High-resolution analysis of the treated coeliac disease microbiome reveals increased inter-individual variability

Authors:

Jelle Slager[1], Hanna L. Simpson[1], Ranko Gacesa[1,2], Lianmin Chen[1,3,4,5], Jody Gelderloos[1], Astrid Maatman[1], Cisca Wijmenga[1], Alexandra Zhernakova[1], Jingyuan Fu[1,3], Rinse Weersma[2], Gieneke Gonera[1,6], Iris H. Jonkers[1], Sebo Withoff[1,*]

* Corresponding author: s.withoff@umcg.nl

1. Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
2. Department of Gastroenterology and Hepatology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
3. Department of Pediatrics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
4. Department of Cardiology, Nanjing Medical University, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
5. Cardiovascular Research Center, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, China
6. Department of Pediatrics, Wilhelmina Hospital Assen, Assen, the Netherlands

Highlights:

- Shotgun metagenomic sequencing of the largest coeliac disease (CeD) patient cohort to date provides insights into gut microbiome composition down to the strain level.
- Novel associations between tCeD (treated CeD; adhering to a gluten-free diet) were identified.
- tCeD patients have a less uniform microbiome structure.
- Bacteria display higher mutation frequency in tCeD, compared to controls.
- The European subspecies of *Eubacterium rectale* is almost absent in tCeD patients.

Abstract:

Background: Coeliac disease (CeD) is a prevalent immune-mediated disorder primarily affecting the small intestine, characterised by a breakdown of tolerance to dietary gluten. Its onset results from the multifaceted interplay of genetic and environmental factors. Recent data show that alterations in gut microbiome composition could play an important role in CeD. However, many current studies have been constrained by small sample sizes and limited resolution.

Methods: To address these limitations, faecal gut microbiota from two Dutch patient cohorts, CeDNN (129 treated CeD patients; tCeD and 106 non-CeD controls) and the Lifelines Dutch Microbiome Project (24 self-reported treated CeD and 659 non-CeD controls), was analyzed using shotgun metagenomic sequencing. Controls were matched by age, sex, and sequencing depth. Self-reported IBS and IBD were used as comparative conditions of the gastrointestinal tract. Inter-individual variation among cases and controls was calculated for the whole microbiome and on the strain-level. Finally, species-specific gene repertoires were analysed in tCeD patients and controls.

Results: Alpha diversity was decreased in patients with self-reported IBS and IBD, but not in tCeD. Each condition displayed a unique microbial pattern, with minimal overlap in differentially abundant species. In addition to confirming existing studies, we identified an increase in the abundance of previously unreported species, including *Clostridium sp. CAG:253*, *Roseburia hominis*, and *Clostridium lavalense*. We demonstrate that changes in tCeD gut microbiome can be, at least partly, explained by adherence to a GFD and report an increased inter-individual variation in gut microbiome composition among tCeD patients. More detailed analyses revealed that a higher bacterial mutation frequency in CeD also leads to a higher inter-individual variation on the strain-level. In line with this, we found that the immotile European subspecies of *Eubacterium rectale*, with distinct carbohydrate metabolism potential, was nearly absent in tCeD patients.

Conclusion: Our study sheds light on the complex interplay between the gut microbiome and CeD, revealing unique microbial patterns and increased inter-individual variation among tCeD patients. These findings expand our understanding of the role of the microbiome in intestinal health and disease.

Keywords:

Coeliac disease, gut microbiome, metagenomic sequencing, gluten, gluten-free diet

Introduction:

Coeliac disease (CeD) is a multifactorial immune-mediated disorder characterised by an inflammatory response in the small intestine (1). CeD arises in genetically-susceptible individuals and is triggered by a loss of tolerance to dietary gluten—a storage protein found in grains such as wheat, barley, and rye (1–3). Approximately 1-2% of the global population is affected by CeD, with prevalence varying depending on the carriage of risk human leukocyte antigen (HLA) haplotypes and the extent of gluten consumption (4,5). In CeD patients, an inflammatory immune response is initiated by the translocation of partially digested gluten peptides from the intestinal lumen into the lamina propria. The peptides then undergo deamidation by tissue transglutaminase (tTG2), conferring an increased binding affinity to class II HLA-DQ2 or -DQ8 molecules on antigen-presenting cells which leads to the potent activation of gluten-specific CD4+ T cells upon peptide recognition (6–8). Subsequently, a complex cytokine-mediated cascade ensues, ultimately resulting in the impairment of the intestinal epithelial barrier by intraepithelial lymphocytes (9). Currently, strict life-long adherence to a gluten-free diet (GFD) is the only available treatment option, which requires considerable patient education and monitoring (10).

CeD is a complex disorder influenced by a combination of genetic and environmental factors. While the presence of the HLA-DQ2/DQ8 haplotypes constitutes the primary genetic risk factor, approximately 30% of the general population carries these haplotypes without developing CeD. Moreover, genome-wide association studies (GWAS) have identified an additional 39 non-HLA CeD risk loci, primarily associated with T cell activation and innate immunity (11–13). However, these non-HLA loci contribute to only 14% of the genetic variance in CeD (14). These findings indicate that while genetic susceptibility is necessary, it alone is insufficient for the development of CeD. Thus, other non-genetic factors likely play a role in the propagation of an inflammatory response to gluten and the subsequent development of CeD.

The gut microbiome has been proposed to be one of these factors. Multiple studies on the CeD microbiome have shown alterations in gut microbiome composition in CeD patients, such as reduced levels of beneficial bacteria including *Bifidobacterium* and *Lactobacillus* species, and increased levels of *Proteobacterium* species, compared to healthy controls (15–19). Several mechanisms have been proposed that implicate changes in microbiome composition to CeD pathophysiology including, but not limited to, a) specific microbial proteases could be involved in processing immunogenic gluten peptides (20–22), b) toxins or other microbe-derived metabolites secreted by pathobionts may affect intestinal barrier permeability thereby influencing the passage of gluten peptides into the lamina propria c) microbial proteins may contain sequences similar to immunogenic gluten peptides (mimicry), potentially initiating an inflammation immune response (23). Viral infections have also been implicated in a decrease in tolerance to gluten (24–26). Together, these findings underscore

the intricate interplay among the gut microbiota, gluten, and the immune system in CeD development.

Although these findings provide interesting insights, further analysis is necessary as many of the CeD microbiome studies carried out thus far have been constrained by small patient cohorts and/or relied upon 16S sequencing which limits the statistical power and taxonomic resolution to the genus level, respectively. Furthermore, most studies do not provide information on the functionality of the gut microbiota.

Our results, in addition to confirming previous studies (27,28), revealed novel associations between tCeD and the microbiome, explored the contribution of a GFD to gut microbiome composition and reported a less uniform microbiome composition, as well as increased strain-to-strain variation in patients with tCeD. Upon taxonomic analyses, we found both tCeD and self-reported IBD to present a unique microbiome signature, with limited overlap among changes in microbial abundance. Lastly, we reported shifts in subspecies distribution of *Eubacterium rectale* and *Firmicutes bacterium CAG:83* among tCeD patients, as compared to controls.

Materials and Methods

Study population

Treated CeD

The Celiac Disease Northern Netherlands cohort (CeDNN; unpublished as of yet) comprises 129 CeD patients and 106 controls that provided faecal samples at the project's onset and had complete metadata including age and sex. Most CeD patients in CeDNN had already been clinically diagnosed and were adhering to a GFD at the time of inclusion. Since nearly all controls were relatives or partners of the included patients, sharing genetic and/or lifestyle factors, we chose to enhance our study by merging this dataset with a subset of participants from the Lifelines Dutch Microbiome Project (LL-DMP) (29): 24 self-reported CeD-patients from LL-DMP were included, as well as 659 controls. Controls were selected in such a way, that the combined dataset contained 5 matched (age, sex, sequencing depth) controls for every CeD patient (Table 1). Matching was performed with the R program *MatchIt* (v4.5.5) (30). Self-reported IBD- and IBS-patients were excluded from the control group.

IBD and IBS

For comparative conditions, 93 self-reported IBD patients (Crohn's disease or ulcerative colitis) from LL-DMP were matched with 5 controls each (age, sex, sequencing depth; 465 controls total). 570 self-reported IBS patients from LL-DMP were matched with 3 controls each (age, sex, sequencing depth; 1710 controls total). Self-reported CeD, IBD- and IBS-patients were excluded from both control groups (Table 1).

152 Table 1: Description of the study population

Variable	Cohort	CeD		IBD		IBS	
		CeD	Controls	IBD	Controls	IBS	Controls
Sample size	CeDNN	129	106				
	LL-DMP	24	659	93	465	570	1710
Age (range)		30 (3-82)	30 (2-82)	53 (33-74)	54 (13-81)	49 (19-84)	49 (13-83)
Sex female (%)		98 (64%)	450 (59%)	56 (60%)	267 (57%)	411 (72%)	1233 (72%)
GFD (%)		93%	N/A				

153
154 Summary statistics of the data used in this study. The average age is indicated. Adherence to a gluten-free (GFD) diet was
155 largely unknown for controls.
156
157 **Ethical approval**
158 The LL-DMP study and the CeDNN study were approved by the medical ethical committee of
159 the University Medical Center Groningen (METc numbers 2017/152 and 2013/440,
160 respectively). Additional written consent was signed by all participants, their parents or legal
161 representatives (for participants aged under 18).
162
163 **Phenotypic data**
164 The CeDNN cohort uses the same questionnaires as Lifelines. Additionally, clinical information
165 is available for many of the participants, including date of diagnosis and adherence to a GFD.
166 The time between the date of diagnosis and the date of faecal sampling was used as a proxy
167 for the duration of GFD.
168
169 **Faecal sample collection, DNA extraction and sequencing**
170 Participants from both CeDNN and LL-DMP provided stool samples from home. Participants
171 were asked to freeze stool samples within 15 minutes of stool production. Frozen samples
172 were collected by personnel/students of the University Medical Center of Groningen, (CeDNN)
173 and Lifelines (LL-DMP), transported to their respective biorepositories on dry ice and stored
174 at −80 °C until DNA extraction. Microbial DNA was isolated with the QIAamp Fast DNA Stool
175 Mini Kit (Qiagen), according to the manufacturer's instructions, using the QIAcube (Qiagen)
176 automated sample preparation system. Library preparation for samples with total DNA yield
177 lower than 200 ng (as determined by Qubit 4 Fluorometer) was performed using NEBNext
178 Ultra DNA Library Prep Kit for Illumina, while libraries for other samples were prepared using
179 NEBNext Ultra II DNA Library Prep Kit for Illumina. Metagenomic sequencing was performed
180 at Novogene, China using the Illumina HiSeq 2000 platform to generate approximately 8 Gb
181 of 150 bp paired-end reads per sample.
182
183 **Metagenomic data processing**
184 *Trimming and filtering of reads*
185 Metagenomes were profiled as follows, consistent with previous analyses of 1000IBD (31),
186 Lifelines-DEEP (32) and the Lifelines-DMP cohort (29). Illumina adapters and low-quality reads

(Phred score < 30) were filtered out using *KneadData* (trimmomatic options: “LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50”; v0.10.0) (33). Following trimming, the KneadData-integrated *Bowtie2* tool (v2.5.0) (34) was used to remove reads that aligned to the human genome (GRCh37/hg19).

Taxonomy and pathway analysis

The taxonomic composition of metagenomes was profiled by *MetaPhlAn3* (v3.0.1) (35) using the MetaPhlAn database of marker genes *mpa_v30*. None of the included samples had a eukaryotic or viral abundance >25% of total microbiome content or a total read depth <10 million). MetaCyc pathways profiling was performed with *HUMAnN3* (default settings; v3.0.1) (35) integrated with the *DIAMOND* alignment tool (v2.0.15) (36), UniRef90 protein database (v0.1.1) (37) and *ChocoPhlAn* pan-genome database (*ChocoPhlAn_201901*) (35). Since single genes can be mapped to multiple pathways, pathway abundances were instead normalized to the sum of gene family abundances per sample.

To address the scarcity of microbial data during subsequent analysis, we focused on bacterial taxa/pathways that were present at a minimum abundance of 0.001% in at least 5% (taxa) or 2.5% (pathways) of participants. This yielded 301 taxa (6 phyla, 14 classes, 20 orders, 33 families, 70 genera and 158 species) and 338 pathways for tCeD vs. controls; 310 taxa (7 phyla, 15 classes, 20 orders, 31 families, 71 genera and 166 species) and 349 pathways for IBD vs. controls; and 316 taxa (9 phyla, 17 classes, 23 orders, 35 families, 73 genera and 159 species) and 331 pathways for IBS vs. Controls.

UpSet plot

To visualize the extent of overlap among observed differences in the three comparisons (tCeD, IBD and IBS vs. their respective matched controls), the data was plotted using the *UpSetR* package in R (38).

Alpha and beta diversity and inter-individual variation

Based on the abundance profiles of the taxa that passed the filtering process, we calculated alpha diversity, as measured by Shannon entropy, at the genus and species level using the *diversity* function in R package *vegan* (v.2.6.4) (39). Beta diversity was assessed through Bray-Curtis dissimilarity indices, as determined with the *vegdist* function from the R package *vegan*.

Strain- and gene-level analyses

For the comparison between individuals with tCeD and controls, we focused on studying differences at a sub-species level. Initially, we conducted strain-level profiling using StrainPhlAn3 (v3.0) (35), starting with default parameters for every species. For species for which default parameters were found to be overly stringent, and that were significantly associated with CeD according to the MetaPhlAn output, we adjusted the StrainPhlAn3 as follows: a) "--marker_in_n_samples 50", and b) "--marker_in_n_samples

228 20 --sample_with_n_markers 10". Among these adjustments, three species
229 (*Eisenbergiella tayi*, *Haemophilus parainfluenzae* and *Anaerotruncus colihominis*) were
230 successfully analyzed only with the most lenient parameter set. Only *Enterococcus faecium*
231 failed with all attempted settings. Distance matrices were constructed from the multiple
232 alignments, using the command line tool *distmat*. Subsequently, we used *PanPhlAn3* (35) to
233 analyze the gene repertoire of 9 species with at least a nominally significant increase in
234 interindividual strain distance among CeD patients. Heatmaps were created with the
235 *ComplexHeatmap* package in R (40).
236
237 Analyses were performed using locally installed tools and databases on the high-performance
238 computing infrastructure available at our institution and using the MOLGENIS data platform
239 (41).
240
241 **Statistical analyses**
242 *Associations between alpha diversity and sample group*
243 We conducted a comparison of Shannon diversity between disease groups (tCeD, IBD, IBS)
244 and controls using a linear model. Specifically, when comparing tCeD to controls, we included
245 sequencing batch (CeDNN_batch1, CeDNN_batch2, and LL-DMP) as a covariate.
246
247 *Associations of sample group with bacterial taxa and MetaCyc pathways*
248 We tested for differentially abundant bacterial taxa and MetaCyc pathways by modelling their
249 abundance using linear models with FDR correction. Differences in prevalence between
250 sample groups were tested using logistic regression, followed with a c^2-test using the *step*
251 function from the *stats* package in R (with FDR correction). For the comparisons of tCeD versus
252 controls, sequencing batch (CeDNN_batch1, CeDNN_batch2 and LL-DMP) was used as a
253 covariate. The correlation between taxonomic and pathway abundance was quantified using
254 Kendall's t coefficient.
255
256 *Associations between duration of GFD and bacterial abundance*
257 The duration of GFD was set to 0 for patients, not on a GFD. Using a linear model, effects of
258 age, sex, sequencing depth, and sequencing batch on bacterial abundance were regressed
259 out. Subsequently, Spearman's r tests were used to assess the effect of duration of GFD on
260 the residuals of bacterial abundance.
261
262 *Beta diversity*
263 To evaluate the relationship between microbiome composition and disease groups, we
264 conducted a PERMANOVA test using the adonis2 function from the vegan package. We utilized
265 Bray-Curtis distance metrics and performed 999 permutations for each test. When comparing
266 tCeD to controls, sequencing batch (CeDNN_batch1, CeDNN_batch2, and LL-DMP) was
267 incorporated as a covariate.
268

To measure the heterogeneity within each sample group (tCeD, IBD, IBS, and their respective control groups), we computed the distribution of all pairwise Bray-Curtis distances, excluding sample pairs that were related to each other. Significance was determined using permutation tests with 1000 permutations.

Strain-level analyses

To investigate whether increased inter-individual differences occur at a strain level, we measured strain-to-strain variation using the distance matrices generated by distmat. These matrices served as proxies for the mutational distance between strains of the same species in different samples within either the control group or tCeD group. Sample pairs that were related were excluded from this analysis. Significance was assessed using permutation tests with 1000 permutations.

We then conducted PAM-clustering with the pam function from the cluster package in R. The *prediction.strength* function from the *fpc* package (*clustermethod=claraCBI*) was employed to determine the optimal number of clusters that best described the data. Correlation between tCeD and strain-clusters was examined using Fisher's exact tests, with FDR-correction for multiple comparisons.

Gene prevalence and gene enrichment analysis

Associations between the prevalence of UniRef90 gene families, as determined by PanPhlAn, and tCeD were determined using logistic regression, followed with a c^2-test using the *step* function from the *stats* package in R (with FDR correction). Age, sex, sequencing depth and sequencing batch were used as covariates here. Genes that were present in less than 20% or more than 80% of all samples were excluded from this analysis. Protein family (Pfam) and Gene Ontology annotations were added to UniRef90 gene families and gene enrichment analysis was performed using Fisher's exact tests with FDR correction. Nominally significant gene families were used as input for the enrichment tests.

Correlation between RadC-like JAB domain proteins and inter-individual strain distance

To investigate whether the presence of proteins with the RadC-like JAB domain (PF04002), as identified by PanPhlAn, influences the mutational load of associated bacterial species (specifically, Bacteroides species *B. ovatus*, *B. vulgatus*, *B. uniformis*, *B. xylanisolvens)*, we conducted the following analysis:

For each sample analyzed by both StrainPhlAn and PanPhlAn, we recorded the median mutational distance (as determined by StrainPhlAn) to all other samples within the same sample group (either tCeD or control), as well as the number of proteins with a RadC-like JAB domain identified by PanPhlAn. We assessed trends within sample groups using a linear model. Additionally, we examined any negative trend irrespective of the sample group using

309 a linear mixed-effects model (implemented with the *lmer* function from the *lmerTest* package
310 in R), with the sample group as a random effect.
311

312 Results

313

314 Alpha diversity is decreased in self-reported IBS and IBD subjects, but not in CeD patients 315 on GFD

316 Although contradicting data can be found in literature, most often it is described that within-
317 sample microbiome diversity (alpha diversity) in untreated CeD is higher than that in controls
318 (27,42) and that adherence to a GFD restores alpha diversity to normal levels (27,28). Indeed,
319 using the Shannon index for evenness of the microbiome, we did not observe a significant
320 difference in alpha diversity on either genus or species level, when comparing tCeD (adhering
321 to a GFD) patients and matched controls (Table 2). Self-reported IBS and IBD patients display
322 a small but significant ($p < 0.05$) decrease in alpha diversity at the genus level when compared
323 with matched controls. This is in line with studies that observe a marked decrease in diversity,
324 especially in more inflamed IBD-like samples (43–45). Although tCeD patients show a
325 comparable alpha diversity to controls, several differences are found when studying
326 microbiome composition in more detail.
327

328 tCeD, IBD and IBS show distinct microbial patterns

329 Next, we analyzed the levels of faecal bacteria in individuals with tCeD, IBD, and IBS, along
330 with their matched control groups. Our findings revealed strikingly different trends among
331 these sample sets (Table 2, Figure 1). While all groups were characterized by an increase of
332 several *Bacteroides* species, even within that genus several differences can be observed: *B.*
333 *intestinalis*, *B. stercoris* and *B. vulgatus* are exclusively altered in IBD, while *B. xylanisolvens* is
334 only enriched in tCeD. Only *B. ovatus* is enriched in more than one subject group (tCeD and
335 IBD). All other species with changed abundance in tCeD are unaltered in IBD and IBS. Only
336 three taxa show significant differential abundance in self-reported IBS-patients, compared to
337 controls (*Bacteroides* and *Veillonella* genera and *P. merdae*), which are in line with earlier
338 studies as reviewed by Pittayanon et al. (46). More overlap between the diseases can be seen
339 when looking at prevalence (Table 3), but several bacteria were identified that are affected
340 exclusively in tCeD.

341

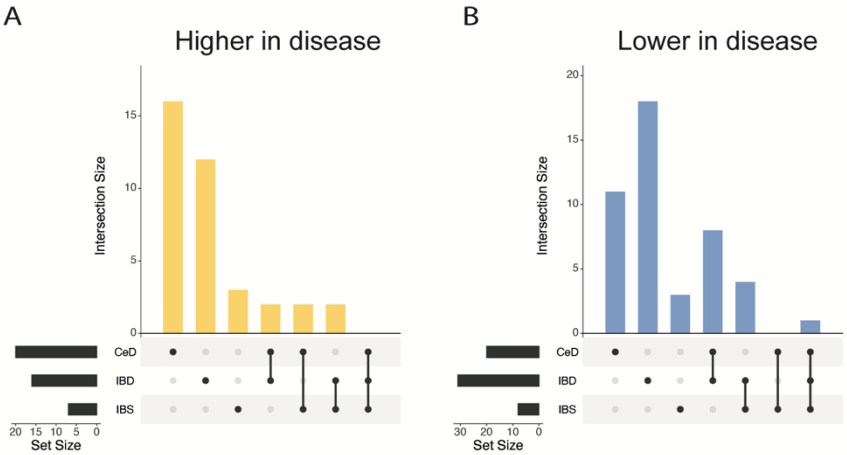


Figure 1: Overlap in microbiome changes between tCeD, IBD and IBS. Bacterial species with A. higher abundance or prevalence in disease, and B. lower abundance or prevalence in disease.

Taxonomic analyses reveal known and novel associations with tCeD

Multiple of the observed differences in bacterial abundance and/or prevalence between tCeD and controls (Table 2 and Table 3) are in accordance with earlier reports, such as the enrichment of *H. parainfluenzae* (28), *R. inulinivorans* (28), *E. coli* (18,47), and *Bacteroides* species (18,27,28), as well as the decreased prevalence of *R. bicirculans* , *C. eutactus* (27), *C. fastidiosus* (28,48), *Roseburia sp. CAG:309* (28), and several *Bifidobacterium* species (18,28,47). In addition, however, several novel species were identified in our analysis, including an increase in *Clostridium sp. CAG:253*, *R. hominis*, and *C. lavalense*.

Table 2: Taxa with significantly altered abundance in tCeD, IBD and IBS

Name	Previously reported	tCeD	IBD	IBS
Shannon index - Genus (control/case)		1.86 / 1.84	1.87 / 1.70 *	1.87 / 1.83 *
Shannon index - Species (control/case)		2.44 / 2.48	2.44 / 2.35	2.45 / 2.44
Bacterial taxa				
Acidaminococcus intestini			3.2	
Anaerotruncus colihominis		2.5[b]		
Bacteroides (genus)			1.0	0.3
Bacteroides fragilis	(27)	1.3		
Bacteroides intestinalis			2.4	
Bacteroides ovatus	(49)	1.5	1.2	
Bacteroides stercoris			1.7	
Bacteroides thetaiotaomicron	(50)	0.9		
Bacteroides uniformis			0.7	
Bacteroides vulgatus			1.4	
Bacteroides xylanisolvens	(51)	1.5		
Bifidobacterium dentium			2.3	
Bifidobacterium pseudocatenulatum			2.5	
Clostridium bolteae			5.2	
Clostridium clostridioforme			5.3	
Clostridium sp. CAG:253		2.3		
Eisenbergiella tayi	(28)	3.2		
Enterococcus faecium		5.1[a,b]		

Firmicutes bacterium CAG:145			2.9	
Firmicutes bacterium CAG:94			1.2	
Flavonifractor plautii			1.1	
Fusobacterium (genus)			3.4	
Haemophilus parainfluenzae	(28)	1.2		
Parabacteroides distasonis			0.8	
Parabacteroides merdae				0.4
Paraprevotella xylaniphila	(28)	1.0		
Roseburia hominis		1.7		
Roseburia inulinivorans	(28)	1.4		
Roseburia sp. CAG:182		2.7[a,b]		
Ruthenibacterium lactatiformans			1.1	
Veillonella (genus)				3.2

Yellow and blue cells indicate an increase and decrease, respectively, in the Shannon index or bacterial abundance in the respective disease. An asterisk (*) indicates a significant difference in the Shannon index. Numbers in the bacterial taxa section represent ^{2}log(fold change) that are significant after multiple testing correction. Coloured cells without text indicate nominally significant differences. Previously reports of associations with CeD are included in a separate column. [a]Significant batch effect within the CeDNN cohort. [b]Significant difference between observed effect in the CeDNN and Lifelines cohorts.

Table 3: Taxa with significantly altered prevalence in tCeD, IBD and IBS

Name	Previously reported	tCeD	IBD	IBS
Acidaminococcus intestini			4 > 12%	
Akkermansia muciniphila			70 > 49%	
Alistipes finegoldii	(51)	99 > 95%		
Alistipes indistinctus			69 > 51%	
Alistipes shahii			89 > 78%	
Anaerotruncus colihominis		20 > 46%		13 > 18%
Bacteroides dorei	(50)	72 > 84%		
Bacteroides galacturonicus		8 > 5%		
Bacteroides sp. CAG:144		28 > 16%	43 > 26%	
Barnesiella intestinihominis			87 > 73%	
Bifidobacterium adolescentis	(16)	85 > 64%		
Bifidobacterium bifidum		37 > 23%		
Bifidobacterium catenulatum		29 > 6%[a]		
Bifidobacterium dentium				5 > 9%
Bifidobacterium longum	(52)	91 > 76%		
Butyrivibrio crossotus				16 > 10%
Clostridium bolteae			3 > 14%	
Clostridium clostridioforme			2 > 10%	
Clostridium lavalense		6 > 16%		
Clostridium leptum	(53)	23 > 54%[a]		
Clostridium sp. CAG:253		7 > 19%[a]		
Clostridium sp. CAG:58				28 > 36%
Coprobacter fastidiosus	(48)	46 > 29%		
Coprococcus eutactus	(54)	49 > 25%	47 > 25%	
Desulfovibrio piger			43 > 22%	
Desulfovibrionaceae (family)		88 > 76%	89 > 76%	
Dialister invisus			31 > 45%	
Eggerthella lenta		25 > 51%	17 > 37%	
Eisenbergiella massiliensis		3 > 16%		
Eisenbergiella tayi	(28)	4 > 32%[a]		
Enterococcus faecium		1 > 14%[a]		

Escherichia coli	(17)	44 > 72%		
Eubacterium rectale			92 > 83%	
Eubacterium siraeum	(48)	70 > 37%	71 > 54%	
Eubacterium sp. CAG:180				30 > 23%
Eubacterium sp. CAG:251			25 > 13%	
Faecalibacterium prausnitzii			100 > 96%	
Firmicutes bacterium CAG:110			71 > 41%	65 > 55%
Firmicutes bacterium CAG:170		34 > 16%	49 > 19%	
Firmicutes bacterium CAG:238		11 > 5%	22 > 6%	
Firmicutes bacterium CAG:83			87 > 66%	
Firmicutes bacterium CAG:95		27 > 9%	36 > 15%	
Flavonifractor plautii			67 > 85%	
Fusicatenibacter saccharivorans			98 > 85%	
Gemmiger formicilis			93 > 67%	93 > 88%
Gordonibacter pamelaeae				31 > 40%
Haemophilus parainfluenzae	(28)	21 > 45%		
Harryflintia acetispora		6 > 13%		8 > 13%
Hungatella hathewayi		5 > 11%		
Odoribacter splanchnicus		96 > 89%	96 > 87%	
Oscillibacter sp. 57_20		92 > 85%		
Oscillibacter sp. CAG:241			84 > 61%	
Parabacteroides distasonis				92 > 85%
Parabacteroides merdae			77 > 62%	
Paraprevotella clara			28 > 14%	
Paraprevotella xylaniphila			52 > 33%	48 > 38%
Phascolarctobacterium succinatutens			30 > 12%	
Prevotella sp. CAG:279				17 > 10%
Roseburia sp. CAG:182		24 > 43%		
Roseburia sp. CAG:309	(28)	14 > 2%		
Ruminococcaceae (family)	(55)	80 > 91%	71 > 86%	76 > 82%
Ruminococcus bicirculans	(28)	64 > 20%[a]		
Ruminococcus gnavus			6 > 17%	8 > 12%
Ruminococcus torques			51 > 37%	
Selenomonadaceae (family)		8 > 3%	12 > 3%	12 > 6%
Selenomonadales (order)			17 > 5%	
Veillonella atypica		4 > 10%		
Veillonella parvula				5 > 9%
Veillonellaceae (family)	(56)	13 > 24%	8 > 18%	9 > 14%
Victivallis vadensis		40 > 16%	49 > 23%	

Yellow and blue cells indicate an increase and decrease, respectively, in prevalence in the respective disease. Numbers represent shifts in prevalence (control > case) that are significant after multiple testing correction. Coloured cells without text indicate nominally significant differences. [a]Significant interaction with batch and/or cohort observed.

Duration of GFD explains part of the observed differences

Although we are primarily reporting on the microbiome composition of tCeD patients, we attempted to assess the contribution of GFD to the increased abundance of species shown in Table 2. After correction for covariates, we found that the duration of adherence to a GFD positively correlated with the increased abundance of *Paraprevotella xylaniphila* and *Roseburia sp. CAG:182* (Spearman rank correlation, $p_{adj} < 0.05$; Table 4). Moreover, given that the majority of the 13 tested species exhibit a positive correlation, these findings imply that GFD indeed plays a role in several species-level associations observed in this study. Notably,

we were unable to replicate the adverse impact of the GFD on the abundance of *H. parainfluenzae* as seen in other research (28,57).

Table 4: Spearman correlation between duration of gluten-free diet and species abundance

Name	Spearman's ρ
Bacteroides fragilis	0.15
Bacteroides ovatus	0.20
Bacteroides thetaiotaomicron	0.13
Bacteroides xylanisolvens	0.05
Paraprevotella xylaniphila	0.28*
Enterococcus faecium	0.04
Clostridium sp. CAG:253	0.12
Eisenbergiella tayi	0.23
Roseburia hominis	-0.11
Roseburia inulinivorans	0.20
Roseburia sp. CAG:182	0.36*
Anaerotruncus colihominis	-0.03
Haemophilus parainfluenzae	0.16

Spearman correlation coefficients between GFD duration and species abundance, for all species significantly enriched in tCeD patients. *$p_{adj} < 0.05$.

Differences in metabolic pathways are largely driven by specific bacterial taxa

After establishing the taxonomic landscape in each of the subgroups analyzed in this study, we analyzed the functional potential of the associated microbiota as can be inferred by mapping of identified gene families to metabolic pathways in the MetaCyc database. Although several differences were found between tCeD patients and controls, very few of these differences were unique to tCeD, as there was a striking concordance with differences observed in IBD and IBS (Table 5). These similarities can probably be explained by the fact that most differentially abundant metabolic pathways are strongly associated to differentially abundant taxa, with *Bacteroides* species and *E. coli* explaining most of the observed changes (Table 5 and Supp. Figure 1). Furthermore, even pathways with lower abundance in tCeD than controls (Table 5, blue) appear to be primarily driven by an increase in bacteria from the *Bacteroidetes* phylum that replaces bacteria carrying this metabolic potential, reflecting the compositional nature of these data. Since pathway abundance seems to mainly reflect taxonomic changes, rather than functional enrichment across several taxa, the identified pathways are, in our assessment, unlikely to be relevant in the context of the disease studies.

Table 5: Metabolic pathways differentially abundant in tCeD

Pathway	CeD	IBD	IBS	Top correlated taxa ($\tau > 0.3$)	τ
PENTOSE-P-PWY: pentose phosphate pathway	0.3	0.3	0.1	*B. uniformis*	0.47
POLYISOPRENSYN-PWY: polyisoprenoid biosynthesis	0.6	0.7	0.2	*B. uniformis* *B. vulgatus*	0.59
PWY-2941: L-lysine biosynthesis II	0.4[a]	0.3	0.1	*Bacteroidaceae* (family)	0.32

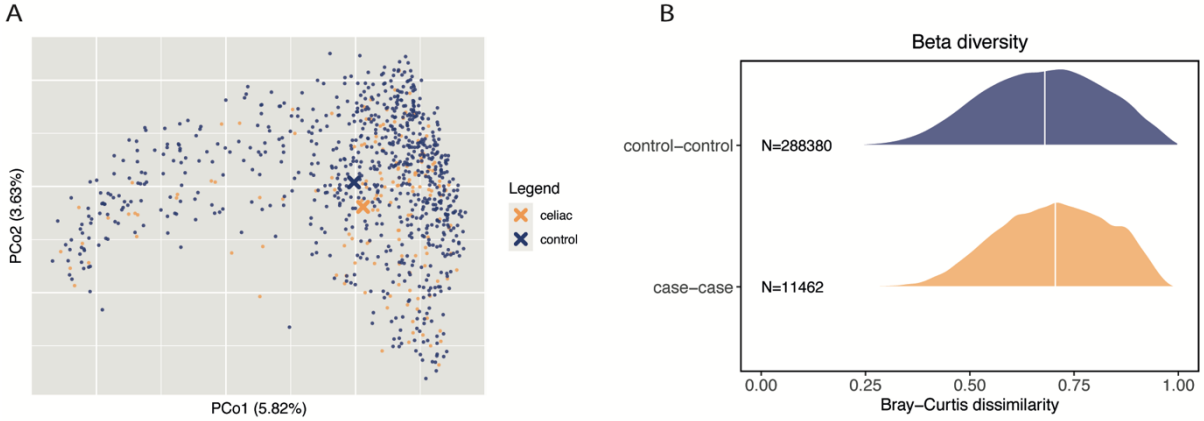
PWY-5083: NAD/NADH (de)phosphorylation	1.4	1.1		*B. ovatus*	0.41
PWY-5838: superpathway of menaquinol-8 biosynthesis I	1.3[a,b]	1.1		*E. coli*	0.54
PWY-5941: glycogen degradation II	0.9[a]	0.8	0.3	*B. ovatus* *B. thetaiotaomicron* *B. uniformis*	0.46
PWY-5971: palmitate biosynthesis II	1.3[b]	1.1		*E. coli*	0.58
PWY-6353: purine nucleotide degradation II (aerobic)	0.7[a,b]	0.5			
PWY-6531: mannitol cycle	1.1	0.8		*B. ovatus*	0.40
PWY-6572: chondroitin sulfate degradation I	0.8[a]	0.7	0.3	*B. thetaiotaomicron*	0.58
PWY-6608: guanosine nucleotide degradation III	0.6	0.4			
PWY-6859: all-trans-farnesol biosynthesis	0.6[a]	0.7	0.2	*B. vulgatus, B. uniformis*	0.55
PWY-6953: dTDP-3-acetamido-3,6-dideoxy-α-D-galactose biosynthesis	3.0			*B. cellulosilyticus*	0.35
PWY-7316: dTDP-N-acetylviosamine biosynthesis	1.7			*B. fragilis*	0.42
RHAMCAT-PWY: L-rhamnose degradation I	0.4	0.3		*B. uniformis*	0.36
ARGSYNBSUB-PWY: L-arginine biosynthesis II (acetyl cycle)	-0.1[a,b]	-0.2	-0.1	*Bacteroidetes* (phylum) *F. prausnitzii*	-0.44 0.32
PWY-7219: adenosine ribonucleotides de novo biosynthesis	-0.1[a]	-0.1	-0.03	*Bacteroidaceae* (family)	-0.39
PWY-7383: anaerobic energy metabolism	-0.3	-0.4		*Bacteroidetes* (phylum) *A. muciniphila*	-0.40 0.31
PWY66-399: gluconeogenesis III	-0.2	-0.3		*Bacteroidetes* (phylum)	-0.43
TRNA-CHARGING-PWY: tRNA charging	-0.1	-0.1		*Bacteroidetes* (phylum)	-0.36

Yellow and blue cells indicate an increase and decrease, respectively, in abundance in the respective disease. Numbers represent ^{2}log(fold change) that are significant after multiple testing correction. Coloured cells without text indicate nominally significant differences. Correlated taxa are reported with associated Kendall's τ coefficient. [a]Significant batch effect within the CeDNN cohort. [b]Significant difference between observed effect in the CeDNN and Lifelines cohorts.

tCeD faecal microbiome is less uniform than that of controls

Since we identified several bacterial species specifically enriched in tCeD, we suspected that tCeD patients' microbiomes would display a certain (t)CeD signature and be more uniform than those of non-CeD controls. To test this hypothesis, we first calculated Bray-Curtis dissimilarity between samples, a measure of beta diversity, and performed principal coordinate analyses (Figure 2A). Not surprisingly, the average overall composition of the tCeD microbiome differed significantly from that of controls (PERMANOVA; R^2=0.003; $p < 0.05$). However, when we tested which group was the most homogeneous, i.e. which group displayed the lowest median within-group dissimilarity (Figure 2B), we found that the microbiome composition of two arbitrary, unrelated tCeD patients tended to be more dissimilar than that of two unrelated controls (permutation test, $p < 0.05$). Interestingly, the same phenomenon, where patients have a less uniform microbiome structure than controls,

421 was observed among self-reported IBD patients (Supp. Figure 2A-B; $p < 0.05$). In contrast, the
422 distribution of Bray-Curtis dissimilarity scores among controls and self-reported IBS patients
423 were indistinguishable (Supp. Figure 2C-D).
424



425
426 Figure 2: Beta diversity in tCeD vs. controls, using Bray-Curtis dissimilarity as a distance metric. A. Principal
427 coordinate plot of tCeD samples and matched controls. B. Within-group dissimilarity histograms. Top: pairs of
428 unrelated controls. Bottom: pairs of unrelated tCeD patients. Indicated are the number of different permutations
429 of two samples that fall into each group.

430 **tCeD dissimilarity is also present on sub-species level**

431 Next, we wanted to assess whether the increased diversity among tCeD patients was merely
432 a global phenomenon, acting on the level of the full microbiome, or whether this effect was
433 present on multiple levels. Using StrainPhlAn to estimate the pairwise evolutionary distance
434 between bacterial strains of the same species from two different participants, we were able
435 to analyze strain-level diversity for a total of 67 bacterial species [Supp. Table S9]. For 9 species
436 (including 4 *Bacteroides* species; Figure 4), the inter-individual evolutionary strain distance
437 among tCeD patients tended (nominal $p < 0.05$) to be higher than among controls, while only
438 1 showed the tendency toward more diversity among controls. After multiple testing
439 correction, *Eubacterium rectale* (Figure 3B) and *Bacteroides* species *B. ovatus, B. uniformis,*
440 and *B. xylanisolvens* remained significant and were all more diverse among tCeD patients.
441

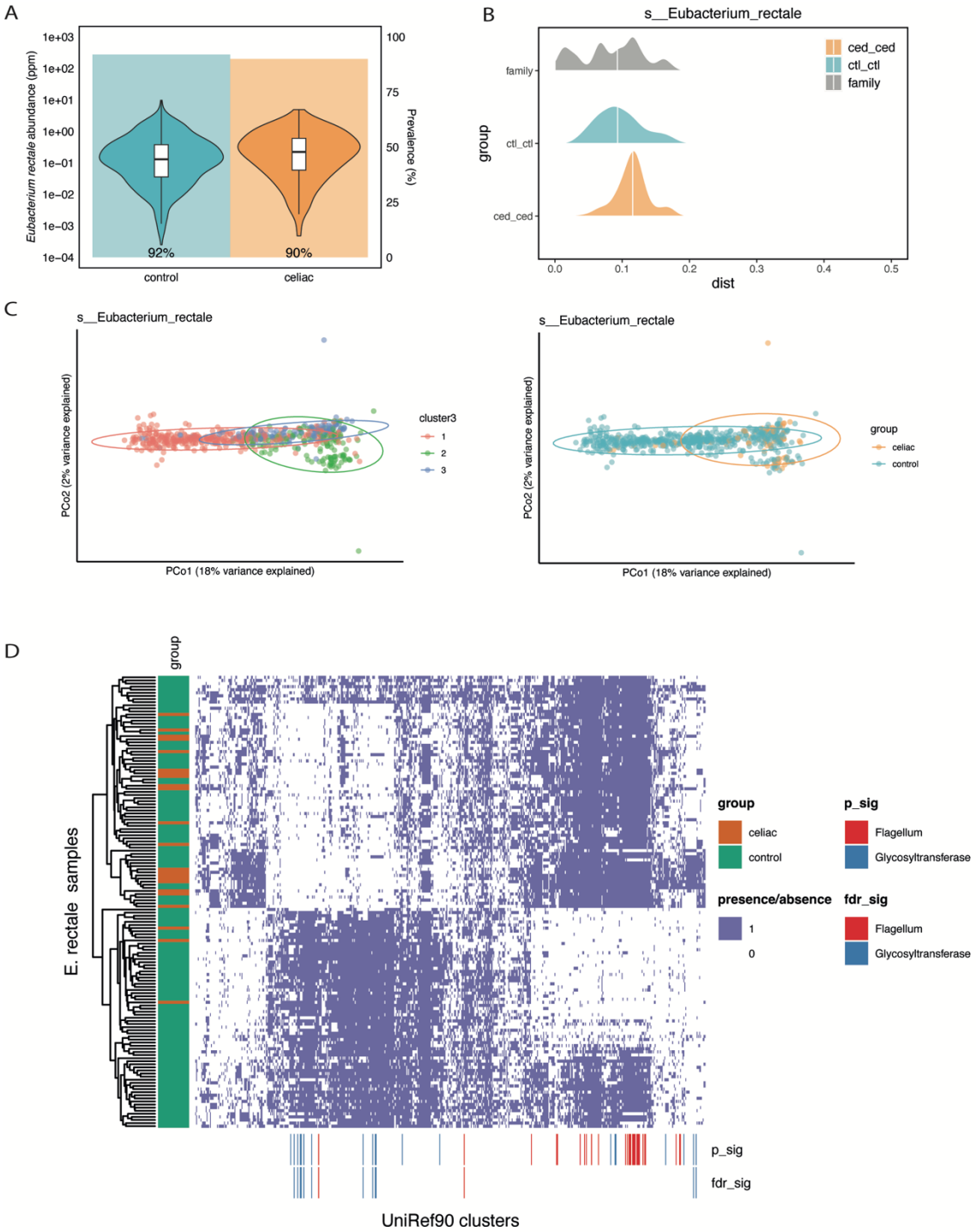
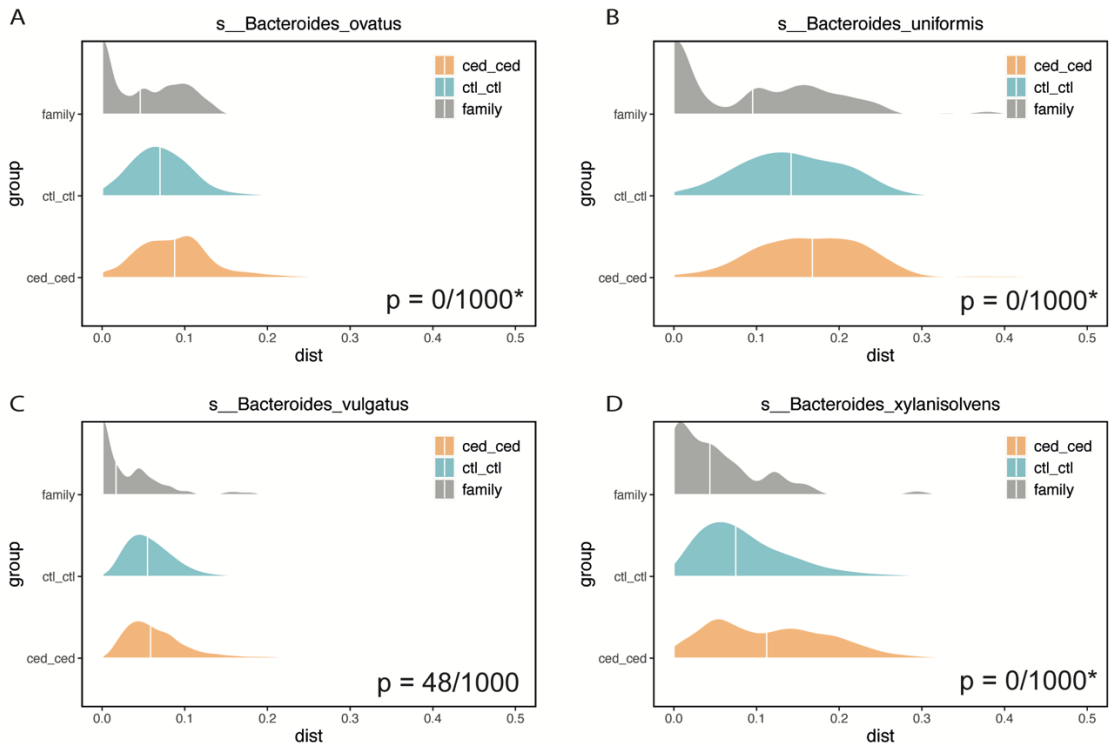


Figure 3: Strain-level differences of *Eubacterium rectale* between tCeD patients and controls. A. Abundance and prevalence of *E. rectale* in patients and controls. Violin plots represent the distribution of abundance (log-scale) in the two groups. Samples with absent *E. rectale* are excluded from this plot. Shaded bars on the background represent prevalence among all samples. B. Distribution of pairwise genetic distances between *E. rectale* strains. Control vs. control (top) and tCeD vs. tCeD (bottom). Pairs of relatives were excluded from both comparisons. C. Non-metric multidimensional scaling plots of pairwise genetic distances between *E. rectale* strains, coloured by cluster-ID (PAM clustering; left) and disease status (right). Ellipses include 95% of the samples in the indicated group. D. Heatmap of presence/absence of *E. rectale* genes in all analyzed *E. rectale* strains. Each row represents the *E. rectale* strain from one participant. A phylogenetic tree based on pairwise genetic distances is shown on the left, as well as the participant's disease status. Each column represents a UniRef90 cluster (gene) that is differentially prevalent in tCeD patients (nominal $p < 0.05$). Coloured bars in the bottom indicate whether the genes are predicted to encode glycosyltransferases or flagellum-related proteins.

455 **European subspecies of *Eubacterium rectale* is reduced among tCeD patients**
456 To further investigate strain-to-strain variation, we performed principal coordinate analyses,
457 combined with PAM (partition around medoids) clustering. This approach revealed distinct
458 sub-species clustering for 46 out of the 67 species studied. For *E. rectale* (Figure 3C) and
459 *Firmicutes bacterium CAG:83* (Supp. Figure 3; likely part of the *Oscillibacter* genus (58)), the
460 dominant clusters among controls (61% and 69%, respectively) were significantly reduced
461 among tCeD patients to 33% and 44% (Fisher's exact test, $p_{adj} < 0.05$).
462
463 Finally, we performed gene-level comparison of cases and controls (PanPhlAn), followed by
464 enrichment analysis of protein family (Pfam) and gene ontology (GO) terms among nominally
465 significant ($p < 0.05$) genes. With this analysis, we aimed to identify gene-level explanations
466 of the strain-level observations (increased inter-individual differences and depletion of the
467 dominant *E. rectale* cluster). The only class enriched among *E. rectale* genes with lower
468 prevalence in tCeD was PF00535 (glycosyl transferase family 2), while over-prevalent genes
469 were enriched for terms related to flagellum assembly and organization, chemotaxis, protein
470 targeting and secretion, and phosphoenolpyruvate-dependent sugar phosphotransferase
471 systems (Figure 3D). Karcher et al. described an *E. rectale* subspecies almost exclusively found
472 in Europe (ErEurope) that is characterized specifically by the presence of several genes from
473 glycosyl transferase family 2 and the absence of motility operons, including chemotaxis and
474 flagellum genes (59). We conclude, therefore, that specifically ErEurope is less prevalent in
475 tCeD patients.
476



477
478 Figure 4: Distribution of pairwise genetic distances between strains of A. *Bacteroides ovatus, Bacteroides*
479 *uniformis*, *Bacteroides vulgatus*, *Bacteroides xylanisolvens*. Control vs. control (top) and tCeD vs. tCeD (bottom).
480 Pairs of relatives were excluded from both comparisons. The significance of increased distance among tCeD
481 patients is indicated by p-values from permutation test (1000 iterations). Asterisks (*) indicate $p_{adj} < 0.05$.

482 **Gene families related to transmembrane transport and DNA repair are less prevalent in**
483 **multiple species in tCeD patients**
484 Besides *E. rectale*, we performed gene-level analyses for 6 other bacterial species. When
485 analyzed separately, no enriched terms could be identified for any of these species. To assess
486 whether there were any bacterial gene or protein classes enriched across species, we also
487 performed the analysis after pooling gene-level data of all 7 bacteria. Among under-prevalent
488 genes in tCeD, we identified two new enriched classes. Firstly, 23 of 150 genes from
489 GO:0055085 (transmembrane transport) were under-prevalent in tCeD. Many of the
490 contributing genes, from 6 different species, were annotated to encode permeases, with some
491 of them having a predicted specificity for carbohydrates. We suggest that differences in
492 nutrient availability during active CeD or a gluten-free diet may play a role in the loss of these
493 genes encoding carbohydrate uptake proteins.
494
495 Secondly, 9 of 21 genes encoding proteins from PF04002 (RadC-like JAB domain) were under-
496 prevalent, across all tested *Bacteroides* species. Proteins from this family have been
497 speculated to have proteolytic and/or nuclease activity and to be involved in DNA repair (60).
498
499 **Presence of DNA repair genes in *Bacteroides* species results in lower mutational drift**
500 Next, sought to assess whether the reduced prevalence in tCeD of DNA repair genes
501 containing the RadC-like JAB domain could be associated with the increased inter-individual
502 variation of strains of *Bacteroides* species. Indeed, we found that the number of genes in a
503 specific strain encoding a RadC-like JAB domain negatively correlated (Figure 5) with the
504 median distance of that strain to strains of the same species within the same sample group
505 (tCeD and control). This negative correlation was observed in all species tested (*B. ovatus*, *B.*
506 *uniformis*, *B. vulgatus* and *B. xylanisolvens*) and was significant in the two species with most
507 data points, *B. uniformis* and *B. vulgatus* (linear mixed-effects model, $p < 0.05$). Moreover, for
508 all four species, there was a negative correlation both in the tCeD group and in controls, in
509 line with our hypothesis that these DNA repair proteins are limiting the mutation rate in these
510 *Bacteroides* species.
511

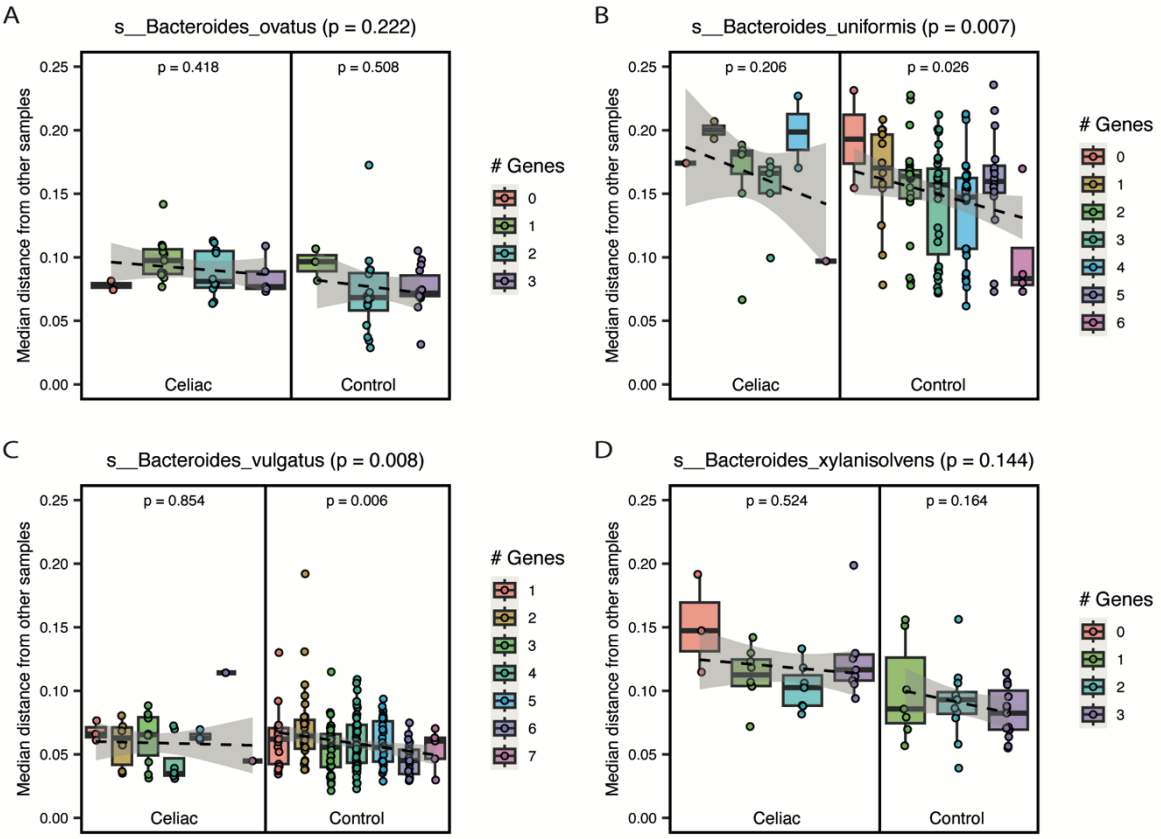


Figure 5: Correlation between prevalence, in a specific sample, of genes encoding the RadC-like JAB domain ('# Genes') and the median mutational distance (StrainPhlAn) from other strains within the same sample group, for each of A. *Bacteroides ovatus*, B. *Bacteroides uniformis*, C. *Bacteroides vulgatus*, and D. *Bacteroides xylanisolvens*. Significance within sample groups is indicated with p-values inside the plot, while p-values in the plot title across both groups report on the association in both groups combined (linear mixed-effects model).

Individual gene families from *Eubacterium rectale, Alistipes putredinis* and *Bifidobacterium adolescentis* are differentially prevalent in tCeD.

Besides gene families that were part of above-mentioned functional classes, 413 more gene families were differentially prevalent in tCeD patients, compared to controls ($p_{adj} < 0.05$; Supp. Table S10). These gene families are derived from *Eubacterium rectale* (38 up, 101 down), *Alistipes putredinis* (168 up, 96 down) and *Bifidobacterium adolescentis* (4 up, 6 down).

Interestingly, *B. adolescentis* strains found in tCeD less frequently encode a LacI/GalR-family transcriptional regulator and galactose-1-phosphate uridyltransferase GalT, part of the Leloir pathway for galactose metabolism. This finding constitutes another potential implication of altered carbohydrate availability during inflammation and/or GFD (61).

Discussion

In this paper, we performed a comprehensive analysis of the largest cohort of CeD microbiota to date, comparing the faecal microbiota of 153 tCeD patients with 765 controls, matched by age, sex, and sequencing depth. Our approach utilises shotgun metagenomic sequencing

535 allowing for microbiome classification at the strain level and providing insights into bacterial
536 genes and pathways.
537
538 *Taxonomic and functional profiling*
539 We aimed to compare taxonomic profiles at the species level with those of previous CeD
540 microbiome studies. In summary, our data (Table 2 and Table 3) confirms previous findings in
541 (treated) CeD patients, demonstrating an increase in several *Bacteroides* species, *H.*
542 *parainfluenzae, R. inulinivorans*, and *E. coli*, along with a decrease of various bifidobacteria,
543 *R. bicirculans*, *C. eutactus,* and *C. fastidiosus* (18,27,28,47). Despite a lower prevalence of the
544 major butyrigenic bacterium *Butyrivibrio crossotus* (62) (Table 3), we observed novel
545 associations with increased abundances of other butyrate-producing bacteria including
546 *Clostridium sp. CAG:253, R. hominis*, and *C. lavalense* in tCeD patients in our tCeD cohort.
547 Interestingly, these bacteria, as well as some of the previously reported bacteria (e.g. *R.*
548 *inulinivorans*) all produce butyrate (62,63). Higher production of butyrate by the CeD
549 microbiome would be in line with observations by Baldi et al. that CeD patients display slightly
550 higher butyrate fractions (64). However, Tjellström et al. showed that one year of adherence
551 to a GFD restored butyrate levels in pediatric CeD patients to normal levels (65). We speculate
552 that the negative effect of a GFD on butyrate production does not necessarily reflect the
553 restoration of the microbiome to a pre-CeD state, but may rather be mediated by other
554 butyrigenic species, such that effects of disease and diet largely cancel each other out.

555
556 Further taxonomic analyses revealed that tCeD, IBD and IBS were all characterised by an
557 increase in the *Bacteroides* genus, which drives remarkably similar changes in metabolic
558 pathways observed in each disease group. Moreover, it seems that even all (moderately)
559 under-abundant pathways are explained by the (moderate) increase of bacteria from the
560 *Bacteroidetes* phylum, therefore likely an artefact caused by the compositional nature of the
561 data (Table 5, Supp. Figure 1).
562
563 Besides the genus *Bacteroides*, however, the different disease groups studied each show
564 highly distinct microbiome profiles (Table 2, Table 3, Figure 1). Although we showed that the
565 GFD contributes to an altered faecal microbiome (Table 4), the current data does not allow us
566 to separate the consequences of CeD and/or GFD from characteristics of the microbiome that
567 contribute to the onset of CeD.

568
569 Although most metabolic pathways are seemingly altered because the bacteria encoding
570 them happen to be altered in tCeD, we cannot dismiss the possibility that certain bacteria are
571 enriched *because* they encode those specific metabolic pathways. Indeed, other studies
572 report several *Bacteroides* species to be enriched in active (18,42,66) and treated CeD patients
573 (18,27,42), as well as in at-risk individuals (51,67,68), and in these studies, *B. fragilis* was
574 speculated to be involved in pathogenesis (49). Although none of the previous studies report

575 to have performed multiple testing correction, compromising their true discovery rate, taken
576 together they do suggest that the overabundance of *Bacteroides* species is not solely caused
577 by GFD and that some of the pathways we have identified might therefore be relevant in CeD
578 pathogenesis. Additionally, the 1.3-fold enrichment of genes involved in the rhamnose
579 degradation pathway (Table 5) could instead be related to differences in carbohydrate supply
580 from the diet. Furthermore, two pathways were uniquely affected in tCeD, also representing
581 the two largest differences with controls: the dTDP-N-acetylviosamine biosynthesis and dTDP-
582 3-acetamido-3,6-dideoxy-a-D-galactose biosynthesis pathways are 3.3-fold and 7.8-fold
583 enriched in tCeD. Both these pathways use α-D-glucopyranose 1-phosphate, a product of
584 glycogen degradation, as a substrate. Since both these pathways and, as in IBD, that of
585 glycogen degradation are overabundant in tCeD and associated with various *Bacteroides*
586 species, it could be that the ability to survive in harsh conditions, with low carbohydrate
587 availability, allows these bacteria to outcompete other bacteria (69). Only two pathways
588 enriched in tCeD do not appear to be driven by specific bacterial taxa, PWY-6353 and PWY-
589 6608 (Table 5), both of which are related to purine nucleotide degradation.
590
591 *Inter-individual variation*
592 Besides species-specific differences between tCeD patients and controls, we found an
593 increased inter-individual variation among patients as compared to the diversity among
594 controls (Figure 2). While we expected to find that patients might share a certain microbiome
595 signature, either related to the disease or GFD, the inverse appears to be true: their
596 microbiome, more than that of controls, deviates from the average non-CeD microbiome (not
597 shown), but these changes apparently occur in different directions between patients.
598 Although we want to refrain from using the term *dysbiosis* (70), this observation could reflect
599 the loss of certain healthy components of the microbiome. On the other hand, it could be that
600 a gluten-*containing* diet, rather than disease or GFD, induces a microbiome signature. While
601 the possibility of a loss of specific 'healthy' bacteria characterizing CeD seems incompatible
602 with the fact that we exclusively observed over-abundant bacteria in tCeD, presence-absence
603 analyses did reveal some bacteria with reduced prevalence in tCeD, such as several species of
604 bifidobacteria, generally regarded as 'healthy' bacteria (71), along with *R. bicirculans*, *E.*
605 *siraeum* and *C. eutactus* (Table 3).
606
607 Interestingly, the increased diversity among patients is also visible in strain-level analyses,
608 where, for several bacteria, the average evolutionary distance between strains of the same
609 species in two unrelated tCeD patients is larger than that between strains in two control
610 individuals (Supp. Table S9). This phenomenon was observed most clearly ($p_{adj} < 0.05$) in
611 *Eubacterium rectale* (Figure 3B) and *Bacteroides ovatus, Bacteroides uniformis*, and
612 *Bacteroides xylanisolvens* (Figure 4), but appears to be more general, as 48 out of 67 analysed
613 species (72%) show the same trend. The most plausible explanation, in our opinion, is that
614 inflammation and oxidative stress accompanying CeD and gluten intake (72) lead to a higher

615 mutational burden in the gut microbiome. In line with this hypothesis, transcriptome analysis
616 of intestinal epithelial cells recently revealed the upregulation of pathways related to DNA
617 double-strand break repair in active CeD, as compared to controls (73). Moreover, we find that
618 *Bacteroides* strains (multiple species) that lack the RadC-like JAB domain, involved in DNA
619 repair (60), are less prevalent in tCeD than in controls (Supp. Table S10 and Supp. Table S11),
620 and that the presence of these repair proteins in *Bacteroides* strains limits their mutation rate
621 and thereby inter-strain variation (Figure 5). These observations together suggest it may be
622 beneficial to these bacteria to increase the genetic diversity of their offspring under stressful
623 conditions. It is, in fact, an established strategy among bacteria to upregulate their evolution
624 rate in response to stress (74). Examples of this strategy are the widespread SOS-response,
625 utilizing an error-prone DNA polymerase (75), and the activation of competence for
626 transformation during DNA replication stress (76,77). It is unclear whether the observed
627 increase in mutational load in the tCeD microbiota is an indication of residual inflammation or
628 merely a result of periods of inflammation in the past.
629
630 Intriguingly, while *Eubacterium rectale*, too, shows clear evidence of what seems to be random
631 mutagenesis (Figure 3B, Supp. Figure 4), we also observed that the ErEurope subspecies (59)
632 was almost absent in tCeD, while nearly half of the controls carried this subspecies.
633 Importantly, ErEurope lacks many motility genes and has a distinct carbohydrate metabolism
634 potential. Karcher et al. suggested that ErEurope utilizes complex plant-derived carbohydrates
635 more efficiently (59). As no interaction with diet was previously observed for ErEurope, most
636 of the different *E. rectale* strains do not co-colonize (59), and no differences in prevalence or
637 abundance were observed between tCeD and controls, it is possible that the strain-level
638 difference in *E. rectale* may exist prior to GFD. It remains unclear whether the lack of motility
639 limits the ability of ErEurope to access carbohydrates during inflammation, whether their
640 carbohydrate metabolism potential plays a role in gluten processing, or whether pre-existent
641 genetic or environmental risk factors for CeD also affect *E. rectale* substrain colonization
642 success.
643
644 *Limitations*
645 Since the patients with CeD in our study strictly followed a GFD, our findings do not allow us
646 to definitively ascertain the role of the microbiome in the onset or progression of the disease.
647 This challenge is common in much of the existing research, where data interpretation is
648 hindered by diet and the possibility that observed differences stem from studying patients
649 with active CeD, potentially influenced by associated intestinal inflammation. Furthermore,
650 our study is limited by the use of faecal samples, which may not accurately reflect conditions
651 in the small intestine.
652
653 Despite these limitations, the current findings are a starting point for future work, addressing
654 questions on which differences do, in fact, reflect changes in the small intestine, which of
655 these precede the onset of disease, and what mechanisms are responsible for these

differences in microbial composition. Additional future data, ideally obtained from prospective and longitudinal studies - including those of active CeD patients, healthy individuals adhering to a GFD, and/or at-risk individuals - will help pull apart the various influences of genetics, diet, and disease on the gut microbiome of CeD patients. A longitudinal study, where data is available before CeD onset, was published in 2021 and reported. However, that study had limited taxonomic resolution (16S sequencing) and statistical power, with only 10 cases and 10 controls (51). Furthermore, as the significance levels in this study were, apparently, not corrected for multiple testing, it is difficult to judge whether the reported differences are beyond what would be expected by chance.

Conclusion

We characterized the microbiota of tCeD patients, as compared to controls, leading to both the confirmation of previously reported CeD-microbiome associations and the identification of novel associations, that may be linked to either CeD, GFD, or both. Strain- and gene-level comparisons revealed a higher mutation frequency in several species, which may be induced by inflammation or oxidative stress. Data on the prevalence of the RadC-like JAB domain in multiple *Bacteroides* species suggests that the increased mutational load may be beneficial for bacteria in stressful conditions. Finally, we find that specifically the European subspecies of *Eubacterium rectale* is virtually absent in tCeD patients, a finding that may, in the future, help to shed light on some of the mechanisms that shape the microbiome in health and disease.

677 # References

678 1. Ludvigsson JF, Leffler DA, Bai JC, *et al.* The Oslo definitions for coeliac disease and related
679 terms. Gut 2013;62:43–52.

680 2. Lebwohl B, Sanders DS, Green PHR. Coeliac disease. Lancet 2018;391:70–81.

681 3. Lebwohl B, Ludvigsson JF, Green PHR. Celiac disease and non-celiac gluten sensitivity.
682 BMJ 2015;351:h4347.

683 4. Gujral N, Freeman HJ, Thomson ABR. Celiac disease: prevalence, diagnosis, pathogenesis
684 and treatment. World J Gastroenterol 2012;18:6036–6059.

685 5. Rewers M. Epidemiology of celiac disease: what are the prevalence, incidence, and
686 progression of celiac disease? Gastroenterology 2005;128:S47-51.

687 6. Kårhus LL, Thuesen BH, Skaaby T, *et al.* The distribution of HLA DQ2 and DQ8 haplotypes
688 and their association with health indicators in a general Danish population. United
689 European Gastroenterol J 2018;6:866–878.

690 7. Bodd M, Ráki M, Tollefsen S, *et al.* HLA-DQ2-restricted gluten-reactive T cells produce IL-
691 21 but not IL-17 or IL-22. Mucosal Immunol 2010;3:594–601.

692 8. Di Sabatino A, Vanoli A, Giuffrida P, *et al.* The function of tissue transglutaminase in
693 celiac disease. Autoimmun Rev 2012;11:746–753.

694 9. Troncone R, Jabri B. Coeliac disease and gluten sensitivity. J Intern Med 2011;269:582–
695 590.

696 10. Al-Toma A, Volta U, Auricchio R, *et al.* European Society for the Study of Coeliac Disease
697 (ESsCD) guideline for coeliac disease and other gluten-related disorders. United
698 European Gastroenterol J 2019;7:583–613.

699 11. Heel DA van, Franke L, Hunt KA, *et al.* A genome-wide association study for celiac disease
700 identifies risk variants in the region harboring IL2 and IL21. Nat Genet 2007;39:827–829.

701 12. Hunt KA, Zhernakova A, Turner G, *et al.* Newly identified genetic risk variants for celiac
702 disease related to the immune response. Nat Genet 2008;40:395–402.

703 13. Dubois PCA, Trynka G, Franke L, *et al.* Multiple common variants for celiac disease
704 influencing immune gene expression. Nat Genet 2010;42:295–302.

705 14. Kumar V, Wijmenga C, Withoff S. From genome-wide association studies to disease
706 mechanisms: celiac disease as a model for autoimmune diseases. Semin Immunopathol
707 2012;34:567–580.

708 15. Collado MC, Calabuig M, Sanz Y. Differences between the fecal microbiota of coeliac
709 infants and healthy controls. Curr Issues Intest Microbiol 2007;8:9–14.

710 16. Sanz Y, Sánchez E, Marzotto M, *et al.* Differences in faecal bacterial communities in
711 coeliac and healthy children as detected by PCR and denaturing gradient gel
712 electrophoresis. FEMS Immunology & Medical Microbiology 2007;51:562–568.

713 17. Nadal I, Donant E, Ribes-Koninckx C, *et al.* Imbalance in the composition of the duodenal
714 microbiota of children with coeliac disease. J Med Microbiol 2007;56:1669–1674.

715 18. Collado MC, Donat E, Ribes-Koninckx C, *et al.* Specific duodenal and faecal bacterial
716 groups associated with paediatric coeliac disease. Journal of Clinical Pathology
717 2009;62:264–269.

718 19. Collado MC, Donat E, Ribes-Koninckx C, *et al.* Imbalances in faecal and duodenal
719 Bifidobacterium species composition in active and non-active coeliac disease. BMC
720 Microbiol 2008;8:232.

721 20. Caminero A, McCarville JL, Galipeau HJ, *et al.* Duodenal bacterial proteolytic activity
722 determines sensitivity to dietary antigen through protease-activated receptor-2. Nat
723 Commun 2019;10:1198.

724 21. Constante M, Libertucci J, Galipeau HJ, *et al.* Biogeographic Variation and Functional
725 Pathways of the Gut Microbiota in Celiac Disease. Gastroenterology 2022;163:1351-
726 1363.e15.

727 22. Caminero A, Galipeau HJ, McCarville JL, *et al.* Duodenal bacteria from patients with
728 celiac disease and healthy subjects distinctly affect gluten
729 breakdown and immunogenicity. Gastroenterology 2016;151:670–683.

730 23. Petersen J, Ciacchi L, Tran MT, *et al.* T cell receptor cross-reactivity between gliadin and
731 bacterial peptides in celiac disease. Nat Struct Mol Biol 2020;27:49–61.

732 24. Welch NL, Welch T, Songtanin B. Viral Triggered Celiac Disease: A Case Report. Cureus
733 15:e40429.

734 25. Kagnoff MF, Austin RK, Hubert JJ, *et al.* Possible role for a human adenovirus in the
735 pathogenesis of celiac disease. J. Exp. Med. 1984;160:1544–1557.

736 26. Stene LC, Honeyman MC, Hoffenberg EJ, *et al.* Rotavirus infection frequency and risk of
737 celiac disease autoimmunity in early childhood: a longitudinal study. Am J Gastroenterol
738 2006;101:2333–2340.

739 27. Palmieri O, Castellana S, Bevilacqua A, *et al.* Adherence to Gluten-Free Diet Restores
740 Alpha Diversity in Celiac People but the Microbiome Composition Is Different to Healthy
741 People. Nutrients 2022;14:2452.

742 28. Francavilla A, Ferrero G, Pardini B, *et al.* Gluten-free diet affects fecal small non-coding
743 RNA profiles and microbiome composition in celiac disease supporting a host-gut
744 microbiota crosstalk. Gut Microbes 2023;15:2172955.

745 29. Gacesa R, Kurilshikov A, Vich Vila A, *et al.* Environmental factors shaping the gut
746 microbiome in a Dutch population. Nature 2022;604:732–739.

747 30. Ho D, Imai K, King G, *et al.* MatchIt: Nonparametric Preprocessing for Parametric Causal
748 Inference. Journal of Statistical Software 2011;42:1–28.

749 31. Imhann F, Van der Velde KJ, Barbieri R, *et al.* The 1000IBD project: multi-omics data of
750 1000 inflammatory bowel disease patients; data release 1. BMC Gastroenterol
751 2019;19:5.

752 32. Tigchelaar EF, Zhernakova A, Dekens JAM, *et al.* Cohort profile: LifeLines DEEP, a
753 prospective, general population cohort study in the northern Netherlands: study design
754 and baseline characteristics. BMJ Open 2015;5:e006772.

755 33. McIver LJ, Abu-Ali G, Franzosa EA, *et al.* bioBakery: a meta'omic analysis environment.
756 Bioinformatics 2018;34:1235–1237.

757 34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods
758 2012;9:357–359.

759 35. Beghini F, McIver LJ, Blanco-Míguez A, *et al.* Integrating taxonomic, functional, and
760 strain-level profiling of diverse microbial communities with bioBakery 3. Elife
761 2021;10:e65088.

762 36. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat
763 Methods 2015;12:59–60.

764 37. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids
765 Res. 2017;45:D158–D169.

766 38. Lex A, Gehlenborg N, Strobelt H, *et al.* UpSet: Visualization of Intersecting Sets. IEEE
767 Trans Vis Comput Graph 2014;20:1983–1992.

768 39. Dixon P. VEGAN, a package of R functions for community ecology. Journal of Vegetation
769 Science 2003;14:927–930.

770 40. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
771 multidimensional genomic data. Bioinformatics 2016;32:2847–2849.

772 41. Swertz MA, Dijkstra M, Adamusiak T, *et al.* The MOLGENIS toolkit: rapid prototyping of
773 biosoftware at the push of a button. BMC Bioinformatics 2010;11 Suppl 12:S12.

774 42. Schippa S, Iebba V, Barbato M, *et al.* A distinctive "microbial signature" in celiac pediatric
775 patients. BMC Microbiol 2010;10:175.

776 43. Pisani A, Rausch P, Bang C, *et al.* Dysbiosis in the Gut Microbiota in Patients with
777 Inflammatory Bowel Disease during Remission. Microbiology Spectrum 2022;10:e00616-
778 22.

44. Bourgonje AR, Roo-Brand G, Lisotto P, *et al.* Patients With Inflammatory Bowel Disease Show IgG Immune Responses Towards Specific Intestinal Bacterial Genera [Internet]. Front. Immunol. 2022;13[cited 2024 Mar 6] Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2022.842911/full

45. Vila AV, Hu S, Andreu-Sánchez S, *et al.* Original research: Faecal metabolome and its determinants in inflammatory bowel disease. Gut 2023;72:1472.

46. Pittayanon R, Lau JT, Yuan Y, *et al.* Gut Microbiota in Patients With Irritable Bowel Syndrome—A Systematic Review. Gastroenterology 2019;157:97–108.

47. Di Cagno R, De Angelis M, De Pasquale I, *et al.* Duodenal and faecal microbiota of celiac children: molecular, phenotype and metabolome characterization. BMC Microbiol 2011;11:219.

48. El Mouzan M, Al-Hussaini A, Serena G, *et al.* Microbiota profile of new-onset celiac disease in children in Saudi Arabia. Gut Pathog 2022;14:37.

49. Sánchez E, Laparra JM, Sanz Y. Discerning the Role of Bacteroides fragilis in Celiac Disease Pathogenesis. Appl Environ Microbiol 2012;78:6507–6515.

50. Sánchez E, Donat E, Ribes-Koninckx C, *et al.* Intestinal Bacteroides species associated with coeliac disease. Journal of Clinical Pathology 2010;63:1105–1111.

51. Leonard MM, Valitutti F, Karathia H, *et al.* Microbiome signatures of progression toward celiac disease onset in at-risk children in a longitudinal prospective cohort study. Proceedings of the National Academy of Sciences 2021;118:e2020322118.

52. De Palma G, Nadal I, Collado MC, *et al.* Effects of a gluten-free diet on gut microbiota and immune function in healthy adult human subjects. Br. J. Nutr. 2009;102:1154–1160.

53. Valitutti F, Cucchiara S, Fasano A. Celiac Disease and the Microbiome. Nutrients 2019;11:2403.

54. Lionetti E, Dominijanni V, Iasevoli M, *et al.* Effects of the supplementation with a multispecies probiotic on clinical and laboratory recovery of children with newly diagnosed celiac disease: A randomized, placebo-controlled trial. Digestive and Liver Disease 2023;55:1328–1337.

55. Dieterich W, Schuppan D, Schink M, *et al.* Influence of low FODMAP and gluten-free diets on disease activity and intestinal microbiota in patients with non-celiac gluten sensitivity. Clin Nutr 2019;38:697–707.

56. Bonder MJ, Tigchelaar EF, Cai X, *et al.* The influence of a short-term gluten-free diet on the human gut microbiome. Genome Med 2016;8:45.

57. Sample D, Fouhse J, King S, *et al.* Baseline Fecal Microbiota in Pediatric Patients With Celiac Disease Is Similar to Controls But Dissimilar After 1 Year on the Gluten-Free Diet

[Internet]. JPGN reports 2021;2[cited 2023 Nov 24] Available from: https://pubmed.ncbi.nlm.nih.gov/37206457/

58. Gophna U, Konikoff T, Nielsen HB. Oscillospira and related bacteria – From metagenomic species to metabolic features. Environmental Microbiology 2017;19:835–841.

59. Karcher N, Pasolli E, Asnicar F, *et al.* Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol 2020;21:138.

60. Iyer LM, Zhang D, Rogozin IB, *et al.* Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic Acids Res 2011;39:9473–9497.

61. Ruiz L, Bottacini F, Boinett CJ, *et al.* The essential genomic landscape of the commensal Bifidobacterium breve UCC2003. Sci Rep 2017;7:5648.

62. Amiri P, Hosseini SA, Ghaffari S, *et al.* Role of Butyrate, a Gut Microbiota Derived Metabolite, in Cardiovascular Diseases: A comprehensive narrative review [Internet]. Frontiers in Pharmacology 2022;12[cited 2024 Jan 10] Available from: https://www.frontiersin.org/articles/10.3389/fphar.2021.837509

63. Liu H, Wang J, He T, *et al.* Butyrate: A Double-Edged Sword for Health? Adv Nutr 2018;9:21–29.

64. Baldi S, Menicatti M, Nannini G, *et al.* Free Fatty Acids Signature in Human Intestinal Disorders: Significant Association between Butyric Acid and Celiac Disease. Nutrients 2021;13:742.

65. Tjellström B, Högberg L, Stenhammar L, *et al.* Faecal short-chain fatty acid pattern in childhood coeliac disease is normalised after more than one year's gluten-free diet. Microb Ecol Health Dis 2013;24:10.3402/mehd.v24i0.20905.

66. De Palma G, Nadal I, Medina M, *et al.* Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. BMC Microbiol 2010;10:63.

67. Sánchez E, De Palma G, Capilla A, *et al.* Influence of Environmental and Genetic Factors Linked to Celiac Disease Risk on Infant Gut Colonization by Bacteroides Species ▿ . Appl Environ Microbiol 2011;77:5316–5323.

68. De Palma G, Capilla A, Nova E, *et al.* Influence of Milk-Feeding Type and Genetic Risk of Developing Coeliac Disease on Intestinal Microbiota of Infants: The PROFICEL Study. PLoS One 2012;7:e30791.

69. Esteban-Torres M, Ruiz L, Rossini V, *et al.* Intracellular glycogen accumulation by human gut commensals as a niche adaptation trait. Gut Microbes 15:2235067.

70. Brüssow H. Problems with the concept of gut microbiota dysbiosis. Microb Biotechnol 2019;13:423–434.

851 71. Tojo R, Suárez A, Clemente MG, *et al.* Intestinal microbiota in health and disease: role of
852 bifidobacteria in gut homeostasis. World J Gastroenterol 2014;20:15163–15176.

853 72. Ferretti G, Bacchetti T, Masciangelo S, *et al.* Celiac Disease, Inflammation and Oxidative
854 Damage: A Nutrigenetic Approach. Nutrients 2012;4:243–257.

855 73. Ramírez-Sánchez AD, Zühlke S, Aguirre-Gamboa R, *et al.* Gene expression and eQTL
856 analysis reflect the heterogeneity in the inflammatory status of the duodenal epithelial
857 lining in coeliac disease [Internet]. 2024;2024.02.29.582756.[cited 2024 Mar 7] Available
858 from: https://www.biorxiv.org/content/10.1101/2024.02.29.582756v1

859 74. Galhardo RS, Hastings PJ, Rosenberg SM. Mutation as a Stress Response and the
860 Regulation of Evolvability. Critical Reviews in Biochemistry and Molecular Biology
861 2007;42:399.

862 75. Radman M. SOS repair hypothesis: phenomenology of an inducible DNA repair which is
863 accompanied by mutagenesis. Basic Life Sci 1975;5A:355–367.

864 76. Prudhomme M, Attaiech L, Sanchez G, *et al.* Antibiotic stress induces genetic
865 transformability in the human pathogen *Streptococcus pneumoniae*. Science
866 2006;313:89–92.

867 77. Slager J, Kjos M, Attaiech L, *et al.* Antibiotic-induced replication stress triggers bacterial
868 competence by increasing gene dosage near the origin. Cell 2014;157:395–406.

869