Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Investigation of iModulons in Clostridium thermocellum to characterize condition-specific changes in co-regulated genes

Sanjeev Dahal
The Pennsylvania State University
Costas D. Maranas
costas@psu.edu

The Pennsylvania State University

Research Article

Keywords: Clostridium thermocellum, Acetivibrio thermocellus, Machine learning, Transcriptional regulatory network (TRN), Independent component analysis, lignocellulose, consolidated bioprocessing, redox balance, iModulons

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11542608/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Investigation of iModulons in *Clostridium thermocellum* to characterize condition-specific changes in co-regulated genes

Sanjeev Dahal[a,b], Costas D. Maranas[a,b]

[a] Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, 16802, USA

[b] DOE Center for Advanced Bioenergy and Bioproducts Innovation, The Pennsylvania State University, University Park, PA, 16802, USA

Abstract

Clostridium thermocellum has been extensively studied for its capacity to deconstruct lignocellulose and convert the resulting sugars into ethanol; however, incomplete knowledge of its transcriptional regulatory network (TRN) limits rational strain engineering. Here, we applied independent component analysis (ICA), an unsupervised machine learning approach to a compendium of 223 RNA-seq samples spanning 10 independent studies to systematically resolve transcriptional regulatory structure. The analysis identified 67 independently modulated gene sets called iModulons (IMs) that explain 80.4% of transcriptomic variance to capture regulatory programs. Next, statistical analysis and genome-scale metabolic modeling uncovered a pathway for crosstalk between sulfate assimilation and ethanol production mediated by intracellular redox balance. Finally, comparative analysis across carbon substrates such as cellulose, xylose, and xylan revealed substrate-specific regulatory shifts and candidate control points for improving lignocellulosic sugar utilization. These results provide a systems-wide framework for linking transcriptional regulation to metabolic function and identify actionable targets for engineering *C. thermocellum* for desired phenotype.

Keywords

Clostridium thermocellum; *Acetivibrio thermocellus*; Machine learning; Transcriptional regulatory network (TRN); Independent component analysis; lignocellulose; consolidated bioprocessing; redox balance; iModulons

1. Introduction

Clostridium thermocellum is a promising organism for consolidated bioprocessing (CBP) because of its high cellulase activity and endogenous pathways for the production of high-value chemicals such as ethanol [2-5]. Therefore, substantial research has been dedicated to elucidating metabolic engineering strategies to improve the production of desirable compounds [6-10]. Consequently, more investigations are being conducted to comprehend mechanisms of tolerance to the accumulation of these compounds [11-13].

While conventional metabolic engineering strategies have been effective in improving overall ethanol titers, resolving the architecture of transcriptional regulatory networks (TRN) provides insights into how an organism dynamically adapts to genetic and environmental perturbations, thereby unlocking new approaches for strain design [14-16]. However, bottom-up approaches to building TRNs require substantial time and resources, and often such networks are not predictive of gene expression [15, 17]. Top-down approaches that leverage large expression datasets and gene module detection to build TRNs can also be applied, out of which independent component analysis (ICA) has been demonstrated to outperform other similar algorithms [18].

ICA is a machine-learning approach that applies an unbiased signal separation technique to quickly reconstruct and characterize the TRN for organisms of interest, with special application to understudied organisms such as *C. thermocellum* [15, 17, 19]. ICA transforms gene expression profiles into a linear combination of independent set of genes called iModulons (IMs), for

biologically relevant understanding of the original transcriptome data [20]. IMs represent modules under the control of same underlying biological signal including transcriptional factors (TFs) or other regulatory elements or genetic perturbations [21]. The iModulons can be considered data-driven analogs of regulons as they can recapitulate known regulon structure [19]. In addition to the reconstruction of regulatory networks, this approach also measures the activity states of the IMs to quantitatively examine condition-dependent upregulation or downregulation of genetic modules [17].

ICA has been applied to various prokaryotic (e.g., *Escherichia coli* [15], *Bacillus subtilis* [22], *Pseudomonas putida* [21], *Pseudomonas aeruginosa* [23, 24], *Mycobacterium tuberculosis* [17], and *Vibrio natriegens* [16]) and eukaryotic systems (breast [20] and bladder cancer [25]). IMs have been implemented to expand regulatory roles of TFs [19, 21], elucidate shifts in condition-dependent transcriptome activity [17, 23], learn response of genetic modules to various stressors [15, 24], identify regulatory strategies to allocate cellular resources [21], and detect novel gene functions [16, 24]. IMs also find application in enhancing metabolic engineering strategies including reduction of acetate overflow to improve growth [16], cross-species engraftment of desirable pathways [26, 27], repurposing of genetic modules for targeted outcomes and limiting of cellular resources to alleviate trade-offs associated with traditional strain engineering approaches [27].

In this study, we applied ICA on a dataset of 223 RNA-Seq samples, labeled CT-compendium to obtain 67 IMs that explain 80.4% of variance in the compendium. We investigated the IMs in *C. thermocellum*, with specific focus on targeted phenotypes such as ethanol bioproduction and lignocellulose deconstruction. We determined the IMs associated with ethanol tolerance and production in the organism and applied flux balance analysis (FBA) on a genome-

scale model of the organism to demonstrate the crucial role of redox metabolism in mediating the interaction between sulfate assimilation and ethanol production in *C. thermocellum* to propose novel engineering strategies. Moreover, we also researched the regulatory networks involved in the utilization of various carbon sources including cellulose and pentose sugars such as xylose and xylan to unravel mechanisms associated with deconstruction of lignocellulosic substrates.

2. Materials and Methods

2.1 RNA-Seq processing, quality control, data normalization, and generation of high-quality expression compendium

RNA-Seq data was acquired from NCBI [28] SRA (Sequence Read Archive) for *Clostridium thermocellum* (Figure 1A, Supplementary Fig. 1) using the iModulonMiner workflow [29]. The workflow was updated to allow completion of download, processing, and quality control (QC) steps for large RNA-Seq datasets. The Nextflow pipeline [30] was run on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. The updated workflow is provided in a separate Github repository (https://github.com/maranasgroup/iModulonMiner_CT). Briefly, raw FASTQ files were downloaded using *fastq-dl* (https://github.com/rpetit3/fastq-dl) followed by read trimming using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and quality control on trimmed reads through FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were then aligned to *Clostridium thermocellum* DSM 1313 genome (download from NCBI database using accession CP002416.1 on March 4, 2026) using Bowtie [31]. The RSeQC [32] was applied to infer read direction before read counts were generated by featureCounts tool [33]. All the quality control

metrics were compiled using MultiQC [34] followed by generation of expression dataset, reported in log-transformed Transcripts per million (log-TPM) units.

Following this, samples that did not meet certain FASTQC metrics were discarded, which are outlined in a previous study [29]. Next, metadata was curated for the remaining samples in a semi-automated manner using custom Python scripts curating data from both NCBI database [28] and Joint Genome Institute Genomes Online Database (JGI GOLD) [35]. The information includes growth rates, growth phase, media, substrates, supplements, temperature, pH, and other experiment relevant information and allowed characterization of the iModulons. Samples with poor metadata quality, single replicates, and in-between replicate Pearson correlation below 0.95 were discarded to ensure high-quality samples for further processing (Supplementary Fig. 2-3). The curation steps are outlined in Step 3 of the modified workflow (https://github.com/maranasgroup/iModulonMiner_CT).

Sample-source and study-based projects and experimental conditions within each project were selected based on the differences in experimental design, study source, and biological context. A reference condition was then designated for each project, and samples within a project were normalized relative to the corresponding reference to mitigate batch effects associated with data collection across diverse sources. This normalization was intended to reduce technical variation and facilitate the identification of independent components reflecting true biological signals.

2.2 Computation of optimal number of robust independent components

Independent component analysis (ICA) was performed using FastICA (scikit-learn v1.0.2, [36]), implemented within the iModulonMiner workflow [29]. In ICA, the preprocessed gene-expression

matrix, X, with genes as rows and samples as columns, is decomposed as X≈MA, where M matrix contains component-specific gene weights and A matrix contains the corresponding activities across samples (Figure 1A).

The ICA was run several times with random seeds, and to prevent any indefinite wait caused by potential convergence issues, a timeout mechanism was enforced. The resulting independent components (IC) were clustered using DBSCAN (scikit-learn v1.0.2, [37]) to find robust independent components (ICs), using maximal distance threshold of 0.1 and a minimum cluster size of half the number of FastICA runs. To account for identical components with opposite signs, a distance metric was used to compute distance matrix:

$$d_{x,y} = 1 - ||\rho_{x,y}||$$

where $\rho_{x,y}$ is pearson correlation between components x and y.

The final robust ICs are the centroids of the clusters and are used to define iModulons (IMs).

The above procedure was performed multiple times from 30 to 100 dimensions with step size of 5 to determine the number of robust components. To find the optimal dimensionality, the number of independent components (ICs) with single genes was compared to the number of ICs that are correlated (Pearson correlation>0.7) with the final components (or ICs in the largest possible dimension). The optimal dimensionality is the dimension at which the number of final components overtakes the number of non-single-gene components (Supplementary Fig. 4). The robust ICs at optimal dimensionality were retained for IM identification.

The gene membership in each IM was determined by applying a component-specific threshold to the absolute gene weights in M matrix. The thresholds were calculated using D'Agostino's K^2 test for normality [38]. Genes whose absolute weights exceeded the computed threshold were assigned to the corresponding IM.

2.3 Characterization of the iModulons (IMs)

Before performing IM characterization, gene annotations were retrieved from multiple sources including NCBI [28], Uniprot [39], and BioCyc [40]. BioCyc was also the source for operon information. Cluster of Orthologous Groups (COG), KEGG orthologs (KO), and PFAM information were obtained from results acquired from eggNOG-mapper [41] run on Galaxy server [42], whereas gene ontology (GO) was derived from Uniprot database. A known TRN was derived from literature sources including a previously published DAP-Seq study [14].

IM characterization and annotation was performed in a semi-automated manner using tools within PyModulon (v. 0.2.1, [29]) library. First, an *Icadata* object was generated with parameter *optimize_cutoff* turned on. Then, multiple enrichment analyses with both known TRNs and gene annotations using Fisher Exact test (*scipy* v. 1.13.1) were executed to identify IMs enriched in genes found in known regulons or functions or pathways. Based on this information, IMs were manually labeled to the name of the regulon or pathways or functions and assigned categories "regulatory" or "functional," respectively. We also computed orthologous iModulons by using *compare_ica* module within PyModulon library by comparing the IMs of *C. thermocellum* with those of selected five organisms—*Pseudomonas aeruginosa*, *Escherichia coli*, *Streptococcus pyogenes*, *Staphylococcus aureus*, and *Bacillus subtilis*. Based on Pearson correlation ($r>0.5$), IMs in *C. thermocellum* were renamed, and assigned to category "ortho_ica". Lastly, *metadata_boxplot* module that utilizes metadata and employs a decision tree regressor to automatically cluster IM activities was applied to annotate yet uncharacterized IMs and assigned to category "activity_boxplot". Any single gene IM ("single_gene" category) were labeled based on their experimental condition (e.g., hpt KO for gene *hpt* knockout) or highest gene weights within the

IM (e.g., SG-0124 IM for highest weight gene member Clo1313_0124). Finally, all remaining IMs were assigned the prefix UC and classified as "uncharacterized."

2.4 Statistical analyses

To create a preliminary list of IMs associated with ethanol concentrations, spearman correlation analysis (*scipy* v. 1.13.1) was performed between IM activities and ethanol concentration due to a small number of conditions (N) for projects *eh_chemostat* and *evo_ag553_ag601* (Table 1). Any IM that passed the two criteria— Spearman $|r| \geq 0.6$, p-value<0.05, were considered significant ones. Multiple test correction using Benjamini-Hochberg method (*statsmodels* v. 0.14.5) was also implemented but the metric was only used as a directional statistic for subsequent analyses. The significant IMs were then modeled individually or in all possible combinations of two, three, four, and five IMs applying multiple regression modeling approach using ordinary least squares (OLS, *statsmodels* v. 0.14.5). Furthermore, each model was also cross-validated (CV) by leaving-one-sample out and training the model on remaining samples (N-1) before predicting the outcome for the left-out sample repeatedly for N times to create a vector of predicted values (YP_N). The vector YT_N was the vector of true values. Then, coefficient of determination (R^2, *sklearn* v. 1.0.2) was computed between YT_N and YP_N. The models were sorted based on their Akaike Information Criterion (AIC) scores with lowest scores on top. Furthermore, to account for multicollinearity of the IMs, variance inflation factor (VIF, *statsmodels* v. 0.14.5) was computed for the IMs associated with the model under consideration. A high maximum VIF in the model indicated higher likelihood of multicollinearity. The models with low AIC scores, low maximum VIF and high CV R^2 were considered for further investigation.

To compare two IMs or genes using large number of samples (n≥30), Pearson correlation (*scipy* v. 1.13.1) statistic with $p<0.05$ as a test of significance was used. However, for a small number of samples (n<30), Spearman correlation (*scipy* v. 1.13.1) was preferred.

For computing enrichment of genes from one set of genes to another, Fisher exact test was performed from PyModulon library.

2.5 Genome-scale modeling and flux balance analysis

A genome-scale metabolic model (GSM) for *C. thermocellum* [43] was used for flux balance analysis (FBA). To constrain the model, substrate and product concentrations (C_i, g/L) of a metabolite *i* were converted to flux (v_i, mmol/(gDW.hr)) using the formula:

$$v_i = (C_i \times D) / (X \times MW_i) \times 1000$$

where D=dilution rate (hr^{-1}), X=biomass concentration (gDW/L), MW_i=molecular weight of metabolite M (g/mol)

For engineered and evolved strains, gene knockouts were simulated by fixing the reactions catalyzed by the encoding proteins to zero. Furthermore, growth rate was allowed to deviate between $0.95D$ and $1.05D$. Likewise, all the substrate and product rates were allowed between $0.9v_i$ to $1.1v_i$ except ethanol ($0.95v_i$ to $1.05v_i$) to ensure lower deviation for the targeted phenotype, i.e., ethanol production flux. Parsimonious flux balance analysis (pFBA, [44]) was implemented from *cobrapy* (v. 0.29.1) library to simulate the model after constraints were applied with *fraction_of_optimum* option set to 0.9 and objective set to biomass maximization. The sulfate fluxes were normalized by predicted growth rate to compare the normalized values across various genetic conditions.

To compute the flux-sum (Φ_i) of a metabolite i, formula provided in a previous study was used [45].

$$\Phi i = 0.5 * (\sum_1^j |S_{ij}.v_j|)$$

where S_{ij}= stoichiometric coefficient of metabolite i in reaction j and v_j = flux of reaction j

The fluxes were compared between two selected strains – reference (LL1004, wild-type) and LL1011 (*Δhpt Δldh Δpta* evolved pH auxostat) by computing the differences of the absolute fluxes of the respective reactions to identify reactions that changed the most between the two strain conditions.

2.6 Differential iModulon Activity (DiMA) analysis

To determine the IMs that were differentially active between two different conditions, DiMA analysis was performed as described previously [15]. The analysis was run using *plot_dima* function in PyModulon library. Default parameters for the module (*fdr*=0.1, *threshold*=5) were used in this study.

2.7 Data visualization

Data visualization was performed using R- and Python-based software packages. Treemap visualizations were generated in RStudio (v. 2024.12.1) using the *treemap* package (v. 2.4-4). Input data—including IM names, variance explained, and assigned IM category—were extracted from the *IcaData* object and formatted to align with the data requirements of the package.

DiMA plots were generated using the *plot_dima* function implemented in PyModulon. Comparisons of IM activities and gene-expression levels across experimental conditions were visualized using the *plot_activities* and *plot_expression* functions, respectively, from PyModulon.

Gene-weight distributions were visualized using the PyModulon's *plot_gene_weights* function. These figures were subsequently customized using *matplotlib* (v. 3.10.0) package.

Heatmaps of IM activities were constructed using *seaborn* (v. 0.13.2) package. Pathway maps were created and annotated in Microsoft PowerPoint for Mac (v. 16.113.3). All remaining visualizations, including custom bar plots, were generated using *matplotlib* (v. 3.10.0) package.

3. Results

We begin by implementation of ICA to the CT-compendium, a dataset comprising 223 RNA-seq samples to identify 67 IMs that collectively explain 80.4% of variance in the compendium. Next, the IMs were used to investigate phenotypes of interest including ethanol bioproduction and lignocellulose deconstruction. We first identified candidate IMs associated with ethanol tolerance and production and employed FBA to elucidate the critical role of cofactor balancing in coordinating the crosstalk between sulfate assimilation and ethanol biosynthesis pathways. Next, we examined the IMs related to the utilization of diverse carbon sources, including cellulose, xylose, and xylan, to uncover mechanisms underlying lignocellulosic substrate deconstruction.

3.1 CT-compendium is a high-quality RNA-Seq dataset from which sixty-seven iModulons were identified

We implemented an updated iModulonMiner workflow (https://github.com/maranasgroup/iModulonMiner_CT) to download and process RNA-Seq datasets for *Clostridium thermocellum.* 376 datasets were available in the NCBI [28] database (March 25, 2026) of which 325 samples were downloaded and processed through iModulonMiner's RNA-Seq Nextflow pipeline [29]. After first iteration of quality-filtering

(Supplementary Fig. 2), 254 samples (~78.2% of downloaded samples) remained. In a semi-automated manner, metadata for samples were retrieved using custom scripts, literature curation, and manual search of relevant databases before proceeding to second iteration of quality control in which conditions with single replicates or low replicate correlation ($r<0.95$) were removed (Supplementary Fig. 3). After these steps, ~68.6% of downloaded samples representing 53 unique conditions and 10 sets of experiments/projects remained (Figure 1B, Table 1, Supplementary Table 1). The curated dataset spans diverse experimental conditions, including two non-engineered strains (DSM 1313 and ATCC 27405), targeted gene deletions, varying pH and temperature

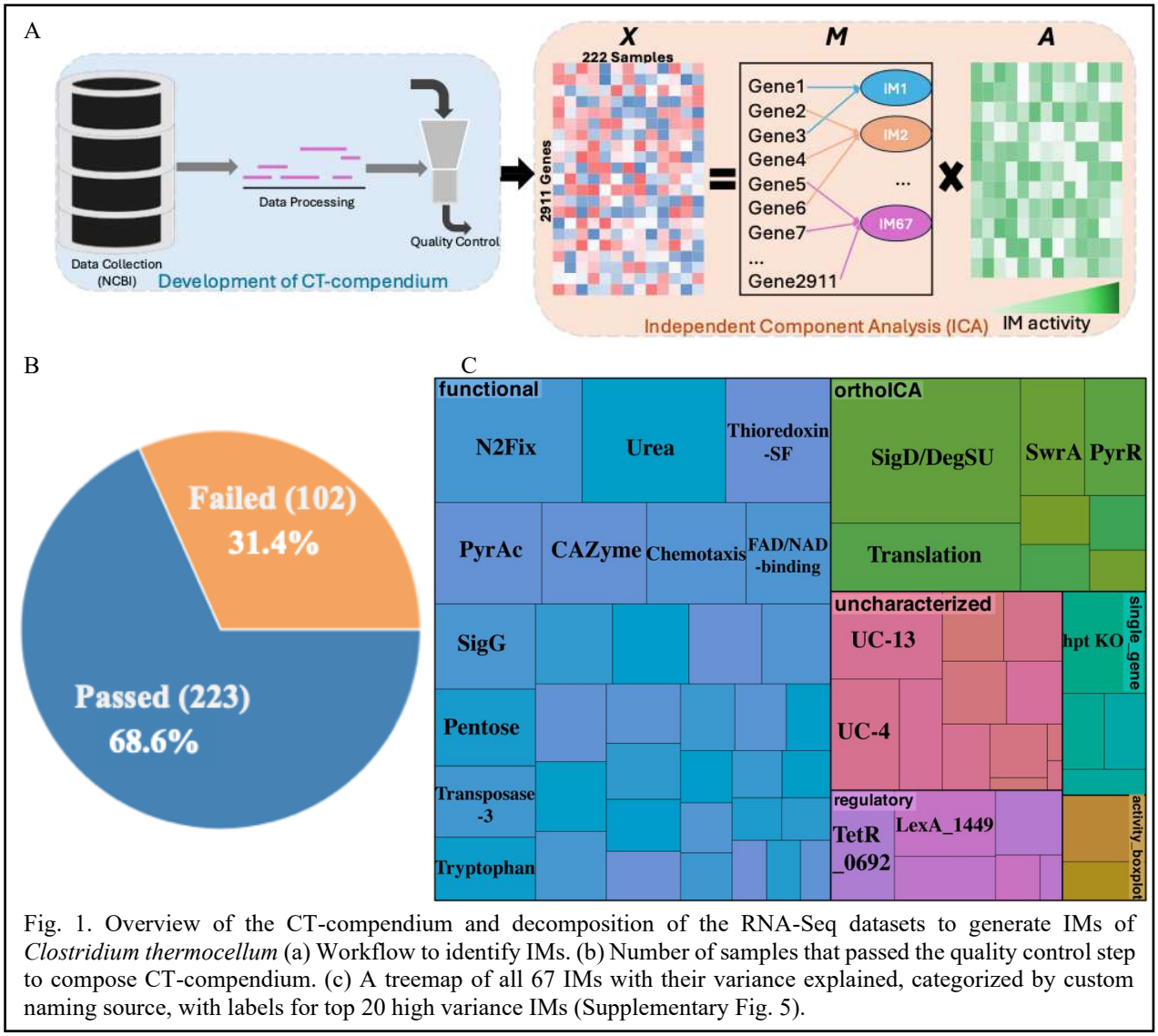


Fig. 1. Overview of the CT-compendium and decomposition of the RNA-Seq datasets to generate IMs of *Clostridium thermocellum* (a) Workflow to identify IMs. (b) Number of samples that passed the quality control step to compose CT-compendium. (c) A treemap of all 67 IMs with their variance explained, categorized by custom naming source, with labels for top 20 high variance IMs (Supplementary Fig. 5).

conditions, strains derived through adaptive laboratory evolution (ALE), different substrates with or without supplements, and planktonic and biofilm growth states (Supplementary Data 1). The curated expression profile will henceforth be called the CT-compendium (Figure 1A, Supplementary Data 2).

We next implemented independent component analysis (ICA) through FastICA (scikit-learn v1.0.2, [36] multiple times from 30 to 100 dimensions with step size of 5 within the iModulonMiner workflow [29]. With ICA, at optimal dimensionality of 90, sixty-seven independent components were identified (Supplementary Fig. 4). The gene membership in each component or iModulons (IMs) was determined by gene-weights that passed threshold computed using D'Agostino's K^2 test for normality [38]. The characterization of IMs was performed using

Table 1. Summary of projects and conditions associated with the CT-compendium. Expanded table in Supplementary Table 1.

Projects	Total Conditions	Total Samples	Submission IDs	Brief Description of Projects
biofilm	2	7	SRA480534	Expression differences in sessile and planktonic populations during growth on cellulose
cell_stressor	5	26	SRA817681	Growth at different temperatures and media conditions including cellulose
eh_chemostat	16	107	SRA1000062-SRA1000063, SRA1000066, SRA1000079, SRA1000084-SRA1000085, SRA1000090, SRA1000146, SRA610607, SRA831131, SRA831133-SRA831136 SRA831138	Expression and metabolic fluxes in unevolved and evolved engineered strains
evo_ag553_ag601	4	10	SRA357389	Physiological impact of the metabolic engineering, adaptive evolution, and metabolite supplementation
lacI	9	23	SRA264897	Compare expression data to unravel three LacI regulons
ll1210	4	12	SRA592804	Assess high ethanol strain, LL1210 at different pH
pentose_sugar_1	5	15	SRA1959963	Engineer and evolve strains for simultaneous co-utilization of pentose and hexose sugars
pentose_sugar_2	2	6	SRA962562	Engineered strains on xylose versus cellobiose
sulfate	4	11	SRA574630	Expression changes for two sulfur sources—sulfate and cysteine
xyl_evo	2	6	SRA310863	Physiological and gene expression changes due to non-fermentable pentose sugars
Total	**53**	**223**		

the procedure outlined in the Materials and Methods. Overall, the 67 IMs cumulatively explain 80.4% of variance in the expression of the CT-compendium (Figure 1C) with six IMs enriched in genes from known regulons out of nineteen (Supplementary Fig. 6). None of the six IMs passed threshold of 0.6 for both regulon and iModulon recalls, which likely suggests either 1) literature-based regulons or 2) predicted IMs reflect secondary regulatory programs. It is possible that with more data available, these IMs will become well-matched with respective regulons. All the IMs and their member genes are provided in Supplementary Data 3.

3.2 A two IM Ordinary least squares (OLS) model explains 61% variance in ethanol concentrations

As IMs represent co-regulated gene sets, they facilitate the systematic identification of gene groups associated with phenotypes of interest, including ethanol biosynthesis. We analyzed two projects—*eh_chemostat* and *evo_ag553_ag601*—comprising multiple samples with matched ethanol concentration measurements to identify iModulons associated with ethanol production. In the *eh_chemostat* dataset, knockout and evolved strains were cultivated in low-carbon (modified LCmed) medium supplemented with cellobiose under chemostat conditions (Holwerda et al., 2020). In the *evo_ag553_ag601* project, knockout and evolved strains were grown in batch culture in MTC_5 medium with or without formate supplementation (Papanek et al., 2018). First, an exploratory set of IMs significantly correlated (absolute value of Spearman $|r| \geq 0.6$, p-value<0.05) with ethanol levels were identified using spearman correlation analysis (*scipy* v. 1.13.1, Table 2). Five IMs – Sulfate, AbrB-Related, SigI6, Fe-S-Related, and Co/Ni-tp-1 (Table 2, Supplementary Data 4) were uncovered as candidates for multiple regression modeling applying ordinary least squares approach (OLS, *statsmodels* v. 0.14.5). All combinations of candidate IMs were modeled

296 to determine the best ranking model based on the lowest Akaike Information Criterion (AIC)
297 scores (Supplementary Data 5).

298 Using the OLS approach, a two-predictor model containing the Sulfate and AbrB-Related
299 IMs provided the best overall fit (adjusted R^2=0.61, F-test p<0.05) with the lowest AIC (Table 2,
300 Supplementary Data 5). Only these two were significant in the correlation analysis after accounting
301 for multiple testing corrections applying Benjamini-Hochberg procedure (q-value<0.05),
302 signifying the two-component model contains robust features. However, only the Sulfate IM was
303 statistically significant in the best scoring model. Furthermore, the leave-one-out cross validation
304 (CV) R^2 (0.51) of the model was marginally higher than that of the single predictor model
305 containing only the Sulfate IM (0.48). Together, these results suggest that the Sulfate IM is a strong
306 predictor while the AbrB-Related IM is likely a weak indicator even though it improves the in-
307 sample fit and overall AIC (Table 2).

308 The inclusion of another IM, Co/Ni-tp-1, to the best scoring two-predictor model modestly
309 increased CV R^2 (0.53). However, accompanying higher AIC (−5.95), loss of statistical
310 significance for all predictors in the three-predictor model (Supplementary Data 5), and high
311 correlation between the Co/Ni-tp-1 and Sulfate IMs (Figure 2A) implied that its addition did not
312 yield a robust, independently interpretable effect. The addition of other IMs (SigI6, Fe-S-Related)
313 did not improve fit of the best scoring model, either. An elevation in variance inflation factor (VIF)
314 score was observed upon the addition of SigI6, indicating collinearity with existing predictors
315 (Supplementary Data 5), which is further supported by high correlation with other IMs (Figure
316 2A). In the case of the Fe-S-Related IM, its univariate model itself performed poorly (adjusted R^2
317 = 0.024, LOO-CV R^2 = −0.18, F-test p = 0.25), suggesting that Fe-S-Related carries little to no
318 predictive signal. Consequently, the two-predictor model (Sulfate + AbrB-Related), based on

lowest AIC and greater parsimony is the best ranking model. Upon further investigation, both the activities of Sulfate and AbrB-Related IMs were determined to be affected by sulfur source indicating a possible interaction between the two IMs and pointing to a likely reason for AbrB-Related's weak contribution to the best model (Figure 2B).

Table 2. Comparison of the effects of different IMs on ethanol production by applying multiple linear regression model (OLS) approach following spearman correlation analysis.

IM	Spearman		OLS				
	p-val	q-val	AIC	p-val	Significant IMs	Adjusted R^2	LOO CV R^2
Sulfate	*	*	-4.26	*	1	0.54	0.48
AbrB-Related	*	*	-1.47	*	1	0.47	0.39
SigI6	*	0.08	-0.46	*	1	0.44	0.36
Co/Ni-tp-1	*	0.08	5.85	*	1	0.22	0.09
Fe-S-Related	*	0.08	9.99	0.25	0	0.02	-0.18
Sulfate + AbrB-Related	NA	NA	-6.36	*	1	0.61	0.51

* refers to statistically significant ($p<0.05$ and $q<0.05$). NA refers to OLS model as spearman correlation analysis was only performed for individual IMs.

3.3 Sulfate assimilation and hydrogenase pathways work together for redox balance

The Sulfate IM contributes the largest weight to the best OLS model. The IM consists of four genes (RS00605, *cysT*, RS00630, and *cooS*), three of which encode for a transporter (RS00605 and *cysT*) and an enzyme (RS00630-- phosphoadenosine phosphosulfate reductase) in the upper part of the sulfate assimilation pathway. Based on these genes, we inferred that the sulfate assimilation pathway is a key metabolic route that *C. thermocellum* reprograms to facilitate higher ethanol production. In the literature, high ethanol producing strains Δ*hydG* and Δ*hydG*Δ*ech* have been observed to downregulate sulfate transport and assimilation pathway [46, 47]. However, the impact of gene knockouts, relevant to the experiments used in this work, on the sulfate pathway have not been studied, to the best of our knowledge. We implemented genome-scale metabolic modeling [43] to study the impact of gene deletions on sulfate transport in the organism for all the

strains in the two aforementioned projects. Since gene expression is poorly correlated to enzyme activity [48], we opted not to apply gene expression as proxy for proteome concentrations within the genome-scale model (GSM). For the simulations, experimental growth rates and

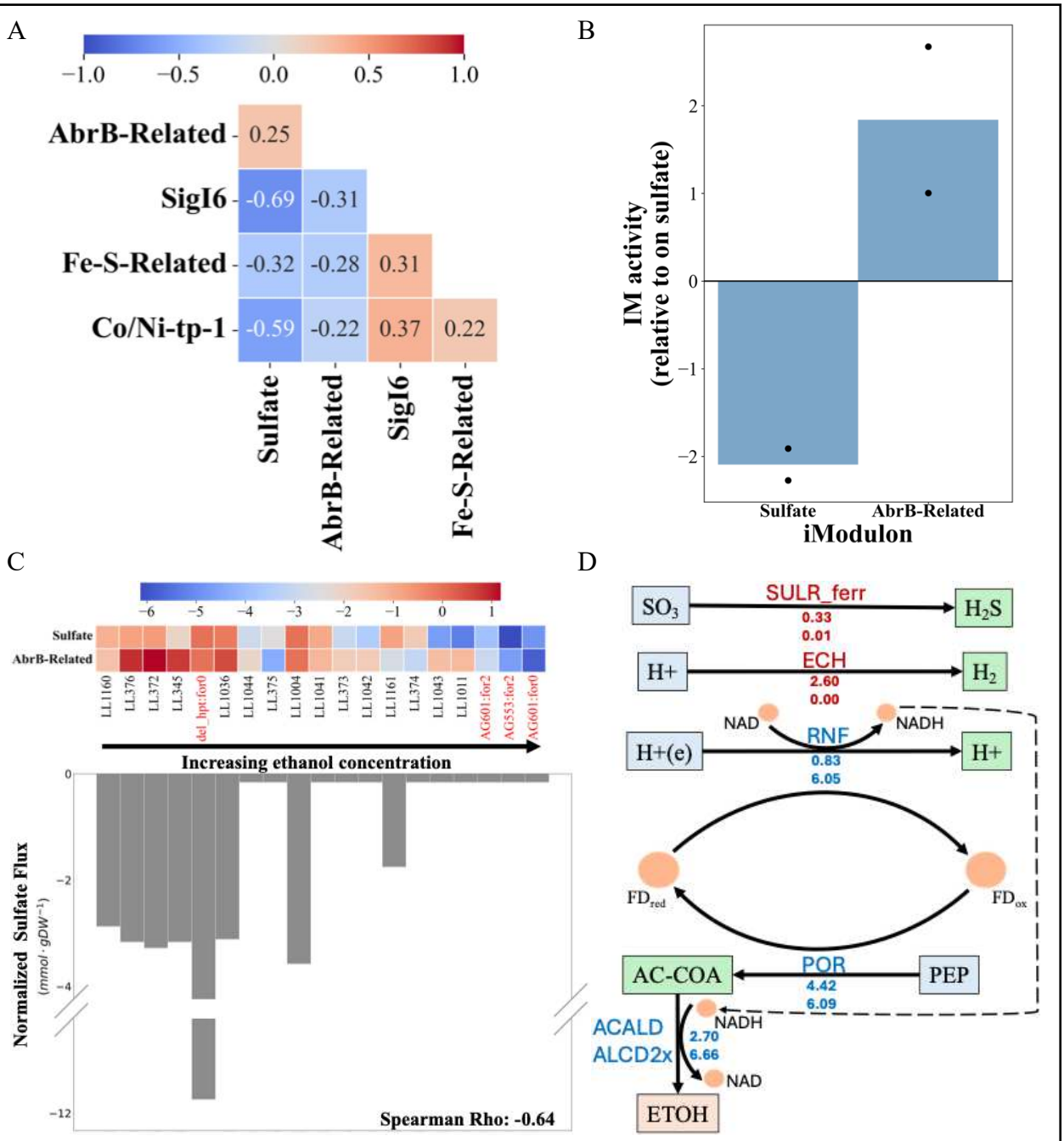


Fig. 2. Crosstalk between sulfate assimilation pathway and ethanol production. (a) Correlation matrix for IMs that are significantly correlated to ethanol levels. (b) Impact of sulfur source (cysteine) on activities of IMs of the best ranking model in the reference strains (*Δhpt*). The activity values are relative to respective IM activities for the reference on sulfate. (c) IM activity for different reference, engineered, and evolved strains producing ethanol at various concentrations (in increasing order), and comparison of predicted sulfate import fluxes across the strains. All samples are from *eh_chemostat* project except the ones in red font are from *evo_ag553_ag601* project. (d) GSM predicted relevant pathways that change between LL1004 and LL1011 to couple sulfate assimilation and ethanol production. Red and blue colors refer to lower and higher flux in LL1011, respectively. The numbers on the reaction edges are reaction fluxes for LL1004 (top) and LL1011 (bottom). Dotted line indicates the oxidation of NADH produced by RNF and other reactions during ethanol biosynthesis.

concentrations of substrates and products were used to constrain the GSM model, and any applicable gene deletions were also applied to the model. A nitrogen-limited condition in the *eh_chemostat* project was excluded from this analysis, leaving 19 conditions for assessment. We performed parsimonious FBA [44] and normalized simulated sulfate import fluxes by predicted growth rates to compare values across the strain conditions (Figure 2C). The absolute normalized sulfate fluxes were negatively correlated with predicted ethanol yield (mmol ethanol per mmol cellobiose) with Spearman r=-0.64, implying a crosstalk between sulfate assimilation and ethanol production pathways (Figure 2C).

We next compared the predicted fluxes between a high ethanol producing engineered and evolved strain LL1011 (*Δhpt Δldh Δpta* evolved pH auxostat) and reference strain LL1004 (wild-type) to explore the overall metabolic reprogramming in LL1011, leading to lower flux through sulfate pathway. The reactions with the biggest change in predicted fluxes between the two strains are in reactions ferredoxin:NAD oxidoreductase (RNF) and ferredoxin-dependent NiFe-hydrogenase (ECH), which both use ferredoxin cofactor. In LL1004, ECH is active, whereas in LL1011, RNF is preferred (Figure 2D). Overall, more ferredoxin cofactor is available in the cell in LL1011, based on flux-sum analysis [45], but the ferredoxin is channeled to RNF for the production of NADH, diverting flux away from reactions such as sulfite reductase (SULR_ferr) reaction, for which ferredoxin is required. These changes in LL1011 likely provide the cofactors-- oxidized ferredoxin and NADH for the enzymes pyruvate ferredoxin oxidoreductase (*pfor*) and acetaldehyde/ethanol dehydrogenase (*adhE*), respectively for the ultimate conversion of pyruvate to ethanol (Figure 2D).

We note that upon further investigation, we found LL1011 to contain the D494G mutation in the *adhE* gene, which expands its cofactor specificity to both NADH and NADPH. However,

studies have demonstrated that even in strains containing this specific *adhE* mutation, overexpression of *rnf* has resulted in increased ethanol concentrations, supporting the predictions in this work [49]. Furthermore, in strains containing D494G mutation, NADH-linked ADH activity does not decrease [12]. Regardless of cofactor specificity predicted by the simulations, our results demonstrate the relationship between ethanol and sulfate assimilation, such that in high ethanol yielding strains, ferredoxin is diverted away from multiple reactions including SULR_ferr to the hydrogenases to produce required cofactors—oxidized ferredoxin, NADH, and/or NADPH. Together, the metabolic reprogramming of sulfate pathway (and possibly other ferredoxin utilizing reactions) is a likely co-occurring phenomenon in higher ethanol producing strains, caused by ferredoxin balancing. Hence, to improve ethanol titers, the results from this study recommend engineering strategies to increase flux through the FNOR (ferredoxin-nicotinamide oxidoreductase) reactions and minimize usage of ferredoxin in other reactions including SULR_ferr. For instance, overexpression of *rnf* in the strains used in the *eh_chemostat* project can possibly further improve ethanol yield as observed in other strains in a previous work [49]. Likewise, overexpression of *pfor* to divert oxidized ferredoxin to increase ethanol production might also be another promising engineering strategy.

3.4 Deconstruction of cellulose is likely regulated by multiple IMs

We performed two complementary analyses to identify iModulons associated with cellulose deconstruction. First, we applied differential iModulon activity (DiMA) analysis [15] to compare IM activities between growth conditions with and without cellulose. This analysis utilized the *cell_stressor* project, in which *C. thermocellum* was cultivated at three temperatures (42°C, 50°C, and 60°C) in MTC medium with and without cellulose. Because growth data on cellulose substrate

was unavailable at the optimal temperature of 60°C, DiMA analysis was conducted independently at 42°C and 50°C, followed by identification of IMs exhibiting consistent differential activity across both conditions to remove any confounding secondary effects caused by growth at sub-optimal temperatures (Figure 3A). Second, we performed an enrichment analysis using Fisher exact test to determine IMs enriched in cellulosome and cellulase genes. For this purpose, genes associated with cellulose deconstruction were collected from a published study (Xu et al., 2016) to compare to the member genes within all sixty-seven IMs. From the DiMA analysis, two IMs—CAZyme and LexA_1449, and from the enrichment analysis, two IMs—CAZyme and SigI6, were identified to be associated with cellulose deconstruction.

The CAZyme IM is dominated by genes involved in cellulose deconstruction (Figure 3B). In enrichment analysis, its F1 score was 0.34, signifying substantial presence of cellulosome and cellulase genes in the IM with genes such as *cipA, olpB,* and *celS* as members of the IM. Within the IM, we identified only one gene in the COG category Transcription, Clo1313_3021, annotated as "accessory gene regulator B" (*agrB*), which was correlated positively with IM activity (Pearson r=0.62). The AgrB protein is a part of Agr system, a quorum sensing machinery, typically encoded in an *agrBDCA* operon, but exceptions have been studied [50, 51]. We investigated the neighboring genes of *agrB* and did not find resemblance to any partial or full *agr* operon. A study has shown that another *agrB* gene along with neighboring gene, annotated as "Radical SAM domain protein" are expressed in the presence of cellulosic substrates in *C. thermocellum* [52].

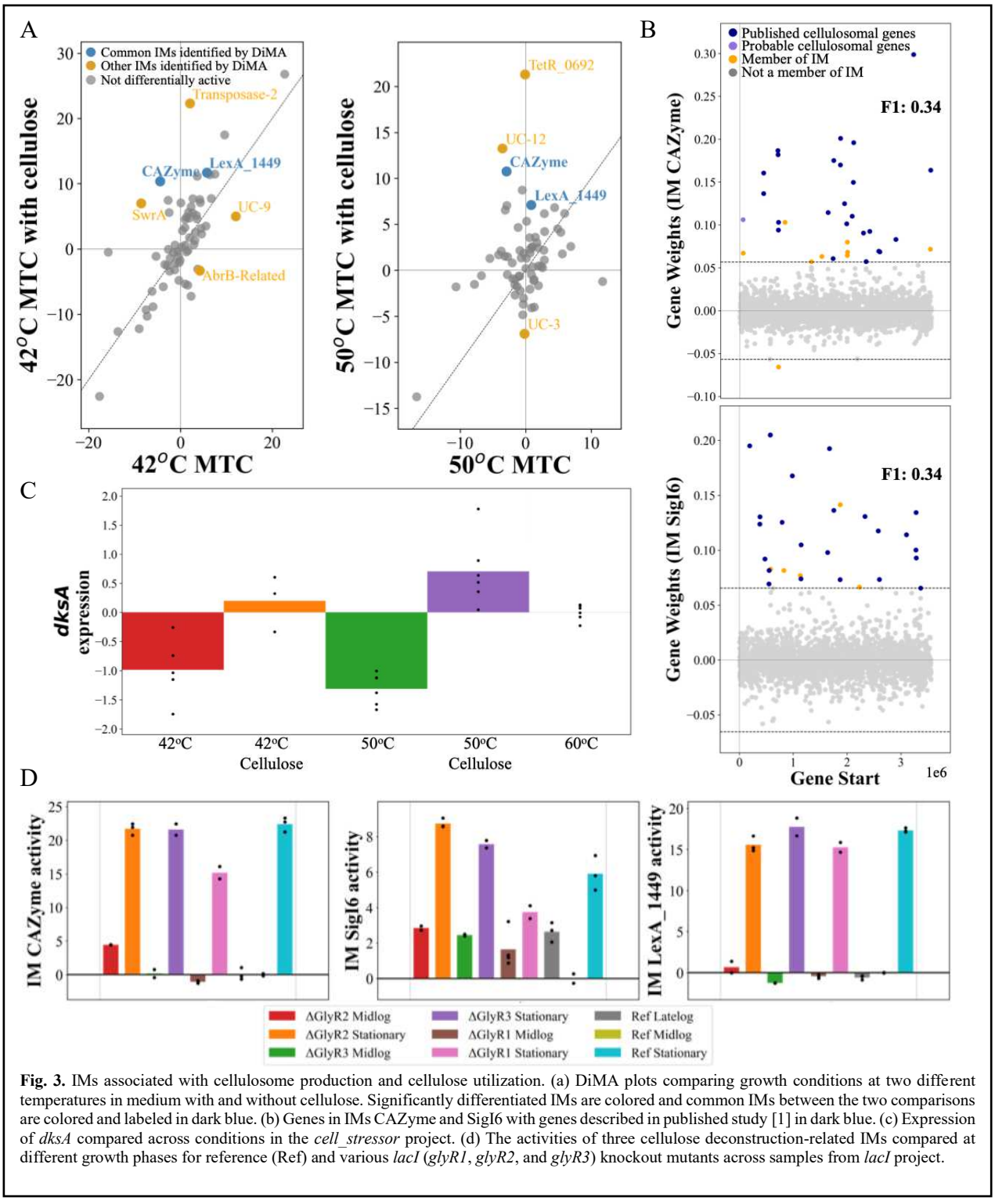


Fig. 3. IMs associated with cellulosome production and cellulose utilization. (a) DiMA plots comparing growth conditions at two different temperatures in medium with and without cellulose. Significantly differentiated IMs are colored and common IMs between the two comparisons are colored and labeled in dark blue. (b) Genes in IMs CAZyme and SigI6 with genes described in published study [1] in dark blue. (c) Expression of *dksA* compared across conditions in the *cell_stressor* project. (d) The activities of three cellulose deconstruction-related IMs compared at different growth phases for reference (Ref) and various *lacI* (*glyR1*, *glyR2*, and *glyR3*) knockout mutants across samples from *lacI* project.

Within the IM, a gene Clo1313_1704, annotated as Radical SAM domain protein is present. The gene Clo1313_1704 is correlated to *agrB* gene (Pearson r=0.58). Therefore, it is possible that these two genes act in concert to the availability of cellulose, even though they are far apart within the

genome. However, when the expression of the gene *agrB* was compared for the project, it was only active in cellulose at 50ºC, suggesting that the response of *agrB* to cellulose is temperature-dependent, and that it might not be a global regulator of cellulosome and cellulase genes (Supplementary Fig. 7).

The SigI6 IM is also considerably dominated by cellulosome and cellulase genes, with an enrichment F1 score of 0.34 (Figure 3B). The gene in COG category Transcription with the highest correlation to the IM activity is Clo1313_1605, annotated as RNA polymerase-binding protein DksA (Pearson r=0.74), and it is activated in media supplemented with cellulose (Figure 3C). DksA interacts with RNA polymerase (RNAP) directly and enhances the inhibitory effects of alarmone (ppGpp) on transcription and aids in directing cellular resources to respond to environmental cues [53]. In *E. coli*, transcription of alternative sigma factor (σ^S) and possibly σ^S-dependent genes is governed by DksA and ppGpp to induce global transcriptional response for stress alleviation. This is accomplished by freeing RNAP from occupying ribosomal DNA to make it available for transcription of σ^S-dependent genes [54]. Interestingly, the IM possesses two alternative sigma factor genes—*sigI3* (Clo1313_1911) and *sigI8* (Clo1313_0525) along with gene encoding anti-sigma factor RsgI5 (Clo1313_0985). The SigIs have been demonstrated to regulate cellulosomal genes [55-58]. The correlation in expression of *dskA* with *sigI8* is high (Pearson r=0.71) and is suggestive of co-expression of these genes. Together, we propose a crucial role of DksA as a global regulator of cellulosome and cellulase genes, which possibly acts by inducing expression of various sigma factors. Targeted genetic perturbations to enhance the expression of *dksA* or its downstream regulon may improve cellulose deconstruction.

The LexA_1449 IM appears to function in stress mitigation as it possesses several hypothetical proteins and proteins associated with stress, including two sigma-24 transcription

435 factors (Clo1313_0320 and Clo1313_0811), DNA polymerase IV (Clo1313_2742) and phage
436 shock protein (Clo1313_2160, *pspC*), with *pspC* being highly correlated to IM activity (Pearson
437 r=0.81). The gene cluster (Clo1313_2156-Clo1313_2160) within which *pspC* resides, except
438 Clo1313_2158, is part of the IM. None of the genes in the IM are directly related to cellulose
439 deconstruction. Unlike the IMs CAZyme and SigI6, LexA_1449 is more active at sub-optimal
440 temperature (42ºC) conditions regardless of substrate, implying that the IM possibly is related to
441 cellulose deconstruction through stress response (Supplementary Fig. 8).

442 The expression of cellulose deconstruction genes has been demonstrated to be dependent on
443 growth phase [59] and growth rate [60]. Therefore, changes to LexA_1449 activity could be caused
444 by differences to growth rates or growth phase, as demonstrated by elevated activity of the IM in
445 stationary phase (Figure 3D). The IMs CAZyme and SigI6 are also upregulated in stationary phase,
446 suggesting activation of cellulolytic machinery in early stationary phase or lower growth rate
447 conditions (Figure 3D), corroborating conclusions from previous studies [59, 60]. Collectively, the
448 results suggest that stress response pathways are activated concomitantly with cellulosome and
449 cellulase production during stationary phase or lower growth conditions.

450

451 **3.5 Evolved strains grow well on hemicellulosic substrates due to their adaptation to redox**
452 **stress and increased gene expression in pentose phosphate pathway**

453 We implemented DiMA analysis on samples from the project *pentose_sugar_1* in which *C.*
454 *thermocellum*, which natively cannot utilize pentose sugars such as xylose, was engineered to
455 strain KJC335 (Δ*hpt::xylAB*), that can grow on xylose. The modified strain was then subsequently
456 evolved to strain KJC19-9 (evolved on xylose) and strain KJCXn4 (evolved on xylose and xylan)

and were grown on CTFUD media containing either cellobiose or xylose with and without xylan as carbon substrates [61, 62].

When comparing IM activities of KJC335 on cellobiose versus xylose, we determined an uncharacterized IM, UC-3, as differentially repressed in xylose growth condition. The same IM was, however, differentially activated between KJC335 on xylose versus evolved strain KJC19-9 on xylose (Figure 4A). We investigated the UC-3 IM to study its importance in xylose utilization. All five genes within UC-3 were found to be differentially expressed in KJC335 between xylose and cellobiose in a previous study [61] including Clo1313_1964 (RS09935, most downregulated) and Clo1313_0528 (RS02705, most upregulated). However, the expression of these genes was found to reverse in the evolved strain KJC19-9 on xylose (Figure 4B,C). This finding corroborates and provides molecular explanation of the findings of a previous study. The growth rate of KJC335 on xylose at exponential phase is low, but that of KJC19-9 in the same condition and growth phase is close to that of non-engineered *C. thermocellum* on cellobiose, indicating overall restored metabolism (Chou et al., 2024). Between cellobiose and xylose, the mutant strain KJC335 evidently undergoes redox alterations, leading to the repression of UC-3 activity. The proposed mechanistic basis for the upregulation of stress response genes in KJC335 on xylose has been discussed in a previous study [61]. The slower growth rate in xylose in part could contribute to the changes in expression of genes [61]. In addition to their finding, we hypothesize redox state in xylose is affected by the availability of pyridoxal 5'-phosphate (PLP).

The reciprocal blast hit of gene RS09935 to *B. subtilis 168* is *pucG*, encoding (S)-ureidoglycine--glyoxylate transaminase (UGT) and RS02705 is annotated as 4Fe-4S ferredoxin. UGT requires PLP as a cofactor [63]. In *B. subtilis*, PLP is produced by deoxyxylulose-5-

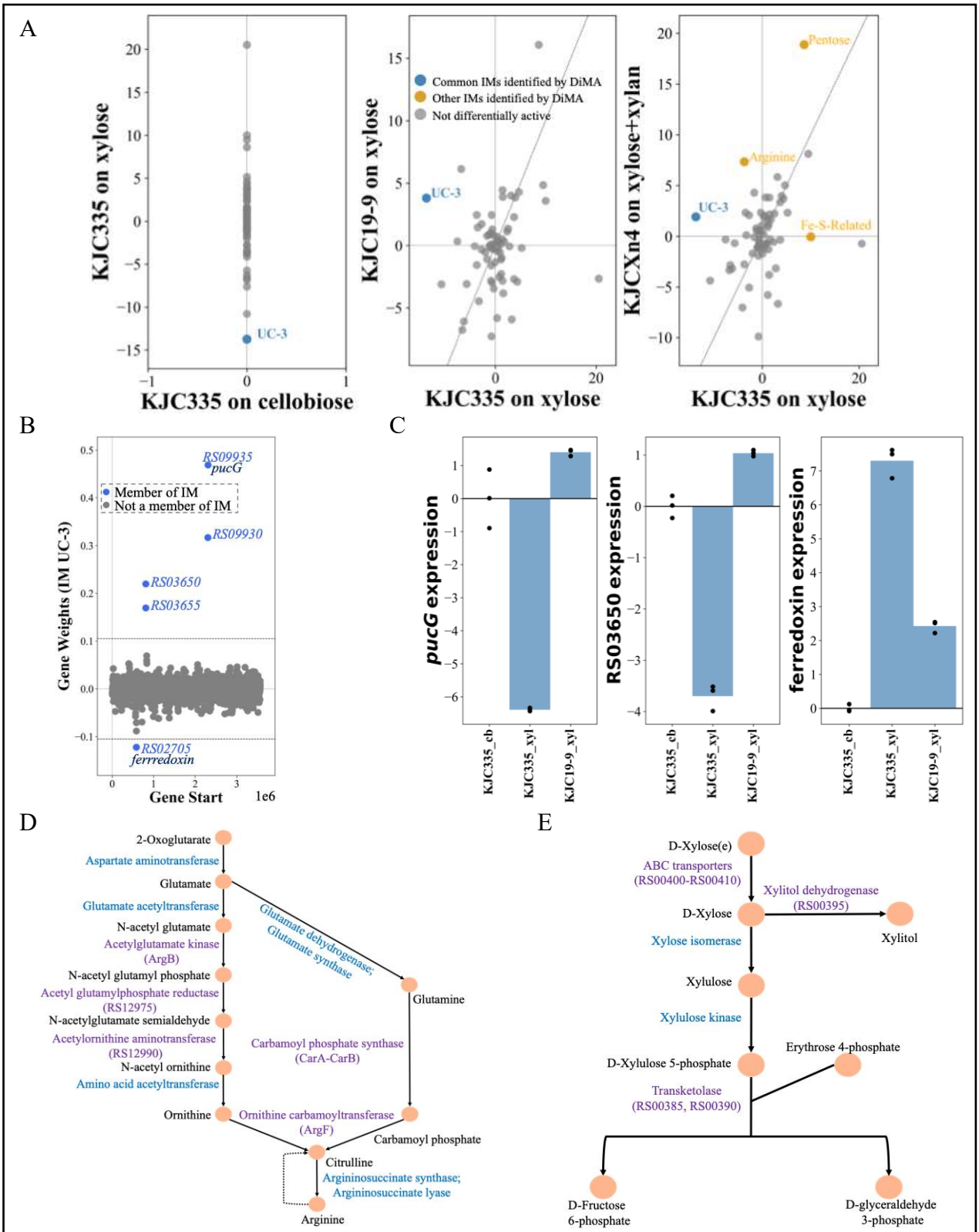


Fig. 4. IMs affected by pentose sugar utilization in *C. thermocellum*. a) DiMA plots comparing growth conditions for different mutants in cellobiose, xylose and xylose with xylan growth conditions. Statistically significantly differentiated IMs are colored and common IM is colored and labeled in dark blue. (b) Gene Weight plot for UC-3 IM with member genes colored and labeled in blue. (c) Expression of selected member genes of UC-3 IM compared across mutant and substrate conditions (cb: cellobiose, xyl: xylose). (d) Arginine pathway under the regulation of Arginine IM with member genes present highlighted in purple. Dotted arrow indicates absence of the pathway in *C. thermocellum*. (e) Xylose transport and utilization through pentose phosphate pathway under the control of Pentose IM (member genes highlighted in purple).

479 phosphate (DXP)-dependent pathway with starting compound erythrose 4'- phosphate (E4P) [64].

However, in the presence of xylose, E4P is channeled to the conversion of xylulose 5'-phosphate (XU5P) to produce fructose 6'-phosphate (F6P) and glyceraldehyde 3'-phosphate (G3P) (Figure 4E). Therefore, due to lower availability of E4P, the organism likely downregulates the expression of PLP-dependent enzymes, such as UGT. This type of metabolic reprogramming can induce considerable alteration in redox state and stress on the organism. Furthermore, another significantly downregulated gene Clo1313_0705 (RS03650) within the UC-3 IM possesses the PFAM domain PF07155, which is the substrate-binding component of energy coupling-factor (ECF)-type riboflavin transporter. Its repression also suggests significant redox alterations in the mutant organism when grown on non-native substrate, xylose. To reach homeostasis, *C. thermocellum* likely must induce the production of ferredoxin and neighboring ferredoxin-binding protein— RS02710 (Supplementary Fig. 9).

Nevertheless, these same genes show drastic, including reversal of, expression changes in strain KJC19-9 on xylose (Figure 4C). Genes *pucG*, RS03650, and RS02705 (ferredoxin) are upregulated in KJC19-9 with the latter showing lower expression than compared to KJC335 on xylose. These expression alterations indicate the evolved strain's demonstrated ability to grow well on xylose, as observed in a previous study [62] likely because KJC19-9 contains mutations to lower redox stress within the organism to restore metabolism.

We also compared the IM activities for KJC335 and KJCXn4 on xylose to find that none of the IMs were significantly differentially active (Supplementary Fig. 10), suggesting strains evolved to grow on xylose and xylan probably do not have significant alterations in gene regulation for growth in xylose alone. A previous study revealed that the evolution on xylose and xylan does not confer any significant advantage to the strain's ability to grow on xylose as KJCXn4 does not display growth rates as high as KJC19 strains on xylose [62]. Our finding, in agreement with the

503 previous study, provides molecular basis for the observed phenotype in the two strains when grown
504 on xylose.

505 Four IMs—UC-3, Arginine, Fe-S-Related, and Pentose—were differentially altered
506 between KJC335 on xylose and KJCXn4 on xylose and xylan (Figure 4A). UC-3, being active in
507 xylose and xylan suggests that this is an important regulatory module in the adaptation to redox
508 stress to restore metabolism for growth on hemicellulosic sugars in *C. thermocellum.* The Arginine
509 IM was also highly active in mixed pentose substrates, and this IM contains gene cluster
510 (Clo1313_2555-Clo1313_2562 except Clo1313_2557) encoding enzymes in the arginine
511 biosynthesis pathway (Figure 4D). Arginine biosynthesis genes have been found to be upregulated
512 in xylose growth condition in a related thermophilic and anaerobic bacterial strain
513 *Thermoanaerobacter* sp. X514 [65]. In the study, the researchers proposed that the arginine
514 produced during mid-exponential phase serves as an ATP source through arginine deiminase
515 (ADI) pathway in the stationary phase. Since no known ADI encoding gene is present in *C.*
516 *thermocellum*, exact mechanism for the activation of the IM is unknown. Likewise, the activation
517 of arginine pathway in mixed substrate hemicellulosic conditions and in *C. thermocellum* has not
518 been investigated. It is possible that *C. thermocellum* either directs arginine flux for ATP
519 production through other pathways or arginine is a nitrogen sink, which needs to be produced to
520 balance nitrogen within the cell during growth on xylose and xylan. Further investigation on the
521 link between arginine metabolism and hemicellulosic substrate utilization could lead to improved
522 uptake and growth on xylose and xylan substrates. Another IM, Fe-S-Related, was substantially
523 repressed compared to KJC335 on xylose, suggesting the organism's adaptation to stress and
524 restored metabolism on non-native hemicellulosic substrates. The IM possesses genes RS02705
525 (ferredoxin) and RS02710 (ferredoxin-binding protein) with high gene weights, which were also

present in the UC-3 IM. The lower activity of Fe-S-Related IM reflects lower expression of constituent genes and demonstrates adaptation of cell as was the case in KJC19 strain on xylose.

Lastly, the Pentose IM is highly active in xylose and xylan mixed substrate condition. The IM is composed of gene cluster (Clo1313_0073-Clo1313_0080) of which the genes Clo1313_0073 (glycerol kinase, *glpK*) and Clo1313_0080 (phosphoglycerate mutase, *gpmA*) are not part of xylose transport and utilization through pentose phosphate pathway (Figure 4E). The activation of the pentose phosphate pathway is evident as the strains utilize xylose and xylan via the heterologous xylose isomerase pathway, with higher availability of xylose in mixed substrate condition. Likewise, the upregulation of *gpmA* is an indication that flux through lower glycolytic channel is more active in mixed pentose sugar growth condition. Finally, the higher expression of *glpK* appears to be a mechanism to direct flux of glycerol, which likely accumulates because of cofactor imbalance, as has been shown to be the case in *S. cerevisiae* strains engineered to utilize xylose [66]. To the best of our knowledge, no such findings have been previously demonstrated in *C. thermocellum* or in mixed hemicellulosic growth conditions. However, it is possible that rationale strain engineering strategies to redirect carbon flux away from glycerol and/or improve cofactor balancing could be critical in improving ethanol yield from hemicellulosic substrates in *C. thermocellum*, as has been demonstrated in *S. cerevisiae* [66].

4. Conclusion

This study provides a systems-level framework for understanding transcriptional regulation in *Clostridium thermocellum* and uncovers candidate engineering targets for improved ethanol production and lignocellulosic substrate utilization. Applying ICA to a compendium of

223 high-quality RNA-seq samples identified 67 iModulons that collectively explain more than 80% of the observed gene expression variance.

Integrating iModulon activities with statistical analyses, metabolomic data, and mechanistic modeling suggested that cofactor balancing is an important constraint on ethanol tolerance and production. In high ethanol-yielding strains, the analyses support a model in which altered redox cofactor allocation toward NAD(P)H generation reduces ferredoxin availability for sulfate assimilation, highlighting redox balancing and sulfur metabolism as potential targets for further engineering. Differential iModulon activities also demonstrated the activation of CAZyme, SigI6, and LexA_1449 IMs and linked cellulose utilization to cellular stress responses. Likewise, changes in activities of multiple IMs implicated pathways related to PLP synthesis, arginine biosynthesis, and pentose metabolism as candidate processes underlying adaptation to growth on xylose and xylan. These findings recommend regulatory targets, including *dksA*, and metabolic strategies addressing cofactor imbalance and glycerol accumulation for experimental evaluation.

The CT-compendium, iModulon resource, updated iModulonMiner pipeline, and accompanying analysis scripts provide a foundation for extending this approach in *C. thermocellum* and related bioprocessing organisms. Additional expression datasets, direct characterization of regulators, and more physiological experiments will further refine the proposed iModulon-based network, advancing rational strain development for consolidated bioprocessing.

Author Contributions

SD: conceptualization, methodology, investigation, and writing—original draft preparation. CDM: mentorship and writing—review and editing.

Funding

The Center for Bioenergy Innovation (CBI), U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number ERKP886.

Acknowledgements

This material is based upon work supported by the Center for Bioenergy Innovation (CBI), U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number ERKP886. The authors of this work recognize the Penn State Institute for Computational and Data Sciences (RRID: SCR_025154) for providing access to computational research infrastructure within the Roar Core Facility (RRID: SCR_026424).

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Xu Q, Resch MG, Podkaminer K, Yang S, Baker JO, Donohoe BS, Wilson C, Klingeman DM, Olson DG, Decker SR: **Dramatic performance of Clostridium thermocellum explained by its wide range of cellulase modalities**. *Science advances* 2016, **2**(2):e1501254.
2. Zhang Y-HP, Lynd LR: **Cellulose utilization by Clostridium thermocellum: bioenergetics and hydrolysis product assimilation**. *Proceedings of the National Academy of Sciences* 2005, **102**(20):7321–7325.
3. Schroeder WL, Kuil T, van Maris AJ, Olson DG, Lynd LR, Maranas CD: **A detailed genome-scale metabolic model of Clostridium thermocellum investigates sources of pyrophosphate for driving glycolysis**. *Metabolic engineering* 2023, **77**:306–322.
4. Olson DG, Hörl M, Fuhrer T, Cui J, Zhou J, Maloney MI, Amador-Noguez D, Tian L, Sauer U, Lynd LR: **Glycolysis without pyruvate kinase in Clostridium thermocellum**. *Metabolic engineering* 2017, **39**:169–180.
5. Olson DG, Maloney MI, Lanahan AA, Cervenka ND, Xia Y, Pech-Canul A, Hon S, Tian L, Ziegler SJ, Bomble YJ: **Ethanol tolerance in engineered strains of Clostridium thermocellum**. *Biotechnology for Biofuels and Bioproducts* 2023, **16**(1):137.
6. Holwerda EK, Olson DG, Ruppertsberger NM, Stevenson DM, Murphy SJL, Maloney MI, Lanahan AA, Amador-Noguez D, Lynd LR: **Metabolic and evolutionary responses of Clostridium thermocellum to genetic interventions aimed at improving ethanol production**. *Biotechnol Biofuels* 2020, **13**:40.
7. Thompson RA, Dahal S, Garcia S, Nookaew I, Trinh CT: **Exploring complex cellular phenotypes and model-guided strain design with a novel genome-scale metabolic model of Clostridium thermocellum DSM 1313 implementing an adjustable cellulosome**. *Biotechnology for biofuels* 2016, **9**(1):194.
8. Hon S, Jacobson T, Stevenson DM, Maloney MI, Giannone RJ, Hettich RL, Amador-Noguez D, Olson DG, Lynd LR: **Increasing the thermodynamic driving force of the phosphofructokinase reaction in Clostridium thermocellum**. *Applied and Environmental Microbiology* 2022, **88**(22):e01258–01222.
9. Sharma BD, Hon S, Thusoo E, Stevenson DM, Amador-Noguez D, Guss AM, Lynd LR, Olson DG: **Pyrophosphate-free glycolysis in Clostridium thermocellum increases both thermodynamic driving force and ethanol titers**. *Biotechnology for biofuels and bioproducts* 2024, **17**(1):146.

10. Chiarelli DP, Lynd LR, Bergamo LW, Olson DG: **Effects of heterologous expression of glycolysis enzymes on product formation in Clostridium thermocellum**. *Metabolic Engineering Communications* 2026:e00279.
11. Tian L, Papanek B, Olson DG, Rydzak T, Holwerda EK, Zheng T, Zhou J, Maloney M, Jiang N, Giannone RJ: **Simultaneous achievement of high ethanol yield and titer in Clostridium thermocellum**. *Biotechnology for biofuels* 2016, **9**(1):116.
12. Pech-Canul A, Hammer SK, Ziegler SJ, Richardson ID, Sharma BD, Maloney MI, Bomble YJ, Lynd LR, Olson DG: **The role of AdhE on ethanol tolerance and production in Clostridium thermocellum**. *Journal of Biological Chemistry* 2024, **300**(8):107559.
13. Whitham JM, Moon J-W, Rodriguez Jr M, Engle NL, Klingeman DM, Rydzak T, Abel MM, Tschaplinski TJ, Guss AM, Brown SD: **Clostridium thermocellum LL1210 pH homeostasis mechanisms informed by transcriptomics and metabolomics**. *Biotechnology for biofuels* 2018, **11**(1):98.
14. Hebdon SD, Gerritsen AT, Chen Y-P, Marcano JG, Chou KJ: **Genome-wide transcription factor DNA binding sites and gene regulatory networks in Clostridium thermocellum**. *Frontiers in Microbiology* 2021, **12**:695517.
15. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, Choudhary KS, Yang L, King ZA, Palsson BO: **The Escherichia coli transcriptome mostly consists of independently regulated modules**. *Nat Commun* 2019, **10**(1):5536.
16. Shin J, Zielinski DC, Palsson BO: **Deciphering nutritional stress responses via knowledge-enriched transcriptomics for microbial engineering**. *Metabolic Engineering* 2024, **84**:34–47.
17. Yoo R, Rychel K, Poudel S, Al-Bulushi T, Yuan Y, Chauhan S, Lamoureux C, Palsson BO, Sastry A: **Machine learning of all Mycobacterium tuberculosis H37Rv RNA-seq data reveals a structured interplay between metabolism, stress response, and infection**. *MSphere* 2022, **7**(2):e00033–00022.
18. Saelens W, Cannoodt R, Saeys Y: **A comprehensive evaluation of module detection methods for gene expression data**. *Nature communications* 2018, **9**(1):1090.
19. Chauhan SM, Poudel S, Rychel K, Lamoureux C, Yoo R, Al Bulushi T, Yuan Y, Palsson BO, Sastry AV: **Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile Sulfolobus acidocaldarius**. *Front Microbiol* 2021, **12**:753521.
20. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C: **Elucidating the altered transcriptional programs in breast cancer using independent component analysis**. *PLoS computational biology* 2007, **3**(8):e161.
21. Lim HG, Rychel K, Sastry AV, Bentley GJ, Mueller J, Schindel HS, Larsen PE, Laible PD, Guss AM, Niu W *et al*: **Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network**. *Metab Eng* 2022, **72**:297–310.
22. Rychel K, Sastry A, Palsson B: **Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. Nat Commun 11: 6338**. In.; 2020.
23. Rajput A, Tsunemoto H, Sastry AV, Szubin R, Rychel K, Sugie J, Pogliano J, Palsson BO: **Machine learning from Pseudomonas aeruginosa transcriptomes identifies independently modulated sets of genes associated with known transcriptional regulators**. *Nucleic Acids Res* 2022, **50**(7):3658–3672.
24. Rajput A, Tsunemoto H, Sastry AV, Szubin R, Rychel K, Chauhan SM, Pogliano J, Palsson BO: **Advanced transcriptomic analysis reveals the role of efflux pumps and media composition in antibiotic responses of Pseudomonas aeruginosa**. *Nucleic Acids Res* 2022, **50**(17):9675–9688.
25. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, López-Bigas N, Kamoun A, Neuzillet Y, Gestraud P: **Independent component analysis uncovers the landscape**

of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell reports* 2014, **9**(4):1235–1245.
26. Choe D, Olson CA, Szubin R, Yang H, Sung J, Feist AM, Palsson BO: **Advancing the scale of synthetic biology via cross-species transfer of cellular functions enabled by iModulon engraftment**. *Nature Communications* 2024, **15**(1):2356.
27. Shin J, Zielinski DC, Palsson BO: **Modulating bacterial function utilizing A knowledge base of transcriptional regulatory modules**. *Nucleic Acids Research* 2024, **52**(18):11362–11377.
28. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, Feldgarden M, Fine AM, Funk K, Hoffman J: **Database resources of the National Center for Biotechnology Information in 2025**. *Nucleic acids research* 2025, **53**(D1):D20–D29.
29. Sastry AV, Yuan Y, Poudel S, Rychel K, Yoo R, Lamoureux CR, Li G, Burrows JT, Chauhan S, Haiman ZB *et al*: **iModulonMiner and PyModulon: Software for unsupervised mining of gene expression compendia**. *PLoS Comput Biol* 2024, **20**(10):e1012546.
30. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C: **Nextflow enables reproducible computational workflows**. *Nature biotechnology* 2017, **35**(4):316–319.
31. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome biology* 2009, **10**(3):R25.
32. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments**. *Bioinformatics* 2012, **28**(16):2184–2185.
33. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**(7):923–930.
34. Ewels P, Magnusson M, Lundin S, Käller M: **MultiQC: summarize analysis results for multiple tools and samples in a single report**. *Bioinformatics* 2016, **32**(19):3047–3048.
35. Mukherjee S, Stamatis D, Li CT, Ovchinnikova G, Kandimalla M, Handke V, Reddy A, Ivanova N, Woyke T, Eloe-Fardosh EA: **Genomes OnLine Database (GOLD) v. 10: new features and updates**. *Nucleic acids research* 2025, **53**(D1):D989–D997.
36. Hyvarinen A: **Fast and robust fixed-point algorithms for independent component analysis**. *IEEE transactions on Neural Networks* 1999, **10**(3):626–634.
37. Ester M, Kriegel H-P, Sander J, Xu X: **A density-based algorithm for discovering clusters in large spatial databases with noise**. In: *kdd: 1996*. 226–231.
38. Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO: **A multi-scale expression and regulation knowledge base for Escherichia coli**. *Nucleic Acids Res* 2023, **51**(19):10176–10193.
39. **UniProt: the universal protein knowledgebase in 2025**. *Nucleic acids research* 2025, **53**(D1):D609–D617.
40. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q: **The BioCyc collection of microbial genomes and metabolic pathways**. *Briefings in bioinformatics* 2019, **20**(4):1085–1093.
41. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P: **Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper**. *Molecular biology and evolution* 2017, **34**(8):2115–2122.
42. **Galaxy for accessible, reproducible, and collaborative data analyses: 2026 update**. *Nucleic Acids Research* 2026:gkag469.
43. Willis TC, Schroeder WL, Khana DB, Qian X, Dahal S, Amador-Noguez D, Maranas CD: **Reconstruction of a resource balance analysis model of Clostridium thermocellum examines the metabolic cost of glycolytic and cellulosome enzymes**. *Metabolic engineering* 2025.

44. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D: **Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models**. *Molecular systems biology* 2010, **6**:390.
45. Kim P-J, Lee D-Y, Kim TY, Lee KH, Jeong H, Lee SY, Park S: **Metabolite essentiality elucidates robustness of Escherichia coli metabolism**. *Proceedings of the National Academy of Sciences* 2007, **104**(34):13638–13642.
46. Biswas R, Wilson CM, Giannone RJ, Klingeman DM, Rydzak T, Shah MB, Hettich RL, Brown SD, Guss AM: **Improved growth rate in Clostridium thermocellum hydrogenase mutant via perturbed sulfur metabolism**. *Biotechnology for biofuels* 2017, **10**(1):6.
47. Garcia S, Thompson RA, Giannone RJ, Dash S, Maranas CD, Trinh CT: **Development of a genome-scale metabolic model of Clostridium thermocellum and its applications for integration of multi-omics datasets and computational strain design**. *Frontiers in bioengineering and biotechnology* 2020, **8**:772.
48. Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM: **Multi-omics quantification of species variation of Escherichia coli links molecular features with strain phenotypes**. *Cell systems* 2016, **3**(3):238–251. e212.
49. Lo J, Olson DG, Murphy SJ-L, Tian L, Hon S, Lanahan A, Guss AM, Lynd LR: **Engineering electron metabolism to increase ethanol production in Clostridium thermocellum**. *Metabolic engineering* 2017, **39**:71–79.
50. Gray B, Hall P, Gresham H: **Targeting agr-and agr-like quorum sensing systems for development of common therapeutics to treat multiple gram-positive bacterial infections**. *Sensors* 2013, **13**(4):5130–5166.
51. Kodori M, Ghalavand Z, Yadegar A, Eslami G, Azimirad M, Zali MR: **In-depth characterization of accessory gene regulator loci and associated virulence factors in tcdA+ B+ Clostridioides difficile isolates**. *Current Research in Microbial Sciences* 2025, **9**:100435.
52. Wilson CM, Rodriguez Jr M, Johnson CM, Martin SL, Chu TM, Wolfinger RD, Hauser LJ, Land ML, Klingeman DM, Syed MH: **Global transcriptome analysis of Clostridium thermocellum ATCC 27405 during growth on dilute acid pretreated Populus and switchgrass**. *Biotechnology for biofuels* 2013, **6**(1):179.
53. Ma X, Ma L, Huo Y-X: **Reconstructing the transcription regulatory network to optimize resource allocation for robust biosynthesis**. *Trends in biotechnology* 2022, **40**(6):735–751.
54. Girard ME, Gopalkrishnan S, Grace ED, Halliday JA, Gourse RL, Herman C: **DksA and ppGpp regulate the σS stress response by activating promoters for the small RNA DsrA and the anti-adapter protein IraP**. *Journal of bacteriology* 2018, **200**(2):10.1128/jb. 00463–00417.
55. Nataf Y, Bahari L, Kahel-Raifer H, Borovok I, Lamed R, Bayer EA, Sonenshein AL, Shoham Y: **Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors**. *Proceedings of the National Academy of Sciences* 2010, **107**(43):18646–18651.
56. Kahel-Raifer H, Jindou S, Bahari L, Nataf Y, Shoham Y, Bayer EA, Borovok I, Lamed R: **The unique set of putative membrane-associated anti-σ factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation**. *FEMS microbiology letters* 2010, **308**(1):84–93.
57. Munoz-Gutierrez I, Ortiz de Ora L, Rozman Grinberg I, Garty Y, Bayer EA, Shoham Y, Lamed R, Borovok I: **Decoding biomass-sensing regulons of Clostridium thermocellum alternative sigma-I factors in a heterologous Bacillus subtilis host system**. *Plos one* 2016, **11**(1):e0146316.
58. Ichikawa S, Ito D, Asaoka S, Abe R, Katsuo N, Ito T, Ito D, Karita S: **The expression of alternative sigma-I7 factor induces the transcription of cellulosomal genes in the cellulolytic bacterium Clostridium thermocellum**. *Enzyme and microbial technology* 2022, **156**:110002.

59. Raman B, McKeown CK, Rodriguez Jr M, Brown SD, Mielenz JR: **Transcriptomic analysis of Clostridium thermocellum ATCC 27405 cellulose fermentation**. *Bmc Microbiology* 2011, **11**(1):134.
60. Riederer A, Takasuka TE, Makino S-i, Stevenson DM, Bukhman YV, Elsen NL, Fox BG: **Global gene expression patterns in Clostridium thermocellum as determined by microarray analysis of chemostat cultures on cellulose or cellobiose**. *Applied and environmental microbiology* 2011, **77**(4):1243–1253.
61. Tafur Rangel AE, Croft T, González Barrios AF, Reyes LH, Maness P-C, Chou KJ: **Transcriptomic analysis of a Clostridium thermocellum strain engineered to utilize xylose: responses to xylose versus cellobiose feeding**. *Scientific Reports* 2020, **10**(1):14517.
62. Chou KJ, Croft T, Hebdon SD, Magnusson LR, Xiong W, Reyes LH, Chen X, Miller EJ, Riley DM, Dupuis S: **Engineering the cellulolytic bacterium, Clostridium thermocellum, to co-utilize hemicellulose**. *Metabolic Engineering* 2024, **83**:193–205.
63. Ramazzina I, Costa R, Cendron L, Berni R, Peracchi A, Zanotti G, Percudani R: **An aminotransferase branch point connects purine catabolism to amino acid recycling**. *Nature Chemical Biology* 2010, **6**(11):801–806.
64. Richts B, Rosenberg J, Commichau FM: **A survey of pyridoxal 5'-phosphate-dependent proteins in the gram-positive model bacterium Bacillus subtilis**. *Frontiers in molecular biosciences* 2019, **6**:32.
65. Lin L, Song H, Tu Q, Qin Y, Zhou A, Liu W, He Z, Zhou J, Xu J: **The Thermoanaerobacter glycobiome reveals mechanisms of pentose and hexose co-utilization in bacteria**. *PLoS genetics* 2011, **7**(10):e1002318.
66. Hou J, Suo F, Wang C, Li X, Shen Y, Bao X: **Fine-tuning of NADH oxidase decreases byproduct accumulation in respiration deficient xylose metabolic Saccharomyces cerevisiae**. *Bmc Biotechnology* 2014, **14**(1):13.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- InvestigationofiModulonsinClostridiumthermocellumtocharacterizeconditiondraftSupplementary09302026.docx
- supplementarydatacthermimodulons.zip
- InvestigationofiModulonsinClostridiumthermocellumtocharacterizeconditiondraftSupplementarytables.docx