Metabolic stress constrains fermentative production of L-cysteine in Escherichia coli by accelerating transposition through mobile genetic elements in synthetic plasmid constructs

Kevin Heieck
Technical University of Munich
Nathanael David Arnold
Technical University of Munich
Thomas Bartholomäus Brück (brueck@tum.de)
Technical University of Munich

Research Article

Keywords: L-cysteine, Fermentation, E. coli, Metabolic stress, Insertion sequence elements, Transcriptome analysis, Plasmid deep sequencing, Minimal genome strain, Metabolic engineering

Posted Date: October 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2186912/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Microbial Cell Factories on January 16th, 2023. See the published version at https://doi.org/10.1186/s12934-023-02021-5.

Abstract

Background

L-cysteine is an essential chemical building block in the pharmaceutical-, cosmetic-, food and agricultural sector. Conventionally, L-cysteine production relies on the conversion of keratinous biomass mediated by hydrochloric acid. Today, fermentative production based on recombinant *E. coli*, where L-cysteine production is streamlined and facilitated by synthetic plasmid constructs, is an alternative process at industrial scale. However, space-time yields and process stability are still to be optimised for improved economic viability. We simulate an industrial fermentation process with *Escherichia coli* harbouring various L-cysteine production plasmid constructs.

Results

In a comparative experimental design, the*E. coli* K12 production strain W3110 and the reduced genome strain MDS42, almost free of insertion sequences, were used as hosts. Data indcates that W3110 populations acquire growth fitness at the expense of L-cysteine productivity within 60 generations, while production in MDS42 populations remains stable. For the first time, the negative impact of predominantly insertion sequence family 3 and 5 transposases on L-cysteine production is reported, by combining differential transcriptome analysis with NGS based deep plasmid sequencing. Furthermore, metabolic clustering of differentially expressed genes supports the hypothesis, that metabolic stress induces rapid propagation of plasmid rearrangements, leading to reduced L-cysteine yields in evolving populations over industrial fermentation time scales.

Conclusion

The results of this study implicate how selective deletion of insertion sequence families could be a new route for improving industrial L-cysteine or even general amino acid production using recombinant *E. coli* hosts. Instead of using minimal genome strains, a selective deletion of certain IS families could offer the benefits of adaptive laboratory evolution (ALE) while maintaining enhanced L-cysteine production stability.

Background

The amino acid L-cysteine, harbouring a thiol group, provides a high redox activity in cell metabolism, plays a crucial role in protein folding, functions as a catalytic residue of several enzymes and serves as a building block of 5-L-glutamyl-L-cysteinylglycine (GSH) and as a donor compound of sulphur, which is required for the synthesis of Fe/S clusters, biotin, coenzyme A and thiamine (1, 2).

Besides the essential function in metabolism, L-cysteine is also of considerable industrial importance, with applications ranging from pharmaceutical products and cosmetics over food production to feed additives in livestock farming.

Its market size is projected to reach a value of 683.5 million US-Dollars by 2028 at a compound annual growth rate (CAGR) of 6.4% during 2022–2028 with an annual global production volume of 14.000 t in 2015 (3). To date, the cheapest and thereby most prevalent means of L-cysteine production involves chemical hydrolysis of - and extraction from - keratinous biomass, such as feathers, pig bristles and animal hair by means of electrolysis (4). Up to 27 tons of hydrochloric acid are required to obtain 100 kg of a racemic mixture of cysteine from 1.000 kg raw material (4, 5). In order to circumvent negative impacts upon the environment associated with hydrochloric waste disposal, alternative technologies such as fermentation and enzymatic conversion have been explored and rapidly gained significance since their implementation. In 2004, 12% of the globally manufactured L-cysteine global originated from fermentation (6).

The enzymatic conversion of DL-2-amino-Δ2-thiazoline-4-carboxylic acid (D-ATC) to L-cysteine with *Pseudomonas* spp. derived enzymes is limited by product inhibition (7, 8). For biotechnological L-cysteine production, the bacteria *C. glutamicum, and E. coli* harbouring optimised plasmids represent the dominant expression organisms. Since yields from *C. glutamicum* are low (approx. 950 mg/L), *E. coli* is the preferred host for L-cysteine production by fermentation (9).

However, there are still major obstacles in upscaling fermentation processes with engineered microorganisms. The stability of strains with synthetic production is highly fragile and presents a challenge when implementing bioprocesses on a large scale (10, 11). Declining productivity affects the economic feasibility of fermentations over extended time periods.

Cells possess a finite pool of resources that are required for growth and homeostasis such as replication, transcription, translation, and numerous enzymatic reactions. Depending on signals from the environment and growth conditions, cells must economise on these resources to streamline their vitality and survival (12, 13). The introduction of designed plasmid constructs and the upregulation of the genetic elements for recombinant L-cysteine production pose a defiance to the tightly regulated homeostasis within host cells (14, 15). This metabolic load hinders the expression of other genes, thereby negatively affecting growth rate and promoting evolutionary pressure (16–18). In microorganisms, several concepts are reported that can lead to a selection advantage and thus to both phenotypic and genotypic variation within populations (19, 20). In bacteria, activation of mobile genetic elements, such as insertion sequences (IS) and corresponding transposons can lead to mutagenesis-based inactivation of synthetic constructs (21). In addition, expression of regulatory elements of the SOS response has been shown to increase the expression of error-prone DNA polymerases, which can indirectly induce further mutations in recombinant gene elements, such as plasmids (22, 23). These effects have a negative impact on the time dependent-productivity (space-time yield) and consequently the total yield of the biomolecule to be produced.

Fitness and productivity of a producing organism can be improved by several means, including rational metabolic engineering, adaptive laboratory evolution (ALE) (24, 25), the use of reduced and minimal genomes (26, 27), and the improvement of fermentation conditions (28). Metabolic engineering of *E. coli*

strains optimising L-cysteine production, targets the overexpression of specific bottleneck genes (Fig. 1). In order to capture 3-P-glycerate from glycolysis and feed it into the synthesis of the precursor amino acid L-serine to finally convert it to L-cysteine, the two feedback-resistant genes *serA* and *cysE* are overexpressed (29, 30). An L-cysteine production pathway uncoupled from glycolysis involves the assimilatory reduction of sulphate. With the expression of *cysM*, assimilated thiosulphate is converted to L-cysteine via an intermediate step. Since large amounts of L-cysteine have an inhibitory effect on *E. coli* cells or are even toxic, the overexpression of an L-cysteine efflux gene (*eamA*) is essential (31).

With the advent of high-resolution omics technologies, system-wide characterization of metabolic stress during recombinant production was facilitated. Despite the necessity of establishing these technologies, high costs, methodological effort as well as the required bioinformatic knowhow still constrains the wide-spread applicability of these technologies (28).

This study investigates the genotypic and phenotypic characteristics of engineered L-cysteine producing *E. coli* strains by simulating an industrial fermentation process. Within 60 generations, L-cysteine productivity was observed to collapse by up to 85%. Hence, parallelly to the *E. coli* K-12 production strain W3110, a reduced genome strain almost free of any IS (MDS42), was selected as a control which showed stable L-cysteine productivity and growth fitness throughout the whole simulation. For the first time, using both comparative RNA- and plasmid deep sequencing on early and late populations, strong differences in L-cysteine and sulphur metabolism were uncovered. Moreover, predominantly IS3 and 5 family transposases appear to induce rapid plasmid rearrangements, which we investigated in an industrially relevant *E. coli* L-cysteine production system. Both observations, presumably in combination, disrupt L-cysteine production in late W3110 populations.

Results And Discussion

Stability of L-cysteine-producing phenotypes

To study phenotypic and genetic diversities in L-cysteine producing *E. coli* populations over timescales relevant on industrial levels, we simulated a gradual scale-up growth process. To identify potential effects of transposable elements that disrupt synthetic constructs, we cultivated the minimal genome strain MDS42, almost free of any insertion sequences (IS), alongside the traditional K-12 strain W3110. Specifically, we cultivated three different clones of two *E. coli* strains, each with an L-cysteine producing plasmid (Fig. 2A). In order to prevent plasmid loss, we kept the cultures at a constant antibiotic selection. By serially transferring the W3110 production strains every 10 h and the MDS42 production strains every 5 h (due to its higher growth rate), we achieved populations > 60 cell generations (Supplementary Table 2). Hence, an exponential growth phase was guaranteed at all times throughout the experiment. At each transfer step, we sampled the growing populations and immediately cryopreserved them, in order to study time dependent phenotypic and genetic variability.

L-cysteine titres of each sampled population were determined throughout the simulated fermentation. While an initial increase in relative L-cysteine titres of transformed W3110 strains towards a maximum was measured, the titres at 60 accumulated generations only reached 15–35% of the maximal product concentration. The cysteine titres of the W3110 strain with integrated pCYS_i plasmid remained at a constant high level until a drop at 63 generations, while the other plasmids showed a steady decline from generation 15–20 onwards (Fig. 2B). Transformed MDS42 strains showed no steady decline in generated L-cysteine titres with accumulating generations. Instead, the L-cysteine space time yields were stable at 75–100% (Fig. 2D). However, total L-cysteine titres in MDS42 populations were overall lower compared to W3110 populations (Supplementary Fig. 1).

Furthermore, we tested strain viability by determining growth rates of the sampled populations. As a function of generation numbers, growth rates of W3110 differed depending on the transformed plasmid (Fig. 2C). W3110 populations with integrated pCYS_i plasmid showed no considerable changes in growth rate. In contrast, we observed a substantial change of fitness in case of W3110 with integrated pCYS_m and pCYS. When comparing initial and final growth rates, we detected an increase of 13% for pCYS_m, and 27% for pCYS. There was no considerable effect on fitness in transformed MDS42 populations (Fig. 2D).

Overall, the decline in L-cysteine titres tended to correlate with the increase in growth rates in an inversely proportional pattern, indicating an acquisition of fitness at the expense of L-cysteine production.

Transcriptome analyses reveals strong differences regarding expression levels in W3110 populations and few differences in MDS42 populations over the course of simulated fermentation

To determine potential causes for the collapse of L-cysteine production, we examined the transcriptomes of W3110 and MDS42 populations subjected to the simulated fermentation. We sequenced and compared transcriptomes from early generation populations (EGP) with later generation populations (LGP). Interestingly, when assessing global relationships between samples, we found that EGPs and LGPs of the MDS42 strains hardly differed from each other, while there were larger deviations in W3110 populations (Fig. 3). This result is consistent with previous findings of a limited evolutionary adaptability of the MDS42 strain and illustrates why we observed negligible changes in L-cysteine production and fitness in MDS42 populations (32, 33).

Metabolic clustering helps understanding the burden on cell populations during L-cysteine production

To gain a deeper insight into the differences in expression levels in W3110 and MDS42 populations, we examined differentially expressed genes (DEGs) with a p-value cut-off < 0.05 (Supplementary tables 5–

10). After clustering the DEGs according to their metabolic function, we subsequently plotted the number of DEGs in a cluster against the mean fold change of those in the cluster (Fig. 4). Overall, we noticed a lower number of different metabolic clusters in MDS42 populations than in W3110 populations. In general, stress features such as nitrogen starvation, acetyl-coA-, nitrate-, and carbon assimilation were up-regulated in EGPs, indicating a resource-draining L-cysteine production process. However, most up-regulated features in EGPs, where L-cysteine production was highest, belonged to the cluster of sulphur and L-cysteine starvation (Table 1). The metabolic flux towards L-cysteine coupled with overexpression of its corresponding exporter gene *eamA* most likely caused the intracellular deficiency of sulphur and the sulphuric amino acid.

In LGPs, additionally to up-regulated clusters such as biofilm formation, iron starvation and anaerobic respiration, we also identified features belonging to an L-cysteine degradation operon (Table 2). The corresponding D/L-serine membrane transporter CyuP was potentially protecting the cell from increasing sulphur starvation by shuttling serine out of the cell before this precursor amino acid was further metabolised to L-cysteine. CyuA, on the other hand, may have acted as an L-cysteine desulfidase releasing sulphur by degradation. The observation of the decrease in measured L-cysteine titres in LGPs supports this hypothesis. Furthermore, we suspected that the expression of the *cyuAP* operon was triggered by sulphur starvation.

In the further course of the transcriptome analysis, we found clusters that indicated genetic instability of the production strains. Transposases for insertion sequence families 3 and 66 as well as mismatch repair genes were upregulated in LGPs (Table 2). In the sulphur and L-cysteine starvation cluster of EGPs we detected *cysD*, a gene that facilitates stress-induced mutagenesis (SIM) (34). Therefore, we examined the impact of this genetic instability on the plasmid level.

Table 1
List of sulphur and L-cysteine metabolism related operons detected as clusters in transcriptome analysis.

DEGs of sulphur/ L-cysteine metabolism related operons	Metabolic function	Literature
cyuAP	L-cysteine desulfidase, L-cysteine utilization permease	(35), (36)
aslB	Putative anaerobic sulfatase maturation enzyme	(37)
tauABCD	Taurine utilization proteins	(38)
ssuEADCB	Aliphatic sulfonates utilization proteins	(39)
cysPUWA, sbp	Sulphate/thiosulphate transport proteins	(40), (41)
cysDNC	Sulphate activation proteins	(42)
cysJIH	Sulphite reductase proteins	(43)
tsuAB	Thiosulphate transport proteins	(44)

Table 2
List of genes related to genetic instability detected in transcriptome analysis.

DEGs belonging to genetic instability	Metabolic function	Literature
insJK	IS3 family transposase	(45)
yjgZ	Putative IS66 family transposase	InterPro
ybcN, rusA, iprA	DNA repair	(46), (47), (48)

Deep sequencing reveals an accumulation of insertion sequences (IS) in production plasmids of evolving populations

After extraction and deep sequencing of plasmids from EGPs and LPGs, we mapped the sequenced reads against the associated plasmid sequence reference (Methods). The average per base coverage depth was > 140.000x (Supplementary Figs. 2–4). Consistent with other studies, we did not find genetic variance in production plasmids by single nucleotide polymorphism (SNP) analyses (data not shown) (20, 21). Interestingly, we noticed an increase in unmapped reads in LGPs of W3110 compared to EGPs (Fig. 5A). In a next step, we mapped these reads against an insertion sequence database from *E. coli* revealing higher coverages of aligned reads to IS in LGPs (Fig. 5A). Surprisingly, we also identified reads, that could

be mapped to IS in plasmids from MDS42 populations, although the number was reduced by 10-fold. In order to assign the mapped reads to the different IS families, we used NCBI's megablast algorithm (Methods).

The distribution of the different IS families was very similar for all three production plasmids. This observation was the same for EGPs compared to LGPs (Fig. 5B). Nevertheless, the total number of reads that could be mapped to IS doubled on average when comparing W3110 EGPs with LGPs, while the number remained the same for MDS42 populations. This indicates, that transpositions of different IS families propagated in similar frequencies. For plasmids extracted from W3110 populations, we most frequently identified reads that could be mapped to the IS3 and IS5 family, which supports the result of a high expression of *insJK*, an IS3 family transposase, in LGPs of W3110_pCYS. In addition, reads mapped to the ISAs1 family were very abundant. Among plasmids from MDS42 populations, the reads could be mapped mainly to sequences belonging to IS200- and IS110 families. Due to deletion of most IS, the few transposition events in MDS42 populations are in line with the stable L-cysteine production and growth rates, rendering *E. coli* MDS42 as a stable host for industrial L-cysteine fermentations.

Conclusions

The data in this study suggest that insertion sequence (IS) transposition, triggered by the metabolic stress of L-cysteine production, leads to structural genetic rearrangements in production plasmids, resulting in a collapse of L-cysteine production capacities. L-cysteine titres in W3110 populations decreased by 65–85% within 60 generations while growth rates increased up to 27% depending on the L-cysteine plasmid construct. Contrarily, MDS42 populations exhibited nearly stable space-time yields of 75–100% with no considerable effect on growth related fitness.

Transcriptome based metabolic clustering revealed upregulation of general stress response genes, indicating a heavy burden on cells. A sulphur starvation response most likely triggered L-cysteine degradation in LGPs and promoted stress-induced mutagenesis. The upregulation of insertion sequence families 3 and 66 transposases as well as mismatch repair genes revealed genetic instability of L-cysteine producing phenotypes.

With a deep sequencing approach of plasmid constructs, an accumulation of predominantly IS3 and IS5 could be detected in W3110 LGPs, whereas IS accumulation was neglectable in MDS42 populations. This result is consistent with other reports where transposition of IS and the associated genetic instability has an inhibitory effect on large-scale biomolecule production (21, 49). The negative impact of IS elements on plasmid driven L-cysteine production was here experimentally demonstrated for the first time.

Consequently, the genetically stable minimal genome *E. coli* strain MDS42 rendered as a promising host for L-cysteine production. However, MDS42, while showing a superior growth rate and stability compared to W3110, exhibits an overall lower L-cysteine space-time yield (Supplementary Fig. 1). Therefore, a more

targeted approach would be the selective deletion of predominant IS3 and 5 with its corresponding transposases, which we have demonstrated to be putatively responsible for disruption of production plasmid structure. This would potentially provide an off-trade between increased plasmid stability over the entire production phase while maintaining evolutionary adaptability of the cell system linked to high cell density fermentations at maximal growth rates and L-cysteine productivity. This study postulates, that selective deletion of IS and the corresponding transposases in genomes of industrially relevant production strains, such as W3110 can be used as a new approach to optimise the *E. coli* based industrial L-cysteine production. The aim here is always the optimization of space-time and total L-cysteine yields to minimise energy expenditure in large scale, batch industrial (100–300 m^3) fermentations, where particularly cost of oxygen input by mixing the fermentation broth is a major cost effector for the entire process.

Materials And Methods

Strains and plasmids

Two parental *E. coli* K12 strains were used to construct the strains with three specific plasmids (Table 3):

E. coli W3110: *F- λ- rph-1 INV(rrnD, rrnE)*

E. coli MDS42 (Scarab Genomics): MG1655 genome almost free from any ISs, *fhuACDB*, *endA* and more (48).

For introduction of the plasmids into the strains, standard transformations with chemical competent cells were performed. For the fermentation simulation experiments, single colonies of freshly transformed cells were used to inoculate precultures.

Table 3
Plasmids used in this study

Plasmid	Relevant features	References
pCYS	p_{GAPDH}:*ydeD*, $p_{serAp1,2}$:*serA317*, p_{cysE}: *cysE-XIV*, tet^R, p15A	(29, 30, 50, 51)
pCYS_i	p_{GAPDH}:*cysE-XIV-ydeD-serA317*, tet^R p15A	This study
pCYS_m	p_{GAPDH}:*ydeD*, $p_{serAp1,2}$:*serA317*, p_{fic}:*cysM*, p_{cysE}: *cysE-XIV*, tet^R, p15A	This study

PCYS_i was generated by ligation of four amplified and digested amplicons in an equimolar ratio in a 40 µl reaction. The fragments contained the metabolic pathway genes of pCYS (*ydeD*, *cysE-XIV* and *serA317*) and the empty vector as a backbone. The digestions were conducted with SacI, XhoI, BamHI, PacI, and NcoI. Fast digest restriction enzymes, buffer, the T4 DNA Ligase as well as the Phusion DNA polymerase were used from Thermo Fisher Scientific. For amplification of different fragments, specific

PCR primer were used (Supplementary Table 1). PCYS_m was generated by introducing the stationary phase promoter *pfic* and the *cysM* gene into the backbone of pCYS via Gibson cloning standard procedure (NEB). PCRs were performed with specifically designed primers (Supplementary Table 1).

Media

A minimal medium (MM) adapted to the production of L-cysteine was used for all cultivations. This MM consisted of 10 g/L glucose 5 g/L KH_2PO_4, 5g/L $(NH_4)_2SO_4$, 1 g/L Na_3Citrat x 2 H_2O, 0.9 g/L L-isoleucine, 0.6 g/L D, L-methionine 0.5 g/L NaCl, 2 g/L ammonium thiosulphate, 1.2 g/L $MgSO_4$ x 7 H_2O, 18 mg/L thiamine-HCl, 9 mg/L pyridoxine-HCl, 15 mg/L tetracycline and 100 ml/L LB-medium. Additionally, 10 ml/L of a trace element solution was added. This trace element solution included 3.75 g/L H_3BO_4, 1.55 g/L $CoCl_2$ x 6 H_2O, 0.55 g/L $CuSO_4$ x 5 H_2O, 3.55 g/L $MnCl_2$ x 4 H_2O, 0.65 g/L $ZnSO_4$ x 7 H_2O, and 0.33 g/L Na_2MoO_4 x 2 H_2O. The trace element solution is adjusted to pH 4.0 with HCl before autoclaving. Everything else got sterile filtrated after pH adjustment to 7-7.05. Tetracycline, vitamins, $CaCl_2$ and $MgSO_4$ were added immediately before inoculation.

Simulated long-term fermentation

Single colonies of freshly transformed cells were inoculated into aerated cultures with 25 ml medium and horizontal shaking of 150 r.p.m (New Brunswick Innova 44). W3110 strains were cultivated at 32°C for 10 h. Due to a higher growth rate of the MDS42 strain, cultures were kept at 32°C just for 5 h. After each time point cultures were inoculated into 25 ml fresh medium (starting OD600 = 0.05) and incubated under the same conditions for another 10 h and 5 h. At each passage, sample's OD_{600} were recorded to determine the accumulated generations (Supplementary Table 2) and 1 ml got snap-frozen with 1ml 50% glycerol in liquid N2 and stored at -80°C. Each strain harbouring one out of three plasmids got cultivated in triplicates.

Determination of L-cysteine titre by spectrophotometry

Each population sample from the glycerol stock was used for inoculation of 25 ml medium and the culture got cultivated at 32°C at 150 r.p.m for 72 h (New Brunswick Innova 44). After incubation, 1 ml culture was centrifuged at 15.000 x g for 1 min. Both the pellet and the supernatant were further treated according to a modified method of Gaitonde to determine the L-cysteine titres (Supplementary Note 1) (52).

Measurement of population growth rates

In order to measure population growth rates, the OD_{600} values of cultures grown for L-cysteine productivity analysis (as described in the previous section) were recorded every hour. Un-inoculated minimal medium was used for background subtraction. Calculation of growth rates were performed with the following formula:

$$r = \frac{N(t)^{\frac{1}{t}}}{N(0)} - 1$$

where: N(t): cell number at time t, N(0): cell number at time 0, r: growth rate and t: time passed.

RNA sequencing and analysis

RNA extractions were performed from exponentially growing cultures, which were also grown for growth rate and cysteine titre determination (as described in previous sections) using a standard RNA extraction kit (Promega SV total RNA isolation system). Depletion of ribosomal RNA, transcriptome library construction, and sequencing of the corresponding samples were performed by Eurofins Genomics. Sequencing was conducted with the technology of Illumina NovaSeq 6000 with a 150 bp paired-end reading. Raw read counts were created using featureCounts (version 1.5.1) (53). Only reads overlapping “CDS” features were counted. All reads mapping to features with the same meta-feature attribute were summed. Only reads with unique mapping positions and a mapping quality score of at least 10 were considered for read counting. Supplementary alignments were ignored. Instead of counting paired-end reads twice, they were counted only once, i.e. as single fragment. Reads mapping to multiple features were assigned to the feature that has the largest number of overlapping bases. A Trimmed Mean of M-values (TMM) normalization was performed using the edgeR package (version 3.16.5) (54, 55). Mapping of reads to reference sequences of *E. coli* K-12 W3110 and *E. coli* K-12 MDS42 was performed using BWA-NEM (version 0.7.12-r1039).

To determine which features are significantly differentially expressed, a small threshold ($<<0.01$) was applied to the false discovery rate (FDR) values, which is the p-value adjusted for multiple testing using Benjamini-Hochberg procedure. Fold changes were calculated by dividing values of the later generation population (LGP) by values of the early generation population (EGP).

The metabolic gene clustering was carried out with the EcoCyc *E. coli* database regarding all differentially expressed features with p-values < 0.05. Mapping results and expression profiling statistics are shown in supplementary tables 3–4.

Plasmid analysis of early and late generation population of E. coli **W3110 and MDS42**

The Extraction of plasmids was performed using a standard kit (ThermoFisher Scientific GeneJET Plasmid Miniprep Kit). The subsequent library preparation and deep sequencing was performed by Eurofins Genomics for Illumina NovaSeq 6000 S4 paired-end 2 x 150 bp. The per base coverage depth was > 140.000x. Adapter trimming, quality filtering and per-read pruning was done to retain only high quality bases. Reads were then mapped to the corresponding reference plasmid sequence, followed by a single nucleotide variant calling. Reads which could not be mapped to the plasmid reference sequences, got then mapped against an insertion sequence database (ISfinder_Nucl) with the Burrows-Wheeler Aligner. The reads, which could then be mapped to insertion sequences, were blasted against each

insertion sequence family in the *E. coli* W3110 genome with NCBI's megablast algorithm. Only highly similar sequences were selected and displayed.

Abbreviations

ALE: Adaptive laboratory evolution

C. Glutamicum: Corynebacterium glutamicum

DEG: Differentially expressed gene

E. coli: Escherichia coli

EGP: Early generation population

IS: Insertion sequence

LGP: Late generation population

NGS: Next generation sequencing

R.P.M: Rounds per minute

SNP: Single-nucleotide polymorphism

Declarations

Author Contributions

TB and KH conceived and designed the study. KH carried out the experiments and drafted the manuscript. KH, NA and TB edited and finalised the manuscript.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data and materials are available as described in the study and its additional supplementary file.

Competing interests

There are no conflicts to declare

Funding

The study was financed by the industrial contract no #2018112610005964.

Author Contributions

TB and KH conceived and designed the study. KH carried out the experiments and drafted the manuscript. KH, NA and TB edited and finalised the manuscript.

Acknowledgements

Not applicable

References

1. Cicchillo RM, Baker MA, Schnitzer EJ, Newman EB, Krebs C, Booker SJ. Escherichia coli L-serine deaminase requires a [4Fe-4S] cluster in catalysis. J Biol Chem. 2004;279(31):32418-25.
2. Johnson DC, Dean DR, Smith AD, Johnson MK. Structure, function, and formation of biological iron-sulfur clusters. Annu Rev Biochem. 2005;74:247-81.
3. L-Cysteine Market Share and Size Forecast 2022-2028 Comprehenisve Research by Future Growth, Business Demand, Opportunities and Challenges. 2022.
4. Ismail NI. Production of Cysteine: Approaches, Challenges and Potential Solution. International Journal of Biotechnology for Wellness Industries. 2014;3:95-101.
5. Renneberg R. High grade cysteine no longer has to be extracted from hair. Biotechnology for beginners: Academic Press: Amsterdam; 2008. p. 132 pp.
6. Berehoiu RMT, Popa CN, Popescu S. Assessment of the E 920 additive (L-cysteine) in relation to some problems of modern food industry. Economic Engineering in Agriculture and Rural Development. 2013;13(1).
7. Shiba T, Takeda K, Yajima M, Tadano M. Genes from Pseudomonas sp. strain BS involved in the conversion of L-2-amino-Delta(2)-thiazolin-4-carbonic acid to L-cysteine. Appl Environ Microbiol. 2002;68(5):2179-87.
8. K.H N. Effects of Anoxic Conditions on the Enzymatic Conversion of D,L-2-Amino-Thiazoline-4-Carboxylic Acid to L-Cystine. Acta Biotechnology. 1997;17:185-93.
9. Wei L, Wang H, Xu N, Zhou W, Ju J, Liu J, et al. Metabolic engineering of Corynebacterium glutamicum for L-cysteine production. Appl Microbiol Biotechnol. 2019;103(3):1325-38.
10. Nielsen J, Keasling JD. Engineering Cellular Metabolism. Cell. 2016;164(6):1185-97.

11. Borkowski O, Ceroni F, Stan GB, Ellis T. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. Curr Opin Microbiol. 2016;33:123-30.
12. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. Proc Natl Acad Sci U S A. 2014;111(23):8488-93.
13. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R. Proteome reallocation in Escherichia coli with increasing specific growth rate. Mol Biosyst. 2015;11(4):1184-93.
14. Carbonell-Ballestero M, Garcia-Ramallo E, Montanez R, Rodriguez-Caso C, Macia J. Dealing with the genetic load in bacterial synthetic biology circuits: convergences with the Ohm's law. Nucleic Acids Res. 2016;44(1):496-507.
15. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015;112(51):15690-5.
16. Ceroni F, Algar R, Stan GB, Ellis T. Quantifying cellular capacity identifies gene expression designs with reduced burden. Nat Methods. 2015;12(5):415-8.
17. Glick BR. Metabolic load and Heterologous Gene Expression. Biotechnology Advanced. 1995;13:247-61.
18. Matthew Scott CWG, Eduard M. Mateescu, Zhongge Zhang, Terence Hwa. Interdependence of Cell growth and Gene Expression, Origins and Consequences. Science. 2010;330:1099-102.
19. Susann Müller HHaTB. Origin and analysis of microbial population heterogeneity in bioprocesses. Analytical Biotechnology. 2010;21:100-13.
20. Xiao Y, Bowen CH, Liu D, Zhang F. Exploiting nongenetic cell-to-cell variation for enhanced biosynthesis. Nat Chem Biol. 2016;12(5):339-44.
21. Rugbjerg P, Myling-Petersen N, Porse A, Sarup-Lytzen K, Sommer MOA. Diverse genetic error modes constrain large-scale bio-based production. Nat Commun. 2018;9(1):787.
22. Podlesek Z, Zgur Bertok D. The DNA Damage Inducible SOS Response Is a Key Player in the Generation of Bacterial Persister Cells and Population Wide Tolerance. Front Microbiol. 2020;11:1785.
23. R. Napolitano RJ-B, J.Wagner and R.P.P.Fuchs. All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. The EMBO Journal. 2000;19:6259-65.
24. Conrad TM, Lewis NE, Palsson BO. Microbial laboratory evolution in the era of genome-scale science. Mol Syst Biol. 2011;7:509.
25. Shi A, Fan F, Broach JR. Microbial adaptive evolution. J Ind Microbiol Biotechnol. 2022;49(2).
26. Lee JH, Sung BH, Kim MS, Blattner FR, Yoon BH, Kim JH, et al. Metabolic engineering of a reduced-genome strain of Escherichia coli for L-threonine production. Microb Cell Fact. 2009;8:2.
27. Tarnopol RL, Bowden S, Hinkle K, Balakrishnan K, Nishii A, Kaczmarek CJ, et al. Lessons from a Minimal Genome: What Are the Essential Organizing Principles of a Cell Built from Scratch? Chembiochem. 2019;20(20):2535-45.

28. Carneiro S, Ferreira EC, Rocha I. Metabolic responses to recombinant bioprocesses in Escherichia coli. J Biotechnol. 2013;164(3):396-408.
29. al. Le, inventor; Consortium für elektrochemische Industrie GmbH, assignee. Process for preparing O-Acetylserine, L-cysteine and L-cysteine-related products. USA2001.
30. Bell JK, Pease PJ, Bell JE, Grant GA, Banaszak LJ. De-regulation of D-3-phosphoglycerate dehydrogenase by domain removal. Eur J Biochem. 2002;269(17):4176-84.
31. Daßler T. Identification of a major facilitator protein from Escherichia coli involved in efflux of metabolites of the cysteine pathway. Molecular Microbiology. 2000;36:1101-12.
32. Umenhoffer K. Reduced evolvability of Escherichia coli MDS42, an IS-less cellular chasis for molecular and synthetic biology applications. Microb Cell Fact. 2010;9.
33. Csorgo B, Feher T, Timar E, Blattner FR, Posfai G. Low-mutation-rate, reduced-genome Escherichia coli: an improved host for faithful maintenance of engineered genetic constructs. Microb Cell Fact. 2012;11:11.
34. Mamum AAMA. Identity and function of a large gene network underlyiing mutagenic repair of DNA breaks. Science Reports. 2012;338:1344-8.
35. Loddeke M, Schneider B, Oguri T, Mehta I, Xuan Z, Reitzer L. Anaerobic Cysteine Degradation and Potential Metabolic Coordination in Salmonella enterica and Escherichia coli. J Bacteriol. 2017;199(16).
36. Nonaka G, Takumi K. Cysteine degradation gene yhaM, encoding cysteine desulfidase, serves as a genetic engineering target to improve cysteine production in Escherichia coli. AMB Express. 2017;7(1):90.
37. Berteau O. A New Type of Bacterial Sulfatase Reveals a Novel Maturation Pathway in Prokaryotes. The Journal of biological chemistry. 2006;281:22464-70.
38. Ploeg JRvd. Identification of Sulfate Starvation-Regulated Genes in Escherichia coli, a Gene Cluster Involved in the Utilization of Taurine as a Sulfur Source. Journal of bacteriology. 1996:5438-46.
39. van Der Ploeg JR, Iwanicka-Nowicka R, Bykowski T, Hryniewicz MM, Leisinger T. The Escherichia coli ssuEADCB gene cluster is required for the utilization of sulfur from aliphatic sulfonates and is regulated by the transcriptional activator Cbl. J Biol Chem. 1999;274(41):29358-65.
40. Sirko A. Sulfate and Thiosulfate Transport in Escherichia coli K-12: Nucleotide Sequence and Expression of the CysTWAM Gene Cluster. Journal of bacteriology. 1990:3351-7.
41. Kertesz MA. Proteins Induced by Sulfate Limitation in Escherichia coli, Pseudomonas putida, or Staphylococcus aureus. Journal of bacteriology. 1993:1187-990.
42. Leyh TS, Vogt TF, Suo Y. The DNA sequence of the sulfate activation locus from Escherichia coli K-12. Journal of Biological Chemistry. 1992;267(15):10405-10.
43. Ostrowski J, Wu JY, Rueger DC, Miller BE, Siegel LM, Kredich NM. Characterization of the cysJIH Regions of Salmonella typhimurium and Escherichia coli B. Journal of Biological Chemistry. 1989;264(26):15726-37.

44. Tanaka Y. Crystal structure of a YeeE-YedE family protein engaged in thiosulfate uptake. Science Advances. 2020;6.
45. Haas M, Rak B. Escherichia coli insertion sequence IS150: transposition via circular and linear intermediates. J Bacteriol. 2002;184(21):5833-41.
46. Chen CS, Korobkova E, Chen H, Zhu J, Jian X, Tao SC, et al. A proteome chip approach reveals new DNA damage recognition activities in Escherichia coli. Nat Methods. 2008;5(1):69-74.
47. Kumar A, Beloglazova N, Bundalovic-Torma C, Phanse S, Deineko V, Gagarinova A, et al. Conditional Epistatic Interaction Maps Reveal Global Functional Rewiring of Genome Integrity Pathways in Escherichia coli. Cell Rep. 2016;14(3):648-61.
48. Posfai G, Plunkett G, 3rd, Feher T, Frisch D, Keil GM, Umenhoffer K, et al. Emergent properties of reduced-genome Escherichia coli. Science. 2006;312(5776):1044-6.
49. Park MK, Lee SH, Yang KS, Jung SC, Lee JH, Kim SC. Enhancing recombinant protein production with an Escherichia coli host strain lacking insertion sequences. Appl Microbiol Biotechnol. 2014;98(15):6701-13.
50. al. We, inventor; Consortium für Elektrochemische Industrie GmbH, assignee. Microorganisms and processes for the fermentative preparation of L-cysteine, L-cystine, N-acetylserine or Thiazolidine derivates. USA1999.
51. Tobias Daßler TM, Christoph Winterhalter and August Böck. Identification of a major facilitator protein from Escherichia coli involved in efflux of metabolites of the cysteine pathway. Molecular Microbiology. 2000;36:1101-12.
52. Gaitonde MK. A spectrophotometric method for the direct determination of Cysteine in the presence of other naturally occuring amino acids. Biochem J. 1967;104:627-33.
53. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
54. Robsinson MD. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11.
55. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

Figures

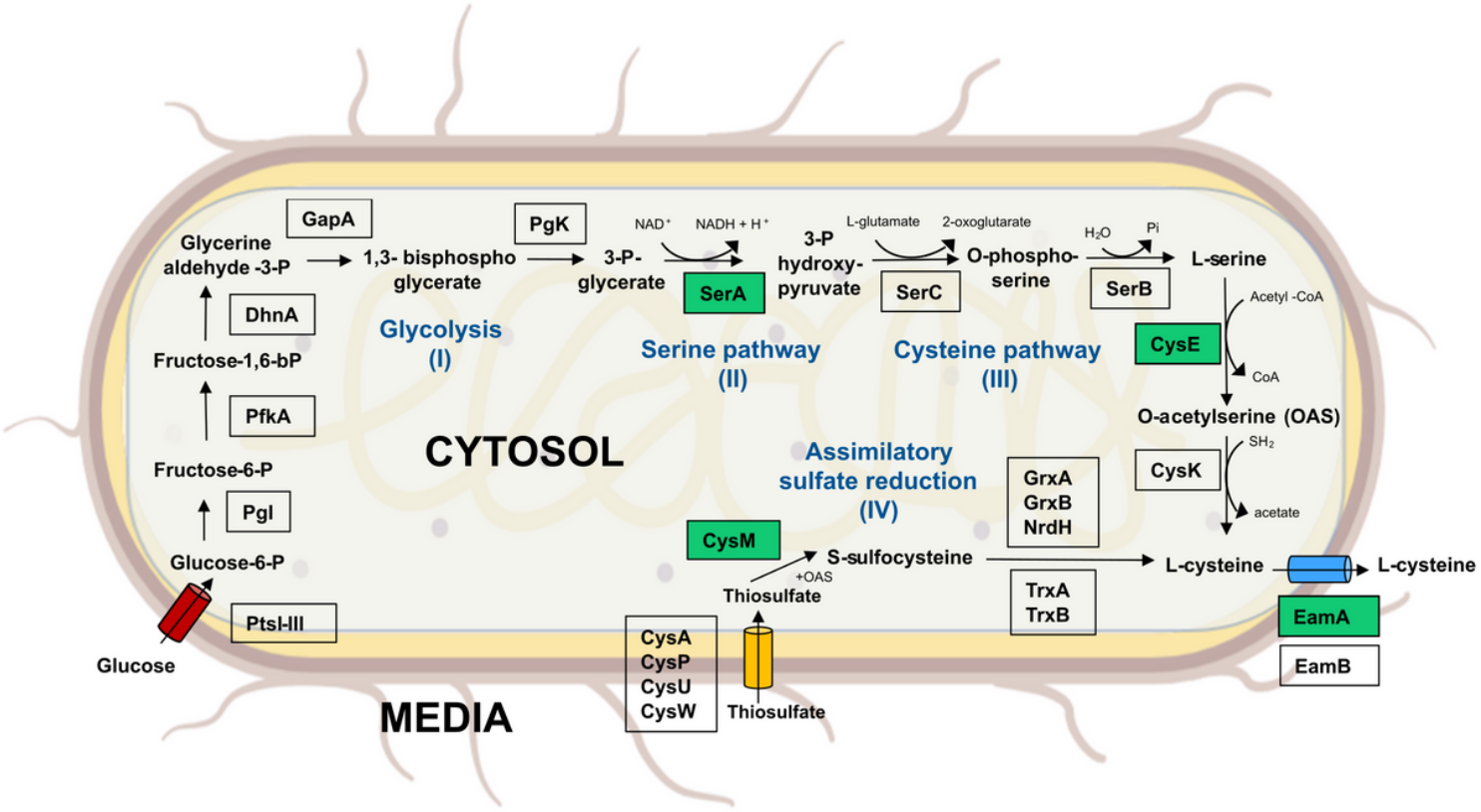

Figure 1

Metabolic engineering strategies for L-cysteine biosynthesis in *Escherichia coli*. Through glycolysis, the essential intermediate 3-phopsho-glycerate is produced (I), which is incorporated into L-serine biosynthesis (II). As a precursor amino acid, L-serine is used for the formation of L-cysteine (III). Alternatively, L-cysteine can be synthesised by the assimilatory sulphate reduction pathway (IV). Finally, L-cysteine is transported out of the cell. Corresponding proteins for conversion of substrates are shown in black boxes. Boxes with green background represent expressed proteins within the synthetic plasmid constructs in this work. PtsI-III: Phosphotransferase system I-III; PgI: Glucose-6-phosphate isomerase; PfkA: Phosphofructokinase 1; DhnA: Fructose-bisphosphate aldolase; GapA: Glyceraldehyde-3-phosphate dehydrogenase A; PgK: Phosphoglycerate kinase; SerA: D-3-phosphoglycerate dehydrogenase; SerC: Phosphoserine aminotransferase; SerB: Phosphoserine phosphatase; CysE: Serine acetyltransferase; CysK: Cysteine synthase A; CysA,P,U,W: ATP-dependent sulphate/thiosulphate uptake system; CysM: Cysteine synthase B; GrxA,B: Glutaredoxin 1,2; NrdH: Glutaredoxin-like protein; TrxAB: Thioredoxin 1,2; EamA,B: Cysteine exporter

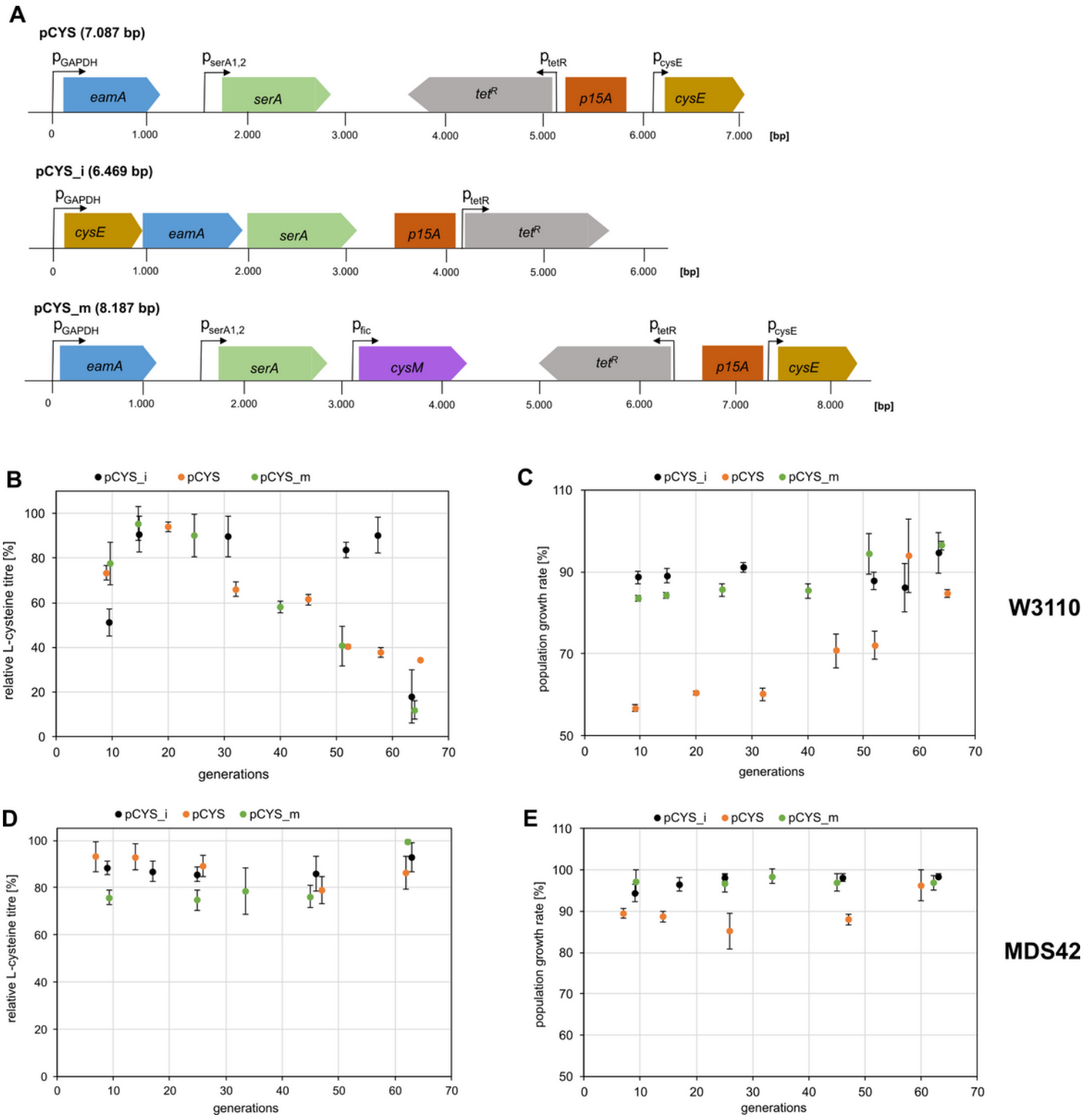

Figure 2

Stability of L-cysteine producing phenotypes. A: Plasmids used for transformation in *E. coli* W3110 and MDS42 to study the stability of L-cysteine production. Boxed arrows indicate genes located on the plasmid, whereas bended arrows display corresponding promoters. **B-E:** Plots of relative L-cysteine titres in % (B+D) and the population growth rates (C+E) as a function of the number of accumulated generations. Cultivation and subsequent L-cysteine titre measurement was carried out in biological

triplicates of W3110 and MDS42 with the three different plasmids shown in Fig. 1A. Relative L-cysteine titres were normalised based on the cell number and the highest titre of the corresponding biological replicate. Population growth rates were normalised based on growth rates of the corresponding non-producing strains harbouring empty vectors.

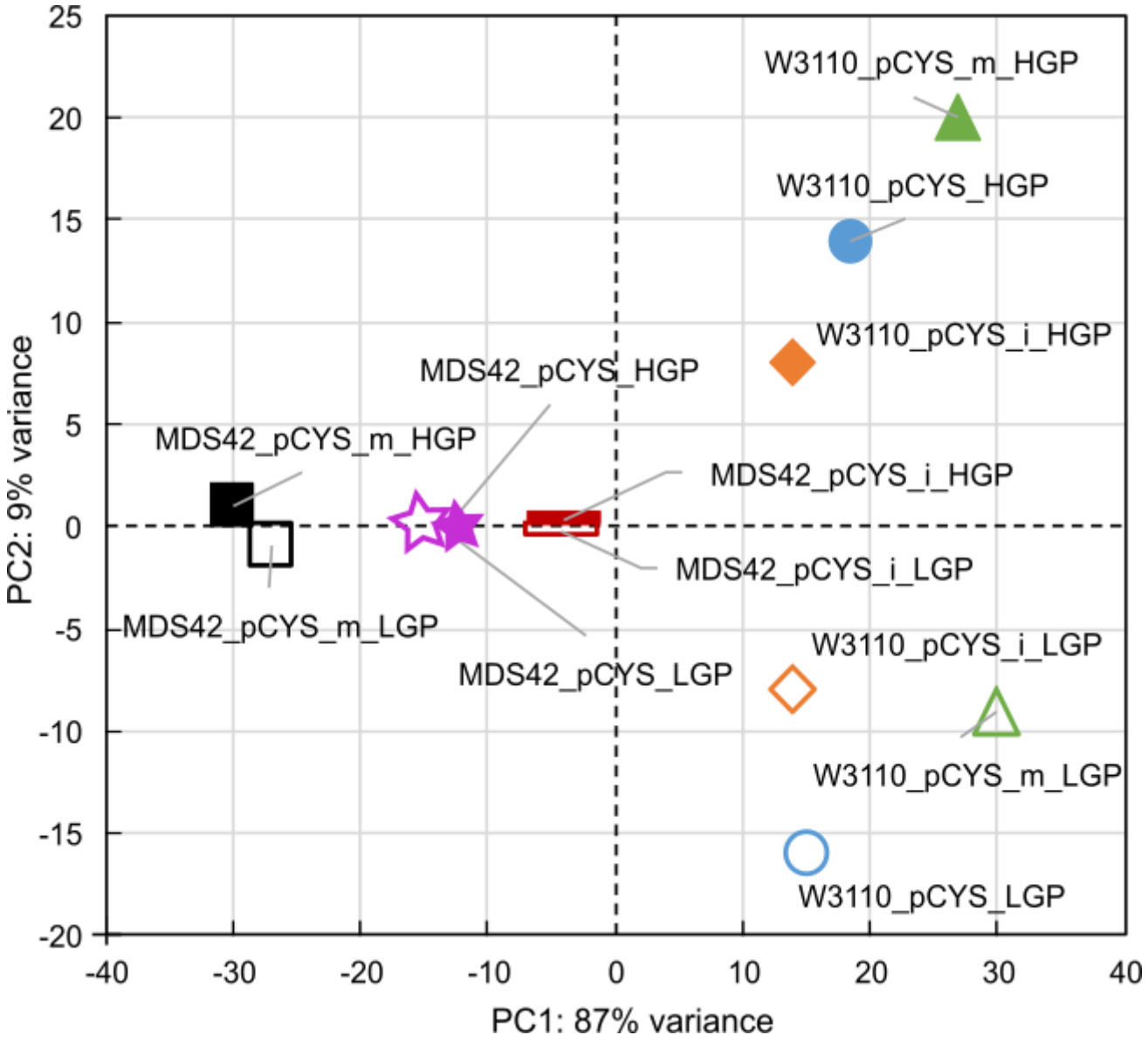

Figure 3

Principal component plot of all samples subjected to transcriptome analysis. Later generation populations (LGPs) are marked with filled-in shapes, while shapes of early generation populations (EGPs) are unfilled. LGPs correspond to populations with 60-65 evolved generations, while EGPs correspond to populations with 7-10 evolved generations. Related samples are marked with the same colour. Samples displayed in close proximity indicate very few differences between them while samples placed more distantly suggest bigger variations.

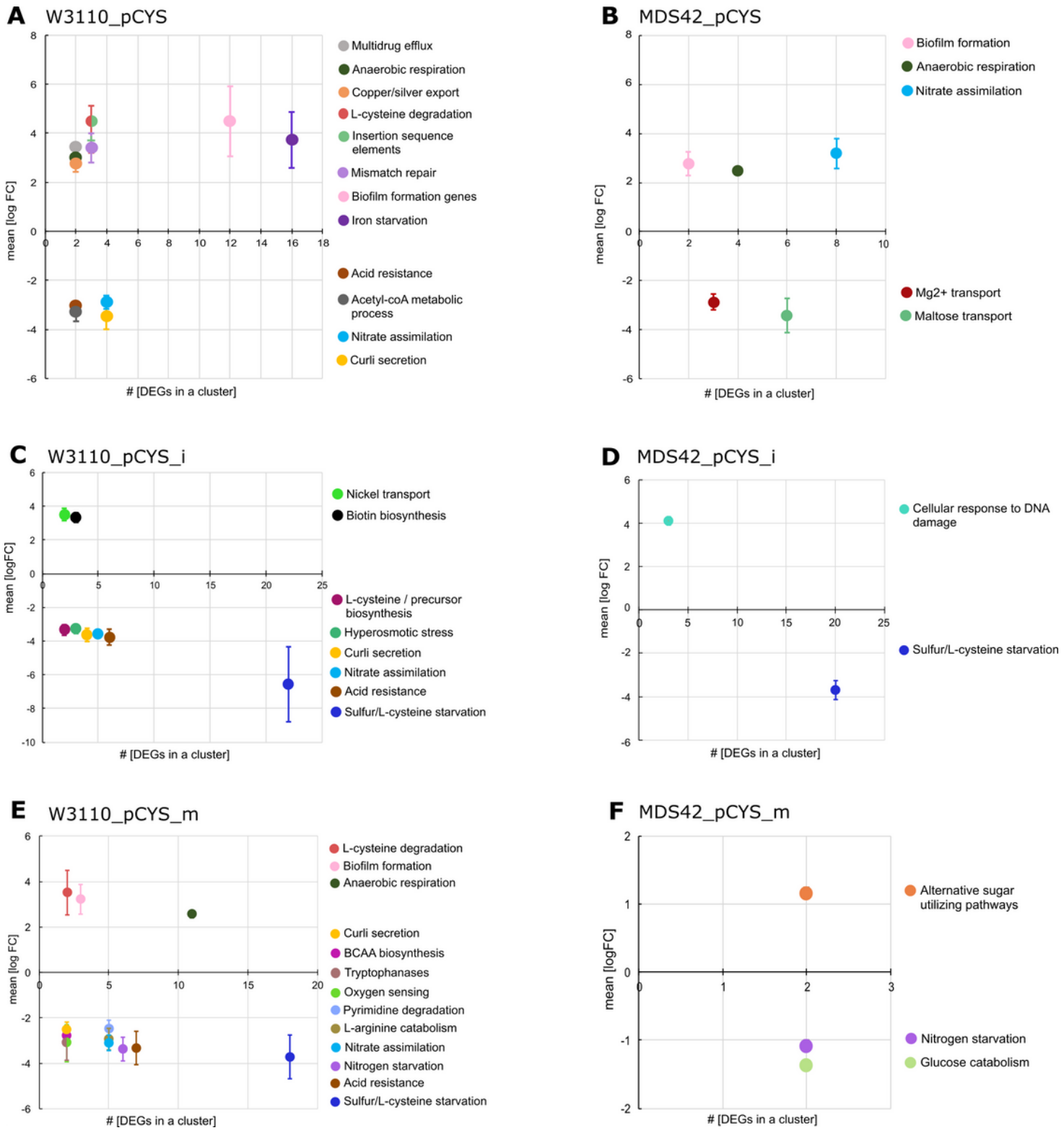

Figure 4

Clustering of differentially expressed genes (DEGs) based on the metabolic function in *E. coli*. For this, logarithmic fold change (logFC) medians of DEGs with the same metabolic function were plotted against the number of DEGs within this group. Fold changes were calculated by dividing values of the later generation population (LGP) by values of the early generation population (EGP). LGPs correspond to populations with 60-65 evolved generations, while EGPs correspond to populations with 7-10 evolved

generations. *E. coli* strains W3110 (A, C, E) and MDS42 (B, D, F) transformed with one of three plasmids engineered for L-cysteine production (pCYS: A, B; pCYS_i: C, D; pCYS_m: E, F) were analysed. A p-value cut-off <0.05 was selected. Genes with unknown function as well as clusters mapped with only one gene were disregarded here, but can be found in supplementary tables 5-10. **A:** 81 genes in total (57 genes within 12 cluster, 20 genes with unknown function and 4 single gene clusters). **B:** 34 genes in total (34 genes within 5 cluster, 6 genes with unknown function and 2 single gene clusters). **C:** 65 genes in total (47 genes within 8 cluster, 10 genes with unknown function and 8 single gene clusters). **D:** 23 genes within 2 cluster. **E:** 105 genes in total (70 genes within 12 cluster, 17 genes with unknown function and 18 single gene clusters. **F:** 14 genes in total (6 genes within 3 cluster, 4 genes with unknown function and 4 single gene clusters).

A

Plasmid	pCYS				pCYS_i				pCYS_m			
Strain	W3110		MDS42		W3110		MDS42		W3110		MDS42	
Population	EGP	LGP	EGP	LGP	EGP	LGP	EGP	LGP	EGP	LGP	EGP	LGP
Unmapped reads to plasmid sequence [%]	4.4	8.2	2.7	2.5	3.3	6	2.5	2.4	3.8	7.1	2	2.1
Reads mapped to IS [%]	0.13	0.24	0.01	0.01	0.11	0.18	0.01	0.01	0.12	0.23	0.01	0.01
Reads mapped to IS [total]	3803	5852	271	249	2361	5253	261	245	2101	6070	180	196

B

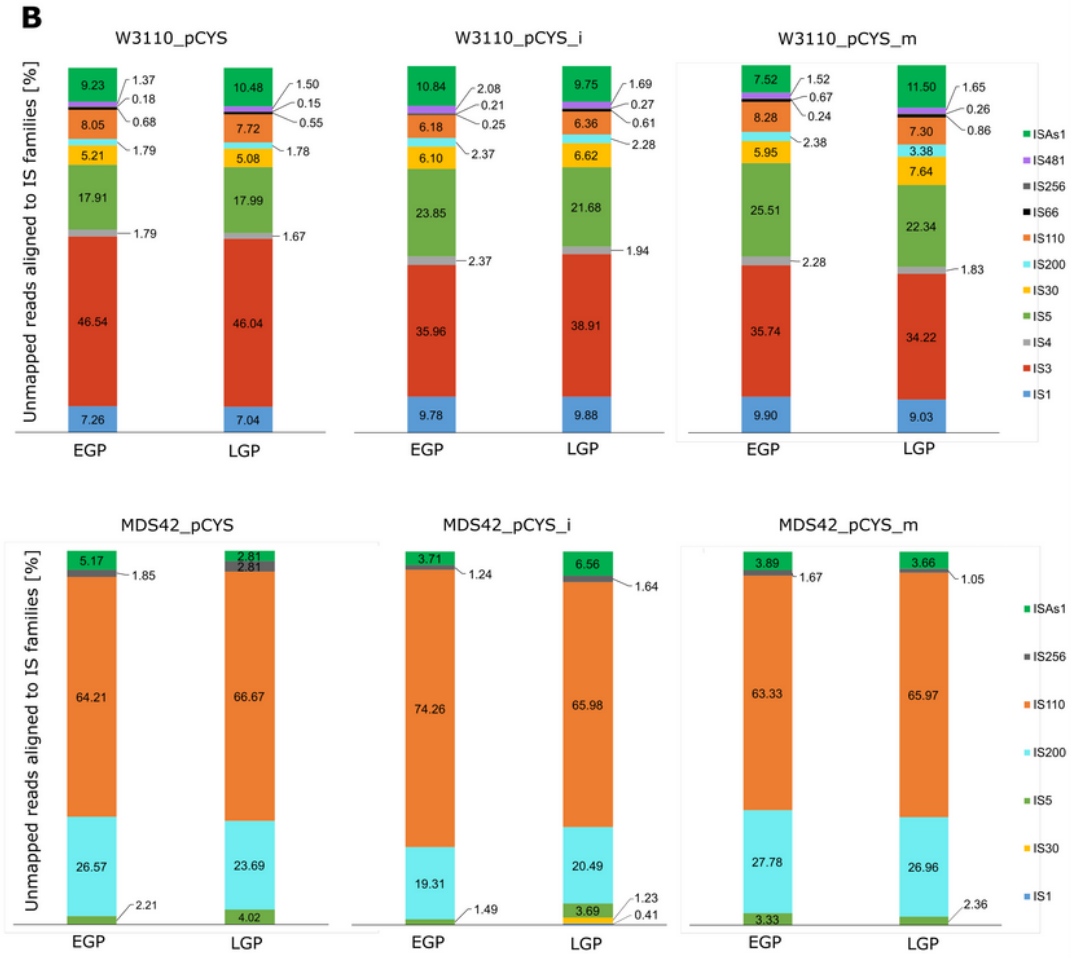

Figure 5

Mapping of sequenced reads against insertion sequence (IS) families. A: Table showing the percentages of reads that did not map to the plasmid sequences as well as the percentages of those unmapped reads mapped to IS. Plasmids pCYS, pCYS_i and pCYS_m extracted from early and late generation populations of W3110 and MDS42 got deep sequenced (Methods). **B:** Proportions of unmapped reads aligned to the different IS families in percent. The caption is arranged in the same order as the stacked bars.

Supplementary Files

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

- Supplementarydata.docx