Supplementary materials

Sequence search and protein characterization

To obtain a comprehensive database of O-GlcNAcylation regulating proteins, we downloaded amino acid sequences that matched the keywords "O-GlcNAcase" or "O-GlcNAc transferase" from UniRef90,[1] and the taxonomy was restricted to "bacteria". All the results were manually inspected to exclude mismatches or fragment sequences.

Taxonomic information regarding these proteins was directly extracted from the FASTA name. The conserved domains of proteins were predicted using website tool PFAM with default parameter: http://pfam.xfam.org/. The proteins sub-cellular localization predictions were done with PSORTb v.3.0. Multiple sequences similarities comparisons were calculated with Clustal Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/.

Taxonomic distribution of OGA in bacteria from human gut

The protein coding sequences of 1520 reference genomes from cultivated human gut bacteria were download from NCBI.[2] The 225 OGA sequences retrieved from UniRef90 were searched against the protein coding sequences of the 1520 reference genomes using BLASTP with cut-off of identity > 80%, coverage >80% and e-value < 1e-5.

Phylogenetic tree

To investigate differences in OGA sequence types, we clustered OGA sequences at 70% sequence identity using CD-hit (version 4.8.1).[3] The resulting represent proteins were aligned with Mafft (v7.450)[4] and phylogenetic tree was built using Fasttree (version 2.1.10) with default parameters.[5]

To investigate if UniRef90_A0A139TME1 was horizontal transferred among different phylum, we search for UniRef90_A0A139TME1 homologues (identity: 40%, coverage 70%) in NCBI non-redundant protein database, which were downloaded

from: ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz on 2020-09-05. The resulting sequences were used to build the phylogenetic tree as described above. The taxonomic information of these sequences was extracted from the corresponding taxonomy database (https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/) using Taxonkit (version 0.5.0).

Genomic Compositional Analysis

GC content, GC3s content, and codon usage bias of each gene in *A.muciniphila* were calculated using CodonW (version 1.3). Four indices of codon usage bias included CAI (codon adaptation index), CBI (codon bias index), FOP (frequency of optimal codon), and ENC (effective number of codons).

Calculate the lineage probability index (LPI) of OGA genes in different genomes

To investigate the possibility of horizontal transfer of other OGA genes, we download 82 genomes of OGA gene-encoding strains from the 1520 reference genomes described in section "Taxonomic distribution of OGA in bacteria from human gut", which covered 9 genus from Bacteroidetes and Firmicutes. Analysis of these genomes for horizontal gene transfer candidates was conducted with the program DarkHorse (version: 2.0_rev09)[6]: BLAST queries for all predicted proteins encoded in the genomes were carried out using the diamond. A maximum of 500 target sequences with e-values < 10^{-5} were accepted for each search. The non-self taxonomies were defined at phylum level.

Dataset in this study

The public metagenomic sequence data of individuals were collected from 11 cohorts covering samples from six different countries and six diseases.[7-17] The detailed information was shown in **Supplementary Table 6**.

To construct metagenomic datasets of healthy individuals from each datasets, we screened out the data for individuals free of any recorded diseases, including impaired glucose tolerance and hypertension **(Supplementary Table 6)**.

Identification and quantification of OGA genes in human gut microbiota

The above UniRef90 FASTA sequence were aligned to the IGC using BLASTP (version 2.2.29+),[16,18] Matching queries were filtered to include only alignments with > 80% identity, > 80% coverage, and an e-value < 1e-5. The IGC database was downloaded from http://meta.genomics.cn/meta/home on 2019-09-01. Since more than half of the OGA genes from IGC didn't have a taxonomy assignment at the phylum level, the taxonomy information of these OGA genes from IGC were assigned as its homologues in Uniref90 in the subsequent analysis.

Two methods are employed to quantify OGA gene abundance in human metagenomic datasets.

Method 1: After quality control, reads were mapped to all 9878647 gene catalog from IGC using Bowtie2 (version 2.2.3)[19]with settings:--sensitive-local. Alignments with < 95% nucleotide identity or where the read was covered by < 90% of it's length were discarded. The read-depth of each reference gene was quantified based on mapped reads and the gene relative abundance was estimated by scaling abundance to sum to 1.0 across genes per sample.

Method 2: ShortBred[20] (Version 0.9.5) were employed to target quantify OGA gene family in two IBD cohorts. Firstly, short sequences markers of OGA family were identified using UniRef90 database with default parameters, and then quantified these markers in the metagenomic samples, normalizing by the number of RPKM.

Taxonomic quantification of metagenomic datasets

Quality-filtered metagenomic datasets were taxonomically profiled using MetaPhlAn2 (version 2.7.7) with default parameters. Linear Discriminant Analysis (LDA) Effect Size (LEfSe) analysis was performed with website tool http://huttenhower.sph.harvard.edu/galaxy/root, to identify bacterial taxa that were differentially abundant in groups (UC vs Control). In detail, bacterial taxa that were differentially abundant in groups were first identified and tested using the Kruskal Wallis test (p<0.05).The identified features were then subjected to the linear discriminant analysis (LDA) model with a threshold logarithmic LDA score set at 2.0

and ranked. Respective cladograms were generated with genus at the lowest level. Quantitative plots of differential features were generated showing means differences with standard deviation between groups.

Sequence specific analysis of metagenomic OGA sequences

To look into the sequence specific differences among OGAs from different samples, we download all of the metagenomic gene datasets from USA-IBD cohort (https://ibdmdb.org/tunnel/public/HMP2/WGS/1818/products), which were predicted using IGS Metagenomics Pipeline. BLASTP was employed to search against the metagenomic protein coding sequences for OGA protein sequences (cut off: >80% identity, >80% coverage, e-values <1e-5). The NAGIdase domains of these metagenomic OGA sequences were extracted and the Shannon entropy of each amino acid was calculated using Oligotype (version 3): the higher entrophy value for each amino acids means the higher variation between samples.[21] The 3D domain structure of NAGIdase were download from the PDB database (2VVN) and visualized with PyMol (version:2.3.2).

O-GlcNAcylated peptides enrichment and liquid chromatography with tandem mass spectrometry analysis

Jurkat cells were cultured for 2 days and treated with PBS (Con), LPS (100 ng/mL), BtGH84 (100 μg/mL) + LPS (100 ng/mL) or AkkGH84 (100 μg/mL) + LPS (100 ng/mL). BtGH84 or AkkGH84 were added 2 h before LPS treatment. After treatment, cells were lysis with RIPA buffer (50 mM Tris-HCl, 150 mM NaCl, 2 mM NaF, 1 mM EDTA, 1 mM EGTA, 1 mM NaVO4, and 1% Triton X-100). Protein samples were collected and digested into peptides using trypsin, chymotrypsin and Glu-C according to the filter aided sample preparation protocol. The dried digested peptides were resuspended with 80% acetonitrile containing 2% formic acid and loaded onto a 8 mg hydrophilic interaction liquid chromatography (HILIC) material-packed pipet tip to enrich the O-GlcNAcylated peptides. Nonspecifically adsorbed non-O-GlcNAc peptides were removed with 80% acetonitrile containing 5%

formic acid. Afterward, O-GlcNAcylated peptides were eluted and subjected to mass spectrometry for identification and quantification.

LC-MS/MS analysis was carried out using an Ultimate 3000 system coupled with an Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer (Thermo Fisher Scientific). The mobile phase A was 0.1% formic acid and 2% acetonitrile; the mobile phase B was 0.1% formic acid and 80 % acetonitrile; Then, 500 ng enriched peptides were injected and separated on a reversed-phase column (150 mm×150 μm inner diameter) packed with 1.9 μm C18 particles using a 120 min acetonitrile gradient in 0.1% formic acid at a flow rate of 600 nL/min. The mass spectrometer was set to a full scan range of 350-1550 m/z with a resolution of 70 000. The top 20 most intensive ions were selected for MS/MS with higher-energy collision dissociation (HCD) using a resolution of 17500. For data processing, all MS and MS/MS raw spectra were searched by Byonic v.2.82 (protein metercs, San Carlos, CA) using a human protein database. A maximum of three missed cleavages and a minimum peptide length of 7 amino acids were allowed. Carbamidomethyl (Cys) was set as the fixed modification. O-GlcNAcylated Ser/Thr, Acetyl (protein N-term) and oxidation (Met) were set as the variable modifications. The score >100 of identified O-GlcNAc proteins were collected and use to delineate the role of O-GlcNAcylated proteins in bacterial OGAs-mediated protective effect.

Label-free intensity-based quantification was performed to determine the quantitative changes of O-GlcNAc sites in Jurkat during PBS, LPS, BtGH84+LPS or AkkGH84+LPS treatment. A confidence index based on at least two technical replicates was built for each glycopeptide.

Immunoprecipitation, sWGA pull down, immunoblot and ELISA

Caco-2 or Jurkat treated with Con (PBS) LPS, BtGH84+LPS, AkkGH84+LPS were collected and lysed in RIPI lysis buffer containing 50 mM Tris-Cl (pH 8.0), 150mM NaCl, 5 mM EDTA, 0.5% Triton X-100, 0.1% sodium deoxycholate, protease and phosphatase inhibitor cocktail (Roche). Lysates were incubated with rotation at 4 °C for 30 min and then centrifuged at 12000 rpm at 4 °C for 15 min,

followed by collection of supernatant for further pre-clear by protein A beads (Santa Cruz). For immunoprecipitation, 5 μg of anti-NF-κB p65 antibody (Abcam) or anti-IKKβ antibody (Abcam) were incubated with 500 μg precleared cell lysates at 4 °C for overnight and then protein A agarose beads (Santa Cruz) were added to incubate for another 2 h to precipitate the protein-antibody complex. For the purpose of enriching O-GlcNAcylated protein, 50 μl succinylated wheat germ agglutinin agarose (sWGA, Vector Laboratories) were incubated with 250 μg cell lysates at 4 °C for 12 h. Proteins pulled down by succinylated wheat germ agglutinin agarose were then eluted by 5 × sample buffer containing 250 mM Tris-Cl (pH 6.8), 4 % bromophenol blue, 2 % SDS, 50 % glycerol and 10 % β-mercaptoethanol at 95 °C for 5 min for subsequent immunoblotting analysis.

Protein samples were separated on a SDS-polyacrylamide gel and transferred onto polyvinylidene difluoride membranes (Millipore). The membranes were blocked and incubated with anti-O-GlcNAc antibody (CTD110.6) (Cell Signaling Technology); anti-NF-κB p65 antibody (Abcam); anti-IKKβ antibody (Abcam); anti-IκBα antibody (Abcam) or anti-Histone-H3 antibody (Abcam). Expression of primary antibodies was visualized using HRP-coupled second antibodies and enhanced chemiluminescence reagent kit (Bio-Rad Laboratories, USA). β-actin and Histone-H3 were used as loading control. The concentrations of TNF-α, IL-6, IL-1β, IL-8 and IL-17A were evaluated in cell lysates or colonic extracts using specific ELISA kits (BD Biosciences) according to the manufacturer's instructions.

NF-κB transcriptional activity assay

NF-κB-luciferase promoter-reporter system were used to measure the transcriptional activity of NF-κB. To evaluated the NF-κB transcriptional activity, Jurkat or Caco-2 were transfected with 1 μg of NF-κB-luciferase promoter-reporter construct (pGL4.32 [luc2P/NF-κB-RE/Hygro]) using Lipofectamine 3000 (Invitrogen). The pRL-TK vector expressing wild-type Renilla luciferase was used as a control reporter. Twenty-four hours after transfection, the cells were treated with PBS, LPS, BtGH84+LPS or AkkGH84+LPS as described above. Luciferase activity

was assessed using the Dual-Luciferase Reporter Assay System (Promega) after the treatment. Relative luciferaseunit (RLU) was the ratio of NF-κB luciferase activity to Renilla activity. All experiments were performed in triplicate and relative luciferase activity was reported as the fold induction after normalization for transfection efficiency.

Pectin/zein hydrogel drug delivery system

In order to avoid the protease attack, pectin/zein beads delivery system was used to delivery of protein to animal colon.[22] Briefly, zein (Sigma-Aldrich) was dissolved in 85% alcohol solution (containing 0.5% calcium chloride, w/v) at a concentration of 10 mg/mL. Recombinant protein was mixed with 60 mg/mL pectin solution to prepare pectin/protein solution. Then, the pectin/protein solution was inhaled into a syringe with a 23 G needle, and dropped (approximately 50 μl/drop) into the zein solution gently. Pectin/protein drop was hardened into a bead in zein solution immediately. The beads were collected and washed using distilled water several times. Bovine serum albumin (BSA) beads were prepared simultaneously and used as control. The beads with an average size of 2 mm and contained 5 μg protein.

The induction of TNBS- and OXA-colitis

The procedure for induction of TNBS colitis and OXA-colitis model was similarly, as described in **Supplementary Figure 12**. Briefly, age- and sex-matched C57BL/6J mice were presensitized by epicutaneous application of 150 μl of 1% (wt/vol) TNBS or 3% (wt/vol) oxazolone on the day 1. On day 8, anesthetized the mouse by i.p. dosing of 80 μl/10 g body weight of ketamine/xylazine solution, and followed by intrarectal administration of 100 μl of 2.5% (wt/vol) TNBS solution or 1% oxazolone in 50% ethanol using a 1 ml syringe to a 3.5 F catheter. BtGH84 and AkkGH84 were administrated on the day 8 daily to day 13. Body weight was determined daily. Colon length was measured at the end of the experiment. Colon tissues were fixed in 4% paraformaldehyde and embedded in paraffin, followed by

hematoxylin-eosin staining. Severity of inflammation was scored based on body weight loss, occult blood and stool consistency in a blind manner.

Statistical analysis

Statistical analyses were performed in the R language. In metagenomic analysis, differences between the two groups were determined by the Manne-Whitney U test, followed by false discovery rate correction. The relation analysis of OGA abundance and other factors was done with Spearman correlation (package psych) and multiple linear regression (package stats). Logistic regression (function glm) (family = binomial) were used to analyse the influence of OGA abundance on disease status. In bio-experiments analysis, differences between two groups were determined using unpaired two-tailed Student's t-test. Differences involving more than two groups were using two-tailed, one-way analysis of variance with multiple comparison post hoc analysis. $P < 0.05$ was considered as statistically significant.

References:

1. Apweiler R, Bairoch A, Wu CH, *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;32:D115-9.
2. Yuanqiang Z, Wenbin X , Guangwen L, *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 2019; 37:179-85.
3. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150-2.
4. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772-780. doi:10.1093/molbev/mst010.
5. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;26:1641-50.
6. Sheila Podell, Terry Gaasterland. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* 2007;8:R16.

7. HMP Consortium H. A framework for human microbiome research. *Nature* 2012;486:215-21.
8. Qin N, Yang F, Li A, *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* 2014;513:59-64.
9. Le Chatelier E, Nielsen T, Qin J, *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* 2013;500:541-46.
10. Zhang X, Zhang D, Jia H, *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 2015;21:895-905.
11. Qin J, Li Y, Cai Z, *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55-60.
12. Karlsson FH, Tremaroli V, Nookaew I, *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;498:99-103.
13. Nielsen HB, Almeida M, Juncker AS, *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;32:822-28.
14. Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;569:655-62.
15. Feng Q, Liang S, Jia H, *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat Commun* 2015;6:6528.
16. Li J, Jia H, Cai X, *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;32:834-41.
17. Qin J, Li R, Raes J, *et al*.A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464(7285):59-65.
18. Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
19. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nat Meth* 2012;9:357-9.

20. Kaminski J, Gibson MK, Franzosa EA, *et al*. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol* 2015;11(12):e1004557.
21. Eren, A. M, Maignien, L, Sul, W. J *et al.* Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 2013;4:1111-9.
22. Jie Gao, Yubin Li, Yu Wan, *et al.* A Novel Postbiotic From Lactobacillus rhamnosus GG With a Beneficial Effect on Intestinal Barrier Function. *Front Microbiol* 2019;10:477.