data matrix | gap-filled data matrix |
| ms2 | fused mgf | fused mgf | fused mgf |
| | | | |
| algorithm | ADAP | ADAP | ADAP |
| noise level ms1 | 1000 | 1000 | 1000 |
| noise level ms2 | 1000 | 1000 | 1000 |
| | | | |
| SN* | 4.57 | 17.38 | 15.6 |
| coefficient area threshold* | 57.94 | 108.56 | 39.05 |
| ms2 mz tol | 0.005 | 0.005 | 0.005 |
| ms2 rt tol | 0.1 | 0.1 | 0.1 |
| noise level | 10000 | 10000 | 10000 |
| peak width (min/max)* | 0.018 / 0.718 | 0.030 / 0.434 | 0.083 / 0.581 |
| rt wavelet (min/max)* | 0.001 / 0.081 | 0.002/ 0.110 | 0.012 / 0.103 |
| peaktable filter | absolute intensity top 30000 | absolute intensity top 30000 | absolute intensity top 30000 |
| | | | |
| dmz* | 0.005 | 0.004 | 0.002 |
| min scan* | 6 | 4 | 4 |
| ppm* | 6 | 12 | 6 |

*value determine via SLAW optimization

**Supplemental Table 3.4.** Pre-processing steps, input parameters, and set values used for LC-MS data are listed in the order of execution**.**

# References

1. Shaver, A.O., Gouveia, G.J., Kirby, P.S., Andersen, E.C. & Edison, A.S. Culture and Assay of Large-Scale Mixed-Stage Caenorhabditis elegans Populations. *J Vis Exp* (2021).
2. Yoshimura, J. et al. Recompleting the Caenorhabditis elegans genome. *Genome Res* **29**, 1009-1022 (2019).
3. Lin, D.Y. & Zeng, D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97**, 321-332 (2010).
4. Liu, Z., Vol. Master of Science 61 (University of Florida, Gainesville, FL; 2021).
5. Hall, J.A. & Rosenthal, R. Interpreting and evaluating meta-analysis. *Eval Health Prof* **18**, 393-407 (1995).
6. Rosenthal, R. & DiMatteo, M.R. Meta-analysis: recent developments in quantitative methods for literature reviews. *Annu Rev Psychol* **52**, 59-82 (2001).
7. Liu, Q. et al. Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Sci Rep* **10**, 13856 (2020).
8. Walejko, J.M., Chelliah, A., Keller-Wood, M., Gregg, A. & Edison, A.S. Global Metabolomics of the Placenta Reveals Distinct Metabolic Profiles between Maternal and Fetal Placental Tissues Following Delivery in Non-Labored Women. *Metabolites* **8** (2018).

# CO-AUTHORED PUBLICATIONS

# C.1 Long-term Metabolomics Reference Material

Goncalo J. Gouveia,[1,4] Amanda O. Shaver,[2,4] Brianna M. Garcia,[3,4] Alison M. Morse,[5] Erik C. Andersen,[6] Arthur S. Edison,[1,2,4]* Lauren M. McIntyre [5] *.

[1]Department of Biochemistry & Molecular Biology, Green Street, University of Georgia, Athens, Georgia, 30602

[2] Department of Genetics, University of Georgia, Green Street, Athens, Georgia, 30602

[3] Department of Chemistry, University of Georgia, 140, Cedar Street, Athens, Georgia, 30602

[4] Complex Carbohydrate Research Center, University of Georgia, 315, Riverbend Road, Athens, Georgia, 30602.

[5] Department of Molecular Genetics and Microbiology and University of Florida Genetics Institute, Mowry Road, University of Florida, Gainesville, Florida, 32610.

[6] Department of Molecular Biosciences, Northwestern University, 2205, Tech Drive, Evanston, Illinois, 60208.

*Corresponding authors: aedison@uga.edu and mcintyre@ufl.edu

This manuscript was published in Analytical Chemistry in June 2021. DOI: https://pubs.acs.org/doi/10.1021/acs.analchem.1c01294

## C.1.1 Abstract

The use of quality control samples in metabolomics ensures data quality, reproducibility and comparability between studies, analytical platforms and laboratories. Long-term, stable and sustainable reference materials (RMs) are a critical component of the QA/QC system, however, the limited selection of currently available matrix- matched RMs reduce their applicability for widespread use. To produce a RM in any context, for any matrix that is robust to changes over the course of time we developed IBAT (Iterative Batch Averaging meThod). To illustrate this method, we generated 11 independently grown *E. coli* batches and made a RM over the course of 10 IBAT iterations. We measured the variance of these materials by NMR and showed that IBAT produces a stable and sustainable RM over time. This *E. coli* RM was then used as a food source to produce a *Caenorhabditis elegans* RM for a metabolomics experiment. The metabolite extraction of this

material, alongside 41 independently grown individual *C. elegans* samples of the same genotype, allowed us to estimate the proportion of sample variation in pre-analytical steps. From the NMR data, we found that 40% of the metabolite variance is due to the metabolite extraction process and analysis and 60% is due to sample-to-sample variance. The availability of RMs in untargeted metabolomics is one of the predominant needs of the metabolomics community that reach beyond quality control practices. IBAT addresses this need by facilitating the production of biologically relevant RMs and increasing their widespread use.

### C.1.2 Contributions

I aided in sample generation, provided intellectual contributions and feedback, and provided manuscript edits.

## C.2 Taguchi Design of Experiments Approach for Untargeted Metabolomics Sample Preparation Optimization

Brianna M. Garcia[a], Goncalo J. Gouveia[a], **Amanda O. Shaver[a]**, I. Jonathan Amster[a], Arthur S. Edison[a], and Franklin E. Leach III[a]*

[a]University of Georgia, Athens, GA 30602, United States

This manuscript was submitted to Analytical Chemistry in December 2021 and is under review.

### C.2.1 Abstract

Metabolomics commonly uses analytical techniques such as nuclear magnetic resonance (NMR) and liquid chromatography coupled to mass spectrometry (LC-MS) to quantify and identify metabolites associated with biological variation. Metabolome coverage from non-targeted LC-

MS studies relies heavily on the pre-analytical protocols (*e.g.,* homogenization and extraction) used. Chosen protocols impact which metabolites are successfully measured, which in turn impacts biological conclusions. Different homogenization and extraction methods produce significant variability in metabolome coverage, sample reproducibility, and extraction efficiency. Herein we describe an efficient Taguchi method design of experiments (DOE) approach to optimize the extraction solvent and volume, extraction time, and LC reconstitution solvent for a sequential non-polar and polar *Caenorhabditis elegans* extraction. DOE is rarely used in metabolomics yet provides a systematic approach for optimizing sample preparation while simultaneously decreasing the number of experiments required to obtain high-quality data.

### C.2.2 Contributions

I aided in sample generation, provided intellectual contributions and feedback, and provided manuscript edits.

# BIOGRAPHICAL SKETCH

Amanda Shaver graduated with a degree in Biology from the University of Kansas in 2011. During this time, she was able to work with Dr. Paulyn Cartwright studying the evolutionary development of an enigmatic medusozoan, *Polypodium hydriforme*. Between 2011 and 2015, before attending graduate school, she held several laboratory technician positions to refine her research interests. In 2015, she joined the Integrated Life Sciences Program at the University of Georgia where she obtained her Ph.D. in Genetics in the spring of 2022. While at the University of Georgia, she joined Dr. Edison's laboratory where she spearheaded method development and approaches to identify unknown metabolites in *Caenorhabditis elegans* as part of a large consortium project. In the future, she hopes to merge her long-standing interest in host-parasite interactions with her biochemical training.