nature portfolio

Peer Review File

Integrative genomic reconstruction reveals heterogeneity in carbohydrate utilization across human gut bifidobacteria

Corresponding Author: Professor Andrei Osterman

Version 0:

Decision Letter:

27th March 2024

Dear Jeff,

Thank you very much for your enquiry about submitting your manuscript "Bifidobacteria efficacy" to Nature Microbiology. It certainly sounds interesting, and we would be happy to consider it for publication and as part of our Microbiome and Nutrition call for papers. However, I'm sure you'll understand that we cannot make a firm decision about whether to send the paper out to review until we have carefully read the full paper (and appropriate background literature).

In order to submit your complete manuscript to Nature Microbiology, please use the link below:

Link Redacted

If you have any questions, please feel free to contact me.

Yours sincerely,

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Integrative genomic reconstruction of carbohydrate utilization networks in bifidobacteria: global trends, local variability, and dietary adaptation.
Arzamasov et al.

In this manuscript the authors first perform a reconstruction of C-metabolic pathways in over 250 reference bifidobacterial strains based on in silico and literature data to produce 70 pathways for the utilization of a variety of carbohydrates. This information was then used to construct a so-called Binary Phenotype Matrix for gene-trait matching purposes (and validated using published carbohydrate utilization phenotypes) and applied to explore the variable carbohydrate metabolic abilities of a wider range of bifidobacterial (sub)species/strains. Strain level diversity was noticed (particularly among the B. longum species) while (predicted) carbohydrate utilization phenotypes did not always follow phylogeny. A new B. longum subspecies was proposed based on distinct (predicted) C-utilization abilities. Furthermore, apparent geographical (Bangladeshi) specific C-utilization properties were discovered among B. catenulatum strains based on a presumed xyloglucan metabolism gene cluster (as corroborated by RNAseq data). The predicted utilization abilities were in part experimentally assessed using a subset of bifidobacterial strains and employing a specific range of carbohydrates as well as using a HMO glycoprofiling approach. The obtained experimental validation was then taken to infer a genus wide reconstruction of carbohydrate utilization profiles (including 306 genomes of isolates and 2404 MAGs), while also incorporating pathways for B vitamin/amino avid biosynthesis and urea utilization. Again, species and strain level variations were observed, while findings were also investigated for age and host life style differences.
Overall, the paper constitutes a very significant amount of bioinformatic analysis, complemented with relevant experimental data. It's a timely paper that may have value beyond those focused on bifidobacterium. However, I am not convinced that the knowledge advances and novel findings are sufficient to justify publication of this manuscript in Nature Microbiology. Many of their observations confirm (perhaps in a wider context) what was published before at species or strain level.

Some other comments
L66-74: This statement is perhaps too simplistic and does not take into account possible cross-feeding in the gut. HMO utilisation in the gut is not necessarily promoted by a single strain/species, but can also be achieved by cooperative cross-feeding.

The authors propose the establishment of a novel species within B. longum, but the authors don't indicate which genes they used for phylogenetic analysis, it is not clear how they manually refine taxonomy using ANI. I am not sure what difference between extended figures and supplemental figures are but this tree should be given more prominence).

Along the same line: L124-134: Average Nucleotide Identity provides an indication of species-level clustering (based on % identity of common regions between two sequences). An initial ANI clustering should be refined using an appropriate phylogenetic methodology (based on multiple sequence alignment of single-copy core genes). If the sequences selected for phylogenetic analysis are appropriate and so is the methodology for tree construction, there should be no need to manually correct phylogeny using ANI. One would argue why using phylogenetics at all!

In its current presentation the paper is difficult to interpret, the narrative is not logical and it jumps from figure to figure. For example Figure 2 is discussed, Figure 3 is skipped and then Figure 4a and 5a are next (line 186)

Similarly, why go from making predictions in 263 genomes in Figure 2 to 2973 genomes in Figure 6? A more logical approach would be to use 263 genomes to make predictions and then test this on the larger data-set to ensure it is scalable and comparable. Then it should be phenotypically validated before moving onto the species-specific results.

The authors should validate their predictions upfront in vitro using geographically distinct strains and species on different substrates. Can they replicate Binary Phenotpye Matrix (line 158) predicted strain level differences in vitro? I would not expect perfect match but reader needs to be convinced their predictions are accurate before presenting results.

Figure 2 is too detailed and hard to take away main findings from. As just one example (line 211): "Compared to other B. longum subspecies, this distinct clade had fewer predicted host glycan utilization pathways, suggesting a lineage specific loss of genes and gene clusters". If not highlighted in the text it would not be possible for reader to see this in figure. One option would be to move this figure to supplemental and replace with a simplified version that perhaps presents broad categories of substrates together with their full names (not abbreviations) or present by species and colour boxes by % of strains of species that metabolise substrate to show strain-level variability.

Minor comments/remarks
L171: Based on previous reports, not all species are able to utilise sucrose.

L176-180: How these results add to previously published reports? These are not novel findings, as metabolism of HMOs, host-derived, mono- di- and polysaccharides in B. longum, B. Breve and B. Bifidum have been extensively described in literature.

L181-182: Species-level differences have been described in literature before.

L187-191: has this been experimentally proven or is only a speculation?

L232-38: These are not novel findings.

L47-51: Has this been proven experimentally?

L255-58: The strain level heterogeneity of bifidobacterial species is not a novel finding. Has been discussed extensively in literature.

L440-44: This have been extensively described in literature.

L504: Given the a high level of strain specialisation in utilising carbohydrates exist in bifidobacteria, predictive approaches applied to MAGs may not be reflective of strain-level abilities.

Supplementary Figure1: Maybe this tree is missing bootstrap values? Are the authors sure that all nodes returned a 100 bootstrap value? Also, a list of genes used to compute this phylogeny should be indicated.

(Remarks on code availability)
I did not review the code, but a colleague of mine had a look at it (and did not seem to find any issues)

Reviewer #2

(Remarks to the Author)
In this manuscript Arzamasov et. al. conduct a comprehensive investigation into the carbohydrate utilization potential of bifidobacteria. The authors perform an expert-led and manual curation of carbohydrate utilization across a few hundred reference genomes, verify the accuracy of their metabolic predictions using both previously generated and newly generated data, and then extend their predictions to a large set of a few thousand lower-quality genomes to assess bifidobacteria diversity across the globe.

This work is exceptionally thorough, expertly executed, and fills a significant knowledge gap in the field. Bifidobacteria are increasingly recognized as important to human health, and this work presents a much-needed foundational understanding of bifidobacteria genomics. The manuscript stands out for its meticulous attention to detail in the manual curation of bifidobacteria carbohydrate utilization pathways, as well as for its successful application of these detail-oriented methods to several large genomic databases. This paper will undoubtedly serve as a foundational text in the field of bifidobacteria genomics for years to come, providing both newcomers with a thorough introduction and experts with a valuable reference for their ongoing work.

I have no critiques or concerns about any of the methods or results, and the level of detail presented in the methods section and supplementary tables is exemplary. However, I do have a few minor questions, the clarification of which might be helpful for future readers.

Q1) Would it be feasible to provide a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper? It would be great to enable others to take advantage of your manually curated pathways.

Q2) Around lines 412-415, could you describe why you need a random forest model to predict carbohydrate utilization pathways in the full set of genomes? Why couldn't you use the same carbohydrate utilization pathway prediction method used for the 263 reference genomes?

Q3) In Figure 2 and Figure 6a,b, it's very difficult to match the colors to the species in the legend. In Figures 2 and 6a I suggest adding additional text in the figures themselves to point out the major species, and in Figure 6b I suggest adding taxonomic labels along the y-axis.

Sincerely,
Dr. Matthew O

(Remarks on code availability)
As stated in the main review, a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper would be a welcome addition to the code.

Reviewer #4

(Remarks to the Author)
Bifidobacteria are common commensals in the human gut microbiome and were one of the earliest commercialized probiotics. Their consistent association with health and their importance in the early establishment of the infant microbiome has generated significant interest in understanding their niche in the human gut microbiome. This has been hampered by the large metabolic heterogeneity in the Bifidobacteria genus, especially their metabolism of human milk oligosaccharides (HMOs) which are still not well characterized. Arzamasov provide a detailed reconstruction of carbohydrate metabolism in Bifidobacteria based on 253 reference genomes. Experimental validation in 16 Bifidobacteria strains showed good accuracy in recapitulating growth on specific carbon sources. The authors also predicted carbohydrate usage patterns in more than 2,700 additional genomes. I do feel that the presented curated reconstruction of carbohydrate utilization expands the current knowledge significantly and the manuscript includes some convincing applications. Unfortunately, there is very little description how this reconstruction is achieved and the relevant data is not provided to the scientific community. I also have some concerns regarding the prediction of carbohydrate utilization for the ~2,700 which is currently lacking validation. This could be alleviated with some better description of the methodology, providing additional data, and some experimental validation of the prediction pipeline. All of this is definitely in the realm of the authors' capabilities and would improve the manuscript significantly while also providing a valuable resource for other researchers.

Major comments

For me, one key part of the manuscript is the metabolic reconstruction of Bifidobacteria carbon metabolism. Thus, is was a bit surprising to find very little description how this was achieved. The description in lines 135-140 is too superficial and the Methods only mention the use of the "mcSEED" platform without providing a reference URL to the method/platform (the references used 46/47 used in the text do not describe mcSEED). I would ask the authors for the following:

1. Provide a detailed description of how the subsystems were generated in mcSEED. Also describe how orthologous gene clusters were established (method to infer homology and propagate functional roles).
2. Provide the specific gene clusters (sequences) for the functional roles along with the annotations (description, subsystems, substrates and products) so this can be used by other researchers.

The validation with the 16 reference strains is quite comprehensive and the additional transcriptomics does support the claims that the metabolic reconstruction does indeed recover carbohydrate utilization adaptations on the strain level. However, it should also be noted that the experiments on those 16 strains is the current extent of validation in the manuscript, despite all the additional predictions. It seems like those 16 strains were also part of the manually reconstructed set of 263 genomes, so it is essentially a best-case evaluation. I wonder how well those predictions would extend to genomes that are geographically or genomically different from the used reference (strain) genomes.

It would be helpful if the authors could also test some more distant strains from the 2,700 genome prediction set. This would help to establish a bound on the prediction accuracy which is currently hard to evaluate as also argued below. Right now, the observed increased heterogeneity in carbohydrate utilization phenotypes in that genome set could also be a consequence of

model bias.

Regarding the random forest prediction of phenotypes it would be great to provide some motivation for the chosen methodology in lines 675-714. Why were only 27 subsystems chosen even though 70 were reconstructed initially? How exactly were the outgroup proteins chosen? DIAMOND is capable of performing local alignments which are often used in functional annotation pipelines such as EGGNOG etc. So I wonder what the motivation for the additional domain clustering was.
In a similar vein, I did not understand the necessity for the Machine Learning. Presence/absence of the subsystems inferred in the reconstruction can be translated directly into a BPM matrix as done before in the manuscript in lines 662-669. Was this necessary to overcome missing gene clusters/functional roles? A machine learning model would also be somewhat less mechanistic because the predicted carbohydrate utilization could in theory be inferred from unrelated subsystems/gene clusters which would be hard to interpret.
In case the machine learning is indeed necessary, it might be a bit more stringent to use a train-test-validation setup where the hyperparameter training is done on the train-test set combination and the final model is tested on a hold-out validation set. As argued above, experimental validation for some strains from the validation step would strengthen the manuscript as well.

Lines 610-613 do not specify how consent for the fecal sampling from Bangladeshi children was obtained and whether the legal guardians consented to the additional use in this study. Please also indicate the local collaborators handling the sampling.

Minor comments

It was not obvious to me which proportion of the 525 functional roles/gene clusters are currently *not* represented in the common protein annotation pipelines and databases. It would be helpful if the manuscript could point out what fraction is indeed new or is already covered by databases such as EGGNOG, ModelSEED, KEGG, etc.

In the manuscript subsystems and pathways are used interchangeably, however not all subsystems are valid pathways IMHO. Pathways usually require some kind of linear chain of substrate-product conversions or modification. Were all subsystems completely connected biochemical pathways?

There are a few typos in the figures, for instance "Bangaldesh" and "pawthays" in Fig. 1. I would recommend to revise the figures again.

In lines 596-600: Genome were pooled from different studies here which used different sequencing and assembly pipelines. Was there some check that this did not show biases in the phenotype matrix?

The phrase in line 85-87 can be misinterpreted to mean that all Bangladeshi children do not receive enough breast milk. Maybe reformulate to "[...] may not be sufficiently stable in a sub-population of Bangladeshi children who did not receive [...]".

Please provide a reference for the statements made in lines 98-100.

Line 590: Please indicate whether checkM version 1 or 2 was used since they give very different results.

Line 683: How was the limit of 50 hits chosen?

Line 774: Did the human milk donors match the geographic region or population the strains were obtained from?

(Remarks on code availability)
The repository is well documented and includes the analyses *after* the annotation/reconstruction with mcSEED (see related comments above).

The provided R markdown notebook (`Sup_code_file.Rmd`) in the GitHub repository could be run and did generate the figures shown in the paper. The renv environment setup did not work for me, but was not necessary to run the notebook.

Decision Letter:

3rd September 2024

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Andrei,

Thank you for your patience while your manuscript "Integrative genomic reconstruction of carbohydrate utilization networks in bifidobacteria: global trends, local variability, and dietary adaptation" was under peer-review at Nature Microbiology. It has now been seen by 3 referees, whose expertise and comments you will find at the end of this email. Although they find your work of some potential interest, they have raised a number of concerns that will need to be addressed before we can consider publication of the work in Nature Microbiology.

In particular, we will require you to validate using strains that are distinct from your reference set of genomes, add more methods information and justifications for some approaches, caveat that cross-feeding will play an important role, and provide detailed ethics information.

Should further experimental data allow you to address these criticisms, we would be happy to look at a revised manuscript.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We strongly support public availability of data. Please place the data used in your paper into a public data repository, if one exists, or alternatively, present the data as Source Data or Supplementary Information. If data can only be shared on request, please explain why in your Data Availability Statement, and also in the correspondence with your editor. For some data types, deposition in a public repository is mandatory - more information on our data deposition policies and available repositories can be found at https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data.

Please include a data availability statement as a separate section after Methods but before references, under the heading "Data Availability". This section should inform readers about the availability of the data used to support the conclusions of your study. This information includes accession codes to public repositories (data banks for protein, DNA or RNA sequences, microarray, proteomics data etc…), references to source data published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", mentioning any restrictions on availability. If DOIs are provided, we also strongly encourage including these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

If revising your manuscript:

* Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

* If you have not done so already we suggest that you begin to revise your manuscript so that it conforms to our Article format instructions at http://www.nature.com/nmicrobiol/info/final-submission. Refer also to any guidelines provided in this letter.

* Include a revised version of any required reporting checklist. It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. A revised checklist is essential for re-review of the paper.

When submitting the revised version of your manuscript, please pay close attention to our href="https://www.nature.com/nature-portfolio/editorial-policies/image-integrity">Digital Image Integrity Guidelines. and to the following points below:

-- that unprocessed scans are clearly labelled and match the gels and western blots presented in figures.
-- that control panels for gels and western blots are appropriately described as loading on sample processing controls
-- all images in the paper are checked for duplication of panels and for splicing of gel lanes.

Finally, please ensure that you retain unprocessed data and metadata files after publication, ideally archiving data in perpetuity, as these may be requested during the peer review and production process or after publication if any issues arise.

Please use the link below to submit a revised paper:

Link Redacted

Note: This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first.

Nature Microbiology is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision, even if a similar study has been accepted for publication at Nature Microbiology or published elsewhere (up to a maximum of 6 months).

In the meantime we hope that you find our referees' comments helpful.

Yours sincerely,

****

Reviewer Expertise:

Referee #1: Bifidobacteria, gut microbiome
Referee #2: gut microbiome, multi-omics, bioinformatics
Referee #3: gut microbiome, modelling

Reviewer Comments:

Reviewer #1 (Remarks to the Author):

Integrative genomic reconstruction of carbohydrate utilization networks in bifidobacteria: global trends, local variability, and dietary adaptation.
Arzamasov et al.

In this manuscript the authors first perform a reconstruction of C-metabolic pathways in over 250 reference bifidobacterial strains based on in silico and literature data to produce 70 pathways for the utilization of a variety of carbohydrates. This information was then used to construct a so-called Binary Phenotype Matrix for gene-trait matching purposes (and validated using published carbohydrate utilization phenotypes) and applied to explore the variable carbohydrate metabolic abilities of a wider range of bifidobacterial (sub)species/strains. Strain level diversity was noticed (particularly among the B. longum species) while (predicted) carbohydrate utilization phenotypes did not always follow phylogeny. A new B. longum subspecies was proposed based on distinct (predicted) C-utilization abilities. Furthermore, apparent geographical (Bangladeshi) specific C-utilization properties were discovered among B. catenulatum strains based on a presumed xyloglucan metabolism gene cluster (as corroborated by RNAseq data). The predicted utilization abilities were in part experimentally assessed using a subset of bifidobacterial strains and employing a specific range of carbohydrates as well as using a HMO glycoprofiling approach. The obtained experimental validation was then taken to infer a genus wide reconstruction of carbohydrate utilization profiles (including 306 genomes of isolates and 2404 MAGs), while also incorporating pathways for B vitamin/amino avid biosynthesis and urea utilization. Again, species and strain level variations were observed, while findings were also investigated for age and host life style differences.
Overall, the paper constitutes a very significant amount of bioinformatic analysis, complemented with relevant experimental data. It's a timely paper that may have value beyond those focused on bifidobacterium. However, I am not convinced that the knowledge advances and novel findings are sufficient to justify publication of this manuscript in Nature Microbiology. Many of their observations confirm (perhaps in a wider context) what was published before at species or strain level.

Some other comments
L66-74: This statement is perhaps too simplistic and does not take into account possible cross-feeding in the gut. HMO utilisation in the gut is not necessarily promoted by a single strain/species, but can also be achieved by cooperative cross-feeding.

The authors propose the establishment of a novel species within B. longum, but the authors don't indicate which genes they used for phylogenetic analysis, it is not clear how they manually refine taxonomy using ANI. I am not sure what difference between extended figures and supplemental figures are but this tree should be given more prominence).

Along the same line: L124-134: Average Nucleotide Identity provides an indication of species-level clustering (based on % identity of common regions between two sequences). An initial ANI clustering should be refined using an appropriate phylogenetic methodology (based on multiple sequence alignment of single-copy core genes). If the sequences selected for phylogenetic analysis are appropriate and so is the methodology for tree construction, there should be no need to manually correct phylogeny using ANI. One would argue why using phylogenetics at all!

In its current presentation the paper is difficult to interpret, the narrative is not logical and it jumps from figure to figure. For example Figure 2 is discussed, Figure 3 is skipped and then Figure 4a and 5a are next (line 186)

Similarly, why go from making predictions in 263 genomes in Figure 2 to 2973 genomes in Figure 6? A more logical approach would be to use 263 genomes to make predictions and then test this on the larger data-set to ensure it is scalable and comparable. Then it should be phenotypically validated before moving onto the species-specific results.

The authors should validate their predictions upfront in vitro using geographically distinct strains and species on different substrates. Can they replicate Binary Phenotpye Matrix (line 158) predicted strain level differences in vitro? I would not expect perfect match but reader needs to be convinced their predictions are accurate before presenting results.

Figure 2 is too detailed and hard to take away main findings from. As just one example (line 211): "Compared to other B. longum subspecies, this distinct clade had fewer predicted host glycan utilization pathways, suggesting a lineage specific loss of genes and gene clusters". If not highlighted in the text it would not be possible for reader to see this in figure. One option would be to move this figure to supplemental and replace with a simplified version that perhaps presents broad categories of substrates together with their full names (not abbreviations) or present by species and colour boxes by % of strains of species that metabolise substrate to show strain-level variability.

Minor comments/remarks
L171: Based on previous reports, not all species are able to utilise sucrose.

L176-180: How these results add to previously published reports? These are not novel findings, as metabolism of HMOs, host-derived, mono- di- and polysaccharides in B. longum, B. Breve and B. Bifidum have been extensively described in literature.

L181-182: Species-level differences have been described in literature before.

L187-191: has this been experimentally proven or is only a speculation?

L232-38: These are not novel findings.

L47-51: Has this been proven experimentally?

L255-58: The strain level heterogeneity of bifidobacterial species is not a novel finding. Has been discussed extensively in literature.

L440-44: This have been extensively described in literature.

L504: Given the a high level of strain specialisation in utilising carbohydrates exist in bifidobacteria, predictive approaches applied to MAGs may not be reflective of strain-level abilities.

Supplementary Figure1: Maybe this tree is missing bootstrap values? Are the authors sure that all nodes returned a 100 bootstrap value? Also, a list of genes used to compute this phylogeny should be indicated.

Reviewer #1 (Remarks on code availability):

I did not review the code, but a colleague of mine had a look at it (and did not seem to find any issues)

Reviewer #2 (Remarks to the Author):

In this manuscript Arzamasov et. al. conduct a comprehensive investigation into the carbohydrate utilization potential of bifidobacteria. The authors perform an expert-led and manual curation of carbohydrate utilization across a few hundred reference genomes, verify the accuracy of their metabolic predictions using both previously generated and newly generated data, and then extend their predictions to a large set of a few thousand lower-quality genomes to assess bifidobacteria diversity across the globe.

This work is exceptionally thorough, expertly executed, and fills a significant knowledge gap in the field. Bifidobacteria are increasingly recognized as important to human health, and this work presents a much-needed foundational understanding of bifidobacteria genomics. The manuscript stands out for its meticulous attention to detail in the manual curation of bifidobacteria carbohydrate utilization pathways, as well as for its successful application of these detail-oriented methods to several large genomic databases. This paper will undoubtedly serve as a foundational text in the field of bifidobacteria genomics for years to come, providing both newcomers with a thorough introduction and experts with a valuable reference for their ongoing work.

I have no critiques or concerns about any of the methods or results, and the level of detail presented in the methods section and supplementary tables is exemplary. However, I do have a few minor questions, the clarification of which might be helpful for future readers.

Q1) Would it be feasible to provide a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper? It would be great to enable others to take advantage of your manually curated pathways.

Q2) Around lines 412-415, could you describe why you need a random forest model to predict carbohydrate utilization pathways in the full set of genomes? Why couldn't you use the same carbohydrate utilization pathway prediction method used for the 263 reference genomes?

Q3) In Figure 2 and Figure 6a,b, it's very difficult to match the colors to the species in the legend. In Figures 2 and 6a I suggest adding additional text in the figures themselves to point out the major species, and in Figure 6b I suggest adding taxonomic labels along the y-axis.

Sincerely,
Dr. Matthew O

Reviewer #2 (Remarks on code availability):

As stated in the main review, a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper would be a welcome addition to the code.

Reviewer #4 (Remarks to the Author):

Bifidobacteria are common commensals in the human gut microbiome and were one of the earliest commercialized probiotics. Their consistent association with health and their importance in the early establishment of the infant microbiome has generated significant interest in understanding their niche in the human gut microbiome. This has been hampered by the large metabolic

heterogeneity in the Bifidobacteria genus, especially their metabolism of human milk oligosaccharides (HMOs) which are still not well characterized. Arzamasov provide a detailed reconstruction of carbohydrate metabolism in Bifidobacteria based on 253 reference genomes. Experimental validation in 16 Bifidobacteria strains showed good accuracy in recapitulating growth on specific carbon sources. The authors also predicted carbohydrate usage patterns in more than 2,700 additional genomes. I do feel that the presented curated reconstruction of carbohydrate utilization expands the current knowledge significantly and the manuscript includes some convincing applications. Unfortunately, there is very little description how this reconstruction is achieved and the relevant data is not provided to the scientific community. I also have some concerns regarding the prediction of carbohydrate utilization for the ~2,700 which is currently lacking validation. This could be alleviated with some better description of the methodology, providing additional data, and some experimental validation of the prediction pipeline. All of this is definitely in the realm of the authors' capabilities and would improve the manuscript significantly while also providing a valuable resource for other researchers.

Major comments

For me, one key part of the manuscript is the metabolic reconstruction of Bifidobacteria carbon metabolism. Thus, is was a bit surprising to find very little description how this was achieved. The description in lines 135-140 is too superficial and the Methods only mention the use of the "mcSEED" platform without providing a reference URL to the method/platform (the references used 46/47 used in the text do not describe mcSEED). I would ask the authors for the following:

1. Provide a detailed description of how the subsystems were generated in mcSEED. Also describe how orthologous gene clusters were established (method to infer homology and propagate functional roles).
2. Provide the specific gene clusters (sequences) for the functional roles along with the annotations (description, subsystems, substrates and products) so this can be used by other researchers.

The validation with the 16 reference strains is quite comprehensive and the additional transcriptomics does support the claims that the metabolic reconstruction does indeed recover carbohydrate utilization adaptations on the strain level. However, it should also be noted that the experiments on those 16 strains is the current extent of validation in the manuscript, despite all the additional predictions. It seems like those 16 strains were also part of the manually reconstructed set of 263 genomes, so it is essentially a best-case evaluation. I wonder how well those predictions would extend to genomes that are geographically or genomically different from the used reference (strain) genomes.

It would be helpful if the authors could also test some more distant strains from the 2,700 genome prediction set. This would help to establish a bound on the prediction accuracy which is currently hard to evaluate as also argued below. Right now, the observed increased heterogeneity in carbohydrate utilization phenotypes in that genome set could also be a consequence of model bias.

Regarding the random forest prediction of phenotypes it would be great to provide some motivation for the chosen methodology in lines 675-714. Why were only 27 subsystems chosen even though 70 were reconstructed initially? How exactly were the outgroup proteins chosen? DIAMOND is capable of performing local alignments which are often used in functional annotation pipelines such as EGGNOG etc. So I wonder what the motivation for the additional domain clustering was.
In a similar vein, I did not understand the necessity for the Machine Learning. Presence/absence of the subsystems inferred in the reconstruction can be translated directly into a BPM matrix as done before in the manuscript in lines 662-669. Was this necessary to overcome missing gene clusters/functional roles? A machine learning model would also be somewhat less mechanistic because the predicted carbohydrate utilization could in theory be inferred from unrelated subsystems/gene clusters which would be hard to interpret.
In case the machine learning is indeed necessary, it might be a bit more stringent to use a train-test-validation setup where the hyperparameter training is done on the train-test set combination and the final model is tested on a hold-out validation set. As argued above, experimental validation for some strains from the validation step would strengthen the manuscript as well.

Lines 610-613 do not specify how consent for the fecal sampling from Bangladeshi children was obtained and whether the legal guardians consented to the additional use in this study. Please also indicate the local collaborators handling the sampling.

Minor comments

It was not obvious to me which proportion of the 525 functional roles/gene clusters are currently *not* represented in the common protein annotation pipelines and databases. It would be helpful if the manuscript could point out what fraction is indeed new or is already covered by databases such as EGGNOG, ModelSEED, KEGG, etc.

In the manuscript subsystems and pathways are used interchangeably, however not all subsystems are valid pathways IMHO. Pathways usually require some kind of linear chain of substrate-product conversions or modification. Were all subsystems completely connected biochemical pathways?

There are a few typos in the figures, for instance "Bangaldesh" and "pawthays" in Fig. 1. I would recommend to revise the figures again.

In lines 596-600: Genome were pooled from different studies here which used different sequencing and assembly pipelines. Was there some check that this did not show biases in the phenotype matrix?

The phrase in line 85-87 can be misinterpreted to mean that all Bangladeshi children do not receive enough breast milk. Maybe reformulate to "[...] may not be sufficiently stable in a sub-population of Bangladeshi children who did not receive [...]".

Please provide a reference for the statements made in lines 98-100.

Line 590: Please indicate whether checkM version 1 or 2 was used since they give very different results.

Line 683: How was the limit of 50 hits chosen?

Line 774: Did the human milk donors match the geographic region or population the strains were obtained from?

Reviewer #4 (Remarks on code availability):

The repository is well documented and includes the analyses *after* the annotation/reconstruction with mcSEED (see related comments above).

The provided R markdown notebook (`Sup_code_file.Rmd`) in the GitHub repository could be run and did generate the figures shown in the paper. The renv environment setup did not work for me, but was not necessary to run the notebook.

****

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript has been substantially revised and has addressed all my comments in a substantive and satisfactory manner. The authors have really gone out of their way to add further experimental and in silico data.

(Remarks on code availability)
I did not review the code myself, but a colleague kindly did this and found all in order.

Reviewer #2

(Remarks to the Author)
The authors have adequately addressed all of my concerns, including the large task of developing and publishing a bioinformatics pipeline. In my opinion this manuscript is now ready for publication.

(Remarks on code availability)
Publishing the code in this way significantly enhances the reproducibility and usability of these annotations. While the code may not be of professional software engineering quality and may present challenges for some non-bioinformatic end users, I believe that is entirely acceptable in this context. It's unrealistic to expect all biologists to have expertise in developing polished, user-facing software. By releasing the code openly, the authors allow those who care deeply about the methods to inspect and understand them, while also laying a foundation for computer science experts to build upon and improve the tools moving forward.

Reviewer #4

(Remarks to the Author)
I thank the authors for the thorough replies to my comments and questions. Everything that I brought up has been addressed by the authors.

I also feel that the additional validations and specifically the addition of the "glycobif" tool make this much more relevant for the research community now.

The observation that this manuscript will more than triple the amount of protein annotations for Bifidobacteria carbohydrate utilization functions is also quite impressive (Fig. 1b, only 30% of the annotations were previously known).

(Remarks on code availability)
I already tested the companion code in the first revision and provided some feedback that was incorporated by the authors.

I tested the glycobif tool as well. The software could be installed following the provided instructions and generated the individual carbohydrate utilization predictions and the BPM matrix as expected.

Decision Letter:

Our ref: NMICROBIOL-24030909B

23rd April 2025

Dear Dr. Osterman,

Thank you for submitting your revised manuscript "Integrative genomic reconstruction of carbohydrate utilization networks in bifidobacteria: global trends, local variability, and dietary adaptation" (NMICROBIOL-24030909B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Microbiology, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Microbiology Please do not hesitate to contact me if you have any questions.

Sincerely,

Reviewer #1 (Remarks to the Author):

The manuscript has been substantially revised and has addressed all my comments in a substantive and satisfactory manner. The authors have really gone out of their way to add further experimental and in silico data.

Reviewer #1 (Remarks on code availability):

I did not review the code myself, but a colleague kindly did this and found all in order.

Reviewer #2 (Remarks to the Author):

The authors have adequately addressed all of my concerns, including the large task of developing and publishing a bioinformatics pipeline. In my opinion this manuscript is now ready for publication.

Reviewer #2 (Remarks on code availability):

Publishing the code in this way significantly enhances the reproducibility and usability of these annotations. While the code may not be of professional software engineering quality and may present challenges for some non-bioinformatic end users, I believe that is entirely acceptable in this context. It's unrealistic to expect all biologists to have expertise in developing polished, user-facing software. By releasing the code openly, the authors allow those who care deeply about the methods to inspect and understand them, while also laying a foundation for computer science experts to build upon and improve the tools moving forward.

Reviewer #4 (Remarks to the Author):

I thank the authors for the thorough replies to my comments and questions. Everything that I brought up has been addressed by the authors.

I also feel that the additional validations and specifically the addition of the "glycobif" tool make this much more relevant for the research community now.

The observation that this manuscript will more than triple the amount of protein annotations for Bifidobacteria carbohydrate utilization functions is also quite impressive (Fig. 1b, only 30% of the annotations were previously known).

Reviewer #4 (Remarks on code availability):

I already tested the companion code in the first revision and provided some feedback that was incorporated by the authors.

I tested the glycobif tool as well. The software could be installed following the provided instructions and generated the individual carbohydrate utilization predictions and the BPM matrix as expected.

Version 3:

Decision Letter:

5th June 2025

Dear Professor Osterman,

I am pleased to accept your Resource "Integrative genomic reconstruction reveals heterogeneity in carbohydrate utilization across human gut bifidobacteria" for publication in Nature Microbiology. Thank you for having chosen to submit your work to us and many congratulations.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Microbiology style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. You will not receive your proofs until the publishing agreement has been received through our system

Due to the importance of these deadlines, we ask you please us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/nmicrobiol/editorial-policies). In particular your manuscript must not be published elsewhere.

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Microbiology as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Microbiology logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

Best wishes and congratulations!

P.S. Click on the following link if you would like to recommend Nature Microbiology to your librarian
http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**

Reviewer comments

Reviewer #1 (Remarks to the Author):

In this manuscript the authors first perform a reconstruction of C-metabolic pathways in over 250 reference bifidobacterial strains based on in silico and literature data to produce 70 pathways for the utilization of a variety of carbohydrates. This information was then used to construct a so-called Binary Phenotype Matrix for gene-trait matching purposes (and validated using published carbohydrate utilization phenotypes) and applied to explore the variable carbohydrate metabolic abilities of a wider range of bifidobacterial (sub)species/strains. Strain level diversity was noticed (particularly among the B. longum species) while (predicted) carbohydrate utilization phenotypes did not always follow phylogeny. A new B. longum subspecies was proposed based on distinct (predicted) C-utilization abilities. Furthermore, apparent geographical (Bangladeshi) specific C-utilization properties were discovered among B. catenulatum strains based on a presumed xyloglucan metabolism gene cluster (as corroborated by RNAseq data). The predicted utilization abilities were in part experimentally assessed using a subset of bifidobacterial strains and employing a specific range of carbohydrates as well as using a HMO glycoprofiling approach. The obtained experimental validation was then taken to infer a genus wide reconstruction of carbohydrate utilization profiles (including 306 genomes of isolates and 2404 MAGs), while also incorporating pathways for B vitamin/amino acid biosynthesis and urea utilization. Again, species and strain level variations were observed, while findings were also investigated for age and host life style differences.

Overall, the paper constitutes a very significant amount of bioinformatic analysis, complemented with relevant experimental data. It's a timely paper that may have value beyond those focused on bifidobacterium. However, I am not convinced that the knowledge advances and novel findings are sufficient to justify publication of this manuscript in Nature Microbiology. Many of their observations confirm (perhaps in a wider context) what was published before at species or strain level.

We sincerely appreciate the reviewer's detailed and constructive feedback, which has helped us further refine and strengthen our manuscript

To enhance the novelty of our study, we expanded the metabolic reconstruction by increasing the number of analyzed genomes from 2,973 to 3,083, including nine novel genomes of Bangladeshi and Malawian isolates (sequenced by us during preparation for the resubmission). This broader dataset improves the representation of genomes recovered from non-Westernized samples.

Additionally, we generated new experimental data on the growth of 14 *Bifidobacterium* strains, including strain belonging to a novel subspecies-level clade within *Bifidobacterium longum* (*Bl. nov.*; Fig. 5). Our results demonstrate that *Bl. nov.* uniquely metabolizes soluble starch and pullulan—unlike other *B. longum* subspecies—but is unable to utilize human milk oligosaccharides (HMOs). We also provide the first phenotypic characterization of three *Bifidobacterium hominis* isolates (*Bifidobacterium sp002742445* in the previous iteration), a recently delineated species based solely on genomic criteria (https://doi.org/10.1101/2024.06.20.599854). Furthermore, we validated the ability of the Malawian isolate *Bifidobacterium adolescentis* M56B_1C3 to metabolize fucosylated HMOs, specifically 2'-fucosyllactose (2'FL), a phenomenon previously not described for this species.

Finally, leveraging growth data and HMO glycoprofiling, we confirmed that closely related strains within *Bifidobacterium longum* subsp. *suis* exhibit distinct carbohydrate preferences, representing different "ecotypes". Specifically, strain Bg131.S11_17.F6 efficiently metabolizes

multiple neutral (LNT, LNnT, LNH), long-chain neutral fucosylated, and sialylated HMOs. In contrast, strain Bg41121_2E1 preferentially utilizes short-chain neutral fucosylated HMOs (e.g., 2′FL) but does not actively consume LNnT, LNH, or sialylated HMOs.
Together, these additions substantially strengthen our findings, further demonstrating inter- and intraspecies metabolic diversity within *Bifidobacterium* and highlighting novel metabolic capabilities relevant to host adaptation and ecological specialization.

Some other comments

L66-74: This statement is perhaps too simplistic and does not take into account possible cross-feeding in the gut. HMO utilisation in the gut is not necessarily promoted by a single strain/species, but can also be achieved by cooperative cross-feeding.

We fully acknowledge that HMO metabolism in the gut is often a cooperative process involving multiple species rather than being mediated by a single strain. Indeed, a recent study has demonstrated that various species within the human gut microbiota can metabolize HMOs (PMID: 39920790), suggesting potential cross-feeding interactions.

To address this, we have added a sentence to the discussion section highlighting the importance of cross-feeding in HMO and other dietary glycan utilization:

L525-531*: A limitation of our study is that we did not account for cross-feeding interactions between bifidobacteria and other members of the gut microbial community. The lower prevalence of catabolic pathways for polysaccharides compared to their low-molecular constituents indicates that many Bifidobacterium strains might rely on syntrophic interactions — both within the genus and with keystone degraders such as Bacteroides and Segatella (Prevotella)[105–107]. Recent findings also point to a widespread capacity for HMO degradation across gut microbes[108], implying potential HMO cross-feeding between bifidobacteria and other taxa.*

The authors propose the establishment of a novel species within *B. longum*, but the authors don't indicate which genes they used for phylogenetic analysis, it is not clear how they manually refine taxonomy using ANI. I am not sure what difference between extended figures and supplemental figures are but this tree should be given more prominence).

In the revised manuscript, we have added Supplementary Table 18, which lists the 487 core genes identified using through a pangenome analysis. The concatenated nucleotide sequences of these genes were aligned, and the resulting alignment was used to construct a maximum-likelihood phylogenetic tree.

Regarding the ANI analysis, we would like to clarify that it was applied selectively to species with complex subspecies structures (*B. longum* and *B. catenulatum*) to complement the phylogenomic analysis. The methodological details are provided in Methods and Supplementary Note 1. ANI values offer a standardized genome-wide percentage identity metric with widely used thresholds for species and subspecies delineation (PMIDs: 30504855; 38228587).

Finally, we acknowledge the Reviewer's concern about the visibility of the phylogenetic tree. Due to space constraints, the tree depicting 263 reference *Bifidobacterium* genomes was placed in Supplementary Figures rather than the main text or Extended Data Figures. This decision was made to preserve phylogenetic distances and strain labels, which would have been compromised if the tree was resized or reformatted into a circular representation to fit within the main text.

Along the same line: L124-134: Average Nucleotide Identity provides an indication of species-level clustering (based on % identity of common regions between two sequences). An initial ANI clustering should be refined using an appropriate phylogenetic methodology (based on multiple sequence alignment of single-copy core genes). If the sequences selected for phylogenetic analysis are appropriate and so is the methodology for tree construction, there should be no need to manually correct phylogeny using ANI. One would argue why using phylogenetics at all!

Our primary method for taxonomic validation was a maximum-likelihood phylogenetic tree constructed from 263 *Bifidobacterium* genomes. The purpose of the additional ANI analysis for *B. longum* and *B. catenulatum* was to refine the subspecies delineation.
Tree-based branch lengths, while valuable for phylogenetic inference, do not provide universal thresholds for species/subspecies classification, as they are influenced by the alignment and parameter (e.g., evolutionary model) choices. In contrast, ANI offers a standardized, genome-wide percentage identity metric with widely used thresholds for species/subspecies delineation (PMIDs: 30504855; 38228587). Thus, ANI was used as a complementary approach to help clarify the complex subspecies structure within *B. longum* and *B. catenulatum*, ensuring robust taxonomic classification based on widely used numerical thresholds (Supplementary Note 1).

To provide more clarity, we have modified the text:

L110-115*: We first constructed a maximum-likelihood phylogenetic tree based on 487 core genes identified through a pangenome analysis to refine the taxonomic assignments of 25 strains (Supplementary Fig. 1; Supplementary Table 1). Additional pairwise average nucleotide identity (ANI) comparisons for select Bifidobacterium longum and Bifidobacterium catenulatum genomes allowed us to delineate the subspecies structure of these species (see Supplementary Note 1; Extended Data Fig. 1).*

In its current presentation the paper is difficult to interpret, the narrative is not logical and it jumps from figure to figure. For example Figure 2 is discussed, Figure 3 is skipped and then Figure 4a and 5a are next (line 186)

We appreciate the reviewer's feedback regarding the manuscript's structure and figure presentation. In the revised version, we have restructured the manuscript to improve the logical flow and readability (please refer to our response to the next comment for further details). Additionally, we have rearranged the order of the figures (main, extended, and supplemental) to ensure they appear in a better ordered sequence within the text.

Similarly, why go from making predictions in 263 genomes in Figure 2 to 2973 genomes in Figure 6? A more logical approach would be to use 263 genomes to make predictions and then test this on the larger data-set to ensure it is scalable and comparable. Then it should be phenotypically validated before moving onto the species-specific results.

In the revised version, we first describe the technical details of metabolic reconstruction in 263 reference genomes and the automated pathway prediction in 2,820 additional genomes. Within each of these sections, we now explicitly mention that our predictions align well with experimental data (please refer to our response to the following comment for further details). Following this, we present a species-to-strain funnel, incorporating the full dataset of 3,083 genomes,

To improve clarity and coherence, we have also reorganized the figures:

- The previous Fig. 6 is now Fig. 2 to reflect the broader dataset early in the manuscript.
- The previous Fig. 2 has been moved to Extended Data Fig. 4

Finally, the manuscript concludes with a detailed section on experimental validation, where we present specific examples of validated growth phenotypes, HMO consumption, and transcriptome analyses, supporting the confidence of our metabolic predictions.

The authors should validate their predictions upfront in vitro using geographically distinct strains and species on different substrates. Can they replicate Binary Phenotpye Matrix (**line 158**) predicted strain level differences in vitro? I would not expect perfect match but reader needs to be convinced their predictions are accurate before presenting results.

In the revised manuscript, we compare 38 predicted carbohydrate utilization phenotypes with *in vitro* growth data for 30 geographically distinct strains, including isolates from Bangladesh, Malawi, the US, and Europe. This dataset includes 14 additional strains analyzed during the revision process.

- Predictions for 20 out of 30 strains were obtained via manual curation, while the remaining 10 strains were analyzed using our automated pipeline termed glycobif
- The overall prediction accuracy was 95% for manually curated strains and 94% for those processed via the automated pipeline

We have modified the manuscript accordingly:

L144-148*: To assess the robustness of our reconstruction approach, we compared the predicted carbohydrate utilization phenotypes with in vitro growth data for 33 strains from previous studies[10,35,46] and performed experiments to generate similar data for 20 reference strains (see the last section). The overall prediction accuracy for both comparisons was 95% (Supplementary Table 12).*

L162-164*: The predicted carbohydrate utilization capabilities for 10 additional strains from this work were compared with in vitro growth data, resulting in a prediction accuracy of 94%* ***(Supplementary Table 12e,f)***

The detailed description experimental validation is provided in the section "Experimental assessment of predicted carbohydrate utilization capabilities in bifidobacterial isolates".

Figure 2 is too detailed and hard to take away main findings from. As just one example (line 211): "Compared to other B. longum subspecies, this distinct clade had fewer predicted host glycan utilization pathways, suggesting a lineage specific loss of genes and gene clusters". If not highlighted in the text it would not be possible for reader to see this in figure. One option would be to move this figure to supplemental and replace with a simplified version that perhaps presents broad categories of substrates together with their full names (not abbreviations) or present by species and colour boxes by % of strains of species that metabolise substrate to show strain-level variability.

To improve clarity and readability, we have:

- Moved the original Fig. 2 to Extended Data Fig. 4
- Replaced Fig. 2 with Fig. 6
- Figure 2c displays the percentage of genomes per species/subspecies encoding a specific metabolic pathway, effectively illustrating both inter- and intraspecies variability, which is now clarified in the revised figure legend

Minor comments/remarks

L171: Based on previous reports, not all species are able to utilise sucrose.

We have revised ***L186–L189*** to emphasize that our analysis focuses on the genomic potential for sucrose metabolism rather than direct experimental validation across all species:

The pathways for sucrose (Scr), maltose (Mal), malto- and isomaltooligosaccharide (MOS and IMO), melibiose (Mel), raffinose-family oligosaccharide (RFO), and short-chain fructooligosaccharide (scFOS) utilization were found in >84% of all genomes and most Bifidobacterium species, except B. bifidum.

A sucrose utilization pathway was considered present if a genome encoded at least one of the three combinations of functional roles: (ScrP AND ScrT) OR (CscA AND CscB) OR (ScrA AND ScrB) (Supplementary Tables 4;5). Based on these rules, we identified the sucrose utilization pathways in all studied species except *B. bifidum*.

We then validated the growth on sucrose for *B. adolescentis, B. breve, Bc. kashiwanohense, Bc. kashiwanohense_A, B. dentium, B. hominis, Bl. infantis, Bl. longum, Bl. suis, Bl. nov., B. pseudocatenulatum, B. scardovii*, and the lack of growth for *B. bifidum* (Fig. 5a). Sucrose utilization for the remaining major species/subspecies has been reported in the literature or databases: *B. angulatum* (bacdive.dsmz.de/strain/1685), *Ba. animalis* (bacdive.dsmz.de/strain/1736), *Ba. lactis* (PMID: 12513973), *Bc. catenulatum* (PMID: 37188713), *B. gallicum* (PMID: 2223594), *B. globosum* (PMID: 30737347), *B. ruminantium* (bacdive.dsmz.de/strain/1723), *B. tsurumiense* (PMID: 21484167).

The lack of sucrose metabolism in certain species (or, rather, strains) mentioned by the Reviewer may be attributed to strain-specific disruptive mutations or low basal expression levels of sucrose utilization genes rather than an absolute absence of the pathway.

L176-180: How these results add to previously published reports? These are not novel findings, as metabolism of HMOs, host-derived, mono- di- and polysaccharides in B. longum, B. Breve and B. Bifidum have been extensively described in literature.

AND

L181-182: Species-level differences have been described in literature before.

We appreciate the reviewer's comment and acknowledge that species-level differences within *Bifidobacterium* (including specific glycan preferences of *Bl. longum*, *B. breve*, and *B. bifidum*) have been described in the literature. However, our work considerably expands on previous reports in several key ways:

- Comprehensive pathway analysis – we extend the analysis to 68 carbohydrate utilization pathways (some of which, such as the xyloglucan degradation pathway, were reconstructed for the first time in this study) across >3,000 genomes. This provides a more detailed and nuanced view of metabolic differentiation across bifidobacterial species
- Genotype-phenotype integration – while many prior studies have largely focused on phenotypic differences (e.g., growth data), we go further by establishing direct genotype-phenotype associations. Our approach integrates genomic, functional, and experimental evidence to elucidate pathway variants responsible for metabolic differences

Thus, while species-level variations have been noted before, our study provides a much deeper, genome-informed perspective, revealing novel insights into metabolic capabilities that were previously underexplored.

L187-191: has this been experimentally proven or is only a speculation?

The specific prediction for *Bl. infantis* Bg064.S07_13.C6 was indeed experimentally validated (Fig. 5).

- This strain was confirmed to grow on xylotriose (XOS) and long-chain fructooligosaccharides (lcFOS), supporting our computational predictions
- To our knowledge, this represents the first report of a *Bl. infantis* strain growing on XOS, while the ability of *Bl. infantis* to metabolize fructooligosaccharides with dp > 4 has been previously described (PMID: 16204533)

However, our study provides new insights into the lcFOS phenotype:

- Previous reports did not unequivocally link lcFOS metabolism to a specific gene cluster. In contrast, we predict that utilization of lcFOS is dictated by the presence of a gene cluster encoding a Beta-2,1/6-fructofuranosidase (GH32) and a putative ABC-transport system for lcFOS (Supplementary Tables 4;5).
- We now differentiate the lcFOS phenotype from the more conserved ability to metabolize short-chain fructooligosaccharides (scFOS, dp 3)

L232-38: These are not novel findings

In this section, we explicitly and extensively acknowledge the studies describing the functional roles (e.g., glycoside hydrolases) constituting the pathways driving the characteristic carbohydrate utilization preferences of *Bl. longum*. However, the novel aspect of our work includes the analysis of pathway representation in 961 *Bl. longum* genomes

L247-51: Has this been proven experimentally?

Yes, we experimentally characterized the growth of *Bl. suis* Bg131.S11_17.F6 and its closely related strains (Bg41121_2E1 and M257B_A6) on 43 substrates, testing 38 predicted carbohydrate utilization phenotypes (Fig. 5).

- Our results confirmed that the fermentation profile of *Bl. suis* Bg131.S11_17.F6 was more similar to *Bl. infantis* than to the other two *Bl. suis* strains
- Specifically, unlike Bg41121_2E1 and M257B_A6, *Bl. suis* Bg131.S11_17.F6 did not grow on xylose, arabinose, xylotriose, arabinotriose, and arabinan, but grew on LNnT and 3′/6′-sialyllactoses, consistent with our predictions

Additionally, we conducted a detailed glycoprofiling analysis of HMO utilization for Bg131.S11_17.F6 (H1-cluster positive strain) and Bg41121_2E1 (FHMO-cluster positive strain, Fig. 4).

- Bg131.S11_17.F6 efficiently metabolized multiple neutral HMOs (including LNT, LNnT, and LNH), long-chain neutral fucosylated HMOs, and sialylated HMOs
- Bg41121_2E1, in contrast, preferentially metabolized short-chain neutral fucosylated HMOs (e.g., 2′FL), which were poorly utilized by Bg131.S11_17.F6. This strain also did not actively consume LNnT, LNH, or sialylated HMOs

L255-58: The strain level heterogeneity of bifidobacterial species is not a novel finding. Has been discussed extensively in literature.

We appreciate the reviewer's comment and acknowledge that strain-level heterogeneity in *Bifidobacterium* has been previously described, which we duly acknowledged in the (revised) manuscript. However, our study considerably expands on this concept in several key ways:

- Scope: previous studies have examined strain-level heterogeneity in *Bifidobacterium*, but typically for a limited set of strains and phenotypes. Our study

extends this analysis to >3,000 genomes, including novel isolates from non-Westernized populations, and 68 pathways. This allowed us to document previously undescribed examples of strain-level variability, for example, the unexpected lack of sialic acid utilization in certain Bangladeshi *B. breve* strains (**L318-322**)

- Beyond general strain heterogeneity, we identified cases where differences between closely related strains are comparable to interspecies differences. For instance, B*l. suis* Bg131.S11_17.F6 and Bl. suis Bg41121_2E1 exhibit drastically different carbohydrate utilization profiles (Fig. 4,5) a level of variability not previously reported within bifidobacterial species/subspecies

Thus, while the existence of strain-level heterogeneity is well established, our study provides a quantitative and qualitative expansion of this concept, revealing previously unrecognized metabolic diversity and ecological adaptations within the *Bifidobacterium* genus.

L440-44: This have been extensively described in literature.

With respect to *B. bifidum*, while the ability to degrade mucin-O-glycan and HMOs has been extensively studied (and we acknowledge that in the revised version), statements about the metabolism of plant glycans by this species have been controversial. While some studies report the ability of *B. bifidum* strains to grow on plant-derived di- and oligosaccharides (PMIDs: 25523018; 22562993), these findings do not fully align with our genomics-based reconstruction or with experimental data from both our work and other publications (PMIDs: 20974960; 32872165). To better reflect these discrepancies, we have modified our initial statement as follows:

L196-202: *For example, B. bifidum exhibited a notable specialization towards the metabolism of host mucin O-glycans, HMOs, and their degradation products lacto-N-biose (LNB), galacto-N-biose (GNB), and N-acetyl-D-glucosamine (GlcNAc) — consistent with previous reports[50] (Fig. 2c). Additionally, 91% of genomes of this species possessed a novel putative pathway for utilizing the rare monosaccharide L-xylulose (L-Xul). At the same time, B. bifidum genomes lacked complete pathways for utilizing plant-derived mono-, di-/oligo- and polysaccharides, even though some studies have reported growth on these substrates[37,51,52].*

L504: Given the high level of strain specialization in utilizing carbohydrates exist in bifidobacteria, predictive approaches applied to MAGs may not be reflective of strain-level abilities.

We agree with the reviewer that estimating strain-level variability using MAGs has inherent limitations. First, MAGs do not always represent individual strains, as they are often composite assemblies derived from mixed microbial populations, making direct comparisons to cultured isolates challenging. Second, MAGs reconstructed from short-read sequencing data are more susceptible to incompleteness and contamination, which can introduce artificial variability—missing genes may lead to an underestimation of metabolic potential, while contamination can inflate pathway predictions. To minimize these biases, we applied more strict selection criteria than recommended by MIMAG standards, retaining only MAGs with very high completeness (>97%) and low contamination (<3%) to encourage reliable metabolic reconstructions. Additionally, we performed dereplication to prioritize the most representative MAGs, maximizing the likelihood that they correspond to individual strains rather than mixed populations. We have modified the discussion section to explicitly state these limitations

***L522-524**: A portion of observed strain-level variability may be artificially inflated by incompleteness or contamination of MAGs assembled from short-read data, underscoring the need to generate more high-quality hybrid MAGs and isolate genomes.*

Supplementary Figure1: Maybe this tree is missing bootstrap values? Are the authors sure that all nodes returned a 100 bootstrap value? Also, a list of genes used to compute this phylogeny should be indicated.

The bootstrap values are already reflected in Supplementary Fig. 1 using color coding, where nodes with poor bootstrap support are marked in red. Most nodes in the tree indeed have high bootstrap support, which is consistent with previous studies. For example, a prior analysis reported a phylogenetic tree of the *Bifidobacterium* genus (built using a similar approach) where almost all nodes had bootstrap values 100 (PMID: 33777340). We have included the list of genes used to construct the phylogenetic tree in Supplementary Table 18 in the revised manuscript.

Reviewer #1 (Remarks on code availability):I did not review the code, but a colleague of mine had a look at it (and did not seem to find any issues)

Besides the revised code for the manuscript, we have also made the annotation and pathway prediction pipeline publicly available as a standalone tool "glycobif" (https://github.com/Arzamasov/glycobif). Glycobif uses amino acid FASTA as input and outputs two tables: (i) a binary phenotype matrix (BPM) that contains the binary representation of carbohydrate utilization pathways (1 = pathway is present, 0 = pathway is absent; for isolates, these values can be directly interpreted as phenotypes), (ii) a table with the representation of functional roles (1 = role is present, 0 = role is absent) for each pathway. The annotation files with locus tags of annotated proteins can be found in the tmp folder.

Reviewer #2 (Remarks to the Author):
In this manuscript Arzamasov et. al. conduct a comprehensive investigation into the carbohydrate utilization potential of bifidobacteria. The authors perform an expert-led and manual curation of carbohydrate utilization across a few hundred reference genomes, verify the accuracy of their metabolic predictions using both previously generated and newly generated data, and then extend their predictions to a large set of a few thousand lower-quality genomes to assess bifidobacteria diversity across the globe.
This work is exceptionally thorough, expertly executed, and fills a significant knowledge gap in the field. Bifidobacteria are increasingly recognized as important to human health, and this work presents a much-needed foundational understanding of bifidobacteria genomics. The manuscript stands out for its meticulous attention to detail in the manual curation of bifidobacteria carbohydrate utilization pathways, as well as for its successful application of these detail-oriented methods to several large genomic databases. This paper will undoubtedly serve as a foundational text in the field of bifidobacteria genomics for years to come, providing both newcomers with a thorough introduction and experts with a valuable reference for their ongoing work.
I have no critiques or concerns about any of the methods or results, and the level of detail presented in the methods section and supplementary tables is exemplary. However, I do have a few minor questions, the clarification of which might be helpful for future readers.

We sincerely thank the reviewer for their positive feedback and appreciation of our manual curation efforts.

Q1) Would it be feasible to provide a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper? It would be great to enable others to take advantage of your manually curated pathways.

We thank the reviewer for this suggestion. To enable others to annotate *Bifidobacterium* genomes using our manually curated database, we have made our annotation and pathway prediction pipeline available as a standalone tool, "glycobif" (GitHub: https://github.com/Arzamasov/glycobif). This tool uses amino acid FASTA files (.faa) as input and performs DIAMOND-based annotation against a manually curated database. It then predicts the representation of 68 curated carbohydrate utilization pathways and 589 functional roles, providing a genotype-based inference of carbohydrate metabolism. For genomes from cultured isolates, the predicted pathways can be directly interpreted as carbohydrate utilization phenotypes.

The pipeline requires DIAMOND, Python, and R libraries, and we provide a YML file for conda/mamba environment setup, along with detailed instructions on installation, usage, and known limitations. The best results can be obtained when genomes of human-colonizing bifidobacteria with high completeness and low contamination (ideally, complete or draft genomes of cultured isolates or MAGs assembled using a combination of short and long reads) are used as input.

Additionally, for users interested in manual curation and comparative genomic analysis, we suggest using public tools such as BV-BRC (https://www.bv-brc.org/) and The SEED database (https://pubseed.theseed.org/FIG/seedviewer.cgi), which allow for a detailed examination of genomic loci across genomes.

Q2) Around lines 412-415, could you describe why you need a random forest model to predict carbohydrate utilization pathways in the full set of genomes? Why couldn't you use the same carbohydrate utilization pathway prediction method used for the 263 reference genomes?

The primary limitation of applying the mcSEED-based reconstruction approach used for the 263 reference genomes is its poor scalability. One of the most time-intensive aspects of this approach is the manual refinement of gene annotations, which includes verifying the correct propagation of annotations to orthologous genes in other genomes. This process often requires detailed examination of gene neighborhoods across multiple genomes (e.g., to separate paralogs), making it impractical for application to thousands of genomes.

To address this challenge, we implemented a custom pipeline ("glycobif"), which leverages our manually curated reference dataset to predict carbohydrate utilization pathways and associated metabolic phenotypes in an automated manner. We believe that our reference set of genomes captures the functional and phylogenetic diversity of human-associated bifidobacteria, ensuring reliable predictions. Notably, glycobif performance was validated against 10 non-reference strains, achieving an overall phenotype prediction accuracy of 94%, which is comparable to the 95% accuracy obtained for the 20 manually curated reference strains.

Q3) In Figure 2 and Figure 6a,b, it's very difficult to match the colors to the species in the legend. In Figures 2 and 6a I suggest adding additional text in the figures themselves to point out the major species, and in Figure 6b I suggest adding taxonomic labels along the y-axis.

We have modified the respective figures to improve clarity. Figure 6a (now Figure 2a in the revised manuscript) has been updated to display "spiders" (centroids with lines connecting them to individual points) instead of individual points, making species-level patterns more apparent. Additionally, we have added text labels to the centroids corresponding to major species for better readability. For Figure 6b (now Figure 2b), due to space limitations, we were unable to include text annotations directly on the figure. Instead, we have adjusted the colors and reordered the boxplots (top to bottom) to match the taxonomy order in the legend. Furthermore, the old Figure 2 has been reassigned as Extended Data Figure 4, where we have added text labels as suggested.

Q4) Reviewer #2 (Remarks on code availability):

As stated in the main review, a tutorial guiding users on how to annotate their own bifidobacteria genomes using the methods described in this paper would be a welcome addition to the code.

As mentioned previously, our annotation and phenotype prediction pipeline is now available as a standalone tool "glycobif" (https://github.com/Arzamasov/glycobif). We have provided instructions on the GitHub repository. Briefly, Glycobif uses faa (amino acid FASTA) as input and outputs two tables: (i) a binary phenotype matrix (BPM) that contains the representation of carbohydrate utilization pathways and corresponding phenotypes (1 = pathway is present, 0 = pathway is absent; for isolates, these values can be directly interpreted as phenotypes), (ii) a table with the representation of functional roles (1 = role is present, 0 = role is absent) for each pathway. The annotation files with locus tags of annotated proteins can be found in the tmp folder.

Reviewer #4 (Remarks to the Author):

Bifidobacteria are common commensals in the human gut microbiome and were one of the earliest commercialized probiotics. Their consistent association with health and their importance in the early establishment of the infant microbiome has generated significant interest in understanding their niche in the human gut microbiome. This has been hampered by the large metabolic heterogeneity in the Bifidobacteria genus, especially their metabolism of human milk oligosaccharides (HMOs) which are still not well characterized. Arzamasov provide a detailed reconstruction of carbohydrate metabolism in Bifidobacteria based on 253 reference genomes. Experimental validation in 16 Bifidobacteria strains showed good accuracy in recapitulating growth on specific carbon sources. The authors also predicted carbohydrate usage patterns in more than 2,700 additional genomes. I do feel that the presented curated reconstruction of carbohydrate utilization expands the current knowledge significantly and the manuscript includes some convincing applications. Unfortunately, there is very little description how this reconstruction is achieved and the relevant data is not provided to the scientific community. I also have some concerns regarding the prediction of carbohydrate utilization for the ~2,700 which is currently lacking validation. This could be alleviated with some better description of the methodology, providing additional data, and some experimental validation of the prediction pipeline. All of this is definitely in the realm of the authors' capabilities and would improve the manuscript significantly while also providing a valuable resource for other researchers.

We sincerely thank the reviewer for the detailed and valuable feedback, which greatly helped us refine our manuscript. In response:

- We have expanded the manuscript to clarify the workflow behind our curated reconstruction of carbohydrate utilization. We emphasize that, by building on our reference dataset and propagation pipeline, users need not re-do the extensive manual reconstruction and literature review (spanning five years) that underpins this work
- We now provide “glycobif,” a standalone, publicly accessible tool (available at https://github.com/Arzamasov/glycobif) that allows predict carbohydrate utilization pathways in *Bifidobacterium* genomes
- To address concerns about our predictions for ~2,700 additional genomes, we tested “glycobif” on 10 strains not included in our original reference set. The resulting predictions showed an accuracy (94%) comparable to that for the reference strains (95%), supporting the robustness of our phenotype projection approach

Major comments

For me, one key part of the manuscript is the metabolic reconstruction of Bifidobacteria carbon metabolism. Thus, is was a bit surprising to find very little description how this was achieved. The description in lines 135-140 is too superficial and the Methods only mention the use of the "mcSEED" platform without providing a reference URL to the method/platform (the references used 46/47 used in the text do not describe mcSEED). I would ask the authors for the following:

1. Provide a detailed description of how the subsystems were generated in mcSEED. Also describe how orthologous gene clusters were established (method to infer homology and propagate functional roles).

mcSEED is a private clone of the publicly available pubSEED platform and has been maintained separately by our group. The metabolic reconstruction was performed using mcSEED through an iterative process of annotation refinement and subsystem curation, as follows:

- Genomes were first annotated using RAST (presently hosted at BV-BRC https://www.bv-brc.org/app/Annotation) and then uploaded into mcSEED. An empty subsystem was created, which can be conceptualized as a large table where rows represent genomes and columns represent functional roles within a given pathway or set of pathways
- We then started creating a list of functional roles by mining the literature and updating the initial RAST-generated annotations for a subset of reference *Bifidobacterium* genomes (e.g., type strains) where these functions had been characterized
- Once the initial set of functional roles within a subsystem was created, we propagated the respective annotations to all 263 reference *Bifidobacterium* genomes. Orthologous genes were first assigned automatically using predefined protein family classifications such as PGFams (cross-genus protein families) and PLFams (genus-specific protein families; PMIDs: 26903996; 31667520), which group similar proteins together to aid in ortholog identification. We then manually checked and refined the initial ortholog assignments by analyzing genomic context for each functional role and using sequence identity as supplementary information. Generally, genes with conserved genomic neighborhoods were classified as orthologous, while paralogs were split and assigned distinct annotations when appropriate
- After capturing the representation of characterized functional roles, we tried to fill potential pathway gaps by predicting novel functional roles using three genome context techniques: (i) clustering of genes on the chromosome (operons), (ii) co-regulation of genes by a common regulator, and (iii) co-occurrence of genes in a set of related genomes. For example, this allowed us to predict the function of uncharacterized transporters that are co-localized in operons with characterized glycoside hydrolases. We also used regulon reconstruction to annotate genes that are potentially co-regulated with characterized carbohydrate metabolism genes. During this iterative process we created the final list of functional roles in each subsystem
- We then designed the rules for assigning detailed pathway variants based on the representation of specific sets of functional roles (i.e., genomic signatures): (i) transporter + catabolic pathway ("U"), (ii) catabolic pathway without transporter ("P"), (iii) no catabolic pathway ("N") (Extended Data Fig. 2; Supplementary Table 5)
- Finally, detailed pathway variants were converted to simplified predicted carbohydrate utilization phenotypes: "U" to a numeric binary phenotype "1" (utilization or degradation), "P" and "N" to a numeric binary phenotype "0" (no utilization or degradation)

We have amended the “Subsystems-based annotation and *in silico* metabolic reconstruction of glycan utilization pathways” subsection in Methods.

2. Provide the specific gene clusters (sequences) for the functional roles along with the annotations (description, subsystems, substrates and products) so this can be used by other researchers.

We have added an additional sheet in Supplementary Table 7, titled “all annotations”, which provides a comprehensive list of 39,589 protein-coding genes and their functional annotations from 263 reference *Bifidobacterium* genomes. These genes correspond to 589

functional roles curated across 25 subsystems. The protein sequences associated with these functional roles have been made publicly available in the "Bif_263_subsystem_proteins.zip" file, which has been uploaded to Figshare (doi.org/10.6084/m9.figshare.26053936). Other sheets of this table depict 25 metabolic subsystems and contain: (i) the distribution of 39,589 genes across 25 subsystems, (ii) detailed pathway variants for all pathways present within each subsystem, (iii) predicted binary carbohydrate utilization phenotypes.

The detailed descriptions of functional roles are given in Supplementary Table 4.

- Column "pathway" lists pathway(s) to which the functional role contributes to. Further details on the metabolic pathways and corresponding phenotypes can be found in Supplementary Table 3
- Column" function_description" includes the detailed information about the function (e.g. substrates and linkage specificity for glycoside hydrolases).
- Columns "example_mcSEED_id" and "example_locus_tag" list representative gene IDs (searchable within Bif_263_subsystem_proteins.faa) and locus tags (searchable via NCBI) for all functional roles of interest

The validation with the 16 reference strains is quite comprehensive and the additional transcriptomics does support the claims that the metabolic reconstruction does indeed recover carbohydrate utilization adaptations on the strain level. However, it should also be noted that the experiments on those 16 strains is the current extent of validation in the manuscript, despite all the additional predictions. It seems like those 16 strains were also part of the manually reconstructed set of 263 genomes, so it is essentially a best-case evaluation. I wonder how well those predictions would extend to genomes that are geographically or genomically different from the used reference (strain) genomes.

It would be helpful if the authors could also test some more distant strains from the 2,700 genome prediction set. This would help to establish a bound on the prediction accuracy which is currently hard to evaluate as also argued below. Right now, the observed increased heterogeneity in carbohydrate utilization phenotypes in that genome set could also be a consequence of model bias.

We thank the reviewer for the suggestion. In the revised manuscript, we additionally measured the growth of 14 additional *Bifidobacterium* strains on 43 substrates *in vitro* to test 38 predicted carbohydrate utilization phenotypes (Fig. 5a). Among these, four strains belonged to the reference collection analyzed to further support the predicted phenotypic heterogeneity within *Bifidobacterium longum*, including the experimental characterization of a proposed novel subspecies-level clade (*Bl. nov*).

The remaining 10 strains were of Malawian and Bangladeshi origin, isolated in previous studies (PMIDs: 26898329, 31296738, 31296739), but their genomes were sequenced during this revision. This set included three *Bifidobacterium hominis* isolates (GTDB: *Bifidobacterium sp002742445* in the previous iteration), a recently delineated species based on genomic criteria (https://doi.org/10.1101/2024.06.20.599854) and not yet phenotypically characterized. Additionally, we tested *Bifidobacterium adolescentis* M56B_1C3, a strain of Malawian origin predicted to encode fucosylated HMO utilization pathways (e.g., 2′-fucosyllactose (2′FL)) not previously described in this species, making this genome unique among 22 reference and 478 additional *B. adolescentis* genomes.

We performed the phenotype prediction for these 10 strains using the "glycobif" pipeline. The overall accuracy prediction of 94% was comparable to the accuracy for 20 reference strains

(95%), for which phenotypes were predicted via manual reconstruction (Supplementary Table 12f). The detailed comparison between predicted phenotypes observed *in vitro* growth data is shown in Fig. 5b and Supplementary Table 12e. As part of it, we confirmed the predicted ability of *B. adolescentis* M56B_1C3 to grow on 2'FL.

The *B. adolescentis* M56B_1C3 strain example demonstrates that we may underpredict the strain-level heterogeneity within bifidobacterial species. Different understudied populations worldwide may likely contain other region-specific strains with a unique representation of carbohydrate utilization pathways.

Regarding the random forest prediction of phenotypes it would be great to provide some motivation for the chosen methodology in lines 675-714. Why were only 27 subsystems chosen even though 70 were reconstructed initially? How exactly were the outgroup proteins chosen? DIAMOND is capable of performing local alignments which are often used in functional annotation pipelines such as EGGNOG etc. So I wonder what the motivation for the additional domain clustering was.

We have provided the answers to the respective questions below.

1. Why were only 27 subsystems chosen even though 70 were reconstructed initially?

We apologize for any confusion regarding the terminology of "subsystems" and "pathways." In our manuscript, subsystems are artificial constructs—structured tables that represent functional roles and pathway variants—and **most subsystems encompass multiple pathways** (please also see a response regarding your second question about subsystems below). This distinction explains the discrepancy between the number of subsystems and the total number of reconstructed pathways. In the revised manuscript, we have revised our metabolic reconstruction, resulting in an updated count of 25 subsystems (previously 27) and 72 reconstructed pathways/inferred "metabolic phenotypes" (previously 70) across the 263 reference genomes.

Among the 72 reconstructed pathways, 68 were incorporated into the propagation pipeline, while the remaining four pathways (ManAc, GalNAc, Man, and GalA) were excluded. These pathways could not be propagated because all reference genomes contained only downstream catabolic pathways but lacked characterized or predicted transporters for the corresponding carbohydrates (pathway variant P converted to a binary phenotype 0)

2. How exactly were the outgroup proteins chosen?

We defined outgroup proteins as all proteins (encoded in the 263 reference genomes) that were not captured in the 25 curated subsystems.

3. DIAMOND is capable of performing local alignments which are often used in functional annotation pipelines such as EGGNOG etc. So I wonder what the motivation for the additional domain clustering was.

We implemented a clustering procedure to minimize the number of false positive and false negative annotations. Unlike domain-based approaches such as Pfam, our annotation method assigns functions based on whole amino acid sequences rather than identifying individual domains. This can introduce misannotations in two ways: (i) if the best DIAMOND hit corresponds to only part of a multidomain protein, leading to an incorrect assignment of multiple functions, or (ii) if the best hit lacks certain domains present in the query sequence, resulting in a loss of some functions. We observed this issue during our previous work (PMID: 38093016) for certain groups of proteins, for example, PTS (phosphotransferase system) components, which can be encoded as either a single gene (multidomain protein) or separate genes (individual functional units IIA, IIB,

and IIC). If a multidomain PTS component (e.g., IIABC) is present in the database but only a portion of the protein aligns with the query, this can lead to incorrect functional assignment, but our pipeline will assign the correct function (e.g., IIA) function rather annotating it as a full IIABC unit.

In our previous study (PMID: 38093016), we empirically observed that this clustering procedure becomes less relevant as the reference database grows larger and contains more complete annotations. However, even if these misannotations affect only a small subset of proteins, they often systematically impact specific functional categories, such as transporters.

In a similar vein, I did not understand the necessity for the Machine Learning. Presence/absence of the subsystems inferred in the reconstruction can be translated directly into a BPM matrix as done before in the manuscript in lines 662-669. Was this necessary to overcome missing gene clusters/functional roles? A machine learning model would also be somewhat less mechanistic because the predicted carbohydrate utilization could in theory be inferred from unrelated subsystems/gene clusters which would be hard to interpret.

In case the machine learning is indeed necessary, it might be a bit more stringent to use a train-test-validation setup where the hyperparameter training is done on the train-test set combination and the final model is tested on a hold-out validation set. As argued above, experimental validation for some strains from the validation step would strengthen the manuscript as well.

The primary goal of using machine learning (ML) in our study was to replicate the logic of our manual curation process while, as the Reviewer correctly points out, making it slightly less stringent to account for the lower genome completeness and potential contamination often observed in MAGs. We aimed to make our models as mechanistic as possible, ensuring that the ML predictions closely mirrored the manual pathway assignment criteria used in our curation process. To achieve this, we manually curated the lists of functional roles used for training the ML model for each pathway. These curated lists were designed to resemble the genomic signatures we used to assign pathway variants during manual curation (Supplementary Table 5). For example, in the xylose (Xyl) pathway, we included functional roles driving xylose catabolism (XylA, XylB) and relevant transporters (XylFGH, AraFGH, FruEFGK), while excluding functional roles related to xylooligosaccharide and xylan metabolism, even though they were part of the same subsystem. This ensured that the model captured pathway-specific functional signatures rather than overgeneralizing based on all functional roles present in the subsystem.

We did not implement a more strict approach (e.g., “if roles A, B, C are present, then assign 1; otherwise, assign 0”), as was done in a previous study (PMID: 38093016), because some pathways were too complex to encode in a rule-based manner, and the time required for troubleshooting scripts and updating annotations was very high. Additionally, we avoided the neighbor-based approach (i.e. use the information about pathway variants in closely related genomes) used in the previous publication, as it sometimes incorrectly treated rare pathway variants that were absent or highly uncommon in the reference model. While this ML approach is not ideal, we have experimentally established that its prediction accuracy (94%) is comparable to manual curation (95%). Furthermore, glycobif provides detailed output on the representation of functional roles, allowing users to manually refine pathway predictions using their own expertise if needed.

Lines 610-613 do not specify how consent for the fecal sampling from Bangladeshi children was obtained and whether the legal guardians consented to the additional use in this study. Please also indicate the local collaborators handling the sampling.

We have modified the following section to:

L574-587: *Fecal samples that were used for culturing Bangladeshi bifidobacterial strains were collected during the course of studies conducted by the International Centre for Diarrhoeal Disease Research (icddr,b). These were (i) the MAL-ED birth cohort study of children aged 0-24 months (Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development; ClinicalTrials.gov identifier NCT02441426) and (ii) a cohort of healthy 12-24 month-old Bangladeshi children enrolled in parallel with children with acute malnutrition in a study of microbiota-directed complementary food (MDCF) prototypes (ClinicalTrials.gov identifier NCT0308473)[47,48] These studies were approved by the Ethical Review Committee of the icddr,b. Bifidobacterial strains were also isolated from fecal samples collected in a previously reported study of Malawian twins discordant for acute malnutrition[49]. The protocol was approved by the College of Medicine Research Ethics Committee of the University of Malawi, and by the Human Research Protection Office of Washington University in St. Louis. Written informed consent including provisions for future use of materials was provided by the parents or guardians of participating children before enrolment.*

Minor comments

It was not obvious to me which proportion of the 525 functional roles/gene clusters are currently *not* represented in the common protein annotation pipelines and databases. It would be helpful if the manuscript could point out what fraction is indeed new or is already covered by databases such as EGGNOG, ModelSEED, KEGG, etc.

We thank the reviewer for the suggestion. To assess the extent to which our manually curated annotations are already represented in common protein annotation pipelines, we conducted a comparative analysis of the 39,589 genes curated from 263 reference *Bifidobacterium* genomes. This comparison was performed against annotations assigned by Prokka and EggNOG-mapper. Given that EggNOG-mapper outputs multiple annotation fields, including description, EC number, and KEGG_ko numbers, we aggregated information from these fields, prioritizing external text descriptions of EC and KEGG_ko numbers. The results of this analysis are presented in Fig. 1b and Supplementary Table 19.

We categorized the comparison into four groups:

- *new:* manually curated annotations provide novel functional insights beyond general or non-specific automated annotations. For example, we use annotation 2′FL, 3FL ABC transporter substrate-binding component (TC 3.A.1.1), whereas Prokka labeled it as a hypothetical protein, and EggNOG assigned it as a multiple sugar transport system substrate-binding protein
- *corrected*: manually curated annotations correct incorrect specific annotations assigned by automated tools. For example, D-mannose isomerase (EC 5.3.1.7) was misannotated by Prokka as Sulfoquinovose isomerase and by EggNOG as N-acylglucosamine 2-epimerase
- *refined*: manually curated annotations increase specificity, such as specifying linkage preference or substrate specificity for glycoside hydrolases (GHs). For example, we refined an annotation from alpha-L-fucosidase [EC:3.2.1.51] (EggNOG) to alpha-1,6-L-fucosidase (GH29)

- *same* = cases where automated annotations and manually curated annotations were identical.

We have added the following lines to the manuscript:

L125-131: *The curated set of 589 roles — comprising 235 components of glycan-specific transporters, 197 catabolic CAZymes, 72 downstream catabolic enzymes, and 85 transcription factors — was used to functionally annotate 39,589 out of 541,418 protein-coding genes in 263 reference Bifidobacterium genomes (Supplementary Tables 4,7). Manual curation improved 76.6% and 69% of functional annotations compared to the automated tools Prokka and EggNOG-mapper, respectively, including the improvement of over 90% of transporter and transcriptional regulator annotations (Fig. 1b).*

The proportion of curated functional roles not present in common annotation pipelines was particularly high for transporters and transcriptional regulators, which is expected given the insufficient incorporation of experimental data into public databases. However, many of the newly assigned annotations have been previously described or proposed in previously published studies, which is why we extensively mined the literature. Information regarding the experimental validation status of functional roles—whether they have been experimentally characterized in *Bifidobacterium*, predicted in previous studies, or newly predicted in this work — is provided in Supplementary Table 4 (columns "evidence" and "evidence info").

In the manuscript subsystems and pathways are used interchangeably, however not all subsystems are valid pathways IMHO. Pathways usually require some kind of linear chain of substrate-product conversions or modification. Were all subsystems completely connected biochemical pathways?

We apologize for any confusion caused by our use of terminology regarding subsystems and pathways. In the original paper that introduced the subsystem concept (PMID: 16214803), subsystems were designed as a generalized framework for biological processes, encompassing collections of functional roles that implement a specific biological function or structural complex. While many early subsystems had a one-to-one correspondence with individual metabolic pathways, the subsystems created in this study are more complex, often encompassing multiple pathways and, in some cases, alternative pathway variants (e.g., extracellular degradation vs. intracellular utilization).

The primary reason for grouping multiple pathways within a single subsystem is that many carbohydrate utilization pathways exhibit a nested structure, where shared functional roles exist between different metabolic processes. For example, the xylan degradation pathway consists of three distinct stages: extracellular degradation of the xylan backbone by endo-1,4-xylanase, uptake and hydrolysis of released xylooligosaccharides (XOS), and downstream conversion of released xylose into xylulose-5-phosphate. Since the xylan degradation pathway shares functional roles with both XOS and xylose utilization pathways, grouping these pathways within a single subsystem alleviates curation and helps prevent false positive pathway assignments. For instance, if a genome encodes endo-1,4-xylanase but lacks XOS transporters and downstream catabolic enzymes, it should not be classified as encoding a complete xylan degradation pathway.

To summarize, subsystems in this study are not equivalent to pathways but instead function as artificial constructs designed to facilitate carbohydrate utilization pathway reconstruction and improve pathway variant assignment accuracy.

There are a few typos in the figures, for instance "Bangaldesh" and "pawthays" in Fig. 1. I would recommend to revise the figures again.

We thank the Reviewer for catching these typos. We have fixed these and additional typos found while proofreading the Supplementary Tables.

lines 596-600: Genome were pooled from different studies here which used different sequencing and assembly pipelines. Was there some check that this did not show biases in the phenotype matrix?

We did not implement a specific bias check to account for differences in sequencing and assembly pipelines across studies. However, we acknowledge that such biases are most likely to affect MAGs assembled from short-read sequencing data, as differences in the resulting genome completeness and contamination may influence downstream metabolic predictions. To partially mitigate this issue, we applied strict completeness and contamination thresholds (97% and 3%, respectively), which are much higher than the commonly used standards for high-quality MAGs (90% completeness and 5% contamination).

Despite these stringent selection criteria, we recognize that they may still not be entirely ideal, as we observed anecdotal cases where CheckM overestimated completeness and underestimated contamination. These inconsistencies became apparent when manually analyzing the outputs of our glycobif pipeline, where certain genomes were missing genes expected to be highly conserved across all studied species (e.g., Fructokinase (EC 2.7.1.4)) or contained gene clusters from different species (such as rare occurrences of orphan oligosaccharide and polysaccharide metabolism genes in *Bifidobacterium bifidum* where downstream catabolic pathways are absent).

We believe that the main limitation of CheckM is that it estimates genome completeness and contamination based on universal marker gene sets, which may not be detailed enough to detect the gain or loss of accessory genes that are species- or subspecies-specific. Certain carbohydrate utilization genes fall into this category, making them more susceptible to assembly-related biases. Given the importance of this issue, we plan to further investigate these biases in future studies, including comparisons with CheckM2, using our reconstruction of isolate genomes as a reference.

The phrase in line 85-87 can be misinterpreted to mean that all Bangladeshi children do not receive enough breast milk. Maybe reformulate to "[...] may not be sufficiently stable in a sub-population of Bangladeshi children who did not receive [...]".

We thank the Reviewer for the suggestion, we have amended the respective sentence:

L77-79: *For instance, a probiotic Bl. infantis strain did not durably engraft in the microbiota of malnourished Bangladeshi infants whose diets were low in breastmilk compared to complementary foods*[28].

Please provide a reference for the statements made in lines 98-100.

We have cited a recent perspective article (https://doi.org/10.3389/fsysb.2024.1394084) that summarizes current challenges in annotating bacterial transporters, providing context for our statements. To our knowledge, most studies on bifidobacterial carbohydrate metabolism primarily focus on the analysis of CAZyme representation, rather than explicitly investigating transporters. For example, we selected references 9, 31–34, which follow this approach, though many additional studies exist that similarly emphasize CAZymes while overlooking transporter annotation.

Although genome-scale metabolic modeling (GEMs) studies typically include transporters in their reconstructions, many annotation errors persist due to insufficient manual curation, as discussed in the cited perspective article. Regarding transcription factors, our statement is primarily based on findings from our own work (e.g., Fig. 1b).

Line 590: Please indicate whether checkM version 1 or 2 was used since they give very different results.

All original publications, from which additional isolated genomes and MAGs were collected, used CheckM (v1.0.7-v1.0.13) to estimate genome completeness and contamination. We believe CheckM, as currently listed in the manuscript, correctly represents the name of the used tool (https://github.com/Ecogenomics/CheckM) since the newer iteration has a different name (CheckM2; https://github.com/chklovski/CheckM2).

Line 683: How was the limit of 50 hits chosen?

The limit of 50 hits was chosen empirically based on a manual analysis of DIAMOND results, particularly in cases where proteins were incorrectly assigned additional functions due to their best hits being multidomain proteins. We observed that this threshold provided the best performance, which we evaluated using two validation approaches.

First, we applied our annotation pipeline to additional genomes that were not used to build the DIAMOND database, for which we had also conducted manual curation of annotations and predicted phenotypes. Second, we performed a leave-one-out approach, where we excluded specific genomes from the DIAMOND database and then attempted to re-annotate them using the remaining database. These tests confirmed that a top 50 hit limit minimized incorrect function assignments while maintaining high annotation accuracy.

Line 774: Did the human milk donors match the geographic region or population the strains were obtained from?

The pooled HMO mixture was collected from multiple individual donors, primarily from North America, with the goal of capturing a broad range of HMO structures. In our HMO glycoprofiling experiments, we tested two American (LFYP), one German (the type strain of *Bl. infantis*), and seven Bangladeshi strains. Therefore, for most strains, the HMO donor population did not match the geographic origin of the strains.

However, we note that the ratios of HMO structures in the pooled mixture were similar to those reported in milk from Bangladeshi mothers with a secretor phenotype (PMID: 34993388). This suggests that our pooled HMO mixture could serve as a reasonable proxy for HMOs encountered by Bangladeshi infants from secretor mothers. We acknowledge, however, that the HMO profile in our mixture differs from that of nonsecretor Bangladeshi mothers. Unfortunately, we did not have access to milk samples from nonsecretor mothers and do not have information regarding the secretor status of the mothers of infants from whom the strains were originally isolated.

Remarks on code availability:

The repository is well documented and includes the analyses *after* the annotation/reconstruction with mcSEED (see related comments above). The provided R markdown notebook (`Sup_code_file.Rmd`) in the GitHub repository could be run and did

generate the figures shown in the paper. The renv environment setup did not work for me, but was not necessary to run the notebook.

Due to the manual nature of annotation and reconstruction in mcSEED, this process cannot be fully formalized into code. However, to facilitate reproducibility and extend our analysis to additional genomes, we have made our annotation and phenotype prediction pipeline available as a standalone tool, “glycobif” (GitHub: https://github.com/Arzamasov/glycobif).

Glycobif uses faa (amino acid FASTA) as input and outputs two tables: (i) a binary phenotype matrix (BPM) that contains the binary representation of carbohydrate utilization pathways (1 = pathway is present, 0 = pathway is absent; for isolates these values can be directly interpreted as phenotypes), (ii) a table with the representation of functional roles (1 = role is present, 0 = role is absent) for each pathway. The annotation files with locus tags of annotated proteins can be found in the tmp folder.

We have also removed the renv environment from the GitHub repository.

Reviewer comments

Reviewer #1:

Remarks to the Author:

The manuscript has been substantially revised and has addressed all my comments in a substantive and satisfactory manner. The authors have really gone out of their way to add further experimental and in silico data.

Remarks on code availability:

I did not review the code myself, but a colleague kindly did this and found all in order.

We sincerely thank the reviewer for the positive and encouraging feedback, as well as for the time and effort they dedicated to evaluating our work. Their input has been invaluable in helping us improve the quality of our manuscript.

Reviewer #2:

Remarks to the Author:

The authors have adequately addressed all of my concerns, including the large task of developing and publishing a bioinformatics pipeline. In my opinion this manuscript is now ready for publication.

Remarks on code availability:

Publishing the code in this way significantly enhances the reproducibility and usability of these annotations. While the code may not be of professional software engineering quality and may present challenges for some non-bioinformatic end users, I believe that is entirely acceptable in this context. It's unrealistic to expect all biologists to have expertise in developing polished, user-facing software. By releasing the code openly, the authors allow those who care deeply about the methods to inspect and understand them, while also laying a foundation for computer science experts to build upon and improve the tools moving forward.

We sincerely thank the reviewer for their positive assessment and thoughtful remarks, which have been invaluable in improving the quality of our manuscript. We fully acknowledge that the code does not meet professional software engineering standards. Our primary goal was to make the code/pipeline publicly available, and we hope that more experienced researchers in the community will be able to build upon and enhance it. We also plan to update the curated compendium of functional roles as new insights emerge in future studies, particularly those related to the functional characterization of monosaccharide and human milk oligosaccharide (HMO) transporters.

Reviewer #4:

Remarks to the Author:

I thank the authors for the thorough replies to my comments and questions. Everything that I brought up has been addressed by the authors.

I also feel that the additional validations and specifically the addition of the "glycobif" tool make this much more relevant for the research community now.

The observation that this manuscript will more than triple the amount of protein annotations for Bifidobacteria carbohydrate utilization functions is also quite impressive (Fig. 1b, only 30% of the annotations were previously known).

Remarks on code availability:

I already tested the companion code in the first revision and provided some feedback that was incorporated by the authors.

I tested the glycobif tool as well. The software could be installed following the provided instructions and generated the individual carbohydrate utilization predictions and the BPM matrix as expected.

We sincerely thank the reviewer for their encouraging feedback and for the time and effort they devoted to evaluating our work. Their thoughtful input has been instrumental in improving the quality of our manuscript.