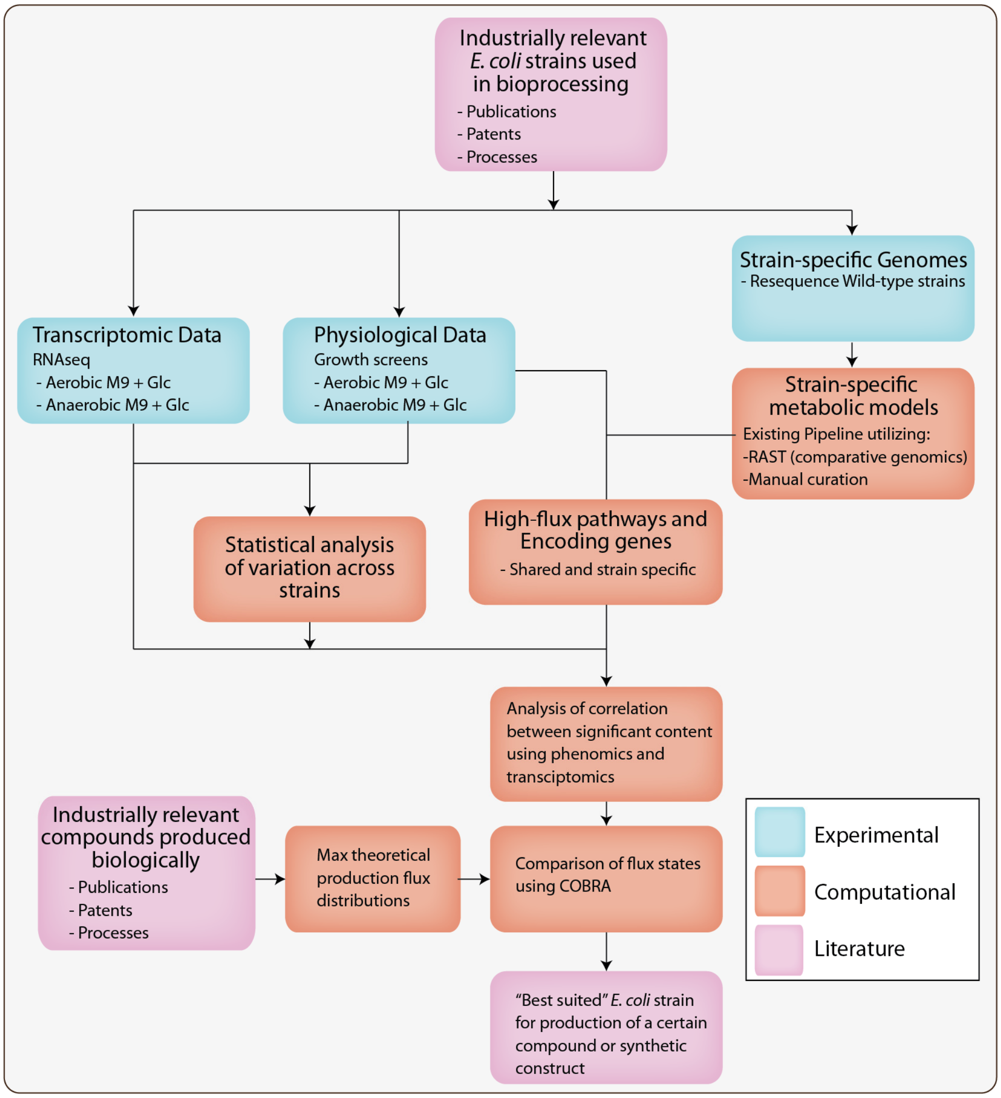
Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes

Supplementary Figures and Tables

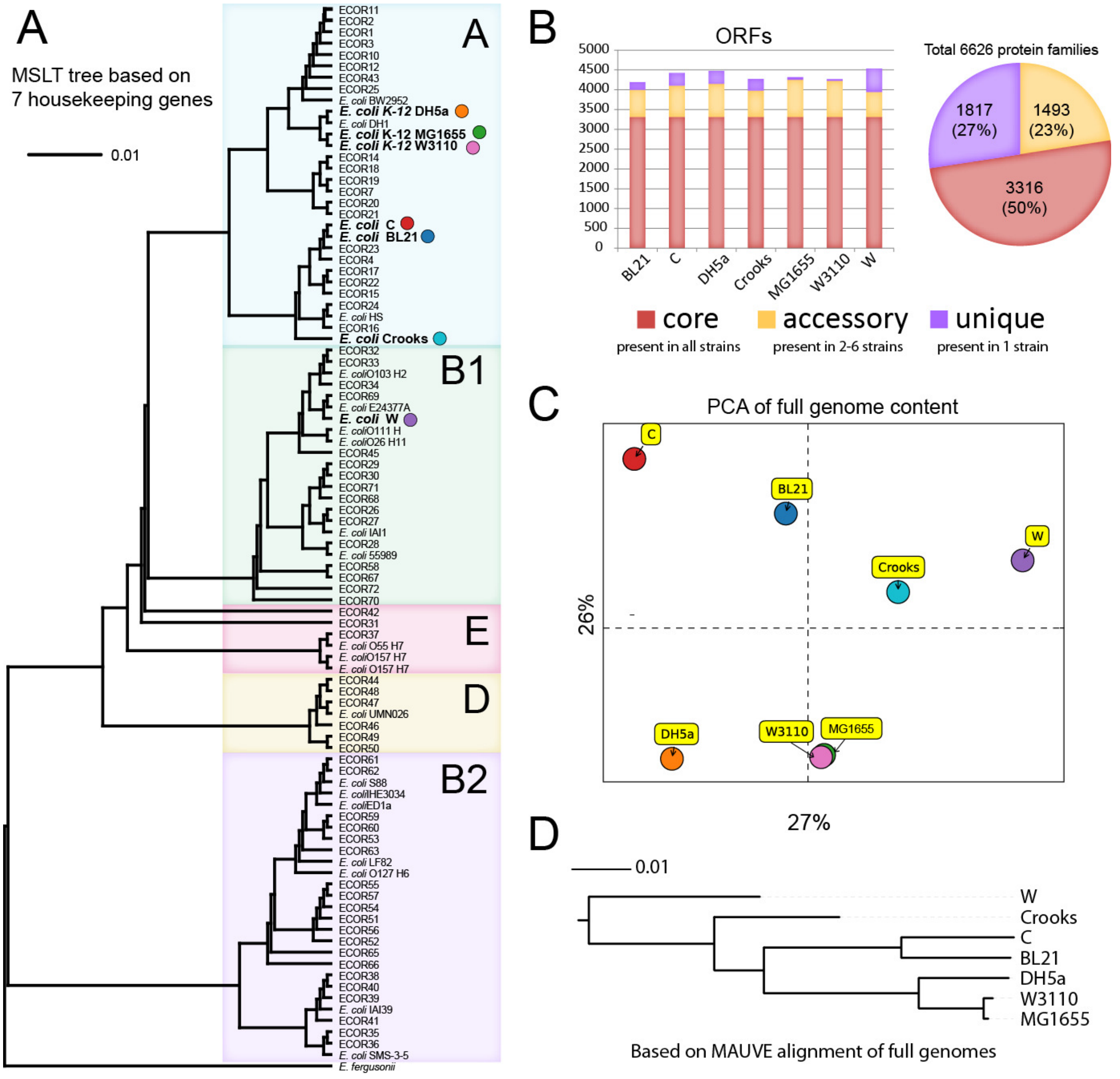
Contents

Supplemental Figures



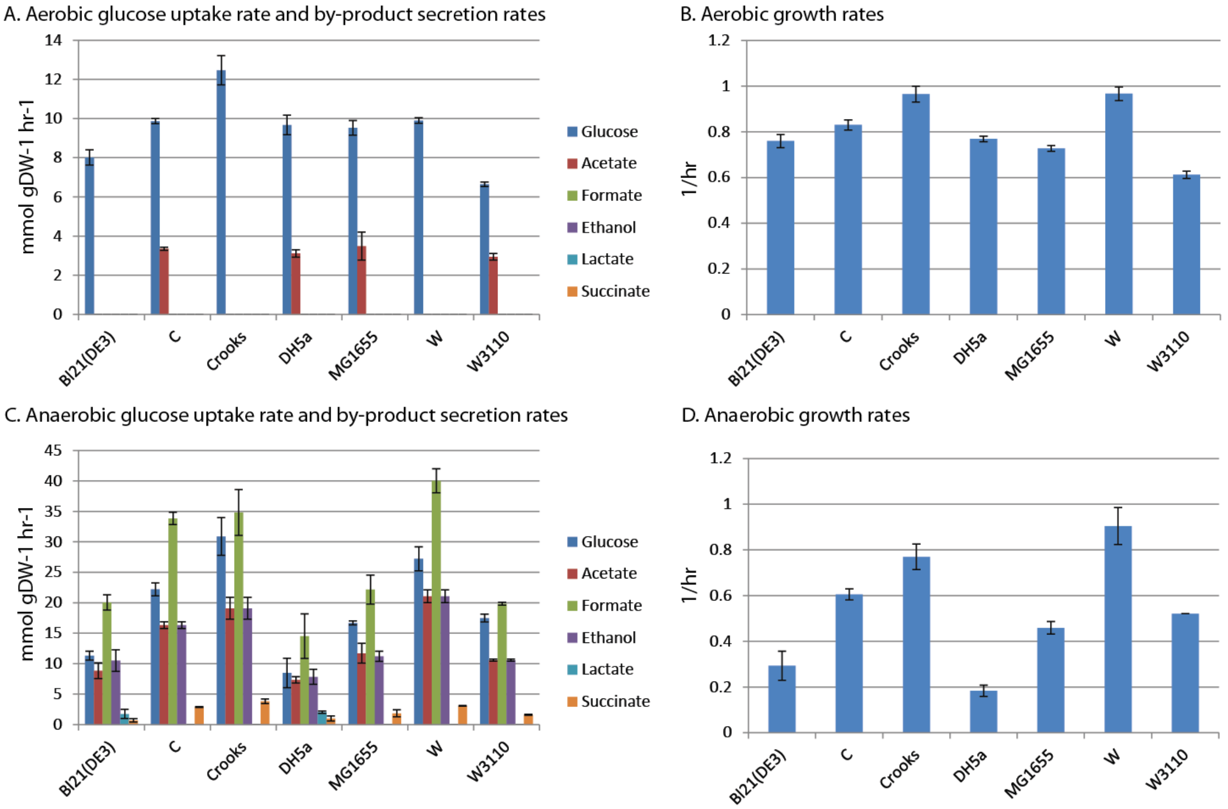
Supplemental Figure 1: Project workflow, Related to Figure 1.

This workflow illustrates step taken to integrate data on the 7 strains based on literature search, computational analysis and experimental data generation.



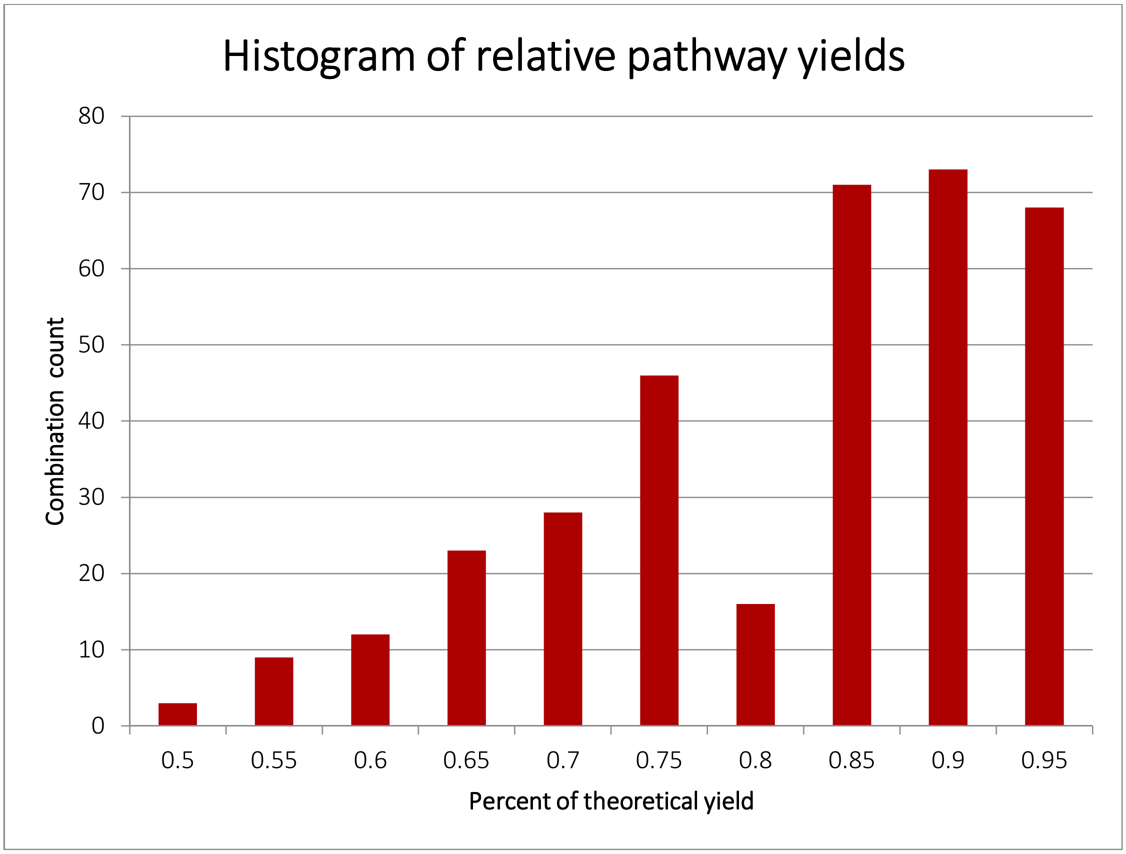
Supplemental Figure 2: Genome content statistics, Related to Figure 1.

A) MLST tree based on 7 housekeeping genes. Letters in the top right corner represent establish *E. coli* phylogroups. **B)** Core genes are those that were present in all 7 strains, genes present in 2-6 strains are labeled accessory and those genes unique to a single strain are labeled as unique. **C)** PCA of full genome content based on binary presence or absence of genes. **D)** MAUVE alignment of full genomes.



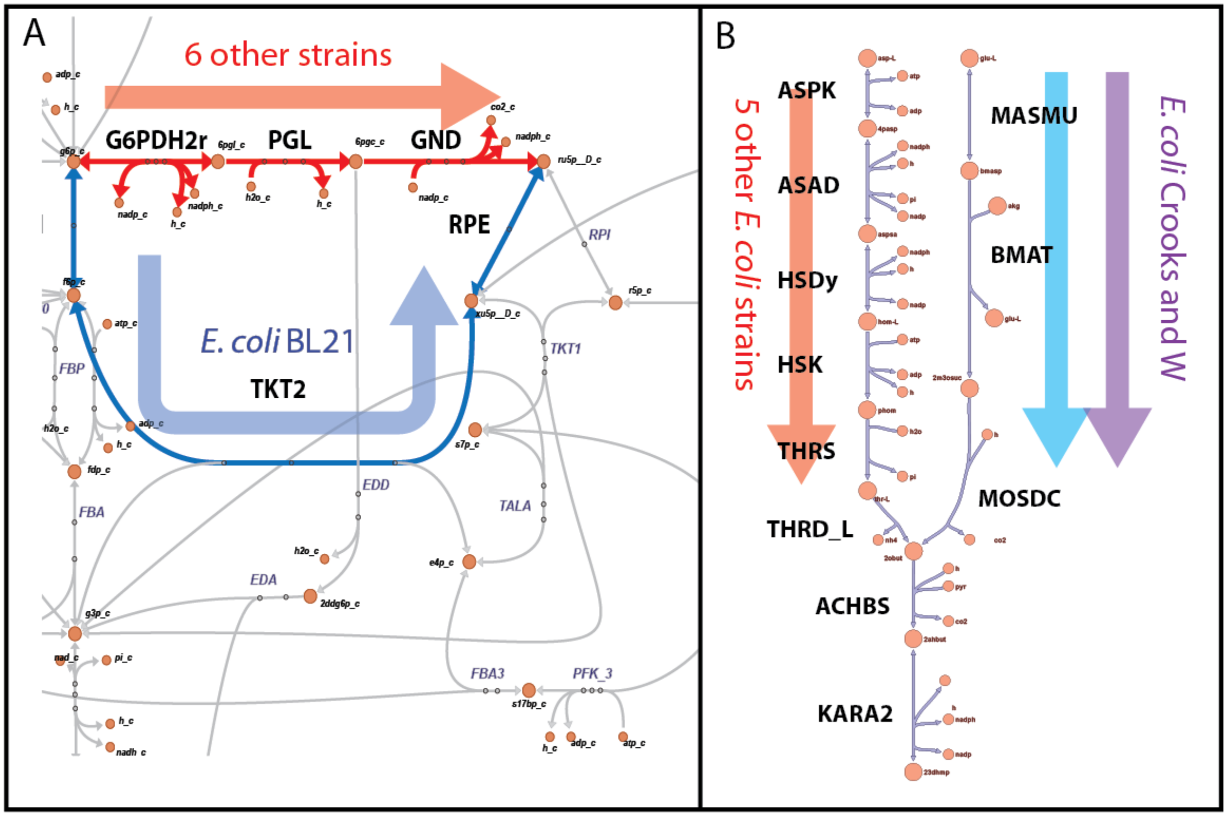
Supplemental Figure 3: Aerobic and anaerobic physiological data, Related to Figure 2.

Glucose uptake rates and by-product secretion rates in aerobic conditions **(A)** and anaerobic conditions **(C)** Growth rates in aerobic conditions **(B)** and anaerobic conditions **(D).**



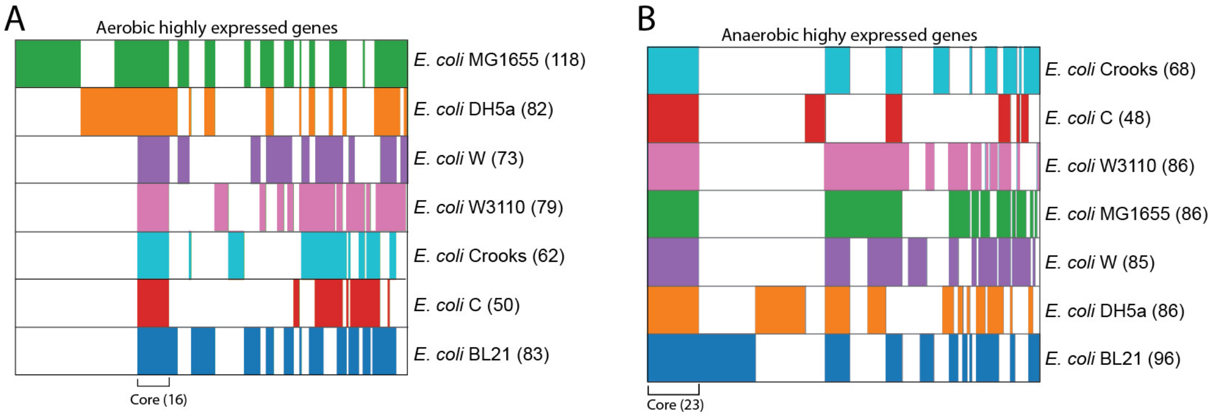
Supplemental Figure 4: Histogram of 341 combinations of strain and condition that yield 45-95% of the highest yield for 245 pathways, Related to Figure 2.

Theoretical yields for production of industrial compounds via heterologous pathways in aerobic and anaerobic conditions. All theoretical yield data is available in **Supplemental Data File 4**. Refer to text for a description of these results.

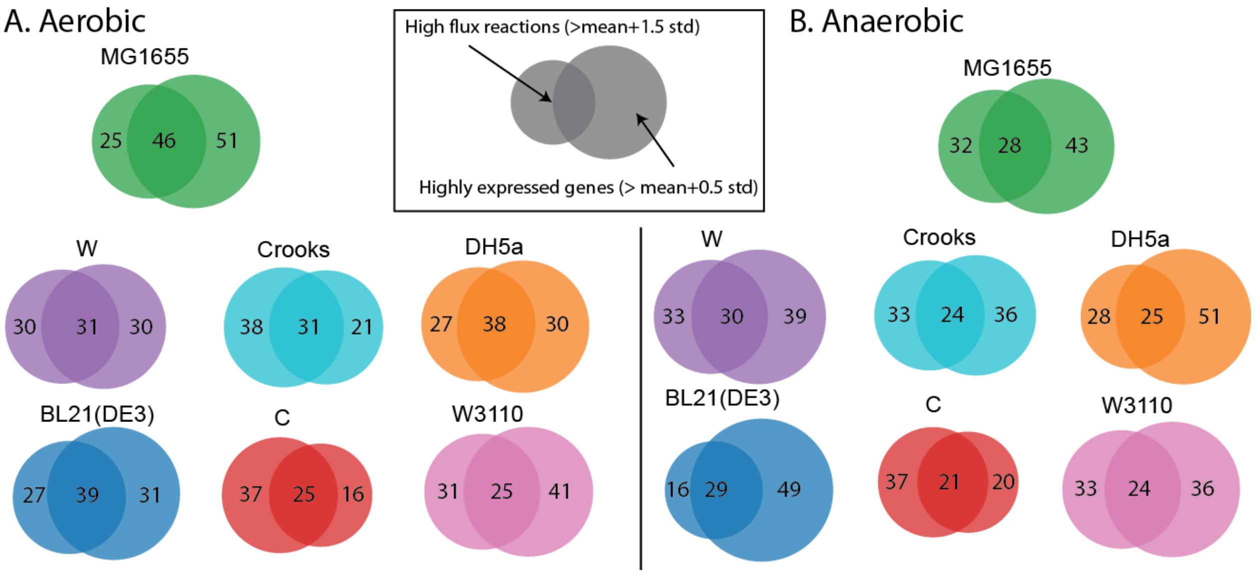


Supplemental Figure 5: Reaction network differences between strains of *E. coli*, Related to Figure 2.

A) *E. coli* BL21 lacks the PGL reaction forcing flux through TALA to generate ribulose-5-phosphate, instead of via the oxidative Pentose Phosphate Pathway. This alternate route does not generate NADPH leading to decreased yield of several compounds. **B)** *E. coli* strains Crooks and W have an alternate isoleucine biosynthesis pathway.

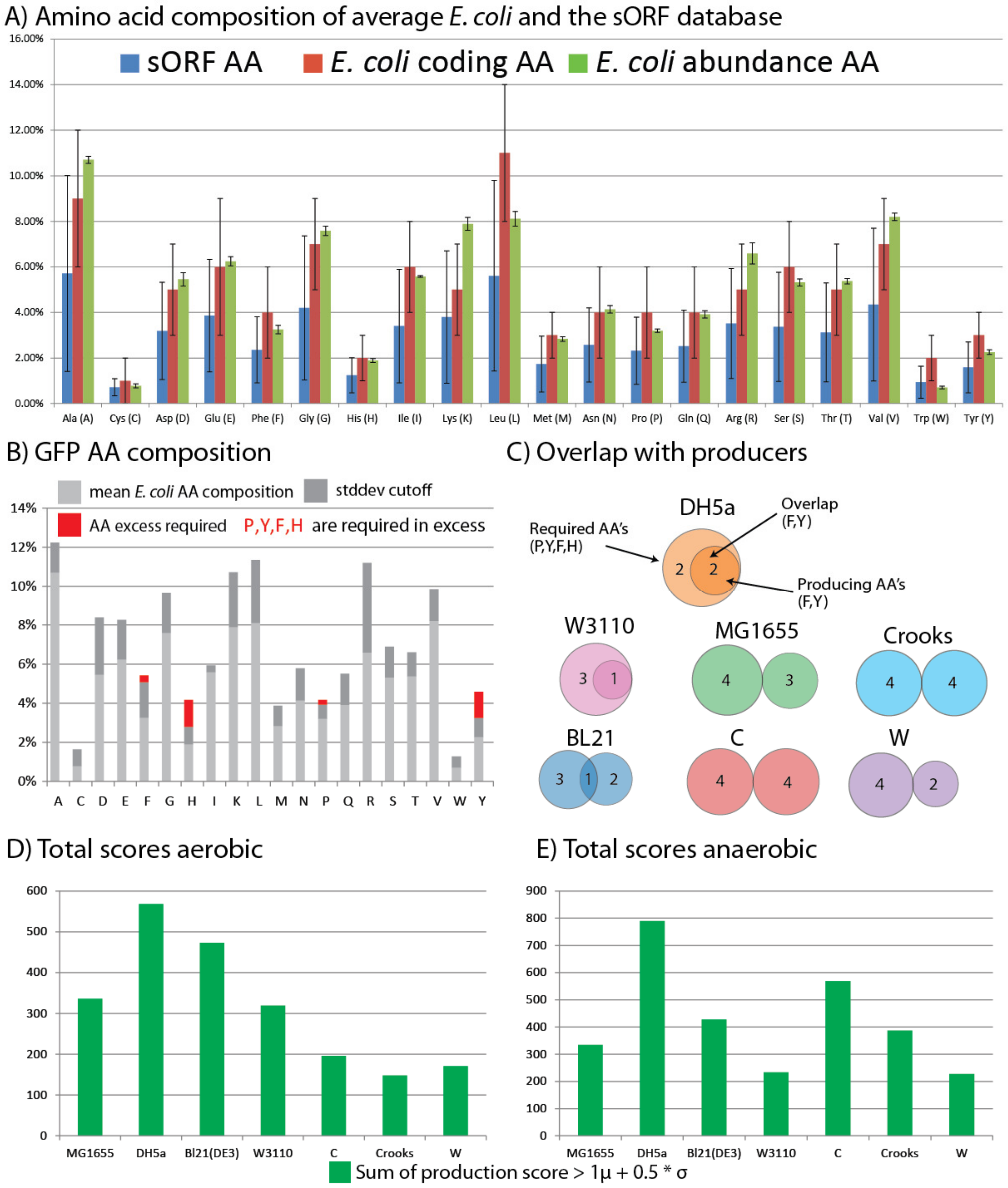


Supplemental Figure 6: Highly expressed genes in aerobic and anaerobic conditions, Related to Figure 3. Highly expressed genes for each strain were clustered and plotted. The counts of highly expressed genes for each strain are indicated next to the strain name in parentheses. Shared (core) and unique high flux reactions for each strain are shown for **A)** aerobic and **B)** anaerobic conditions.



Supplemental Figure 7: Venn diagrams comparing high flux reaction catalyzing genes and highly expressed genes, Related to Figure 4.

A) aerobic and **B)** and anaerobic conditions. The size of each circle is proportional to the number of entries and the standard deviations used to determine each set were determined such that sample sets were comparable in size.


A) Amino acid composition of average *E. coli* and the sORF database

B) GFP AA composition

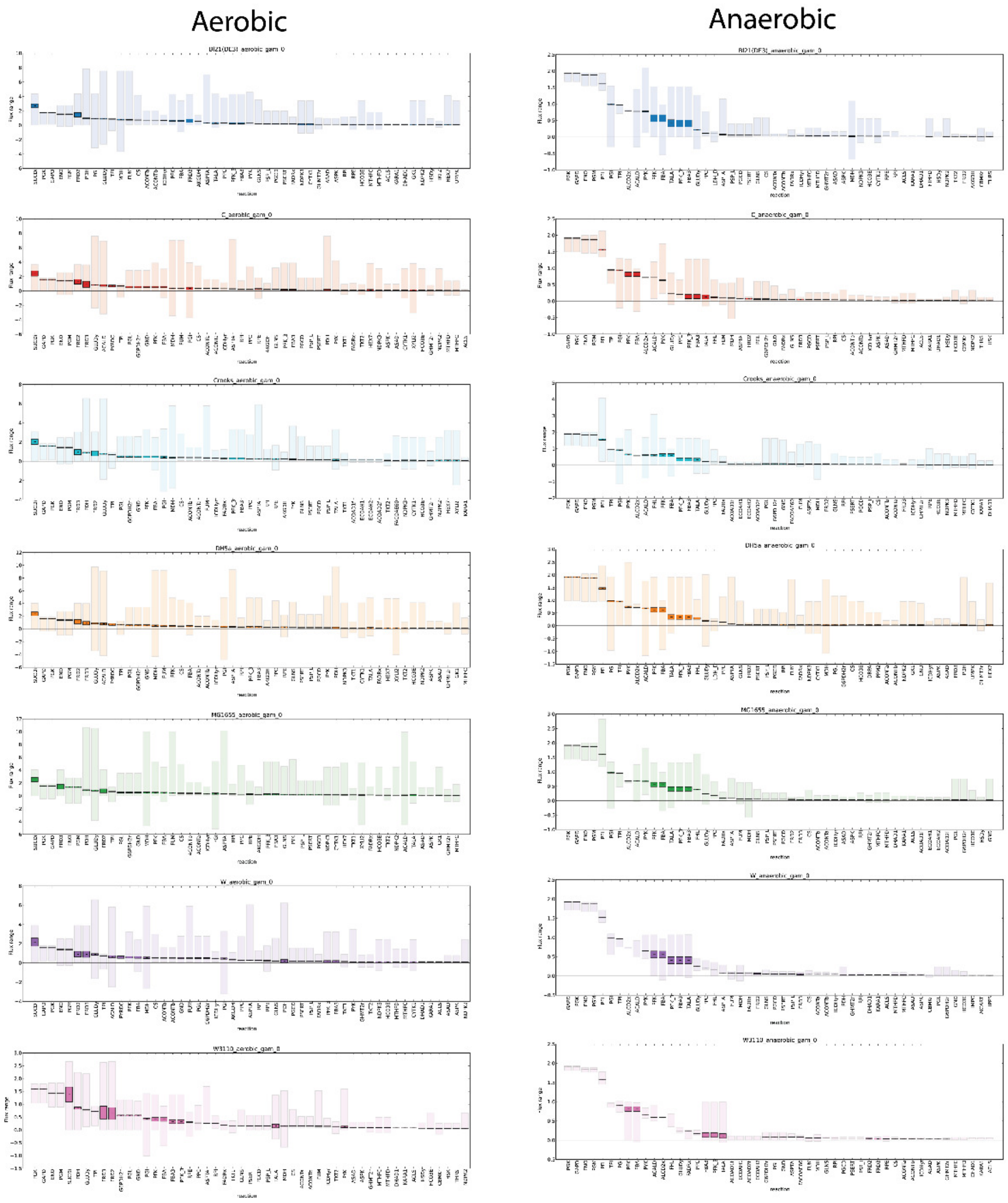
C) Overlap with producers

D) Total scores aerobic

E) Total scores anaerobic

Supplemental Figure 8: Production potential of synthetic biology constructs, Related to Figure 5.

A) A comparison of the average amino acid composition in three datasets: 1) the average amino acid composition frequency (%) and standard deviation of each sORF (synthetic ORF) in the Registry of Standard Biological Parts according to the sequence (blue) 2) the average amino acid composition frequency (%) of each ORF in *E. coli* K-12 MG1655 according to sequence (red), 3) The average amino acid composition frequency (%) of the ORFs expressed in *E. coli* K-12 MG1655 determined by multiplying the count of each ORF from ribosome profiling data by that ORF's amino acid composition (green). The variation for the expressed amino acid frequency is the lowest. **B)** An example of the amino acid composition for green fluorescent protein (GFP). Grey bars represent the amino acid composition of the average *E. coli* cell (panel A, green bar). The dark grey represents a significance cutoff (>μ+10σ). The amino acid composition of GFP is significantly above this cutoff for proline (P), tyrosine (Y), phenylalanine (F), histidine (H) and glycine (G). Thus, these amino acids are considered to be in demand for high expression of this protein. **C)** The amino acids considered to be in demand were compared to each strain's best amino acid production potentials (see R-score). The Venn diagrams indicate overlap of amino acids in demand for GFP and each strain's preferred highly expressed amino acid production profiles in aerobic conditions. In this case, DH5a is predicted to be the best producer because it natively has higher production potential for P, Y, F and G, 4 out of the 5 amino acids needed in greater quantity for expression of GFP. **(D and E)** The fractional overlap was compared for all 3,928 coding constructs in the Registry of Standard Biological Parts database for amino acid production potentials in all strains in aerobic **(D)** and anaerobic **(E)** conditions. The bars indicate the number of instances when a strain was found to be enriched in production of key amino acids for a given construct.



Supplemental Figure 9: Normalized flux ranges for each strain in aerobic and anaerobic conditions, Related to Figure 2.

Darker colors indicate range of sampled flux values, lighter shade indicates all possible flux values calculated using flux variability analysis. All flux ranges are available in **Supplemental Data File 3**.

Supplemental Tables

Supplemental Table 1: Summary of genetic differences between published and *de novo* resequenced *E. coli* genomes, Related to Figure 1.

	Mutation[a]	Bl21(DE3)	Crooks	MG1655	W	W3110	DH5a
	NCBI ID	NC_012971	NC_010468	NC_000913.3	NC_017664	NC_007779	NZ_JRYM00000000 [NCBI Nucleotide]
Base substitution	SNP	1	23	n.d.	12	10	0
Short insertions and deletions	INS short	n.d.	n.d.	2	1	1	0
	DEL short	1	1	1	0	1	0
Total		**2**	**24**	**3**	**13**	**12**	0

[a] Mutations detected by the BreSeq pipeline (Barric *et al.*, 2009, McKenna *et al.*, 2010). Insertion Sequence mutations detected were not included in this analysis. Genes associate to phages were not included in the table. Genes with many mutations close in proximity were not considered in the table, but are reported in the table of all mutations in the supplement.

E. coli C was sequenced and published here with NCBI Bioproject Accession: PRJNA314810

SNP: single base substitution
INS short: single base insertion (< 10 bp)
DEL short: short base deletions (< 10 bp)
n.d.: none determined

Supplemental Table 2: Enrichment analysis for transcription factor control across strains, Related to Figure 3.

Ratios of transcript under anaerobic conditions compared with that under aerobic conditions based on an α-level of 0.001 (*significant difference). Table is provided as a separate Excel file in Data S7.

Supplemental Table 3. Computational prediction of major flux shifts between states as validated with mRNA expression data, Related to Figure 4.

	True Prediction Set			False Prediction Set			
	Inc-Up	Dec-Down	Sum	Inc-Down	Dec-Up	Sum	% True Prediction
Bl21(DE3)	0	11	11	0	4	4	73% (11/15)
C	7	28	35	7	1	8	81% (35/43)
Crooks	7	22	29	7	3	10	74% (29/39)
DH5a	11	36	47	2	3	5	90% (47/52)
MG1655	8	31	39	3	0	3	93% (39/42)
W	6	16	22	4	0	4	85% (22/26)
W3110	7	17	24	5	2	7	77% (24/31)

The genes encoding reactions which were predicted to significantly increase (Inc) or decrease (Dec) in flux were aligned with the experimental genes that significantly increased (Up) or decreased (Down) in expression. Therefore, true predictions were those where Inc-Up or Dec-Down, and vice versa for false predictions.

Supplemental Table 4: Major and minor isozymes that catalyze high-flux reactions present in the metabolic models and their measured transcript ratios, Related to Figure 4.

The ratio represents the average expression ratio that each major isozyme is expressed over the minor isozyme transcript measurements. Some strains exhibited strain-specific isozyme expression ratios, as noted in the table.

Reaction[a]	Condition[b]	Major[c]	Minor[d]	Ratio	Literature Evidence
ACALD	Universal	*adhE*	*mhpF*	>800x	
ACONTa,b	Universal	*acnB*	*acnA*	>2x	
ALAR	Universal	*alr*	*dadX*	1x	
ALATA_D2	Universal	*glyA*	*itaE*	15x	
FBA	Universal	*fbaA*	*fbaB, ydjI*	>8x	(Nakahigashi et al., 2009)
FBP	Universal	*fbp*	*glpX, yggF*	>2x	(Nakahigashi et al., 2009)
FHL	Universal	*fdhF, hycB, hycC, hycD, hycE, hycF, hycG*	*hyfA, hyfB, hyfC, hyfD, hyfE, hyfF, hyfG, hyfH, hyfJ*	31x	
GLNS	Universal	*glnA*	*puuA*	>30x	
NDPK3	Universal	*adk*	*ndk*	>27x	
PFL	Universal	*pflB, pflA*	*pflC, pflD, tdcE, yfiD*	>5x	
PGM	Universal	*gpmA*	*tyjC, gpmM*	>8x	(Nakahigashi et al., 2009)
PGMT	Universal	*pgm*	*yqaB*	12x	
PPA	Universal	*ppa*	*ppx, surE*	>2x	
TALA	Universal	*talB*	*talA*	3x	(Nakahigashi et al., 2009)
ACKr	ANA	*ackA*	*purT, tdcD*	>10x	
ACLS	ANA	*ilvN*	*ilvB*	>10x *	
ALCD2x	ANA	*adhE*	*adhP, frmA*	>150x	
FUM	AER	*fumB*	*fumA, fumC*	Strain specific	
LDH_D	AER	*ldh*	*dld*	1.4x	
PFK	AER	*pfkA*	*pfkB*	>4x	(Nakahigashi et al., 2009)
PYK	AER	*pykF*	*pykA*	1x	(Nakahigashi et al., 2009)

[a] Name of the reaction in the *iJO1366* model.

[b] Condition in which the isozyme serves a major or minor role (AER and ANA does not infer the opposite is true).

[c] Gene encoding the major isozyme of the reaction.

[d] Gene encoding the minor or lesser role of the reaction.

*ilvB expressed highly in DH5a

Supplemental Table 5: Comparison of best production strain for each compound produced via all heterologous pathways, Related to Figure 5.

Aerobic:

	W	W3110	Bl21(DE3)	C	Crooks	DH5a	MG1655	pathways
12ppd	0%	0%	0%	0%	75%	0%	0%	4
13ppd	0%	12%	0%	7%	2%	7%	34%	41
14btd	0%	0%	9%	0%	0%	20%	89%	44
2mbtoh	100%	0%	0%	0%	0%	0%	0%	2
2obut	20%	0%	0%	0%	0%	0%	100%	5
2opntn	0%	0%	0%	0%	0%	0%	0%	1
2phetoh	0%	0%	0%	0%	0%	100%	0%	5
2ppoh	0%	0%	0%	0%	14%	0%	0%	14
3hb	0%	0%	0%	0%	56%	0%	0%	9
3hpp	0%	0%	0%	4%	0%	80%	12%	25
3hpt	0%	0%	0%	0%	0%	0%	67%	3
3mbtoh	100%	0%	0%	0%	0%	0%	100%	2
3mob	100%	0%	0%	0%	0%	0%	0%	3
aa	0%	0%	10%	0%	0%	20%	30%	10
acryl	0%	4%	0%	0%	4%	13%	26%	23
btd	100%	0%	0%	0%	0%	0%	100%	2
btoh	0%	0%	28%	0%	0%	6%	39%	18
ghb	0%	0%	0%	0%	0%	100%	100%	2
ibutoh	85%	0%	0%	0%	0%	0%	23%	13
ppoh	0%	5%	0%	11%	11%	26%	74%	19

Anaerobic:

	Crooks	W3110	MG1655	W	C	Bl21(DE3)	DH5a	pathways
12ppd	0%	25%	0%	50%	0%	0%	0%	4
13ppd	0%	20%	29%	5%	10%	0%	0%	41
14btd	0%	0%	0%	0%	9%	45%	0%	44
2mbtoh	0%	0%	0%	100%	0%	0%	0%	2
2obut	0%	0%	0%	40%	20%	60%	0%	5
2opntn	0%	0%	100%	0%	0%	0%	0%	1
2phetoh	0%	0%	0%	0%	0%	0%	0%	5
2ppoh	14%	0%	0%	29%	50%	0%	21%	14
3hb	44%	0%	0%	11%	44%	0%	33%	9
3hpp	0%	0%	20%	40%	24%	4%	0%	25
3hpt	0%	67%	67%	0%	33%	0%	0%	3
3mbtoh	0%	0%	0%	100%	0%	0%	0%	2
3mob	0%	0%	0%	0%	0%	0%	100%	3
aa	0%	0%	0%	0%	20%	10%	20%	10
acryl	4%	9%	13%	30%	26%	9%	4%	23
btd	0%	0%	0%	0%	0%	0%	100%	2
btoh	0%	0%	39%	0%	17%	0%	0%	18
ghb	0%	0%	100%	0%	0%	0%	0%	2
ibutoh	0%	0%	23%	8%	0%	0%	0%	13
ppoh	0%	11%	26%	42%	21%	42%	0%	19

Supplemental Table 6: Strain-specific findings and key differences elucidated through this work, Related to Figure 1.

Finding	Omics data type(s) utilized
W has a maximum growth rate similar between aerobic and anaerobic conditions	P
W, C, and Crooks have significantly elevated uptake rates of glucose which give it immediate high productivity potential	P
BL21 *pgl* defect causes it to have a reduced production potential related to products requiring significant reducing equivalents	G, M
BL21 has disrupted regulation of the TCA cycle which provides a greater potential for production of products leading form the TCA cycle anaerobically	G, T
K-12 MG1655 has a relatively broad set of highly-expressed metabolic genes aerobically	T
C has a relatively narrow set of highly-expressed metabolic genes aerobically and anaerobically (count of HEGs is < μ-1*σ for HEGs in all strains)	T
K-12 MG1655 has the highest production potential for products examined aerobically	T, M
W has the highest production potential for products examined anaerobically	T, M
DH5a was found to be the preferred producer for a number of synthetic constructs	T, M

G- genomics, P – Phenomics, T – Transcriptomics, M – Genome-scale Modeling

Supplemental Table 7: Literature reported physiological rates for strains examined in this study, Related to Table 1.

Study PMID	*E. coli* Strain	GR	SUR	Substrate	OUR	ac	for	suc	eth	lac
21782859	K-12	0.6	NR	Glucose	NR	NR	NR	NR	NR	NR
21782859	W	0.9	NR	Glucose	NR	NR	NR	NR	NR	NR
23950949	BL21	0.73	NR	Glucose	NR	NR	NR	NR	NR	NR
PMC201879	W3110	0.68	10.5	Glucose	15	NR	NR	NR	NR	NR
PMC201879	W3110	0.43	18.5	Glucose	0	NR	NR	NR	NR	NR
15838044	BW25113	0.65	7.6	Glucose	NR	4.8	NR	NR	NR	NR
PMC3588905	MG1655	0.50	NR	Glucose	NR	NR	NR	NR	NR	NR
23064346	K1060	0.34	7.22	Glucose	NR	NR	NR	NR	NR	NR

Supplemental Data Files

Data S1: Genetic summaries, Related to Figure 1

A Microsoft Excel File containing genetic summary information:

- Tab 1: Resequencing analysis for *E. coli* BL21
- Tab 2: Resequencing analysis for *E. coli* K-12 MG1655
- Tab 3: Resequencing analysis for *E. coli* W
- Tab 4: Resequencing analysis for *E. coli* Crooks
- Tab 5: Resequencing analysis for *E. coli* K-12 W3110
- Tab 6: Resequencing analysis for *E. coli* DH5a
- Tab 7: Shared and unique orthologs between the strains and computation of core/pan genome
- Tab 8: Unique genes present in only one out of the seven strains analyzed
- Tab 9: Comparison of amino acid differences for all shared ORFs between the 7 strains. Differences are identified as mutations between the AA sequence of the K-12 MG1655 protein (WT)

Data S2: Model Contents and Abbreviations, Related to Figure 2

A Microsoft Excel File containing all reactions, genes and metabolites in each model along with their unique identifiers to other databases.

Data S3: Strain-specific Models in SBML format, Related to Figure 2

Strain specific models for all 7 strains including those that are not constrained using any data and those that are constrains with measured uptake rates in aerobic and anaerobic conditions.

Data S4: Theoretical Yields, Related to Figure 2

A Microsoft Excel File containing theoretical yields for native and heterologous pathways in both aerobic and anaerobic conditions

- Tab1: Production yields of all native compounds in aerobic conditions
- Tab 2: Production yields of all native compounds in anaerobic conditions
- Tab 3: Production yields of all heterologous pathways in aerobic conditions
- Tab 4: Histogram of production yields for all heterologous pathways in aerobic conditions
- Tab 5: Production yields of all heterologous pathways in anaerobic conditions
- Tab 6: Histogram of production yields for all heterologous pathways in anaerobic conditions
- Tab 7: Histogram for production yields for all heterologous pathways in aerobic and anaerobic conditions filtering out 0% and 100% yield pathways.

Data S5: Flux Variability Analysis, High Flux Reactions and Correlation with Physiological data, Related to Figure 2

A Microsoft Excel File containing High-flux reactions and highly expressed genes identified in all seven strains in aerobic and anaerobic conditions as well as their correlation with physiological data.

- Tab 1: High flux reactions in aerobic conditions
- Tab 2: High flux reactions in anaerobic conditions
- Tab 3: Highly expressed genes in aerobic conditions
- Tab 4: Highly expressed genes in anaerobic conditions
- Tab 5: Correlation values for aerobic fluxes and aerobic physiological rates
- Tab 6: Correlation values for anaerobic fluxes and measured anaerobic physiological rates
- Tab 7: Correlation values for aerobic expression values and measured aerobic physiological rates
- Tab 8: Correlation values for anaerobic expression values and measured anaerobic physiological rates

Data S6: RNA-seq Expression Values and Transcription Factor Enrichments, Related to Figure 3

A Microsoft Excel File containing triplicate normalized counts for RNA-seq reads in aerobic and anaerobic conditions as well transcription factor enrichment values for all transcription factors in all 7 strains in both aerobic (Tab 2) and anaerobic (Tab 3) conditions.

Data S7: Overlap of Highly Expressed Genes with High Flux Reactions and Flux Shift Predictions, Related to Figure 4

A Microsoft Excel File containing the overlapping genes and reactions that are both high flux and highly expression in aerobic and anaerobic conditions as well as the prediction of flux shifts and gene expression overlap going from aerobic to anaerobic conditions.

Data S8: Heterologous Pathways Scoring, Related to Figure 5

A folder of Microsoft Excel Files containing overlap scores for each heterologous pathway in each of the seven strains for both aerobic and anaerobic conditions.

Data S9: Synthetic Biology Construction Potentials, Related to Figure 5

A Microsoft Excel File containing the mean amino acid abundances in *E. coli* as well as abundances of amino acids in 3,982 synthetic biology constructs.