Video Article

Phage Phenomics: Physiological Approaches to Characterize Novel Viral Proteins

Savannah E. Sanchez[1], Daniel A. Cuevas[2], Jason E. Rostron[1], Tiffany Y. Liang[3], Cullen G. Pivaroff[1], Matthew R. Haynes[1], Jim Nulton[4], Ben Felts[4], Barbara A. Bailey[4], Peter Salamon[4], Robert A. Edwards[1,5,6], Alex B. Burgin[7], Anca M. Segall[1], Forest Rohwer[1]

[1]Department of Biology, San Diego State University

[2]Computational Science Research Center, San Diego State University

[3]Bioinformatics and Medical Informatics Research Center, San Diego State University

[4]Department of Mathematics and Statistics, San Diego State University

[5]Department of Computer Science, San Diego State University

[6]Mathematics and Computer Science Division, Argonne National Laboratory

[7]SPARC Committee, Broad Institute

Correspondence to: Savannah E. Sanchez at sanchez.s.elizabeth@gmail.com

URL: http://www.jove.com/video/52854
DOI: doi:10.3791/52854

Keywords: Immunology, Issue 100, phenomics, phage, viral metagenome, Multi-phenotype Assay Plates (MAPs), continuous culture, metabolomics

Date Published: 6/11/2015

Citation: Sanchez, S.E., Cuevas, D.A., Rostron, J.E., Liang, T.Y., Pivaroff, C.G., Haynes, M.R., Nulton, J., Felts, B., Bailey, B.A., Salamon, P., Edwards, R.A., Burgin, A.B., Segall, A.M., Rohwer, F. Phage Phenomics: Physiological Approaches to Characterize Novel Viral Proteins. *J. Vis. Exp.* (100), e52854, doi:10.3791/52854 (2015).

Abstract

Current investigations into phage-host interactions are dependent on extrapolating knowledge from (meta)genomes. Interestingly, 60 - 95% of all phage sequences share no homology to current annotated proteins. As a result, a large proportion of phage genes are annotated as hypothetical. This reality heavily affects the annotation of both structural and auxiliary metabolic genes. Here we present phenomic methods designed to capture the physiological response(s) of a selected host during expression of one of these unknown phage genes. Multi-phenotype Assay Plates (MAPs) are used to monitor the diversity of host substrate utilization and subsequent biomass formation, while metabolomics provides bi-product analysis by monitoring metabolite abundance and diversity. Both tools are used simultaneously to provide a phenotypic profile associated with expression of a single putative phage open reading frame (ORF). Representative results for both methods are compared, highlighting the phenotypic profile differences of a host carrying either putative structural or metabolic phage genes. In addition, the visualization techniques and high throughput computational pipelines that facilitated experimental analysis are presented.

Video Link

The video component of this article can be found at http://www.jove.com/video/52854/

Introduction

Viruses that infect Bacteria (a.k.a. bacteriophage or phage) are estimated to exist at more than 10^{31} virus like particles (VLPs) globally and outnumber all other organisms in an environment[1,2]. The first metagenomic study investigating the viral communities associated with marine environments focused on quantifying the diversity seen within the viral fraction[3]. Additionally, Breitbart and colleagues found that over 65% of the viral community sequences shared no homology to any sequences available in public databases. Subsequent metagenomic studies found similar evidence: metagenomes from marine sediments in San Diego, California contain 75% unknown viral sequences[4]; metagenomes from hypersaline lakes of the Salton Sea contain 98% unknown viral sequences[5]; and coral–associated metagenomes contain 95 - 98% unknown viral sequences[6]. This accumulation of unannotated information has resulted in phage genetic material being “the dark matter of the biological universe”[7].

Genomic characterization of phage relies on identifying sequence similarity through comparison against existing nucleic acid and protein databases. Because phage-encoded genetic information is predominantly unknown, homology-based methods are ineffective. Within their genome, phages typically encode three major gene types: transcription and replication genes, metabolic genes, and structural genes. The transcription and replication genes (class I/II genes[8]) include polymerases, primases, endo/exo-nucleases, and kinases. These genes are highly conserved due to their importance in phage infection, transcribing and replicating phage genetic material. Phage polymerases are readily identified using traditional sequence homology methods due to their global conservation[9] and have been shown to serve as effective phylogenetic markers[10]. In contrast, phage metabolic and structural genes (class II/III genes[8]) are increasingly divergent and often annotated as hypothetical genes.

Phage metabolic genes affect the metabolic capacity of the host and are not necessarily required for viral replication. These genes, often referred to as auxiliary metabolic genes[11] (AMGs), appear to modulate host metabolism and allow optimal progression of infection and success of virion

maturation. AMGs have been associated with the utilization and uptake of limiting nutrients or in energy production pathways. Some examples include photosystem genes found in the genomes of various cyanophage[12-16], genes connected to and regulated by phosphate metabolism[17,18], and utilization of the pentose phosphate pathway for phage dNTP biosynthesis[18,19]. In comparison, structural genes are among the mid to late genes produced during infection and vary across different phage-host systems. The production of structural proteins are dependent on the availability of viral dNTP, and energy pools for their transcription, translation, and assembly[8]. The capsid and tail fiber structural proteins are deemed as the most divergent of all viral protein-encoding genes and are required for successful virion production. Their divergence is typically attributed to the active role they play in shaping virus-host coevolution[20]. Divergent proteins, regardless of the gene class, are readily overlooked when using traditional homology and sequence alignment techniques. An effort to correct for the limitations seen with strict sequence comparisons has resulted in bioinformatics tools capable of using sequence characteristics to determine association, such as artificial neural nets[21]. Artificial neural nets (ANNs) allow for the prediction of structural and metabolic genes, however, require downstream experimental validation to directly characterize gene function.

The objective of this manuscript is to provide phenomic protocols capable of monitoring both catabolic and anabolic metabolism of a host bacterium during the expression of a novel phage gene, functionally predicted through ANNs. The field of phenomics, the biology associated with cellular phenotypes, is well established in systems biology to aid in the investigation of proteins with unknown or pleiotropic function. Phenomic tools are used to link phenotypic information to genotypic information. We hypothesize for putative phage genes that their function(s) can be determined through observing host physiological effects during phage gene expression. To investigate this hypothesis, two quantitative methods were chosen. Multi-phenotype Assay Plates (MAPs) were used to monitor host substrate utilization and the subsequent biomass formation while metabolomics measured host metabolite diversity and relative abundance during growth in specific environmental conditions. Putative structural and metabolic proteins were overexpressed in *Escherichia coli* and representative results from both experiments are compared. Numerous visual techniques and high throughput processing pipelines are presented to facilitate experimental replication. Lastly, the reproducibility and accuracy of the presented methods are discussed in the context of expected physiological effects for an annotated capsid protein and phage metabolic protein, thioredoxin, plus two putative AMGs.

Protocol

1. Preparation of Multi-phenotype Assay Plate (MAP) Substrates, Basal Media, Pre-growth Media, and Buffer

1. Make 50 ml of 1.0 - 1.25% substrate stocks.
    1. Dissolve 1.0 - 1.25% (w/v) of solid substrate in sterile water (heat if necessary). Filter sterilize solutions with a 0.22 µm filter unit and store stocks at RT. Examples of substrates used in these experiments are provided in **Table 2**.

2. Make 250 ml of 3x basal media for all substrate classes (carbon, nitrogen, sulfur and phosphorus). The MOPS (3-morpholinopropane-1-sulfonate) based basal media[22] contains the following: 1x MOPS (40 mM MOPS + 10 mM Tricine), 0.4% glycerol*, 9.5 mM NH_4Cl*, 0.25 mM $NaSO_4$*, 1.0 mM $MgSO_4$*, 1.32 mM K_2HPO_4*, 10 mM KCl, 0.5 µM $CaCl_2$, 5 mM NaCl, 6 µM $FeCl_3$, and 0.1 % (w/v) L-arabinose** (**Tables 1 and 2**).
Note: *These compounds are not included depending on the basal media. For example, 0.4% glycerol is not used in the carbon basal media and in the sulfur basal media 1.0 mM $MgSO_4$ is replaced by 1.0 mM $MgCl_2$. **L-arabinose is specific for induction of pBAD promoter vector, pEMB11, which was transformed into an ara$^-$*E. coli* K-12 strain (BW 27784)[23].
    1. Add each component of the basal media to a sterile flask and bring solution to volume. Mix well. With a 0.22 µm filter unit, filter sterilize each basal media into a sterile bottle.

3. Make 500 ml of MOPS pre-growth media. MOPS based pre-growth media[22] contains the following: 1x MOPS (40 mM MOPS + 10 mM Tricine), 0.4% glycerol, 9.5 mM NH_4Cl, 0.25 mM $NaSO_4$, 1.0 mM $MgSO_4$, 1.32 mM K_2HPO_4, 10 mM KCl, 0.5 µM $CaCl_2$, 5 mM NaCl, and 6 µM $FeCl_3$.
    1. Add each component of the pre-growth media to a sterile flask and bring to volume. Filter sterilize with a 0.22 µm filter unit, then add ampicillin at a final concentration of 100 µg/ml. Store solution at 4 °C.

4. Make 500 ml of 0.5x Luria-Bertani (LB) agar plates. To a sterile flask, add LB components to make a 0.5x concentration (1.25 g yeast extract, 2.5 g NaCl, 2.5 g tryptone, 7.5 g agar). Mix well and autoclave to sterilize.
    1. Cool agar, then add 100 µg/ml of antibiotic ampicillin with mixing. Pour ~25 ml of agar into Petri dishes, allow to set, and store at 4 °C until needed.

5. Make 250 ml of 10 mM Tris (pH 7.4) plus 10 mM $MgSO_4$ buffer solution. Filter sterilize with a 0.22 µm filter unit, and store solution at RT.

2. Preparation of Bacterial Cell-suspension

1. Prepare fresh colonies. From a 12.5% glycerol frozen stock, plate fresh *E. coli* clones onto 0.5x LB agar containing 100 µg/ml ampicillin. Incubate at 37 °C for 24 hr.
2. Prepare liquid cultures:
    1. Pipette 3 ml of pre-growth media containing 100 µg/ml ampicillin into culture tubes. For each clone, use biological triplicates.
    2. Inoculate independent colonies from section 2.1 into prepared culture tubes. Incubate at 37 °C with shaking for 22 hr.

3. Pellet and wash bacterial cells:
    1. Transfer O/N cultures into 1.7 ml microfuge tubes. Pellet cells via centrifugation at maximum speed (16,900 x g) for 2 min in a microcentrifuge. Decant supernatant and wash cells by re-suspending in 500 µl of 10 mM Tris/10 mM $MgSO_4$ buffer. Repeat.

2. Re-pellet cells, decant supernatant and re-suspend cells in 1 ml of 10 mM Tris/10 mM $MgSO_4$ buffer.

4. Measure cell density and concentrate cells (if necessary):
   1. Dilute cells from section 2.3.3 10-fold in the 10 mM Tris/10 mM $MgSO_4$ buffer and transfer this dilution into a cuvette. Measure optical density ($OD_{600\ nm}$) of this dilution and record.
   2. Concentrate cells to achieve final concentration of 0.07 ($OD_{600\ nm}$) in the MAPs (the cells will be diluted 15 fold when they are added to the MAPs, therefore cells need to be at an initial concentration of $OD_{600\ nm}$ = 1.05). Transfer concentrated cell suspension into a reservoir and save (at RT) until step 3.5 of MAPs preparation.

3. Preparation of Multi-phenotype Assay Plates (MAPs)

1. Label MAPs. Label sterile 96-well micro-titer plates with an *E. coli* clone identification number and MAP schematic type.
2. Aliquot sterile water into MAPs. Aseptically transfer sterile water into a liquid reservoir. Pipette 60 µl into each well of the micro-titer plate using a multi-channel pipette.
3. Aliquot basal media into MAPs. Aseptically transfer 3x basal media into a liquid reservoir. Pipette 50 µl into each well of the micro-titer plate using a multi-channel pipette. Repeat steps 3.2 and 3.3 for each basal media used in the MAP schematic.
4. Aliquot substrates into MAPs. Transfer 30 µl of each substrate into the appropriate well of the MAPs.
5. Add bacterial suspension to the MAPs. Transfer 10 µl of bacterial cells from section 2.4.2 into each well of the MAP using a multi-channel pipette. Change tips before re-introducing pipette back into the culture stock.
6. Cover MAPs with adhesive film. Cover each plate with a single adhesive plate film. Firmly press the film atop the wells of the plate and along the edges to create an even and tight seal. Remove excess film from the edges of the micro-titer plate with a sterile razor blade.
7. Measure optical density of MAPs:
   1. Place prepared MAPs in the “Input” sleeve of a multi-plate spectrophotometer plate reader. Open plate reader software and create a protocol that measures absorbance ($OD_{600\ nm}$) every 30 min, for a total of 32 hr, with 60 sec of shaking between reads.
   2. Set temperature in the apparatus to 37 °C. Start protocol once MAPs have equilibrated to set temperature.
   3. After 32 hr, remove MAPs from the “Output” sleeve of the plate reader. Label each data text file to include: the *E. coli* clone identification number, date the MAP ran, and the MAP schematic type.

4. Multi-phenotype Assay Plates (MAPs) Processing and Parameterization

1. Assess growth curves using an automated QA process, *e.g.,* PMAnalyzer[24].
   1. Verify that growth curves have $OD_{600\ nm}$ < 0.20 in the first 2 hr. Do not use absorbance at the initial time point (t_0) in the filter due to artifacts of condensation, which typically resolve at t_1 (30 min).
   2. Remove growth curves that violate QA filter. Save curve information (sample name, well number, and $OD_{600\ nm}$ values) in a separate output file for future reference. Continue with analysis if growth curve passes QA filter.

2. Calculate median growth curves. Using replicate data, per well, calculate the median growth curve by taking the median optical density ($OD_{600\ nm}$) value at each time point. Record resulting median growth curves in a separate output file for future use.
3. Calculate the best-fit logistic model for each growth curve using an adaptation of the equation proposed by Zwietering[25]. The logistic equation includes three parameters describing bacterial growth: the lag time, λ (hr), the maximum rate of growth, μ_{max} ($OD_{600\ nm}$ 100^{-1}), and the final biomass yield, α ($OD_{600\ nm}$). Use data at time = 30 min (y_1) as the initial value due to artifacts of condensation.

$$\hat{y} = y_1 + \frac{A - y_1}{1 + \exp\left[\left(\frac{\mu_{max}}{A}\right)(\lambda - t_i)\right] + 2} \quad [1]$$

   1. Calculate the maximum growth rate using a direct search. Record the largest rate of change over a 90 min interval within the data.

$$\mu_{max} = \max_{i \in 1,2,\ldots,n} \frac{\log(y_{i+3}) - \log(y_i)}{t_{i+3} - t_i} \quad [2]$$

   2. Estimate the final biomass yield by searching for the upper asymptotic value of the largest moving average over a 90 min window. Specifically, define as the asymptote of the growth curve.

$$A = \max_{i \in 1,2,\ldots,n-2} \frac{\sum_{j=1}^{i+2} y_j}{3} \quad [3]$$

   3. Using the μ_{max} and A values from 4.3.1 and 4.3.2, find the λ value by trying all values λ:

$$\lambda = 0.25K, \quad K = 1, 2, \ldots, 2n \quad [4]$$

      1. Use each λ value to calculate a sum of squared error (SSE). Use the lowest SSE is the SSE (λS):

$$SSE(\lambda) = \sum_{i=1}^{n} (y_i - \hat{y}_i(\lambda_K))^2 \quad [5]$$

5. Phenotype Analysis of MAPs

1. Calculate the growth level (GL) for each growth curve to assess the overall growth per clone per substrate. Define GL as the harmonic mean of the shifted logistic-fitted values:

$$GL = \frac{n}{\sum_{i=1}^{n} \frac{1}{\hat{y}_i + A}}$$ [6]

2. Assign a phenotype to each growth curve based on the GL values. Here, the four phenotypes used are: expected growth, no growth, gain of function, and loss of function.
    1. Determine phenotypes by comparing growth across all clones on a specific substrate in terms of standard deviations away from the mean. Use this statistic and a minimum growth threshold to designate the phenotype presented by the growth curve (**Figure 1A,B**). **Figure 1C** provides examples of phenotype assignment.
    2. Define growth as GL ≥ 0.4 (**Figure 1A,B**). Define Gain of Function (**Figure 1C**, green) as having a GL > two standard deviations above the mean and a mean > the growth threshold. Loss of Function (**Figure 1C**, red) is defined as having a GL < two standard deviations below the mean and a mean > the growth threshold.

3. Create visual representations of the data set based on phenotypes and growth curves.
    1. View growth dynamics portraying $OD_{600\ nm}$ values as color to quickly compare growth curves between multiple clones (**Figure 2**).
    2. View phenotypic distributions among all clones based on growth conditions (**Figure 3**).

6. Building the Continuous Culture Apparatus

1. Reactor port construction (**Figure 4A,B**). Drill three 1/4 in. holes, spaced 3/4 in. apart, in the cap of the continuous culture reactor.
    1. For port α: screw a female luer thread style panel mount to 1/16 in. barb from the bottom of the cap with the barb pointing up, out of the cap. Secure barb with a 1/4 in. panel mount lock nut.
    2. For port β: screw a female luer thread style panel mount to 1/16 in. barb with the barb pointing up, out of the cap. Secure barb with a 1/4 in. panel mount lock nut.
    3. For port γ: screw a female luer thread style panel mount to 1/16 in. barb from the bottom of the cap so the barb is pointing up, out of the cap. Secure barb with 1/4 in. panel mount lock nut. Screw a male luer with integral lock ring to 1/8 in. wide bore hose to the barb fitting on the inside of the cap.
    4. For the outflow port: drill a 5/16 in. hole at the 75 ml mark of the continuous culture reactor. Cut 3/4 in. off the nut end of the 1/4 in. barbed bulkhead fitting. Screw in the 1/4 in. barbed bulkhead fitting (gasket is on the outside of the bottle and the nut is on the inside, **Figure 4B**).

2. Feeding bottle port construction (**Figure 4C**). Drill two 1/4 in. holes spaced 1 in. apart in the cap of the feeding bottle.
    1. For the port δ: screw a female luer thread style panel mount to 1/8 in. barb with the barb pointing up, out of the cap. Secure barb with a 1/4 in. panel mount lock nut.
    2. For the port ε: screw a female luer thread style panel mount to 1/16 in. barb with the barb pointing up, out of the cap. Secure barb with a 1/4 in. panel mount lock nut.
        1. Screw a male luer with integral lock ring to 1/8 in. wide bore hose to the barb fitting, on the inside of the cap.

3. Feeding tubes and tube extensions:
    1. Cut two 1 in. pieces of tubing (1/8 in. outer diameter (OD) x 1/16 in. inner diameter (ID)). Fit pieces onto ports α and γ of the continuous culture reactor made in section 5.2.
    2. Cut a single 1 in. piece of tubing (1/4 in. OD x 1/8 in. ID). Fit piece onto port B of the continuous culture reactor.
    3. Cut a single 3.5 in. piece of tubing (1/8 in. OD x 1/16 in. ID). Fit this piece to the inside of port γ, on the male luer with integral lock ring to 1/8 in. wide bore hose. Port γ is the sampling port of the continuous culture reactor.
    4. Cut a single 1 in. piece of tubing (1/4 in. OD x 1/8 in. ID). Fit this piece on the δ port of the feeding bottle.
    5. Cut a single 11 in. piece of tubing (1/16 in. OD x 1/8 in. ID). Fit this piece to the inside of port ε of the feeding bottle, on the male luer with integral lock ring to 1/8 in. wide bore hose.
    6. Cut two 18 in. pieces of tubing (1/8 in. OD x 1/16 in. ID). Fit one piece to port ε of the feeding bottle. Save the other piece for section 7.

7. Running Continuous Cultures for Metabolomics

1. Sterilize materials:
    1. Autoclave the dry materials for 30 min. Make sure to wrap cap of the feeding bottle made in section 5 in aluminum foil. Keep attached tubing straight.
        1. Wrap cap of the continuous culture reactor, made in section 5, in aluminum foil. Keep attached tubing straight. Wrap 18 in. tubing cut in section 6.3.5 and the smallest peristaltic pump tubing adaptor in aluminum foil. Keep tubing straight. Autoclave for 30 min.
    2. Make 2 L of 0.5x LB broth. In the feeding bottle, make 0.5x LB broth. Cover feeding bottle with foil. Autoclave for 1 hr.
    3. Make 70 ml of 0.5x LB broth. Pour broth into the continuous culture reactor. Place a stir bar in the continuous culture reactor. Cover reactor with foil and autoclave for 30 min.

2. Continuous culture set-up:
    1. Bacterial cell-suspension.
        1. From a 12.5% glycerol frozen stock, plate fresh *E. coli* clones onto 0.5x LB agar containing 100 µg/ml ampicillin. Incubate at 37 °C for 24 hr.
        2. Inoculate a single colony into 3 ml of 0.5x LB broth containing 100 µg/ml ampicillin. For each clone use biological triplicates. Incubate at 37 °C with shaking for 22 hr.

    2. Connect the parts together.
        1. Place all autoclaved materials into a biological safety cabinet and turn ultraviolet light on while media cools. Once media has cooled, add 0.1% L-arabinose and 100 µg/ml ampicillin to all 0.5x LB broth.
        2. Unwrap continuous culture reactor cap and screw onto the reactor. Avoid touching the sampling tube. Unwrap the feeding bottle cap and screw onto the feeding bottle. Avoid touching the sampling tube.
        3. Keep the 18 in. tubing attached to port ε sterile. Un-wrap the 18 in. tubing and peristaltic pump tubing adaptor.
        4. Fit a free end of the 18 in. tubing onto one end of the peristaltic pump tubing adaptor. Fit the other end of the peristaltic pump tubing adaptor to the 18 in. tubing attached to port ε of the feeding bottle cap.
        5. Screw 0.22 µm filter units onto ports β and δ of the continuous culture reactor. Screw a 0.22 µm filter unit on the adaptor of port α. To the 0.22 µm filter unit, fit the free end of the 18 in. tubing.

    3. Start the continuous culture.
        1. In the biosafety hood, inoculate the continuous culture reactor with 100 µl of O/N culture made in section 7.2.1.1.
        2. Close system before moving the continuous culture into a 37 °C incubator. In the incubator have a magnetic stir plate and mini peristaltic pump set up.
        3. Fit the peristaltic pump tubing adaptor into the peristaltic pump. Tubing from the feeding bottle starts at the “In” side and tubing leading to the reactor starts at the “Out” side.
        4. Set the peristaltic pump to: “FAST”. Start the peristaltic pump by switching to “FORWARD”. Check that media begins to flow through the tubing.
        5. Place reactor on magnetic stir plate and begin mixing. Continuous culture reactor must equilibrate for 24 hr before sampling.

    4. Sampling of the continuous culture. Screw a 5 ml luer lok syringe onto port γ and pull up 4.5 ml of culture.
        1. Use 500 µl of the culture to measure the density ($OD_{600\ nm}$). Dilute cells to make 4 x 1 ml cultures at $OD_{600\ nm}$ = 0.35.
        2. Aliquot diluted cultures into 1.7 ml microfuge tubes. Repeat sampling every 12 hr for 4 days.

    5. Prepare samples for GC-TOFMS:
        1. Pellet cells via centrifugation at max speed (16,900 x g) for 2 min. Decant supernatant. Wash cells with 500 µl of phosphate buffered saline (PBS). Repeat this step.
        2. Decant supernatant a final time, and then submerge microfuge tubes with pelleted cells in liquid nitrogen until bubbling stops. Store samples in -80 °C freezer.

8. Running Serial Passage Batch Cultures for Metabolomics

1. Preparing bacterial cell-suspension. From a 12.5% glycerol frozen stock, plate fresh *E. coli* clones onto 0.5x LB agar containing 100 µg/ml ampicillin. Incubate at 37 °C for 24 hr.
    1. Inoculate individual colonies into 3 ml of 0.5x LB broth containing 100 µg/ml ampicillin. Use biological triplicates for each clone.

2. Serial passage batch cultures.
    1. Subculture O/N from 8.1.1 via a 500-fold dilution into 3 ml of LB broth containing 50 µg/ml ampicillin and 0.1% L-arabinose. Subculture should have an $OD_{600\ nm}$ < 0.005. Incubate at 37 °C with shaking.
    2. Check optical density ($OD_{600\ nm}$) at 2 and 2.5 hr. Grow cells to achieve $OD_{600\ nm}$ = 0.35.
        1. Subculture via a 500-fold dilution into 3 ml of LB broth containing 50 µg/ml ampicillin and 0.1% L-arabinose. Subculture should have an $OD_{600\ nm}$ < 0.005.

    3. Check optical density ($OD_{600\ nm}$) at 2 and 2.5 hr. Grow cells to achieve $OD_{600\ nm}$ = 0.35.

3. Sampling serial passage batch cultures.
    1. Using sterile forceps, place a 0.22 µm membrane filter on top of a filter manifold. Aliquot 1 ml of Phosphate Buffered Saline (PBS) onto the filter paper and then turn on vacuum pump.
    2. Repeat addition of 1 ml of PBS. Transfer 1 ml of subculture from section 8.2.3 onto the filter paper. From here, move quickly to get samples into liquid nitrogen immediately. Complete this in 1 min.
    3. Aliquot 1 ml of PBS onto filter paper to wash sample. Flush rapidly without spilling sample over sides of filter paper. Repeat.
    4. Using sterile forceps place filter into a 1.7 ml microfuge tube by carefully folding the filter. Submerge microfuge tube in liquid nitrogen until bubbling stops. Store sample in -80 °C freezer.

9. Metabolite Analysis & Processing

1. Ship 3 samples per clone on dry ice to a metabolomics core facility for sample processing, analysis and normalization of GC-TOFMS.
2. For each sample determine the mean, median, standard deviation and coefficient of variation across metabolites, using replicate data.

3. For statistical analysis, manually code a QA pipeline using conditional statements and repetitive executions[26] to remove metabolites that do not follow defined conditions.
    1. For Condition 1, remove metabolites in which more than 2 samples have zero abundance. Remove internal standards from the metabolomic profiles.
    2. For Condition 2, remove those metabolites that are not found in the cells of interest. Here, remove the following metabolites: indole 3 acetate, dihydroabietic acid, nonadecanoic acid, salicylic acid, salicylaldehyde, cholesterol, phenol, 1-monostearin, octadecanol, 1-monopalmitin, dodecane, dodecanol, 1-hexadecanol, 5-methoxytryptamine, benzoic acid, pentadecanoic acid, phosphoric acid, pelargonic acid, palmitic acid, capric acid, hydroxylamine, stearic acid, myristic acid, maleimide, levoglucosan, and 4-hydroxybutyric acid.
    3. For Condition 3, remove those metabolites whose coefficient of variation is greater than 1.

4. Capture global metabolomics effects by looking at metabolite relative abundances.
    1. Create a visual representation of the median metabolite abundance per clone for each metabolite (**Figure 6**).

5. Identify outliers by observing clone-metabolite pair frequencies.
    1. Calculate the Z score for each median value of every metabolite in a clone. Calculate z scores using the following equation:

$$Zmx = \frac{X_{MC} - \mu_C}{\sigma_C} \quad [7]$$

Note: Term X_{MC} represents the abundance level of a specific metabolite for a specific clone. Term μ_C represents the mean of all clones for a specific metabolite. Term σ_C is the associated standard deviation.

    2. Create a visual representation of the data in which outliers are highlighted (data not provided).

Representative Results

All samples used to determine the open reading frames (ORFs) selected in this study were collected from Starbuck Island, Site 7 (STAR7) and Caroline Atoll, Site 9 (CAR9) of the southern Line Islands. An estimated 100 L of seawater from these sites was collected below the coral boundary layer using bilge pumps, as described previously[5]. Contents of the pumps were subject to fractionalization through large pored filters to remove small eukaryotes and subsequently concentrated using 100 kDa tangential flow filters leaving only microbes and viral like particles (VLPs). To separate the VLPs, remaining seawater was passed through 0.45 µm filters resulting in the virome. Chloroform was introduced to this viral fraction to arrest growth of any remaining cells and stored at 4 °C.

VLPs were purified using the cesium chloride method, in which density gradients separate through centrifugation and allow for the recovery of virions at ~1.35 g/ml to 1.5 g/ml [3]. Viral DNA was extracted using a CTAB/phenol:chloroform protocol and amplified via multiple displacement amplification using Phi29 reagents. Sequencing of the virome was accomplished with commercially available pyrosequencing technology.

Bioinformatics used in the processing and selection of viral ORFs for this study are as follows. Three pre-processing steps were used on the CAR9 and STAR7 viral metagenomes. First, public software was used to remove tag sequences that resulted from amplification of the viral DNA prior to sequencing[27]. Second, common sequencing artifacts such as sequence duplicates and low copy number were filtered out of the dataset via an additional bioinformatics program[28]. Lastly, removal of foreign sequence contamination[29] was performed for those sequences that had ≥ 90% coverage and ≥ 94% identity with sequences in the following databases: RefSeq virus genomes; Human — Reference GRCh37; Human — Celera Genomics; Human — Craig Venter (HuRef); Human — Seong-Jin Kim (Korean); Human — Chromosome 7 version 2 (TCAG); and Human — James Watson, YanHuang (YH; Asian), Yoruba (NA18507; African) reference sequences[21]. Following these processes, sequences from the CAR9 samples totaled 591,600 and STAR7 sequences totaled 939,311. These sequences were uploaded onto MGRAST and assembled via assembler software using default settings. Contigs were translated into 6 reading frames and putative open reading frames (pORFs) were identified using scripts, as previously described[21].

To identify unknown ORFs a number of similarity based searches were performed to remove ORFs of known function. Briefly, the following searches were performed with their corresponding searching criteria[21]:

1. Significant similarity of ≥ 95% identity over ≥ 40 base pair (bp) by MGRAST BLAT in M5NR database.
2. Significant similarity (e-value ≤ 0.001) by TBLASTN against all public metagenomes in My Metagenome Database Resource.
3. Significant similarity (e-value ≤ 0.001) by BLASTP and TBLASTN against NR database.
4. Significant similarity (e-value ≤ 0.001) by RPS-BLAST against the Conserved Domain Database.
5. Protein translations from a select fraction of each dataset compared against solved protein structures in the Protein Data Bank.
6. Calculation of dinucleotide frequencies using Dinucleotide Signatures package.

The resulting pORFs were designed for expression in *E. coli* using publically available gene design software. Back-translation of the amino-acid sequences employed a Universal Codon Usage Table designed to accommodate expression in *E. coli* with a minimum usage threshold of 2%. Restriction enzyme recognition sequences for BamHI and HindIII were excluded from the sequences to facilitate cloning. An outside company synthesized the engineered gene sequences[30] and then ORFs were cloned into a medium-copy number pBAD promoter vector, pEMB11, via standard restriction enzyme cloning. All clones were transformed into *E. coli* K-12 strain BW 27784[23].

Multi-phenotype Assay Plates (MAPs)

A high throughput and robust software pipeline was leveraged for analysis of the MAPs, the PMAnalyzer[24]. The pipeline was developed in a Linux server environment and performs various steps including: parsing of optical density files, formatting data into readable text files, pre-

processing of growth curves for quality assurance (QA), and performing mathematical modeling techniques to analyze growth curves. The primary modeling scripts were developed in Python version 2.7.5 to make use of the **PyLab** module.

MAPs reproducibility was assessed using the standard error (SE) for replicate data (**Figure 2A**). Raw growth curves were compared to logistic growth curves to determine whether the program PMAnalyzer accurately parameterized and modeled clone growth during experimentation (data not shown). For further information on the accuracy and validity of the MAPs and PMAnalyzer see Cuevas *et al.*[24]

Following validation of the method, MAPs data was analyzed using the multiple parameters, such as maximum growth rate (μ_{max}) and growth level (GL), provided by the processing pipeline. Comparative visualization of growth curves is often used for growth data interpretation; however, the number of curves that can be visualized at a time for the comparison has limitations. To analyze numerous growth curves simultaneously, heat map derived plots were implored to compare tens of clones grown on a single substrate against that conditions' average response (**Figure 2B**). The influence placed by overexpression of a novel phage protein is observed through changes in curve parameters, specifically: lag phase, exponential phase and the maximum biomass yield (asymptote). As an example, the steep climb from lag phase into exponential phase modeled in the growth curve for the Capsid protein (**Figure 2A**) is reproduced by a quick change in color intensity from black to white for the same clone in the dynamic plot of **Figure 2B**.

To obtain a global picture of clone distribution across substrates, phenotypic classifications derived from the GL were used (**Figure 3**). Here, the four phenotypes are separate into four charts where the height of each bar represents the number of clones displaying that phenotype for a specific substrate. Outliers in the data are recognized as clones falling into the "gain of function" or "loss of function" category. Outliers can then be sought out individually and investigated more closely experimentally. In addition, global analysis recognizes substrate biases in the assay. Substrates such as phenylalanine, malic acid, and glycine resulted in a "no growth" classification. Substrates that consistently fall into the no growth classification, across all clones, are not heavily weighted in downstream functional characterization.

Metabolomics

The catabolic products from clones expressing unknown phage genes were identified using metabolomics. Briefly, clones were grown under either a continuous culture or serial passage in batch culture environment prior to being sent out for GC-TOFMS analysis at a metabolomics core facility. For details on sample processing, analysis, and normalization for GC-TOFMS implemented by the chosen core facility see Fiehn *et al.*[31] Briefly, 1 ml of cold extraction solvent is added to each sample, after which samples are vortexed and sonicated in a cold bath for 5 min. Samples are finally centrifuged and half of the sample is decanted and dried down for analysis. Extracts are purified and spiked with internal retention index markers before being loaded onto the gas chromatograph and then subsequently transferred to the mass spectrometer. Data from each sample is analyzed such that signal intensities for all detected signals in the chromatogram are reported. For normalization, the abundance of peaks for each sample is summed and the total peak abundances are averaged between all samples in the set. Metabolite abundances per sample are divided by the sample's peak abundance and then multiplied by the average peak abundance of the sample set. The resulting data is used for the analysis of metabolomics in the research discussed.

Validation of metabolomics reproducibility was required to determine the appropriate sample size for each culturing method. To detect the precision seen within samples and the variation seen across sample sizes the standard error of the mean (S_M) for both n = 3 and n = 6 datasets were reviewed (**Figure 5B**). Regardless of the continuous culture (CC) sample size, less than 1% of the data had a $S_M \leq 1.5$. Median S_Ms were 221 and 300, and values ranged from 0 to 7.55×10^5 and 3.74×10^5, for n = 3 and n = 6 respectively. S_Ms were also calculated for each set of sample replicates in the serial culture (SC) method. Again, less than 1% of the data had a $S_M \leq 1.5$, a median S_M of 137, and a range from 0 to 3.51×10^5. To compare the S_M distributions between each sample set (CC n = 3 *vs.* CC n = 6, CC n = 3 *vs.* SC n = 3, and CC n = 6 *vs.* SC n = 3) a permutation test was performed. The distribution of either continuous culture S_M values dataset was not significantly different from the distribution of the serial culture S_M values (p-value = 0.0). However, the distribution of S_M values for continuous culture n = 3 data was significantly different from that of the S_M values for the continuous culture n = 6 data (p-value = 1.908804×10^{-49}). Lastly, the coefficient of variation per metabolite was compared before and after implementation of a quality assurance (QA) step (**Figure 5C**). In total, 210 metabolites were removed after implementation of the QA pipeline (40% of the data). Less than 1% of the removed data had a metabolite abundance of zero, ~2% was internal standard data, ~5% was data from metabolites never before observed in *E. coli*, and the remaining metabolites (> 30%) had a coefficient of variation greater than 1.

As with the MAPs analysis, global observations provided an initial understanding of the depth of information metabolomics offers. To obtain a global picture, clones were hierarchically clustered based on their relative metabolite abundances providing information on clone-metabolite profiles, potential clones with related functions, and clone-metabolite outliers (**Figure 6**). To highlight protein functions, metabolites are separated and grouped based on common metabolic pathways. Using this analysis with preliminary results, it was evident that metabolomics is capable of separating genes from different classes (**Figure 6**, highlighted clones). In addition, outlier identification with the metabolomics data was determined by calculating standard scores (z scores) for each clone-metabolite pair. To ensure statistical significance, outliers were defined as a clone-metabolite pair with a Z score value of 2, accounting for only 5 percent of the data (data not shown).

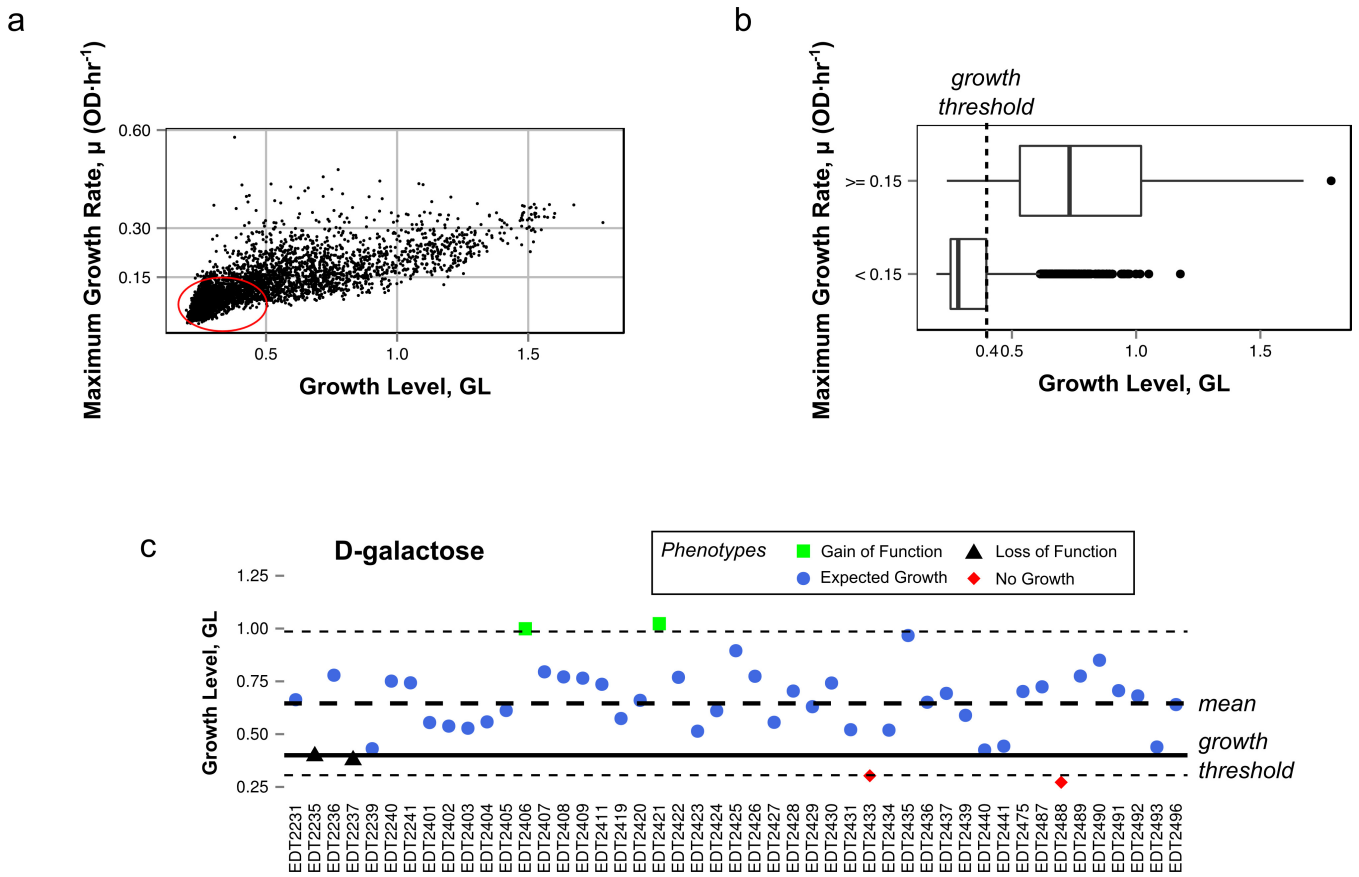

Figure 1. Definitions of phenotype classifications. (**A**) Relationship between the growth level (GL) and maximum growth rate. Data points circled in red represent growth curves showing little to no substrate utilization. (**B**) Boxplot representation defining growth threshold based on the distribution of growth curves with a minimal growth rate (< 0.15 OD/hr). (**C**) The variance and standard deviation of GL is calculated for substrate D-galactose. Short dashed lines represent two standard deviations away from the mean. Please click here to view a larger version of this figure.

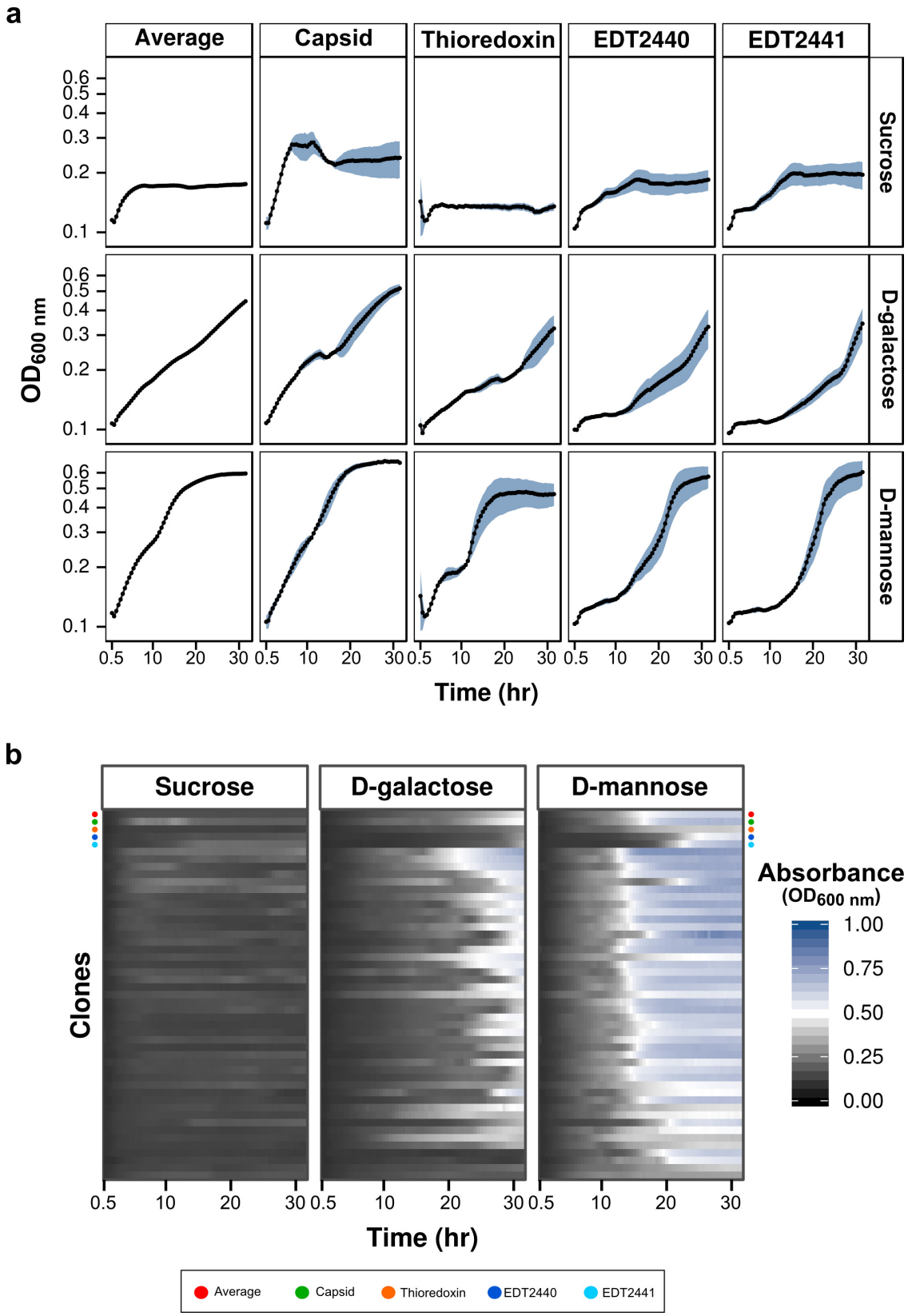

Figure 2. MAPs validation via precision and differentiation. (**A**) Growth curves for annotated structural (Capsid) and metabolic clones (Thioredoxin), two novel metabolic clones (EDT2440, EDT2441), and the average response of clones grown on sucrose, D-galactose and D-mannose in the MAPs. Blue lines indicate the standard error seen between replicate data (n = 3). (**B**) The growth curves for 47 different clones are represented as heat maps for sucrose, D-galactose and D-mannose. The annotated structural (green circle) and metabolic clones (orange circle), two novel metabolic clones (dark and light blue circles), and the average response (red circle) are highlighted. Please click here to view a larger version of this figure.

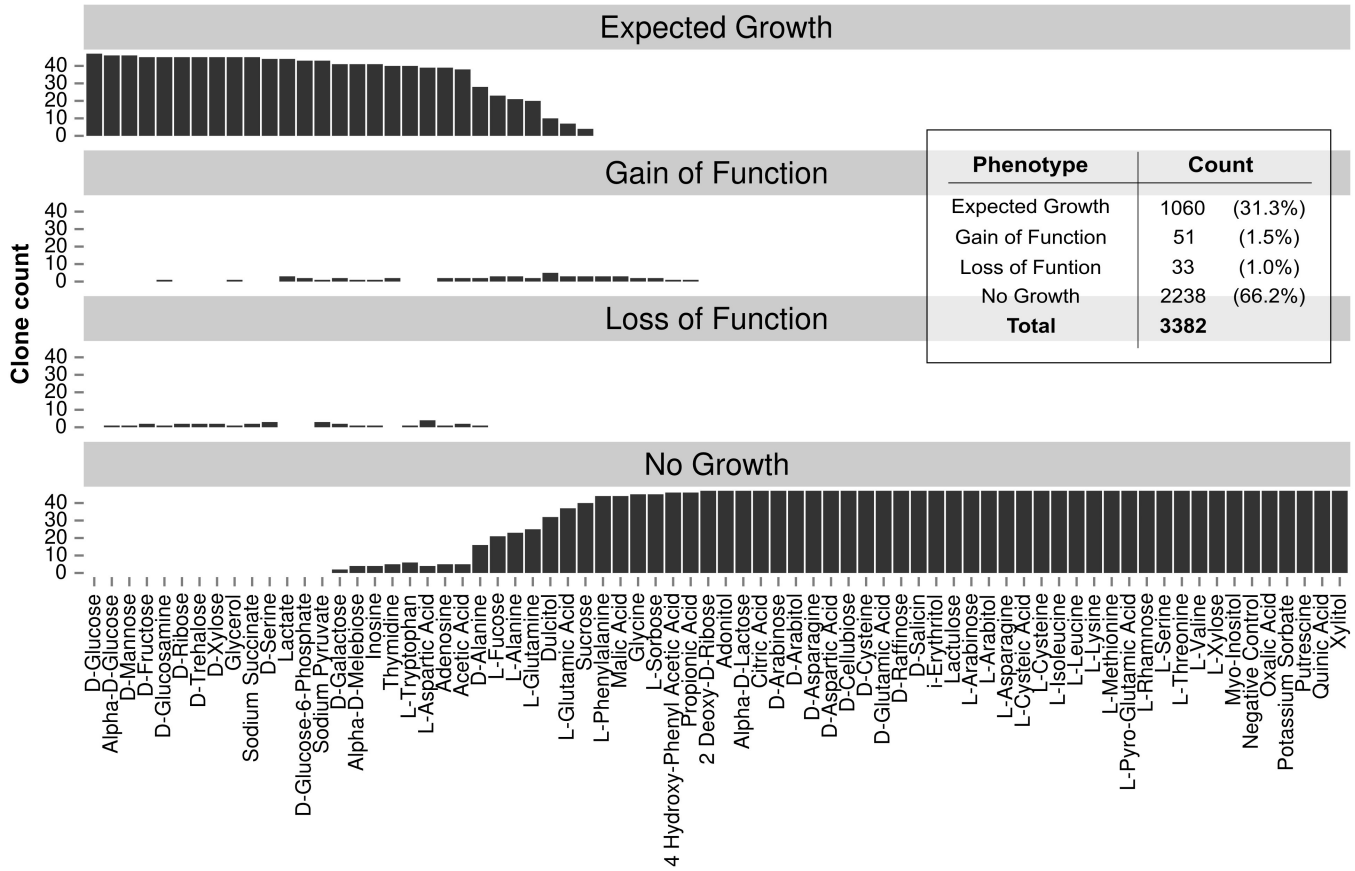

Figure 3. Clone distribution for each phenotype across multiple substrates. The phenotype-clone count for 47 clones across 72 carbon-specific growth conditions. Table provides direct counts for each phenotype. Please click here to view a larger version of this figure.

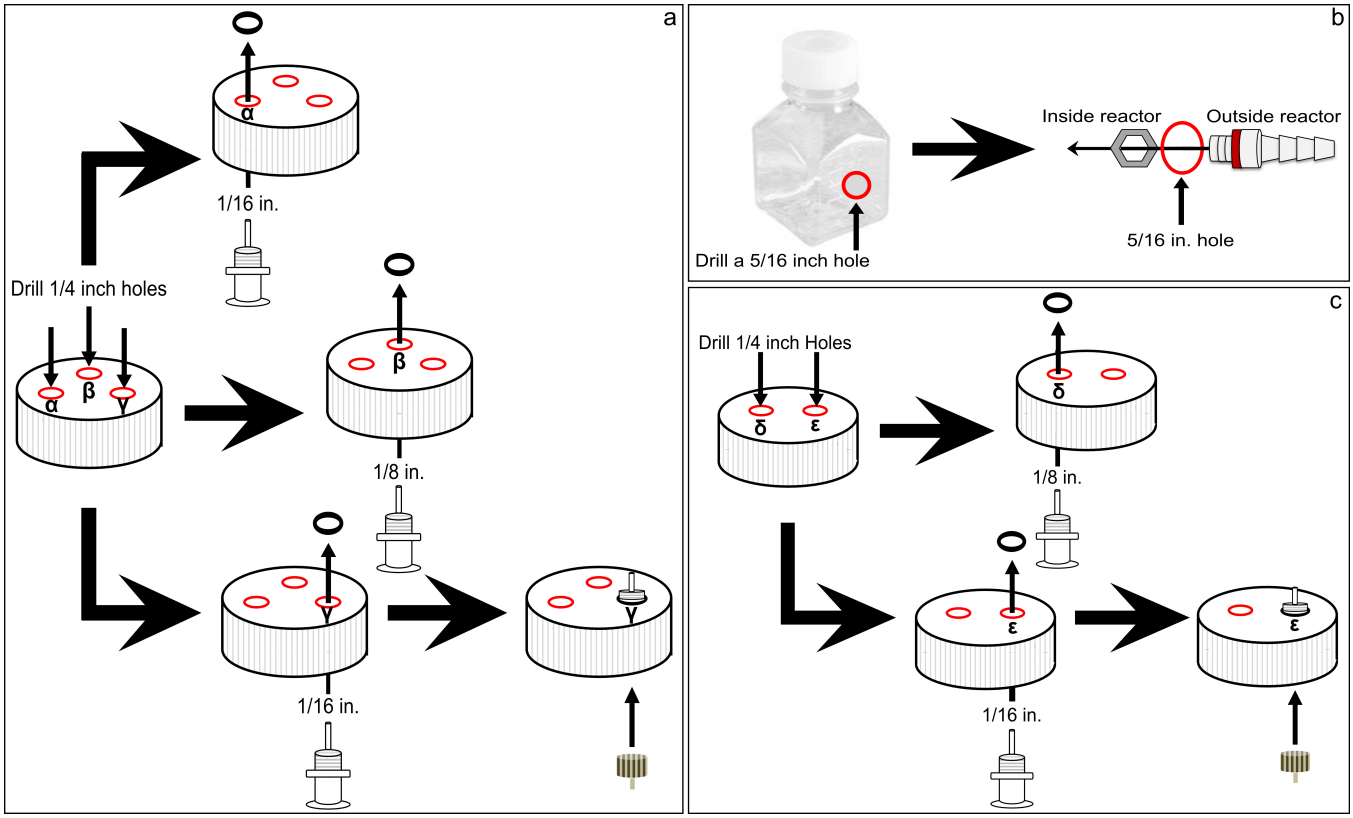

Figure 4. Diagram detailing the construction of the continuous culture apparatus. (**A**) Steps used to build ports α-γ of the continuous culture reactor, (**B**) steps used to build the out flow port of the continuous culture reactor, and (**C**) the steps to build ports δ and ε of the continuous culture feeding bottle. Please click here to view a larger version of this figure.

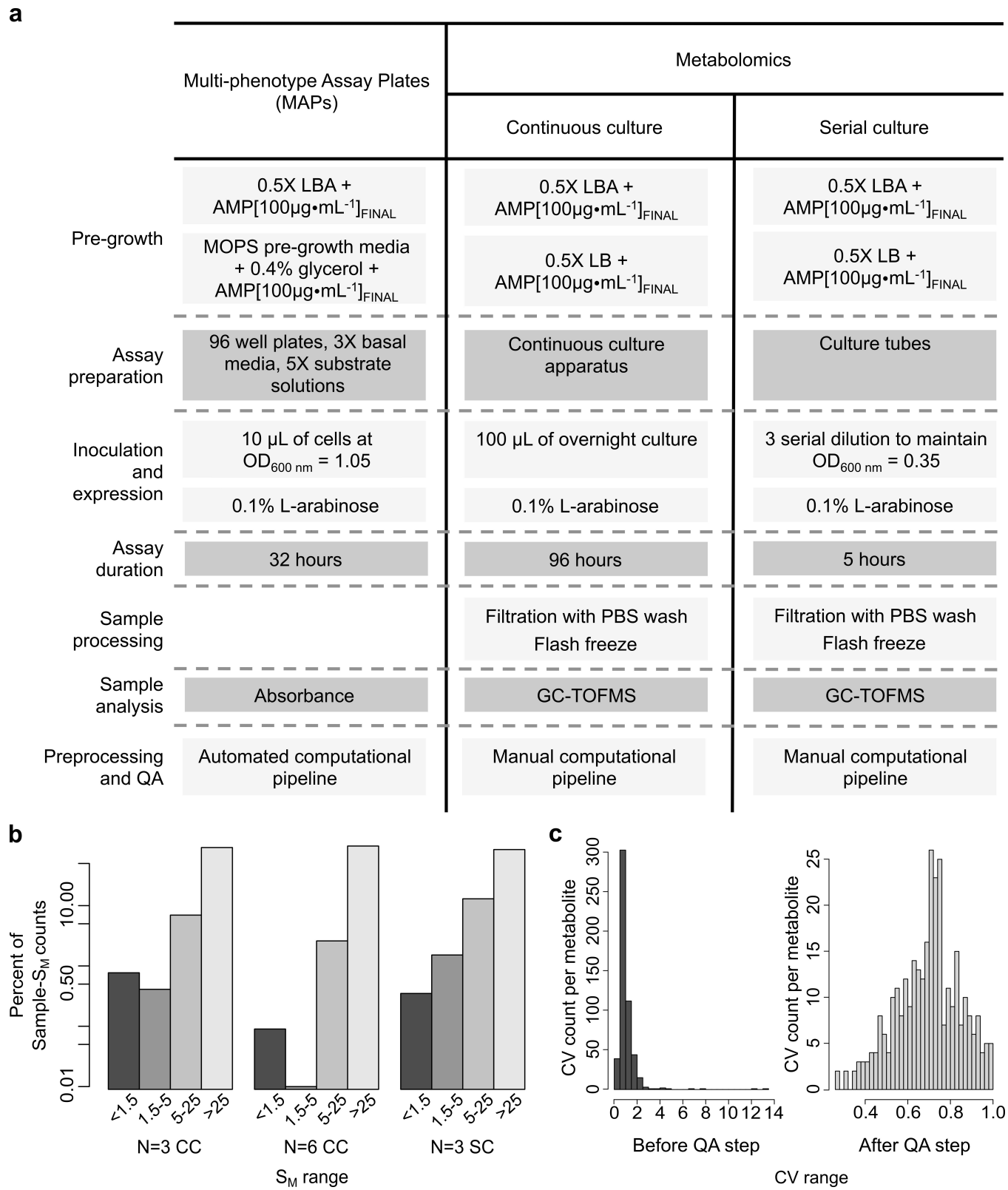

Figure 5. Comparison of the phenomic methods presented. (**A**) Workflow for the preparation of the Multi-phenotype Assay Plates (MAPs), continuous cultures and serial cultures. (**B**) The percentage of Standard error of mean (S_M) counts for both n = 3 and n = 6 sample sizes for the continuous culture (CC) and serial culture (SC) preparation methods for metabolomics. The y-axis is on a log scale. (**C**) The distributions of coefficients of variation (CV) per metabolite, before and after implementation of the QA pipeline for the continuous culture (CC) method. Please click here to view a larger version of this figure.

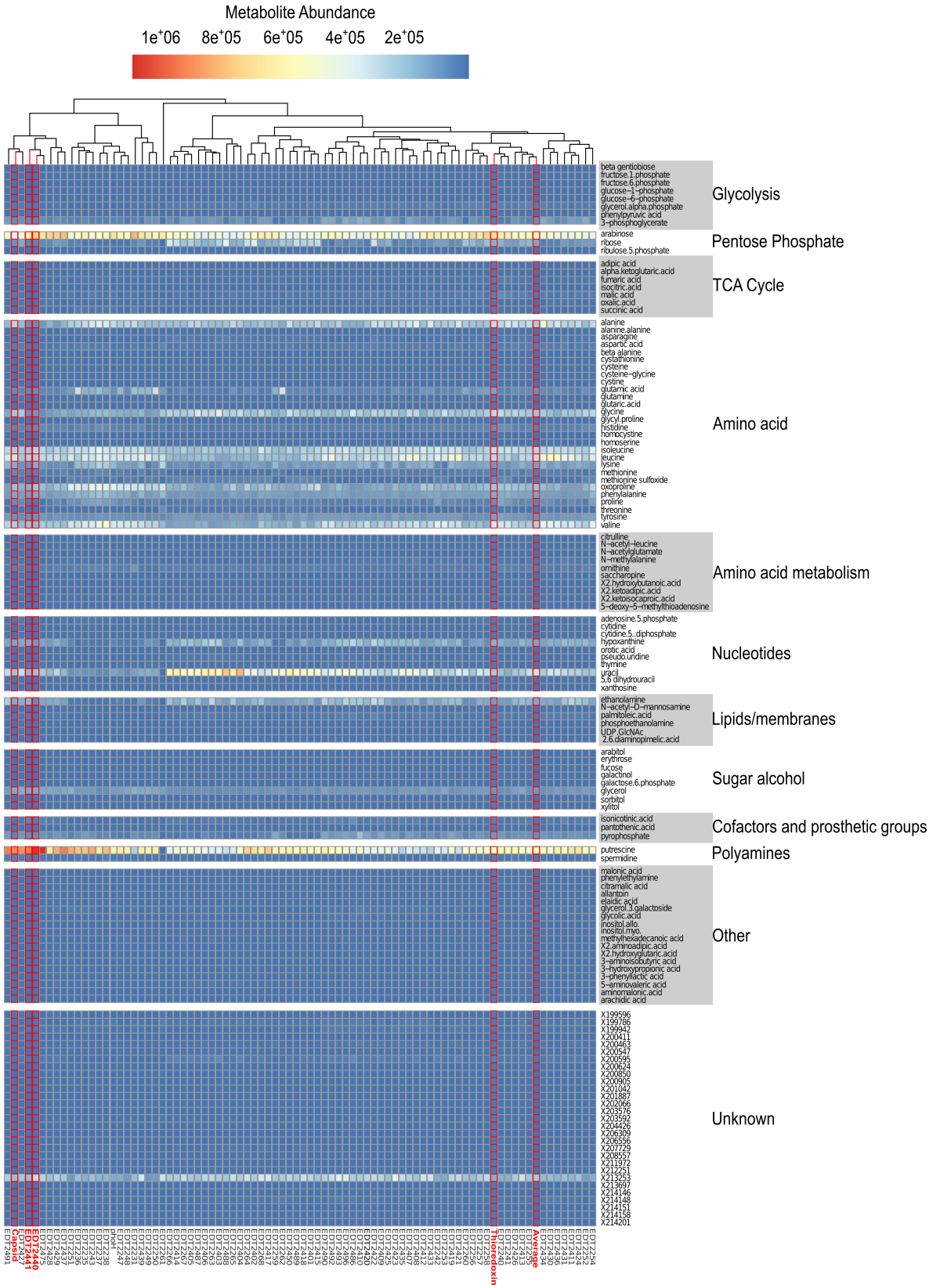

Figure 6. Metabolomic profiles of clones grown in continuous culture. The median metabolite abundances for a set of metabolites are plotted for 84 clones grown in continuous cultures. Metabolite profiles for annotated structural (Capsid) and metabolic clones (Thioredoxin), two novel metabolic clones (EDT2440, EDT2441), and the average metabolic response are highlighted in red. Please click here to view a larger version of this figure.

Compound	Carbon	Nitrogen	Sulfur	Phosphorus
Glycerol	–	0.40%	0.40%	0.40%
Ammonium chloride	9.5 mM	–	9.5 mM	9.5 mM
Sodium sulfate	0.250 mM	0.250 mM	–	0.250 mM
Magnesium sulfate	1.0 mM	1.0 mM	–	1.0 mM
Potassium phosphate	1.32 mM	1.32 mM	1.32 mM	–
Magnesium chloride	–	–	*	–
Potassium chloride	10 mM	10 mM	10 mM	10 mM
Calcium chloride	0.5 µM	0.5 µM	0.5 µM	0.5 µM
Sodium chloride	5 mM	5 mM	5 mM	5 mM
Ferric chloride	6 µM	6 µM	6 µM	6 µM
L- arabinose	0.10%	0.10%	0.10%	0.10%
MOPS pH 7.4	1x	1x	1x	1x

Table 1. The compounds and concentrations of the different basal media used in the MAPs. *1.0 mM of magnesium chloride is substituted. 1x MOPS = 40 mM MOPS, 4 mM Tricine.

Carbon substrates	Nitrogen substrates	Sulfur substrates	Phosphorus substrates
2 deoxy-D-ribose	2-deoxy-D-ribose	1-butane-sulfonic acid	adenosine-5-monophoshate
4 hydroxy-phenyl acetic acid	acetamide	acetyl cysteine	beta-glycerophosphate
acetic acid	adenine	D-cysteine	creatinephosphate
adenosine-5-monophosphate	adenosine	D-methionine	D-glucose-6-phosphate
adonitol	allantoin	diethyl-dithiophosphate	diethyl-dithiophosphate
alpha-D-glucose	beta-phenylethylamine	DL-ethionine	DL-alpha-glycerophosphate
alpha-D-lactose	biuret	glutathione	potassium phosphate
alpha-D-melebiose	cytidine	isethionic acid	sodium pyrophosphate
citric acid	cytosine	L-cysteic acid	sodium thiophosphate
D-alanine	D-alanine	L-Cysteine	
D-arabinose	D-asparagine	L-djenkolic acid	
D-arabitol	D-aspartate	L-methionine	
D-asparagine	D-cysteine	magnesium sulfate	
D-aspartate	D-glucosamine	methane sulfonic acid	
D-cellubiose	D-glutamic acid	N-acetyl-DL-methionine	
D-cysteine	DL-alpha-amino-N-butyric acid	N-acetyl-L-cysteine	
D-fructose	D-methionine	potassium-tetra-thionate	
D-galactose	D-serine	sodium thiosulfate	
D-glucosamine	D-valine	sulfanic acid	
D-glucose	gamma-amino-N-butyric acid	taurine	
D-glucose-6-phosphate	glycine	taurocholic acid	
D-glutamate	guanidine	thiourea	
D-mannose	histamine		
D-raffinose	inosine		
D-ribose	L-alanine		
D-salicin	L-arginine		
D-serine	L-asparagine		
D-trehalose	L-citrulline		
D-xylose	L-cysteine		

dulcitol	L-glutamic acid
glycerol	L-glutamine
glycine	L-glutathione
i-erythritol	L-histidine
inosine	L-isoleucine
L-alanine	L-leucine
L-arabinose	L-lysine
L-arabitol	L-methionine
L-asparagine	L-ornithine
L-aspartate	L-phenyl-alanine
L-cysteic acid	L-proline
L-cysteine	L-pyro-glutamic acid
L-fucose	L-serine; L-threonine
L-glutamic Acid	L-tryptophan
L-glutamine	L-valine
L-isoleucine	N-acetyl-D-glucosamine
L-leucine	putrescine
L-lysine	thiourea
L-methionine	thymidine
L-phenylalanine	thymine
L-pyro-glutamic acid	tyramine
L-rhamnose	tyrosine
L-serine	uridine
L-sorbose	
L-threonine	
L-tryptophan	
L-valine	
L-xylose	
lactate	
lactulose	
malate	
myo-inositol	
oxalic acid	
potassium sorbate	
propionic acid	
putrescine	
quinic acid	
sodium pyruvate	
sodium succinate	
sucrose	
thymidine	
xylitol	

Table 2. List of substrates used in the MAP experiments.

Discussion

Here, we present phenomic approaches for the functional characterization of putative phage genes. Techniques include a developed assay capable of monitoring host anabolic metabolism, the Multi-phenotype Assay Plates (MAPs), in addition to the established method of metabolomics, capable of measuring effects to catabolic metabolism. We provided additional tools to manage the large data sets resulting from these technologies, allowing for high throughput processing and analysis[24]. Lastly, through the comparison of an annotated phage capsid protein, phage thioredoxin, two putative metabolic phage genes, and the average experimental response we propose various strategies to interpret both datasets and gene classes, with emphasis on identification of phenotypic trends and identification of outliers.

As mentioned, both approaches quantitatively measure only half of host metabolism. To interpret the relative function of any of the novel proteins under investigation, data from both methods is required to provide evidence of function. While this is not a focus of our current manuscript, data outputs from each phenomic method is put through combinatory analyses that focus on clustering techniques such as random forest and principal component analysis. Furthermore, hypotheses resulting from the combined analysis must be subsequently validated by traditional genetic methodologies.

Finally, the methods presented are heavily influenced by bacterial physiology and therefore follow the same standards. When undertaking either method, considerations need to be made to ensure independent, clonal groups are experimented with; contamination is prevented; a single variable is being tested; and appropriate controls are being ran simultaneously. Failure to account for these points will result in unclear results, similar to any physiological assay.

Multi-phenotype Assay Plates (MAPs)

The development of MAPs provides a high throughput and adaptable assay compared to technologies currently available (**Figure 5A** and **Tables 1,2**). The assay uses supplies, equipment, and fundamental techniques available in all microbiology labs. The incorporation of a computational pipeline, PMAnalyzer[24], for subsequent data processing and analysis ensures rapid data interpretation. In addition, both experimental and analytical aspects of the approach can be readily adjusted or tuned for customized purposes. For example, if a large proportion of the data fails to pass filtering outlined in section 4, one can manually sift through the growth curves to identify issues. If the problem arises due to stringent filter parameters, adjustments to the script can be made. Alternatively, if problems are associated with the experimental process (*i.e.,* prolonged condensation; improper transferring of bacterial cells, etc.) then additional replicates can be readily repeated.

As described in Cuevas *et al.*[24], the PMAnalyzer is a single bash program written as a wrapper script that executes the parsing and analysis scripts as a cohesive, automated pipeline. All scripts are freely accessible from a Git repository at [25] by taking the median value for each time point across triplicate data, and subsequently parameterizes the logistic curve to obtain the lag time, maximum growth rate, asymptote, and a novel term, Growth Level. The median value was chosen over the mean in our study to reduce the effect of large outliers, however, the script can be readily adapted to calculate the mean of replicate data. Due to reduced variation (SE) seen across replicate data (**Figure 2A**) we maintained the use of the median in the PMAnalyzer for fitting a logistic curve. Additionally, the cut off for growth in this study (GL ≥ 0.4) was determined by comparing how data separated across Growth Level and maximum growth rate (**Figure 1A,B**). Depending on the instruments and model system used this term may vary, requiring redefinition of this cut off value.

A major advantage of our assay is the ability to compare phenotypes using a single parameter characterizing overall microbial growth, which we define as Growth Level (GL). GL is a harmonic mean, and therefore mitigates the effects of large outliers in the data. The use of a harmonic mean with shifted logistic-fitted values to provide a summary of growth was arrived at through trial and error. Other methods attempted to differentiate growth included: time it took to reach specific curve parameters (half μ_{max}, μ_{max}, and carrying capacity), the coefficient of determination (R^2), and combinations of the R^2 multiplied by specific curve parameters. Using a harmonic mean with shifted logistic-fit values for the GL provided the greatest range in evaluating growth, thus it became the method of choice. One consideration to note is that dynamic growth curve patterns have the potential of being lost when using a single parameter or a fitted model. For instance, the individual curve parameters of the logistic curve and the GL are incapable of representing biphasic growth. In a single carbon environment, this effect on growth implies mediation of the viral protein on either conversion of the substrate or shift in substrate utilization. Additional effects potentially lost when not considering multiple growth parameters include: prolonged lag time, proposing an increased burden of viral machinery or products; rapidly accelerating exponential phase, suggesting viral proteins coupled to host energy production pathways; or higher levels of biomass formation, implying viral support in host nutrient uptake and anabolism (data not shown). Thus, plotting nascent growth curves (**Figure 2A,B**) provides information regarding trends over time whereas the GL takes into account the major variables of the logistic model, providing a single quantitative number to represent overall success of a clone.

When considering the different responses contributed by structural and metabolic genes in the MAPs, it is observed that the different substrate classes in question provide the greatest evidence for protein function. For example, metabolic proteins are often associated with acquisition of limiting nutrients, which are unspecific to host central metabolism[16,32]. Preliminary MAP experiments reveal that clones harboring putative metabolic phage genes have an increased lag phase when grown on central metabolism carbon sources (**Figure 2A**). Conversely, clones carrying putative structural genes, which require large proportions of host energy and dNTP pools, result in a false positive response on growth for central and amino acid metabolism carbon substrates. This is likely due to the accumulation of insoluble proteins resulting in host filamentation and/or inclusion bodies, as observed via microscopy (**Figure 2A** and data not shown). While further analysis is required to validate these preliminary results, the MAPs are capable of retrieving phenotypic responses that correlate to hypothesized functions of specific phage gene classes.

In addition to the elucidation of unknown viral proteins, the MAPs are a novel resource to investigate the functional and metabolic diversity of an individual bacterium or a community of bacteria. MAP components are designed for easy alteration to support the growth of a range of bacteria; including marine, auxotrophic, and anaerobic microbes. To facilitate these efforts the defined basal and pre-growth media require additional or

adjusted chemical species before a different bacterial genus can be supported in the MAPs. One note in this use of the MAPs is to maintain defined media, prohibiting the use of ingredients such as tryptone, yeast extract and peptone.

Metabolomics

The field of metabolomics is dependent on metabolite databases, which include isolated metabolites identified by mass spectrometry. The core facility chosen here has one of the largest metabolomics databases. Interestingly, more than half of the metabolites resulting from our experimentations were unidentifiable (~65%), while others had never before been recorded in our host, *Escherichia coli* (examples include: Indole 3 acetic acid[33], salicylic acid[34], and dihydroabietic acid[35]). This fact could be attributed to either a strong bias of the database towards plant metabolites, or the specific proteins under investigation. Regardless, the result is a limited number of known metabolites available for data representation and analysis. In the future, multiple metabolomics methods using various databases would allow for greater metabolite coverage.

Presently, both known and unknown metabolites are used when comparing and contrasting our novel viral proteins. Using this approach, we hypothesize that clones harboring functionally similar proteins will share an increased similarity in their complete metabolomic profile. Preliminary metabolomics analysis revealed that while structural and metabolic genes do not clearly separate from one another, those genes exhibiting similar effects on the host when overexpressed do correlate (**Figure 6**). For example, the annotated Capsid gene clusters closely with the putative metabolic genes highlighted in this study, EDT2440 and EDT2441. Investigations using a publicly available transmembrane topology and signal peptide predictor program showed evidence that both putative metabolic genes harbor a single transmembrane domain. Interestingly 5 out of the 9 clones in the first cluster group (most left portion of the dendrogram) have predicted transmembrane domains using the same topology program. Further investigations are needed, however, it is likely that the metabolites present during the overexpression of these clones are associated with cellular stress response resulting from membrane or structural burdens. This evidence supports that while the metabolomics data possesses an increased amount of noise, the method is capable of highlighting signals that differentiate general effects of genes, both within and across a gene class. To determine whether the method is capable of extracting out specific information of gene function, metabolites were grouped into specific metabolic pathways. The hypothesis being, if a clone affects metabolites specific to a single pathway, then the overexpressed gene is active in that pathway. Prior to the establishment of our metabolomics quality assurance pipeline, preliminary data revealed that over and underrepresented metabolites were typically “unknown”, providing little information on the pathways they are associated with (data not shown). Preprocessed metabolomics data, however, reveals that the majority of the metabolite profiles are similar and only a select number of unknown and known metabolite abundances vary across clones, for instance putrescine and uracil (**Figure 6**). To provide greater resolution of protein function efforts are being made to experimentally compare the novel phage genes against known phage genes, which can be used to fill in the “holes” of metabolite based functional characterization. Using this technique, the assigned function of known viral genes provides a reference for the function of the unknown genes. Nonetheless, the limiting factor of metabolomic analysis is the size and relevance of the database. To correct for these limitations, metabolomic databases relatable to this research need to be developed; such as a database of metabolites and their abundances specific to the ASKA collection of *E. coli* clones in which a single ORF is overexpressed[36]. Evidence for the need of such databases was provided in 2013 when researchers at the Lawerence Berkeley National Laboratory compiled the first comprehensive database of metabolites specific to entire mutant libraries of model bacteria[37]. This research provided novel insight into genes required for utilization of specific metabolites, revealing the clear connection between phenotype and genotype.

When considering metabolomics as a tool, it is important to define the processing regime followed at the core facility. An artifact of most experimental procedures is the day-to-day variance associated with the instruments of use. To date all GC-MS analysis implements the use of internal standards that are included in each analytical run; however, addition of project specific internal *samples* ran each day of experimentation removes additional variance. These considerations must be addressed early to avoid normalization problems and biases. Another solution is to process all samples at a core facility on the same machine and as a single batch, an option available at any core facility.

The various tools both introduced and re-explored in this manuscript provide novel means to screen and characterize functionally unknown phage genes. The simplicity and adaptability of the experimental techniques with the streamline use of computational pipelines assures these methods are applicable to a broad range of research endeavors and fields. Our goal is that the phenomic approaches presented here will aid further investigations of novel phage proteins in addition to systems that are equally functionally undefined.

Disclosures

The authors have nothing to disclose.

Acknowledgements

We thank Benjamin Knowles, Yan Wei Lim, Andreas Haas, and members of the Viral Dark Matter consortium for their help and constructive input on this manuscript. This research is funded by the National Science Foundation (DEB-1046413) and is part of the *Dimensions: Shedding Light on Viral Dark Matter* project.

References

1. Wommack, K. E., Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews.* **64**, (1), 69-114 (2000).
2. Hendrix, R. W. Recoding in Bacteriophages. *Recoding: Expansion of decoding rules enriches gene expression.* **24**, 249-258 (2010).
3. Breitbart, M., *et al.* Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences.* **99**, (22), 14250-14255 (2002).
4. Breitbart, M., *et al.* Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings. Biological sciences / The Royal Society.* **271**, (1539), 565-574 (2004).
5. Dinsdale, E. A., *et al.* Functional metagenomic profiling of nine biomes. *Nature.* **452**, (7187), 629-632 (2008).

6. Vega Thurber, R. L., *et al.* Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proceedings of the National Academy of Sciences of the United States of America.* **105**, (47), 18413-18418 (2008).
7. Pedulla, M. L., *et al.* Origins of highly mosaic mycobacteriophage genomes. *Cell.* **113**, (2), 171-182 (2003).
8. Calendar, R., Abedon, S. T. *The bacteriophages.* Oxford Univeresity Press (2006).
9. Breitbart, M., Miyake, J. H., Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS microbiology letters.* **236**, (2), 249-256 (2004).
10. Klenk, H. -P., Palm, P., Zillig, W. DNA-Dependent RNA Polymerases as Phylogenetic Marker Molecules. *Systematic and Applied Microbiology.* **16**, (4), 638-647 (1993).
11. Breitbart, M., Thompson, L., Suttle, C., Sullivan, M. Exploring the Vast Diversity of Marine Viruses. *Oceanography.* **20**, (2), 135-139 (2007).
12. Mann, N. H., Cook, A., Millard, A., Bailey, S., Clokie, M. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature.* **424**, (6950), 741 (2003).
13. Mann, N. H., *et al.* The genome of S-PM2, a “photosynthetic” T4-type bacteriophage that infects marine Synechococcus strains. *Journal of bacteriology.* **187**, (9), 3188-3200 (2005).
14. Lindell, D., *et al.* Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proceedings of the National Academy of Sciences of the United States of America.* **101**, (30), 11013-11018 (2004).
15. Millard, A., Clokie, M. R. J., Shub, D. A., Mann, N. H. Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. *Proceedings of the National Academy of Sciences of the United States of America.* **101**, (30), 11007-11012 (2004).
16. Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., Chisholm, S. W. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS biology.* **3**, (5), e144 (2005).
17. Rohwer, F., *et al.* The complete genomic sequence of the marine phage Roseophage SIOI shares homology with nonmarine phages. *Limnology and Oceanography.* **45**, (2), 408-418 (2000).
18. Miller, E. S., *et al.* Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *Journal of bacteriology.* **185**, (17), 5220-5233 (2003).
19. Thompson, L. R., *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America.* **108**, (39), E757-E764 (2011).
20. Paterson, S., *et al.* Antagonistic coevolution accelerates molecular evolution. *Nature.* **464**, (7286), 275-278 (2010).
21. Seguritan, V., *et al.* Artificial neural networks trained to detect viral and phage structural proteins. *PLoS computational biology.* **8**, (8), e1002657 (2012).
22. Neidhardt, F. C., Bloch, P. L., Smith, D. F. Culture Medium for Enterobacteria. *Journal of Bacteriology.* **119**, (3), 736-747 (1974).
23. Khlebnikov, A., Datsenko, K. A., Skaug, T., Wanner, B. L., Keasling, J. D. Homogeneous expression of the P(BAD) promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology (Reading, England).* **147**, (Pt 2), 3241-3247 (2001).
24. Cuevas, D. A., *et al.* Elucidating genomic gaps using phenotypic profiles. *F1000Research.* (2014).
25. Zwietering, M. H., Jongenburger, I., Rombouts, F. M., van’t Riet, K. Modeling of the bacterial growth curve. *Applied and environmental microbiology.* **56**, (6), 1875-1881 (1990).
26. Venables, W. N., Smith, D. M. *An introduction to R.* Available from: http://cran.r-project.org/doc/manuals/R-intro.pdf (2014).
27. Schmieder, R., Lim, Y. W., Rohwer, F., Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC bioinformatics.* **11**, 341 (2010).
28. Schmieder, R., Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England).* **27**, (6), 863-864 (2011).
29. Schmieder, R., Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One.* **6**, (3), e17288 (2011).
30. Welch, M., *et al.* Design parameters to control synthetic gene expression in Escherichia coli. *PloS One.* **4**, (9), e7002 (2009).
31. Fiehn, O., *et al.* Quality control for plant metabolomics: reporting MSI-compliant studies. *The Plant journal: for cell and molecular biology.* **53**, (4), 691-704 (2008).
32. Zeng, Q., Chisholm, S. W. Marine viruses exploit their host’s two-component regulatory system in response to resource limitation. *Current biology: CB.* **22**, (2), 124-128 (2012).
33. Shindy, W. W., Smith, O. E. Identification of plant hormones from cotton ovules. *Plant physiology.* **55**, (3), 550-554 (1975).
34. Lee, H. I., Leon, J., Raskin, I. Biosynthesis and metabolism of salicylic acid. *Proceedings of the National Academy of Sciences of the United States of America.* **92**, (10), 4076-4079 (1995).
35. Chappell, J. The Biochemistry and Molecular Biology of Isoprenoid Metabolism. *Plant Physiology.* **107**, (1), 1-6 (1995).
36. Kitagawa, M., *et al.* Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA research: an international journal for rapid publication of reports on genes and genomes.* **12**, (5), 291-299 (2005).
37. Baran, R., *et al.* Metabolic footprinting of mutant libraries to map metabolite utilization to genotype. *ACS chemical biology.* **8**, (1), 189-199 (2013).