Next-Gen^3: Sequencing, Modeling and Advanced Biofuels

Period Covered September 2014 - August 2017

November 2017

Karsten Zengler
University of California, San Diego
La Jolla, California

PREPARED FOR THE U.S. DEPARTMENT OF ENERGY, BIOLOGICAL AND ENVIRONMENTAL RESEARCH (BER)
UNDER CONTRACT NO. DE-SC0012586

TABLE OF CONTENTS

INTRODUCTION

Successful, scalable implementation of biofuels is dependent on the efficient and near complete utilization of diverse biomass sources. One approach is to utilize the large recalcitrant biomass fraction (or any organic waste stream) through the thermochemical conversion of organic compounds to syngas, a mixture of carbon monoxide (CO), carbon dioxide (CO_2), and hydrogen (H_2), which can subsequently be metabolized by acetogenic microorganisms to produce next-gen biofuels. The goal of this proposal was to advance the development of the acetogen *Clostridium ljungdahlii* as a chassis organism for next-gen biofuel production from cheap, renewable sources and to detail the interconnectivity of metabolism, energy conservation, and regulation of acetogens using next-gen sequencing and next-gen modeling. To achieve this goal we specifically used:

Omics-driven elucidation of the multidimensional genome architecture. A higher level of genome annotation captures the elements responsible for the flow of information encoded by the genome. This genome architecture is comprised of the collection of components including (but not limited to) promoters, transcription start sites, regulatory non-coding regions, untranslated regions, transcription units, and ribosome binding sites. Integration of data from cutting edge experimental methods developed in our lab, such as ChIP-exo and ribosome profiling allowed for characterization of regulation and gene expression patterns and, subsequently, the extraction of novel information such as transcriptional pause sites, translational pause sites, and multi-protein complex stoichiometry. In particular we investigated the role of translation in optimization of carbon and energy utilization through differential translational efficiency.

The control of mRNA translation is vital to all species. We employed RNA-seq, TSS-seq, and Ribo-seq to decipher condition-dependent translational regulation in the model acetogen *Clostridium ljungdahlii*. Integration of multi-omics data obtained from cells grown autotrophically or heterotrophically revealed that pathways critical to carbon and energy metabolism are under strong translational regulation. We showed that major subsystems involved in energy and carbon metabolism are not only differentially transcribed and translated, but their translational efficiencies are differentially elevated in response to resource availability under different growth conditions. Translational efficiency is controlled on the molecular level by a combination of features associated with the coding and the 5′-untranslated regions of mRNA, suggesting that *C. ljungdahlii* prioritizes translation of genes essential for thriving in energy-deprived niches.

Reconstruction and validation of a predictive next-gen model including metabolism and macromolecular synthesis (ME-model). The Wood-Ljungdahl pathway (WLP) in *Clostridium ljungdahlii* enables the use of either H_2 or CO as electron donors with

accompanied reduction of CO_2 thereby making WLP the only known CO_2-fixing pathway coupled to energy conservation. The feasibility of autotrophic growth was poorly understood for a long time as no ATP was gained at the substrate level. Knowledge of how a bacterium completely lacking cytochrome-encoding genes could maintain the proton motive force was lacking. It was then discovered that the RNF complex couples ferredoxin oxidation, NAD^+ reduction and proton exportation by a novel mechanism called “electron bi furcation”. To explore how growth strategies occur, models like constraint-based genome scale models of metabolism (i.e., M-models) have been useful for gaining insight to possible energy flux routes. While M-models have enabled much progress in elucidating cofactor fluxes, critical components of the cell, such as the production of macromolecules and the mechanistic utilization of metals, vitamins, and cofactors, are usually absent in these models thereby limiting in-depth understanding of cellular life.

So-called metabolic and gene expression models (ME-models) include not only metabolic reactions, but they also include explicit representations of major cellular processes such as macromolecular synthesis and basic transcriptional regulation, which significantly broadens the scope and predictability of microbial systems biology. Specifically, the ME-model will: 1) Account for the transcriptional and translational cost of proteins and complex formation; 2) Incorporate the energetics associated with cofactor dependencies and prosthetic group usage; 3) Quantitatively predict transcript and protein levels; 4) Predict optimal codon usage for heterologous pathways. With these ME-models, the optimal molecular constitution of cells can be computed as a function of genetic and environmental parameters. Since both RNA and protein abundances are explicitly predicted, cofactor requirements can now be explored.

We completed the *C. ljungdahlii* ME-model, named iJL965-ME, that captures all major central metabolic, amino acid, nucleotide, lipid, major cofactors, and vitamin synthesis pathways as well as pathways to synthesis RNA and protein molecules necessary to catalyze these reactions. Furthermore, the reconstruction includes the WLP, with updated cofactors, and its associated mechanisms for energy conservation. iJL965-ME was used to reveal how protein allocation and media composition influence metabolic pathways and energy conservation in *C. ljungdahlii*, and to accurately predict secretion of acetate, ethanol, and glycerol during changing carbon.

Trace metals are essential for all living organisms, for they are required for catalytic processes essential to energy conservation, metabolism, replication, and maintenance. Yet metals pose a unique challenge in constraint-based models of metabolism (i.e. M-models) as they are neither produced nor consumed biochemically; instead, metals in M-models are generally treated as a lumped sum in the biomass objective function rather than be

integrated into the network. In M-models, metal availability and growth rate are linearly correlated even though there is contrary experimental evidence. In iHN637, the M-model for *C. ljungdahliii* reconstructed by our group, seven of ten metals (Ca^{2+}, Cu^{2+}, Mg^{2+}, Mn^{2+}, Mo^{2+}, Ni^{2+}, Zn^{2+}, Co^{2+}, Fe^{2+}, Na^{+}) could only be imported or exported (in addition to their inclusion in the biomass objective function, which represents the total composition of the cell, and only Co^{2+} was predicted to participate in flux-carrying reactions that were not a transport reaction or biomass production. Thus, most metal ions were not associated to the reactions they help catalyze. This represents a general fact for M-models.

The next generation of constraint-based genome-scale models change this paradigm. Metabolic and gene expression models (ME-models) cover the processes of transcription, translation, and metabolism, which can also include protein modifications. Protein modifications can account for the presence of metals in biochemical reactions and thus enable predictions of the optimal distribution of resources in response to limited metal availability. Therefore, ME-models provide a robust, genome-wide approach to define how transition metals affect an organism's functional network, which addresses the articulated need to bridge chemistry and biology in a coherent and systematic way. The detailed representation of cofactors and prosthetic groups will enable us to manipulate the cofactor dependency of heterologous pathways to maximize energy conservation, subsequently optimizing chemical production by *C. ljungdahlii*.

Our study substantially enhanced our knowledge about chemolithoautotrophs and their potential for advanced biofuel production. It provides next-gen modeling capability, offers innovative tools for genome-scale engineering, and provides novel methods to utilize next-gen models for the design of tunable systems. The following report contains information about work performed under contract DE-SC0012586.

The report consists of three parts, addressing varies aspects of the work:

A Optimization of carbon and energy utilization through differential translational efficiency

B Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen

C Exploring the evolutionary significance of tRNA operon structure using metabolic and gene expression models

A Optimization of carbon and energy utilization through differential translational efficiency

The metabolic versatility of acetogens for the fermentation of a large number of sugars yields great promise for the production of biofuels and commodity chemicals. In particular the ability to grow autotrophically with H_2:CO_2 or syngas (H_2/CO/CO_2) makes these organisms ideal chassis for sustainable bioproduction and acetogenic clostridia are currently deployed for the commercial conversion of syngas to biofuels. *Clostridium ljungdahlii* is emerging as a promising cell factory for bioproduction (Kopke et al., 2010) as well as a model organism for gaining in-depth knowledge necessary to develop new design strategies for acetogens. *C. ljungdahlii* is readily cultured heterotrophically in the laboratory in simple media, either on a diverse set of five or six carbon sugars, or autotrophically with CO or H_2 as electron donor. Furthermore, metabolic models and genetic manipulation tools already developed and optimized for this organism, make *C. ljungahlii* an ideal candidate for the study of acetogenesis (Nagarajan et al., 2013).

However, in order to harness the full biosynthetic potential, it is important to understand the regulatory mechanisms that orchestrate energy metabolism in *C. ljungdahlii*. These include, but are not limited to, the Wood-Ljungdahl pathway (WLP), the formate dehydrogenase complex, the hydrogenase complex, and the Rnf complex, which are all central to energy equilibrium in *C. ljungdahlii* (Ljungdahl, 2009; 2et al., 2013, Latif et al., 2013). A thorough understanding of all factors that regulate energy metabolism under autotrophic and heterotrophic growth conditions is crucial for the metabolic engineering of acetogens and for optimizing targeted production of desired chemicals.

In recent years, next-generation omics approaches, such as RNA-seq, Ribo-seq, proteomics, and metabolomics have been employed to identify the functionality and organizational structure of the bacterial genome. These approaches directly address the genotype-phenotype relationship in bacteria, providing crucial insights into the design strategies for microbial cell factories. In particular, Ribo-seq in combination with RNA-seq has enabled global measurements of translation and provided new insights into translational regulation (Ingolia et al., 2009). Here we carried out cognitive analysis of RNA-seq and Ribo-seq to understand the translational control underlying energy and metabolism in the model acetogen *C. ljungdahlii*. Furthermore, we integrated transcription start site (TSS) information with RNA-seq to gain insight into the structure of the 5′-untranslated region (5′ UTR) and used Ribo-seq to understand its effect on the translational efficiency (TE). We provide evidence that metabolic pathways required for utilization of carbon and energy are not only regulated at the transcriptional and translational level, but they h1ave evolved to enhance TE

of specific nodes in the network to maintain optimized energy homeostasis in a growth-condition-dependent manner. We show that the AU content of the 5′ UTR, the AU content of the coding region and to a lesser extent, codon adaptation control TE and are crucial factors for acetogens to thrive in energy-deprived environments.

Multi-omics analyses of heterotrophically and autotrophically grown cultures. We carried out RNA-seq and Ribo-seq experiments for autotrophic cultures grown either on CO or H_2:CO_2 and heterotrophic cultures grown on fructose. To enable direct comparison between transcription and translation, strand-specific RNA-seq libraries were prepared from the same lysates used for Ribo-seq experiments in biological duplicates. RNA-seq and Ribo-seq libraries were deeply sequenced and mapped reads were normalized as FPKM and RPKM, respectively. RNA-seq replicates for cultures grown on CO, H_2:CO_2, or fructose were highly reproducible with a Pearson correlation of 0.995, 0.991, and 0.989, respectively. Ribo-seq replicates were also highly correlated with Pearson correlations of 0.995, 0.952, and 0.930, respectively.

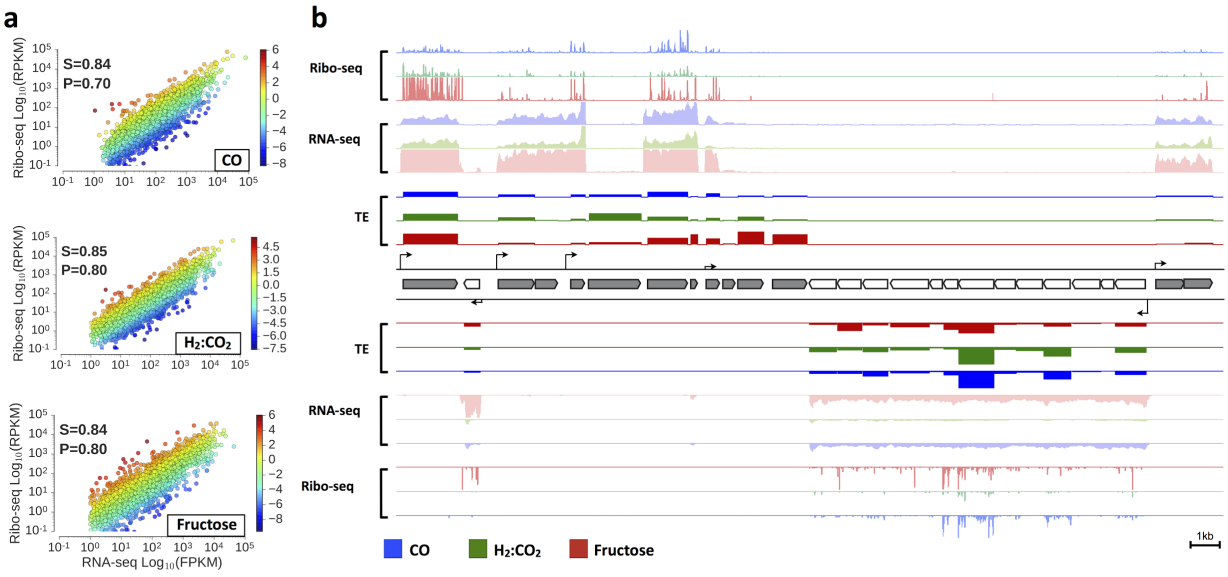

Fig. 1 Overview of omics experiments carried out for each growth condition and the correlation between RNA-seq and Ribo-seq in all growth conditions. (a) Correlations between RNA-seq and Ribo-seq in CO, H_2:CO_2 and fructose. Pearson's and Spearman's coefficients are shown inside each subfigure. Colors represent in the scatter plot represents the translational efficiency values, as depicted in the colorbars. **(b)** An example of Ribo-seq, RNA-seq, TE, and TSS profiles mapped onto genomic region between 4,535,800 to 4,564,000. RNA-seq and Ribo-seq profiles were normalized in RPM. TE of each gene is calculated by Ribo-seq level divided by RNA-seq level. Arrows indicate TSS positions.

While the majority of genes are regulated at the transcriptional level, transcription and translation in bacteria are spatially coupled and many genes are subjected to firm translational control (McCarthy and Gualerzi, 1990; Ingolia et al., 2009; Ingolia, 2014). In line with previous findings in *Escherichia coli* and *Streptomyces coelicolor* (Ingolia et al., 2009, Jeong et al., 2016), RNA-seq and Ribo-seq data from *Clostridium ljungdahlii* were

moderately correlated in all conditions tested (**Fig. 1a**), suggesting widespread translational regulation. We calculated the TE of each gene by dividing the translational level by the transcriptional level and noticed significant discrepancy in TE among different genes (**Fig. 1b**).

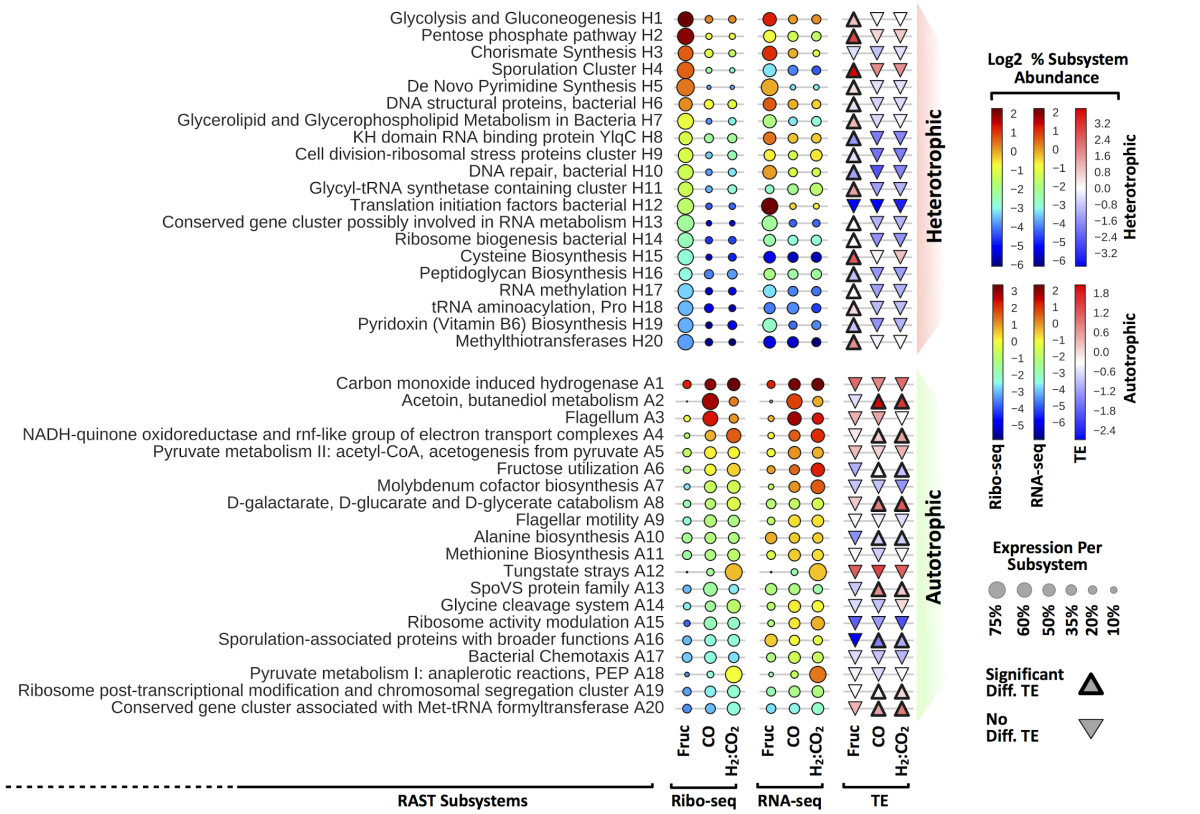

Fig. 2 Differential translation and differential TE of subsystems in fructose-, CO- and H_2:CO_2-grown cultures. Genes were grouped into subsystems and translation and transcription were both percent-normalized per each experiment. Color of bubbles represents the level of translation or transcription per each experiment (% normalized). The size of bubbles represents the level of translation or transcription per each subsystem (% normalized). The top panel represents the top 20 subsystems that are translationally induced in fructose relative to both autotrophic (CO and H_2:CO_2) conditions (P<0.01) and sorted in descending order according to translation levels in fructose. Subsystems with differential TE are depicted by up-pointing triangle with thick edges in fructose (Fruc). The bottom panel represents the top 20 subsystems that are differentially translationally induced in both autotrophic conditions relative to heterotrophic growth in fructose (P<0.01) and are sorted in descending order according to translation in CO. Subsystems with differential TE are depicted by up-pointing triangles with thick edges in CO and H_2:CO_2.

Translational efficiency is differentially controlled in a condition-dependent manner

The variability of TE was observed both within a given condition as well as across different conditions (**Figs. 1a, b**), indicating a functional link between TE and phenotype. Classification of genes into discrete functional units and the measurement of transcription or translation of these units provides insight into the control of gene expression in its functional context. Therefore, we functionally annotated the *C. ljungdahlii* genome using RAST (Aziz et al., 2008), resulting in the classification of 1731 genes into 270 subsystems. Differentially translated subsystems under all growth conditions were determined by DESeq2 (Love et.

2014). To enable direct comparisons between the three conditions, RNA-seq and Ribo-seq data per subsystem for growth on CO, H_2:CO_2 were compared as percent values. The top 20 differentially translated subsystems (P<0.01) in heterotrophic (**Fig. 2** top) and autotrophic conditions (**Fig. 2** bottom) are shown.

The top differentially translated subsystems in heterotrophic and autotrophic conditions were associated with carbon and energy sources present in the corresponding growth media. In heterotrophic growth 16 out of the 20 most differentially translated subsystems were those related to carbon metabolism (**Fig. 2:** H1, H2 and H7) and *de novo* macromolecule synthesis and maintenance (**Fig. 2:** H3, H5, H6, H9-H16, H18 and H19). The remaining clusters (**Fig. 2:** H4, H8, H17 and H20) had no obvious link to heterotrophic metabolism or fast growth. Glycolysis and the pentose phosphate pathways (H1 and H2) were highly enriched followed by the chorismate synthesis subsystem (H3), which is the precursor molecule for *de novo* synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan. The sporulation cluster (H4) was unexpectedly highly enriched. After close inspection, we found that out of four genes in this subsystem, *Clju_c41620* (encoding a putative RNA-binding S1 domain-containing protein) was the only differentially translated gene. This protein weakly interacts with the ribosome and facilitates the recognition of the translation start site (further discussed below).

Under autotrophic growth, subsystems were differentially translated according to the energy and carbon sources provided. 13 out of 20 subsystems were closely connected to carbon fixation and energy conservation (**Fig. 2:** A1 A4, A6-A8, A12, A14 and A18), fermentation (**Fig. 2:** A2 and A5) and motility (**Fig. 2:** A3, A9 and A17). The top four subsystems (A1-A4) consisted of the CODH/AscA cluster, 2,3-butanediol dehydrogenase (BDD), flagellum, and the Rnf complex. The CODH/AscA complex is directly involved in carbon fixation and energy conservation through the Wood-Ljungdahl pathway (WLP). Under CO growth, BDD translation represented 7% of the total translation and the flagellum, flagellar motility, and bacterial chemotaxis (all related to motility and chemotaxis) represented 4% of total translation. Generally, differentially TE subsystems were less frequent under autotrophic growth compared to heterotrophic growth. A2 and A4 had the most differentially TE subsystems. A2 represents the 2,3-butanediol/acetoin fermentation pathway, whereas A4 represents the Rnf complex cluster (**Fig. 2**), which consists of the *rnfCDGEAB* genes and the Rnf transcriptional regulator, *rseC*. The Rnf complex has been shown to be essential for autotrophic growth, but redundant under heterotrophic growth (Tremblay et al., 2012).

Differentially translated subsystem specific to autotrophic or heterotrophic growth

Highly responsive subsystems in H_2:CO_2 encompassed the Rnf complex, flavodoxin, and the aldehyde:ferredoxin oxidoreductase (**Fig. 2:** A12). The Rnf complex, discussed below in

more detail, is under strong translational control. On the other hand, subsystems induced specifically under CO growth were related to acetoin, butanediol metabolism, the flagellum, and one-carbon metabolism (i.e. WLP). Cells growing in CO were conspicuously the most highly motile when examined under the microscope, which supports the measured differential translation. Overall, differentially translated subsystems related to pathways involved in H_2:CO_2 and CO, hint at a regulatory mechanism that specifically accounts for physiological requirements when growing under autotrophic conditions.

To gain insight into how TE is differentially controlled under autotrophic and heterotrophic conditions, we analyzed genes of major carbon and energy subsystems that were significantly enriched (**Fig. 2**). These systems consisted of glycolysis/gluconeogenesis, the WLP, fermentation pathways, the Rnf complex, and the ATPase complex (**Fig. 3**). Genes with redundant functions which are not differentially translated were not included in the analysis. As expected, the majority of genes in glycolysis and gluconeogenesis were differentially enriched during heterotrophic growth (**Fig. 3**, blue arrows), whereby fructose is taken up preferentially via the fructokinase/fructose-6-phosphate isomerase (G1) and the 6-phosphofructokinase (G3) route. Under autotrophic growth, the fructose phosphotransferase system (PTS) and 1-phosphofructokinase (G2) were also significantly enriched. Two enzymes involved in pyruvate metabolism were differentially translated (P4 and B1 in **Fig. 3**). The incomplete TCA cycle exhibited differential translation, whereas genes involved in fermentation were only differentially translated under autotrophic growth. Most notably are E1 (bifunctional aldehyde/alcohol dehydrogenase) and B3 (2,3-butanediol dehydrogenase), both differentially translated with high efficiency in autotrophic conditions (**Fig. 3,** A2 in **Fig. 2**). The WLP is mostly differentially translated under autotrophic growth with W5 (methenyl-THF cyclohydrolase) and W7 (methylene-THF reductase) being the least efficient (**Fig. 3**). All genes encoding the F1F0 ATPase are differentially transcribed and differentially translated under heterotrophic growth condition. The remarkable low-TE of the ATPase cluster implies that its translation is relatively more resilient to transcriptional fluctuations. The Rnf genes (*rnfCDGEAB*) are differentially transcribed, differentially translated, and most genes, including the Rnf regulator *rseC*, exhibit differential TE under autotrophic growth conditions (further discussed below).

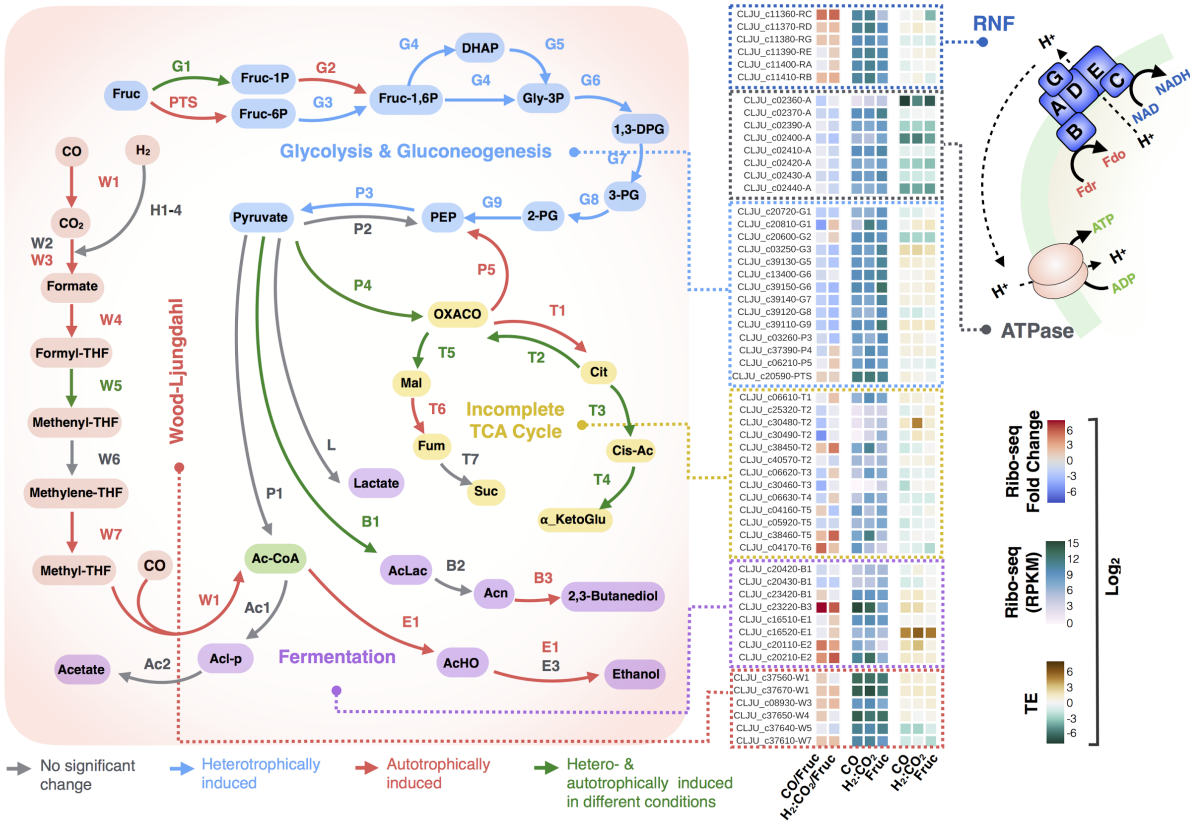

Fig. 3 Metabolic map of major carbon and energy pathways exhibiting differential translation and differential TE. Differential fold change is calculated as the log2 CO/fructose or H_2:CO_2/fructose translation ratio. Heterotrophically induced (red arrows), autotrophically induced (blue arrows), insignificant (grey arrows) and condition-specific (green arrows) translation is depicted in all pathways. **Glycolysis & Gluconeogenesis:** fructose phosphotransferase system (**PTS**); fructokinase /fructose-6-phosphate isomerase (**G1**); 1-phosphofructokinase (**G2**); 6-phosphofructokinase (**G3**); ketose-bisphosphate aldolase (**G4**); triose-phosphate isomerase (**G5**); glyceraldehyde-3-phosphate dehydrogenase (**G6**); phosphoglycerate kinase (**G7**); phosphoglycerate mutase (**G8**); enolase phosphopyruvate hydratase (**G9**); pyruvate:ferredoxin oxidoreductase (**P1**); pyruvate, phosphate dikinase (**P2**); pyruvate kinase (**P3**); pyruvate carboxylase (**P4**); PEP carboxykinase (**P5**). **Fermentation:** phosphotransacetylase (**Ac1**), acetate kinase (**Ac2**), bifunctional aldehyde/alcohol dehydrogenase (**E1**), aldehyde:ferredoxin oxidoreductase (**E2**), additional alcohol dehydrogenases (**E3**), acetolactate synthase (**B1**), acetolactate decarboxylase (**B2**), 2,3-butanediol dehydrogenase (**B3**); lactate dehydrogenase (L). **Incomplete TCA cycle**: citrate synthase (**T1**); citrate lyase (**T2**); aconitase (**T3**); isocitrate dehydrogenase (**T4**); malate dehydrogenase (**T5**); fumarase (**T6**); fumarate reductase (**T7**). **Wood-Ljungdahl pathway:** electron-bifurcating [FeFe] hydrogenase (**H1**); Other [FeFe] hydrogenases (**H2**); [NiFe] hydrogenase (**H3**); hydrogenase maturation factor (**H4**); bifunctional CO dehydrogenase/ acetyl-CoA synthase (CODH/ACS) (**W1**); seleno formate dehydrogenase (**W2**); non-seleno formate dehydrogenase (**W3**); Formyl-THF ligase (**W4**); methenyl-THF cyclohydrolase (**W5**); methylene-THF dehydrogenase (**W6**); methylene-THF reductase (**W7**). **Rnf complex & ATPase:** RnfC (**RC**); RnfD (**RD**); RnfG (**RG**); RnfE (**RE**); RnfA (**RA**); RnfB (**RB**); ATPase (**A**). Fructose (**Fruc**); fructose 1-phosphate/6-phosphate (**Fruc-1P/-6P**); fructose 1,6-bisphosphate (**Fru-1,6P**); dihydroxyacetone phosphate (**DHAP**); glycerol 3-phosphate (**Gly-3P**); 1,3-bisphosphoglycerate (**1,3-DPG**); 3-phosphoglycerate (**3-PG**); 2-phosphoglycerate (**2-PG**); phosphoenolpyruvate (**PEP**); oxaloacetate (**OXACO**); citrate (**Cit**); isocitrate (**Cit-Ac**); a-ketoglutarate (**a-KetoGlu**); malate(**Mal**); fumarate (**Fum**); succinate (**Suc**); acetolactate (**AcLac**); acetoin (**Acn**); acetaldehyde (**AcHO**); acetyl-phosphate (**Acl-p**); tetrahydrofolate (**THF**); reduced ferredoxin (**Fdr**); oxidized ferredoxin (**Fdo**).

Rnf subunits are under strict translational control in heterotrophic growth

The *rnfC* gene is transcribed at a significantly lower level during heterotrophic growth (FPKM= 621.7 under fructose growth compared to 2560.7 and 3080.8 under CO and H2:CO2 growth, respectively; **Figs. 3**, **4a**). Notably, *rnfC* is acutely translationally repressed under heterotrophic condition (TE = 0.1 for fructose compared to 0.9 in CO and 1.3 in H_2:CO_2; **Fig. 4b**), thus contributing only ~1% of the *rnf* gross translation. Under heterotrophic

growth, the Rnf regulator *rseC* is transcribed at a high level in each growth condition (FPKM= 2098.1, 2991.8, 1022.9, for CO, H_2:CO_2 and fructose growth conditions). However, *rseC* translation is highly repressed under heterotrophic growth at the translational level comparable to that of *rnfC* (TE= 20, 2.2, 0.3 for CO, H_2:CO_2 and fructose growth conditions, respectively; **Fig. 4b**). Thus, the Rnf complex is highly translationally repressed especially for *rnfC* and *rseC*.

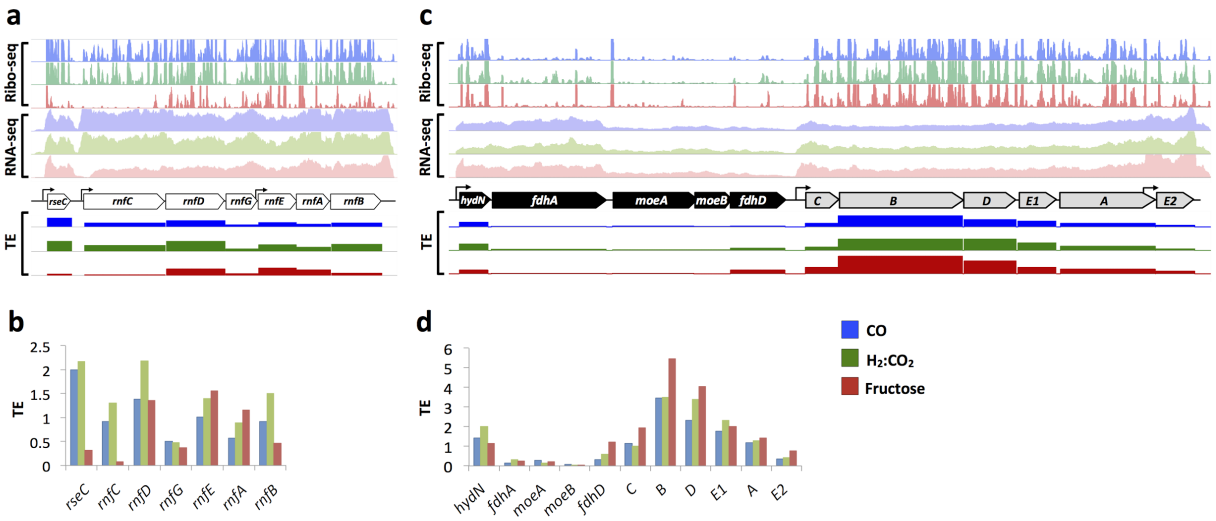

Fig. 4 Transcriptional and translational regulation of the Rnf (white), formate dehydrogenase (black), and hydrogenase (grey) complexes in all growth conditions. Results are shown for CO, H_2:CO_2 and fructose in blue, green and red, respectively. **(a)** The *rnf* complex (*Clju_c11350-Clju_c11410*) has one major TSS upstream of *rnfC*. In addition *rnfEAB* are transcribed from an internal promoter that is positioned at the 3′-end of *rnfG*. *rseC* is transcribed from one TSS and transcription is comparable across all conditions, however it is poorly translated in fructose. **(b)** *rnfC* has the lowest translation and lowest TE in heterotrophic growth. **(c)** Expression of the formate dehydrogenase and the hydrogenase genes (*Clju_c06990-Clju_c07080*). Both clusters are expressed from upstream TSSs. *hydN* and *fdhA* are translated at a much lower efficiency compared to the hydrogenase B and D genes despite having higher transcription. **(d)** The hydrogenase genes have higher TE compared to the formate dehydrogenase genes. The hydrogenase *E2* gene is transcribed at a significantly higher level from an internal promoter, however its TE is the lowest in the hydrogenase gene cluster. The translational regulation of the two clusters is independent of the growth condition.

The formate dehydrogenase operon is inefficiently translated compared to the downstream hydrogenase complex in all growth conditions

The only active hydrogenase (Hyd) in *C. ljungdahlii* is the one orthologous to HytABCDE1E2 in *C. autoethanogenum,* which is the only hydrogenase active under H_2:CO_2 growth (Mock et al., 2015). In *C. ljungdahlii*, Hyd catalyzes the reduction of NADP and ferredoxin and the oxidation of H_2 under H_2:CO_2 growth. Additionally, Hyd interacts with formate dehydrogenase (Fdh) and the resulting complex (Hyd-Fdh) catalyzes the reduction of CO_2 to formate and the oxidation of H_2 (Mock et al., 2015). Under CO growth, the bifurcating carbon monoxide dehydrogenase (CODH) catalyzes the oxidation of CO to CO_2 and the reduction of ferredoxin. The Hyd-Fdh complex then catalyzes the oxidation of ferredoxin and the reduction of CO_2 to formate. Under heterotrophic growth, the pyruvate ferredoxin oxidoreductase catalyzes the oxidation of pyruvate to acetyl-CoA, the reduction of ferredoxin

and the generation of CO_2 as byproduct (Latif et al., 2014). CODH catalyzes the oxidation of ferredoxin and the reduction of CO_2 into CO, whereas Hyd-Fdh catalyzes the reduction of CO_2 into formate using reduced ferredoxin. Fdh and Hyd are both multimeric complexes, both active under all growth conditions tested, and both are essential for the WLP, which plausibly underscores the observed stable TE of both complexes in all conditions (**Fig. 4c**, **d**). Our omics analysis illustrates that at least *hydN* and *fdhA* are transcribed from one upstream TSSs and their transcriptional levels are greater than *hydCBDAE1*. The latter genes are also transcribed from one detectable TSS, whereas *hydE2* is transcribed from an internal TSS positioned at the 3′ end of *hydA* (**Fig. 4c**). Despite higher transcriptional levels of *hydN* and *fdhA*, *hydBDE1* are translated at a much higher level (higher TE). In fact, *hydB* is at least three-fold more translationally efficient than *hydN* and *fdhA*. These results suggest translational regulation is seminal for the regulation of key energy conservation centres in this model acetogen.

TE is governed by a combination of features linked to the 5′ untranslated region as well as the coding region

As illustrated above, a wide spectrum of genes exhibited condition-dependent variability in their TE, suggesting plausible regulation at the translational level. Under the growth conditions tested, the vast majority of genes exhibited stable TE despite differences in the levels of transcription across growth conditions (88.8%, 89.1%, 88.4% for CO, H_2:CO_2 and fructose respectively). Interestingly, we find strong variability in TE of operonic genes, despite having comparable transcriptional level, suggesting that intrinsic mRNA features fine-tune the rate of translation (**Fig. 1b**). To explore these features and determine their influence on translation and TE, we compared RNA-seq and Ribo-seq data of genes with low-TE (<20th percentile) or low-TE (>80th percentile). The difference between the two sets was strikingly more significant at the translational level (Wilcoxon signed-rank test P= $1.7e^{-92}$, $1.2e^{-78}$, $8.2e^{-81}$ for CO, H_2:CO_2, and fructose, respectively) when compared to the transcriptional level (Wilcoxon signed-rank test P= $1.2e^{-9}$, 0.13, $4.4e^{-6}$ for CO, H_2:CO_2, and fructose, respectively), implying pronounced translational regulation (**Fig. 5a**).

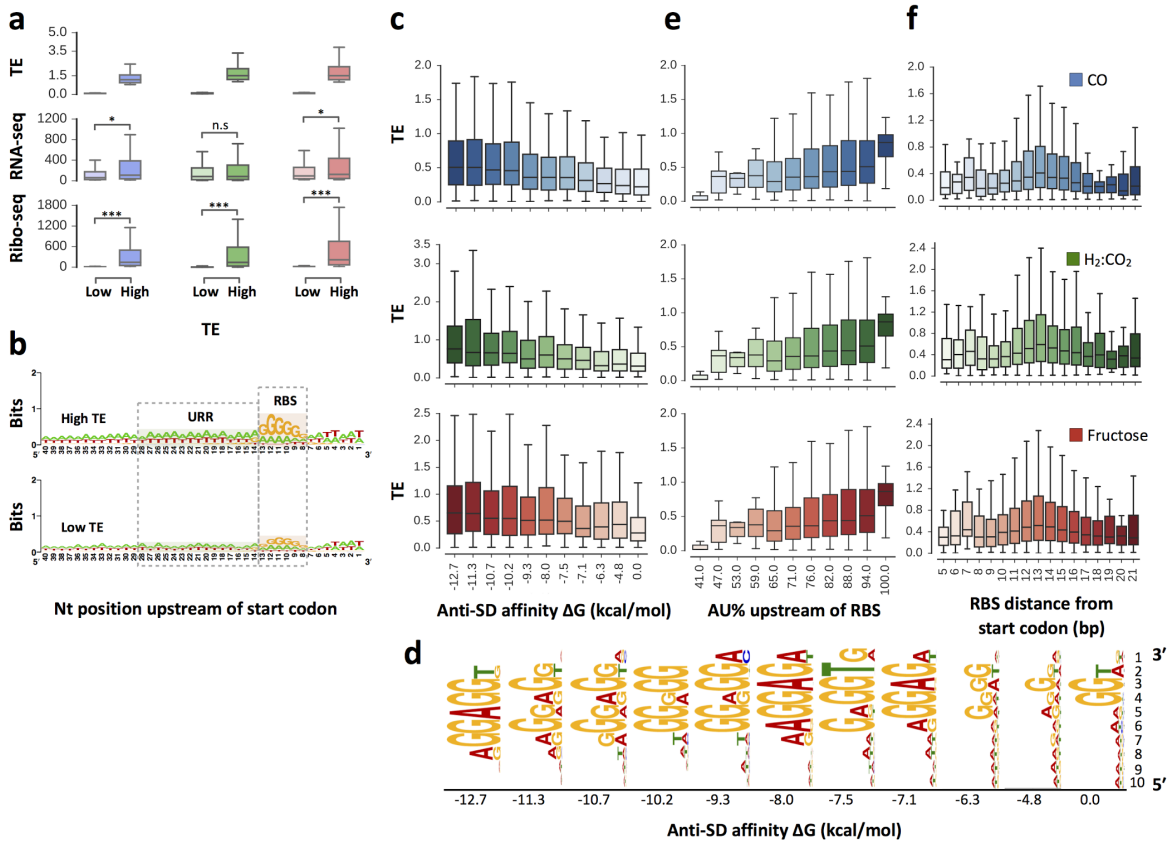

Fig. 5 Influence of UTR features on translational efficiency. (a) Comparison between genes with low and high translational efficiency (TE) in all conditions. Low-TE genes are below 20th percentile, whereas high-TE genes are above 80th percentile in all conditions. P<1e^{-10} are signified with "***", 1e^{-10}≤P<0.01 are signified with "*" and P>0.01 are signified with "n.s" **(b)** Low- and high-TE genes in all conditions have visible differences in their RBS sequence and the AU content in their upper RBS region (URR). RBS and URR are highlighted by boxes at their corresponding regions. **(c)** RBS affinity towards the anti-Shine Dalgarno (anti-SD) sequence (AAGGAGGU) positively affects the translational efficiency in all conditions. The affinities of RBS towards the anti-SD region were grouped into eleven categories ranging from ΔG of 0 to -12.7. **(d)** The RBS motif per each category in CO was determined using MEME. **(e)** Positive effect of the 15 bp AU% content in the URR on TE in all conditions. TSS data were used to ensure that TSS is upstream of the URR. **(f)** The distance of the RBS 5′ end from the start codon is most optimum at 13 bp. Deviation of the RBS position in either direction negatively influences TE.

Previous studies have reported a direct regulation of TE via the 5′-untranslated regions (Migone et al., 2002; Gebauer and Hentze, 2004; Wade and Grainger, 2014). Here we investigated the effect of different features in the 5′ UTR on TE. To accurately determine the 5′ UTR regions, we first performed a comprehensive transcription start site (TSS) analysis using four different growth conditions. We comprehensively determined a total of 1,465 TSSs that correspond to the 5′-end of the primary transcriptome. The TSSs were further categorized by their genomic locations. 1,245 TSSs were annotated as primary TSSs, which cover 29% of total gene content excluding operons and 50% of total gene content including operons. In addition, we detected 116 internal TSSs and 25 antisense TSSs that could manifest potential control of gene expression (Wade and Grainger, 2014). 125 orphan TSSs were also identified at intergenic regions with no associated genes, suggesting the presence of novel transcriptional units. Alignment of 50 bp upstream of TSS revealed conservation of

two motifs at -10 and to a lesser extent at -35 consistent with sigma factor binding motifs, implying high-accuracy detection of TSS. It is worth to note that we could not detect any leaderless genes under the growth conditions tested, which further emphasize the importance of translational regulation via the 5′ UTR in *C. ljungdahlii*.

To investigate *cis*-acting regulatory elements of translational control, we defined the 5′ UTR from the region between primary TSSs and start codon of corresponding genes. The most frequent size range of 5' UTR distribution was 20-39 nt. The median 5′ UTR length was 47 nt, implying that for the vast majority of genes, *cis*-acting elements, and secondary structures play a critical role in translational regulation. The ribosome-binding site (RBS) is one of the critical elements for translational initiation[19] (Li and Weissman, 2012), which in turn directly impacts TE. We compared the composition of the -10 and -35 regions of the 5′ UTR by analyzing 40 nt upstream of the TSS using WebLogo (Crooks et al., 1994). There were two clear differences between low-TE and high-TE genes, namely the high-TE genes had a stronger RBS motif and the upper RBS region (URR) had an increased AU content (**Figs.4, 5b**). Based on these differences, we investigated how TE is influenced by RBS affinity towards the anti-Shine Dalgarno (aSD) sequence (AAGGAGGU), the RBS distance from the TSS, and the AU% content of the URR. We analyzed the affinity of the aSD sequence towards RBS (see Methods) for both low- and high-TE genes. The difference was highly significant between the two groups under all three conditions, suggesting that RBS affinity towards the initiating ribosomes is a key determinant for TE. Further, we organized all genes into eleven categories according to their ∆G affinity and compared their TE (**Fig. 5c**). The gradual decrease in mean TE with increasing ∆G implies that TE is strongly influenced by the RBS affinity towards the aSD (**Fig. 5c**). Furthermore, MEME analysis (http://meme-suite.org/tools/meme) showed that RBS motif conservation increased with TE and those with lowest TE had a hardly recognizable RBS motif, whereas groups with low-TE exhibit an optimal RBS motif (**Fig. 5d**).

RPS1 (*Clju_c41620*)*,* a protein weakly associated with the 30S ribosomal subunit, has strong affinity towards AU-rich regions at the 5′ UTR (Komarova et al., 2005; Nakagawa et al., 2010) and interacts with the 5′ UTR of mRNA through a 10-15 nt motif to facilitates the initiation of translatio (Subramanian, 1983). In addition, RPS1 furnishes the 30S subunit with an RNA chaperone activity that is essential for the binding and unfolding of structured mRNAs, allowing the correct positioning of the initiation codon for translation (Duval et al., 2013). Further, RPS1 competes with RNases for the binding of AU-rich regions, plausibly protecting AU-rich URR from degradation, which leads to increased TE (Hajnsdorf & Boni 2012; Komarova et al., 2005). We reasoned that AU-rich URRs could result in greater TE. To validate this hypothesis, we calculated the AU% in regions 15 nt (15 nt showed strongest difference between low- and high-TE sets in **Fig. 5b**) upstream of each RBS. To eliminate

false positives arising from the high AT content of the *C. ljungdahlii* genome (31.1 % GC), we limited our analysis to promoters that had their TSS at least 15 nt upstream of the URR. Genes associated with transcripts harboring URRs with 100% AU had the highest TE (**Fig. 5e**). TE of low- versus high-AU% groups were statistically significant in all growth conditions, suggesting that the AU content at the URR significantly impacts TE.

We further compared the position of the RBS relative to the translation start site and showed that genes with highest TE were those harboring RBSs 13 nt upstream of the translation start site (**Fig. 5f**). In addition, we found that the most conserved RBS motifs tend to be at optimum distance from the translation start site. Finally, we analyzed the effect of codon usage on TE, using the codon adaptation index (CAI; Sharp & Li 1987). We found that the average AU% has strikingly more influence on TE when comparing low- and high-TE sets than CAI (Mann-Whitney test, P_{CAI}=1.9e-07, $P_{AU\%}$=8.3e-60).

Features that promote low-TE are enriched in differentially transcribed genes involved in condition-specific carbon and energy metabolic pathways

Our results hint towards prioritization of subsystems involved in carbon and energy metabolism by differentially increasing their TE in a condition-dependent manner. Accordingly, we reasoned that genes classified in these subsystems could plausibly be prioritized for higher translation rates through optimization of UTR and coding region features that facilitate higher initiation rates and/or features that promote mRNA stability. In contrast, genes under other subsystems involved in general cell maintenance activities would carry less optimal features. This could be beneficial since less translationally efficient systems are higher translational stability (**Fig. 5a**). Accordingly, if features in the UTRs and in the coding regions have significant influence on TE in a mRNA-level-dependent manner, we expect to find subsystems related to carbon and energy metabolism enriched in genes that are differentially transcribed and show a low-TE, but not enriched in differentially transcribed and low-TE genes.

We created two groups comprising differentially transcribed genes (>1.5 fold change) with high or low-TE, without *a priori* knowledge of the subsystems they are classified under. Group 1 represents our “test” group and consist of low-TE genes (>80th percentile) for both autotrophic growth (blue and green dots in **Fig. 6a, b**) and heterotrophic growth (red dots in **Fig. 6a, b**). group 2 represents our “control” group and consists of low-TE genes (<20th percentile) in both autotrophic (blue and green squares in **Fig. 6a, b**) and heterotrophic conditions (red squares in **Fig. 6a, b**).

We calculated RAST-enrichment as the ratio of the number of genes found in group 1 or group 2 relative to the total number of genes in each subsystem, excluding subsystems that contain only one gene. We sorted the subsystems according their highest ratio and aligned the top 10 subsystems in autotrophic group 1, heterotrophic group1, autotrophic group 2 and

heterotrophic group 2 to RAST categories. We found a clear enrichment of subsystems under the energy (respiration) category in autotrophic group 1 (**Fig. 6c**). Likewise, we found clear enrichment of subsystems under the carbon metabolism category in heterotrophic group 1 (**Fig. 6d**). In contrast, no enrichment of subsystems under carbon or energy categories in autotrophic or heterotrophic group 2 were detected. Instead, we identified subsystems involved in general cell maintenance.

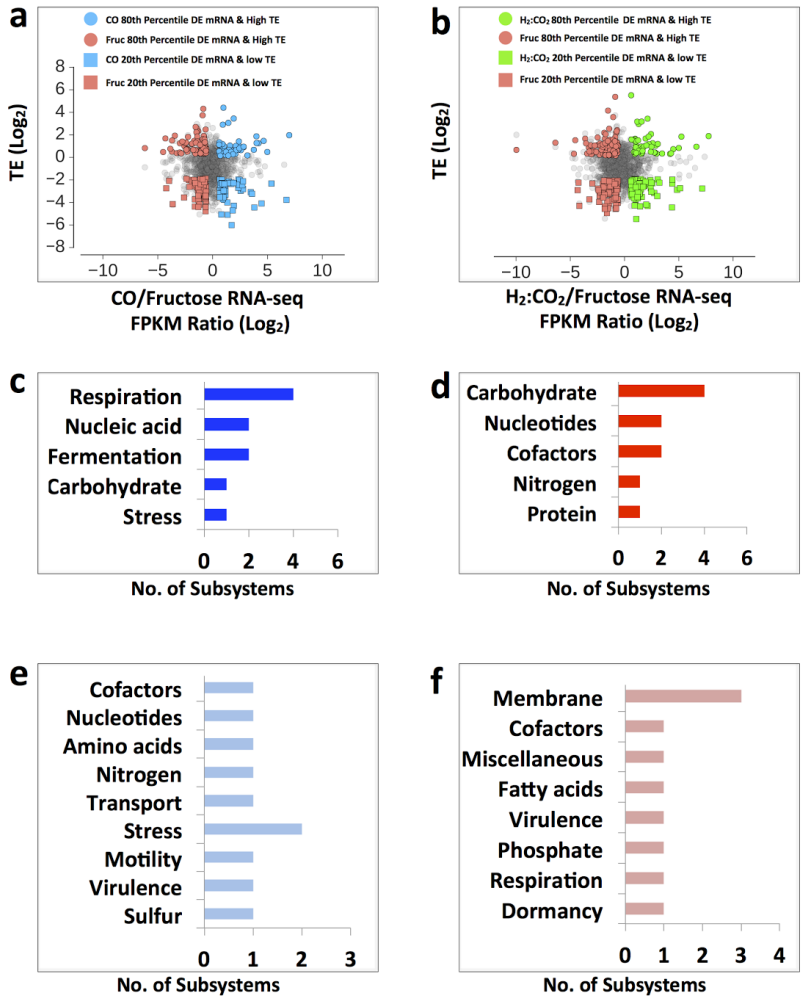

Fig. 6 Analysis of translational efficiency, functional allocation, and genomic features of differentially transcribed mRNA. (**a**) X-axis represents Log2 ratio of differentially induced genes in cells grown with CO or fructose (Fruc). TE above 80th percentile is represented in blue and red dots for CO and fructose, respectively. TE below 20th percentile is represented in blue and red squares for CO and fructose, respectively. (**b**) Similar to (a), with H_2:CO_2 instead of CO. (**c**) RAST subsystem count of low-TE genes that are autotrophically differentially expressed; (**d**) low-TE genes that are heterotrophically DE; (**e**) RAST subsystem count of low-TE genes that are autotrophically differentially expressed; (**f**) RAST subsystem count of low-TE genes that are heterotrophically differentially expressed. Autotrophic and heterotrophic low-TE and low-TE genes are taken from those shown in (a) and (b).

Overall, our results show that carbon and energy metabolism subsystems have evolved to be profoundly translationally efficient, such that they are highly responsive to changes in mRNA levels in a growth-condition-dependent manner. On the other hand, less responsive subsystems involved in general cell maintenance that are growth-condition-independent

have lower TE such that their translation is less affected by fluctuations in transcription, that might result from changes in growth conditions.

We see a clear influence of UTR features, CAI, and coding region AU% on TE. The Mann-Whitney test was used to calculate P values (**Table 1**). The AU% of the coding region had a more significant P value than CAI, suggesting that AU% in the coding region has a stronger impact on TE than CAI. The AU% at the URR was also more significantly different between low- and high-TE groups in autotrophic growth but lower in heterotrophic growth. Thus, the AU% of the URR could be more important in regulating TE in differentially translational efficient genes than RBS strength under autotrophic growth, whereas those genes that are differentially translational efficient under heterotrophic growth could be more influenced by the RBS strength.

Table 1. Comparison of 5′ UTR and coding region features for groups discussed in Fig. 6.

Feature	Autotrophic *P* values*†	Heterotrophic *P* values*‡
URR AU%	5.81E-05	2.23E-03
RBS ΔG	2.60E-03	6.37E-05
RBS distance	8.51E-02	1.54E-02
Codon adaptation index	6.87E-05	3.10E-04
Coding region AU%	6.60E-10	1.20E-07

* Mann-Whitney U test between low- and high-TE mRNA
† N=159
‡ N=196

Here, we carried out a multi-omics approach to study the translational control underlying important carbon and energy metabolism in the model acetogen *C. ljungdahlii*. RNA-seq and Ribo-seq data were combined from identical samples to ensure high robustness. Datasets were highly correlated in comparison to previous studies, in which RNA and ribosome footprints were obtained from different samples. In all growth conditions, we found that a sizable number of genes had TEs markedly above or below the average, implying strong translational regulation. By using RAST functional enrichment at the subsystem level, we showed that carbon and energy pathways were highly regulated at the translational level under autotrophic growth, whereas under heterotrophic growth translational regulation was highest for carbon metabolism subsystems and subsystems involved in fast growth including *de novo* synthesis of amino and nucleic acids.

We provided examples of strong translational control in energy conservation pathways. For example, the Rnf-, the formate dehydrogenase-, and the hydrogenase complexes were all highly regulated at the translational level. The Rnf complex was shown to be translationally repressed in heterotrophic growth, where it has been shown to be dispensable. However, the formate dehydrogenase and the hydrogenase complexes showed no apparent difference in their TE in all growth conditions; the hydrogenase complex on average has higher TE than the formate dehydrogenase in all growth conditions. The ATPase genes were translationally

inefficient regardless of the growth condition, implying strong translational stability that is independent of autotrophic or heterotrophic growth conditions.

We defined multiple features in the 5′ UTR and in the coding region that showed a clear effect on TE. By comparing enrichment of these features in highly translationally efficient and in highly translationally inefficient subsystems, we showed that AU content at the URR as well as at the coding region are very important determinants of TE. In addition, RBS affinity to aSD and the distance of the RBS from the translation start site were also critical determinants.

By analyzing high-TE (group 1) and low-TE (group 2) differentially transcribed mRNA, we demonstrate that genes related to carbon and energy metabolism are enriched in group 1 plausibly because they are required to be translated readily and efficiently to quickly adapt to changes in the relevant growth conditions tested. In contrast to group 1, we demonstrate that translationally inefficient genes in group 2 are involved in housekeeping activities, such as membrane transport and protein synthesis (ribosomal proteins) and tend to have constant, but low-TE. Thereby, we argue that genes in group 1 are very sensitive to changes in mRNA levels and their TE positively correlates with mRNA levels. We further demonstrate that genes important in all growth conditions, including housekeeping genes, have lower TE, which render them less sensitive to fluctuations in mRNA levels. Furthermore, we show that metabolic and energy subsystems specific for growth in autotrophic or heterotrophic conditions are mostly enriched in group 1. Whereas group 2 contained mostly housekeeping subsystems.

Our study uncovers a novel regulatory mechanism for a bacterium that thrives at the energetic limit of life and highlights utilization of scarce resources at optimal efficiency. We propose that pathways involved in carbon and energy metabolism are specifically controlled through optimizing the TE level, allowing for dynamic resource allocation. Our findings have broad implications on how microorganisms control and optimize their metabolic networks. The results provide a new framework for metabolic regulation in this model acetogen, that can readily be extrapolated to other industrially important microbes. Unraveling of regulatory mechanisms lays the foundation for advanced strain design and engineering efforts.

B Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen

Reconstructing an acetogen ME-model. To create an acetogen metabolic and gene expression model (ME-model), an existing genome-scale M-model of *C. ljungdahlii* (iHN637) was first updated (Nagarajan et al., 2013). By using recent literature and genome annotations as reference (Mock et al., 2015; Tan et al., 2015; Seemann, 2104; Kopke et al., 2010; Becker et al., 2005), twentyeight reactions were added and four reactions removed from iHN637. The updated M-model (iJL680) consisted of 43 additional genes and contained updated cofactor stoichiometry and directionality of redox reactions based on experimental

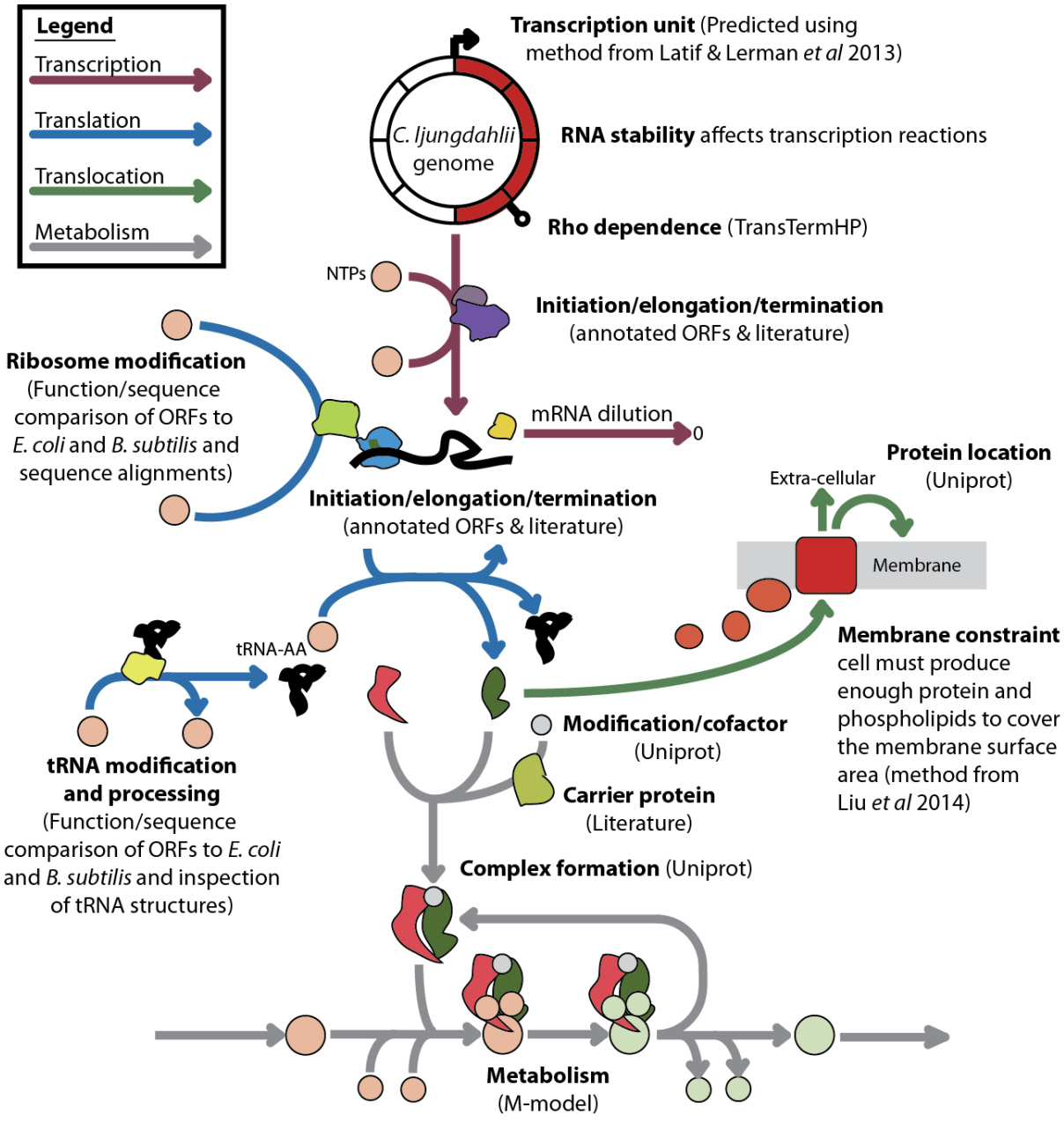

Fig. 7: Representation of the ME-model. The E-matrix reconstruction accounted for transcription, translation, and translocation as well as associated reactions to produce functional enzymes. Integration of the E-matrix (colored arrows) with the M-model (grey arrows) resulted in the ME-model.

data and exhibits comparable predictability.

Following established methods, an acetogen gene expression network (*i.e.*, E-matrix) was reconstructed from *C. ljungdahlii* (Lerman et al., 2012; Thiele and Palsson, 2010; Thiele et al., 2009; Lloyd et al., 2017). This reconstruction included an additional 196 protein-coding open reading frames (ORFs), 89 RNA genes, 576 transcription units (415 of which were rho-dependent and 29 were RNA-stable), 19 types of rRNA modifications, 17 types of tRNA modifications, 735 protein complexes with updated stoichiometry, 219 modified protein complexes, and 134 translocated proteins. The turnover rate for metabolic enzymes (approximated by k_{eff}, a required parameter for ME-models) was set to the average turnover rate of all enzymes found in acetogens in the enzyme database Brenda, 25 s^{-1} (Placzek et al., 2017). Coupling constraints, which link macromolecular synthesis costs with reactions, were calculated using the formulation in COBRAme (O'Brien et al., 2014; Lloyd et al., 2017; Placzek et al., 2017).

Using the COBRAme framework, the acetogen E-matrix was integrated with iJL680 to create the ME-model, iJL965-ME. iJL965-ME accounts for all of the major central metabolic pathways and biomass synthesis pathways as well as transcription, translation, macromolecule modifications, and translocation reactions (**Fig. 7**). Because iJL965-ME covers an extensive scope of cellular processes, we can predict fermentation profiles, including overflow metabolism products, gene expression, and usage of co-factors and metals, which are described in detail below.

Accuracy of predicted growth and yield phenotypes improve with iJL965-ME. Unlike the M-model, iJL965-ME predicted both batch (*i.e.*, maximum nutrient uptake) and nutrient-limited growth conditions for *C. ljungdahlii*. Due to internal constraints on protein production and catalysis, referred to as proteomic limitations (O'Brien et al., 2014), iJL965-ME growth rate was a non-linear function of the substrate uptake rate. Thus, optimal carbon uptake rate and maximum growth rate could be simultaneously predicted, whereas M-models require information of one rate to predict the other (O'Brien et al., 2014). As a result, we identified unique growth rate and yield functions for growth with CO, CO_2+H_2, or fructose (**Fig. 8**).

Overflow metabolism is the seemingly wasteful process in which a substrate is not fully oxidized, resulting in lower energy yields, inefficient metabolism, and fermentation products. Hypotheses for why this phenomenon occurs are varied, which makes characterizing and modelling mixed fermentation production so challenging. Generally, M-models do not predict alternative fermentation products without additional constraints on redox fluxes, oxygen uptake, or the objective function Nagarajan et al., 2013; Valgepea et al., 2017a; Valgepea et

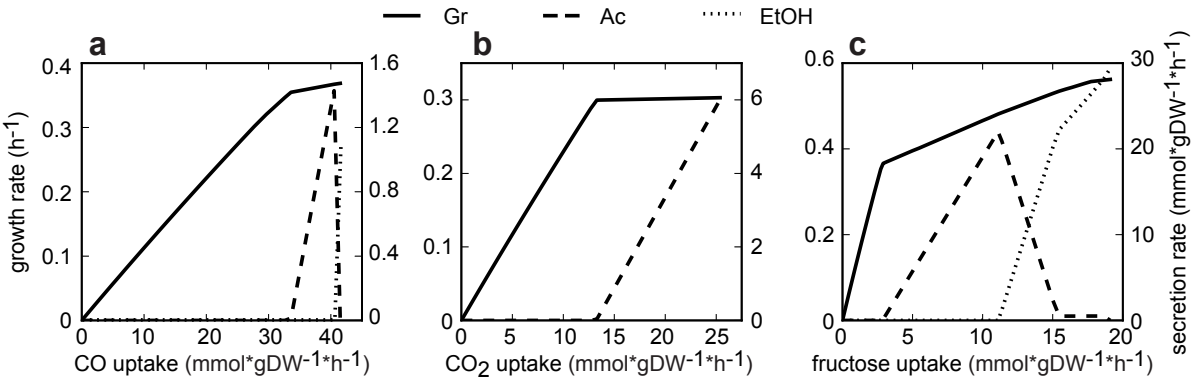

Fig. 8: Predicted growth rate and yield. Maximum growth rate, acetate secretion rate, and ethanol secretion rate changed as a function of **a**, CO, **b**, CO_2, and **c**, fructose uptake rate.

al., 2017b; Dash et al., 2014). However, iJL965-ME was able to predict intrinsically changes in the primary fermentation product as a function of substrate availability for CO and fructose growth. When protein production approached proteome limitations (exemplified by *in silico* maximum growth rate and *in vivo* mid-log phase), iJL965-ME correctly predicted the start of ethanol secretion after acetate secretion due to trade-offs in protein production (**Fig. 8A, C**). Thus, iJL965-ME was able to recapitulate overflow metabolism by accounting for redox balancing and concurrent proteome limitations.

The ME-model also predicted substrate-specific growth rates with high accuracy. Specifically, growth rate predictions from iJL965-ME were more accurate than the M-model, iJL680 (Pearson's r: 0.68 > 0.29; Spearman ρ: 0.60 > 0.091; **Fig. 9A**). Due to distinct resource requirements (the main factor being proteome composition) when metabolizing different substrates, unique *in silico* maximum growth rates for individual substrates can be obtained through iJL965-ME. Unlike the M-model (iJL680), which predicted that glucose and fructose would have identical growth rates, iJL965-ME correctly predicted slower growth on glucose than for fructose. Furthermore, iJL965-ME highly improved predictions of the ratio of maximum acetate secretion rate to substrate uptake rate compared to the M-models iHN637 and iJL680 (**Fig. 9B**).

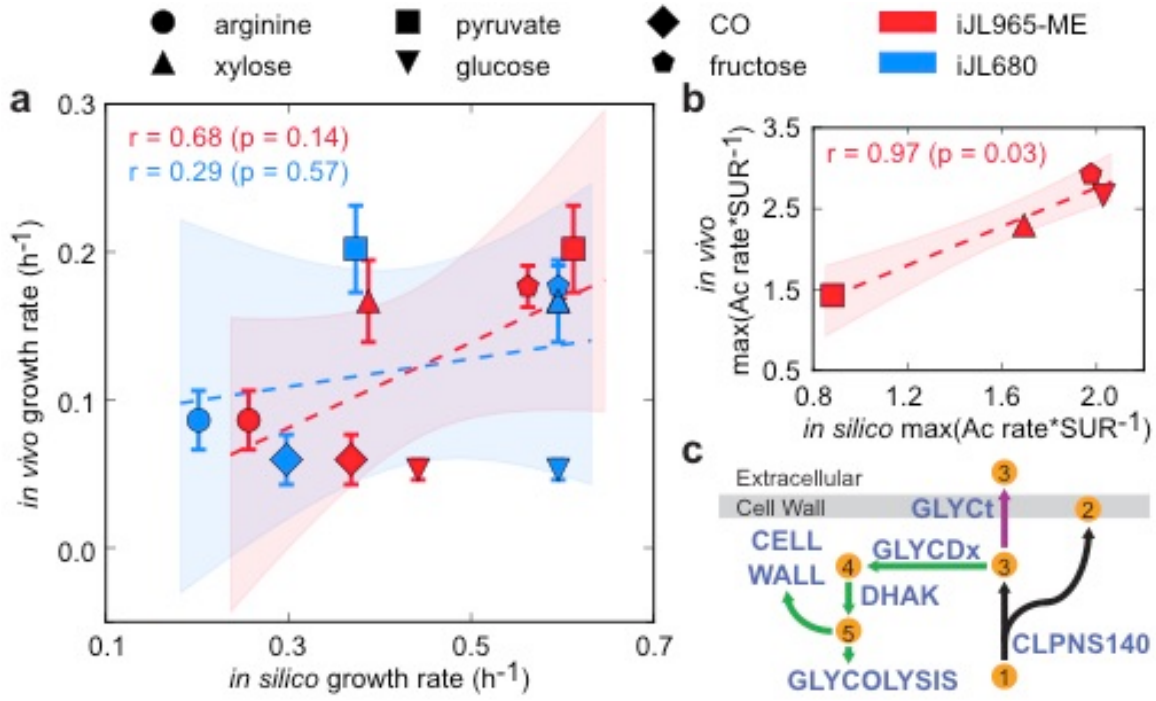

Fig. 9: Predictions of growth rate and product production. a, Two sets of predicted growth rates, from iJL680 and iJL965-ME, were plotted against *in vivo* measured growth rates for arginine, xylose, pyruvate, glucose, CO, and fructose growth conditions (±std, n=3). Linear regressions and 95% confidence intervals were represented by dashed lines and shaded areas, respectively. In iJL680, carbon atom uptake was constrained to 30 mmol*gDW^{-1}*h^{-1}, while in iJL965-ME, the optimal carbon uptake was constrained by inherent proteome limitations. r and p represent Pearson's correlation and p-value. **b,** Predicted maximum acetate secretion rate (Ac; mmol*gDW^{-1}*h^{-1}) to substrate uptake rate (SUR; mmol*gDW^{-1}*h^{-1}) was plotted against measured averaged values. **c,** Predicted pathway mechanism for observed glycerol production in spent media. Glycerol was a byproduct of cell membrane formation during cardiolipin production. While the cell was carbon-limited, glycerol was recycled into biomass using the pathway highlighted in green. When cells were proteome-limited, *C. ljungdahlii* secreted glycerol (purple arrow). Abbreviations: 1 = phosphatidylglycerol (n-C14:0), 2 = cardiolipin (n-C14:0), 3 = glycerol, 4 = dihydroxyacetone, 5 = dihydroxyacetone phosphate, CLPNS140 = cardiolipin synthase (n-C14:0), GLYCt = glycerol transport, GLYCDx = glycerol dehydrogenase, DHAK = dihydroxyacetone kinase.

Interestingly, iJL965-ME predicted previously unknown secretion of glycerol (<2.5e-3 mmol*gDW^{-1}*h^{-1}) following acetate and ethanol production during growth on xylose or glucose, but not on arginine or pyruvate. Like ethanol, glycerol secretion occurred due to trade-offs in proteomic limitations resulting in overflow metabolism, as the cell no longer invested resources to recycle glycerol, a byproduct of cardiolipin production (**Fig. 9C**). In order to verify glycerol production, we carried out HPLC analysis and measured 0.024±0.012 mM and 0.083±0.018 mM of glycerol from cultures grown on either xylose or glucose, respectively.

Predicting gene expression. Because RNA and protein abundance requirements are coupled to reaction fluxes in ME-models instead of a lumped biomass composition like in M-models, ME-models enable *in silico* predictions of transcription and translation (mmol*gDW^{-1}*h^{-1}) (O'Brien et al., 2014, Lloyd et al., 2017). To test the accuracy of our model, genes were

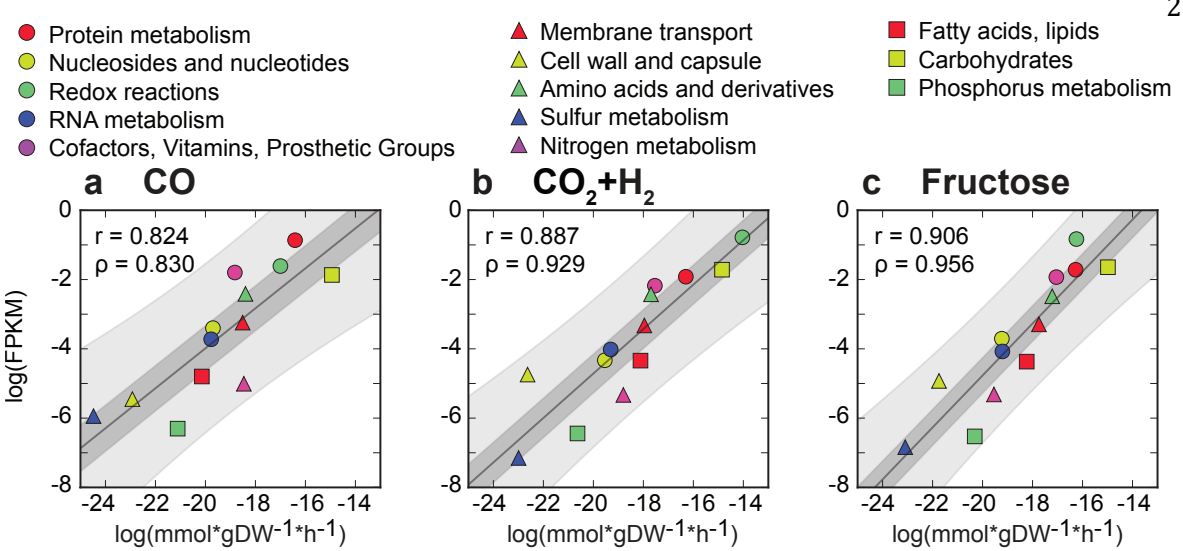

Fig. 10: Predicted and experimental gene expression. Categorized by RAST subsystem and summed, predicted gene expression (transcription flux reactions) was compared to RNA-seq data for *C. ljungdahlii* grown on **a**, CO, **b**, CO_2+H_2, and **c**, fructose. Linear regressions, 95% confidence intervals of the regression, and 95% prediction intervals are represented by lines, dark shaded areas, and light shaded areas respectively. Scatter plots shown are for the highest Pearson r between predicted and experimental data. Normalized total transcription flux (mmol*gDW^{-1}*h^{-1}) of the Wood-Ljungdahl pathway was plotted against carbon substrate uptake rate for **d**, CO, **e**, CO_2+H_2, and **f**, fructose. Pearson r reflects correlation with growth rate.

categorized by RAST subsystems and summed as per predicted transcription flux reactions. The *in silico* results were strongly correlated to RNA-seq data for *C. ljungdahlii* grown on CO, CO_2+H_2, or fructose (r >=0.82). At the highest correlation, all categories fell within the prediction interval of the linear regression (**Fig. 10A-C**), enabling to forecast substrate-specific expression of pathways.

At the gene level, 396 genes could be strongly linked to growth rate (r>0.9, p<0.05*Bonferonni). However, correlation of these genes was dependent on the growth substrate (68 genes for CO, 275 for CO_2+H_2, and 224 for fructose). Growth-correlated genes that were shared between conditions involved genes related to translation (e.g. rRNA and specific tRNAs). Under autotrophic conditions, expression of WLP genes were correlated more with substrate availability than growth rate (r_{CO}: 0.983>0.955, r_{CO2+H2}: 0.996>0.884; **Fig. 10D, E**). In addition, the reactions fluxes of essential WLP reactions carbon monoxide dehydrogenase (CODH4) and 5,10-methylenetetrahydrofolate reductase (MTHFR5) were linearly related to CO uptake during growth on CO, while other non-WLP redox reactive reactions (e.g. RNF) were correlated with growth rate. Similarly, WLP reactions were linearly linked to CO_2 uptake in CO_2+H_2 conditions, in addition to the linear response of ferredoxin:NADPH hydrogenase to H_2, while non-WLP redox reactions were correlated with growth rate.

In heterotrophic conditions, the WLP was more active under nutrient-limitations than proteome limitations, as its activity level was related to acetate secretion (r = 0.993, p<0.01,

Fig. 10F). The WLP was recapturing CO_2 for biomass production using the reducing power gained by metabolizing fructose. At greater than 57% of the optimal fructose uptake (**Fig. 10F**), the primary provider of oxidized ferredoxin switched from WLP to ferredoxin:NADP reductase (FRNDPR2r) and acetaldehyde:ferredoxin oxidoreductase (AOR_CL). Extraneous reducing power captured by NAD^+ from glyceraldehyde-3-phosphate dehydrogenase (GADP) was removed by producing ethanol (alcohol dehydrogenase; ALCD2x). These findings are corroborated by a previous report that *C. ljungdahlii* grows mixotrophically, instead of heterotrophically, when presented with sugar as a carbon source (Jones et al., 2016).

Nickel controls phenotype through Wood-Ljungdahl activity. In M-models, metal availability and growth rate are linearly correlated even though there is contrary experimental evidence (Saxena and Tanner, 2011). In iHN637, seven of ten metals (Ca^{2+}, Cu^{2+}, Mg^{2+}, Mn^{2+}, Mo^{2+}, Ni^{2+}, Zn^{2+} + Co^{2+}, Fe^{2+}, Na^+) could only be imported or exported (in addition to their inclusion in the biomass objective function, which represents the total composition of the cell (Feist and Palsson, 2010)), and only Co was predicted to participate in flux-carrying reactions that were not a transport reaction or biomass production. Thus, most metal ions were not associated to the reactions they help catalyze. This represents a general fact for M-models. Cofactor integration in iJL965-ME, however, allows systematic interrogation of the effects of metal availability. Particularly, iJL965-ME's only nickel-containing proteins, CODH4 and carbon monoxide dehydrogenase:Acetyl-CoA synthase (CODH_ACS), are part of the WLP, which afforded the possibility of controlling this pathway through changes in media composition both *in silico* and *in vivo.* Due to *C. ljungdahlii*'s reliance on WLP for autotrophic growth, nickel was predicted to be

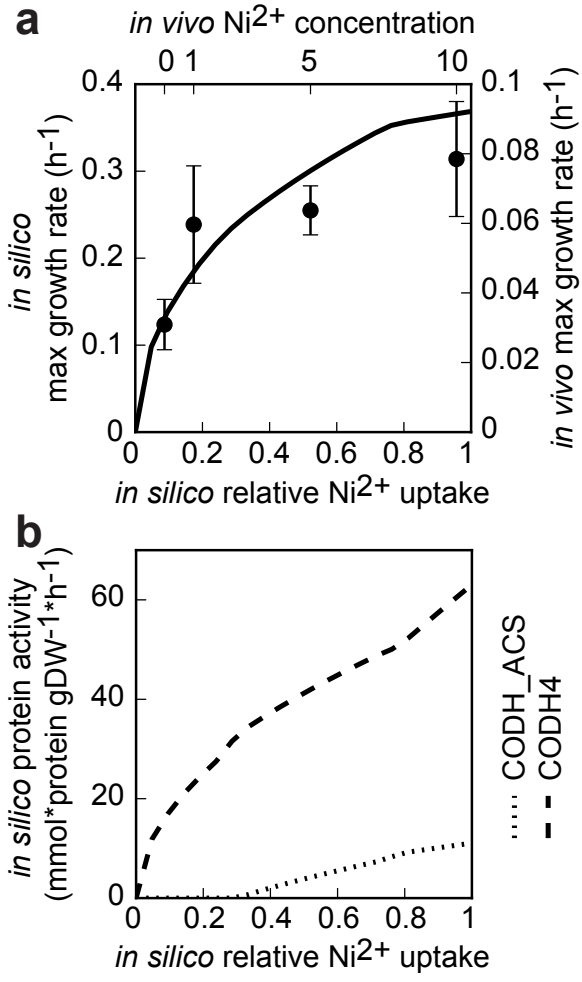

Fig. 11: Effects of nickel availability on *C. ljungdahlii* grown on CO. a, Maximum predicted growth rate was plotted against relative nickel uptake (line), and *in vivo* maximum growth rate verses the concentration of added nickel was plotted on the opposite axes (dot, ±std, n=3). **b**, Predicted protein activity of the nickel-containing enzymes, carbon monoxide dehydrogenase (CODH4) and carbon monoxide dehydrogenase:acetyl-CoA synthase (CODH_ACS), was plotted against relative nickel uptake.

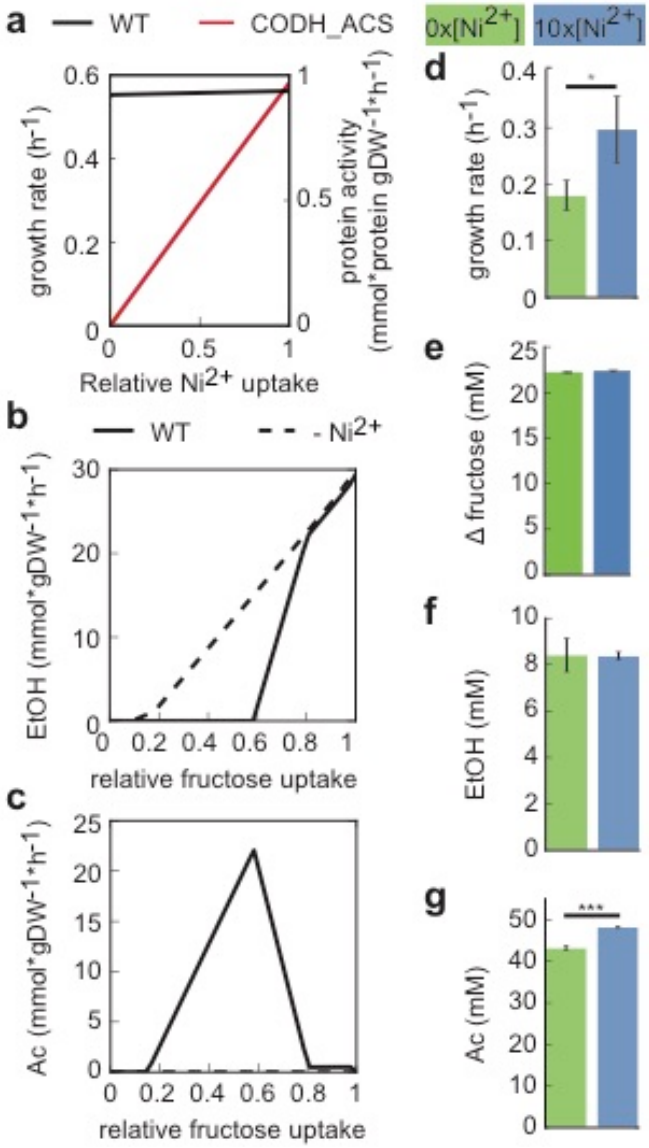

Fig. 12: Effects of nickel availability on *C. ljungdahlii* grown on fructose. a, Predicted growth rate and protein activity of carbon monoxide dehydrogenase:acetyl-CoA synthase (CODH_ACS) were plotted against relative nickel uptake (mmol*gDW^{-1}*h^{-1}). **b**, Predicted ethanol (EtOH) secretion at optimal nickel uptake (WT) and no available nickel (-Ni^{2+}) were plotted against relative fructose uptake (mmol*gDW^{-1}*h^{-1}). **c**, Predicted acetate (Ac) secretion at optimal nickel uptake and no available nickel were plotted against relative fructose uptake (mmol*gDW^{-1}*h^{-1}). Measured **d**, growth rate, **e**, fructose consumption, **f**, final ethanol concentration, and **g**, final acetate concentration of fructose-grown *C. ljungdahlii* without added nickel and with ten times the concentration of nickel were plotted (±std, n=3). Gray asterisk indicates difference significance is p=0.06, and three black asterisk indicates significance of p<0.001.

essential for CO-growth. Although true essentiality could not be tested due to trace nickel in the media, the amount of additional nickel (added as multiples of 0.10 mM) significantly influenced *in vivo* growth rate in a quadratic fashion as predicted (**Fig. 11A**). According to iJL965-ME, the non-linear effects of nickel limitations were caused by an uneven distribution of metal resources between CODH_ACS and CODH4, resulting in different rates of decreasing protein activity (**Fig. 11B**). In turn, the other reactions in WLP were correlated to either CODH_ACS, like MTHFR5 and methyltetrahydrofolate corrinoid/iron-sulfur protein methyltransferase (METR), or CODH4. Finally, iJL965-ME predicted that while nickel availability affected growth rate, protein activity, and acetate and ethanol yield, the acetate-to-ethanol production rate would not change. The acetate:ethanol production rate ratio, as determined by HPLC, remained constant at 1.4 for different nickel concentrations. Acetate:ethanol production rate was unchanged with a ratio of 1.48±0.34, regardless of the nickel concentrations used (0x, 1x, and 5x [10x excluded due to carbon depletion].

iJL965-ME predicted that nickel limitations would have different effects on fructose-grown cells. Removal of nickel was not predicted to affect growth rate or fructose uptake significantly (Δ_{gr}=98%, $\Delta_{fructose}$=99%, Fig 6A). However, there was no CODH_ACS or METR activity under nickel depletion, which reduced the WLP activity and eliminated acetate secretion. Instead, the model predicted that only ethanol secretion would occur (**Fig. 12B,**

C). To test this prediction, *C. ljungdahlii* was grown either without added nickel (0x) or with high nickel concentrations (10x). Both cultures consumed the same amount of fructose (p=0.26) and produced identical amounts of ethanol (p=0.95), but exhibited different growth rates (p=0.062) and final concentrations of acetate (p=2.2e-4) (**Fig. 12D-G**). Increased acetate secretion rate (p=0.016) and final acetate concentrations in the 10x condition were due to the nickel-stimulated WLP consuming more CO_2.

We showed that the incorporation of the E-matrix into constraint-based genome-scale models significantly widens the scope of their application, including prediction of overflow metabolism and optimal expression levels, as well as media optimization strategies. Such capabilities proved useful for exploring and understanding system responses of *C. ljungdahlii*. The reconstructed *C. ljungdahlii* ME-model (iJL965-ME) was not only more accurate than the M-model at predicting growth rates and acetate secretion rates, but was also capable of predicting secretion of ethanol (H_2, as a less effective oxidizing agent than CO, was an exception) and the novel secretion of glycerol (**Figs. 8, 9**). Furthermore, *in silico* predictions of gene/subsystem expression were highly comparable to *in vivo* transcriptomics for three separate conditions, bolstering confidence in predicting macromolecular responses to environmental changes (**Fig. 10A-C**). C1 metabolism under both autotrophic and mixotrophic conditions was examined in more depth, and the potential of controlling WLP activity through media composition was explored (**Figs. 10-12).** Note that acetogens grow mixotrophically while using organic substrates. Although the lack of CODH_ACS activity (achieved by removing nickel from the media) may not cease WLP activity entirely, it may stop acetate production (as *in vivo* nickel depletion results suggest), leading to ethanol production as the main fermentation end product (**Fig. 12**). However, the discrepancy between *in silico* and *in vivo* growth rates of nickel-depleted cells grown on fructose implied that WLP was more important than predicted for maximizing growth in mixotrophic conditions (**Fig. 12**). In contrast, nickel was essential for CO-growth, but had no effect on the acetate:ethanol ratio (**Fig. 11**).

As demonstrated in this study, ME-models like iJL965-ME provide a comprehensive, genome-scale, systems biology approach that links the environment and macronutrient metabolism. In particular, the combination of C1 metabolism, multi-omics predictions, and cofactor integration in iJL965-ME is an important milestone for a holistic understanding of metals in metabolism. Although nickel was the only trace metal to be investigated here, iJL965-ME invites further studies elucidating specific effects of concurrent metal limitations and genetic perturbations. The ME-model represents an inclusive method that unites analysis and integration of multiple data types.

C Exploring the evolutionary significance of tRNA operon structure using metabolic and gene expression models

An operon is a co-regulated cluster of genes that are expressed on the same RNA transcript. These genomic features arise through a variety of means, including horizontal gene transfer that places a gene under another gene's promoter, horizontal gene transfer of whole operons, deletion of intervening sequences, and genome rearrangement. Though the presence of an operon may be a random event, selection pressures can drive the maintenance of operons. For example, potential benefits bestowed by an operon onto the host organism include a reduction in regulation costs (Price et al., 2005), diminished stochastic gene expression through synchronicity of protein ratios (Ray and Igoshin, 2012; Nunez et al., 2013) and insurance that all functional steps in a pathway are produced (Zaslaver et al., 2006). Such theories hint at an evolutionary optimization problem to promote efficiency in gene expression.

In order to optimize cellular efficiency, translation must be carefully controlled because it requires the highest energy and resource expenditure of any process in fast-growing cells. Since the available tRNA pool could be rate-limiting during protein translation (Kurland, 1993), close correspondence between codon usage and the available tRNA pool, often quantified through the tRNA adaptation index (tAI) (dos Reis et al., 2004), must be maintained efficiently. Even though tRNA co-expression explained *E. coli's* tRNA profile better than tRNA gene copy number (widely recognized as a correlated estimate for tRNA profile (Kanaya et al., 1999; McDonald et al., 2015) relatively few papers have investigated the influence of operons on tRNA expression levels (Wald et al., 2014). Yet rRNA and tRNA genes can often be found on the same operon, and 23.8% of all tRNA genes from prokaryotic genomes sequenced by 2014 were found to be located in an operon with another tRNA gene Wald et al., 2014). Such evidence implies that evolutionary pressures may also shape genomic tRNA structure.

Constraint-based modeling offers a biophysically-based approach to estimate tRNA concentrations and usage. In particular, constraint-based metabolic and gene expression models (*i.e.*, ME-models) are well-suited for examining potential insights into operon structure. The scope of predictions that ME-models cover is extensive; these models account for transcription, tRNA charging, translation, and metabolic reactions. Additionally, ME-models incorporate the underlying genome architecture through transcriptional units that account for co-expression of genes. ME-models have been used to successfully recapitulate several levels of phenotypes, from growth rates to pathway expression levels, and even undiscovered operons (Lerman et al, 2012; O'Brien et al., 2013). As of writing, only *E. coli* and *C. ljungdahlii* have completed ME-models that use the COBRAme framework, which

allowed comparisons of model perturbations with the knowledge that the constraints within the models (e.g., coupling constraints) were similarly formulated.

Using the two available COBRAme-based ME-models, one for *Escherichia coli* and one for *Clostridium ljungdhalii,* we examined the systematic importance of tRNA co-expression. We validated the two models for the purposes of this study and examined the tRNA operon structures, thereby identifying unique tRNA operon solutions to two different selective pressures. One solution led to optimization of phenotype through fragmenting operons and the other solution to optimized efficiency through optimal grouping of tRNAs.

tRNA operon structure: Fragmentation versus modularity. Examination of tRNA-containing operons organization in two bacteria, the fast-growing generalist *E. coli* and the slower-growing homoacetogen *C. ljungdahlii*, revealed two different strategies (Caspi et al., 2008). These two strategies will be referred to as fragmentation, where tRNA organization leads to both a high number of singly-transcribed tRNA genes and a minimization of co-transcribed tRNA species, and modularization, which is the tendency towards polycistronic tRNA genes. In *E. coli*, 23% of tRNA genes could be transcribed monocistronically, and 37% could be expressed as polycistronic transcripts that lack other tRNA genes. When considering unique tRNA species by anticodon, the number of single transcripts that can be uniquely expressed increased to 54%, and for tRNA species by amino acid (AA), 56%. Furthermore, *E. coli* appeared to favor less tRNA genes per transcript and did not have an operon containing more than seven tRNAs, while the highest number of unique tRNA species per operon was four. Thus, *E. coli* displays a fragmentation strategy for its tRNA operon structure (blue bars, **Fig. 13**).

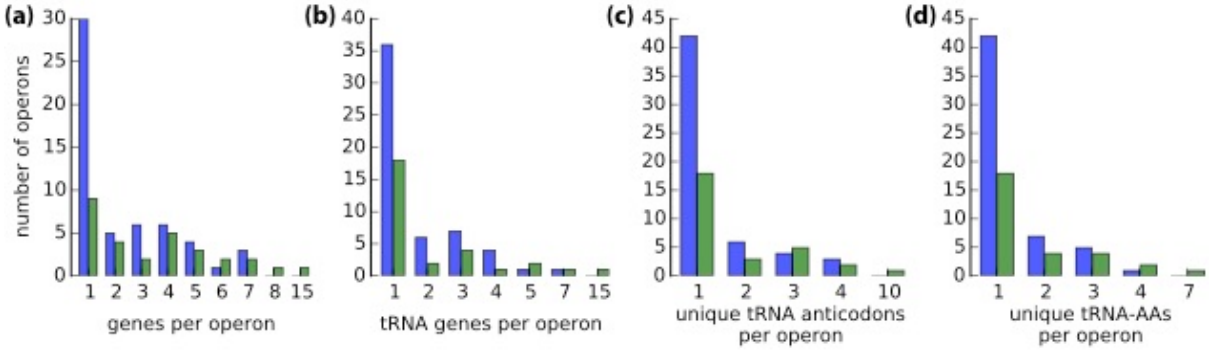

Fig. 13: **Distribution of tRNAs by operon in *E. coli* and *C. ljungdahlii*.** Bar graphs show operon count by (a) the number of genes per operon, (b) the number of tRNAs per operon, (c) the number of unique anticodons as represented by tRNAs per operon, and (d) the number of unique amino acids as represented by tRNAs per operon for *E. coli* (blue) and *C. ljungdahlii* (green). All potential operons, including alternative start and end sites, are included.

In case of *C. ljungdahlii*, the analysis revealed that only 8.4% of tRNA genes could be expressed monocistronically and 26% could be expressed as polycistronic transcripts lacking other tRNA genes. Looking at tRNA species, only 32% of tRNAs by anticodon and 34% of tRNAs by AA were capable of being uniquely expressed on a single transcript. Thus,

C. ljungdahlii had the majority of its tRNA species co-transcribed with another tRNA type, and *C. ljungdahlii* could express fifteen tRNAs, including the only tRNA-his gene, on a single transcript. The bias towards polycistronic tRNA genes means that *C. ljungdahlii* prefers modularization in comparison to *E. coli* (green bars, **Fig. 13**).

Predicted tRNA charging amino acid usage is consistent with amino acid requirements. AA compositions predicted by the *E. coli* ME-model (iLE1678-ME) and the *C. ljungdahlii* ME-model (iJL965-ME) were compared against *in vivo* data. AA composition was calculated from transcriptomic data using RNA-seq (FPKM) data from *E. coli* batch-grown on glucose, glycerol, xylose, and acetate and *C. ljungdahlii* batch-grown on fructose, CO and CO_2 +H_2 as a proxy for protein count. Only proteins reconstructed in the ME-models were considered. For each substrate condition, the ME-models were simulated at maximum growth rate (which was calculated when substrate availability was greater than what can be consumed, and considered to be equivalent to *in vivo* batch growth), half of the maximum substrate uptake rate, and minimal substrate availability (*i.e.*, tenth of maximum substrate uptake rate). Predicted AA compositions were calculated from tRNA charging reactions (mmol*gDW−1*h−1) which reflects the exact AA requirements of the *in silico* cell.

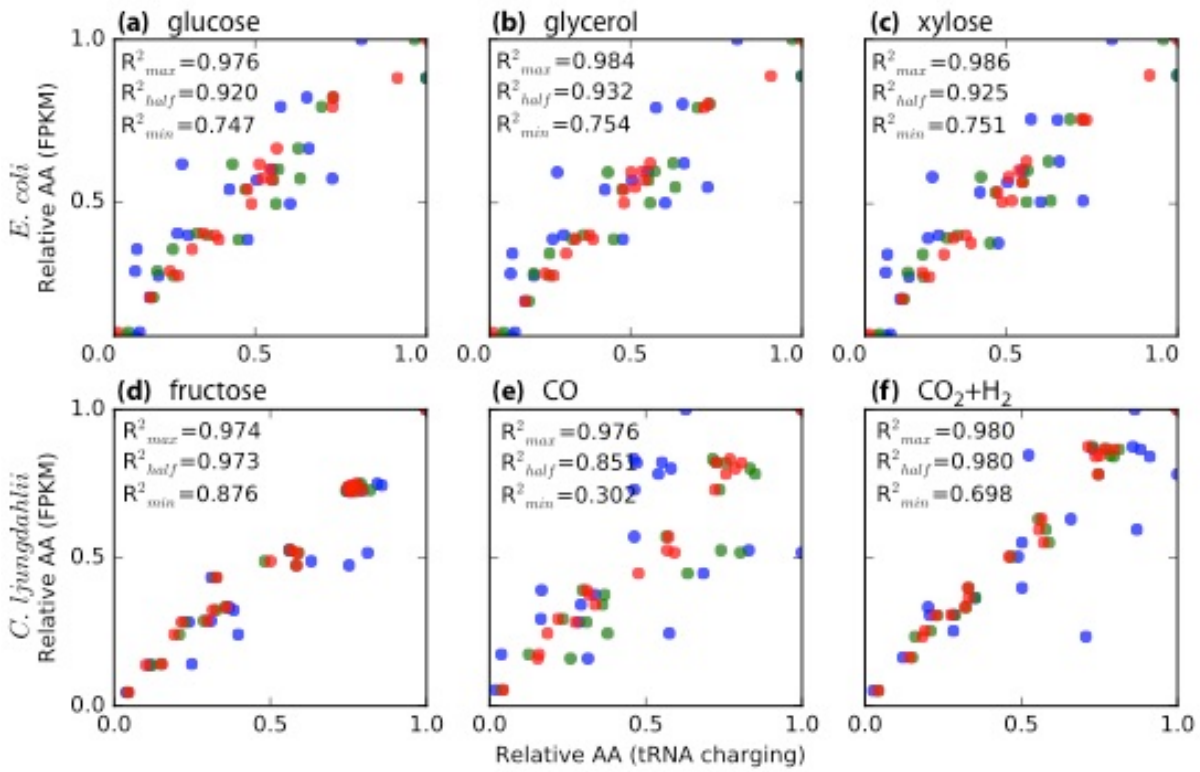

Fig. 14: **Comparing *in silico* and *in vivo* AA composition for *E. coli* and *C. ljungdahlii*.** *In vivo* AA compositions were calculated using RNA-seq, harvested mid-log phase from batch-grown cells, as a proxy for protein count. *In silico* AA compositions were the sum of AA-categorized tRNA charging reactions (mmol*gDW*−1*h−1) at maximum growth rate (red), half of the maximum substrate uptake (green), and minimum (*i.e.*, tenth) of the maximum substrate uptake rate (blue) on glucose, glycerol, or xylose for *E. coli* (top row) and fructose, CO, or CO_2 +H_2 for *C. ljungdahlii* (bottom row). Values are relative to the most AA required, which is alanine for *E. coli* and lysine for *C. ljungdahlii*.

The predicted and measured AA compositions were highly comparable ($R2 \geq 0.964$ for all batch-growth conditions in both models; **Fig. 14**). The high correlation between *in silico* and *in vivo* values continued to hold true for tRNA molecule concentrations (uM) and calculated AA composition from protein expression (ribosome profiling, RPKM) in *E. coli*, both of which were more appropriate comparisons for *in silico* tRNA expression and tRNA charging reactions. With these validations for AA composition and our knowledge of the genome architecture, we have confidence in the output of translation and the underlying structure of transcription in the ME-models for batch conditions. The goodness of fit decreased when *in vivo* batch-grown cells were compared to *in silico* growth on half of the maximum substrate uptake rate and minimal substrate availability. Thereby iLE1678-ME and iJL965-ME demonstrated their capability to predict variable AA compositions dependent on substrate availability. Furthermore, expression values from *in silico* minimal and half substrate availability were able to explain tRNA molecule concentrations in low growth rate (0.4 h−1) better than *in silico* maximum growth rate could. Although the higher correlations imply that ME-models continue to be accurate at lower growth rates, the actual influence of growth rate on tRNA pools is currently inconclusive and requires more investigation. Despite the lack of evidence to support conclusions from non-optimal growth rates, ME-models still provide an opportunity to specifically examine the effects of varying tRNA operon structure.

Optimized tRNA operon structure meets tRNA abundance requirements. To examine whether tRNA gene location and co-transcription influences the cell, 1000 models with all tRNAs randomly shuffled into another tRNA's location, henceforth referred to as Monte-Carlo (MC) tRNA location models, were built for *E. coli* and *C. ljungdahlii* each. The MC tRNA location models were then simulated with validated substrates (glucose, glycerol, xylose, and acetate for *E. coli* and fructose, CO, and CO_2+H_2 for *C. ljungdahlii*) at maximum growth rate, half of the maximum substrate uptake rate, and minimal substrate uptake (**Fig. 15**). With this setup, we can examine whether the two organisms' different tRNA organization strategies, fragmentation and modularization, promote optimization for translational purposes under particular growth conditions.

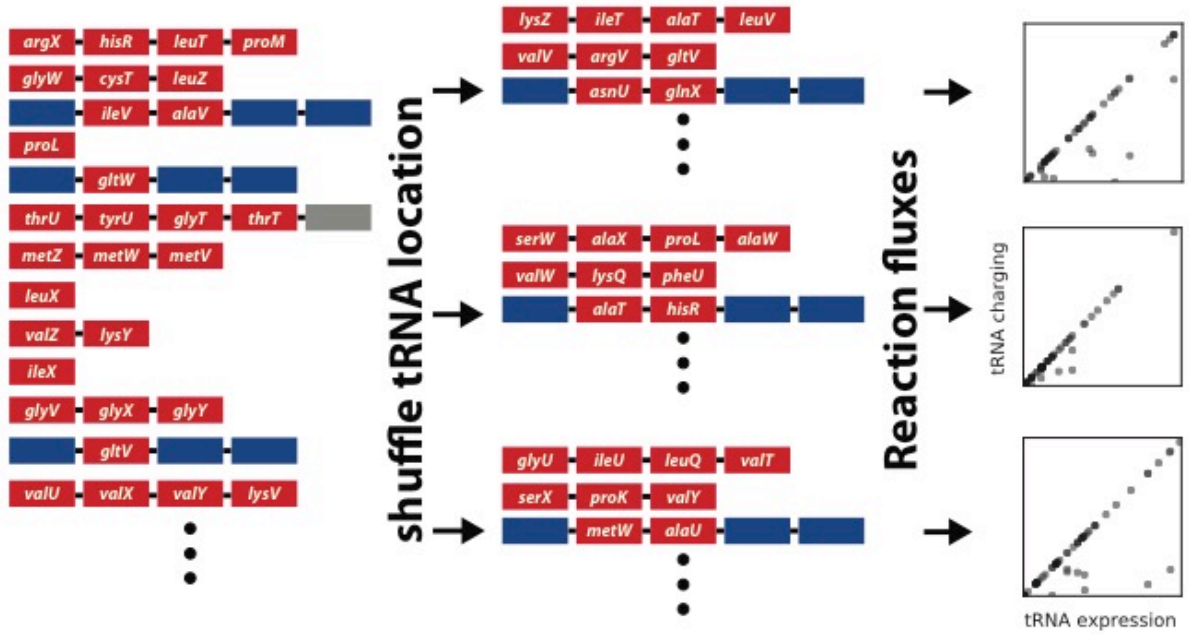

Fig. 15: **Diagram of the Monte-Carlo method for tRNA location shuffling.** Red boxes represent tRNA genes, blue boxes represent rRNA genes, and grey boxes represent open reading frames. Operon diagram is not to scale for gene size and distance.

Shuffling tRNA order and location has a dramatic effect on tRNA expression, as the range of AA-categorized tRNA (tRNA-AA) expression can vary drastically in relation to other tRNA-AA molecules (**Fig. 16**). When tRNA-AA expressions of the MC tRNA location models were compared against the original models' (iLE1678-ME and iJL965-ME which contain published genome architectures), tRNA expression was revealed to be minimized. Both iLE1678-ME and iJL965-ME performed better than the median MC tRNA location model because they expressed less total tRNA for a significant number of tRNA-AA molecules ($p < 0.02$ for all maximum growth rate conditions; **Fig. 16**). Thus, the original tRNA operon structures led to reduced cost of tRNA expression.

In contrast to the flux ranges of tRNA expression, the AA composition of the cell, as represented by tRNA charging reactions, remains relatively constant. Regardless, iLE1678-ME and iJL965-ME revealed that the published tRNA operon structures also promoted utilization of tRNA usage (*i.e.*, tRNA charging reactions) at maximum growth rate. For a significant number of tRNA-AA molecules, iLE1678-ME and iJL965-ME used more tRNA in tRNA charging reactions than the median MC tRNA location model ($p < 0.05$ for all conditions but two; **Fig. 16**). *E. coli* on acetate and *C. ljungdahlii* on fructose were the exceptions, as tRNA expression was minimized, but tRNA usage was not maximized. Thus, the original tRNA operon structures generally led to increased tRNA usage.

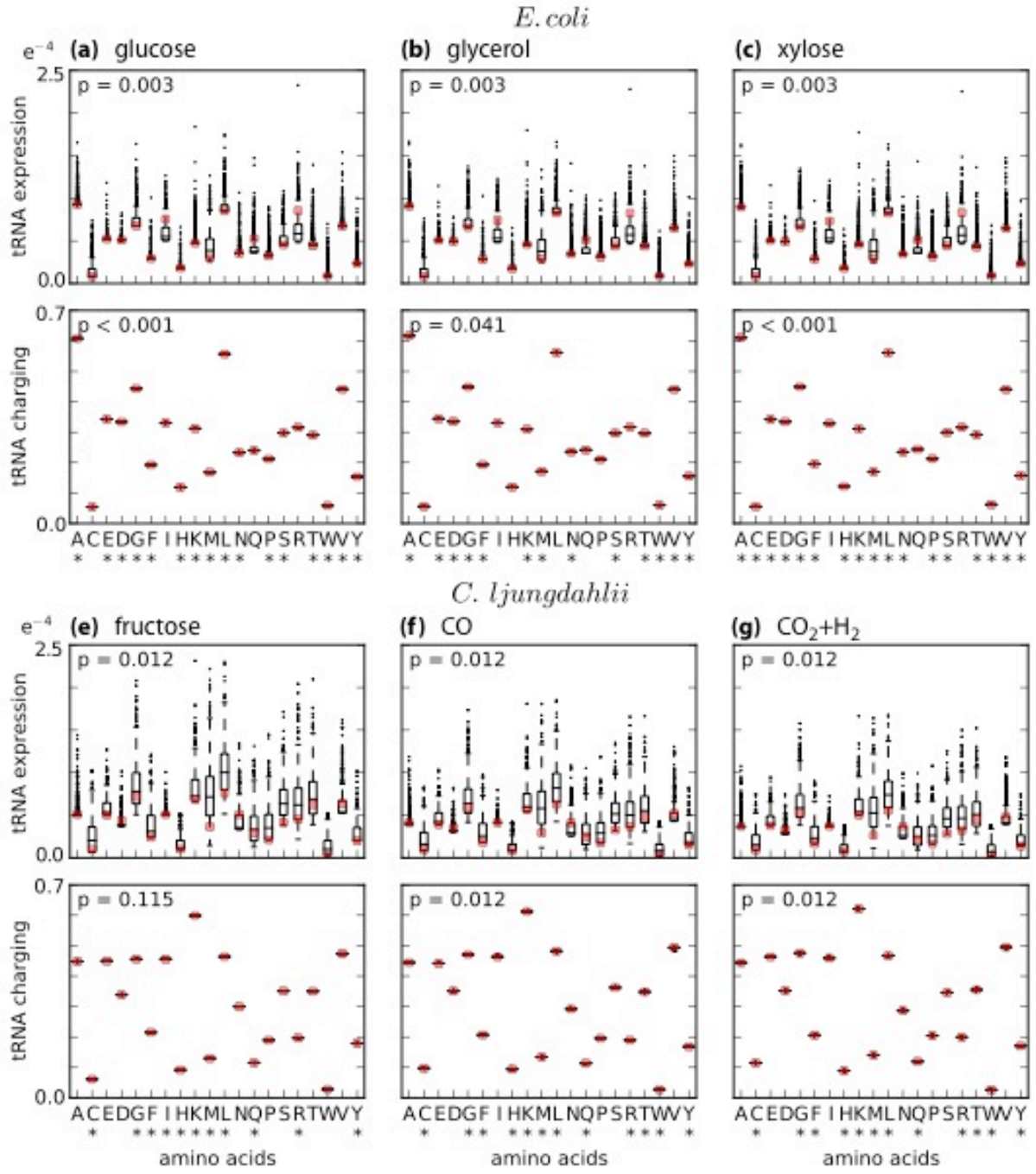

Fig. 16 Comparing tRNA expression and tRNA charging fluxes against the original models'. AA-categorized *in silico* tRNA expression (mmol*gDW−1) and tRNA charging fluxes (mmol*gDW−1) from the MC tRNA location models were plotted as box-plots, and red dots indicate the original models' predictions. *E. coli* was batch simulated on (a) glucose, (b) glycerol, and (c) xylose, and *C. ljungdahlii* on (d) fructose, (e) CO, and (f) CO_2+H_2 . P values are from binomal tests of whether the original models give rise to lower expression levels or higher tRNA usage than the median values from the MC tRNA location models. Asterisks indicate tRNAs by AA that had both less than average tRNA expression and greater than average tRNA usage.

If tRNA expression could be likened to capital costs and tRNA usage to operating costs, then *E. coli* and *C. ljungdahlii* have minimized capital costs by optimizing expression of necessary tRNAs. The operating costs have likewise been maximized, even though tRNA operon structure does not influence operating costs as strongly as it does capital costs, as seen

through the lack of fluctuation in tRNA usage and the non-optimal tRNA usage in acetate-grown iLE1678-ME. Together, these observations suggest that the cells partly control their capital expenses at maximum growth rate though tRNA operon structure. At least half of the tRNA-AA molecules in the original models have both lower expression and higher usage than the median MC tRNA location model (*i.e.*, tRNA-AA optimization) at maximum growth in multiple substrate conditions (**Fig. 16**). However, *E. coli* and *C. ljungdahlii* did not optimize the same tRNA-AA molecules, with only F, G, K, M, and Y being shared between the two models, thereby showing that optimized tRNA-AA molecules may differ by organism.

Both iLE1678-ME and iJL965-ME displayed less efficient tRNA expression and tRNA charging usage as growth rate dropped from maximum, and they were no longer efficient at minimum growth rate with the exception of CO_2 +H_2, implying that tRNA operon structures have been optimized for growth when nutrients were abundant. The number of optimized tRNA-AA molecules also decreased with growth rate. *E. coli* on xylose and *C. ljungdahlii* stood out as retaining the most optimized number of tRNA-AA molecules with 9 AAs and 7 AAs respectively. Perhaps this optimization of tRNA-AA molecules for lower growth rate inducing substrates (gr*glucose* = 0.92 vs gr*xylose* = 0.87; gr*C O*= 0.38 vs gr*C O*2 +*H*2 = 0.31) hints at an evolutionary process that ensured continued resource efficiency in less desirable conditions once preferred substrates are depleted.

Positive selection for high tRNA efficiency. Despite a trend towards minimization in capital expenses, iLE1678-ME (*E. coli*) performed at an average in total tRNA efficiency, as measured by the total tRNA usage to total tRNA expression ratio, compared to the MC tRNA location models (**Fig. 17e**). Its maximum growth rate was also average (**Fig. 17g**). However, when the range of tRNA efficiency values and growth rates of the MC tRNA location models were compared against *C. ljungdahlii's* ranges, *E. coli* has evolved to minimize the potential error around tRNA efficiency, rRNA expression, and growth rate (**Fig. 17**). Fragmentation of the operon structure ensured that regardless of tRNA order or location, potential phenotypes cannot deviate too far from the original value (**Fig. 17a, b**), which may reflect a history of tRNA genes being regularly added and subtracted from the genome to reach its current, optimal state (Wald et al., 2014). The only non-random gene locations in tRNA-containing operons were occupied by rRNA genes, which refers to the set of 16S, 5S, and 23S rRNAs. In iLE1678-ME, all seven rRNA gene sets were co-expressed with tRNA genes, and rRNA expression was driven, in part, by the need for the associated tRNA genes. All three of the tRNAs with anticodon UGC, which codes for tRNA-ala, were on a polycistronic transcript with an rRNA gene set. Since alanine was the most required AA, iLE1678-ME subsequently expressed a significant amount of rRNA genes at maximum growth rate (**Fig. 17f**). The selective maximization of rRNA expression points at growth rate optimization in *E. coli*, as ribosome amount is linearly correlated to growth rate (Scott et al, 2010).

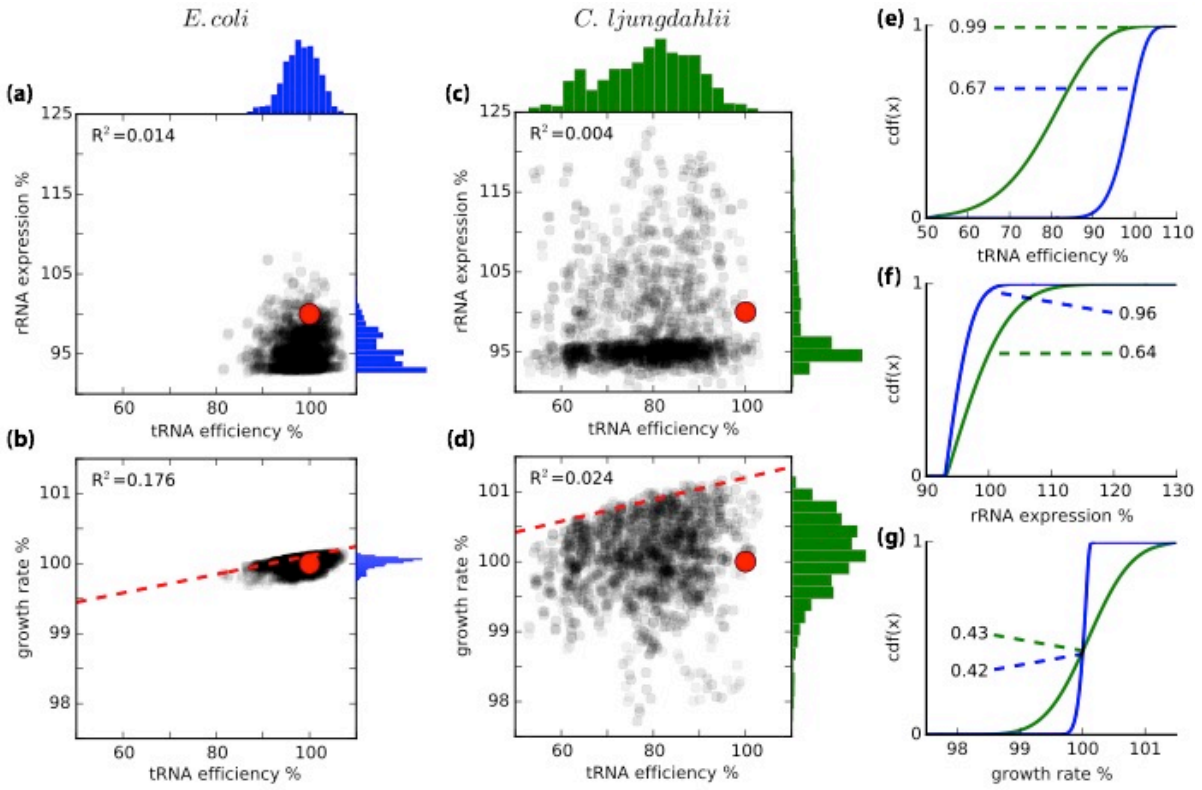

Fig. 17 Comparing efficiencies and growth rates from the MC tRNA location models as a percentage of the original models'. All results are from batch simulations on different substrates, with *E. coli* data coming from glucose, glycerol, or xylose conditions (n=3000), and *C. ljungdahlii* from fructose, CO, or CO2+H2 conditions (n=3000). (a & c) rRNA expression (LrRNA mmol*gDW−1) was plotted against protein:tRNA (LtRNA charging mmol*gDW−1:LtRNA mmol*gDW−1). (b & d) Growth rate (h−1) was plotted against protein:tRNA, and soft upper boundaries (red dashed line) were found. Red dots represent the original models' averaged results. R2 values are from linear regressions. A histograph of each dataset is displayed opposite to its axis in (a- d). Cumulative density functions (cdf) calculated from the histographs in (a-d) were plotted against (e) protein:tRNA, (f) rRNA expression, and (g) growth rate. Dotted lines indicate the probability of obtaining a value less than the original models' prediction for *E. coli* (blue) and *C. ljungdahlii* (green).

While *E. coli* has been optimized for output, particularly rRNA production, *C. ljungdahlii* seemed to be focused on minimizing capital expenditures, as demonstrated by the significantly high tRNA efficiency in iJL965-ME which remained high even as growth rate dropped, while both growth rate and rRNA expression were average compared to the MC tRNA location models (**Fig. 17**). However, average rRNA expression may also point to efficient resource usage. Unlike rRNA arrangement in *E. coli*, seven of iJL965-ME's nine rRNA gene sets were co-expressed with tRNAs. Furthermore, *C. ljungdahlii* does not associate a specific tRNA species with rRNA, which allowed *C. ljungdahlii* the ability to fine tune its rRNA need by expressing operons with the necessary amount of tRNAs per species, thereby minimizing resources spent on producing more rRNA, while *E. coli* has evolved so that an abundant amount of rRNA is available for maximum growth rate. Finally, unlike *E. coli's* tight range of values, shuffling of tRNA locations would lead to drastic changes in tRNA efficiency, rRNA expression, and growth rate. Thus, in contrast to *E. coli's* fragmentation, modularization in *C. ljungdahlii* sacrificed growth rate for tRNA efficiency and resource

frugality.

Although tRNA efficiency, rRNA expression, and growth rate were not correlated ($R2 \leq 0.176$, Fig 3.10a-d), there were operon structures that resulted in higher growth rates. This may not be so important for *E. coli*, since its range of potential growth rates was limited, and the payoff between tRNA efficiency and growth rate was low (slope of the upper soft boundary, m*E.coli* =0.010)

Does tRNA operon structure reflect K/r strategists? Fragmentation and modularization may hint at a deeper understanding of the differences between K- and r-strategists, where K-strategists are typically associated with slow growth due to limitations by density-dependent controls, and r-strategists with fast growth (Note: K and r strategists can be differentiated by their maximum specific growth rate under conditions with excess substrate (*i.e.*, batch growth). *C. ljungdahlii*, as a K-strategist (max *in silico* growth rate on fructose is 0.57 h*−1), evolved to maximize efficiency of resources at the tRNA operon structure level. Thus, *C. ljungdahlii* matches cost to need, which may provide *C. ljungdahlii* with a slight edge over competitors when nutrients are limiting for the ecological community. *E. coli*, an r-strategist (max *in silico* growth rate on glucose is 0.92 h*−1), has evolved to always perform near optimum in regards to its tRNA operon structure, and rRNA expression, which is tied to tRNA expression, is maximized. Furthermore, *E. coli's* fractured tRNA-containing operon structure may allow *E. coli* to quickly match tRNA-demands specific to available substrates, as *E. coli* is a generalist that consumes multiple carbon sources. Thus, *E. coli* has optimized its output, which may allow it to persist in an ecological community through rapid growth.

Although ME-models currently lack other factors that affect tRNA amounts (e.g., regulation, proximity to the origin of replication, leading verses lagging strand, individualized aminoacyl-synthase turnovers), ME-models account for genome architecture (gathered from publicly available databases), transcription, tRNA charging, and translation (validated through a combination of 'omics and Northern blot data), which allowed us the ability to interrogate the importance of tRNA operon structure for two organisms, *E. coli* and *C. ljungdahlii*. Examination of these two organisms' operon structures revealed two different strategies: Fragmentation in *E. coli* and modularization in *C. ljungdahlii*. Using iLE1678-ME (*E. coli*) and iJL965-ME (*C. ljungdahlii*) as a basis, 1000 models with randomly shuffled tRNA locations for each organism were built. Predictions from these MC tRNA location models compared to those from iLE1678-ME or iJL965-ME showed that tRNA operon structure was optimized for tRNA abundance requirements. In iLE1678-ME, the tRNA operon structure also leads to high rRNA expression, while in iJL965-ME, tRNA efficiency was optimized. These

conclusions regarding optimization primarily hold strong for batch growth conditions, which implies that tRNA operon structure is a nonrandom result of selective pressures for maximizing growth rate.

SUMMMARY

Successful, scale-able implementation of biofuels is dependent on the efficient and near complete utilization of diverse biomass sources. Lignocellulosic biomass holds great promise for achieving renewable fuel standards set forth by the US and EU. However, a major limitation in the production of biofuels from lignocellulosic biomass is conversion of the lignin fraction. A promising approach to utilize this recalcitrant biomass (or any organic waste stream) is through thermochemical conversion of organic compounds to syngas, a mixture of CO, CO_2, and H_2. Subsequently, syngas can be metabolized by acetogenic microorganisms and converted to multi-carbon organics such as acetate, ethanol, butanol, butyrate, and 2,3-butanediol. Acetogens are comprised of a physiologically diverse panel of organisms. In addition to being able to ferment a variety of organic molecules, acetogens offer several attractive metabolic features absent in model microorganisms currently used for biofuel production, such as *Escherichia coli* and yeast. Chief among these is their ability to reduce CO_2 as an electron acceptor via the Wood-Ljungdahl pathway to produce multi-carbon organic molecules. This capacity for CO_2-reduction makes it feasible to achieve near stoichiometric conversion of biomass to desired organic products and fuel molecules via the recovery of low-potential electrons. Production of biofuels with acetogens has been stymied so far by a poor understanding of their metabolic, energetic, and regulatory networks that govern its physiology. Thus, desirable physiological properties of acetogens for biofuel production cannot, at present, be introduced into *E. coli* or other chassis organisms. Next-generation omics approaches, e.g. RNA-seq, ChIP-exo, and Ribosome profiling, enable researchers to rapidly decipher genome architecture and deeply characterize organisms. Generation of such data in acetogens is not only beneficial to better understanding these microorganisms but also is the basis upon which systems level analysis can be performed. Recently, the development of genetic manipulation tools for the acetogen *Clostridium ljungdahlii* has opened the window for establishing this species as a new chassis for biofuel production.

Computational modeling is a prerequisite for rational genome-scale engineering for biofuel production. In addition to genome-scale models of metabolism, next generation models, so called ME-models, have recently been developed, expanding the scope of models to include major cellular processes such as macromolecular synthesis and transcriptional regulation. These next-generation models enable engineering of multiple cellular processes resulting in the advanced design of tunable systems for bioproduction.

In work performed under the contract contract DE-SC0012586 we elucidated how the model acetogen *Clostridium ljungdahlii* regulates energy and carbon metabolism not only at a transcriptional level but that major pathways related to energy conservation and product formation are controlled at a translational level. Furthermore, we identified major genomic features that control carbon and energy flow in different growth conditions this Gram-positive bacterium. While postranscriptional control of gene expression plays a major role in gene regulation, we currently lack a deeper understanding of posttranscriptional control of gene expression in industrial relevant bacteria. Using a multi-omics approach and detailed analysis of the mRNA 5' untranslated regions, we discovered that translational efficiency is not only affected by the strength of RBS, but also depends on the AU content upstream of RBA and the distance of the RBS from the translation start site. These findings provide novel mechanistic insights into posttranscriptional regulation, resulting in differential translational efficiency (TE) in a growth-dependent manner. Our work uncovers a novel regulatory mechanism for the model acetogen *C. ljundahlii* that thrives at the energetic limit of life and highlights utilization of scarce resources at optimal efficiency. We propose that energy and carbon metabolism pathways are specifically controlled at the TE level, allowing for dynamic resource allocation. Our findings have broad implications on how microorganisms control and optimize their metabolic networks. The results provide a new framework for metabolic regulation in this model acetogen that can readily be extrapolated to other industrially important microbes, and will thus lay the foundation for advanced strain design and engineering efforts.

To use this new knowledge for a system biology-based design, we developed a novel genome-scale model of metabolism and macromolecular synthesis is deployed to gain new insights into the biology of the model acetogen *C. ljungdahlii.* Metabolic and gene expression models (ME-models) include more than metabolic reactions; they also contain representations of major cellular processes like macromolecular synthesis and basic transcriptional regulation, which significantly broadens the scope and predictability of microbial systems biology. The model of *C. ljungdahlii* reconstructed represents the first ME-model of a gram-positive bacterium and captures all major central metabolic, amino acid, nucleotide, lipid, major cofactors, and vitamin synthesis pathways as well as pathways to synthesis RNA and protein molecules necessary to catalyze these reactions. The model was used to reveal how protein allocation and media composition influence metabolic pathways and energy conservation in acetogens. While

standard metabolic models only predict formation of a single product, the ME-model allows for the first time to accurately predict secretion of acetate, ethanol, and glycerol during changing carbon and metal availability. Predicting overflow metabolism is of particular interest since it enables new design strategies, e.g. the formation of glycerol by *C. ljungdahlii* (which was experimentally confirmed) had not been described and describes new metabolic capability of this microbe. Furthermore, prediction and experimental validation of changing secretion rates based on metal availability opens the window into fermentation optimization and provides new knowledge about the proteome utilization and carbon flux in acetogens.

Lastly, we investigated the effect of the operon structure of *C. lungdahlii* on its growth phenotype, using the newly reconstructed ME-model. An operon is a co-regulated cluster of genes that are expressed on the same RNA transcript. Though the presence of an operon may be a random event, selection pressures can drive the maintenance of operons. In order to optimize cellular efficiency, translation must be carefully controlled because it requires the highest energy and resource expenditure of any process in fast-growing cells. Using the COBRAme-based ME-model for *C. ljungdhalii,* we examined the systematic importance of tRNA co-expression. We validated the two models for the purposes of this study and examined the tRNA operon structures, thereby identifying unique tRNA operon solutions to two different selective pressures. One solution led to optimization of phenotype through fragmenting operons and the other solution to optimized efficiency through optimal grouping of tRNAs.

This study substantially enhanced our knowledge about chemolithoautotrophs and their potential for advanced biofuel production. It provides next-generation modeling capability, offer innovative tools for genome-scale engineering, and provide novel methods to utilize next-generation models for the design of tunable systems that produce commodity chemicals from inexpensive sources.

ACKOWLEDGEMENT

This report is based upon work performed in the laboratory of Prof. Karsten Zengler (Department of Pediatrics, UC San Diego) by Mahmoud M. Al-Bassam, Joanne Liu, Ji-Nu Kim, and Livia Zaramela, the laboratory of Prof. Bernhard Palsson (Department of Bioengineering, UC San Diego) by Colton Lloyd, Ali Ebrahim, and Connor Olson, and Prof. Nathan Lewis (Department of Pediatrics, UC San Diego). We also like to acknowledge members of the Zengler lab (Cristal Zungia, Nafeesa Khan, Clifford Wright, Katrina Hung, Nien-Chen Wang, Daniela Galzerani, Harish Nagarajan, Haythem Latif, Nathan Chapin, Janna Tarasova, Cameron Martino, Kristine Ly, Julian Alberni, and Catherine Frusetta) for technical assistance.

BIBLIOGRAPHY

Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).

Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43,** W39–W49 (2015).

Becker, S. A. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Microbiol.* **5,** 8 (2005).

Caspi, R. et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res **36**, D623-631, (2008).

Crooks, G. E. *et al.* WebLogo: A Sequence Logo Generator. *Genome Res.* **14,** 1188–1190 (2004).

Dash, S., Mueller, T. J., Venkataramanan, K. P., Papoutsakis, E. T. & Maranas, C. D. Capturing the response of Clostridium acetobutylicumto chemical stressors using a regulated genome-scale metabolic model. *Biotechnol. Biofuels* **7,** 144 (2014).

dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res **32**, 5036-5044, (2004).

Duval, M. *et al. Escherichia coli* Ribosomal Protein S1 Unfolds Structured mRNAs Onto the Ribosome for Active Translation Initiation. *PLoS Biol.* **11,** (2013).

Feist, A. M. & Palsson, B. O. The biomass objective function. *Curr. Opin. Microbiol.* **13,** 344–349 (2010).

Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology* **5,** 827–835 (2004).

Hajnsdorf, E. & Boni, I. V. Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie* **94,** 1544–1553 (2012).

Heap, J. T., Pennington, O. J., Cartman, S. T., Carter, G. P. & Minton, N. P. The ClosTron: A universal gene knock-out system for the genus Clostridium. *J. Microbiol. Methods* **70,** 452–464 (2007).

Huang, H. *et al.* RISPR/Cas9-Based Efficient Genome Editing in *Clostridium ljungdahlii*, an Autotrophic Gas-Fermenting Bacterium. *ACS Synth. Biol.* **5,** 1355–1361 (2016).

Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8,** (2014).

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).

Jeong, Y. *et al.* The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). *Nat. Commun.* **7,** 11605 (2016).

Jones, S. W. *et al.* CO2 fixation by anaerobic non-photosynthetic mixotrophy for improved carbon conversion. *Nat. Commun.* **7,** 12800 (2016).

Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**, 143-155 (1999).

Komarova, A. V., Tchufistova, L. S., Dreyfus, M. & Boni, I. V. AU-rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J. Bacteriol.* **187,** 1344–1349 (2005).

Kopke, M. *et al. Clostridium ljungdahlii* represents a microbial production platform based on syngas. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 13087–92 (2010).

Kurland, C. G. Major codon preference: theme and variations. Biochem Soc Trans **21**, 841-846 (1993).

Latif, H. *et al.* A streamlined ribosome profiling protocol for the characterization of microorganisms. *Biotechniques* **58,** 329–32 (2015).

Latif, H., Zeidan, A. a., Nielsen, A. T. & Zengler, K. Trash to treasure: Production of biofuels and commodity chemicals via syngas fermenting microorganisms. *Curr. Opin. Biotechnol.* **27,** 79–87 (2014).

Lerman, J. A. et al. *In silico* method for modelling metabolism and gene product expression at

genome scale. Nat Commun 3, 929,(2012).

Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484,** 538–541 (2012).

Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

Ljungdahl, L. G. A life with acetogens, thermophiles, and cellulolytic anaerobes. *Annu. Rev. Microbiol.* **63,** 1–25 (2009).

Lloyd, C. J. *et al.* COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. *bioRxiv* (2017).

Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6,** 26 (2011).

Love, M. I., Anders, S. & Huber, W. Differential analysis of count data - the DESeq2 package. *Genome Biol.* **15,** (2014).

Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: A database for prokaryotic operons. *Nucleic Acids Res.* **37,** (2009).

McCarthy, J. E. G. & Gualerzi, C. Translational control of prokaryotic gene expression. *Trends Genet.* **6,** 78–85 (1990).

McDonald, M. J., Chou, C. H., Swamy, K. B., Huang, H. D. & Leu, J. Y. The evolutionary dynamics of tRNA-gene copy number and codon-use in *E. coli*. BMC Evol Biol **15**, 163, (2015).

McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24,** 422–430 (2014).

Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol.* **3,** REVIEWS0004 (2002).

Mock, J. *et al.* Energy conservation associated with ethanol formation from H_2 and CO_2 in *Clostridium autoethanogenum* involving electron bifurcation. *J. Bacteriol.* **197,** 2965–2980 (2015).

Nagarajan, H. *et al.* Characterizing acetogenic metabolism using a genome-scale metabolic

reconstruction of *Clostridium ljungdahlii*. *Microb. Cell Fact.* **12,** 118 (2013).

Nakagawa, S., Niimura, Y., Miura, K. -i. & Gojobori, T. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc. Natl. Acad. Sci.* **107,** 6382–6387 (2010).

Nunez, P. A., Romero, H., Farber, M. D. & Rocha, E. P. Natural selection for operons depends on genome size. Genome Biol Evol **5**, 2242-2254, (2013).

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9,** 693 (2014).

Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42,** (2014).

Placzek, S. *et al.* BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45,** D380–D388 (2017).

Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15**, 809-819, (2005).

Ray, J. C. & Igoshin, O. A. Interplay of gene expression noise and ultrasensitive dynamics affects bacterial operon organization. *PLoS Comput Biol* **8**, e1002672, (2012).

Saxena, J. & Tanner, R. S. Effect of trace metals on ethanol production from synthesis gas by the ethanologenic acetogen, *Clostridium ragsdalei*. *J. Ind. Microbiol. Biotechnol.* **38,** 513–521 (2011).

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* 330, 1099-1102, (2010).

Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

Sharp, P. M. & Li, W. H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).

Singh, N. & Wade, J. T. Identification of regulatory RNA in bacterial genomes by genome-scale mapping of transcription start sites. *Methods Mol. Biol.* **1103,** 1–10 (2014).

Subramanian, A. R. Structure and Functions of Ribosomal Protein S1. *Prog. Nucleic Acid Res. Mol. Biol.* **28,** 101–142 (1983).

Tan, Y., Liu, Z.-Y., Liu, Z. & Li, F.-L. Characterization of an acetoin reductase/2,3-butanediol dehydrogenase from *Clostridium ljungdahlii* DSM 13528. *Enzyme Microb. Technol.* **79–80,** 1–7 (2015).

Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).

Thiele, I., Jamshidi, N., Fleming, R. M. T., Palsson, B. O. & Stothard, P. Genome-Scale Reconstruction of *Escherichia coli's* Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization. *PLoS Comput. Biol.* **5,** e1000312 (2009).

Tremblay, P. L., Zhang, T., Dar, S. A., Leang, C. & Lovley, D. R. The Rnf complex of *Clostridium ljungdahlii* is a proton-translocating ferredoxin:NAD+ oxidoreductase essential for autotrophic growth. *mBio* **4,** (2012).

Valgepea, K. *et al.* Arginine deiminase pathway provides ATP and boosts growth of the gas-fermenting acetogen *Clostridium autoethanogenum*. *Metab. Eng.* **41,** 202–211 (2017a).

Valgepea, K. *et al.* Maintenance of ATP Homeostasis Triggers Metabolic Shifts in Gas-Fermenting Acetogens. *Cell Syst.* **4,** 505–515.e5 (2017b).

Wade, J. T. & Grainger, D. C. Pervasive transcription: Illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* **12,** 647–653 (2014).

Wald, N. & Margalit, H. Auxiliary tRNAs: large-scale analysis of tRNA genes reveals patterns of tRNA repertoire dynamics. Nucleic Acids Res **42**, 6552-6566, (2014).

Zaslaver, A., Mayo, A., Ronen, M. & Alon, U. Optimal gene partition into operons correlates with gene functional order. Phys Biol **3**, 183-189, (2006).