

Here are the answers to the comments of the reviewers.

In addition, I want to state that

- we have more stringently clarified the nomenclature of cellulose biosynthesis gene clusters in this manuscript. In particular, we stick now tightly to the terms used in Römling and Galperin, 2015 for the reference cellulose biosynthesis gene clusters, naming as type I, II, III and IV. To discriminate, we termed the different cellulose biosynthesis gene clusters in *E. coli* EcType 1, 2, 3 and 4. Thereby, EcType 1 belongs to type IIa gene cluster, while EcType 4 belongs to the type Ib gene cluster.
- we have revised the text throughout for clarity.
- we have revised the figures for spelling errors and quality of display

Reviewer 2

This manuscript presents a comprehensive bioinformatic and experimental investigation of horizontally transferred cellulose biosynthesis gene clusters in *E. coli*. The authors identify three novel classes of BcsA cellulose synthases (BcsA2-4) that can be found on plasmids or at alternative chromosomal locations, with the BcsA4 cluster in strain 730V1 demonstrated to be functionally active. The work is timely and addresses important questions about the mobility and diversification of biofilm matrix components. However, there are several areas where the manuscript could be strengthened.

Major Concerns

1. Functional characterization is superficial for the most interesting finding

The demonstration that the plasmid-encoded bcs type 4 cluster in *E. coli* 730V1 is functional is arguably the most significant experimental finding, yet the characterization remains quite preliminary. The authors show Calcofluor white binding and Congo red staining, but:

What is the actual cellulose yield? Quantitative measurements (e.g., anthrone/sulfuric acid assay or HPLC quantification of hydrolyzed glucose) would substantially strengthen the claim that this cluster produces cellulose rather than just binding to the dyes non-specifically.

The Updegraff and KOH extractions are mentioned in Methods but the results are not clearly presented in Results. I could not find the TLC data showing glucose or cellobiose as the digestion product. This is critical evidence for cellulose identity.

What is the morphology/physical properties of the cellulose produced by BcsA4 compared to BcsA1? Electron microscopy of the extracted material would be valuable, especially since the authors discuss potential differences in polymer properties.

While these are valuable questions raised by the reviewer, they go beyond the research questions that are addressed in this manuscript. However, we acknowledge that the reviewer has a major point here. While we and others have demonstrated the Congo red and Calcofluor white binding in the context of the *E. coli* and *Salmonella* system is quite specific, there can be novel components introduced that challenge the established model framework.

In order to provide proof of cellulose biosynthesis we therefore deleted the *bcsA4* cellulose biosynthesis gene in two different independent experimental approaches using the kanamycin and chloramphenicol resistance cassette and one step gene inactivation. We observed that upon deletion of *bcsA4* we abolish dye binding on Congo red and Calcofluor white plates and cell aggregation in liquid M9 minimal medium. Therefore, the phenotypes and dye binding capacities that we observed previously are due to cellulose biosynthesis and *bcsA4* activity. These results are shown as a new Figure 5.

2. The functional significance of multiple operons remains unresolved

The manuscript identifies that *E. coli* can harbor alternative *bcs* clusters, but the fundamental question of "why?" is not adequately addressed:

Are there conditions where the plasmid-encoded cluster is preferentially expressed over the chromosomal copy?

The authors show expression under microaerophilic conditions in M9, but comparative expression analysis (qRT-PCR or promoter fusions) between the native and acquired clusters in strains that retain both would be transformative.

Does the *BcsA4* cluster confer a fitness advantage under any specific condition?

The authors mention thermotolerance of 730V1 due to tLST, but do not test whether cellulose production contributes to this phenotype. Survival assays under various stresses (heat, UV, desiccation, chlorine) with and without induced cellulose production would help address the ecological rationale.

Is there functional redundancy or specialization? The presence of multiple clusters suggests non-redundant functions, but this is not explored.

First of all, there is no chromosomal copy in *E. coli* 730V1 only remnants of the *bcs* operon on the chromosome, only the plasmid copy exists as a functional entity. This we have confirmed by PCR. So we cannot compare differential expression in the native context in a direct way.

We can now unambiguously show with the help of the *bcsA4* deletion mutant and using Calcofluor white staining, that cellulose biosynthesis occurs at 42C, but to a minor extent at 28 or 37 C concomitant with the increasing, but unconventionally low', but consistent coloration on Congo red plates (Figure 5 and data not shown).

Functional studies, which have all been performed and demonstrated for chromosomally encoded cellulose production as mentioned by the reviewer, are otherwise beyond the scope of this manuscript, although, of course, relevant to be asked.

Functional redundancy/specialization: Again relevant question, but beyond the scope of this work as also subtle differences in already reported functionalities for cellulose biosynthesis can be expected.

3. Bioinformatic analysis lacks rigor in some areas

Query coverage threshold (47% and 40%) is unusually low for identifying homologs. The authors should justify why such low thresholds were chosen, as this may include many

false positives or distantly related glycosyltransferases that are not true cellulose synthases.

This is the empirical value that was (afterwards) identified as the threshold for identification of cellulose synthases. BLAST analysis does not align homologous regions when the sequence similarity is low. Therefore the low value for the query coverage. Of note, the amino acid identity/similarity of experimentally verified cellulose synthases can be quite low. For example, the identity/similarity of cellulose synthase from *Cereibacter sphaeroides* and *E. coli* is 24.81/40.35. To indicate that this value is not the true query coverage we named it now 'apparent' query coverage.

Phylogenetic trees (Figure 1A, S1) are presented without clear bootstrap support values on the main branches. Please include bootstrap values on the main nodes to support the proposed relationships.

Bootstrap values >49 have been maintained.

The categorization into BcsA2-4 appears somewhat arbitrary. The authors should provide more explicit criteria (identity/similarity cutoffs) for assigning a protein to a particular class.

This information has now been included in the manuscript. Besides defining the identity/similarity cutoffs, BcsA2, BcsA3 and BcsA4 form a homogenous type with high amino acid sequence identity and characteristic amino acid motifs. BcsA2 and BcsA3 have most diverged in the cyclic di-GMP binding PilZ domain with the residue motifs for binding conserved.

For the GGDEF/EAL domain proteins associated with these clusters, the authors claim these have "not previously been recognized to post-translationally regulate cellulose biosynthesis." This is a strong statement that requires literature verification and more careful phrasing. Are there truly no reports linking YliF/DosC (YddV) homologs to cellulose regulation in any organism?

We have checked the literature one more time and according to the best of our knowledge, DosC has not been associated with the stimulation of cellulose biosynthesis although it has been associated with the production of curli another biofilm extracellular matrix component of *E. coli* often co-expressed with cellulose. Indeed, chromosomally encoded DosC is found in a complex with the phosphodiesterase DosP and the polynucleotide polymerase PNP to locally regulate its activity under different oxygen levels. DosC has also been reported to positively regulate curli and the exopolysaccharide N-acetylglucosamine. We now cite the respective references in the manuscript. Whether homologous, but distinct DosC on the plasmid forms a complex with DosP remains to be investigated.

4. Statistical analysis of plasmid sizes is insufficient

The comparison of plasmid sizes (Figure S3) is presented as a box plot with means, but: No statistical test is shown (e.g., Mann-Whitney U test given the non-normal distribution). The statement that BcsA-bearing plasmids are "larger" than non-BcsA plasmids needs statistical support.

What is the n for each group? This should be stated.

Are these data corrected for sampling bias? The database of complete plasmids is not necessarily representative.

All the requested information has been added in the Figure legend and Method section. However, we can only assess the plasmids as they are in the database. There is certainly a sampling bias (perhaps against clinical isolates) that we can neither influence, define nor estimate at this point.

Specific Technical Comments

Abstract

Line 11-12: "phosphoethanolamine modified exopolysaccharide cellulose" - this is stated as fact, but the manuscript describes multiple classes of cellulose synthases, including those (like BcsA2/3) that lack bcsG and thus likely produce unmodified cellulose. Please clarify that phosphoethanolamine modification is specific to certain classes.

Specified more precisely that phosphoethanolamine modified cellulose biosynthesis in *Salmonella typhimurium* and *Escherichia coli* is conducted by the type IIa nanomachine.

Introduction

Line 45-46: "chemically inert cellulose I macromolecule" - "chemically inert" is an overstatement. Cellulose can be derivatized and is susceptible to enzymatic hydrolysis. Perhaps "chemically stable" or "recalcitrant" would be more accurate.

revised

The introduction could benefit from a brief mention of what "rdar" and "pdar" morphotypes are, as these terms are used extensively later. Consider adding a sentence early on defining these abbreviations.

Introduction and Materials and Methods have been revised to include the 'rdar', 'pdar', 'bdar' and 'saw' terms and their definitions.

Results

Section 3.1: The statement that BcsA2 and BcsA3 are "restricted to the genus *Escherichia*" is based on BLAST analysis, but the methods don't mention screening all genera. Please clarify the scope of the search or soften the claim.

We have repeated our unbiased BLAST analysis with the currently available dataset.

Section 3.6: "We hypothesize that the co-localisation of these genes indicates connected functionality" - This hypothesis is reasonable but should be tested experimentally, at least by showing that deletion of the adjacent GGDEF protein affects cellulose production. As written, it remains speculative.

We have made this statement as a reasonable basis for future follow-up studies as these experiments are beyond the scope of this work.

Section 3.9: The experiments showing cellulose production at 42°C are interesting given the thermotolerant nature of 730V1. However, the authors should show that cellulose is indeed produced by the plasmid-encoded cluster and not by an unidentified chromosomal copy. The *bcsA1* deletion is mentioned, but was this confirmed by PCR/sequencing?

E. coli 730V1 had been sequenced and its sequence showed deletion of almost the entire cellulose biosynthesis operon besides parts of *bcsC* and *bcsG* as displayed in Figure 2B. We had confirmed lack of *bcsA* by PCR as in Table 4. Upon receiving the strain from our collaborators, we confirmed that the chromosomal operon is grossly deleted by PCR. Primers are shown in Table 1.

Figures

Figure 1: The phylogenetic tree is difficult to read in the current format. Consider using higher resolution and including branch length scale.

Figure 1 has been revised accordingly for clarity. A size bar and its definition scaling the branch length had been previously provided.

Figure 4: The colony morphology images are informative but the quality is suboptimal. Please provide higher resolution images or include magnification information.

This and other figures have been revised to include size bars and/or dimensions.

Figure S2: The plasmid maps are too small to read the gene annotations. Consider splitting into multiple panels or providing a higher-resolution version.

Plasmid maps and phylogenetic trees have been split into multiple panels.

Methods

Section 2.2: The rationale for selecting a nalidixic acid-resistant variant should be explained. Does this mutation affect biofilm formation or cellulose production? The authors mention they "subsequently investigated for *rdar* morphotype expression" but the results of this verification are not shown.

The rationale for the creation of a nalidixic acid resistant strain and the results of the assessment of the *rdar* morphotype expression are reported in the result section. Basically nalidixic acid resistance provides a marker for counterselection upon triparental mating. The corresponding figure S8A has been previously shown in the supplementary data.

Section 2.6: The authors state they used "LB without salt (LB-NS)" but standard LB has ~10 g/L NaCl. Please clarify what is meant by "salt" - is this NaCl-free? This is important for reproducibility.

Revised as it means LB without the NaCl.

Section 2.9: The TLC method is described but no standards are mentioned. What was used as a positive control? What were the expected R_f values for glucose and cellobiose?

D-glucose and D-cellobiose has been used as standards. Crystalline cellulose has now been used as substrate as well to be digested by cellulase. The R_f values for glucose and cellobiose have been determined experimentally as no R_f values are available for this solvent system in the literature to our knowledge.

It is known that the cellulase from *Trichoderma viride*, Sigma has promiscuous substrate specificity, however, in combination with the isolation protocol and the use of alkaline KOH and Updegraff reagent to isolate amorphous and crystalline and crystalline cellulose respectively, the unambiguous identification of the substrate is ensured.

Minor Issues

Writing and Clarity

Throughout: The manuscript is overly dense in places with long, complex sentences that obscure the meaning. For example, Section 3.3 (lines 360-370) could be broken into multiple sentences for clarity.

Revised as suggested.

Line 170-171: "The gene neighborhood ... was assessed manually" - this should specify how many genomes were examined and give an estimate of the effort involved.

We state 'for each non-redundantly retrieved BcsA cellulose synthase by evaluating the annotated plasmids.' In total this means that 62 *E. coli* and 80 non-*E. coli* with non-redundant BcsA proteins were examined. This information is now not only shown in the Figures, but also stated in the manuscript as numbers.

Line 251-252: "BcsA4 members 35% identity and 50% similarity" - for context, what is the typical identity threshold for defining a new class of cellulose synthase? This is not provided.

In our previous classification of the types of cellulose biosynthesis gene clusters (Römling, 2001 and Römling and Galperin, 2015), no thresholds for identity/similarity have been defined.

Line 338: "Strikingly, the bcs type 4 operon harbors additionally bcsD" - this is not "striking" given that class I operons commonly have bcsD. Perhaps "notably" would be more appropriate.

Revised as suggested. However *bcsG* is of type IIa (Figure 1, Figure S4) in the otherwise type Ib gene cluster. This fact is now mentioned in the discussion. Appreciate it has been noted.

References

Reference [32] is a self-flocculation study in *Zymomonas* - the connection to the present work is not obvious. Please explain why this reference is included or consider removing it.

The cellulose biosynthesis operon(s) of *Zymomonas* spp. is now suggested to be a new subtype I gene cluster in the text.

Reference [71] appears to be about *E. coli* O157:H7 and salicylate/decanoate control - I could not see the connection to this manuscript. Please verify all references are appropriate.

We appreciate that the reviewer spotted this miss. We replace this reference with the correct one as mentioned in the Supplementary Material.