A targeted proteomics toolkit for high-throughput absolute quantification of *Escherichia coli* proteins

Tanveer S. Batth^{1,2,‡,§}, Pragya Singh^{1,2,§}, Vikram R. Ramakrishnan^{1,3}, Mirta M. L. Sousa^{1,4,5}, Leanne Jade G. Chan^{1,2}, Huu M. Tran^{1,6}, Eric. G. Luning², Eva H. Y. Pan^{1,2}, Khanh M. Vuu^{1,2}, Jay D. Keasling^{1-3,7}, Paul D. Adams¹⁻³, Christopher J. Petzold^{1,2,*}

¹Joint BioEnergy Institute, Emeryville, California 94608, USA.

²Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720 USA.

³Department of Bioengineering, University of California, Berkeley, California 94720, USA.

⁴Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.

⁵The Proteomics and Metabolomics Core Facility (PROMEC) at NTNU and the Central Norway Regional Health Authority, Trondheim, Norway.

⁶Sandia National Laboratories, Livermore CA 94550 USA

⁷Department of Chemical & Biomolecular Engineering, University of California, Berkeley, California 94720, USA.

[‡]Current address: Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Faculty of Health and Medical Sciences, Blegdamsvej 3b, building 6.1, DK-2200 Copenhagen N

[§]These authors contributed equally to this work.

*Correspondence should be addressed to: C.J.P. (cjpetzold@lbl.gov)

Keywords: Targeted proteomics, High-throughput proteomics, E. coli, Peptide standards, Absolute protein quantification

Abstract

Transformation of engineered *Escherichia coli* into a robust microbial factory is contingent on precise control of metabolism. Yet, the throughput of 'omics technologies used to characterize cell components have lagged far behind our ability to engineer novel strains. To expand the utility of quantitative proteomics for metabolic engineering, we validated and optimized targeted proteomics methods for over 400 proteins from more than 21 major pathways in *E. coli* metabolism. Complementing these methods, we constructed a series of synthetic genes to produce concatenated peptides (QconCAT) for absolute quantification of the proteins and made them available through the Addgene plasmid repository (www.addgene.org). To facilitate high sample throughput, we developed a fast, analytical-flow chromatography method using a 5.5-minute gradient (10 minute total run time). Overall this toolkit provides an invaluable resource for metabolic engineering by increasing sample throughput, minimizing development time and providing peptide standards for absolute quantification of *E. coli* proteins.

1. Introduction

Recent developments in metabolic engineering such as automated DNA assembly (Hillson et al., 2011) and techniques for predictable gene expression (Mutalik et al., 2013) have accelerated pathway construction and optimization, respectively, producing novel engineered organisms more rapidly than ever before. Complementing these developments have been advances in analytical methods such as targeted proteomics to assess the success or failure of the pathway design and inform new design decisions (Alonso-Gutierrez et al., 2013; Juminaga et al., 2012; Redding-Johanson et al., 2011; Singh et al., 2012). Yet, once a pathway is optimized significant effort is still needed to engineer carbon and energy metabolism to produce commercially viable microbial factories (Van Dien, 2013). Tools capable of genome wide searches, such as MAGE (Wang et al., 2009) and TRMR (Warner et al., 2010), have greatly outpaced the throughput of technologies to measure their outputs (Boyle and Gill, 2012). And while product assays or screens provide data about the entire system, predictive knowledge to inform rational design of specific traits is severely limited. Enabling progress to predictive knowledge are systems biology methods, such as transcriptomics, proteomics, and metabolomics (Lanza et al., 2012; Oud et al., 2012; Smolke and Silver, 2011), however, low sample throughput and complex data acquisition limit their utility; consequently, they are generally used on an *ad hoc* basis.

Technological innovation has increased throughput for transcript analysis dramatically, yet for proteomics, most methods development focuses on depth of proteome coverage not sample throughput. However, Percy *et al.* showed that analytical-flow

chromatography, instead of the more commonly used nano-flow chromatography, provides minimal loss of sensitivity with the additional benefits of shorter gradients, reproducible retention times, and method robustness (Percy et al., 2012); characteristics that are essential for biotechnology research. Thus, to maximize the usefulness of targeted proteomics for metabolic engineering of *E. coli* we have designed methods based on analytical-flow chromatography mass spectrometry for quantification of pathway-specific proteins. Despite its importance as a host for a vast number of metabolic engineering studies and industrial processes there has been little attention paid to extensive targeted proteomics development for *E. coli*. We tested, validated, and optimized methods to targeted proteins from major metabolic pathways including glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway, mixed acid fermentation, many amino acid biosynthesis pathways, and fatty acid biosynthesis. Use of analytical-flow chromatography permitted development of a rapid method using a 5.5-minute gradient (10 minute total run time) with scheduled SRMs to quantify proteins for large numbers of samples. Complementing the toolkit, we constructed a series of concatenated peptides for absolute quantification of *E. coli* proteins (QconCATs) (Beynon et al., 2005; Brownridge et al., 2011; Pratt et al., 2006). The QconCATs, consisting of genetic sequences for over 800 tryptic peptides, were roughly organized by pathway and are publicly available through Addgene (Herscovitch et al., 2012) (www.addgene.org), as a resource for standard peptides. With this toolkit, absolute quantification of more than 400 proteins from *E. coli* can be achieved with little development time and rapid LC-MS/MS acquisitions. Here, we demonstrate the utility of the toolkit by characterizing proteins in various pathways from exponentially growing *E.*

coli cultured with glucose or xylose as the carbon source in aerobic and anaerobic conditions.

2.1 Materials and Methods

2.1 E. coli culture conditions

Escherichia coli DH1 (ATCC), a common cloning strain, was selected as base strain for all SRM experiments. Luria broth (LB) agar plates were used to grow single colonies which were picked for 10 ml LB cultures grown overnight at 37 °C shaking at 200 RPM. Cultures were diluted to 0.1 OD₆₀₀ (optical density measured at a wavelength of 600 nm) grown overnight in M9-MOPS medium (Neidhardt et al., 1974) with 1 % carbon source (Glucose, Xylose) at 37 °C with shaking. A second passage of the cultures, started at 0.1 OD₆₀₀ in M9-MOPS with 1% carbon source, was used for methods development. Semi-anaerobic cultures were grown similarly with the exception that the passages were initially sealed in glass culture tubes in an anaerobic chamber. For proteomics analysis, all cells were collected during growth phase at an OD₆₀₀ between 0.6-0.7 followed by centrifugation (5,000 x g) for 5 minutes at 9 °C. The medium was decanted and the cell pellet was frozen until further processing. For methods validation experiments, selected knockout strains from the KEIO knockout collection were picked for 1-ml cultures grown in EZ Rich medium with 2% glucose. The cultures were grown overnight in 96 deep well plates at 37 °C shaking at 900 RPM. Cultures were harvested after 18 hours at an OD₆₀ between 2.0-2.5. Cells were pelleted by centrifugation (4,000g) for 20 minutes at 4 °C.

The medium was decanted and the cell pellets were frozen at -80 °C until further processing.

2.2 Proteomic sample preparation

E. coli cell lysis and protein precipitation was accomplished using Chloroform/Methanol precipitation method as previously described (Wessel and Flugge, 1984). Briefly, 100 µL of concentrated cells were added to a fresh 1.7-mL tube to which 400 µL methanol was added followed by brief vortexing. 100 µL of chloroform, 300 µL of LC/MS grade water were subsequently added with vortexing after each addition. Tubes were then centrifuged at max speed on a benchtop centrifuge for 1 minute to allow for phase separation. The layer consisting of methanol and water was removed followed by the addition of 300 µL of methanol and vortexing to dislodge protein pellet. Tubes were then centrifuged at max speed for 1 minute and the chloroform and methanol layer was removed. The pellet was dried using a speed vac. One hundred (100) mM ammonium bicarbonate buffer with 10% methanol was used to re-suspend the pellet. The DC assay (BioRad, Hercules, CA) based on Lowry chemistry was used to determine the protein concentration. The protein mixture was diluted to a concentration of 0.5 mg/mL for protease digestion. Protein reduction was accomplished using 5 mM tris 2-(carboxyethyl)phosphine (TCEP) for 30 minutes at room temperature. Alkylation was performed with 10 mM iodoacetamide (IAM; final concentration) for 30 minutes at room temperature in the dark. Overnight digestion with trypsin was accomplished with a 1:10 trypsin:total protein ratio in 10% methanol and ammonium bicarbonate buffer at 37 °C. Following digestion, the protein

concentration was repeated on the peptide mixture by using the DC assay for added confidence.

2.3 Liquid Chromatography-Mass Spectrometry

All samples were injected and separated via an Agilent 1100 (SRM developmental and validation experiments) or 1260 (quantification experiments) liquid chromatography system (Agilent, Santa Clara, CA) operating in normal flow mode at 400 $\mu\text{L}/\text{min}$ or 800 $\mu\text{L}/\text{min}$ respectively. In both cases, ten μg of peptides were separated on an Ascentis Express Peptide C18 column (Schuster et al., 2012) (2.7 μm particle size, 160 Å pore size, 5 cm length x 2.1 mm i.d., coupled to a 5 mm x 2.1 mm i.d. guard column with similar particle and pore size, operating at 60 °C; Sigma-Aldrich, St. Louis, MO). For the Agilent 1100, a 22-minute method with the following gradient was used: 95% Buffer A (2 % acetonitrile (Honeywell Burdick & Jackson, Morristown, NJ), 0.1 % Formic Acid), 5 % Buffer B (98 % Acetonitrile, 0.1 % Formic Acid). Buffer B is increased to 40% over 17 minutes, followed by an increase to 95% B in 30 seconds, where it is held for 1 minute. Buffer B is ramped back down to 5% in 30 seconds and equilibrated for 3 minutes prior to the next injection. Similar buffers were used for separation on the Agilent 1260 experiment with a faster gradient and shorter total LC run time. Briefly, 95 % Buffer A as starting condition followed by an increase to 35 % B over 5.5 minutes. Subsequently, B was increased to 80 % in 0.5 minutes where it was held for 2 minutes followed by a rapid ramp back to 5 % B for a re-equilibration time of 2 minute(s).

For SRM methods development experiments the Agilent 1100 LC system was coupled to a 5500 Q-Trap Mass Spectrometer (AB Sciex, Foster City, CA) equipped with a Turbo V source and running Analyst version 1.5.2. For quantitative analysis and demonstration of multi-system utility an Agilent 1260 LC system was coupled to an Agilent 6460QQQ Mass Spectrometer equipped with an Agilent Jet Stream source and running MassHunter version B.05.00. For the 5500 Q-Trap, the source parameters for temperature (650 °C) and ion spray voltage (5000 V) were determined to be the biggest factors in obtaining optimal sensitivity and good peak resolution, parameters for both Q1 and Q3 were set to unit resolution (0.7 Da). The flow rates utilized here required a high source temperature, nebulizing gas (GS1 = 45), heating gas (GS2 = 45) and curtain gas (CUR = 25), to effectively desolvate the ions. For the 6460QQQ, the peptides were ionized by using an Agilent Jet Stream source operating in positive-ion mode with the following parameter settings: Sheath Gas flow = 11 l/min, Sheath Gas Temperature = 350 C, Nozzle Voltage = 1000 V, Nebulizing Pressure = 30 psi, Chamber Voltage = 4500 V. For SRM method development a general template method was utilized consisting of 60 transitions per method with 30-ms dwell time per transition and Q1 and Q3 resolution set to unit. For scheduled SRM methods, the same chromatography conditions were used with a retention time detection window of 0.2 minutes with a target cycle time of 1 - 1.5 seconds to yield fewer than 200 concurrent transitions. Peptide quantification was achieved by summing the integrated peak areas of two or three validated SRMs. Peptide abundances of the same protein were averaged to assign abundance to that protein. Protein copy number per cell was determined by using the OD-specific cell concentration of 10^9 cells/ml/ OD600 in the overnight cultures. The data were plotted by using Spotfire

3.2.1 (TIBCO Software, <http://sporfire.tibco.com>), a data analysis and visualization package. Box plots generated by Spotfire were used to display differences between the cultures grown under different conditions. Briefly, the upper and lower limits of the boxes correspond to the upper and lower quartiles, respectively. The solid line within each box represents the median value. The IQR is defined as the 3rd quartile value minus the first quartile value and the values greater than the 3rd quartile plus the inter quartile range (IQR) are considered outliers.

2.4 In silico SRM selection

Skyline software (MacLean et al., 2010) was used for SRM screening, peak selection, method development, analysis, and data processing purposes. Selection criteria excluded peptides with Met/Cys residues, tryptic peptides followed by additional cut sites (KK/RR), and peptides with proline adjacent to K/R cut sites. All possible doubly charged peptides were screened for y-series ions to establish the peptide identity and the most sensitive transitions. Methods generated by Skyline were set for AB 5500Q or Agilent 6460QQQ which included instrument-specific collision energies, declustering potential (for the 5500Q-Trap), and maximum mass-to-charge ratio were exported to an instrument method file using a template file which contained LC conditions as described above. Acquired SRM data were imported into Skyline where peptides were manually curated into a subset meeting the criteria described above, from which the top two peptides were selected for QconCAT generation. Pathway-specific Skyline methods, which contain proteotypic peptides and SRM information, are available via a Panorama

repository located at
<https://panoramaweb.org/labkey/project/JBEI/Escherichia%20coli%20pathways%20manuscript/begin.view?>. Protein quantification was based on the summed peak areas of the transitions for each peptide.

2.5 QconCAT design and expression

Proteotypic peptides from the targeted proteins, selected based on their performance in SRM validation experiments, were pooled into 18 different QconCAT sequences. Each QconCAT gene (with a N-terminal Glu¹-fibrinopeptide B coding sequence) was codon optimized, synthesized, and subcloned into a pET29b vector with a C-terminal His-tag by GenScript (Piscataway, NJ). Either *E. coli* strain BL21(λ)DE3 (*E. coli* B F^- *dcm ompT hsdS_B(r_B⁻ m_B⁻) gal, λ (DE3)) or BLR(DE3), a BL21 derivative, were transformed with 1 μ L of the pET29b-QconCAT vectors. Single colonies were picked from LB-Agar-Kanamycin plates and cultured overnight in 10 mL of LB medium. The cultures were then diluted to an OD₆₀₀ of 0.075 in 50 mL of LB medium. Once the cultures reached mid log phase ($A_{600} = 0.6 - 0.8$), expression was induced by adding isopropyl β -D-1-thiogalactopyranoside (IPTG) to a final concentration of 3 mM. Cells were harvested ($A_{600} = 2.5 - 3.0$) by centrifugation at 10,000g for 10 minutes. His-tagged QconCATs were purified under denaturing conditions using a protocol adapted from one previously described³⁵. The main difference between these protocols is the use of 8M Urea instead of 6M guanidinium chloride. Briefly, the cell pellet was resuspended in binding buffer and samples were sonicated until the mixture was translucent. Samples were then centrifuged at 10,000g for 20 minutes. The supernatant was incubated with Ni-NTA*

beads for 45 minutes, and the samples were applied to a polypropylene column (Qiagen, Venlo, Netherlands), fractions were collected and ran on an SDS page gel to confirm which fraction contained a band the same size as the expressed QconCAT. The purified QconCATs were purified using Centrifugal Filter Devices with MWCO 3,000 Da (Millipore, Billerica, Massachusetts) and the QconCAT concentrations were measured using the DC protein assay kit (Bio-Rad, Hercules, CA) with bovine serum albumin used as a standard. The final QconCATs were aliquoted and kept frozen at -80°C until further analysis.

3. Results and discussion

Optimization of complex networks, such as those found in cell factories, typically requires many iterative steps to reach the goal. To maximize the usefulness of targeted proteomics for this process we focused on pathway specific proteins to inform design of follow-up experiments. The work described here aims to address three significant barriers to broad adoption of proteomic characterization of *E. coli*: targeted proteomic methods, availability of standards, and sample throughput.

3.1 Building the SRM toolkit: SRM selection, validation, and method optimization

Proteins often don't work in isolation; they function as a part of a molecular network or a biological pathway to produce changes in cells. Pathway-level resources and information

derived from them can aid our understanding of these networks and how metabolic engineering impacts these processes. To aid quantitative analysis of *E. coli* metabolism more than 200 proteins involved in major pathways in *E. coli* (e.g., glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway, mixed acid fermentation, many amino acid biosynthesis pathways, and fatty acid biosynthesis) were selected for methods development. The largest two pathways in our set, lipid anabolism and purine and pyrimidine biosynthesis, consisted of 33 and 30 proteins, respectively. The smallest pathways, glutathione and acetyl-CoA biosynthesis, consisted of only 3 proteins each. In addition to the proteins involved in various pathways, we also characterized approximately 200 proteins in *E. coli* that are abundant under common growth conditions and are related to overall cell function. The later set included various chaperones, elongation factors and other enzymes known to be important for heterologous protein expression and can provide further opportunities for fine tuning of protein production or folding. As determined by previous shotgun proteomic studies (Ishihama et al., 2008; Lu et al., 2007), the abundance of proteins in these two sets ranged from less than a hundred to more than a million copies per cell.

Advances such as accurate quantification, reduced mass spectral acquisition time, and tools for assay development (Addona et al., 2009; Farrah et al., 2012; MacLean et al., 2010; Malmstrom et al., 2012; Picotti and Aebersold, 2012; Picotti et al., 2010; Reiter et al., 2011; Teleman et al., 2012) have led to increased use of targeted proteomics in biotechnology research. Recent developments have been successfully used to compile SRM transitions for 97% of the *Saccharomyces cerevisiae* proteome (Picotti et al., 2013)

and building compendia such as the PeptideAtlas (Brownridge et al., 2011; Costenoble et al., 2011; Kiyonami et al., 2011; Picotti et al., 2009; Picotti et al., 2008). However, care must be exercised to identify peptides and SRM transitions suitable for quantification. The accuracy of targeted proteomics is contingent on elimination of false-positive identifications, preclusion of chemical and biological modifications, and complete and specific digestion procedures. To address these issues a series of *in silico* and experimental validations were performed on the target peptides. Target peptides were selected based on size, sensitivity, the lack of modifications either post-translational or due to sample preparation, and favorable tryptic digestion characteristics (Brownridge and Beynon, 2011) (*i.e.*, no RP or KP sequences; Fig. 1a). Peptides containing proline were preferentially selected due to their propensity to produce intense y-ion fragments. Confirmation of peptide identity from the *in silico* SRM transitions, was accomplished by several metrics including observation of at least five co-eluting y-series transitions from the precursor ion (Fig. 1b). Peptides were also evaluated based on y-series ion intensities (*i.e.*, dotp scores) from an *E. coli* MS spectral library (<http://peptide.nist.gov>) and the measured elution time was compared to the Skyline-generated linear regression of peptides (Fig. 1b, bottom graph). Initial efforts to increase SRM sensitivity for selected peptides by collision energy optimization experiments (data not shown) did not yield significantly improved sensitivity from the Skyline-generated settings; consequently CE optimization was not performed.

Over 10,000 SRM transitions in more than 250 methods were screened to develop this refined SRM toolkit. Problems such as poor elution profiles, multiple elution peaks, or

poor alignment of transition retention times were common for low abundance proteins or those with few tryptic peptides. Subsequent retesting of transitions and changing sample growth conditions were used to improve detection to identify the best peptides from these proteins. In some cases, we were not able to detect peptides for a particular protein under our conditions, potentially due to the absence of the protein in the sample or expression at such low level that none of the peptides were above our level of detection. Efforts to synthesize peptides for every *E. coli* protein were cost prohibitive and beyond the scope of this work.

Once peptides were validated and selected (Supplementary Table 1), a comprehensive method utilizing scheduled SRMs was used to analyze samples in a short gradient LC-MS method (Fig. 2) via fast analytical-flow chromatographic methods (0.8 mL/min flow rate; 5.5-minute gradient) instead of nano-flow chromatography. To maximize the total number of proteins in a single method we limited the number of peptides per protein to two, however in certain cases (*e.g.*, small protein, low expression) only one peptide was available. For shotgun proteomics, two peptides are generally required for confident protein identification, however, for this SRM toolkit the limited number of peptides per protein is supported by peptide standards (see Section 3.2) to confirm the peptide identity and test for linear quantitative response. The resulting MS analysis consisted of >1200 transitions, 689 peptides from 409 proteins and the most intense transitions were selected along with consideration of the elution time to reduce the number of possible concurrent transitions. A retention time window of 0.2 minutes yielded fewer than 200 concurrent transitions for the scheduled SRM method with SRM dwell times from 2.1 – 219 msec.

For such a short chromatographic method the possibility of SRM transitions by background *E. coli* peptides overlapping the peptide of interest was a serious concern. Consequently, the candidate peptides and SRM transitions were screened by SRM-collider (Rost et al., 2012) for potential interferences and target transitions were adjusted accordingly. Additionally, selected strains from the KEIO single gene knockout collection were grown and tested for background signal in the SRM transitions corresponding to the peptides produced from the deleted gene (data not shown). The selected SRM transitions were adjusted when significant background signal was observed in the KEIO knockout sample for the given protein.

One major concern for a very short chromatography method is poor sensitivity due to shorter transition dwell times or ion suppression from co-eluting peptides. When using scheduled SRM methods the size of the retention time window plays an important role in the number of concurrent transitions and the resulting dwell times (for a given cycle time). Consequently, the retention time reproducibility of the LC-SRM method was examined by running three different concentrations of standard mixtures in triplicate on multiple days. The variability (CV%) of retention times for 99% the QconCAT peptides were calculated to be less than 0.03 minutes. This consistency permitted use of very narrow SRM transition scheduling windows (0.2 minutes) and aided quantification of the large number of peptides in a short LC-SRM method.

3.2 Absolute quantification by using QconCAT peptides

Absolute quantification of central pathway proteins benefits flux balance analysis, genome-scale modeling studies, metabolic engineering, and comparison of results from different laboratories. However, quantitative proteomics is significantly restricted by the lack of peptide standards to account for variations in instrument performance (*i.e.*, electrospray stability, ion suppression effects, injection variability) and ionization efficiency. We addressed this issue by generating reference peptides that were concatenated into polypeptide gene constructs known as QconCATs (Fig. 3a). For each protein, two or three of the best peptides were distributed into 18 different QconCAT sequences and, when possible, grouped by pathway (Supplementary Table 2). Current strategies to produce absolute quantification of proteins often focus on synthetic peptides, yet the QconCAT method offers several advantages. Importantly, public plasmid repositories enable distribution of constructs at a fraction of the cost of peptide synthesis while offering the researcher the choice of how much to produce and the option to generate heavy-isotope labeled peptides via production in appropriate media (Fig. 3b). The QconCATs were purified on Ni:NTA columns via the hexa-histidine tag appended at the C-terminus of each construct. Each QconCAT has the nucleotide sequence of Glufibrinopeptide B (GluFib) appended at the N-terminus to enable accurate quantification by comparison to a standard curve of commercially available GluFib (Fig. 3c). We interrogated the completeness of digestion of the QconCATs via shotgun analysis and SRM analysis of peptides resulting from single missed cleavages. Both experiments demonstrated undetectable or negligible signal from missed cleavages, indicating complete digestion of the QconCAT constructs (data not shown). For the 5.5-minute gradient LC-SRM method the lower limit of detection (LLOD; $S/N > 3:1$) was

determined to be 10 fmol and 20 fmol for 86% and 93% of the peptide standards, respectively. The linearity of the QconCAT peptide standards were assayed by measuring an external calibration curve for each available peptide. For individual peptides, their calibration curves showed good linearity across a wide range of concentrations from 50 fmol to 1 pmol (Supplementary Fig. 1). As a whole, these data indicate good sensitivity and reproducibility of the targeted proteomics method.

3.3 Quantitative proteomics of E. coli cultures grown on glucose and xylose

E. coli is metabolically adept at transitioning between aerobic and anaerobic lifestyles. Understanding differences in the *E. coli* proteome in response to oxygen availability can aid efforts to efficiently produce biofuels and high-value chemicals. Likewise, characterization of pathways and protein levels for *E. coli* cultured on glucose and xylose will benefit engineering efforts to optimally grow on the two major fermentable sugars present in cellulosic biomass, a promising source of cheap sugars (Blanch et al., 2008). We therefore applied the SRM toolkit to measure protein levels of biological triplicate samples of *E. coli* cultured with different carbon sources (glucose, xylose) in aerobic and anaerobic conditions. A total of 689 peptides were targeted in the method, of which 470 peptides were detected and quantified in this analysis and 368 were detected in all four biological conditions (Supplementary Table 3). Protein concentrations were determined by averaging the calculated concentrations of the peptides derived from each protein and standard error was used to estimate the accuracy of the quantification. For the majority of proteins (>56 %) the standard error of the concentration was less than 25 % and outliers

were highly correlated across samples. Importantly, there was no significant difference between the total abundance and average amount of protein for the four culture conditions, consistent with the observed growth state (Supplementary Table 4). While we did not observe a noticeable change in the total abundance of proteins in our analysis, some pathways changed significantly across conditions (Fig. 4a). The heat map shown in Figure 4a displays absolute abundance of the sum of the proteins involved in each pathway, thus reflecting both pathway length and abundance. As expected, the most abundant pathways were glycolysis and TCA cycle. Glycolysis was the most up-regulated pathway in this analysis, with Enolase (Eno, 52,568 copies/cell) and Glyceraldehyde-3-phosphate dehydrogenase A (GapA, 102,759 copies/cell) being the most abundant proteins among the pathway proteins. As seen in Figure 4b, proteins associated with the TCA cycle were also expressed in high abundance during aerobic culture conditions for both sugars, but not under anaerobic growth conditions. Both the breadth of pathway protein levels and the median are significantly higher for aerobic cultures than anaerobic, regardless of the sugar. Average concentrations of the most abundant TCA cycle proteins, isocitrate dehydrogenase (Icd) and malate dehydrogenase (Mdh), were 54,198 copies/cell and 44,321 copies/cell, respectively, in the presence of oxygen; but decreased to 14,332 copies/cell and 13,489 copies/cell, respectively, in the absence of it.

In the presence of oxygen, aerobic respiration is the preferred metabolic state in *E. coli*, since it allows complete oxidation of the carbon source to produce maximum energy. Consequently, we compared the protein levels between cultures grown aerobically and

anaerobically for both sugars (Fig. 5). Absolute levels of proteins associated with central carbon metabolism are grouped by pathways and shown via box plots in Figure 5 and relative levels of proteins for the two sugars (aerobic/anaerobic) are shown via heat maps. These data show that glycolysis was up-regulated under anaerobic growth for both sugars compared to aerobic culture conditions. Interestingly, the median protein level in glycolysis was similar for all four conditions, yet the distribution of protein levels was much greater in anaerobically-grown cultures (Fig. 5, Glycolysis box plot). Conversely, levels of proteins in the TCA cycle were much lower and were distributed over a narrow range in anaerobic cultures than in aerobic cultures. The reduced range of values is most easily seen in highly differentially expressed proteins such as SucD and AcnA. As expected, the amounts of proteins involved in the mixed acid fermentation and acetate formation pathways were much higher under anaerobic growth regardless of the carbon source. Protein levels for the pentose phosphate pathway diverged for glucose and xylose cultures. For this pathway, the median protein levels were higher for xylose cultures grown under anaerobic conditions than under aerobic conditions while the opposite was observed for glucose-grown cultures. Previous studies in yeast have shown that anaerobic growth in xylose results in up-regulation of pentose phosphate pathway, most likely to compensate for the reduced production of NADPH via the TCA cycle. Under anaerobic conditions, in the absence of external electron acceptors, pyruvate and its derivatives act as electron acceptors. Pyruvate is converted to ethanol and a mixture of organic acids, such as acetate, lactate, and formate. This is consistent with our observation of higher levels of alcohol dehydrogenase (AdhE), pyruvate formate lyase (PflB) and other mixed acid fermentation enzymes from anaerobic cultures regardless of the carbon source. For

the proteins not directly involved in the central metabolism, a transcription factor, Iron-sulfur cluster Regulator (IscR), and nucleoside diphosphate kinase (Ndk) were among the most differentially expressed. IscR was 6-7 fold more abundant for both sugars during aerobic growth than during anaerobic growth, consistent with the increased assembly of iron-sulfur clusters under aerobic conditions. Nucleoside diphosphate kinase (Ndk) transfers a phosphate group from ATP to other nucleoside diphosphates as part of the purine and pyrimidine biosynthesis pathway, was six-fold more abundant in glucose-containing cultures and three-fold more abundant in xylose-containing cultures under aerobic conditions than under anaerobic conditions. Previous studies that have shown that Ndk levels are highly reduced in anaerobic conditions and this reduction is coupled with increased levels of pyruvate kinase, which replaces Ndk. Indeed, both pyruvate kinase 1 (PykF) and pyruvate kinase 2 (PykA) were over-expressed under anaerobic conditions in our data. These results demonstrate the utility of a fast, accurate method of quantitative proteomics to aid understanding of *E. coli* metabolism.

4. Conclusions

As DNA construction becomes faster and cheaper the speed of pathway engineering will accelerate enabling construction of many modified strains shifting the majority of the effort of development to strain analysis. Likewise, the multiplex capabilities of new genome engineering tools are stimulating progress toward desired traits (Boyle and Gill, 2012), yet *predictive* transformation of engineered organisms into robust microbial factories remains contingent on precise understanding and control of dynamic biological

systems. Just as significant energy has been directed to establish methods capable of predictable protein production (Mutalik et al., 2013) comparable effort is needed to characterize the interactions of biological systems and the dynamic behaviors that come from them. With quantitative proteomic data, previously unforeseen effects of engineering can be deduced thus reducing the number of iterative design cycles necessary to reach this end.

In this work, we present resources to simplify and advance targeted proteomics of *E. coli* metabolic pathways through publically available repositories, both physical (QconCATs via Addgene) and operational (SRM parameters via Panorama). The method developed here relies on increased sample loading (10 – 50 ug) to achieve detection sensitivity that nano-flow chromatography methods achieve with 1-2 ug of sample. Fortunately, analysis of *E. coli* is seldom limited by the amount of protein available. Here, we showed how this workflow quickly identifies differences in protein levels for *E. coli* grown on different carbon sources and under different aeration conditions. Utilizing this workflow, protein levels for over 400 proteins from 96 samples can be analyzed in less than 24 hours. This level of throughput, coupled with absolute quantification, is especially important for analysis of genome engineering and dynamic processes that require high time resolution, such as changes to the rate of protein turnover. Future work will focus on expanding the repertoire of validated SRMs to proteins that were not detected in this study, implementing automated sample preparation protocols, and expanding the set of QconCAT-derived peptide standards for accurate quantification. Overall, this toolkit provides an invaluable resource for metabolic engineering of *E. coli* by increasing sample

throughput, minimizing methods development time, and offering a means for absolute quantification of proteins.

Acknowledgements

This work was part of the DOE Joint BioEnergy Institute (<http://www.jbei.org>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. Targeted proteomics methods in Skyline format and SRM information are available via Panorama located at: (<https://panoramaweb.org/labkey/project/JBEI/Escherichia%20coli%20pathways%20manuscript/begin.view?>). The QconCAT constructs used in this work are available through Addgene (www.addgene.org/).

Disclosure Statement

Jay D. Keasling has financial interests in Amyris Biotechnologies, LS9 Inc., and Lygos Inc.

REFERENCES

- Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J., Mani, D. R., Bunk, D. M., Spiegelman, C. H., et al., 2009. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* 27, 633-41.
- Alonso-Gutierrez, J., Chan, R., Batth, T. S., Adams, P. D., Keasling, J. D., Petzold, C. J., Lee, T. S., 2013. Metabolic engineering of *Escherichia coli* for limonene and perillyl alcohol production. *Metab. Eng.* 19, 33-41.
- Beynon, R. J., Doherty, M. K., Pratt, J. M., Gaskell, S. J., 2005. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods.* 2, 587-9.
- Blanch, H. W., Adams, P. D., Andrews-Cramer, K. M., Frommer, W. B., Simmons, B. A., Keasling, J. D., 2008. Addressing the need for alternative transportation fuels: the Joint BioEnergy Institute. *ACS Chem. Biol.* 3, 17-20.
- Boyle, N. R., Gill, R. T., 2012. Tools for genome-wide strain design and construction. *Curr. Opin. Biotechnol.* 23, 666-71.
- Brownridge, P., Beynon, R. J., 2011. The importance of the digest: proteolysis and absolute quantification in proteomics. *Methods.* 54, 351-60.
- Brownridge, P., Holman, S. W., Gaskell, S. J., Grant, C. M., Harman, V. M., Hubbard, S. J., Lanthaler, K., et al., 2011. Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy. *Proteomics.* 11, 2957-70.
- Costenoble, R., Picotti, P., Reiter, L., Stallmach, R., Heinemann, M., Sauer, U., Aebersold, R., 2011. Comprehensive quantitative analysis of central carbon and amino-acid metabolism in *Saccharomyces cerevisiae* under multiple conditions by targeted proteomics. *Mol. Syst. Biol.* 7, 464.
- Farrah, T., Deutsch, E. W., Kreisberg, R., Sun, Z., Campbell, D. S., Mendoza, L., Kusebauch, U., et al., 2012. PASSEL: The PeptideAtlas SRMexperiment library. *Proteomics.* 12, 1170-5.
- Herscovitch, M., Perkins, E., Baltus, A., Fan, M., 2012. Addgene provides an open forum for plasmid sharing. *Nat. Biotechnol.* 30, 316-7.
- Hillson, N. J., Rosengarten, R. D., Keasling, J. D., 2011. j5 DNA Assembly Design Automation Software. *ACS Synth. Biol.* 1, 14-21.
- Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M. J., Frishman, D., 2008. Protein abundance profiling of the *Escherichia coli* cytosol. *BMC genomics.* 9, 102.
- Juminaga, D., Baidoo, E. E., Redding-Johanson, A. M., Batth, T. S., Burd, H., Mukhopadhyay, A., Petzold, C. J., et al., 2012. Modular engineering of L-tyrosine production in *Escherichia coli*. *Appl. Environ. Microbiol.* 78, 89-98.
- Kiyonami, R., Schoen, A., Prakash, A., Peterman, S., Zabrouskov, V., Picotti, P., Aebersold, R., et al., 2011. Increased selectivity, analytical precision, and throughput in targeted proteomics. *Mol. Cell. Proteomics.* 10, M110 002931.
- Lanza, A. M., Crook, N. C., Alper, H. S., 2012. Innovation at the intersection of synthetic and systems biology. *Curr. Opin. Biotechnol.*

- Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25, 117-24.
- MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., et al., 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics.* 26, 966-8.
- Malmstrom, L., Malmstrom, J., Selevsek, N., Rosenberger, G., Aebersold, R., 2012. Automated workflow for large-scale selected reaction monitoring experiments. *J. Proteome Res.* 11, 1644-53.
- Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A., Tran, A. B., et al., 2013. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods.* 10, 354-60.
- Neidhardt, F. C., Bloch, P. L., Smith, D. F., 1974. Culture medium for enterobacteria. *J. Bacteriol.* 119, 736-47.
- Oud, B., van Maris, A. J., Daran, J. M., Pronk, J. T., 2012. Genome-wide analytical approaches for reverse metabolic engineering of industrially relevant phenotypes in yeast. *FEMS Yeast Res.* 12, 183-96.
- Percy, A. J., Chambers, A. G., Yang, J., Domanski, D., Borchers, C. H., 2012. Comparison of standard- and nano-flow liquid chromatography platforms for MRM-based quantitation of putative plasma biomarker proteins. *Anal. Bioanal. Chem.* 404, 1089-101.
- Picotti, P., Aebersold, R., 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods.* 9, 555-66.
- Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., Aebersold, R., 2009. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell.* 138, 795-806.
- Picotti, P., Clement-Ziza, M., Lam, H., Campbell, D. S., Schmidt, A., Deutsch, E. W., Rost, H., et al., 2013. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature.* 494, 266-270.
- Picotti, P., Lam, H., Campbell, D., Deutsch, E. W., Mirzaei, H., Ranish, J., Domon, B., et al., 2008. A database of mass spectrometric assays for the yeast proteome. *Nat. Methods.* 5, 913-4.
- Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., et al., 2010. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods.* 7, 43-6.
- Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J., Gaskell, S. J., Beynon, R. J., 2006. Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* 1, 1029-43.
- Redding-Johanson, A. M., Batth, T. S., Chan, R., Krupa, R., Szmidt, H. L., Adams, P. D., Keasling, J. D., et al., 2011. Targeted proteomics for metabolic pathway optimization: application to terpene production. *Metab. Eng.* 13, 194-203.
- Reiter, L., Rinner, O., Picotti, P., Huttenhain, R., Beck, M., Brusniak, M. Y., Hengartner, M. O., et al., 2011. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods.* 8, 430-5.
- Rost, H. L., Malmstrom, L., Aebersold, R., 2012. A computational tool to detect and avoid redundancy in selected reaction monitoring. *Mol. Cell. Proteomics.*

- Schuster, S. A., Boyes, B. E., Wagner, B. M., Kirkland, J. J., 2012. Fast high performance liquid chromatography separations for proteomic applications using Fused-Core(R) silica particles. *J Chromatogr A*. 1228, 232-41.
- Singh, P., Batth, T. S., Juminaga, D., Dahl, R. H., Keasling, J. D., Adams, P. D., Petzold, C. J., 2012. Application of targeted proteomics to metabolically engineered *Escherichia coli*. *Proteomics*. 12, 1289-99.
- Smolke, Christina D., Silver, Pamela A., 2011. Informing Biological Design by Integration of Systems and Synthetic Biology. *Cell*. 144, 855-859.
- Teleman, J., Karlsson, C., Waldemarson, S., Hansson, K., James, P., Malmstrom, J., Levander, F., 2012. Automated Selected Reaction Monitoring Software for Accurate Label-Free Protein Quantification. *J. Proteome Res.*
- Van Dien, S., 2013. From the first drop to the first truckload: commercialization of microbial processes for renewable chemicals. *Curr. Opin. Biotechnol.* 24, 1061-8.
- Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., Church, G. M., 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature*. 460, 894-8.
- Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B., Gill, R. T., 2010. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat. Biotechnol.* 28, 856-62.
- Wessel, D., Flugge, U. I., 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* 138, 141-3.

Table and Figure Legends:

Table 1: Optimized SRM parameters for proteins in the acetate formation pathway. A full summary of the SRM toolkit information can be found in Supplementary Table 1.

Figure 1: An overview schematic of SRM toolkit method development. (a) All proteins for a selected pathway are digested *in silico* with Skyline to filter for proteotypic peptides and preferred MS parameters.(*i.e.*, y-series fragments) (b) Peptide retention time selection is based on overlapping of multiple y-series ions, DOTP scores (if available) (top), and retention time based on hydrophobicity calculator (bottom). Additional RT validation was subsequently accomplished with QconCAT peptides.

Figure 2: Extracted ion chromatograms for the validated peptides from the optimized 5.5 minute gradient (10 minute total run time).

Figure 3: A schematic for QconCAT development. (a) Two validated peptides from each protein are selected and a His-tag for affinity enrichment for QconCAT construction. (b) The *E. coli* host is transformed with the plasmid containing the QconCAT sequence and is used to express the QconCAT in cell culture. Following expression cells are lysed and protein purified on Ni-NTA beads. (c) The purified QconCAT is digested with trypsin, quantified by comparison to a Glu-fib standard and used to generate calibration curves for absolute quantification. A summary of the QconCAT constructs and measured limit of detection can be found in Supplementary Table 2.

Figure 4: (a) Heat map visualization of the sum of protein abundances grouped by pathway for glucose and xylose cultures grown aerobically and anaerobically. (b) Box-whisker plot of TCA cycle protein abundances from four separate cultures. The bottom and top of the box represent the first and third quartiles and the line inside the box is the median. The whiskers represent 1.5 times the inter quartile range. Individual protein levels for replicate analyses are overlaid on the box-whisker plot.

Figure 5: A map representing the pathways and proteins involved in *E. coli* central metabolism. Heat map values correspond to the Log10 of protein levels for aerobic/anaerobic culture conditions (Glucose, Xylose). x = not detected. Box-whisker plot of *E. coli* pathway protein abundances from four separate cultures (Carbon source, Condition) colored by pathway and organized from left to right: Glucose, aerobic; Glucose, anaerobic; Xylose, aerobic; Xylose, anaerobic. The bottom and top of the box represent the first and third quartiles and the band inside the box is the median. The whiskers represent 1.5 times the inter quartile range. cpc = copies per cell.

Table 1

Protein			SRM Transitions		Collision Energy	
Uniprot ID	Name	Peptide Sequence	Q1 m/z	Q3 m/z	AB Sciex 5500	Agilent 6460
P27550	ACSA	LVITSDEGVR	544.8	876.44(y8), 763.4(y7), 662.3(y6)	28.5	12.2
		IAEAAVVGIPHNK	716.4	877.5(y8), 778.5(y7), 608.4(y5)	34.6	21
P77218	EUTD	TPPDALEK	435.7	672.4(y6), 460.3(y4), 389.2(y3)	24.5	6.7
		QTEVNR	373.7	618.3(y5), 517.3(y4), 388.2(y3)	22.3	3.5
P0A9M8	PTA	TGGDAPDQTTIVR	716.4	1030.6(y9), 818.5(y7), 690.4(y6)	34.6	21
		LNAPVDEQGR	549.8	871.4(y8), 800.4(y7), 604.3(y5)	28.6	12.5
P0A6A3	ACKA	YTSSVVIDESVIQGIK	869.5	1101.6(y10), 988.5(y9), 873.5(y8)	40.2	28.8
		LGVLGFEVDHER	685.9	1101.5(y9), 988.4(y8), 784.4(y6)	33.5	19.4

a

SELECT PROTEIN

GENERATE PROTEOTYPIC PEPTIDES *IN-SILICO*

- 350 - 1250 m/z
- No Met/Cys
- No modifications
- No miss trypsin cleavage sites
- No consecutive cleavage sites

b

Peptide validation

Intensity

Retention time

Peptide hydrophobicity comparison (measured vs predicted)

Measured time

Sequence specific retention calculator

Figure 2

Figure 3

Figure 4

Figure 5

Figure 1

Figure 2
[Click here to download high resolution image](#)

Figure 3

Figure 4

Figure 5