Article │Received 6 December 2024; Accepted 16 May 2025; Published 26 May 2025
https://doi.org/10.55092/bi20250003

Chemical features of proteins in microbial genomes associated with body sites and gut inflammation

Jeffrey M. Dick

Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring (Ministry of Education), School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; E-mail: jeff@chnosz.net.

Highlights:

- Chemical analysis quantifies water and oxygen content of microbial protein sequences.
- Lower water content in gut bacterial proteins compared to other body sites.
- Gut inflammation reduces oxygen content of bacterial community proteins.
- Obligate anaerobes have more oxidized proteins than aerotolerant bacteria in the gut.

Abstract: Human bodies host complex communities of microorganisms that adapt to different environments, from the gut to other body sites. This study leverages new chemical information from multi-omics datasets to understand how bacterial proteins change in response to two critical factors: oxygen and water availability. Chemical features of proteins were quantified by a computational approach that combines reference genomes with microbial abundances to assess community-level trends. We discovered that microbial proteins vary across different body sites, with the gut presenting unique characteristics. First, gut bacterial proteins have lower water content compared to bacteria in other body areas. This suggests that the intestinal environment drives specific evolutionary adaptations. Second, in patients with inflammatory conditions like COVID-19 and inflammatory bowel disease (IBD), gut bacterial proteins show distinctive chemical changes. Despite the oxidizing conditions associated with gut inflammation, bacterial proteins become more chemically reduced due to the shifting abundances of different types of bacteria. This unexpected result leads to the insight that some bacteria that typically thrive in oxygen-free environments (anaerobic bacteria such as *Faecalibacterium*) have more oxidized proteins than those in aerotolerant bacteria. This can help anaerobes survive and compete when the gut's chemical conditions become more challenging during inflammation. By applying advanced computational techniques to a large collection of microbial community datasets, this research reveals that bacterial genomes actively evolve to survive in specific chemical conditions.

Keywords: host physiology; microbial genomics; multi-omics; chemical features; oxygen; water content; gut microbiome; IBD; COVID-19; inflammation

1. Introduction

Understanding how microorganisms adapt to our body's unique chemical environments provides insights into health and disease. Microbial communities must deal with and adapt to their chemical environments. Importantly, the chemical environment matters for host-associated communities and not only free-living communities [1]. Variations in oxygen concentration and water availability create unique selective pressures across different host body sites, driving microbial adaptation strategies. The human gut, in particular, maintains a complex oxygen gradient that can be disrupted during inflammatory conditions, potentially triggering shifts in microbial community composition sometimes referred to as "dysbiosis" [2–4].

Current understanding of microbial genomic adaptation remains limited, especially regarding how chemical factors shape protein-level transformations. Oxygen and water, critical substrates in metabolic processes, play fundamental roles in determining microbial survival and evolutionary trajectories. Previous research has suggested that environmental oxygen levels influence protein elemental composition [5,6], but comprehensive investigations across different physiological contexts have been sparse. Furthermore, water has many roles in human physiology, as reflected in decreasing organismal water content during development from embryo to adult [7,8], higher water content in cancer compared to normal tissue [9], and gains and losses of H_2O from eukaryotic cells during transitions to cellular proliferation and dormancy [10,11]. A major function of the intestine is absorption of water from the digesta [12], and lower water content in the colon is associated with greater mucosal thickness toward the rectum [13]. Therefore, monitoring oxidation and hydration state at a molecular level may be important for understanding adaptation of microbial genomes to human host habitats.

This study aims to characterize genomic differences in microbial communities by examining chemical metrics of protein sequences across various body sites and inflammatory conditions. Specifically, we investigate how water content ($n\mathrm{H_2O}$) and oxygen content ($n\mathrm{O_2}$) reflect microbial adaptations to host environments. These values are the numbers of water and oxygen molecules in a theoretical reaction to form a protein from thermodynamic components, normalized by the number of amino acid residues. This is a new representation of sequence information, referred to here as "geochemical biology", that bridges biological data with environmental parameters.

Our research bridges multiple analytical domains, combining taxonomic abundance data, reference proteomes, and advanced chemical analysis to unveil nuanced mechanisms of microbial genomic evolution. A large collection of publicly available 16S rRNA datasets were curated for a chemical analysis by combining taxonomic abundances with reference proteomes for taxa in order to generate community reference proteomes as described previously [14,15]. We show that the trends of chemical metrics produced using reference proteomes are consistent with metagenomic data for the same samples. Then, community reference proteomes were used to observe chemical trends for communities in different body sites and differences between communities associated with two inflammatory diseases, COVID-19 and IBD.

The research addresses critical questions: How do chemical environments shape microbial protein characteristics? What evolutionary strategies do bacterial communities employ when confronting changing physiological conditions? By answering these questions, we provide novel insights into the dynamic relationship between host environments and microbial genomic adaptation.

1.1. Terminology and scope

The word "proteomics" refers to experimental characterization of the expression levels of proteins, but the specific term "reference proteome" is commonly used for genomically predicted protein sequences without expression levels. For instance, the authors of a study on taxonomic distribution of the opsin protein family [16] state that "UniProt Reference Proteomes are protein coding sequences derived from genome sequences". Similarly, the authors of a review paper on mass spectrometry-based proteomics for small protein discovery [17] use the term "reference proteome" in this context: "Sequences of the proteolytic peptides are inferred from their MS/MS spectra by matching the fragmentation patterns to theoretical spectra of a reference proteome database that contains the sequences of all annotated proteins of the target organism."

The compound term "community reference proteome" is used here to denote the genomically predicted protein sequences in a community. This compound term is not used in the proteomics literature, so it should not be confused with protein expression levels in proteomic or metaproteomic experiments.

The scope of this study is a description of chemical features of protein sequences at the community level. The methods are chosen to quantify how communities may be shaped by environmental factors. This approach aligns with guild-based microbiome studies to identify diverse groups that use environmental resources in a similar way [18]. For instance, part of this study looks at how anaerobic and aerotolerant subsets of communities, which are composed of various phylogenetic lineages, nevertheless exhibit convergent chemical features of their reference proteomes.

The second part of the scope, community-level differences, implies a high degree of aggregation of protein sequences from individual genomes. This method uses phylogenetic signal (e.g. average protein sequence compositions for genera) to distill genomic differences into a single quantity for each community. This approach aligns with the idea of using taxonomic information to predict trait differences between bacteria [19]. While this study is concerned with the aggregate manifestation of phylogenetic differences between communities, explaining the distribution of particular phylogenetic groups is outside the scope of this study.

The lack of species or subspecies-level taxonomic assignments represents limitations of the aggregation strategy. To address this limitation, comparisons are made with independent data for metagenomes. The comparisons show that community reference proteomes, which use only taxonomic classifications and reference genomes, have similar values for chemical features as proteins in metagenomes, which are based on shotgun DNA sequencing. This shows that the methodology is appropriate for the scope of this study.

No attempt is made here to explain the observed chemical variation in terms of other genomic features such as GC content. Previous authors have found that GC content doesn't capture the diversity of amino acid composition in different ecological settings [20] or the chemical properties of amino acid thought to be targeted by selection [21]. Documenting the chemical features of whole-community protein sequences as done in this study highlights environmental factors that may shape protein evolution, offering a foundation for future studies of how GC content or other genetic factors modulate these patterns.

Several datasets for metagenomes, metagenome-assembled genomes, and metaproteomes are analyzed in this study, and they are referred to by those names or abbreviations (MG, MAG, MP) to distinguish them community reference proteomes. In particular, the metaproteomic datasets analyzed

here use reported protein abundances from mass spectrometry experiments. These metaproteomic datasets are analyzed for comparison with community reference proteomes, with the caveat noted below that metaproteomes represent short time scales of expression level changes, unlike longer time-scale changes of protein sequences in genomes due to evolutionary processes.

2. Methods

2.1. Base assumptions of this study

This study probes how chemical features of microbiomes (specifically, bacterial communities) might reflect different chemical environments of microbial habitats in the human body. Reference proteomes for individual taxa (species, genus, *etc.*) do not vary by environment, but taxonomic abundances do. Multiplying the amino acid compositions of reference proteomes by taxonomic abundances produces community reference proteomes from which chemical features can be calculated. The logic is similar to existing methods for using taxonomic abundances to infer functional potential of communities (e.g. Tax4Fun [22]), but the target variables in this study are chemical features of protein sequences rather than metabolic or functional features.

2.2. Theory and calculation of chemical features

Thermodynamic theory provides a framework for linking elemental stoichiometry of biomolecules and the environment. A key concept is thermodynamic components, also known as basis species [23]. The basis species are a minimal set of molecular species that contain as many species as chemical elements, and which may be used to represent the composition of all other species (such as proteins). Expressing the composition of proteins in terms of basis species, or components, denotes a change in coordinates from elemental composition to chemical composition. There are five different elements in the primary sequences of proteins, so that is the number of thermodynamic components that is used. Because water and oxygen are major factors in metabolic reactions, they are the first two components chosen here. The remaining components, namely glutamine, glutamic acid, and cysteine (QEC) were chosen for several reasons. First, they represent nitrogen, carbon, and sulfur in biologically available forms, *i.e.* amino acids. Second, glutamine and glutamic acid are central metabolites with a relatively high degree of metabolic network connectivity [24]. Third, compared to other possible combinations of amino acids, QEC yields a strong positive correlation between oxygen content (the number of O_2 needed to compose a given protein from the basis species) and carbon oxidation state (Z_C) [24]. Notably, carbon oxidation state is an electronegativity-based measure that is independent of the choice of components, so the correlation between alternative metrics for oxidation state (Z_C and nO_2) reflects an underlying consistency. Finally, for all the proteins in a given genome, there is very little correlation between Z_C and nH_2O (*i.e.* stoichiometric water content) [25]. This important result shows that oxygen and water content computed using the chosen basis species represent distinct chemical features of proteins.

It should be noted that projecting the elemental composition of proteins in terms of basis species does not represent the actual mechanism of protein formation. Nevertheless, thermodynamic components represent a mathematical transformation from elemental composition to chemical

composition. Therefore, elemental compositions of protein sequences inferred from multi-omics datasets can be combined with methods of chemical analysis to test the hypothesis that oxygen and water availability shape the genomic adaptation of the human microbiota.

$n\mathrm{H_2O}$ and $n\mathrm{O_2}$ were calculated according to

$$\begin{bmatrix} 5 & 5 & 3 & 0 & 0 \\ 10 & 9 & 7 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 \\ 3 & 4 & 2 & 1 & 2 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} n\mathrm{Gln} \\ n\mathrm{Glu} \\ n\mathrm{Cys} \\ n\mathrm{H_2O} \\ n\mathrm{O_2} \end{bmatrix} = \begin{bmatrix} c \\ h \\ n \\ o \\ s \end{bmatrix} \quad (1)$$

where nGln, nGlu, nCys, $n\mathrm{H_2O}$, and $n\mathrm{O_2}$ are the stoichiometric coefficients of the basis species composing a protein with formula $\mathrm{C_cH_hN_nO_oS_s}$, and the matrix on the left represents the number of elements in each of the basis species. The values of $n\mathrm{H_2O}$ and $n\mathrm{O_2}$ reported here were normalized by protein length. Instead of calculating elemental formulas of proteins as an intermediate step, amino acid compositions were combined with precomputed values of $n\mathrm{O_2}$ and $n\mathrm{H_2O}$ for amino acid residues to calculate chemical metrics for proteins as described previously and implemented in the canprot R package [9,24].

Equation (1) encapsulates the underlying logic for turning elemental composition (a vector of abundances of elements) into chemical features such as $n\mathrm{H_2O}$ and $n\mathrm{O_2}$. In turn, the sources of data for elemental composition are protein sequences, specifically their amino acid composition. Amino acid compositions of proteins at the community level were generated in this study for community reference proteomes (made by combining taxonomic abundances and genomic reference databases) or inferred from metagenomes and metaproteomes. In addition to the methods outlined below, see Ref. [14] for a graphical overview of the method for generating community reference proteomes (but the RefSeq database used in that study is replaced by GTDB in this study) and Ref. [15] for a description of the software package used for this purpose.

2.3. 16S rRNA gene sequence processing

Literature searches were used to locate publicly available 16S rRNA gene sequencing datasets (Table 1). Gut microbiome datasets with at least 20 samples (total for controls and patients) available by the end of 2023 were included, while those for oropharyngeal and nasopharyngeal microbiomes available by the end of 2022 were included. Sequence data were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). FASTQ sequence files for paired-end sequences were merged using the "fastq_mergepairs" command of VSEARCH version 2.15.0 [26]. Quality filtering was done with maximum expected error rate of 0.005 and a sequence truncation length specific for each dataset (see Table S1 for details and sequence processing statistics). Singletons were removed and remaining sequences were subsampled at a depth of 10000. Reference-based chimera detection with the VSEARCH command "uchime_ref" was performed using the SILVA SSURef NR99 database version 138.1 [27]. Taxonomic classification was performed using the RDP Classifier version 2.13 [28] retrained with the Genome Taxonomy Database (bac120_ssu_reps and ar53_ssu_reps in GTDB release 220) [29]. Samples with less than 100 classified reads were excluded from the following analysis.

Table 1. Sources of 16S rRNA gene sequence data for COVID-19 and IBD.

No.	Accession	Ref.	*n*Control	*n*Patient	No.	Accession	Ref.	*n*Control	*n*Patient
Nasopharyngeal (COVID-19)					Oral or oropharyngeal (COVID-19)				
1	PRJNA774098	[33]	12	21	1	PRJNA767939	[34]	15	22
2	PRJNA714242	[35]	10	32	2	PRJNA692359	[36]	14	26
3	PRJNA673585	[37]	18	56	3	PRJNA780671	[38]	24	30
4	PRJNA777915	[39]	38	38	4	PRJNA684070	[40]	44	50
5	PRJNA726205	[41]	7	76	5	PRJNA669421	[42]	54	46
6	PRJNA707350	[43]	26	63	6	PRJNA639286	[44]	24	97
7	PRJNA726992	[45]	20	75	7	PRJNA683617	[46]	33	221
	PRJNA726994				8	PRJNA739539	[47]	140	166
8	PRJNA683617	[46]	34	216	9	PRJNA660302	[48]	150	242
9	PRJNA703574	[49]	91	226					
Gut (COVID-19)					Gut (IBD)				
1	PRJNA736160	[50]	14	15	1	PRJNA679275	[51]	10	10
2	PRJNA767939	[34]	15	22	2	PRJNA673073	[52]	10	11
3	PRJNA705797	[53]	14	25	3	PRJNA884507	[54]	10	11
4	PRJNA684070	[40]	32	17	4	PRJEB33711	[55]	9	21
5	PRJNA703303	[56]	30	26	5	PRJNA391149	[31]	22	19
6	PRJNA636824	[57]	30	30	6	PRJNA313074	[58]	13	30
7	PRJNA678695	[59]	34	50	7	PRJNA975689	[60]	6	42
8	PRJDB11949	[61]	61	45	8	PRJNA368966	[62]	32	77
	PRJDB12349				9	PRJEB47161	[63]	96	56
9	PRJNA753792	[64]	22	89		PRJEB47162			
10	PRJNA859805	[65]	52	59	10	PRJNA398089 (*)	[66]	43	112
11	PRJNA769052	[67]	50	78	11	PRJEB18471	[68]	62	124
12	PRJEB61722	[69]	38	100	12	PRJEB42155	[70]	19	186
	PRJEB61723				13	PRJNA237362	[30]	31	224
13	PRJNA660302	[48]	72	100	14	PRJNA431126	[32]	38	286
14	PRJNA758913	[71]	38	141	15	PRJNA398187 (*)	[72]	63	283
15	PRJNA756849	[73]	145	104					

Note: Accessions are NCBI BioProject numbers. Counts for control and patient are based on number of samples available from SRA, not numbers of subjects described in papers. Counts exclude samples with low taxonomic classification rate (see Methods) and therefore may be less than the number of available samples. Datasets are ordered by increasing number of samples in each disease and body site group. Asterisks (*) indicate datasets for mucosal samples; the remaining gut datasets are for fecal samples. Several studies [30–32] reported data for both mucosal and fecal samples, but only the data for fecal samples were analyzed here.

2.4. Reference proteomes for taxa

Reference proteomes of archaeal and bacterial taxa were made as described previously [74], except that GTDB (release 220 dated 2024-04-24) was used instead of the NCBI Reference Sequence database (RefSeq). Briefly, for each genome in GTDB, the amino acid composition of all proteins was summed and divided by the number of proteins. Then, amino acid compositions for all genomes in each genus were summed and divided by the number of genomes to generate the amino acid compositions of genus reference proteomes. Similarly, amino acid compositions for all genomes in each family, order, class,

phylum, and domain were summed and divided by the number of genomes to generate the amino acid compositions of reference proteomes for taxa at those levels.

Reference proteomes for species and genera in the Unified Human Gastrointestinal Genome (UHGG) were made in an analogous fashion. Taxonomic lineages and contamination and completeness values for each of the 4744 species-level clusters present in UHGG version 2.0.1 were obtained from the MGnify website (https://www.ebi.ac.uk/metagenomics/browse/genomes, accessed on 2023-12-30); those with contamination <2% and completeness >95% were used to generate reference proteomes for comparison with GTDB.

2.5. Community reference proteomes

The chem16S package version 1.2.0 [15] was used to compute chemical features of community reference proteomes. The lowest-level taxonomic classification, from genus to domain, for each processed 16S rRNA gene sequence was retained. The relative abundances of taxa in each sample were multiplied by the previously computed amino acid compositions of reference proteomes for taxa and summed to obtain the amino acid composition of the community reference proteome. chem16S was modified during this study to include domain-level classifications instead of being restricted to phylum level as described previously [14,15]. Because the great majority of classifications are resolved to lower taxonomic levels (mostly genus level; Table S1), this change make no discernible difference to the results, but it ensures that all of the available taxonomic information is used in the pipeline.

2.6. Metagenomic data processing: Shotgun metagenomes

Forward sequences from each sequencing run were processed for removal of adapter sequences and dereplication using scripts modified from the MG-RAST pipeline [75]. Human sequences were removed using bowtie2 version 2.5.0 [76] with the GRCh38 reference database (https://genome-idx.s3.amazonaws.com/bt/GRCh38_noalt_as.zip, accessed on 2022-11-17). Then, rRNA sequences were removed using SortMeRNA version 2.1b [77], and partial protein sequences were predicted using FragGeneScan version 1.18 [78]. The amino acid compositions of all protein sequences in each run were summed and used to calculate chemical metrics. Dataset accession numbers and processing statistics are listed in Table S2.

2.7. Metagenome-assembled genomes from COVID-19 patients and controls

Nucleotide sequences of metagenome-assembled genomes (MAGs) generated by [79], which are based on metagenomic data originally reported by [80] and [81] (NCBI BioProjects PRJNA650244 and PRJNA624223, respectively), were obtained from the file MAG.zip (https://figshare.com/s/a426a12b463758ed6a54, accessed on 2022-10-27). The MAGs analyzed here pass completeness and contamination thresholds of ≥50% and ≤5%, respectively [79]. Identifications of MAGs from COVID-19 patients and controls were obtained from BioSample metadata for NCBI BioProject PRJNA650244. For each MAG, protein sequences were predicted using Prodigal [82], and the amino acid compositions of all predicted proteins were summed and used to calculate chemical metrics.

2.8. Metaproteomic data processing

Protein sequences were obtained from GenBank (accessed on 2022-09-06), UniProt (accessed on 2022-08-31), or the Human Oral Microbiome Database (HOMD) [83] version 9.15 (updated date: 2022-02-07; accessed on 2023-02-03). For processing UniProt IDs, the UniProt ID mapping tool [84] was used with the taxonomy filter set to include only bacterial sequences (in order to exclude human proteins), and amino acid composition was computed from canonical protein sequences. For processing HOMD IDs, sequence files in the PROKKA directory were used (Genomes Annotated with PROKKA 1.14.6).

For the following datasets, protein sequences were obtained from the UniProt database. Amino acid compositions of proteins identified in each sample were summed and weighted by abundances (where available) to obtain the amino acid composition of the metaproteome. Starch diet (Maier *et al.* [85]): Metaproteomic abundances and UniProt IDs were obtained from https://zenodo.org/record/838741. Additional mapping to the UniParc database was used to retrieve obsolete sequences. Ulcerative colitis (Thuy-Boun *et al.* [86]): Protein IDs were extracted from PeptideEvidence fields of *.mzid.gz files from accession PXD022433 listed on ProteomeXchange [87]. IDs with "." or "_" were omitted to retain UniProt IDs.

For the saliva microbiome from Granato *et al.* [88], Majority Protein IDs and LFQ intensity for saliva cells were taken from Table S9 of the source publication; data for saliva supernatant were not analyzed here. For each protein, the first Majority Protein ID was used to look up the protein sequence in HOMD. Organism ID SEQF2791 (*Selenomonas* sp. HMT 136) was not found in HOMD at the time of this study, so the second Majority Protein ID was used in this case. For the oral microbiome from Jiang *et al.* [89], Majority Protein IDs and LFQ intensity were taken from the proteinGroups.txt file available downloaded from PXD026727. The first Majority Protein ID was used for each protein, except for some organisms not found in HOMD at the time of this study (SEQF1058, SEQF3075, SEQF1068, SEQF1063, SEQF2480, SEQF2762, SEQF3069, and SEQF2791), for which the second Majority Protein ID was used. For both of these datasets, the amino acid composition of each identified protein was multiplied by its abundance measured by LFQ intensity and summed to obtain the amino acid composition of the metaproteome.

For the COVID-19 gut metaproteome from He *et al.* [90], the file "Table S2. Global metaproteome.xlsx" was downloaded from Supplementary Files on ResearchSquare (https://doi.org/10.21203/rs.3.rs-208797/v1). UniProt IDs were extracted from the "Accession" column by keeping values starting with "tr|" or "sp|". Amino acid compositions of the proteins were multiplied by spectral counts and summed to obtain the amino acid composition of the metaproteome. For the COVID-19 gut metaproteome from Grenga *et al.* [91], GenBank protein IDs were extracted from *.mzid.gz files from accession PXD024990 listed on ProteomeXchange, and the esearch command (part of the NCBI E-utilities; https://www.ncbi.nlm.nih.gov/books/NBK25501/) with the "-organism bacteria" option was used to download bacterial protein sequences.

2.9. Oxygen tolerance of genera

The "List of Prokaryotes according to their Aerotolerant or Obligate Anaerobic Metabolism" was taken from Table S1 of Million and Raoult [92]. The list was modified in this study by the removal of *Photorhabdus*, which was listed in both categories, and the addition of obligately anaerobic genera

Anaerobutyricum, *Phocaeicola*, *Romboutsia*, and *Vescimonas*, according to their descriptions in [93–96]. With these changes, 235 obligately anaerobic and 399 aerotolerant genera are listed. Alphabetic suffixes for polyphyletic groups in GTDB [97] were removed before matching genus names to this list. Any genus listed as "variable" or "unknown" or not present in the list was considered to have unassigned oxygen tolerance.

2.10. Statistics

R [98] was used for data processing, visualization, and statistical analysis. Because they are affected by sample sizes, significance tests (*p*-values) were not used in this study. Instead, the effect size is used to measure the strength of association between chemical features and body site or disease condition. The effect size used here is Cohen's *d*, a standardized mean difference, calculated using the R package effsize [99]. The standardization ensures that metrics with different variance (nH_2O and nO_2) can be compared with each other. No predetermined threshold for small or large effects is used; larger effects are identified by comparison among the datasets.

3. Results

3.1. Correspondence between shotgun metagenomes and community reference proteomes

Samples with high levels of host DNA require special treatment for effective shotgun metagenomic sequencing of microbial communities [100]. Therefore, this study uses techniques to maximize information retrieval from 16S rRNA gene sequences, which are less affected by host DNA contamination. Specifically, 16S rRNA-based taxonomic abundances were combined with reference genomes to generate community reference proteomes. The following analysis establishes the reliability of chemical features inferred from community reference proteomes. For this purpose, we analyzed 16S rRNA sequences and shotgun metagenomes for the same samples available from the Human Microbiome Project (HMP) [101]. The selected HMP samples include 49 samples analyzed by Aßhauer *et al.* [22] together with 52 other samples analyzed by Dick and Tan [74].

The pipeline for processing shotgun metagenomes (hereafter just "metagenomes") includes a screening step to remove human DNA (see Methods for details). In order to assess the contribution of putative human DNA, the pipeline was run twice for each HMP metagenome: once with no screening, and once with screening for human DNA removal. Screening human DNA results in low protein prediction rate for some samples (Table S2). These samples exhibit relatively high scatter of chemical metrics (Figure 1a,b) probably as a result of low amounts of microbial DNA in these samples. Because of this, metagenomic sequencing runs were subject to a protein prediction rate cutoff of 40%. This value was arrived at by trial and error to remove the greatest number of outlier points while at the same time keeping at least one sample for each body site. This procedure left one sample for the nasal cavity, represented by the filled blue square in Figure 1a. A formal sensitivity analysis for this cutoff value was not performed.

Screening human DNA from the HMP metagenomes and omitting samples with low protein prediction rate yields a high correlation of chemical features between metagenomically inferred proteins

and community reference proteomes (right-hand panels in Figure 1a,b). Consistent differentiation of chemical features between body sites is evident in scatter plots of $n\mathrm{H_2O}$ vs $n\mathrm{O_2}$ (Figure 1c).

Community reference proteomes are informative about the chemical features of proteins in communities where shotgun metagenomes are challenged by high levels of human DNA. Specifically, metagenomic samples for the nasal cavity and urogenital tract have relatively high amounts of putative human DNA removed in the screening step (Figure S1), and none of the screened metagenomic samples for the nasal cavity passes the protein-prediction-rate cutoff of 40% (Figure 1c). However, community reference proteomes exhibit relatively high $n\mathrm{O_2}$ for the nasal samples. This is a plausible outcome given the oxygenated environment in the airway and corroborates a previous stoichioproteomic analysis [6].

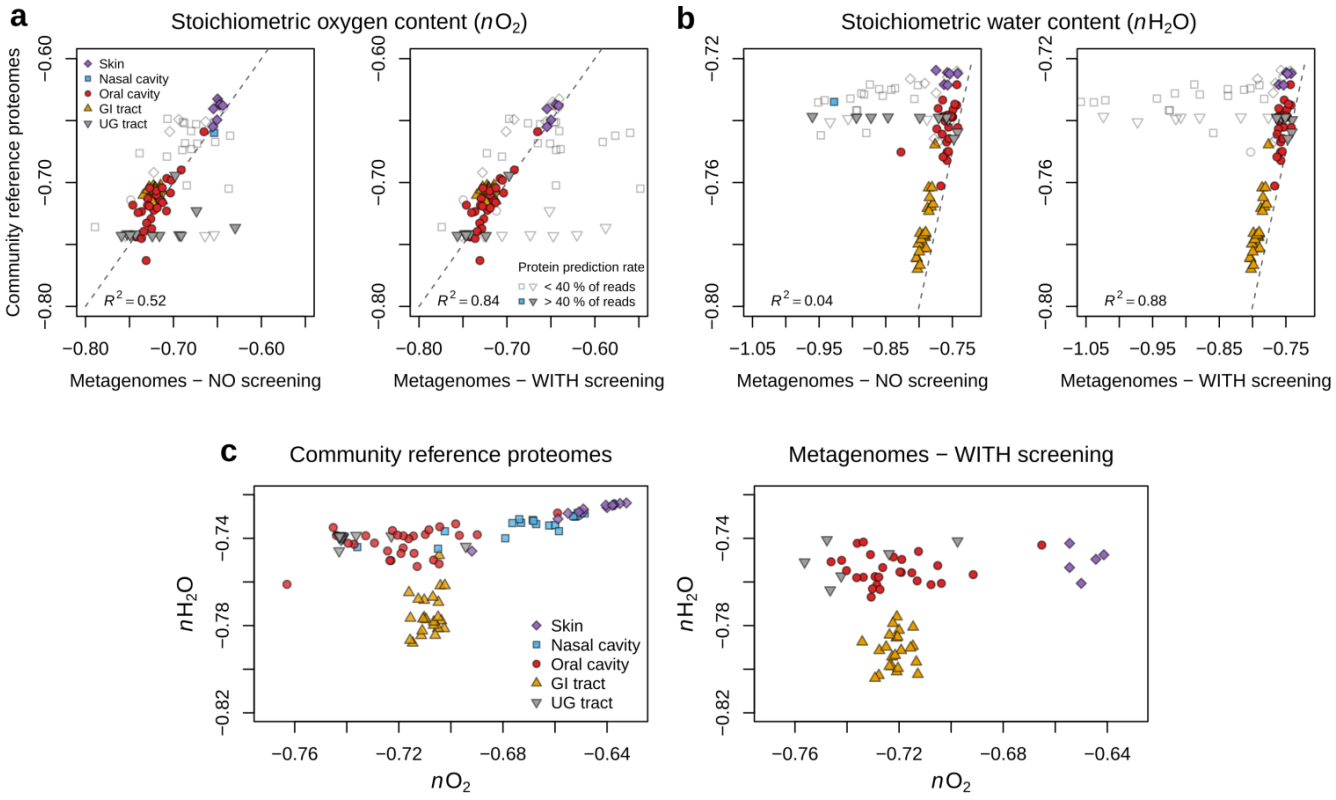

Figure 1. Chemical features of community reference proteomes compared to metagenomic protein sequences. **(a)** Oxygen content ($n\mathrm{O_2}$) and **(b)** water content ($n\mathrm{H_2O}$) for samples from the Human Microbiome Project. Chemical metrics were calculated for metagenomes without a screening step to remove human DNA sequences ((a,b) left) and with the screening step ((a,b) right). Filled symbols indicate runs for which the number of predicted protein sequences is > 40% of the number of metagenomic reads input to the sequence processing pipeline (see Table S2). Unfilled symbols represent metagenomic samples regarded to have potentially high levels of human DNA contamination and were excluded from the calculation of R^2 values. Dashed lines are 1:1 lines, not regression lines. **(c)** Scatter plots of $n\mathrm{H_2O}$ vs $n\mathrm{O_2}$ for community reference proteomes and metagenomes. Metagenomic processing for the latter plot included the screening step to remove human DNA, and only metagenomic sequencing runs with at least 40% protein prediction rate are shown. GI – gastrointestinal; UG – urogenital.

3.2. Ruling out contamination as a major issue

There is potential for contamination (e.g., chimeras) in the GTDB, which is the source of reference proteomes and the 16S rRNA training set for taxonomic classification used in this study. To rule out contamination as a major issue, reference proteomes from GTDB were compared with genomes available in a second database subject to stringent contamination checking. Reference proteomes in this study were derived from 113104 genomes in GTDB release 220, representing a total of 24959 genera. Conversely, theUnified Human Gastrointestinal Genome (UHGG v2.0.1) from MGnify [102,103] consists of 4744 high-quality (CheckM contamination <5% and completeness >50%) species-level clusters representing 1031 genera. In this version of UHGG, genomes likely to contain chimeric sequences and contigs likely to originate from the host genome were removed (https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.1/README_v2.0.1.txt, accessed on 2024-01-02). As reported below, we identified 27 genera with high relative abundance changes in gut datasets for COVID-19 and IBD (see Section 3.5). All but one of these genera is present in the UHGG; the exception is *Vescimonas*, which is an obligately anaerobic bacterium isolated from feces [96]. Therefore, despite the much higher taxonomic diversity of GTDB compared to UHGG, classification of 16S rRNA gene sequences using the GTDB-based training set successfully identifies gut-associated organisms rather than spurious taxonomic groups.

Species-level reference proteomes in GTDB form natural clusters on a plot of water *vs.* oxygen content (Figure 2a), suggesting that chemical metrics for genus reference proteomes represent genuine biological differences. Similar clusters characterize reference proteomes of taxa selected from UHGG using stringent criteria (contamination <2% and completeness >95%, including 2350 species-level clusters representing 643 genera), and chemical metrics for genus-level reference proteomes are highly correlated between GTDB and UHGG (Figure 2b,c). Furthermore, for HMP samples, there is a tight correlation between chemical metrics calculated in this study and a previous study [74], which used the Ribosomal Database Project (RDP) training set and RefSeq-based reference proteomes (Figure 2d). Therefore, we see no evidence that derived chemical features are adversely affected by genome contamination that may be present in the GTDB but not other databases.

There was no minimum number of genomes in a genus for it to be included in the generation of community reference proteomes. The variance of genomes contributing to different genera is visualized in Figure 2a. For instance, despite the large variation and overlapping ranges for species in the *Phocaeicola*, *Bacteroides*, and *Prevotella* genera, they have a lower range of $n\mathrm{H_2O}$ than other genera shown on this diagram. Conversely, ranges of $n\mathrm{O_2}$ for *Bifidobacterium* and *Corynebacterium* overlap with each other but are markedly higher than many others. The uncertainty introduced by using reference proteomes aggregated at taxonomic levels higher than species likely contributes to the scatter of points between community reference proteomes and metagenomes (Figure 1a,b) but does not obscure the big-picture differences between body sites, such as more oxidized communities in the skin and lower $n\mathrm{H_2O}$ of gut communities, both of which are also supported by metagenomic data (Figure 1c).

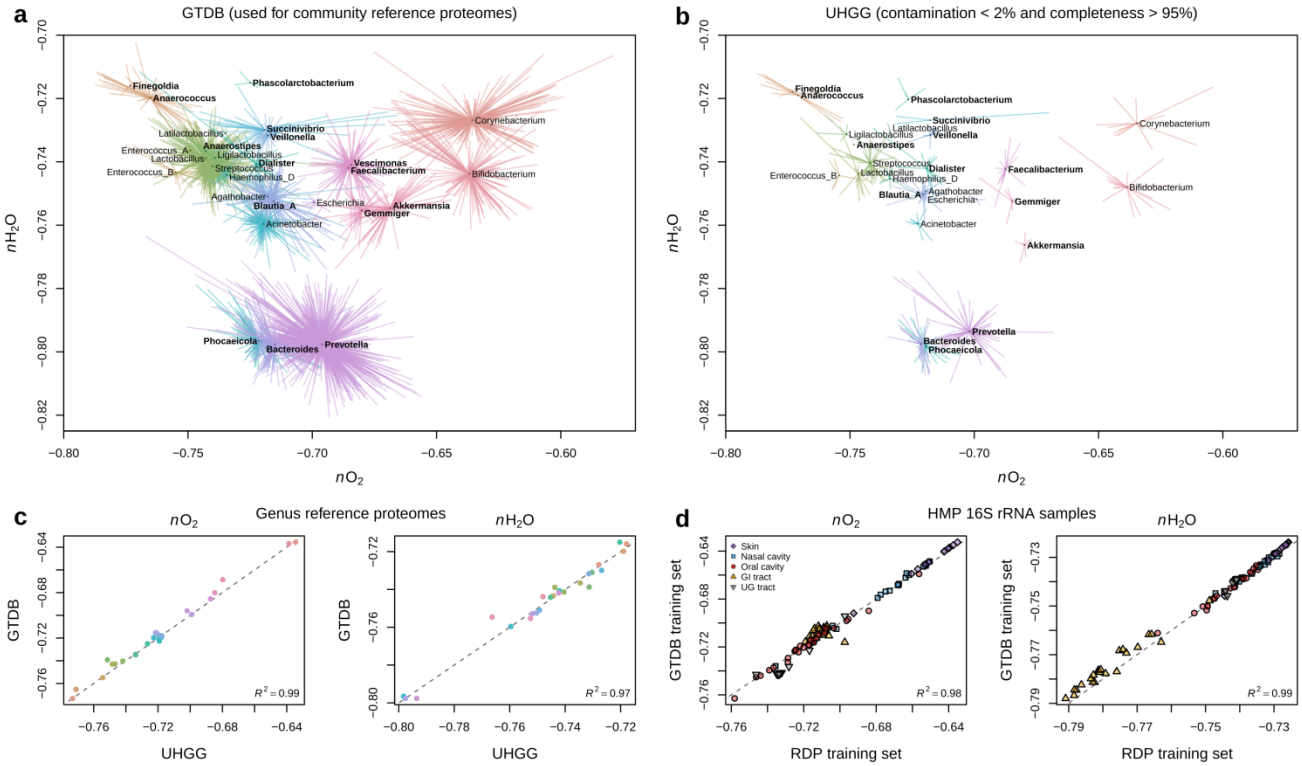

Figure 2. Chemical features for genera from different reference databases. **(a)** Starburst plot of chemical metrics of reference proteomes for selected genera and their species in GTDB. No contamination filtering was applied. The genera shown are those with a relative abundance difference of at least 5% between aggregated samples for controls and patients in one or more COVID-19 or IBD 16S rRNA datasets (see Section 3.5). The starburst patterns are drawn from a central point representing the parent taxon (genus) to each of the children (species). Bold font represents obligately anaerobic genera; **(b)** Chemical metrics of reference proteomes for the same genera in the UHGG computed from species-levels clusters with contamination <2% and completeness >95%; **(c)** Comparison of chemical metrics for genus reference proteomes in GTDB and UHGG; **(d)** Comparison of chemical metrics of community reference proteomes for HMP samples generated with the GTDB-based training set used in this study and with the RDP training set used previously by [74]. Dashed lines in (c) and (d) are 1:1 lines.

3.3. Chemical features of bacterial proteins in different body sites

The dataset of 16S rRNA gene sequences reported by Boix-Amorós *et al.* [104] represents nasal, skin, oral, and fecal (*i.e.*, gut) communities. Reference proteomes for oral and nasal communities exhibit the lowest and highest ranges of nO_2 compared to other body sites. Skin and gut communities have intermediate nO_2 while gut communities have lower nH_2O than other body sites (Figure 3a). Furthermore, Boix-Amorós *et al.* performed viral inactivation experiments by treating samples with ethanol, formaldehyde, heat, psoralen, or trizol. Treatment with ethanol, and to a lesser extent heat or trizol, results in lower nH_2O of community reference proteomes; no systematic change in nO_2 was found (Figure S2).

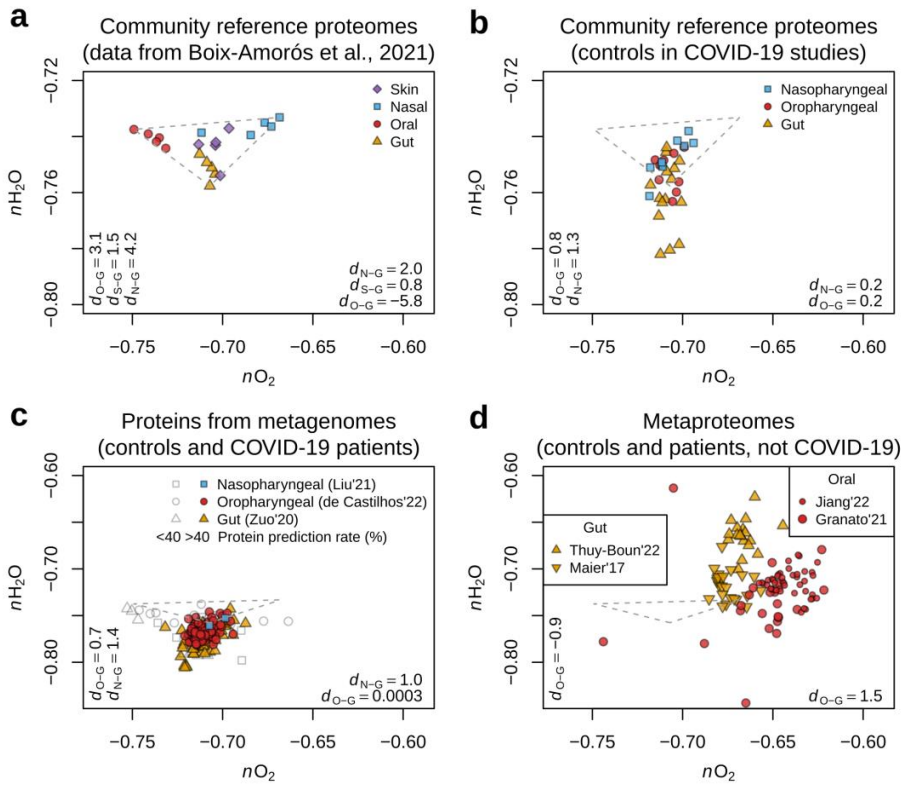

Figure 3. Multi-omics comparison of chemical features for microbiomes in human body sites. **(a)** Oxygen and water content of community reference proteomes for nasal, oral, skin, and gut sites, based on 16S rRNA gene sequences from ref. [104]; **(b)** Community reference proteomes based on 16S rRNA data for controls in COVID-19 studies (Table 1); **(c)** Proteins predicted from metagenomes for controls and COVID-19 patients: nasopharyngeal [105], oropharyngeal [106], and gut [81]; **(d)** Metaproteomes for controls and patients in non-COVID-19 studies: ulcerative colitis [86] and dietary resistant starch [85] for gut microbiomes and oral cancer [88] and lung cancer [89] for oral microbiomes. Each point represents a single sample, except for (b), where each point represents the mean of samples in a particular dataset. The dashed triangle representing the convex hull around the points in (a) is replicated in (b–d) for visual comparison. Effect sizes (Cohen's d) shown on the horizontal and vertical axes were calculated for nO_2 and nH_2O, respectively. O-G, S-G, and N-G refer to differences of oral, skin, and nasal compared to gut samples.

To find out whether the same trends are recapitulated in independent datasets, we generated community reference proteomes for control subjects in COVID-19 studies (Table 1). The trend of relatively low nH_2O in gut communities is also present in these datasets (Figure 3b). Turning to analysis of metagenomes, controls and COVID-19 patients generally exhibit lower protein nH_2O for gut compared to oral communities (Figure 3c). Finally, metaproteomes for gut communities have lower nO_2 and higher nH_2O than those from oral communities (Figure 3d). In summary, community reference proteomes and metagenomes both reveal a tendency for lower nH_2O in gut communities compared to other body sites, whereas metaproteomes of gut communities are reduced compared to oral communities.

3.4. Chemical features inferred from multi-omics data for COVID-19 and IBD

Community reference proteomes for nasopharyngeal and oropharyngeal or oral samples in COVID-19 studies exhibit a range of positive and negative mean differences of chemical metrics between controls and patients (Figure 4a). Although most datasets for oropharyngeal communities have lower mean nO_2 in COVID-19 patients compared to controls, the difference is relatively small as judged by effect size. In contrast, gut datasets on the whole exhibit lower nO_2 in COVID-19 patients. Furthermore, the most extreme points correspond to large negative differences; five gut datasets have $\Delta nO_2 < -0.005$ but none has $\Delta nO_2 > 0.005$.

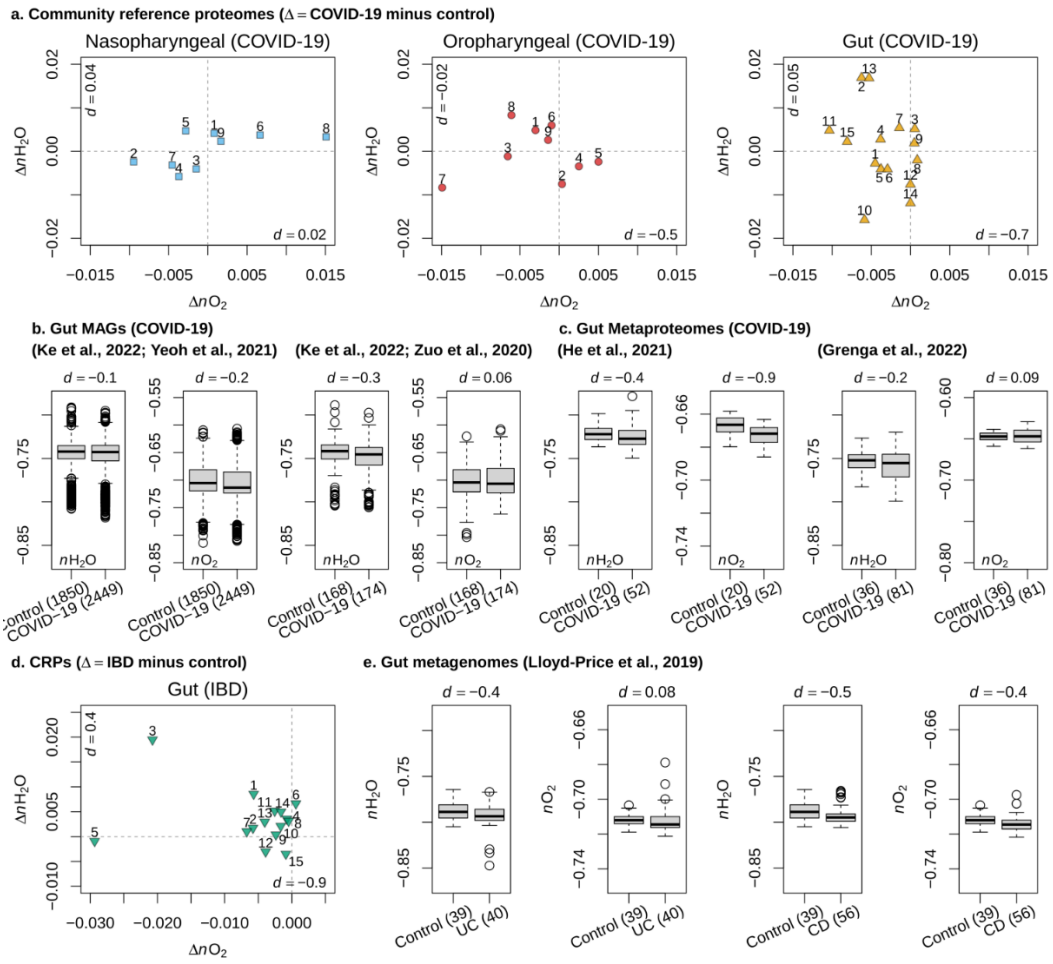

Figure 4. Differences of chemical features for microbial proteins between controls and COVID-19 or IBD patients. **(a)** Mean differences of chemical metrics for reference proteomes for nasopharyngeal, oropharyngeal, and gut communities from 16S rRNA datasets for COVID-19 studies (see Table 1). Effect sizes (Cohen's *d*) are shown for differences of oxygen and water content on the horizontal and vertical axes, respectively; **(b)** MAGs in controls and COVID-19 positive subjects reported by Ke *et al.* [79] based on metagenomic data from Yeoh *et al.* [80] and Zuo *et al.* [81]; **(c)** Bacterial metaproteomes in controls and COVID-19 positive subjects based on data from He *et al.* [90] and Grenga *et al.* [91]; **(d)** Community reference proteomes computed from the 16S rRNA datasets for IBD listed in Table 1; **(e)** Metagenomes for ulcerative colitis (UC) and Crohn's disease (CD) patients and controls from Lloyd-Price *et al.* [66]. In (b), (c), and (e), boxplots show the median (thick line), first and third quartiles (box), most extreme values within 1.5 ×the interquartile range away from the box (whiskers), and outliers (points); effect sizes (Cohen's *d*) for nO_2 or nH_2O are listed above the boxplots.

This chemical reduction trend is supported by multi-omics data. Ke *et al.* [79] reported metagenome-assembled genomes (MAGs) that are in turn based on two previous metagenomic studies. The MAGs based on data from Yeoh *et al.* [80] and Zuo *et al.* [81] are characterized by lower median nO_2 for protein sequences in COVID-19 compared to controls; however, the difference for the former set of MAGs is larger (Figure 4b). Similarly, metaproteomic data reported by He *et al.* [90] yield lower nO_2 for bacterial proteins in COVID-19 patients than controls, but the metaproteomic dataset of Grenga *et al.* [91] shows smaller differences of chemical metrics for bacterial proteins (Figure 4c).

We next analyzed data for IBD to characterize genomic adaptation to a different inflammatory condition. On the whole, community reference proteomes have lower nO_2 in IBD patients compared to controls (Figure 4d). Analysis of metagenomic data from Lloyd-Price *et al.* [66] also yields lower nO_2 of proteins in IBD compared to controls, but the difference is more pronounced for Crohn's disease than for ulcerative colitis (Figure 4e).

In summary, a large majority of 16S rRNA-based community reference proteomes point to lower oxygen content of gut microbial proteins in COVID-19 and IBD patients than in controls. Not all MAGs, metagenomes, and metaproteomes show large differences of protein nO_2, but when they do, they recapitulate the same chemical reduction trend.

3.5. Differential contributions by obligate anaerobes and facultative anaerobes

Dissecting the contributions of obligate anaerobes and aerotolerant organisms can help to understand the community-level trends of nO_2 for COVID-19 and IBD. "Obligate anaerobe" is a widely used term that hides a great deal of variation in actual oxygen tolerance. The functional definition of obligate anaerobes—organisms that do not require O_2 to grow well and whose growth is blocked by exposure to a certain level of O_2 in laboratory tests—means that many such organisms can survive or even grow at low oxygen levels [107]. For instance, some species in the *Bacteroides* genus can tolerate exposure to air for 24 h or more, yet this genus, or the *Bacteroidia* class, is commonly described as obligately anaerobic [107,108]. Using this conventional terminology, we find that reference proteomes of aerotolerant organisms have higher nO_2 than obligate anaerobes (Figure S3), indicating genomic adaptation to oxygen availability in a broad environmental context.

Differences of relative abundance of genera between aggregated samples for controls and COVID-19 or IBD patients are visualized in Figure 5. Only genera with at least a 5% increase or decrease in at least one dataset are shown. Relative abundance differences rather than fold changes are used to visualize taxa with the largest abundance differences. For example, a genus with a 2-fold change from 10% to 20% relative abundance would be shown here, but one with a 2-fold change from 1% to 2% relative abundance would not be shown. *Faecalibacterium* and *Prevotella*, both obligate anaerobes, are notable for large decreases in many datasets for COVID-19 and IBD. Interestingly, *Blautia_A* (obligate anaerobe) decreases by more than 5% and *Escherichia* (aerotolerant) increases by more than 5% in many COVID-19 datasets, in contrast to more subdued changes in most IBD datasets. The relative abundance of the aerotolerant *Streptococcus* increases by at least 5% in several datasets for both COVID-19 and IBD. *Bacteroides* and *Phocaeicola*, both classified as obligate anaerobes, exhibit variable patterns of abundance changes in IBD and COVID-19.

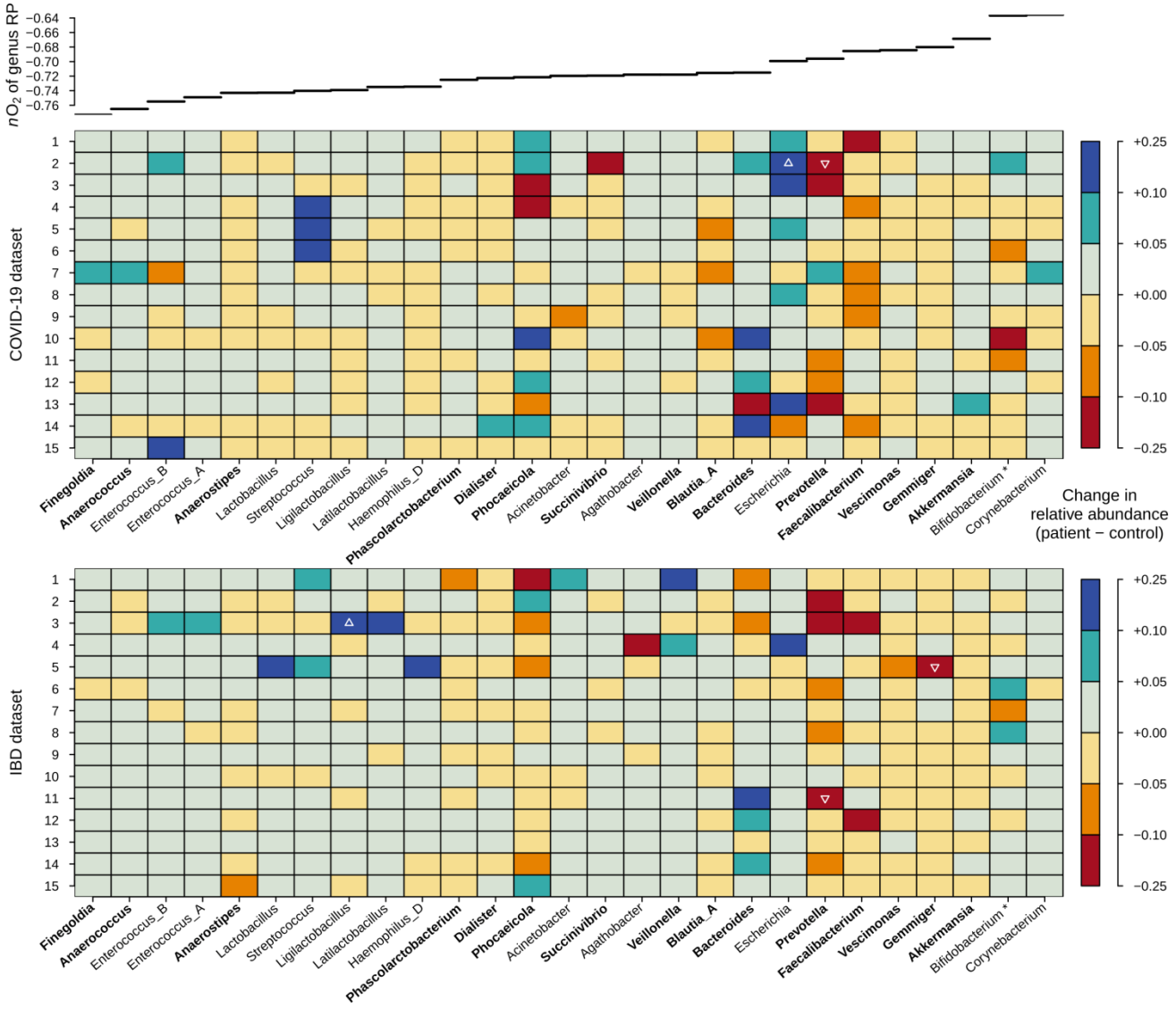

Figure 5. Differences of abundances of genera between controls and COVID-19 or IBD patients. Genus-level classifications were averaged to obtain relative abundances for control and patient groups in each dataset. Genera with a relative abundance change of at least 0.05 (*i.e.*, 5%) between controls and patients in any COVID-19 or IBD dataset are shown in the figure. Blue and red represent increased and decreased relative abundance in patients compared to controls. The most intense colors represent a maximum relative abundance difference of 25%; differences larger than this are indicated by up- or down-pointing triangles for increased or decreased abundances. The genera are ordered by increasing nO_2 of their reference proteomes, plotted at the top. Bold text indicates obligately anaerobic genera. *Bifidobacterium*, which comprises both obligately anaerobic and aerotolerant species and is classified as "variable" in the list from ref. [92], is marked with an asterisk (*).

Despite being obligate anaerobes, important gut bacteria such as *Faecalibacterium* and *Prevotella* have relatively high nO_2 of their reference proteomes, as shown by the upper line plot in Figure 5. Likewise, subcommunities of obligate anaerobes in fecal samples have higher abundance-weighted nO_2 compared to subcommunities of obligate anaerobes in other body sites (Figure 6a). This trend underlies a surprising inversion, in which subcommunities of obligate anaerobes in the gut actually have higher oxygen content of proteins than subcommunities of aerotolerant organisms (see Figure 6b for selected

datasets for COVID-19 and IBD). This inversion is more characteristic of gut communities than those in other body sites (Figure S4).

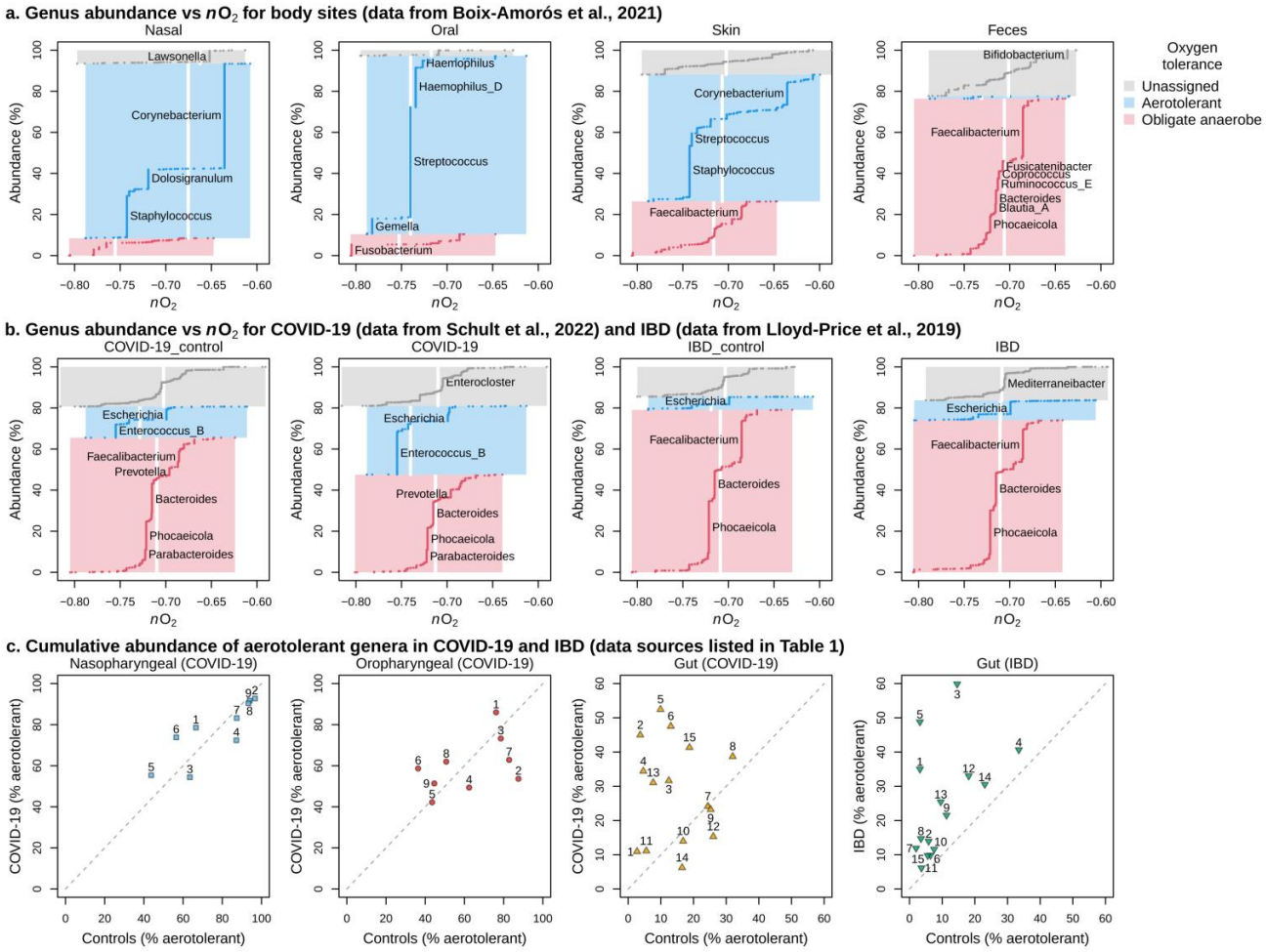

Figure 6. Higher oxygen content of proteins in anaerobic bacteria but higher abundance of aerotolerant bacteria in gut inflammatory diseases. In **(a)** for body sites and **(b)** for gut communities in selected COVID-19 and IBD studies, the length of vertical lines denotes relative abundances of genera; those that are at least 3% abundant are labeled. Within aerotolerance groups (colored boxes), lines are arranged by increasing nO_2 of the reference proteomes for genera, and the width and height of boxes represent the total range of nO_2 and cumulative abundance. Vertical white lines indicate abundance-weighted values of nO_2 for genera in each aerotolerance group. **(c)** Cumulative percent abundance of aerotolerant genera in COVID-19 and IBD studies (Table 1). Points above the diagonal dashed 1:1 line represent datasets with higher abundance of aerotolerant genera in patient compared to control groups.

Among body sites, gut microbiota are uniquely enriched in aerotolerant genera in COVID-19 or IBD patients compared to controls (Figure 6c). Despite the relatively high abundance of aerotolerant genera in patient gut samples, they do not exceed the levels of aerotolerant genera in nasopharyngeal or oropharyngeal samples, reflecting the persistence of obligate anaerobes in the gut even in inflammatory conditions. While an expansion of aerotolerant gut microbes is consistent with previous reports for inflammatory diseases [2,109], the lower nO_2 of proteins for gut communities in COVID-19 and IBD

patients compared to controls is a surprising result that evokes not only environmental factors but also competitive interactions.

4. Discussion

This study bridges multiple scientific domains, combining techniques from microbiology, genomics, and thermodynamics to provide a novel perspective on microbial adaptation to the complex chemical environments within the human body. By developing a method that translates genomic data into chemical features, we've created a new lens for understanding how microorganisms respond to environmental variation, which is important for host-associated and not only free-living communities [1,3]. This new way of measuring molecular characteristics suggests that microorganisms have adapted to our bodies by fine-tuning their protein sequences to survive in specific environments.

Our analysis quantifies the chemical aspects of bacterial proteins across different body sites and during inflammatory conditions (summarized in Table 2). Community reference proteomes for skin and nasal communities are generally highly oxidized. This pattern parallels previous analyses of metagenomic data from the HMP [6,74] and suggests that oxygenated habitats on the skin and in the airway select for genomes with relatively oxidized proteins. However, other results from this analysis are more surprising.

Table 2. Chemical metrics with relatively large effect sizes in this study. CRP—community reference proteome; MG—metagenome; MP—metaproteome.

Comparison	Difference	Data Type	Figure
Nasal vs gut	↑ $n\mathrm{O_2}$	CRP	3a
Oral vs gut	↓ $n\mathrm{O_2}$	CRP	3a
Gut vs other sites	↓ $n\mathrm{H_2O}$	CRP	3a
Gut vs other sites	↓ $n\mathrm{H_2O}$	MG	3c
Gut vs oral	↓ $n\mathrm{O_2}$	MP	3d
Ethanol treatment vs untreated	↓ $n\mathrm{H_2O}$	CRP	S2
Gut COVID-19 vs control	↓ $n\mathrm{O_2}$	CRP	4a
Gut COVID-19 vs control	↓ $n\mathrm{O_2}$	MP	4c
Gut IBD vs control	↓ $n\mathrm{O_2}$	CRP	4d

4.1. Understanding microbial adaptation to chemical variables

Our study highlights two unexpected discoveries about microbial communities. First, we discovered a distinct chemical signature for gut microbial communities, particularly in terms of water content, a metric derived from elemental composition. This finding suggests that bacteria in the intestinal environment have evolved unique genomic adaptations to cope with the gut's specific physiological conditions. The intestine has a physiological function of water absorption [12], and this appears to exert a strong selective pressure on microbial genomic evolution. Moreover, our finding of lower $n\mathrm{H_2O}$ in ethanol-treated communities is consistent with *in vitro* selection for genomes that are adapted to dehydrating environments such as that associated with ethanol treatment [110].

The second major finding concerns how bacterial communities change during inflammatory conditions like COVID-19 and inflammatory bowel disease (IBD). The chemical analysis of proteins

revealed a surprising trend: bacterial proteins become more chemically reduced during these conditions. This result is counterintuitive because inflammation is typically associated with increased oxidation, specifically higher oxygen levels and/or greater availability of other electron acceptors [1,108,111]. This result points to complex survival strategies employed by microorganisms.

During inflammation, the gut environment becomes more hospitable to aerotolerant bacteria that can survive in oxygen-rich conditions. Indeed, we found increased abundances of aerotolerant genera for patients in most COVID-19 and IBD datasets, which recapitulates previous findings for inflammatory diseases [2]. However, *Faecalibacterium*, an obligate anaerobe recognized as having anti-inflammatory associations [112], has a proteome that is more oxidized than many aerotolerant bacteria in the gut. The generally lower abundance of *Faecalibacterium* for COVID-19 and IBD patients compared to controls contributes to a trend of chemical reduction (*i.e.* lower nO_2 of proteins) at the whole-community level. Importantly, this trend is also supported by metagenomic and metaproteomic data.

These findings highlight the intricate relationship between microorganisms and their host environments and suggest that genomic evolution is more nuanced than previously understood. Chemical variables like oxygen levels and water content are not just passive background factors, but active drivers of microbial evolution.

4.2. Limitations and future directions

Several important limitations should be acknowledged. Our analysis primarily used data for fecal samples, which cannot capture the spatial complexity of the gut's oxygen and water gradients. Natural O_2 gradients within the intestine support populations of obligate anaerobes and facultative anaerobes in distinct locations [113]. Furthermore, mucous becomes denser and more continuous toward the rectum and may be associated with a longitudinal gradient of decreasing water content within the gut [13]. Therefore, future research should employ location-specific sampling to more precisely map microbial adaptations along the intestinal tract. Additionally, while we took steps to minimize potential data contamination, further rigorous screening of genomic databases and analysis of laboratory control samples would strengthen our conclusions.

The method of combining 16S rRNA-based taxonomic abundances with reference genomes to generate community reference proteomes overlooks the variations within species. Furthermore, many bacteria are novel and lack reference genomes, making it challenging to include these species in the analysis of this study. This was the reason for using shotgun metagenomes from the Human Microbiome Project (after *in silico* filtering of human DNA) to validate the inferences from community reference proteomes (Figure 1). Nevertheless, we acknowledge that omitting species-level variation could lead to potential errors in the results and even affect the conclusions.The lack of direct quantification of chemical features is a concern for this study. Because the features are derived from elemental composition, a method to analyze elemental compositions of proteins should be used. Many methods for elemental analysis suffer from the lack of quantification of hydrogen [114], which prevents calculating chemical features related to either oxidation or hydration state. It should be noted that data processing for mass spectrometry-based proteomics depends on genomic reference sequences, so that is also not a solution to the problem of direct quantification of chemical features. Alternatively, amino acid analysis of protein samples could be used to check the bioinformatics predictions made in this study. However,

methods to separate microbial and human proteins would be needed to obtain the amino acid or elemental composition of microbial proteins.

We found that metaproteomes are oxidized (Figure 3d) in comparison to community reference proteomes and metagenomically inferred proteins, which have similar ranges of chemical metrics (Figure 1). Higher natural abundances of cytoplasmic than membrane proteins, together with more efficient extraction of cytoplasmic proteins in metaproteomic experiments, can account for the high oxygen content of metaproteomes [14]. Gut metaproteomes are more reduced than oral ones, which is the opposite trend from community reference proteomes. Therefore, low oxygen concentrations in the intestinal lumen [1,115] appear to have a more pronounced effect on protein expression, as detected by metaproteomes, than on genomic adaptation. The contrasting trends reflect different timescales of evolutionary and cellular processes: metagenomes and community reference proteomes probe genomic differences that arise through long-term evolutionary adaptation, whereas metaproteomes reflect not only genomic constraints but also dynamic protein expression on shorter timescales. This suggests future research directions to track microbial protein expression in real time to reveal chemical adaptations at timescales relevant to disease progression.

5. Conclusion

Our study produces new insight on the molecular underpinnings of a well-known phenomenon: that bacterial communities are not static but dynamically responsive to their chemical environments. We used a new bioinformatic approach to examine mechanisms of molecular adaptation in microbial communities. This method bridges multiple analytical domains, combining taxonomic data, protein reference databases, and chemical analysis. This new approach reveals unexpected findings and new perspectives on microbial adaptation. First, gut communities are adapted to restricted water availability compared to other body sites. Second, inflammation in the gut is associated not only with lower abundances of obligate anaerobes (a phenomenon documented in previous studies) but also with relatively oxidized protein sequences in these anaerobes, which provides preliminary evidence that their genomes may have adapted to transiently higher oxygen levels associated with inflammation. A novel implication of this analysis is that anaerobes with highly chemically reduced proteomes may be poorly adapted to oxidative conditions in the intestine and therefore might be predicted to have stronger associations with dysbiosis.

Our findings have broader implications beyond microbiology. The ability to track chemical changes in bacterial proteins offers insights into how living systems adapt to stress and changing conditions. Populations of cells possess sophisticated mechanisms for survival in challenging environments, reshaping themselves at the molecular level through genomic variation. By understanding these intricate evolutionary mechanisms, we move closer to comprehending the complex interdependence between biomolecules and their environments that sustains life.

6. Supplementary data

The author confirms that the supplementary data are available within this article. The Supplementary Information consists of Table S1 (16S rRNA gene sequence processing statistics), Table S2 (Metagenome sequence processing statistics), and Figures S1–S4.

7. Data availability

The original contributions of this study are available in the following repositories. Training files for the RDP Classifier generated from GTDB release 220 are archived at https://doi.org/10.5281/zenodo.12703477. Analysis scripts, processed data files generated in this study, and functions to make the plots are in the "microhum" section of the JMDplots R package version 1.2.22, available at https://github.com/jedick/JMDplots and archived at https://doi.org/10.5281/zenodo.15468698.

Acknowledgments

No funding was received for this work. Declaration of generative AI in the writing process: The author drafted the entire manuscript and then prompted Claude Haiku to make the Abstract, Introduction, Discussion, and Conclusion more accessible to a general audience. The author read and corrected the output and takes full responsibility for the content of the published article.

Conflicts of interests

The author declares no conflict of interest.

Ethical statement

This study involved no human or animal research subjects.

References

[1] Pfister CA, Light SH, Bohannan B, Schmidt T, Martiny A, *et al.* Conceptual exchanges for understanding free-living and host-associated microbiomes. *mSystems* 2022, 7(1):e01374–01321.

[2] Shin N-R, Whon TW, Bae J-W. *Proteobacteria*: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.* 2015, 33(9):496–503.

[3] Lee J-Y, Tsolis RM, Bäumler AJ. The microbiome and gut homeostasis. *Science* 2022, 377(6601):eabp9960.

[4] Zong W, Friedman ES, Allu SR, Firrman J, Tu V, *et al.* Disruption of intestinal oxygen balance in acute colitis alters the gut microbiome. *Gut Microbes* 2024, 16(1):2361493.

[5] Acquisti C, Kleffe J, Collins S. Oxygen content of transmembrane proteins over macroevolutionary time scales. *Nature* 2007, 445(7123):47–52.

[6] Vecchio-Pagan B, Bewick S, Mainali K, Karig DK, Fagan WF. A stoichioproteomic analysis of samples from the Human Microbiome Project. *Front. Microbiol.* 2017, 8:1119.

[7] Moulton CR. Age and chemical development in mammals. *J. Biol. Chem.* 1923, 57(1):79–97.

[8] Toro-Ramos T, Paley C, Pi-Sunyer FX, Gallagher D. Body composition during fetal development and infancy through the age of 5 years. *Eur. J. Clin. Nutr.* 2015, 69(12):1279–1289.

[9] Dick JM. Water as a reactant in the differential expression of proteins in cancer. *Comput. Syst. Oncol.* 2021, 1(1):e1007.

[10] Munder MC, Midtvedt D, Franzmann T, Nüske E, Otto O, *et al.* A pH-driven transition of the cytoplasm from a fluid- to a solid-like state promotes entry into dormancy. *eLife* 2016, 5:e09347.

[11] Marakhova I, Yurinskaya V, Aksenov N, Zenin V, Shatrova A, *et al.* Intracellular K^+ and water content in human blood lymphocytes during transition from quiescence to proliferation. *Sci. Rep.* 2019, 9(1):16253.
[12] Leiper JB. Fate of ingested fluids: Factors affecting gastric emptying and intestinal absorption of beverages in humans. *Nutr. Rev.* 2015, 73(S2):57–72.
[13] Tropini C, Earle KA, Huang KC, Sonnenburg JL. The gut microbiome: connecting spatial organization to function. *Cell Host Microbe* 2017, 21(4):433–442.
[14] Dick JM, Meng D. Community- and genome-based evidence for a shaping influence of redox potential on bacterial protein evolution. *mSystems* 2023, 8(3):e00014–00023.
[15] Dick JM, Kang X. *chem16S*: community-level chemical metrics for exploring genomic adaptation to environments. *Bioinformatics* 2023, 39(9):btad564.
[16] Clarke ND, Taylor JS. Taxonomic distribution of opsin families inferred from UniProt Reference Proteomes and a suite of opsin-specific hidden Markov models. *Front. Ecol. Evol.* 2023, 11:1190549.
[17] Ahrens CH, Wade JT, Champion MM, Langer JD. A practical guide to small protein discovery and characterization using mass spectrometry. *J. Bacteriol.* 2022, 204(1):e00353–00321.
[18] Lee J-Y, Bays DJ, Savage HP, Bäumler AJ. The human gut microbiome in health and disease: time for a new chapter? *Infect. Immun.* 2024, 92(11):e00302–00324.
[19] Ramoneda J, Hoffert M, Stallard-Olivera E, Casamayor EO, Fierer N. Leveraging genomic information to predict environmental preferences of bacteria. *ISME J.* 2024, 18(1):wrae195.
[20] Brbić M, Warnecke T, Kriško A, Supek F. Global shifts in genome and proteome composition are very tightly coupled. *Genome Biol. Evol.* 2015, 7(6):1519–1532.
[21] Teng W, Liao B, Chen M, Shu W. Genomic legacies of ancient adaptation illuminate GC-content evolution in bacteria. *Microbiol. Spectrum* 2023, 11(1):e02145–02122.
[22] Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 2015, 31(17):2882–2884.
[23] Cruise DR. Notes on the rapid computation of chemical equilibria. *J. Phys. Chem.* 1964, 68(12):3797–3802.
[24] Dick JM, Yu M, Tan J. Uncovering chemical signatures of salinity gradients through compositional analysis of protein sequences. *Biogeosciences* 2020, 17(23):6145–6162.
[25] Dick JM. A thermodynamic model for water activity and redox potential in evolution and development. *J. Mol. Evol.* 2022, 90(2):182–199.
[26] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016, 4:e2584.
[27] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2012, 41(D1):D590–D596.
[28] Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014, 42(D1):D633–D642.
[29] Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 2022, 50(D1):D785–D794.

[30] Gevers D, Kugathasan S, Denson Lee A, Vázquez-Baeza Y, Treuren WV, *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014, 15(3):382–392.

[31] Lo Presti A, Zorzi F, Del Chierico F, Altomare A, Cocca S, *et al.* Fecal and mucosal microbiota profiling in irritable bowel syndrome and inflammatory bowel disease. *Front. Microbiol.* 2019, 10:1655.

[32] Weng YJ, Gan HY, Li X, Huang Y, Li ZC, *et al.* Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *J. Digest. Dis.* 2019, 20(9):447–459.

[33] Prasad P, Mahapatra S, Mishra R, Murmu KC, Aggarwal S, *et al.* Long-read 16S-seq reveals nasopharynx microbial dysbiosis and enrichment of *Mycobacterium* and *Mycoplasma* in COVID-19 patients: a potential source of co-infection. *Mol. Omics* 2022, 18(6):490–505.

[34] Rafiqul Islam SM, Foysal MJ, Hoque MN, Mehedi HMH, Rob MA, *et al.* Dysbiosis of oral and gut microbiomes in SARS-CoV-2 infected patients in Bangladesh: elucidating the role of opportunistic gut microbes. *Front. Med.* 2022, 9:821777.

[35] Smith N, Goncalves P, Charbit B, Grzelak L, Beretta M, *et al.* Distinct systemic and mucosal immune responses during acute SARS-CoV-2 infection. *Nat. Immunol.* 2021, 22(11):1428–1439.

[36] Iebba V, Zanotta N, Campisciano G, Zerbato V, Di Bella S, *et al.* Profiling of oral microbiota and cytokines in COVID-19 patients. *Front. Microbiol.* 2021, 12:671813.

[37] Ventero MP, Cuadrat RRC, Vidal I, Andrade BGN, Molina-Pardines C, *et al.* Nasopharyngeal microbial communities of patients infected with SARS-CoV-2 that developed COVID-19. *Front. Microbiol.* 2021, 12:637430.

[38] Gupta A, Bhanushali S, Sanap A, Shekatkar M, Kharat A, *et al.* Oral dysbiosis and its linkage with SARS-CoV-2 infection. *Microbiol. Res.* 2022, 261:127055.

[39] Crovetto F, Selma-Royo M, Crispi F, Carbonetto B, Pascal R, *et al.* Nasopharyngeal microbiota profiling of pregnant women with SARS-CoV-2 infection. *Sci. Rep.* 2022, 12(1):13404.

[40] Wu Y, Cheng X, Jiang G, Tang H, Ming S, *et al.* Altered oral and gut microbiota and its association with SARS-CoV-2 viral load in COVID-19 patients during hospitalization. *npj Biofilms Microbiomes* 2021, 7(1):61.

[41] Hernández-Terán A, Mejía-Nepomuceno F, Herrera MT, Barreto O, García E, *et al.* Dysbiosis and structural disruption of the respiratory microbiota in COVID-19 patients with severe and fatal outcomes. *Sci. Rep.* 2021, 11(1):21297.

[42] Miller EH, Annavajhala MK, Chong AM, Park H, Nobel YR, *et al.* Oral microbiome alterations and SARS-CoV-2 saliva viral load in patients with COVID-19. *Microbiol. Spectrum* 2021, 9(2):e00055–00021.

[43] Gupta A, Karyakarte R, Joshi S, Das R, Jani K, *et al.* Nasopharyngeal microbiome reveals the prevalence of opportunistic pathogens in SARS-CoV-2 infected individuals and their association with host types. *Microbes Infect.* 2022, 24(1):104880.

[44] Xu R, Lu R, Zhang T, Wu Q, Cai W, *et al.* Temporal association between human upper respiratory and gut bacterial microbiomes during the course of COVID-19 in adults. *Commun. Biol.* 2021, 4(1):240.

[45] Shilts MH, Rosas-Salazar C, Strickland BA, Kimura KS, Asad M, *et al.* Severe COVID-19 is associated with an altered upper respiratory tract microbiome. *Front. Cell. Infect. Microbiol.* 2022, 11:781968.

[46] Merenstein C, Liang G, Whiteside SA, Cobián-Güemes AG, Merlino MS, *et al.* Signatures of COVID-19 severity and immune response in the respiratory tract microbiome. *mBio* 2021, 12(4):e01777–01721.

[47] Gao M, Wang H, Luo H, Sun Y, Wang L, *et al.* Characterization of the human oropharyngeal microbiomes in SARS-CoV-2 infection and recovery patients. *Adv. Sci.* 2021, 8(20):2102785.

[48] Ren Z, Wang H, Cui G, Lu H, Wang L, *et al.* Alterations in the human oral and gut microbiomes and lipidomics in COVID-19. *Gut* 2021, 70(7):1253–1265.

[49] Hurst JH, McCumber AW, Aquino JN, Rodriguez J, Heston SM, *et al.* Age-related changes in the nasopharyngeal microbiome are associated with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and symptoms among children, adolescents, and young adults. *Clin. Infect. Dis.* 2022, 75(1):e928–e937.

[50] Zhou Y, Zhang J, Zhang D, Ma W-L, Wang X. Linking the gut microbiota to persistent symptoms in survivors of COVID-19 after discharge. *J. Microbiol.* 2021, 59(10):941–948.

[51] Zakerska-Banaszak O, Tomczak H, Gabryel M, Baturo A, Wolko L, *et al.* Dysbiosis of gut microbiota in Polish patients with ulcerative colitis: a pilot study. *Sci. Rep.* 2021, 11(1):2166.

[52] Al-Amrah H, Saadah OI, Mosli M, Annese V, Al-Hindi R, *et al.* Composition of the gut microbiota in patients with inflammatory bowel disease in Saudi Arabia: a pilot study. *Saudi J. Gastroenterol.* 2023, 29(2):102–110.

[53] Khan M, Mathew BJ, Gupta P, Garg G, Khadanga S, *et al.* Gut dysbiosis and IL-21 response in patients with severe COVID-19. *Microorganisms* 2021, 9(6):1292.

[54] Dahal RH, Kim S, Kim YK, Kim ES, Kim J. Insight into gut dysbiosis of patients with inflammatory bowel disease and ischemic colitis. *Front. Microbiol.* 2023, 14:1174832.

[55] Alam MT, Amos GCA, Murphy ARJ, Murch S, Wellington EMH, *et al.* Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathogens* 2020, 12(1):1.

[56] Chen Y, Gu S, Chen Y, Lu H, Shi D, *et al.* Six-month follow-up of gut microbiota richness in patients with COVID-19. *Gut* 2022, 71(1):222–225.

[57] Gu S, Chen Y, Wu Z, Chen Y, Gao H, *et al.* Alterations of the gut microbiota in patients with coronavirus disease 2019 or H1N1 influenza. *Clin. Infect. Dis.* 2020, 71(10):2669–2678.

[58] Mar JS, LaMere BJ, Lin DL, Levan S, Nazareth M, *et al.* Disease severity and immune activity relate to distinct interkingdom gut microbiome states in ethnically distinct ulcerative colitis patients. *mBio* 2016, 7(4):e01072–01016.

[59] Newsome RC, Gauthier J, Hernandez MC, Abraham GE, Robinson TO, *et al.* The gut microbiome of COVID-19 recovered patients returns to uninfected status in a minority-dominated United States cohort. *Gut Microbes* 2021, 13(1):1926840.

[60] Park H, Yeo S, Lee T, Han Y, Ryu CB, *et al.* Gut microbiota in inflammatory bowel disease: a combined culturomics and metagenomics perspective. *Research Square* 2023, 10.21203/rs.3.rs-3343885/v1.

[61] Mizutani T, Ishizak a A, Koga M, Ikeuchi K, Saito M, *et al.* Correlation analysis between gut microbiota alterations and the cytokine response in patients with coronavirus disease during hospitalization. *Microbiol. Spectrum* 2022, 10(2):e01689–01621.

[62] Bajer L, Kverka M, Kostovcik M, Macinga P, Dvorak J, *et al.* Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. *World J. Gastroenterol.* 2017, 23(25):4548–4558.

[63] Rausch P, Ellul S, Pisani A, Bang C, Tabone T, *et al.* Microbial dynamics in newly diagnosed and treatment naïve IBD patients in the Mediterranean. *Inflamm. Bowel Dis.* 2023, 29(7):1118–1132.

[64] Romani L, Del Chierico F, Macari G, Pane S, Ristori MV, *et al.* The relationship between pediatric gut microbiota and SARS-CoV-2 infection. *Front. Cell. Infect. Microbiol.* 2022, 12:908492.

[65] Wang Y-Z, Zhou J-G, Lu Y-M, Hu H, Xiao F-F, *et al.* Altered gut microbiota composition in children and their caregivers infected with the SARS-CoV-2 Omicron variant. *World J. Pediatr.* 2023, 19(5):478–488.

[66] Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019, 569(7758):655–662.

[67] Al Bataineh MT, Henschel A, Mousa M, Daou M, Waasia F, *et al.* Gut microbiota interplay with COVID-19 reveals links to host lipid metabolism among Middle Eastern populations. *Front. Microbiol.* 2021, 12:761067.

[68] Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* 2017, 2(5):17004.

[69] Galperine T, Choi Y, Pagani J-L, Kritikos A, Papadimitriou-Olivgeris M, *et al.* Temporal changes in fecal microbiota of patients infected with COVID-19: A longitudinal cohort. *BMC Infect. Dis.* 2023, 23(1):537.

[70] Mills RH, Dulai PS, Vázquez-Baeza Y, Sauceda C, Daniel N, *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome link *Bacteroides vulgatus* proteases with disease severity. *Nat. Microbiol.* 2022, 7(2):262–276.

[71] Ferreira-Junior AS, Borgonovi TF, De Salis LVV, Leite AZ, Dantas AS, *et al.* Detection of intestinal dysbiosis in post-COVID-19 patients one to eight months after acute disease resolution. *Int. J. Environ. Res. Public Health* 2022, 19(16):10189.

[72] Ryan FJ, Ahern AM, Fitzgerald RS, Laserna-Mendieta EJ, Power EM, *et al.* Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* 2020, 11(1):1512.

[73] Schult D, Reitmeier S, Koyumdzhieva P, Lahmer T, Middelhoff M, *et al.* Gut bacterial dysbiosis and instability is associated with the onset of complications and mortality in COVID-19. *Gut Microbes* 2022, 14(1):2031840.

[74] Dick JM, Tan J. Chemical links between redox conditions and estimated community proteomes from 16S rRNA and reference protein sequences. *Microb. Ecol.* 2023, 85(4):1338–1355.

[75] Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, *et al.* MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Briefings Bioinf.* 2017, 20(4):1151–1159.

[76] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 2012, 9(4):357–359.

[77] Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012, 28(24):3211–3217.

[78] Rho M, Tang H, Ye Y. FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010, 38(20):e191.

[79] Ke S, Weiss ST, Liu Y-Y. Dissecting the role of the human microbiome in COVID-19 via metagenome-assembled genomes. *Nat. Commun.* 2022, 13(1):5235.

[80] Yeoh YK, Zuo T, Lui GC-Y, Zhang F, Liu Q, *et al.* Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19. *Gut* 2021, 70(4):698–706.

[81] Zuo T, Zhang F, Lui GCY, Yeoh YK, Li AYL, *et al.* Alterations in gut microbiota of patients with COVID-19 during time of hospitalization. *Gastroenterology* 2020, 159(3):944–955.

[82] Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* 2010, 11(1):119.

[83] Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* 2010, 2010:baq013.

[84] Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 2011, 27(8):1190–1191.

[85] Maier TV, Lucio M, Lee LH, VerBerkmoes NC, Brislawn CJ, *et al.* Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. *mBio* 2017, 8(5):e01343–01317.

[86] Thuy-Boun PS, Wang AY, Crissien-Martinez A, Xu JH, Chatterjee S, *et al.* Quantitative metaproteomics and activity-based protein profiling of patient fecal microbiome identifies host and microbial serine-type endopeptidase activity associated with ulcerative colitis. *Mol. Cell. Proteomics* 2022, 21(3):100197.

[87] Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, *et al.* The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* 2020, 48(D1):D1145–D1152.

[88] Granato DC, Neves LX, Trino LD, Carnielli CM, Lopes AFB, *et al.* Meta-omics analysis indicates the saliva microbiome and its proteins associated with the prognosis of oral cancer patients. *Biochim. Biophys. Acta, Proteins Proteomics* 2021, 1869(8):140659.

[89] Jiang X, Zhang Y, Wang H, Wang Z, Hu S, *et al.* In-depth metaproteomics analysis of oral microbiome for lung cancer. *Research* 2022, 2022:9781578.

[90] He F, Zhang T, Xue K, Fang Z, Jiang G, *et al.* Fecal multi-omics analysis reveals diverse molecular alterations of gut ecosystem in COVID-19 patients. *Anal. Chim. Acta* 2021, 1180:338881.

[91] Grenga L, Pible O, Miotello G, Culotta K, Ruat S, *et al.* Taxonomical and functional changes in COVID-19 faecal microbiome could be related to SARS-CoV-2 faecal load. *Environ. Microbiol.* 2022, 24(9):4299–4316.

[92] Million M, Raoult D. Linking gut redox to human microbiome. *Hum. Microbiome J.* 2018, 10:27–32.

[93] Shetty SA, Zuffa S, Bui TPN, Aalvink S, Smidt H, *et al.* Reclassification of *Eubacterium hallii* as *Anaerobutyricum hallii* gen. nov., comb. nov., and description of *Anaerobutyricum soehngenii* sp. nov., a butyrate and propionate-producing bacterium from infant faeces. *Int. J. Syst. Evol. Microbiol.* 2018, 68(12):3741–3746.

[94] Wang C, Li S, Zhang Z, Yu Z, Yu L, *et al.* *Phocaeicola faecalis* sp. nov., a strictly anaerobic bacterial strain adapted to the human gut ecosystem. *Antonie van Leeuwenhoek* 2021, 114(8):1225–1235.

[95] Ricaboni D, Mailhe M, Khelaifia S, Raoult D, Million M. *Romboutsia timonensis*, a new species isolated from human gut. *New Microbes and New Infect.* 2016, 12:6–7.

[96] Kitahara M, Shigeno Y, Shime M, Matsumoto Y, Nakamura S, *et al. Vescimonas* gen. nov., *Vescimonas coprocola* sp. nov., *Vescimonas fastidiosa* sp. nov., *Pusillimonas* gen. nov. and *Pusillimonas faecalis* sp. nov. isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* 2021, 71(11):005066.

[97] Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 2018, 36(10):996–1004.

[98] R Core Team. R: A Language and Environment for Statistical Computing. *Computer software.* R Foundation for Statistical Computing. 2023, Available: https://www.R-project.org/.

[99] Torchiano M. effsize: Efficient Effect Size Computation (version 0.8.1). *Computer software.* 2020, Available: https://CRAN.R-project.org/package=effsize.

[100] Kim M, Parrish RC, Tisza MJ, Shah VS, Tran T, *et al.* Host DNA depletion on frozen human respiratory samples enables successful metagenomic sequencing for microbiome studies. *Commun. Biol.* 2024, 7(1):1590.

[101] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012, 486(7402):207–214.

[102] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 2021, 39(1):105–114.

[103] Richardson L, Allen B, Baldi G, Beracochea M, Bileschi Maxwell L, *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* 2023, 51(D1):D753–D759.

[104] Boix-Amorós A, Piras E, Bu K, Wallach D, Stapylton M, *et al.* Viral inactivation impacts microbiome estimates in a tissue-specific manner. *mSystems* 2021, 6(5):e00674–00621.

[105] Liu J, Liu S, Zhang Z, Lee X, Wu W, *et al.* Association between the nasopharyngeal microbiome and metabolome in patients with COVID-19. *Synth. Syst. Biotechnol.* 2021, 6(3):135–143.

[106] de Castilhos J, Zamir E, Hippchen T, Rohrbach R, Schmidt S, *et al.* Severe dysbiosis and specific *Haemophilus* and *Neisseria* signatures as hallmarks of the oropharyngeal microbiome in critically ill coronavirus disease 2019 (COVID-19) patients. *Clin. Infect. Dis.* 2022, 75(1):e1063–e1071.

[107] Lu Z, Imlay JA. When anaerobes encounter oxygen: mechanisms of oxygen toxicity, tolerance and defence. *Nat. Rev. Microbiol.* 2021, 19(12):774–785.

[108] Winter SE, Bäumler AJ. Gut dysbiosis: ecological causes and causative effects on human disease. *Proc. Natl. Acad. Sci.* 2023, 120(50):e2316579120.

[109] Dubourg G, Lagier J-C, Hüe S, Surenaud M, Bachar D, *et al.* Gut microbiota associated with HIV infection is significantly enriched in bacteria tolerant to oxygen. *BMJ Open Gastroenterol.* 2016, 3(1):e000080.

[110] Lai M, Lü B. Tissue preparation for microscopy and histology. *Compr. Sampling Sample Prep.* 2012, 3(3.04):53–93.

[111] Rigottier-Gois L. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *ISME J.* 2013, 7(7):1256–1261.

[112] Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermúdez-Humarán LG, *et al. Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci.* 2008, 105(43):16731–16736.

[113] Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* 2016, 14(1):20–32.

[114] Karl DM, Grabowski E. The importance of H in particulate organic matter stoichiometry, export and energy flow. *Front. Microbiol.* 2017, 8:826.

[115] Ast T, Mootha VK. Oxygen and mammalian cell culture: are we repeating the experiment of Dr. Ox? *Nat. Metab.* 2019, 1(9):858–860.