Enrichment or depletion?
The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota

Alessandro Tanca[1], Antonio Palomba[1], Salvatore Pisanu[1], Maria Filippa Addis[1] and Sergio Uzzau[1,2]

[1] Porto Conte Ricerche, Tramariglio, Alghero, Italy

[2] Dipartimento di Scienze Biomediche, Università di Sassari, Sassari, Italy`

Correspondence: Dr. Maria Filippa Addis, Porto Conte Ricerche Srl, S.P. 55 Porto Conte/Capo Caccia Km 8.400, Tramariglio, 07041

Alghero (SS), Italy

E-mail: addis@portocontericerche.it

Fax: +39-079-998-567

Additional corresponding author: Professor Sergio Uzzau, E-mail: uzzau@uniss.it

Abstract

To date, most metaproteomic studies of the gut microbiota employ stool sample pretreatment methods to enrich for microbial components. However, a specific investigation aimed at assessing if, how, and to what extent this may impact on the final taxonomic and functional results is still lacking. Here, stool replicates were either pretreated by differential centrifugation (DC) or not centrifuged. Protein extracts were then processed by filter-aided sample preparation, single-run LC, and high-resolution MS, and the metaproteomic data were compared by spectral counting. DC led to a higher number of identifications, a significantly richer microbial diversity, as well as to reduced information on the nonmicrobial components (host and food) when compared to not centrifuged. Nevertheless, dramatic differences in the relative abundance of several gut microbial taxa were also observed, including a significant change in the Firmicutes/Bacteroidetes ratio. Furthermore, some important microbial functional categories, including cell surface enzymes, membrane-associated proteins, extracellular proteins, and flagella, were significantly reduced after DC. In conclusion, this work underlines that a critical evaluation is needed when selecting the appropriate stool sample processing protocol in the context of a metaproteomic study, depending on the specific target to which the research is aimed. All MS data have been deposited in the ProteomeXchange with identifier PXD001573 (http://proteomecentral.proteomexchange.org/dataset/PXD001573).

Keywords: Differential centrifugation / Gut microbiome / Host / Metaproteomics / Microbiology / Sample preparation

Abbreviations: **DC**, differential centrifugation; **GO-BP**, GObiological process; **IU-PF**, InterPro/UniProtKB protein family; **KOG**, Kyoto Encyclopedia of Genes and Genomes orthology group; **LCA**, lowest common ancestor; **LDA**, linear discriminant analysis; **NC**, not centrifuged; **NSAF**, normalized spectral abundance factor; **TMD**, transmembrane domain

1 Introduction

The human gut harbors a complex microbial community, which is responsible for several key physiological functions of the host, including food digestion, provision of substrates to the gut epithelial cells, and immune responses [1–3]. Moreover, a growing amount of data suggests that changes in the microbiota structure and activity are tightly related to the development of metabolic dysfunctions, allergies, chronic inflammatory diseases, autoimmune disorders, and tumors [4–7]. Therefore, uncovering the taxonomic composition and functional capacity within the mammalian gut microbiota can provide fundamental information concerning host health and disease. To this extent, metaproteomics grants the unique ability to determine which functions are actually being changed within the gut microbiota depending on the host genetics or environmental factors [8].

Several papers have been published so far describing the application of the shotgun metaproteomic approach to stool samples collected from human individuals or animal models with the aim of studying the gut microbiota [9,10]. In most cases, stool samples have been subjected to enrichment methods (usually by differential centrifugation (DC) or related procedures, such as ultracentrifugation using a density gradient medium), in order to remove host cells, undigested food and other debris, and thus to enlarge the dynamic range of microbial protein identifications [11–19]. Conversely, a more conservative, "direct" procedure (i.e. not including an enrichment step) has also been used with success in very few cases [20]. Nevertheless, a specific investigation aimed at elucidating if, how, and to what extent sample enrichment steps may impact on the final outcome is still lacking.

Here, we evaluated the influence exerted by the DC of stool on human gut metaproteomic profiling, using a non-centrifuged, directly extracted, sample as a control. Overall performance, technical reproducibility, as well as taxonomic and functional distribution of the identified proteins were investigated. The consequences of sample pretreatment on information concerning microbiota and host proteomes are discussed.

2 Materials and methods

2.1 Stool sample

The human feces used for this study were provided by a healthy volunteer who gave consent to their use for research purposes. Feces were split into ten samples (as illustrated in Fig. 1, top): five (average wet weight 337 mg) were directly subjected to protein extraction, while the remaining five (average wet weight 1191 mg) underwent DC (see below).

2.2 Differential centrifugation

Stool samples were subjected to DC to enrich for microbial cells, according to Verberkmoes et al. [11] and Tanca et al. [21], with minor modifications (see illustration in Fig. 1, bottom). Briefly, samples were resuspended in PBS to reach a final volume of 50 mL, vortexed, shaken in a tube rotator for 45 min, and subjected to low-speed centrifugation at 500 × *g* for 5 min aimed to eliminate particulate and insoluble material. The supernatants were then carefully transferred to a clean polyallomer centrifuge bottle (Beckman Coulter, Brea, CA, USA) and kept at 4C, whereas the pellets were suspended again in PBS. The entire procedure was repeated for a total of three rounds. Finally, the supernatants (one per round, therefore three per sample) were centrifuged at 20 000 × *g* for15min, and the derivative pellets were subjected to protein extraction following the protocol described below.

2.3 Protein extraction, digestion, and quantification

Samples were resuspended by vortexing in extraction buffer (2% SDS, 100 mM DTT, 20 mM Tris-HCl pH 8.8) preheated at 95C. Specifically, a 1:2 (mg/L) sample-to-buffer ratio was used for the stool samples subjected to direct extraction, whereas the three microbial pellets per sample obtained upon DC were first resuspended in the extraction buffer (1:1 ratio) and then pooled, in order to obtain a single tube per sample. Samples were then heated and subjected to a combination of bead-beating and freeze-thawing steps as detailed elsewhere [21]. The protein extract concentration was estimated by whole lane densitometry using QuantityOne software (Bio-Rad, Hercules, CA, USA) after electrophoretic separation through an Any kD Mini-PROTEAN TGX Gel (BioRad) and gel staining with SimplyBlue SafeStain (Invitrogen, Carlsbad, CA, USA).

Protein extracts were subjected to on-filter reduction, alkylation, and trypsin digestion according to the filter-aided sample preparation protocol [22], with slight modifications detailed elsewhere [23] and using Amicon Ultra-0.5 centrifugal filter units with Ultracel-10 membrane (Millipore, Billerica, MA, USA). Peptide mixtures concentration was estimated by measuring absorbance at 280nm with a NanoDrop2000 spectrophotometer (Thermo Scientific, San Jose, CA, USA), using dilutions of the MassPREP *E. coli* Digest Standard (Waters, Milford, MA, USA) to generate a calibration curve.

2.4 LC-MS/MS analysis

LC-MS/MS analysis was carried out using an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific) interfaced with an UltiMate 3000 RSLCnano LC system (Thermo Scientific). The single-run 1D LC peptide separation was performed as previously described [21,24], loading 4 g of peptide mixture per each sample, and the mass spectrometer was set up in a data-dependent MS/MS mode, with Higher Energy Collision Dissociation as the fragmentation method, as illustrated elsewhere [23].

2.5 Data analysis

Peptide identification was performed using Proteome Discoverer (version 1.4.1; Thermo Scientific), with a workflow consisting of the following nodes (and respective parameters): Spectrum Selector for spectra preprocessing (precursor mass range: 350–5000 Da; S/N threshold: 1.5), SEQUEST-HT as search engine (protein database: see below; enzyme: trypsin; maximum missed cleavage sites: 2; peptide length range 5–50 amino acids; maximum delta Cn: 0.05; precursor mass tolerance: 10 ppm; fragment mass tolerance: 0.02 Da; static modification: cysteine carbamidomethylation; dynamic modification: methionine oxidation), and percolator for peptide validation (false discovery rate < 1% based on peptide *q*-value). Results were filtered in order to keep only rank 1 peptides, and protein grouping was allowed according to the maximum parsimony principle.

The protein database was generated based on taxonomic information following an iterative approach, as proposed in a recent paper from our group [25]. Specifically, a preliminary search was performed against the complete UniProtKB database (release 2013_12). Then, the peptide sequences identified in all the samples through the preliminary search were uploaded into the Unipept web application (v.2.4, http://unipept.ugent.be) [26] to carry out a taxonomic assignment based on the lowest common ancestor (LCA) approach. In keeping with this, sequences from 298 detected microbial genera (from Archaea, Bacteria, and Fungi; see Supporting Information 1 for details) retrieved from UniProtKB (release 2013_12) were appended to the *Homo sapiens* sequences retrieved from Swiss-Prot (release 2013_12) in order to generate a customized "host-microbiome" database containing sequences from specific microbial taxa and the hos t(5990075 protein sequences in total). Furthermore, an additional search was carried out using a "food" database containing all UniProtKB sequences belonging to the six most abundant plant genera detected in the preliminary search (namely, *Arachis*, *Musa*, *Corylus*, *Theobroma*, *Glycine,* and *Pisum*; 117 047 total protein sequences), and the results were merged to those obtained with the "host-microbiome" database.

The normalized spectral abundance factor (NSAF) was calculated as described elsewhere [24, 27], and used in order to estimate peptide abundance. The relative abundance of a feature (protein, taxon, functional category, or combined taxonomic-functional feature) was calculated by summing the NSAF values of all peptides matched to that given feature. The NSAF log ratio was calculated as previously described [28] using 2 as correction factor, and employed to estimate the extent of differential abundance between the two pretreatment methods being compared. Statistical significance of differential expression was assessed by applying a *t*-test on logarithmic NSAF values, after replacing missing values with 0.01 (empirically determined as in [27]).

Reproducibility among replicates was measured according to the Pearson correlation coefficient (*r*), as described elsewhere [29]. Pearson correlation coefficient was calculated by plotting the NSAF values measured for each feature in two different replicates of the same method, and then by calculating the mean values among all possible replicate combinations.

Alpha-diversity indexes were calculated according to established methods [30]. InterPro protein families [31] were retrieved from UniProtKB [32]. Kyoto Encyclopedia of Genes and Genomes orthology groups (KOGs) information [33] was gathered using KOBAS (http://kobas.cbi.pku.edu.cn/home.do) [34]. LEfSe was used for linear discriminant analysis (LDA) and generation of cladograms (http://huttenhower.sph.harvard.edu/galaxy/root) [35], considering features with log LDA score > 2 and alpha-value < 0.05 as differentially abundant between sample groups. Protein subcellular localization was predicted using Psort (v.3.0.2, http://www.psort.org/psortb/index.html) [36]. The number of transmembrane domains (TMDs) within protein sequences was predicted using the TMHMM Server (v.2.0, http://www.cbs.dtu.dk/services/TMHMM) [37]. Data were parsed using in-house scripts, and graphs were generated using Microsoft Excel.

The MS proteomics data in this paper have been deposited in the ProteomeXchange Consortium (http:// proteomecentral.proteomexchange.org) via the PRIDE partner repository [38,39] with the dataset identifier PXD001573.

3 Results and discussion

3.1 Overall comparison of performance, reproducibility, and information depth

In quantitative terms, the samples processed without DC (not centrifuged (NC)) gave a mean protein extraction yield estimated in 26 ± 3 g of proteins per milligram of feces, versus 7 ± 0.5 g of proteins per milligram of feces for the samples enriched by DC. In the latter case, the lower protein yield should be due to the fact that most of the proteins contained in the insoluble debris (produced after the first 500 × *g* centrifugation step) and in the final supernatant (produced after the three sequential 20 000 × *g* centrifugation rounds) are removed from the sample, and only the final microbial pellet is subjected to protein extraction. Nevertheless, this issue might be relevant only when limited amounts of sample should be available, which is usually not a problem when dealing with human samples.

In order to compare the two procedures in qualitative terms, which was the purpose of this work, the same amount of peptide mixture was loaded in the LC column for MS analysis for each sample and condition. As a result, a total of 10 536 and 12 418 nonredundant peptides were identified in the NC and DC samples, respectively(18%increaseinDC;histogram in Fig. 2A); similar increments were obtained when considering the number of proteins (3911 for NC vs 4587 for DC, Supporting Information Fig. 1) or peptide-spectrum matches (69 218 for NC vs 81 145 for DC, Supporting Information Fig. 2). Moreover, the percentage of MS/MS spectra reliably matched with peptide sequences was also higher in DC (Supporting Information Fig. 3). Therefore, the DC protocol produces a general increase in the number of identifications.

The reproducibility of the two pretreatment methods was also evaluated using the Pearson correlation coefficient(*r*) as a measure of quantitative reproducibility among replicates. NC and DC exhibited almost identical *r* values (0.79 for peptides and 0.93 for proteins). Run repeatability in similar experimental conditions was measured and described previously (0.87 for peptides and 0.97 for proteins) [21].

An LCA approach was used to assign peptide sequences to specific microbial and nonmicrobial taxa. Accordingly, features unambiguously assigned to a microbial (super)kingdom (Archaea, Bacteria, Fungi) were considered as “microbial,” whereas the other eukaryotic sequences assigned to the phyla

Streptophyta (vegetables) or Chordata (host cells and meat) were considered as "nonmicrobial." According to the taxonomic classification, the percentage of microbial peptides out of the total was measured as 80% for NC versus 89% for DC (8401 vs 11 114 in absolute terms, respectively; histogram in Fig. 2A). Conversely, the number of nonmicrobial peptides was over twofold higher in NC than in DC (1458 vs 652, corresponding to 13.8 and 5.3% of the total, respectively; Supporting Information Fig. 4); among them, peptide sequences of plant (food) origin were over fourfold higher in NC compared to DC (438 vs 104, respectively), while those from the host were about 1.5-fold higher in NC compared to DC (461 vs 302, respectively). These results are consistent with the DC protocol aim of enriching the microbial component by removing host cells, undigested food, fibers, mucus, nonsoluble host proteins, and complexes in the first rounds of low-speed centrifugation. In keeping with this, the increase seen in the number of identified peptides for the DC protocol is likely due to an enrichment in the microbial component versus other host and food proteins, while the increase in the matched spectra is probably dependent on the reduction of interfering nonprotein molecules. Nevertheless, a selective increase in DC of microbial species with better annotated genomic databases might also be a contributing factor, together with the concurrent depletion of proteins from heterogeneous and minor proteinaceous sources from the diet.

The cumulative number of microbial and nonmicrobial peptides detected in five replicate analyses was then calculated, along with the number of identifications common to all replicates (core; line graphs in Fig. 2B and C). Microbial identifications (both "core" and cumulative values) were clearly higher in DC when compared to NC, whereas nonmicrobial identifications followed the opposite trend. It is interesting to notice that by doubling the number of NC replicates analyzed, it is possible to reach a number of microbial identifications similar to that of a single DC replicate; on the contrary, the number of nonmicrobial identifications obtained with five DC replicates does not reach that achieved with a single NC replicate. Nevertheless, at equal numbers of identified proteins the informative content in terms of microbial diversity (i.e. taxonomy and functions) might still be different for the two approaches (see below). The distribution of the core peptides between NC and DC dataset is shown in the Venn diagrams in Fig. 2B (microbial) and C (nonmicrobial). Over 1200 microbial peptides were consistently detected along all replicates with both methods, while 74 and 190 microbial peptides were unique (i.e. found in all replicates of a method and completely undetected with the other method) to NC and DC. Concerning nonmicrobial peptides, NC provided a dramatically higher contribution in terms of unique peptides when compared to DC (160 vs 4, respectively).

3.2 Taxonomic distribution of the gut metaproteome

Almost 80% of the overall microbial peptide sequences were assigned according to an LCA approach to a specific phylum and slightly more than a half to a specific family. Moreover, when comparing NC and DC datasets, no significant variations could be found in the relative amount of microbial peptide sequences assigned to a specific taxon (from the phylum to the genus level), while a slight but statistically significant difference was seen at the species level ($p < 0.05$; Supporting Information Fig. 5).

The "metaproteomic alpha-diversity" was also measured, according to the Simpson and the Shannon–Wiener indexes and using taxonomic family abundances as input data. In both cases, DC showed a much higher diversity when compared to NC ($p < 0.0001$; Supporting Information Table 1).

NC and DC results were also compared based on the relative abundance of the main phyla, according to metaproteomic NSAF data (Fig. 3A). Statistically significant differences were observed for all microbial phyla with an abundance higher than 0.1%. In particular, a marked change in the Firmicutes/Bacteroidetes ratio (1.2 for NC vs 2 for DC) was seen, along with a general increase in the relative abundance of the main phyla in DC when compared to NC (e.g. a twofold increment for Actinobacteria). At the broadest taxonomic level, Firmicutes and Bacteroidetes dominate the gut microbiota in humans and other animals. A lower abundance in Firmicutes has been observed to match with a corresponding increase in Bacteroidetes and vice versa. These phyla include the most abundant variety of bacterial species colonizing the intestine, and a change in their relative abundance has been correlated with a number of metabolic and immunological disorders [40–42]. Therefore, the ability of a method to reliably assess the Firmicutes/Bacteroidetes ratio is crucial, and it should be given careful consideration.

Figure 3B shows the comparison carried out at the family level. Many of the main Firmicutes families were significantly enriched in DC, apart from Clostridiaceae (same percentage as in NC) and Oscillospiraceae (significantly higher in NC); conversely, within Bacteroidetes, Bacteroidaceae were much higher in NC, whereas Prevotellaceae exhibited the opposite trend. Of note, among families belonging to the less-abundant phyla, Desulfovibrionaceae, Bifidobacteriaceae, and Sutterellaceae were enriched almost seven-, three-, and twofold in DC when compared to NC. Therefore, despite the higher alpha-diversity recorded in DC samples, possibly due to a more efficient extraction of microbial proteins when undigested food and host components are depleted, species belonging to Bacteroidaceae and Oscillospiraceae might partition preferentially to the "debris" pellet (see Fig. 1). The cumulative and "core" number of taxonomic families identified in five replicate analyses are given in Supporting Information Fig. 6. Interestingly, no taxonomic families were found to be present in all DC replicates and in none of NC replicates, and vice versa.

Differential NSAF abundances of taxonomic data were also assessed by carrying out an LDA using LEfSe to determine the effect size and to account for the hierarchical structure of the taxonomic ranks. The cladogram in Fig. 4 depicts the hierarchical relationships between the taxa identified in this study; taxa significantly varying between NC and DC (log LDA score > 2 and alpha-value < 0.05) are presented in color. As apparent from the image, each pretreatment method presents differential trends consistently covering the entire taxonomy tree; examples of "class-to-genus axes" are Bacteroidia-*Bacteroides* (represented by many different species) significantly higher in NC, as well as, among those enriched in DC, Clostridia-*Faecalibacterium*, Actinobacteria-*Bifidobacterium,* Betaproteobacteria-*Sutterella,* Deltaproteobacteria-*Desulfovibrio,* and Methanobacteria-*Methanobrevibacter*. Interestingly, however, *Prevotella* (Bacteroidia) and *Ruminococcus* (Clostridia) abundances follow an opposite trend with respect to the related taxa of the same class. To this extent, the most abundant *Ruminococcus* species detected in this study, *R. bromii*, has been previously recognized as specialized in degrading cellulose and in binding tightly and directly to insoluble starch particles in fecal samples [43]. Thus, as considered above, the differential depletion of the Clostridia member *R. bromii* (log LDA score > 3 and alpha-value < 0.01) in DC might depend on its differential substrate colonization in respect to other members of this class, resulting in a preferential localization into the discarded pellet. In addition, while most of the identified species of *Bacteroides* appear to be markedly depleted in DC, two of them (namely, *B. massiliensis* and *B. cellulosilyticus*) show a higher abundance in DC (log LDA score > 3 and alpha-value < 0.01 for both). A possible explanation for such "species-dependent" enrichment/depletion of *Bacteroides* in DC samples might be provided by their species-specific colonization "geography" within the host gut environment [44]. Additional information concerning differentially abundant microbial genera and species is provided in Supporting Information Figs. 7 and 8.

3.3 Functional features of the gut metaproteome

In order to infer functional information on the gut metaproteome, each identified protein was classified according to three different annotations: GO-biological process (GO-BP), InterPro/UniProtKB protein family (IU-PF), KOGs. Diverse annotation methods were employed since no consensus exists on the best functional annotation approach for microbiome analysis. Moreover, the investigation of complementary levels of annotation can contribute to enlarge the information depth of a metaproteomic study, especially considering that databases used for peptide identification usually contain many poorly annotated sequences [45,46].

According to the GO-BP classification, a total of 640 and 730 microbial biological process categories were found in NC and DC, respectively (Supporting Information Fig. 9). Figure 5A illustrates the most abundant GO-BP categories. Among those with abundance >1%, comparable percentages could be observed in most cases for NC and DC, with slight but significant differences, for instance, for translation and transporter activity (higher in NC), as well as for glycolytic process and kinase activity (higher in DC), among others. In addition, 23 categories were found to be significantly differential between NC and DC (log ratio >1 and *p*-value < 0.01; Supporting Information Fig. 10), including proteins related to cell replication and biosynthesis (higher in DC), as well as to pathogenesis and substrate degradation activities (higher in NC, comprising sialidases, collagenases, endopeptidases, and other proteins involved in nutrient degradation).

Taking into account the known functional redundancy among even unrelated taxa [44], these GO-BP functional categories were combined with taxonomic information, in order to assess the taxa-specific contribution and thus to verify whether any of the GO-BP trends was independent from the taxonomic trends described in the previous paragraph. As shown in Fig. 5B, for each of the top 12 GO-BP categories, the abundance values corresponding to the two main phyla (Firmicutes and Bacteroidetes) were investigated. As a result, proteins assigned to Firmicutes and related to flagellum-dependent motility, amino acid metabolism, and polysaccharide catabolism were higher in NC, in clear contrast with the above-mentioned taxonomic trend. In the other cases, the differences in abundance are quite consistent with the taxonomic trend. Supporting Information Figs. 11–13 report the most abundant and differential features combining GO-BP information with phylum/family taxonomic assignment.

According to the IU-PF classification, a total of 299 and 338 microbial protein families were found in NC and DC, respectively (Supporting Information Fig. 14). Supporting Information Figs. 15 and 16 illustrate the most abundant protein families, and those significantly differential between NC and DC, respectively, while Supporting Information Figs. 17–20 show the data concerning the IU-PF functional classes combined with taxonomic assignments. This analysis revealed that the Bacteroidetes TonB-dependent receptor family was significantly higher in NC, and that the Firmicutes (namely, Clostridiaceae) Peptidase S8 was not detectable in DC; conversely, examples of protein families enriched in or unique to DC were histone-like proteins from Firmicutes and sulfate adenylyltransferase from Desulfovibrionaceae, respectively.

According to the KOG classification, a total of 598 and 687 microbial protein families were found in NC and DC, respectively (Supporting Information Fig. 21). Supporting Information Figs. 22 and 23 illustrate the most abundant KOGs, and those significantly differential between NC and DC, respectively, while Supporting Information Figs. 24–27 refer to the combined functional-taxonomic classification. Firmicutes flagellins and glutamate dehydrogenases were significantly higher in NC, in opposition to the general taxonomic behavior. Furthermore, lactocepins and pullulanases (both cell surface-associated enzymes, with possible biotechnological applications) from various families belonging to Firmicutes and Actinobacteria were dramatically depleted in DC, once again in spite of the global taxonomic trend. We chose to employ in parallel both IU-PF and KOG annotations since we observed that some of the main functional categories found with the former were completely absent in the latter (e.g. TonB-dependent receptor), and vice versa (e.g. flagellin), as clearly evident when comparing Supporting Information Figs. 15 and 22).

One of the most striking observations that emerged when comparing the two methods, DC and NC, was the significant change in the Firmicutes/Bacteroidetes ratio, mostly due to the marked reduction of Bacteroidaceae in DC. In parallel, the functional classes that were significantly more depleted in DC were those associated with hydrolase and endopeptidase activities (Supporting Information Fig. 11), accounted for by enzymes that are mainly devoted to degradation of (food) carbohydrates and proteins, respectively. In addition, the most abundant taxonomic-functional class in the whole microbiota was transporter activity associated with the family Bacteroidaceae (Supporting Information Fig. 12). The phylum Bacteroidetes is known for its role in degradation of undigested food residues, mainly represented by dietary glycans. As a further observation, the protein identities assigned to food components were drastically reduced by the DC treatment [47]. When considering all these results, it can be hypothesized that food-degrading functions might undergo a selective depletion in DC, being eliminated together with the undigested food residues in the course of the first centrifugations aimed to remove insoluble debris from the fecal material, and thus leading to the observed variation in the Firmicutes/Bacteroidetes ratio. In addition, it is known that insoluble substrates are colonized by different subsets of fecal bacteria [43,48]; their removal may therefore lead to the introduction of biases depending on the specific composition of the stool sample under examination.

When considering the structural complexity of the bacterial cell and the different protein localization compartments (cytosolic, membrane associated, supramolecular cell-surface associated, or secreted in the extracellular *milieu*), the effect of DC on the final proteomic profile outcome deserves further specific considerations. Based on localization prediction carried out using Psort (Supporting Information Fig. 28), the DC dataset was found to be slightly enriched in cytoplasmic proteins, as well as depleted of membrane and, to a higher extent (over 30% reduction), extracellular/secreted proteins when compared to the NC dataset (significance $p < 10^{-4}$). Moreover, the investigation of microbial proteins containing one or more TMDs revealed that their presence is significantly higher ($p < 10^{-4}$) in NC samples when compared to DC samples. When considering the abundance distribution of TMD-containing proteins among bacterial phyla, differences were observed in DC vs NC for Firmicutes (6 vs 9%, $p < 10^{-5}$), and Actinobacteria (4 vs 28%, $p < 10^{-5}$), while TMDs from Bacteroidetes (11 vs 11%) and Proteobacteria (13 vs 11%) seemed unaffected by the DC protocol. An important conclusion can be drawn from these observations. When applying the DC protocol, there is the risk for the sample to undergo a selective depletion not only in taxonomic terms (i.e. a general depletion in Bacteroidetes, as noted above), but also in structural terms, as observed here for extracellular/secreted and membrane proteins. In fact, adding to the expected loss of highly soluble, secreted proteins due to removal of the final supernatant, other physico-chemical features may favor a differential partitioning of the proteins along the centrifugation steps. For instance, highly hydrophobic or "sticky" proteins that remain attached to the solid surfaces offered by sloughed cells and undigested food, such as bacterial adhesins or enzymes with substrate-binding and degradation functions, would be removed when eliminating the debris in the first steps of the DC protocol. For example, in the case of Actinobacteria (and especially of *Bifidobacterium*), membrane proteins depleted by the DC treatment were mainly pullulanase and subtilisin-like serine protease, which both have a transmembrane anchorage and a surface-exposed catalytic portion. In addition, a contribution to this bias would be provided also by residual cell wall and membrane fragments from dead bacterial cells. Likewise, in clear contrast with the above-mentioned taxonomic trend, Firmicutes flagellins were depleted in DC. As a further consideration, there would also be the possibility of separating bacteria that are actively expressing particular subsets of proteins from those that are not expressing them. Finally, we cannot rule out that proteolytic events or slight changes in protein expression in living microbial cells may occur during the DC process. Therefore, careful scrutiny of this scenario should be given when selecting a sample pretreatment strategy for proteomic characterization of the microbiota.

3.4 Functional features of the host proteome

The main aim pursued when employing a DC pretreatment is the removal of host proteins, which are usually considered as contaminants. However, when investigating gut metaproteome changes related to specific physiological or pathological conditions, preservation of host proteome information may be useful to shed light on the concurrent modifications occurring in the gut environment (e.g. intestinal immune response, cell junctions, mucus layer).

Therefore, in order to investigate qualitative and quantitative differences between the host information achieved using NC and DC protocols, peptide sequences unambiguously assigned to the order Primates (and thus distinguished from food peptides of other mammalian origin, i.e. from meat) were selected for further analysis concerning the host proteome (peptide identification statistics are shown in

Supporting Information Fig. 29). As done for the microbial proteome, host proteins were classified according to GO-BP, IU-PF, and KOG annotations, and relative abundances of all functional categories were comparatively assessed for NC and DC (Supporting Information Figs. 30–32). The main results can be summarized as follows: (1) human glycosyl hydrolases were found as significantly more abundant in NC, as already observed for the microbial enzymatic counterpart; (2) human serine endopeptidase inhibitors (serpins) were higher in NC consistently with the higher abundance in NC of the microbial serine endopeptidases; (3) several proteins related to functions of considerable biological importance in the gut (including some specific members of MHC classes I and II, mucin, antitrypsin, antichymotrypsin e peptidase families) were significantly depleted in the DC dataset; (4) elastase and phospholipase A2 were among protein functions relatively enriched in DC. Taken together, these data highlight, as expected, that the NC protocol maybe preferable for studies that aim to gather microbiome data along with corresponding host information.

4 Concluding remarks

The results presented in this work highlight pros and cons of the stool pretreatment based on DC with regard to the metaproteomic analysis of the human gut microbiome. Among the advantages, samples processed by DC generally achieve a higher number of protein/peptide identifications, with a significantly higher microbial diversity. This is undoubtedly of key importance when conducting a study aimed at assessing subtle changes in the gut microbiota, as well as at identifying very low abundance enzymes involved in specific microbial pathways. However, the elimination of particulate matter, such as food and mucous residues, heavily colonized by specific assortments of microbial taxa, appears to introduce a clear bias toward "free roaming" microbial cells. In addition to taxonomy, functional and structural information is also affected, when considering the depletion observed for specific functional categories, such as flagella or cell surface anchored enzymes. As a further observation, information on the nonmicrobial counterpart (host- and food-derived proteins) is dramatically reduced when applying the DC protocol. Finally, it is also worth noting that the DC procedure is considerably more labor intensive and time-consuming, and that it may be more influenced than NC by the wide variability in feces texture, fiber, and water content. In conclusion, this work clearly underlines that a critical evaluation needs to be made prior to selecting how to process stool samples in the context of a metaproteomic study.

Acknowledgements

The authors wish to thank Alessandro Nigra and Tonina Roggio for their valuable support. The PRIDE team is also acknowledged for the support for MS data deposition. The MS proteomics data in this paper have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [38]: dataset identifier PXD001573. This work was financed by Sardegna Ricerche, program "Art. 26 2012."

The authors have declared no conflict of interest.

Figures

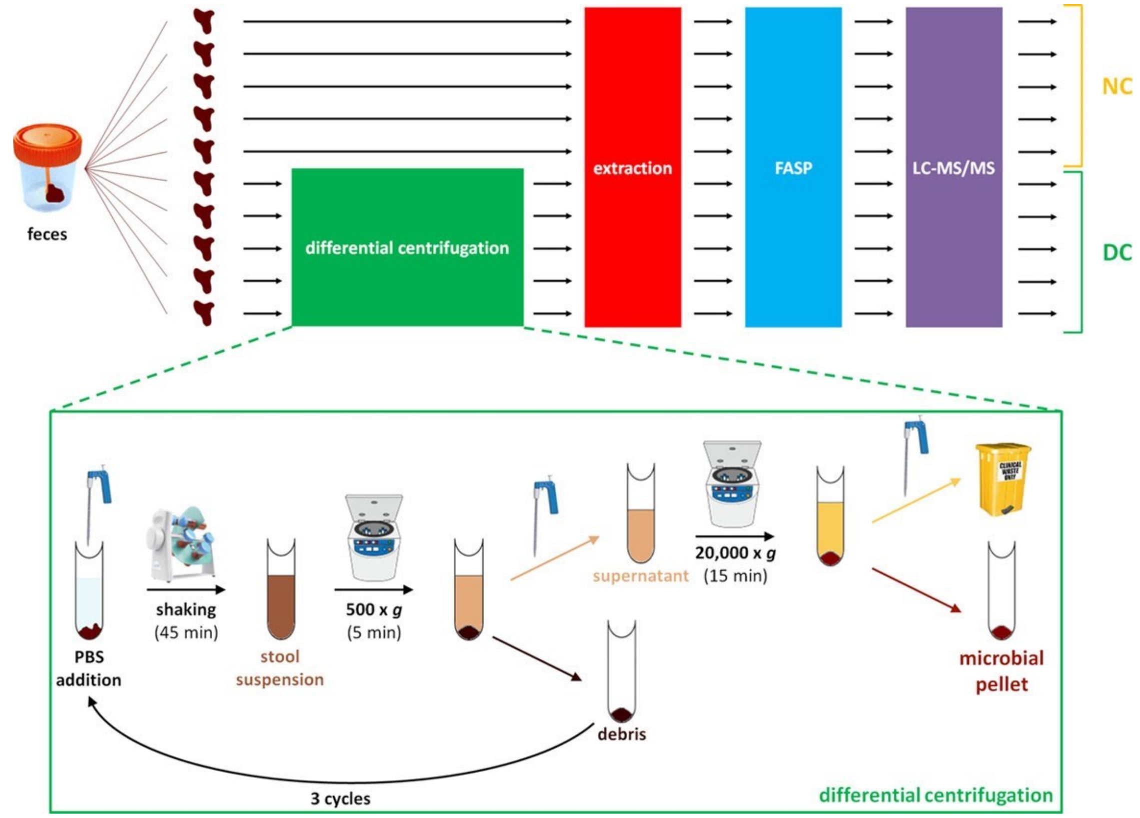


Figure 1. Schematic representation of the study design, with detailed description of the experimental steps comprised in the differential centrifugation pretreatment (green box).

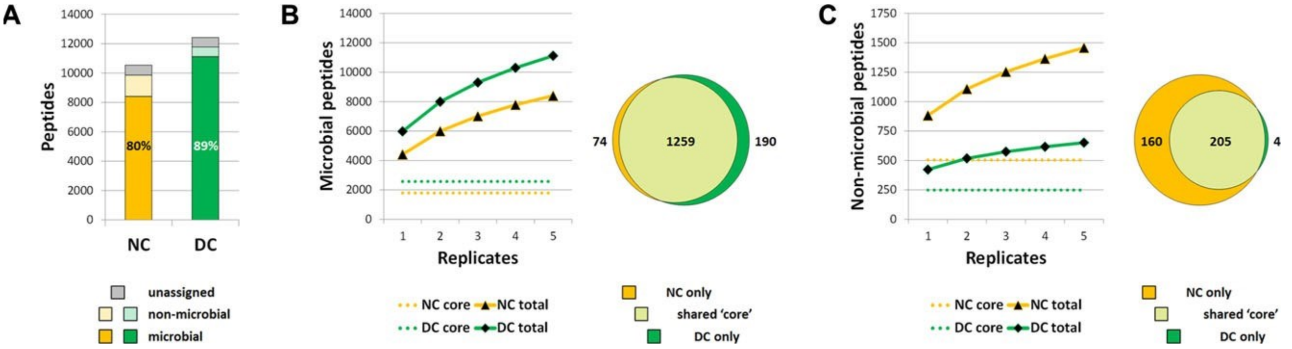


Figure 2. Peptide identification statistics. (A) Histogram comparing the total number of peptides identified without (NC) or with differential centrifugation (DC). Opaque and transparent bars are referred to microbial and nonmicrobial peptides, respectively; gray bars represent peptides with unassigned taxonomy. Percentage values indicate the relative amount of microbial identifications compared to the total. (B) Left, line graph illustrating the cumulative number of microbial peptides detected in five replicate analyses (solid lines), along with the "core" identifications (common to all replicates, dashed lines). Right, Venn diagram depicting the distribution of the microbial "core" peptides in the NC and DC datasets. Specifically, the light green overlapping part comprises peptides detected in all replicates with both methods, while the orange (or dark green) side refers to the peptides found in all NC (or DC) replicates and completely undetected in DC(or NC). (C) The same as in (B), but concerning nonmicrobial peptides.

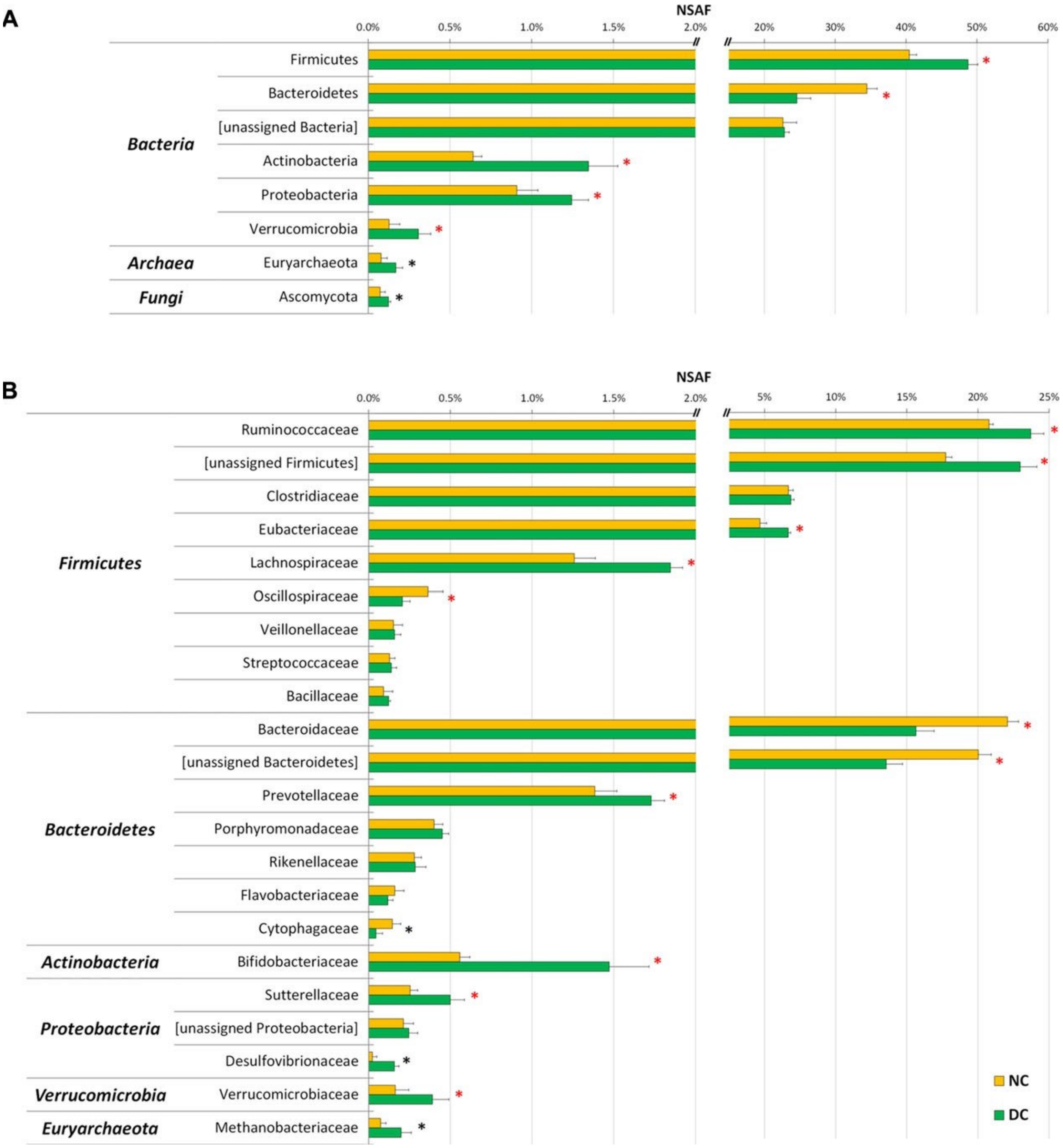


Figure 3. Bar graphs illustrating the microbial phyla (A) and families (B) with a mean abundance higher than 0.1% in NC and/or DC. Peptide taxonomic assignments were carried out based on an LCA approach using Unipept. Phyla (A) and families (B) are grouped based on the (super)kingdom and phylum to which they belong, respectively. Peptide sequences which could not be assigned to a specific phylum (A) or family (B) but only to a higher taxonomic level are shown in square brackets and named as "unassigned" followed by the higher taxonomic level to which they belong. Black and red asterisks indicate a statistically significant difference between groups with $p < 0.05$ and $p < 0.01$, respectively.

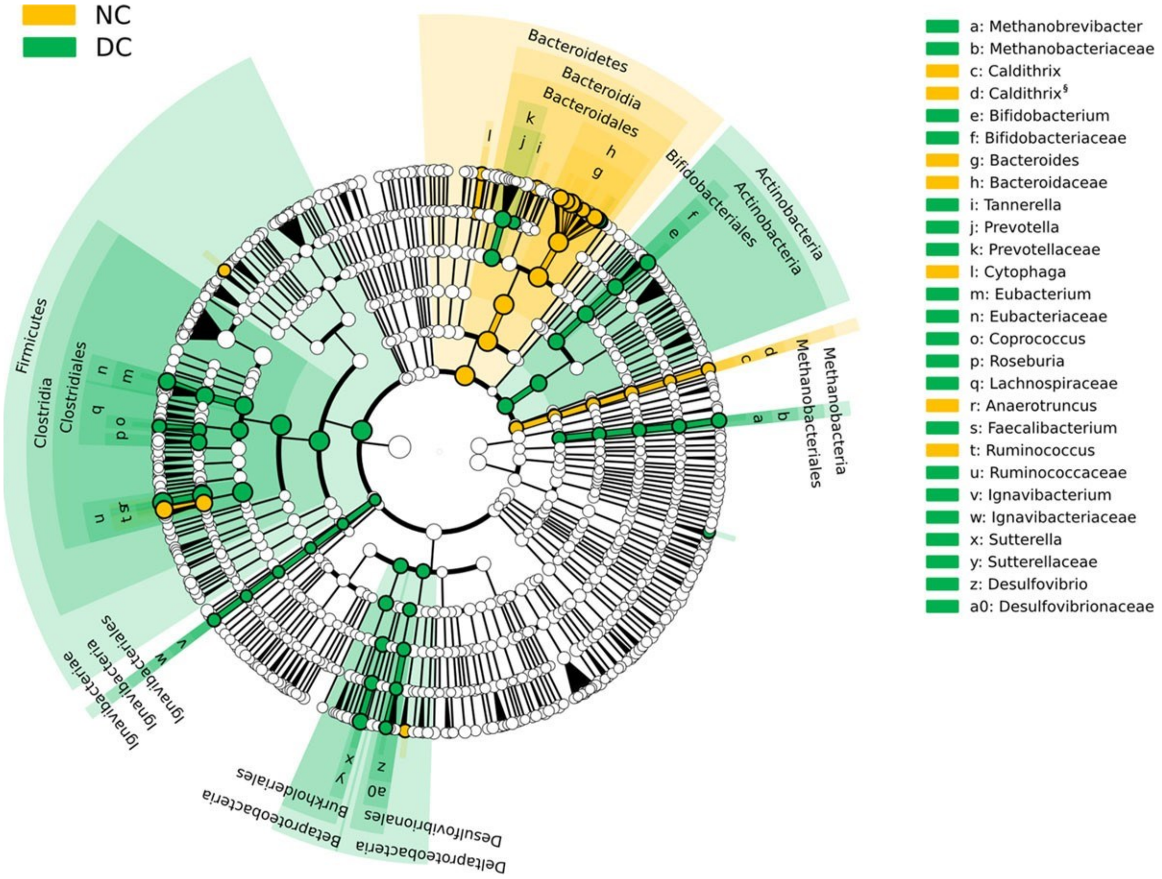


Figure 4. Cladogram showing a hierarchical representation of the taxa identified in this study, generated based on the LEfSe analysis. Each taxon (from the phylum to the species level) is represented by a circle whose size is proportional to the highest logarithmic abundance between the two groups. Taxa with significantly different abundance between NC and DC (Kruskall–Wallis alpha-value < 0.01 and log LDA score > 3) are colored. Phylum, class, and order names are reported within the cladogram, whereas family and genus names are marked with a letter (the legend on the right reports these letters followed by the corresponding taxon name). §Since the family to which the genus Caldithrix belongs is currently unclassified (as well as phylum, class, and order), it has been generically indicated with the same name of the genus.

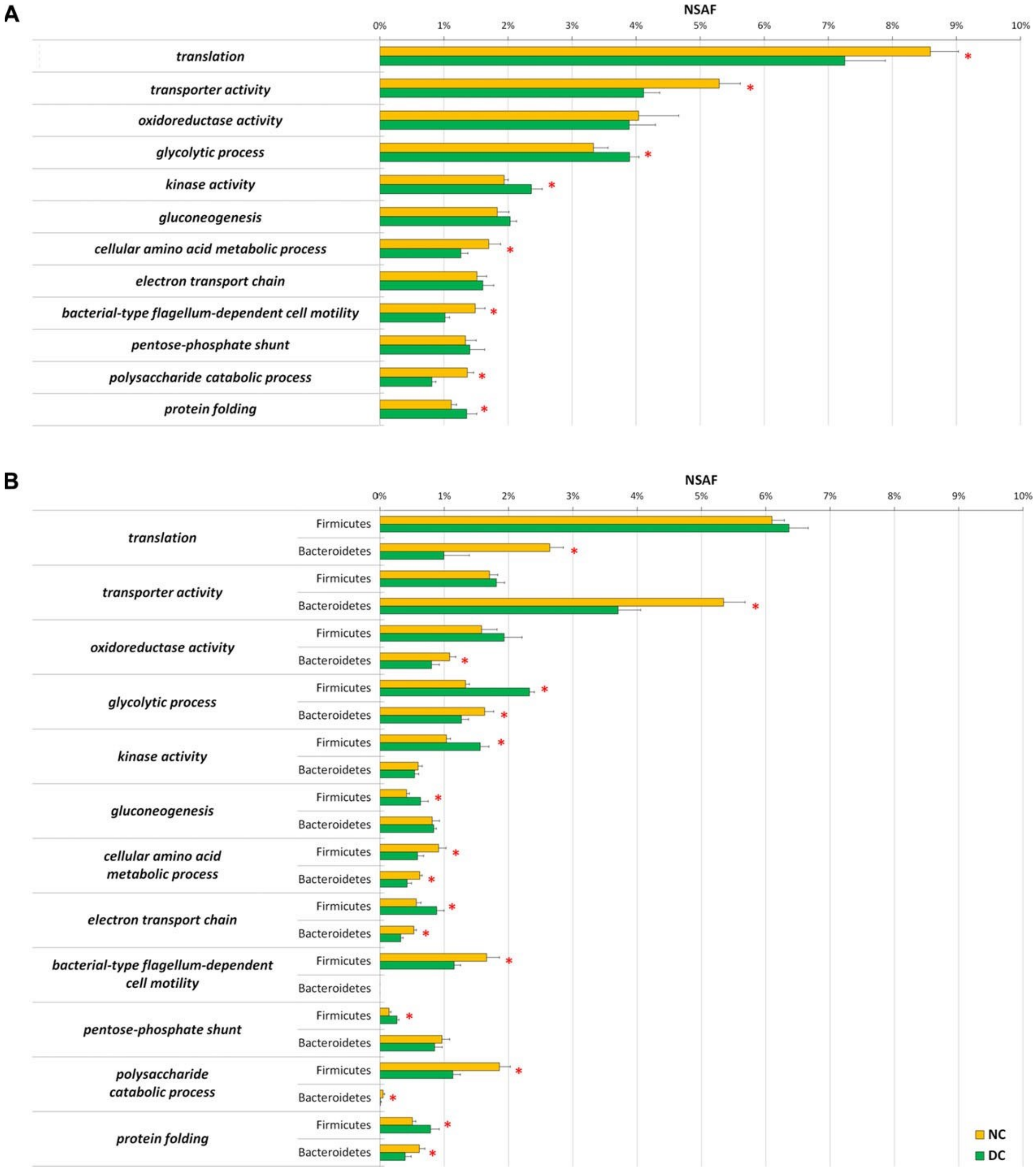


Figure 5. GO-BP classification of the identified proteins. (A) Bar graph illustrating the top 12 GO-BP categories according to the NSAF abundances of the related proteins. Asterisks indicate a statistically significant difference between groups (*p* < 0.01). (B) Taxonomic assignment of functional categories shown in (A): for each GO-BP category, the abundance values corresponding to the two main phyla (Firmicutes and Bacteroidetes) are reported.

References

[1] Hooper, L. V., Littman, D. R., Macpherson, A. J., Interactions between the microbiota and the immune system. *Science* 2012, *336*, 1268–1273.

[2] Tremaroli, V., Backhed, F., Functional interactions between¨ the gut microbiota and host metabolism. *Nature* 2012, *489*, 242–249.

[3] Sommer,F.,Backhed,F.,Thegutmicrobiota—mastersofhost development and physiology. *Nat. Rev. Microbiol.* 2013, *11*, 227–238.

[4] Russell, S. L., Finlay, B. B., The impact of gut microbes in allergic diseases. *Curr. Opin. Gastroenterol.* 2012, *28*, 563– 569.

[5] Collins, S. M., A role for the gut microbiota in IBS. *Nat. Rev.* Gastroenterol. Hepatol. 2014, 11, 497–505.

[6] Tilg, H., Moschen, A. R., Microbiota and diabetes: an evolving relationship. *Gut* 2014, *63*, 1513–1521.

[7] Louis, P., Hold, G. L., Flint, H. J., The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* 2014, *12*, 661–672.

[8] Lamendella, R., VerBerkmoes, N., Jansson, J. K., 'Omics' of the mammalian gut—new insights into function. *Curr. Opin. Biotechnol.* 2012, *23*, 491–500.

[9] Hettich, R. L., Pan, C., Chourey, K., Giannone, R. J., Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem.* 2013, *85*, 4203–4214.

[10] Kolmeder, C. A., deVos, W. M., Metaproteomics of our microbiome—developing insight in function and activity in man and model systems. *J. Proteomics* 2014, *97*, 3–16.

[11] Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A. et al., Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 2009, *3*, 179–189.

[12] Rooijers, K., Kolmeder, C., Juste, C., Dore, J. et al., An it-´ erative workflow for mining the human intestinal metaproteome. *BMC Genomics* 2011, *12*, 6.

[13] Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y. et al., Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 2012, *7*, e49138.

[14] Haange, S. B., Oberbach, A., Schlichting, N., Hugenholtz, F. et al., Metaproteome analysis and molecular genetics of rat intestinal microbiota reveals section and localization resolved species distribution and enzymatic functionalities. *J. Proteome Res.* 2012, *11*, 5406–5417.

[15] Deatherage Kaiser, B. L., Li, J., Sanford, J. A., Kim, Y. M. et al., A multi-omic view of host-pathogen-commensal interplay in *Salmonella*-mediated intestinal infection. *PLoS One* 2013, *8*, e67155.

[16] Ferrer, M., Ruiz, A., Lanza, F., Haange, S. B. et al.., Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ. Microbiol.* 2013, *15*, 211–226.

[17] Perez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H. et al., Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* 2013, *62*, 1591–1601.

[18] Juste, C., Kreil, D. P., Beauvallet, C., Guillot, A. et al., Bacterial protein signals are associated with Crohn's disease. *Gut* 2014, *63*, 1566–1577.

[19] Tang, Y., Underwood, A., Gielbert, A., Woodward, M. J., Petrovska, L., Metaproteomics analysis reveals the adaptation process for the chicken gut microbiota. *Appl. Environ. Microbiol.* 2014, *80*, 478–485.

[20] Kolmeder, C. A., deBeen, M., Nikkila, J., Ritamo, I. et al.,¨ Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One* 2012, *7*, e29913.

[21] Tanca, A., Palomba, A., Pisanu, S., Deligios, M. et al., A straightforward and efficient analytical pipeline for metaproteome characterization. *Microbiome* 2014, *2*, 49.

[22] Wisniewski, J. R., Zougman, A., Nagaraj, N., Mann, M., Universal sample preparation method for proteome analysis. *Nat. Methods* 2009, *6*, 359–362.

[23] Tanca, A., Biosa, G., Pagnozzi, D., Addis, M. F., Uzzau, S., Comparison of detergent-based sample preparation workflows for LTQ-Orbitrap analysis of the *Escherichia coli* proteome. *Proteomics* 2013, *13*, 2597–2607.

[24] Tanca, A., Abbondio, M., Pisanu, S., Pagnozzi, D. et al., Critical comparison of sample preparation strategies for shotgun proteomic analysis of formalin-fixed, paraffin-embedded samples: insights from liver tissue. *Clin. Proteomics* 2014, *11*, 28.

[25] Tanca, A., Palomba, A., Deligios, M., Cubeddu, T. et al., Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* 2013, *8*, e82981.

[26] Mesuere, B., Devreese, B., Debyser, G., Aerts, M. et al., Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* 2012, *11*, 5773– 5780.

[27] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. et al., Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[28] Tanca,A.,Pagnozzi,D.,Burrai,G.P.,Polinas,M.etal.,Comparabilityofdifferentialproteomicsdata generatedfrompaired archival fresh-frozen and formalin-fixed samples by GeLCMS/MS and spectral counting. *J. Proteomics* 2012, *77*, 561– 576.

[29] Robles, M. S., Cox, J., Mann, M., In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS Genet.* 2014, *10*, e1004047.

[30] Hill, T. C., Walsh, K. A., Harris, J. A., Moffett, B. F., Using ecological diversity measures with bacterial communities. FEMS Microbiol. Ecol. 2003, 43, 1–11.

[31] McDowall, J., Hunter, S., InterPro protein classification. Methods Mol. Biol. 2011, 694, 37–47.

[32] Uniprot Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2012, *40*, D71–D75.

[33] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M. et al., Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 2014, *42*, D199–D205.

[34] Xie, C., Mao, X., Huang, J., Ding, Y. et al., KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 2011, *39*, W316–W322.

[35] Segata, N., Izard, J., Waldron, L., Gevers, D. et al., Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011, *12*, R60.

[36] Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G. et al., PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010, *26*, 1608–1615.

[37] Krogh, A., Larsson, B., vonHeijne, G., Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 2001, *305*, 567–580.

[38] Ternent, T., Csordas, A., Qi, D., Gomez-Baena, G. et al., How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* 2014, *14*, 2233–2241.

[39] Vizcaino,J.A.,Cote,R.G.,Csordas,A.,Dianes,J.A.etal.,The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*, D1063–D1069.

[40] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V. et al., An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 2006, *444*, 1027–1031.

[41] Larsen, N., Vogensen, F. K., vanden Berg, F. W., Nielsen, D. S. et al., Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 2010, *5*, e9085.

[42] Man, S. M., Kaakoush, N. O., Mitchell, H. M., The role of bacteria and pattern-recognition receptors in Crohn's disease. Nat. Rev. Gastroenterol. Hepatol. 2011, 8, 152–168.

[43] Leitch,E.C.,Walker,A.W.,Duncan,S.H.,Holtrop,G.,Flint,H. J., Selective colonization of insoluble substrates by human faecal bacteria. *Environ. Microbiol.* 2007, *9*, 667–679.

[44] Lee, S. M., Donaldson, G. P., Mikulski, Z., Boyajian, S. et al., Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* 2013, *501*, 426–429.

[45] Muth, T., Benndorf, D., Reichl, U., Rapp, E., Martens, L., Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.* 2013, *9*, 578– 585.

[46] Seifert, J., Herbst, F. A., Halkjaer Nielsen, P., Planes, F. J. et al., Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics* 2013, *13*, 2786–2804.

[47] Koropatkin, N. M., Cameron, E. A., Martens, E. C., How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* 2012, *10*, 323–335.

[48] Flint, H. J., Scott, K. P., Louis, P., Duncan, S. H., The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* 2012, *9*, 577–589.

Supplementary Files

S1. List of microbial genera used for protein database generation.

Acaryochloris, Acetohalobium, Acholeplasma, Achromobacter, Acidaminococcus, Acidithiobacillus, Acinetobacter, Aeromonas, Agaricus, Agrobacterium, Akkermansia, Algoriphagus, Alistipes, Alkaliphilus, Alteromonas, Aminobacterium, Ammonifex, Amphibacillus, Anabaena, Anaeromyxobacter, Anaerostipes, Anaerotruncus, Anoxybacillus, Arcobacter, Arthromitus, Asticcacaulis, Atopobium, Azobacteroides, Bacillus, Bacteroides, Barnesiella, Bartonella, Batrachochytrium, Bifidobacterium, Blautia, Blochmannia, Borrelia, Brachyspira, Bradyrhizobium, Brevibacillus, Buchnera, Burkholderia, Butyrivibrio, Caedibacter, Caldiarchaeum, Caldicellulosiruptor, Caldilinea, Caldithrix, Candida, Capnocytophaga, Carnobacterium, Carsonella, Catenibacterium, Cellulomonas, Cellulosilyticum, Ceriporiopsis, Chaetomium, Chamaesiphon, Chlorobaculum, Chlorobium, Chloroflexus, Claviceps, Clavispora, Cloacamonas, Clostridium, Colletotrichum, Collinsella, Coprobacillus, Coprococcus, Coriobacterium, Corynebacterium, Cronobacter, Cryptobacterium, Cupriavidus, Cyclobacterium, Cytophaga, Dechloromonas, Dehalobacter, Deinococcus, Desulfitobacterium, Desulfobacca, Desulfocapsa, Desulfosporosinus, Desulfotomaculum, Desulfovibrio, Desulfurivibrio, Desulfurobacterium, Desulfurococcus, Dialister, Dinoroseobacter, Dokdonia, Dorea, Elusimicrobium, Emticicia, Encephalitozoon, Endolissoclinum, Enterococcus, Erysipelothrix, Ethanoligenens, Eubacterium, Eutypa, Exiguobacterium, Faecalibacterium, Ferrimonas, Fervidicoccus, Fibrobacter, Finegoldia, Flavobacterium, Flexistipes, Frankia, Fusobacterium, Geobacillus, Geobacter, Glaciecola, Gloeobacter, Gluconacetobacter, Gordonia, Granulibacter, Granulicella, Halalkalicoccus, Halanaerobium, Haliscomenobacter, Halobacillus, Halobacteroides, Halomonas, Halothermothrix, Halothiobacillus, Helicobacter, Hepatoplasma, Hyphomicrobium, Ignavibacterium, Ignicoccus, Ilyobacter, Janthinobacterium, Kazachstania, Kinetoplastibacterium, Lacinutrix, Lactobacillus, Lactococcus, Legionella, Leptolyngbya, Leptospira, Leptospirillum, Leuconostoc, Liberibacter, Lodderomyces, Macrophomina, Magnetococcus, Magnetospirillum, Mannheimia, Maribacter, Maricaulis, Marinobacter, Megamonas, Megasphaera, Mesorhizobium, Mesotoga, Methanobrevibacter, Methanocaldococcus, Methanocella, Methanococcus, Methanomassiliicoccus, Methanomethylophilus, Methanosarcina, Methanosphaerula, Methylacidiphilum, Methylobacillus, Methylobacterium, Methylomirabilis, Methylomonas, Methylophaga, Methylotenera, Micrococcus, Mixia, Moranella, Mucilaginibacter, Mycobacterium, Mycoplasma, Myroides, Myxococcus, Nasuia, Niabella, Niastella, Nitrosoarchaeum, Nitrosopumilus, Nitrososphaera, Nitrospira, Odoribacter, Olsenella, Opitutus, Oscillibacter, Paenibacillus, Paludibacter, Pantoea, Parabacteroides, Paracaedibacter, Paracoccus, Paraprevotella, Parasutterella, Pediococcus, Pedobacter, Pelagibacter, Pelosinus, Penicillium, Peptostreptococcus, Petrotoga, Phytoplasma, Planctomyces, Polaromonas, Porphyromonas, Portiera, Prevotella, Prochlorococcus, Profftella, Propionibacterium, Providencia, Pseudanabaena, Pseudoalteromonas, Pseudogulbenkiania, Pseudomonas, Pseudonocardia, Puccinia, Pyrobaculum, Pyrococcus, Ralstonia, Regiella, Rhizobium, Rhodococcus, Rhodopirellula, Rhodopseudomonas, Rhodospirillum, Rhodothermus, Roseburia, Roseobacter, Rothia, Ruminococcus, Saccharibacteria, Saccharimonas, Scheffersomyces, Selenomonas, Serpula, Shewanella, Simkania, Singulisphaera, Sinorhizobium, Slackia, Sodalis, Solibacter, Sphaerochaeta, Sphingomonas, Spirochaeta, Staphylococcus, Stenotrophomonas, Streptobacillus, Streptococcus, Streptomyces, Subdoligranulum, Succinatimonas, Sulcia, Sulfobacillus, Sulfurovum, Sutterella, Symbiobacter, Symbiobacterium, Synechococcus, Tannerella, Teredinibacter, Tetragenococcus, Thalassolituus, Thermacetogenium, Thermaerobacter, Thermobacillus, Thermodesulfobium, Thermomicrobium, Thermomonospora, Thermotoga, Thermovirga, Thermus, Thielavia, Thioalkalivibrio, Thiomicrospira, Tremblaya, Treponema, Trichosporon, Uzinura, Veillonella, Verrucosispora, Verticillium, Vibrio, Wallemia, Wolbachia, Yarrowia, Yersinia, Zygosaccharomyces, Zymomonas.

Figure S1. Histogram comparing the total number of proteins identified without (NC) or with differential centrifugation (DC). Opaque and transparent bars are referred to microbial and non-microbial proteins, respectively; grey bars refers indeed to taxonomically unassigned proteins.

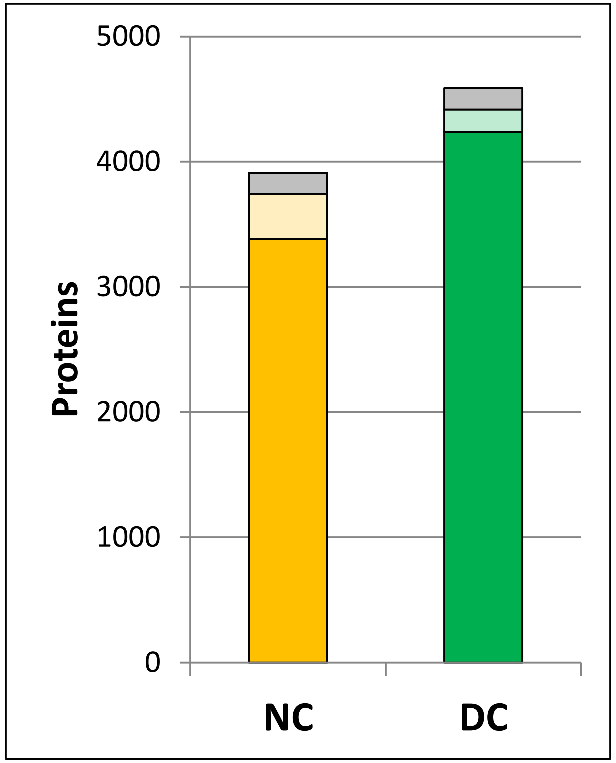


Figure S2. Histogram comparing the total number of peptide-spectrum matches (PSMs) obtained without (NC) or with differential centrifugation (DC). Opaque and transparent bars are referred to microbial and non-microbial PSMs, respectively; grey bars refers indeed to taxonomically unassigned PSMs.

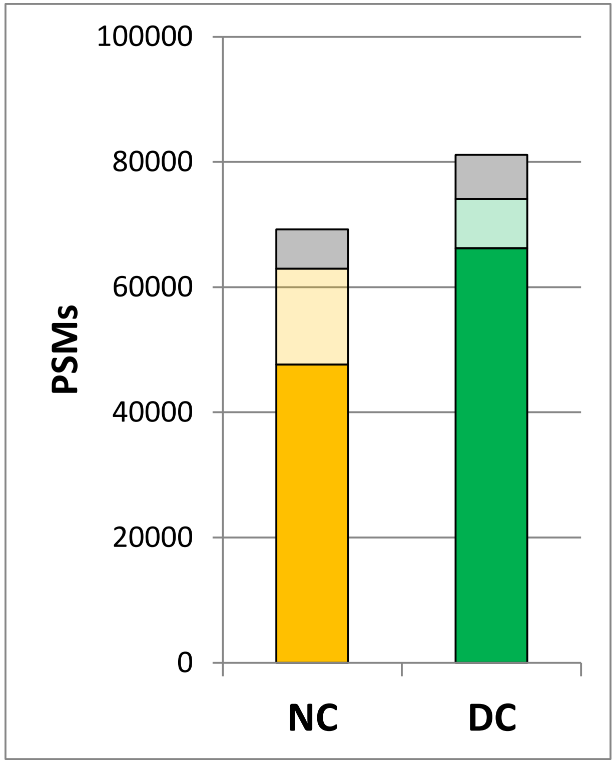


Figure S3. Percentage of overall MS/MS spectra consistently matched with peptide sequences (Percolator *q*-value < 0.01).

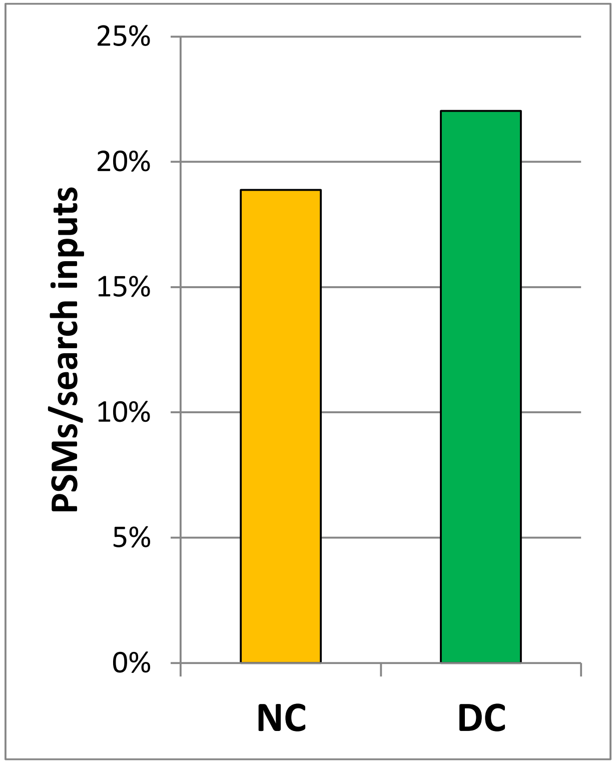


Figure S4. Pie charts illustrating taxonomic distribution of peptides identified in NC and DC samples.

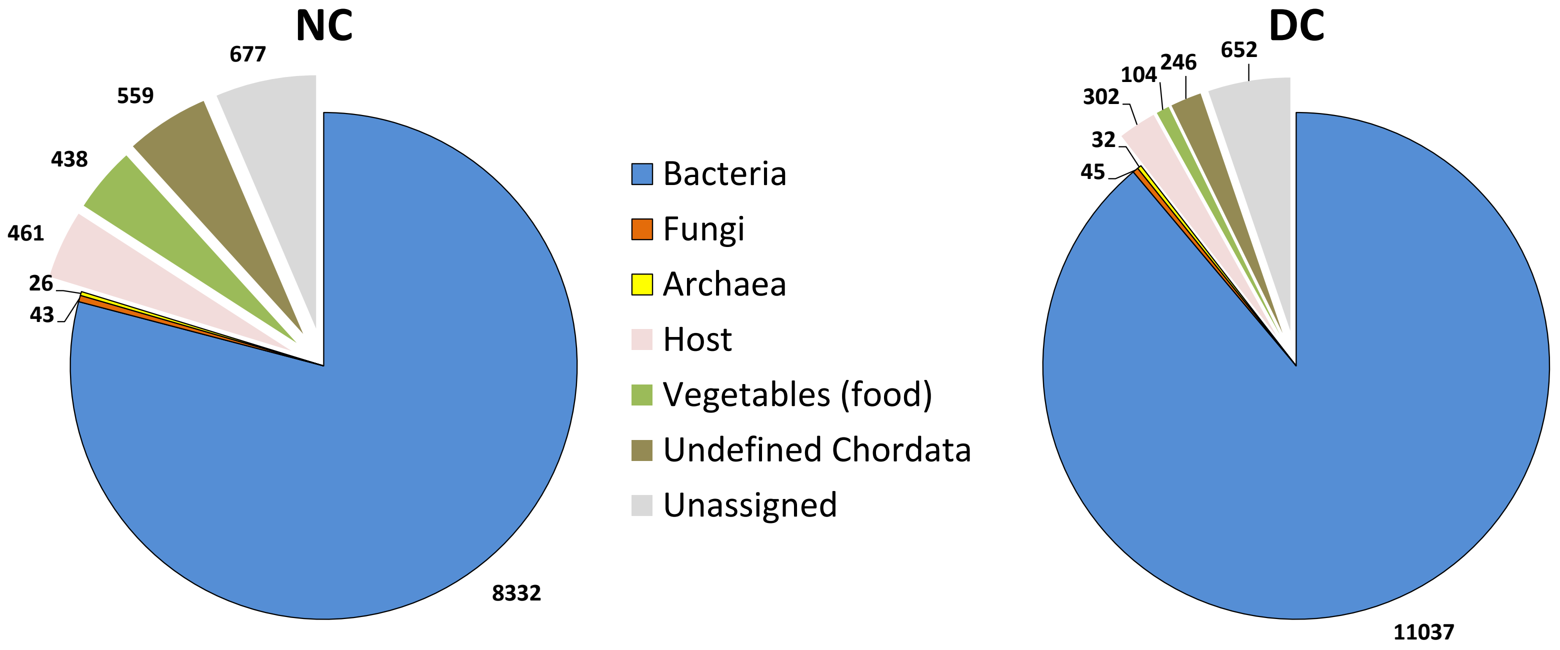
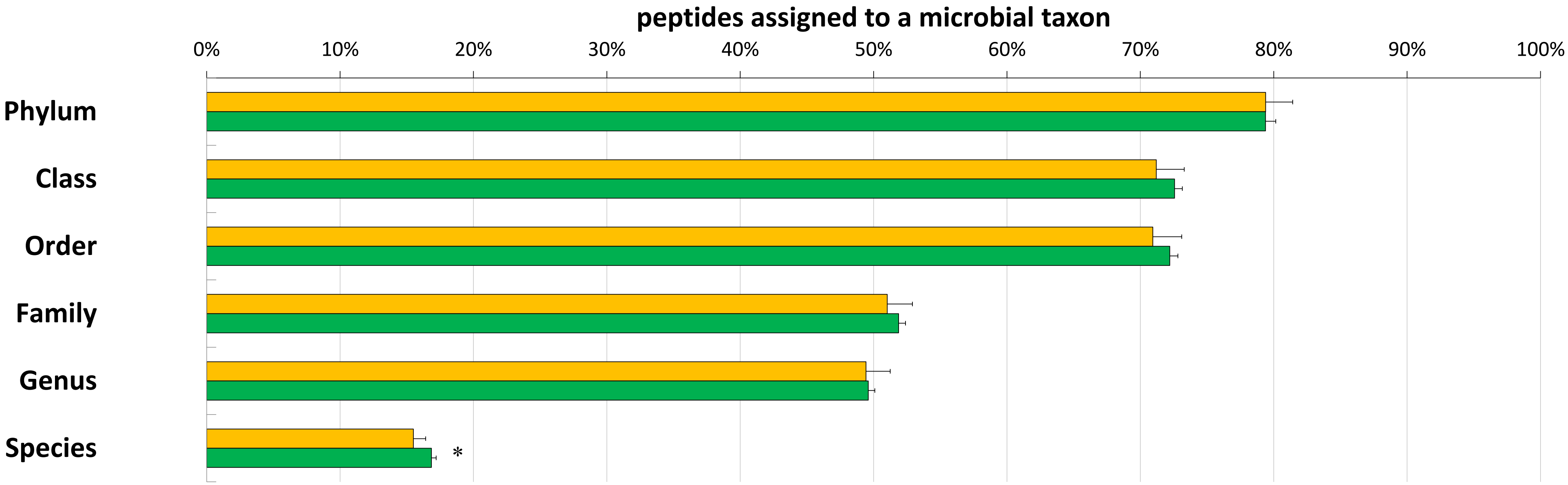


Figure S5. Percentage of microbial peptide sequences consistently assigned to a taxon after LCA analysis with Unipept.



* = $p < 0.05$; orange = NC; green = DC.

Table S1. Evaluation of alpha-diversity.

	Pre-treatment method	mean	SD	NC vs DC (*p*-value) *
Simpson	NC	3.8970	0.0949	1.14E-05
	DC	4.3292	0.0333	
Shannon-Wiener	NC	1.8124	0.0182	7.38E-07
	DC	1.9612	0.0158	

* statistically significant *p*-values (< 0.0001) are shown in bold-type

Figure S6. Cumulative number of microbial taxonomic families detected in five replicate analyses (dashed lines), along with the 'core' identifications (common to all replicates, solid lines).

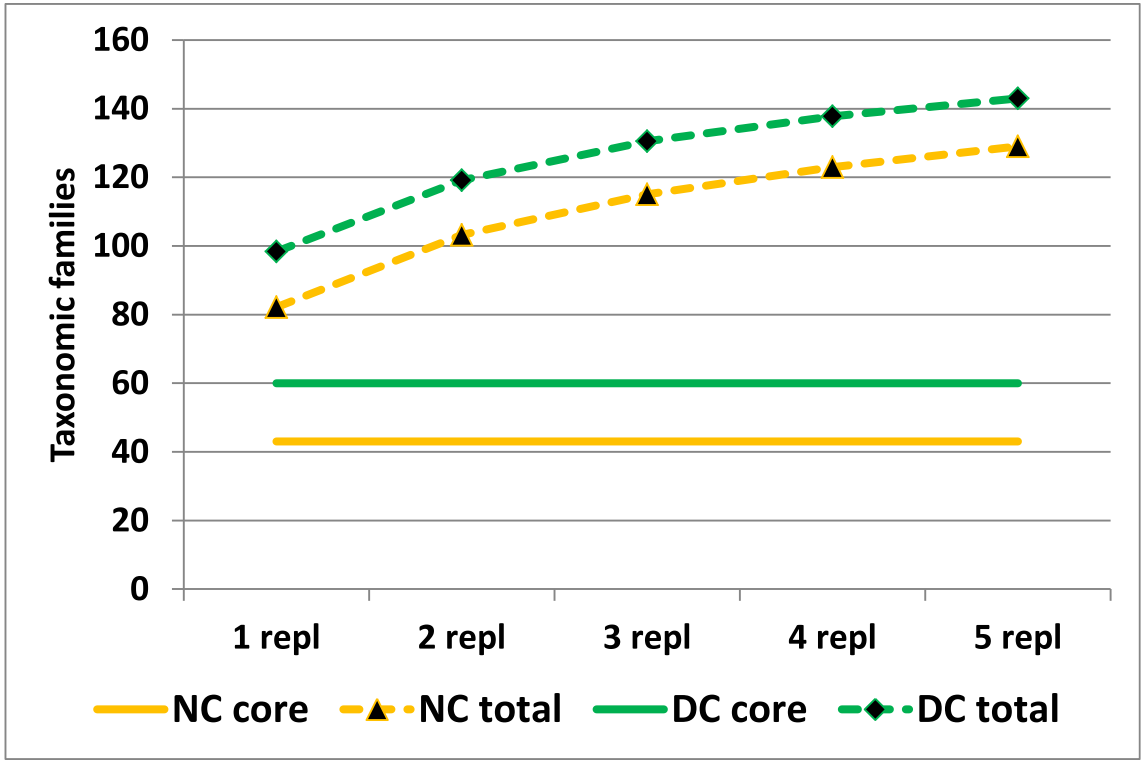


Figure S7. Bar graph illustrating the differential taxonomic genera ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

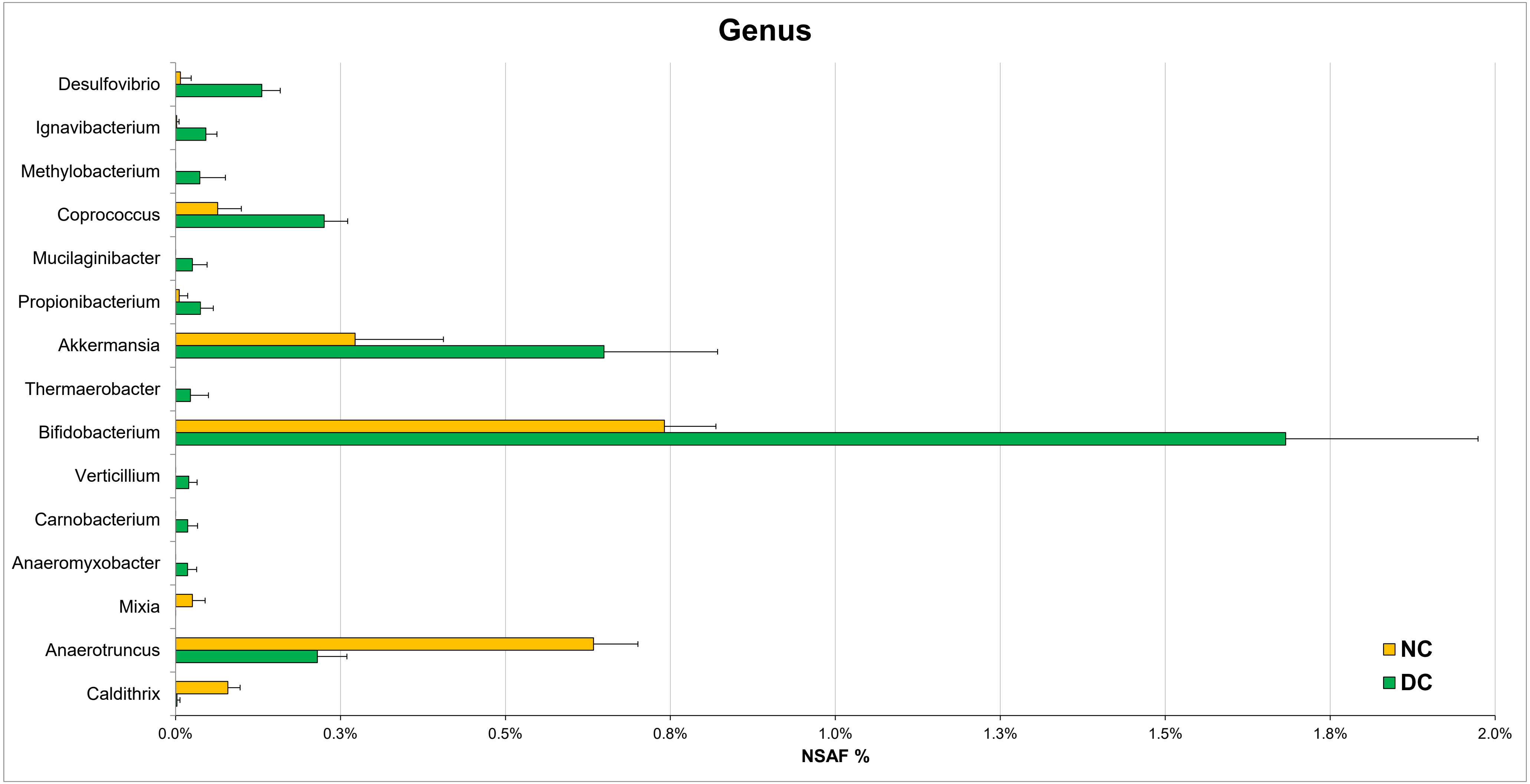


Figure S8. Bar graph illustrating the differential taxonomic species ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

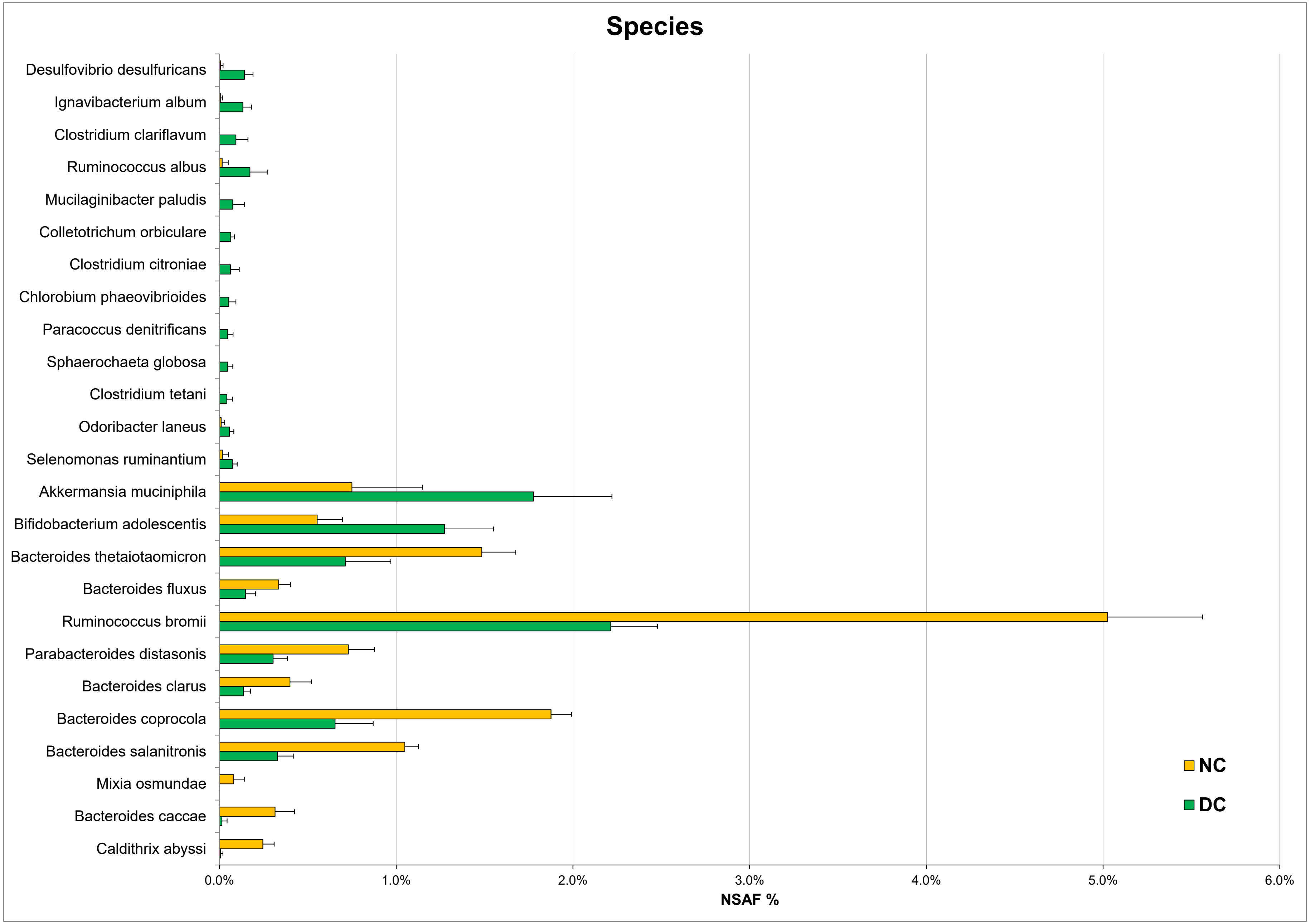


Figure S9. Cumulative number of microbial GO biological processes detected in five replicate analyses (dashed lines), along with the 'core' identifications (common to all replicates, solid lines).

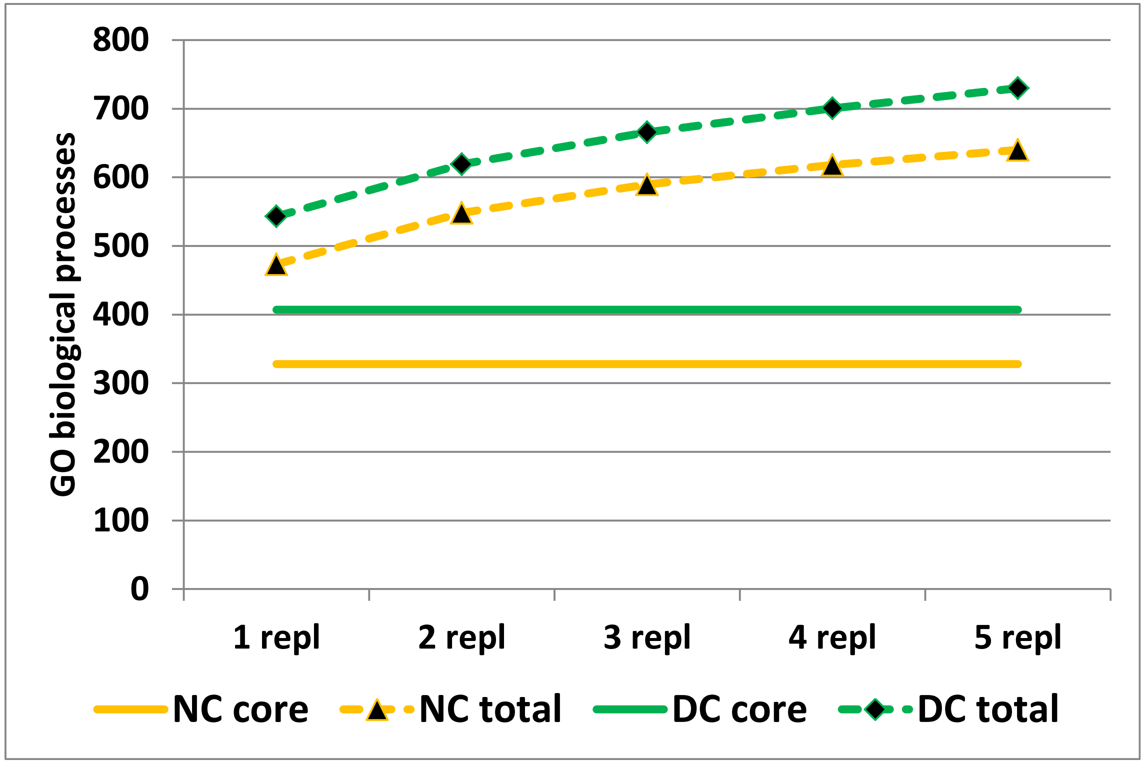


Figure S10. Bar graph illustrating the differential microbial GO biological processes ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

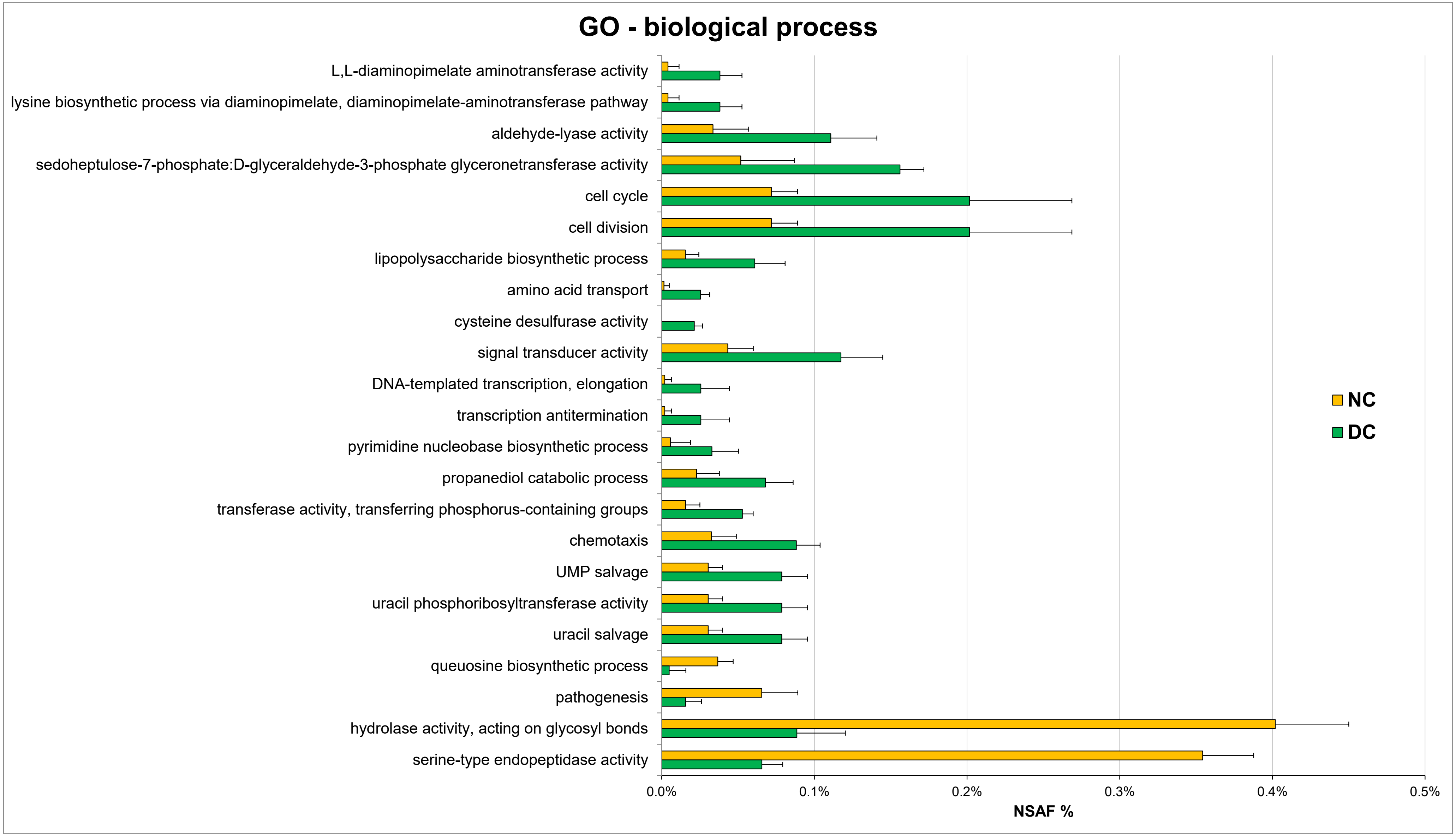


Figure S11. Bar graph illustrating the differential combined microbial biological processes/phyla ($p < 0.01$ and log ratio > 1.5), ordered by decreasing log ratio (DC/NC).

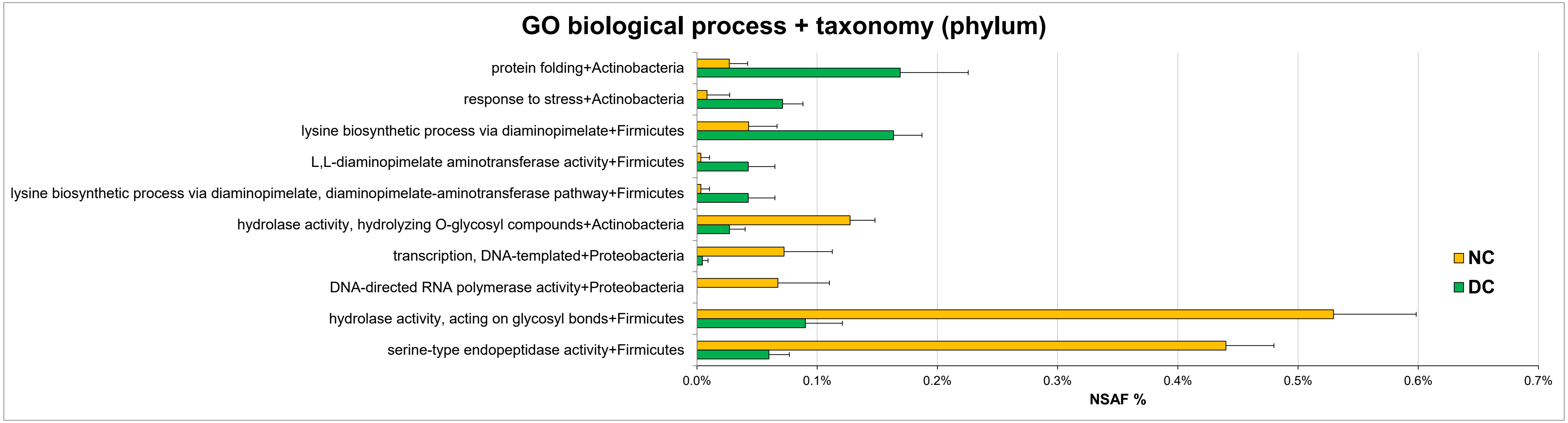


Figure S12. Bar graph illustrating the combined microbial biological processes/taxonomic families identified with a mean abundance higher than 1% in NC and/or DC.

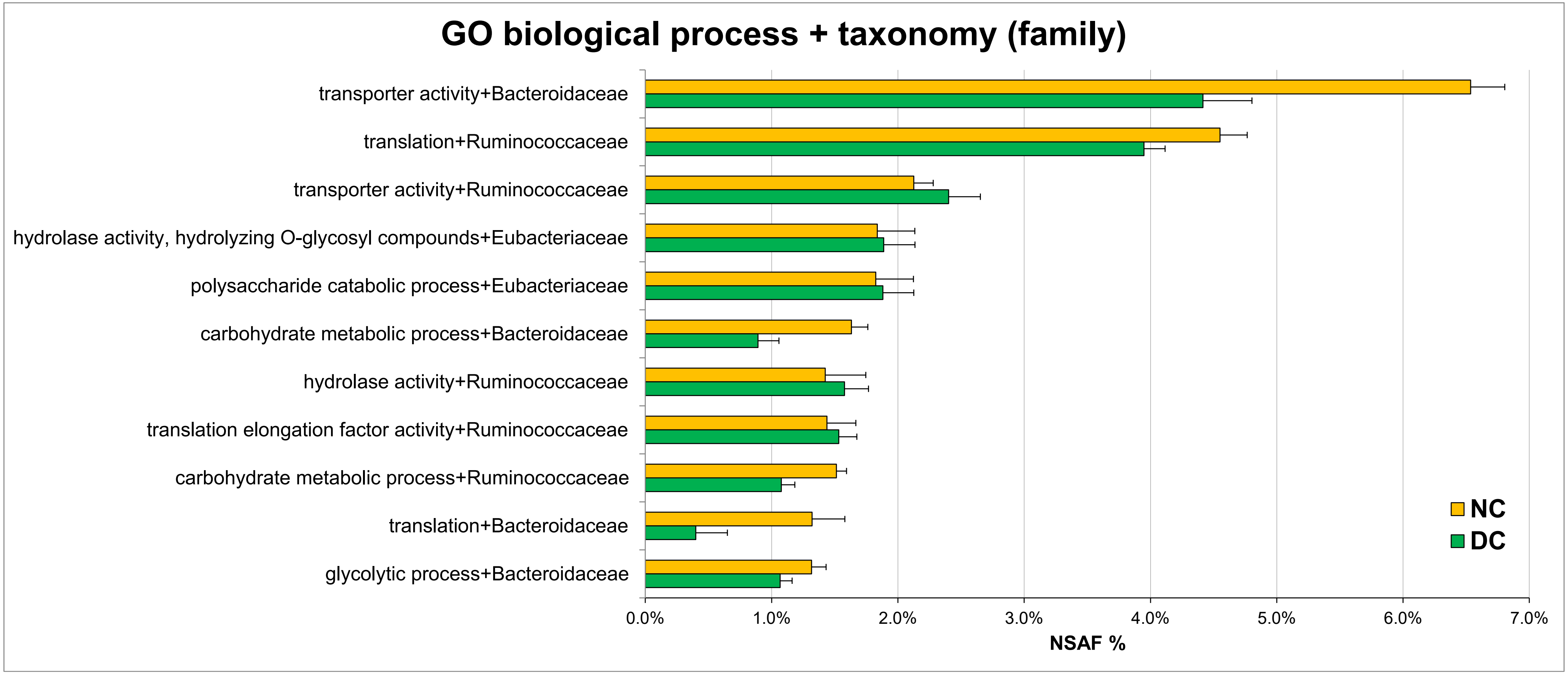


Figure S13. Bar graph illustrating the differential combined microbial biological processes/taxonomic families ($p < 0.01$ and log ratio > 2), ordered by decreasing log ratio (DC/NC).

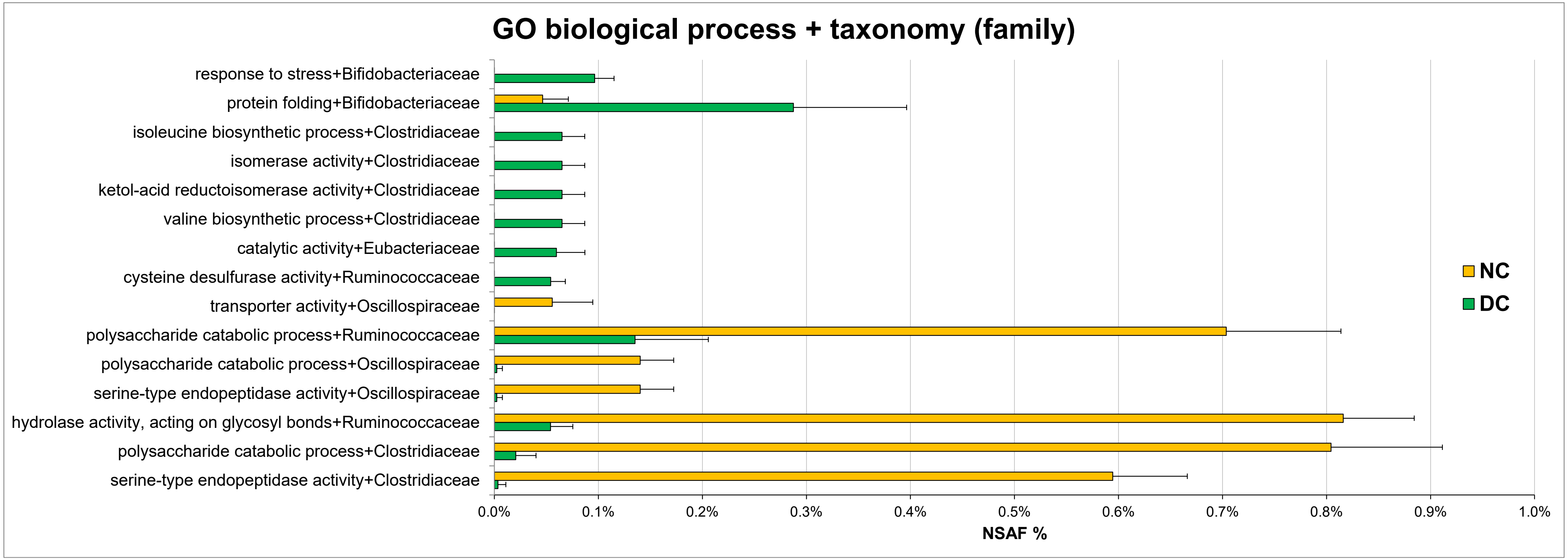


Figure S14. Cumulative number of microbial UniProt protein families detected in five replicate analyses (dashed lines), along with the ‘core’ identifications (common to all replicates, solid lines).

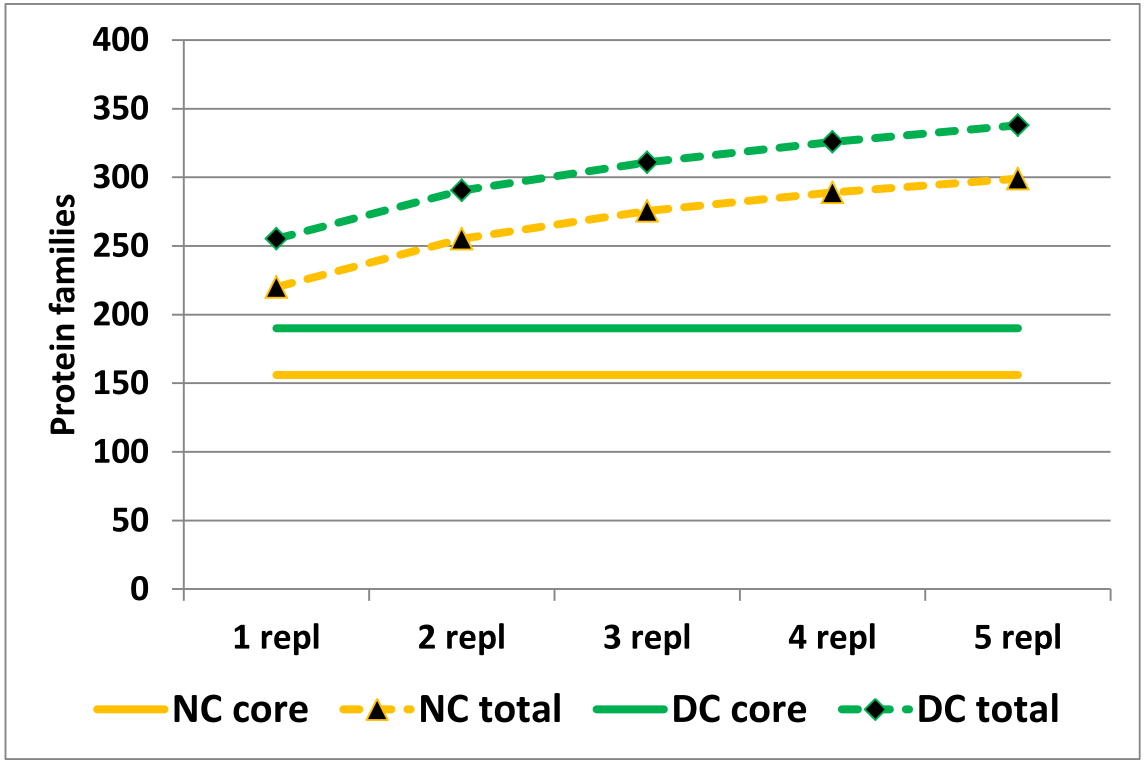


Figure S15. Bar graph illustrating the microbial UniProt protein families identified with a mean abundance higher than 1% in NC and/or DC.

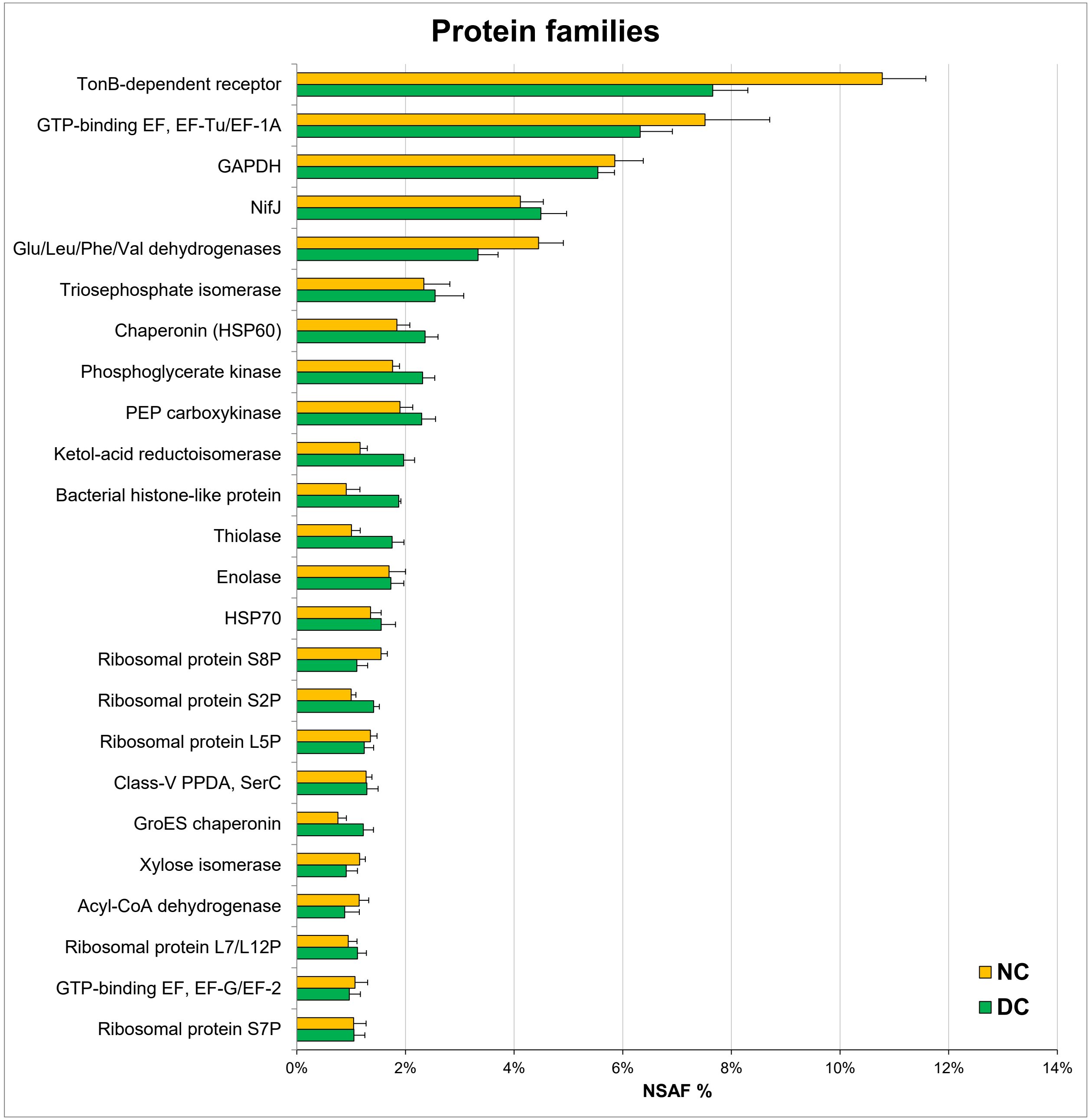


Figure S16. Bar graph illustrating the differential microbial UniProt protein families ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

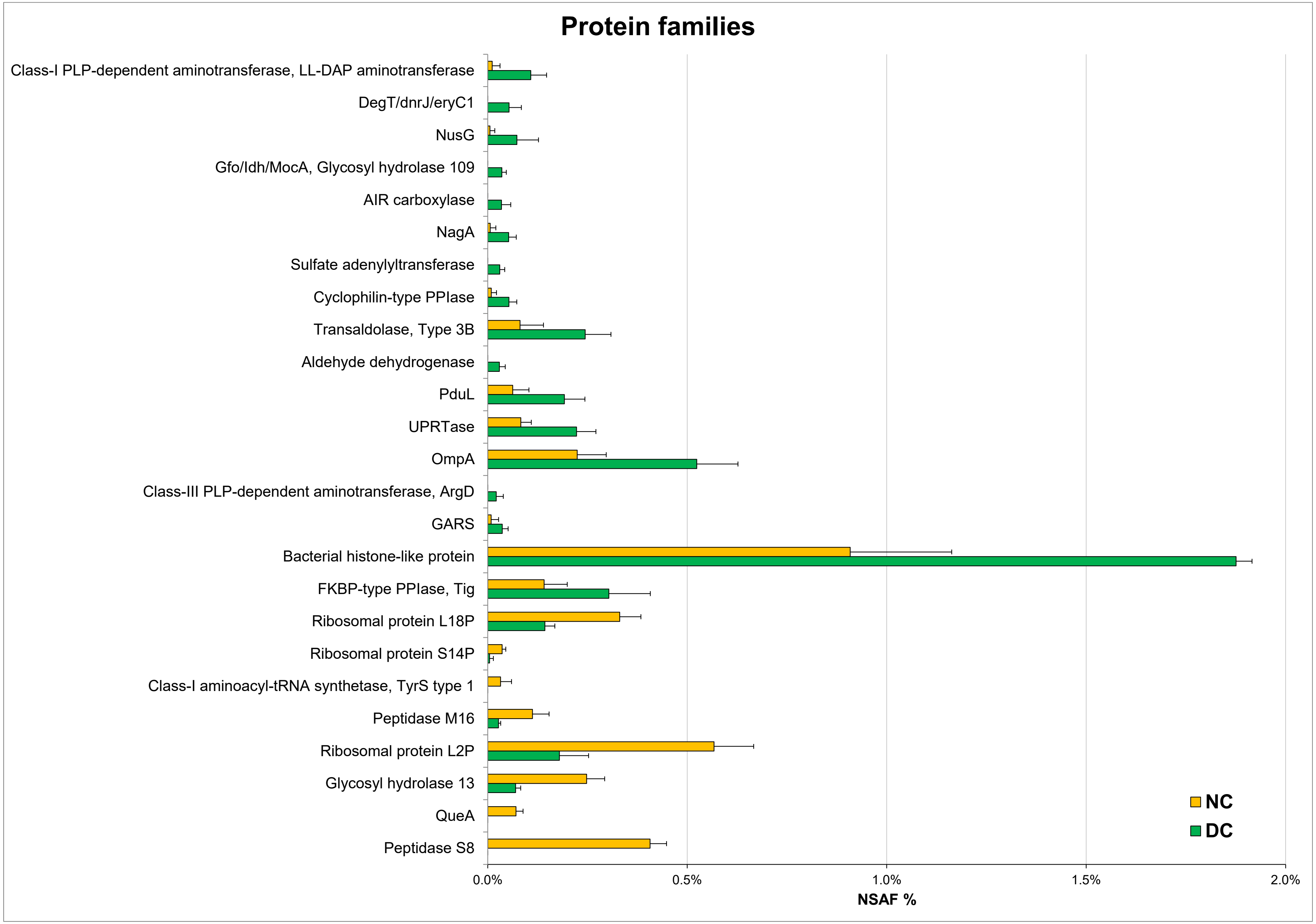


Figure S17. Bar graph illustrating the combined microbial protein families/phyla identified with a mean abundance higher than 1% in NC and/or DC.

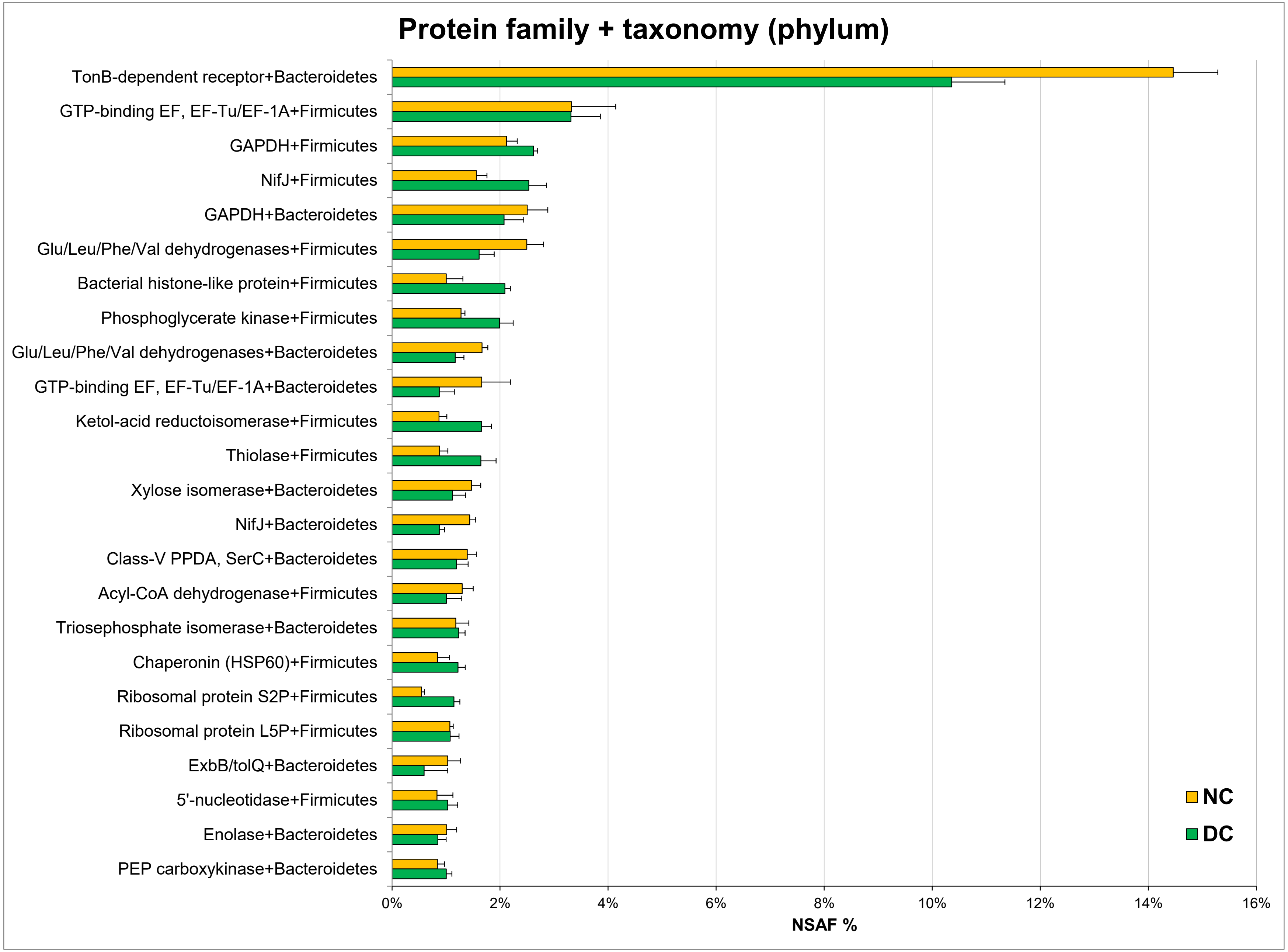


Figure S18. Bar graph illustrating the differential combined microbial protein families/phyla ($p < 0.01$ and log ratio > 1.5), ordered by decreasing log ratio (DC/NC).

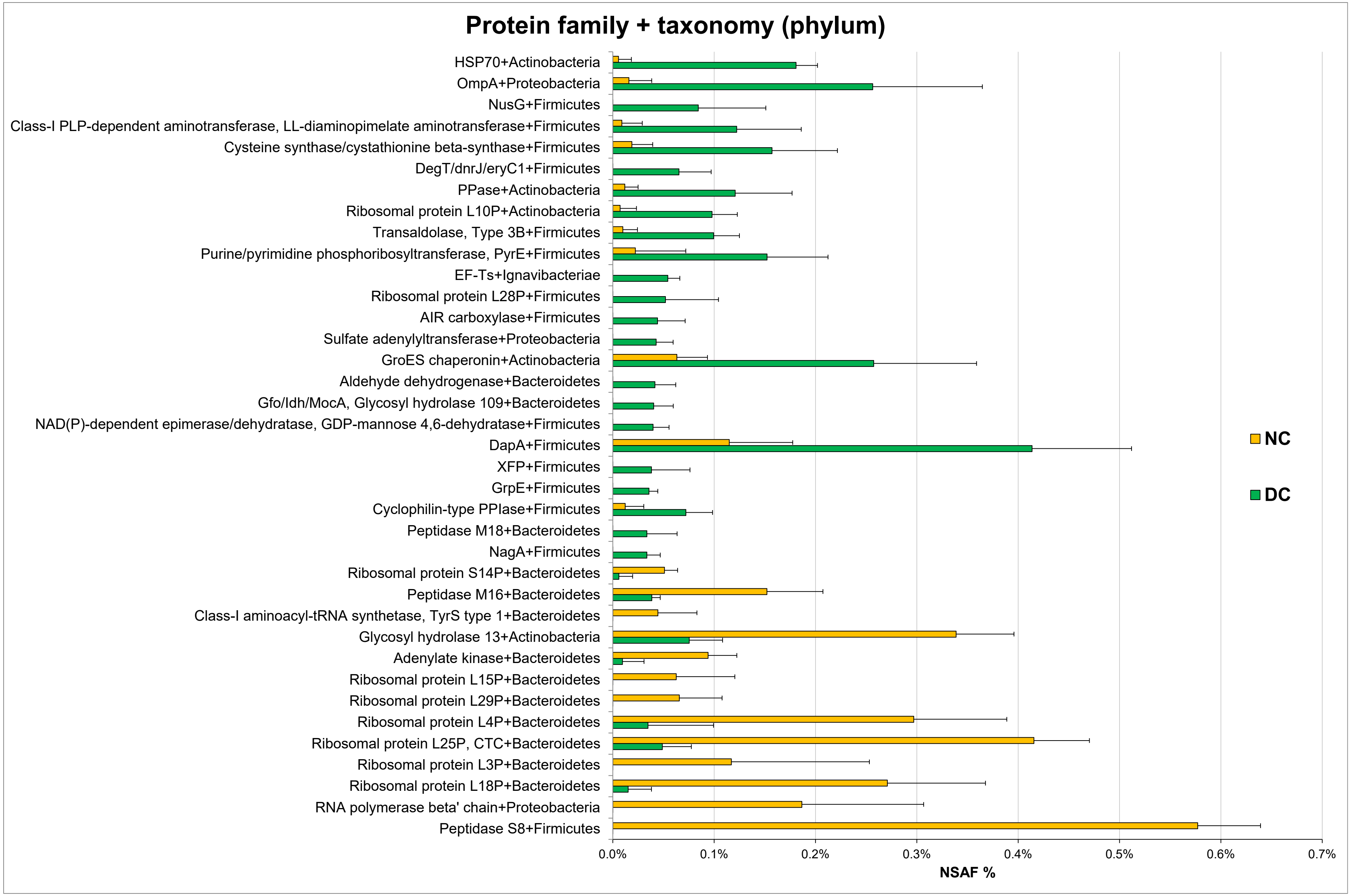


Figure S19. Bar graph illustrating the combined microbial protein families/taxonomic families identified with a mean abundance higher than 1% in NC and/or DC.

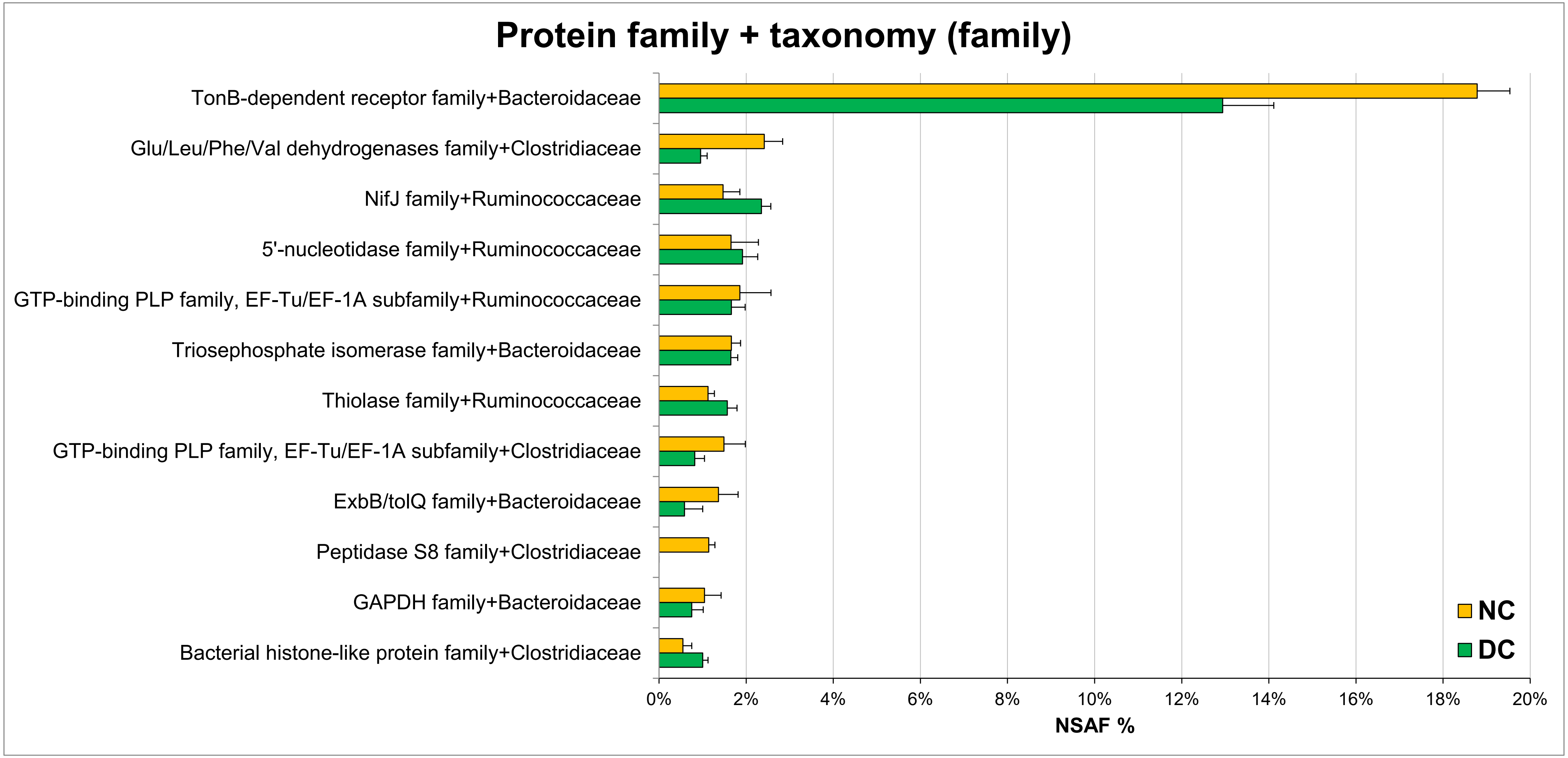


Figure S20. Bar graph with differential combined microbial protein families/taxonomic families ($p < 0.01$ and log ratio > 2), ordered by decreasing log ratio (DC/NC).

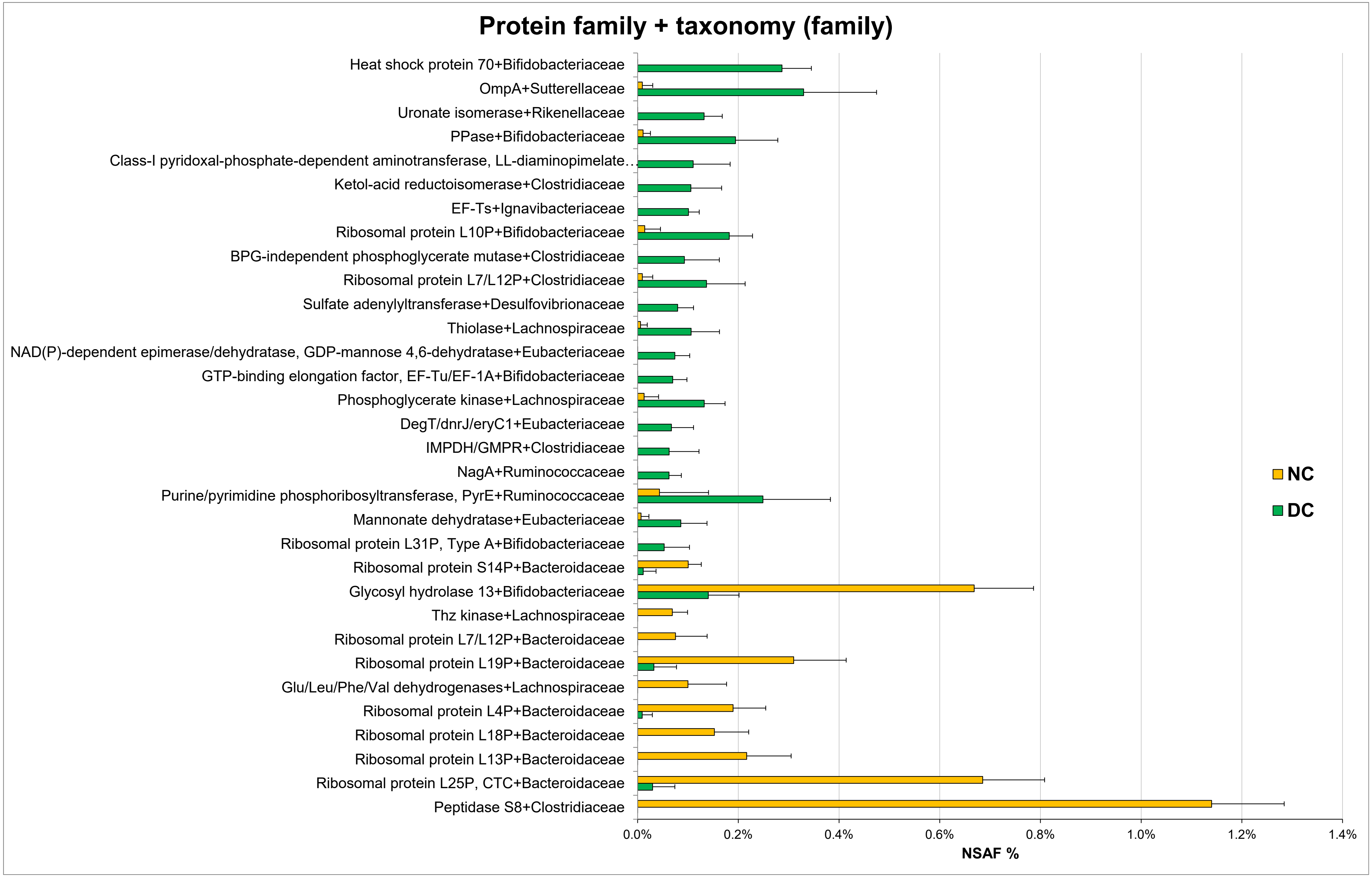


Figure S21. Cumulative number of microbial KEGG ortholog groups detected in five replicate analyses (dashed lines), along with the 'core' identifications (common to all replicates, solid lines).

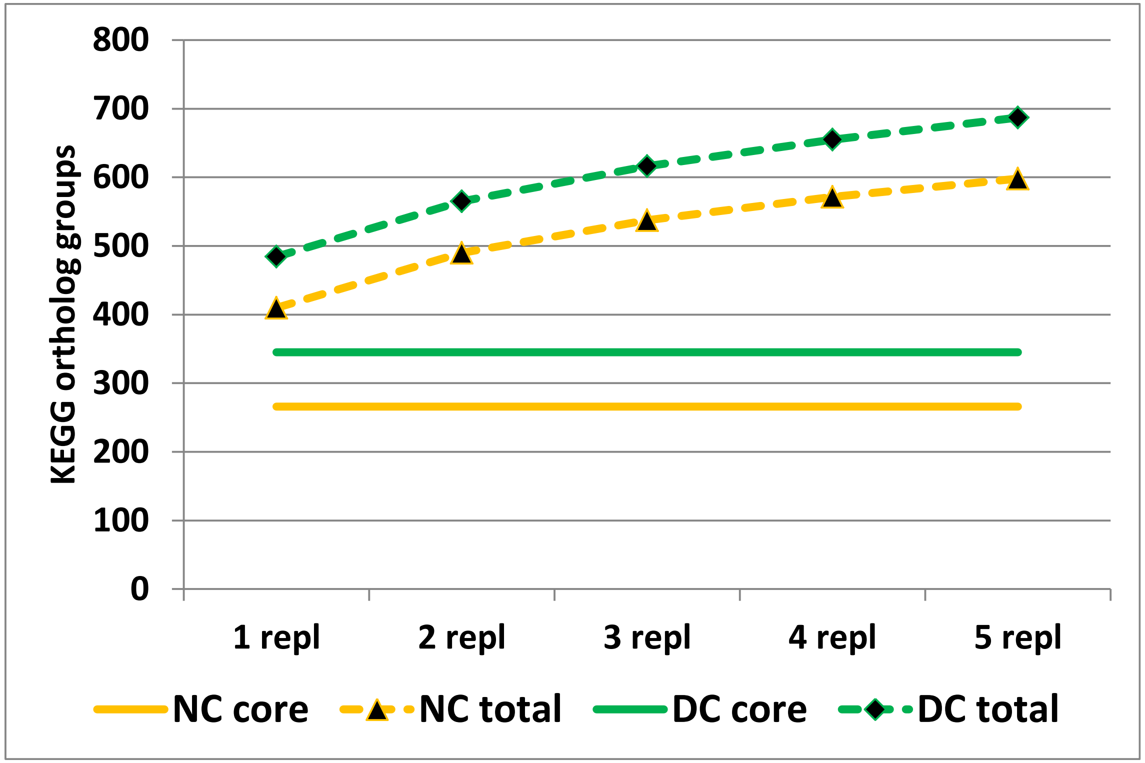


Figure S22. Bar graph illustrating the microbial KEGG ortholog groups identified with a mean abundance higher than 1% in NC and/or DC.

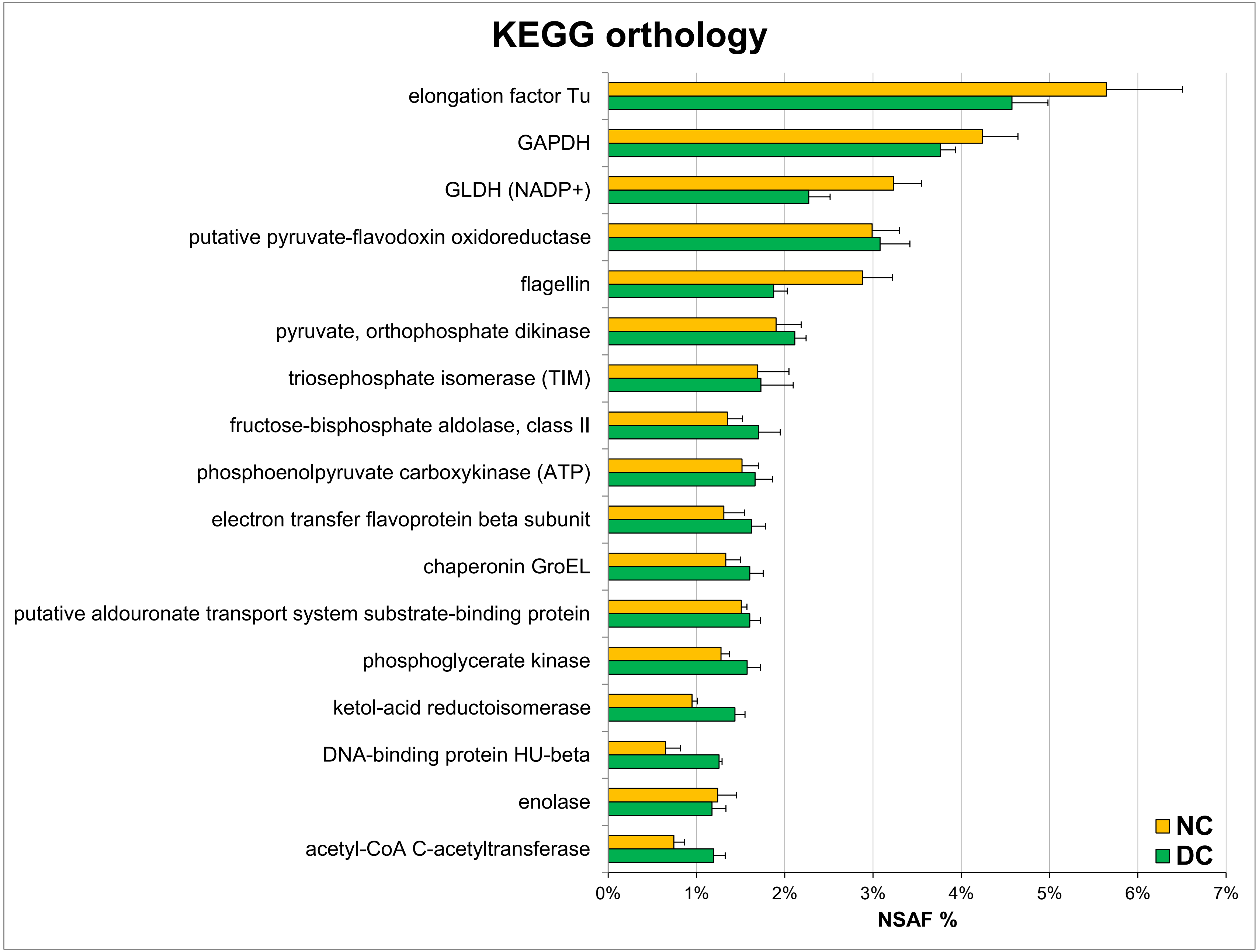


Figure S23. Bar graph illustrating the differential microbial KEGG groups ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

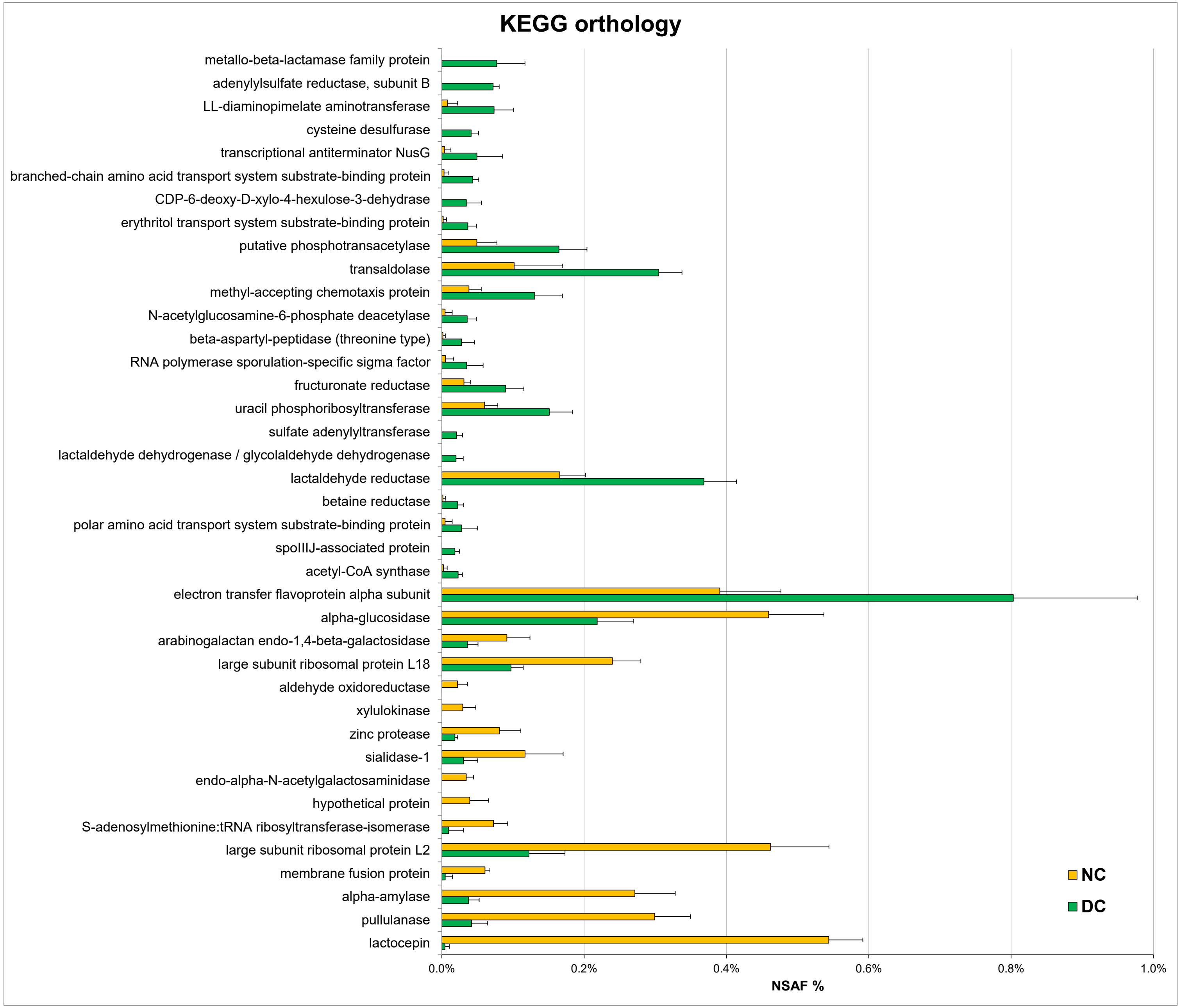


Figure S24. Bar graph illustrating the combined microbial KEGG groups/phyla identified with a mean abundance higher than 1% in NC and/or DC.

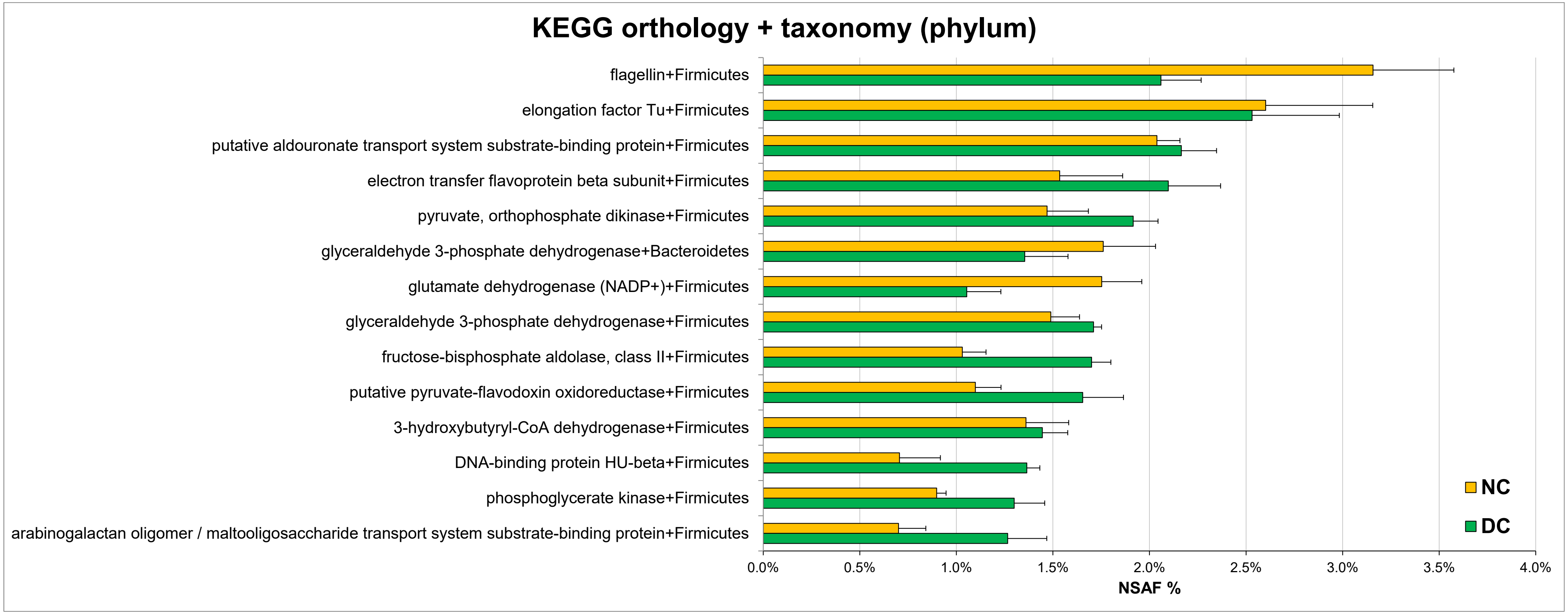


Figure S25. Bar graph illustrating the differential combined microbial KEGG groups/phyla ($p < 0.01$ and log ratio > 1.5), ordered by decreasing log ratio (DC/NC).

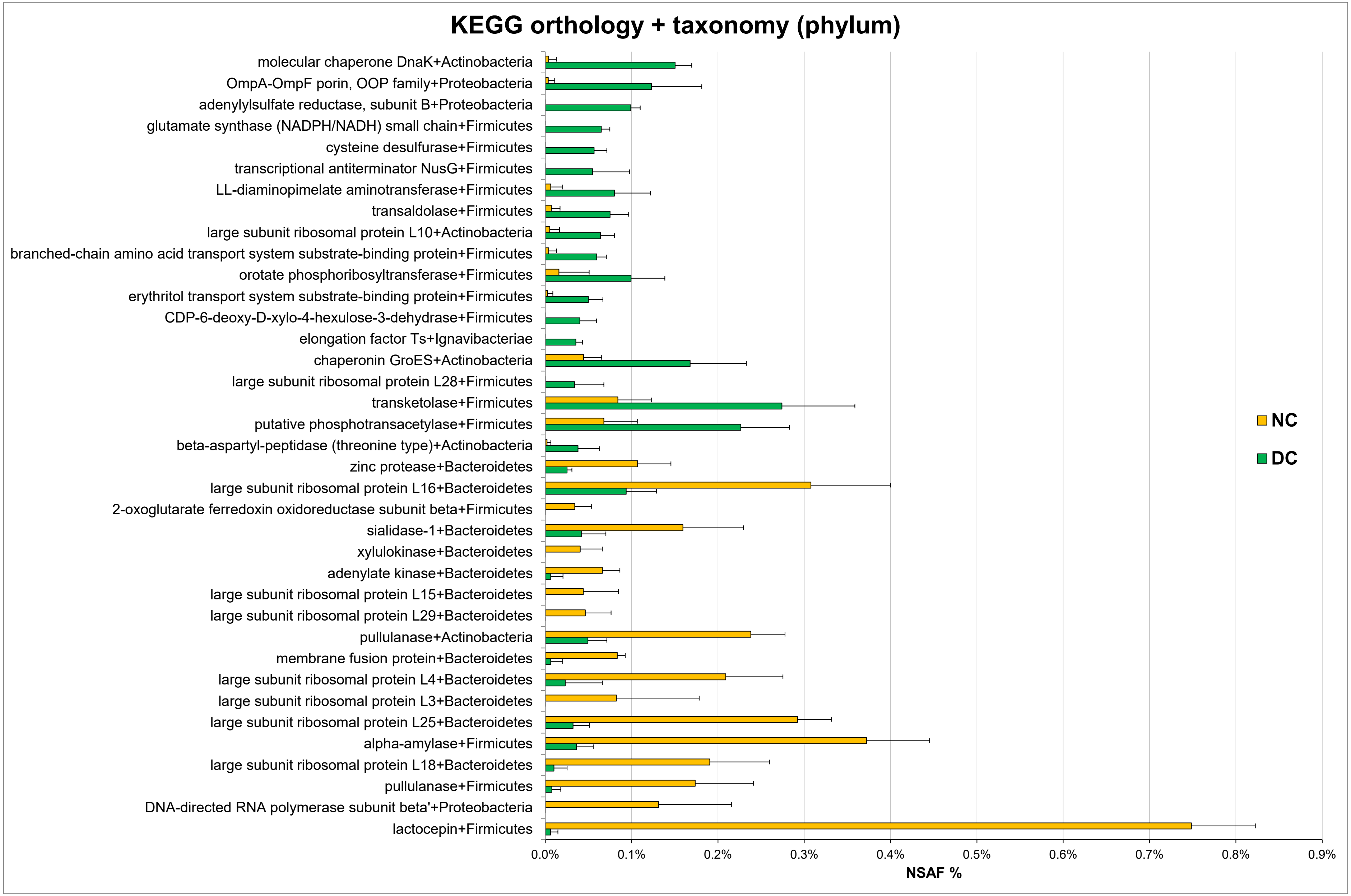


Figure S26. Bar graph illustrating the combined microbial KEGG groups/families identified with a mean abundance higher than 1% in NC and/or DC.

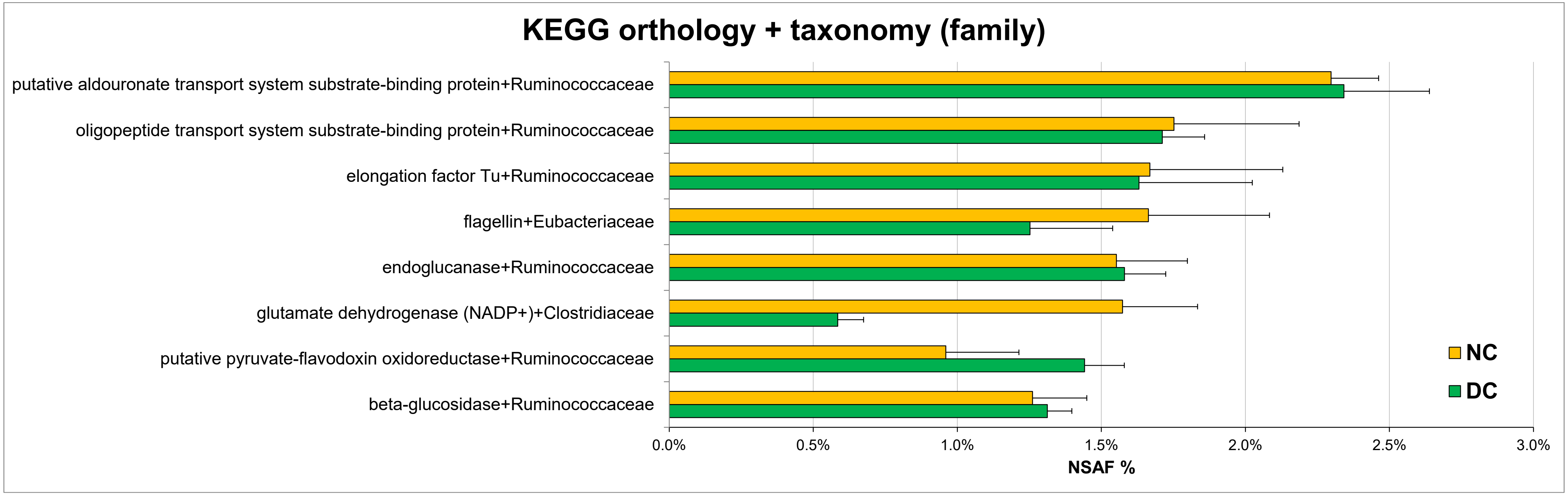


Figure S27. Bar graph illustrating the differential combined microbial KEGG groups/families (p < 0.01 and log ratio > 2), ordered by decreasing log ratio (DC/NC).

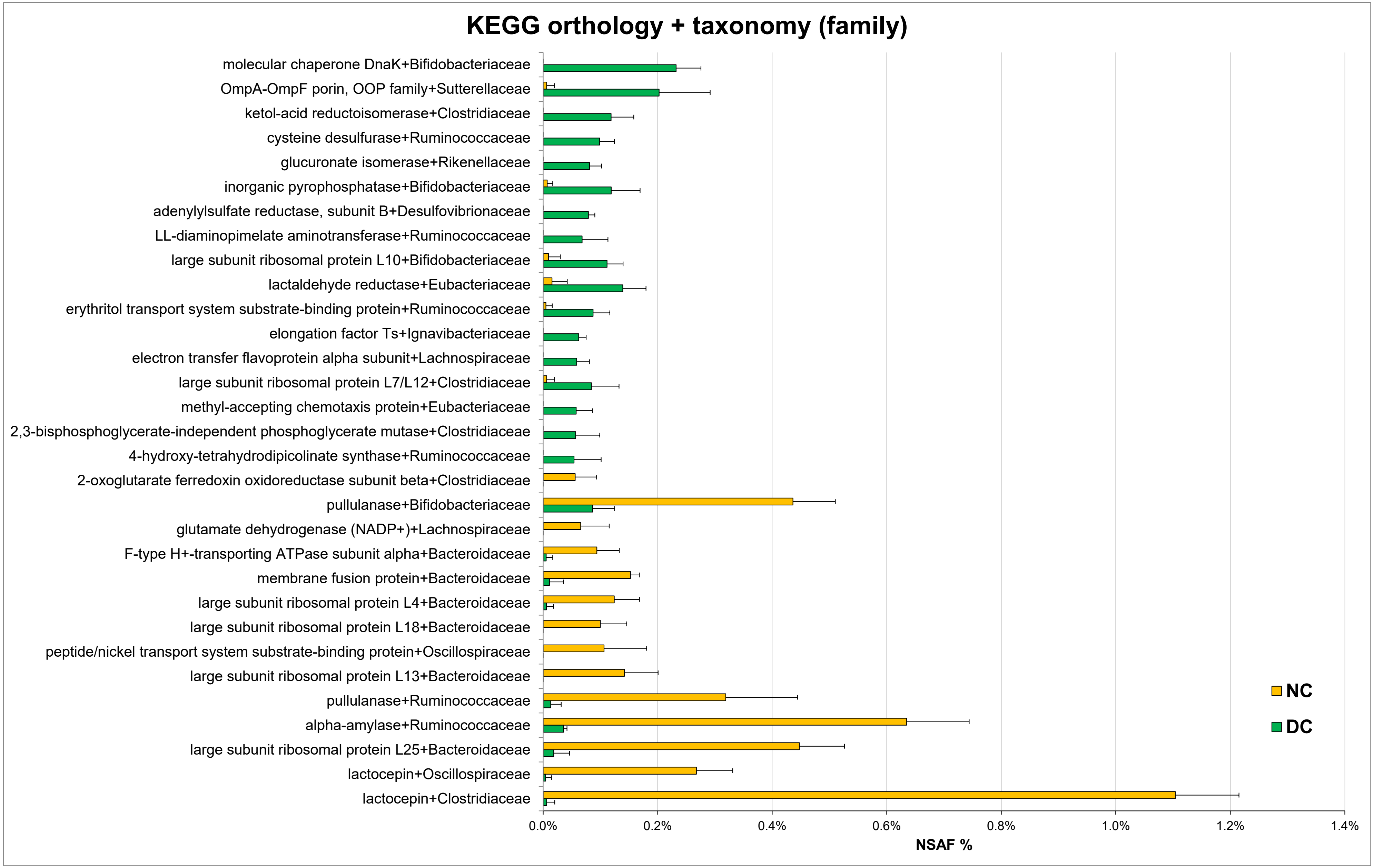


Figure S28. Histogram showing the percentage abundance of microbial proteins with predicted cytoplasmic, extracellular or membrane localization (from left to right), or containing one or more transmembrane domains (TMDs; last graph on the right).

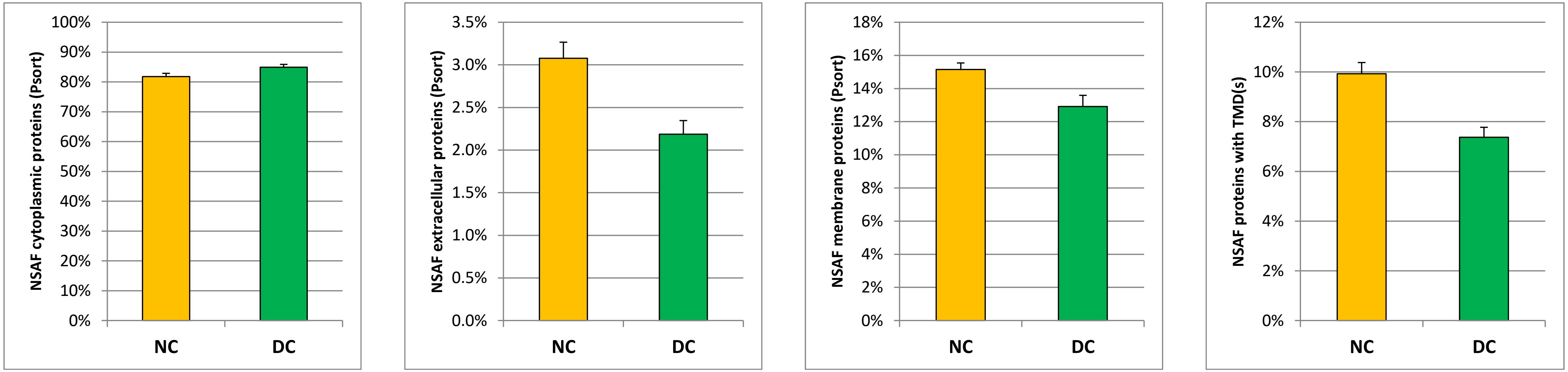


Figure S29. Cumulative number of host peptides detected in five replicate analyses (dashed lines), along with the ‘core’ identifications (common to all replicates, solid lines).

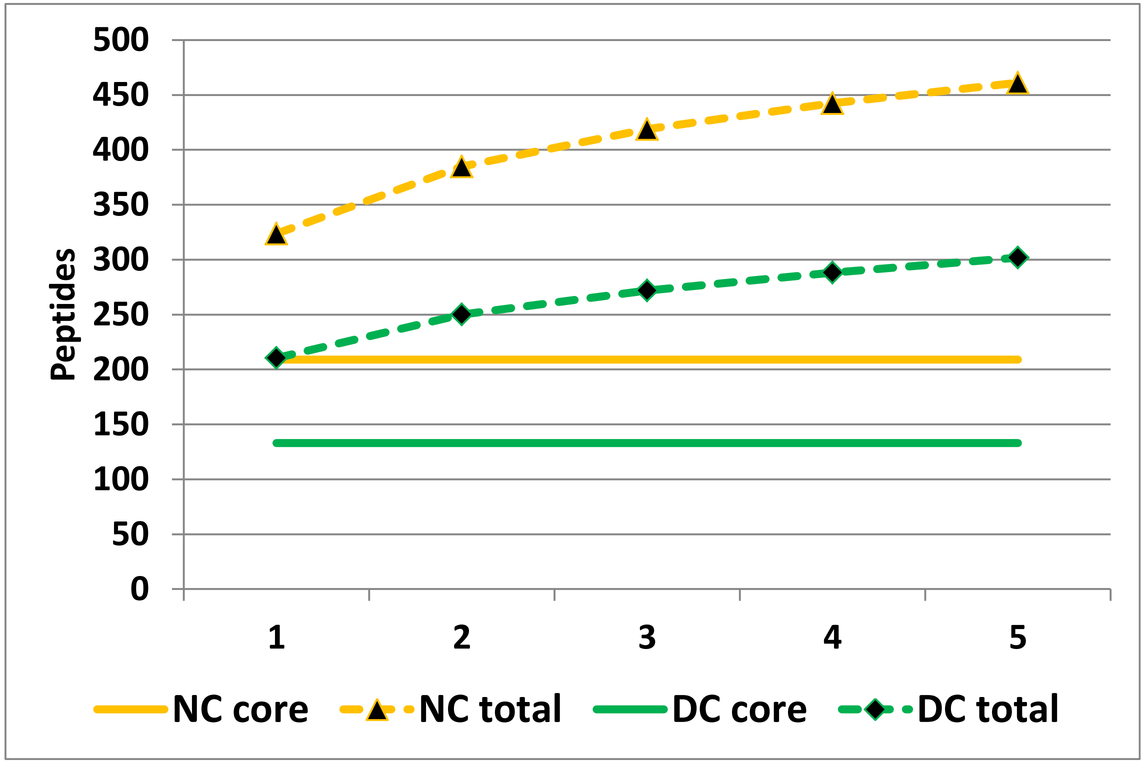


Figure S30. Bar graph illustrating the differential host GO biological processes ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

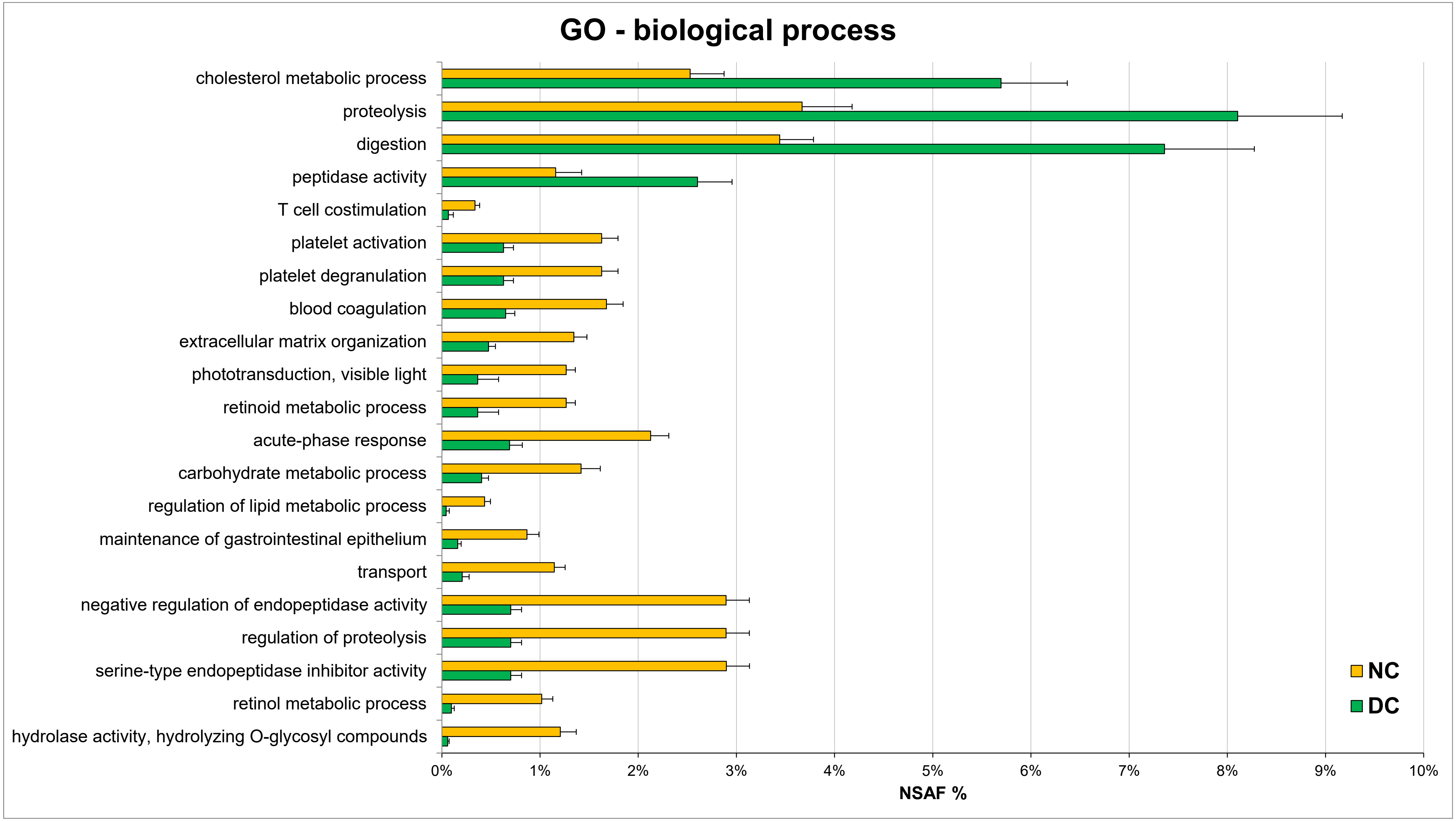


Figure S31. Bar graph illustrating the differential host UniProt protein families ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).

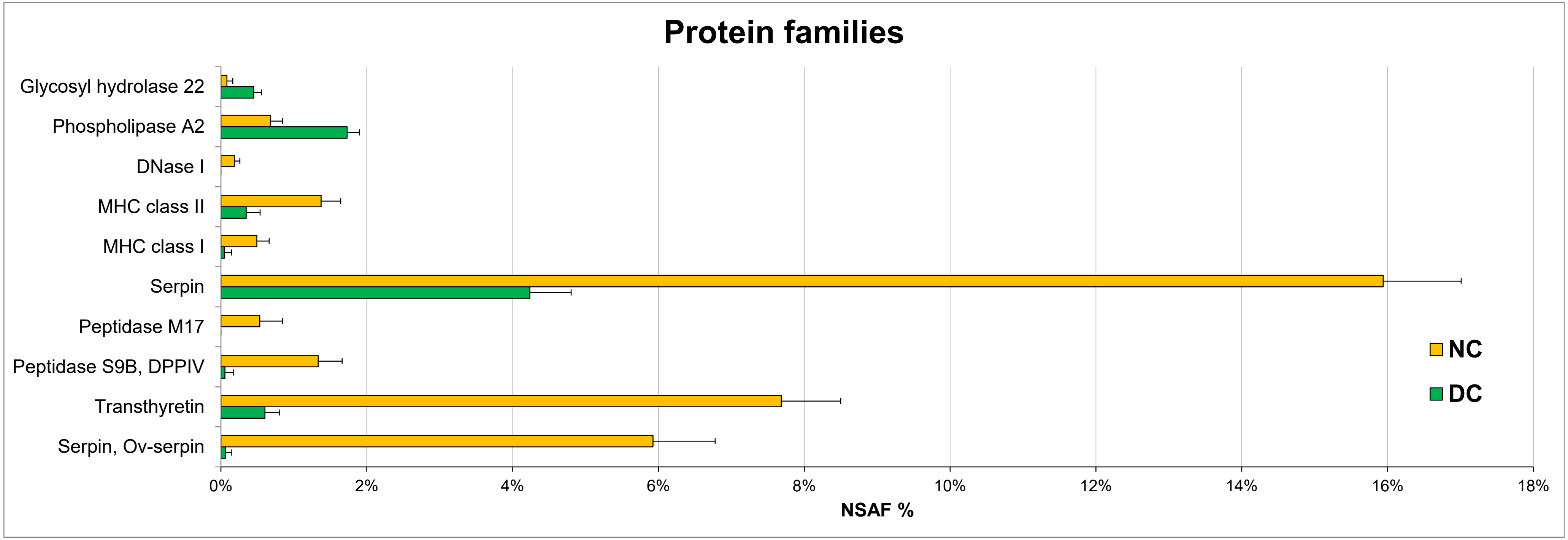
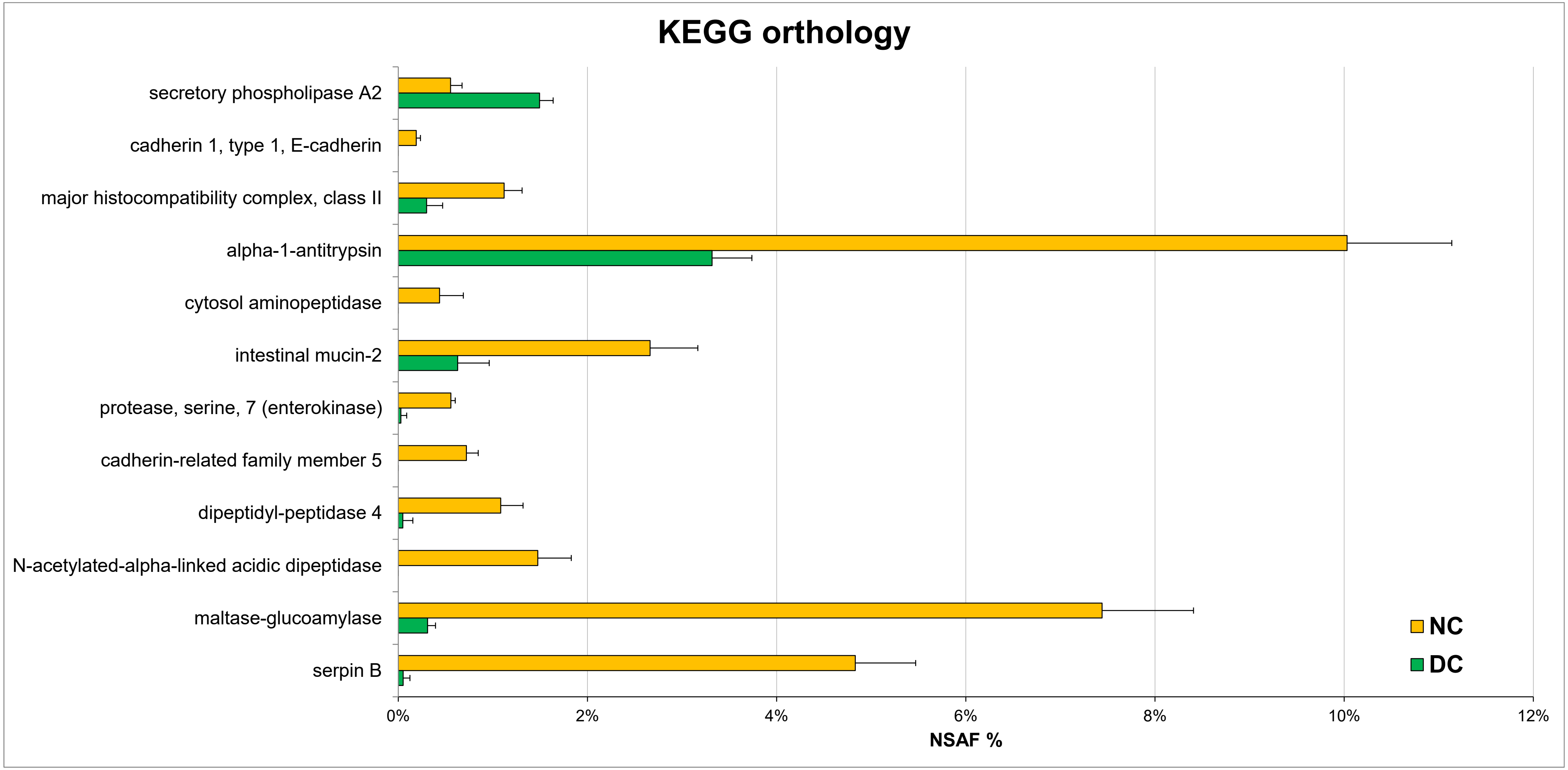


Figure S32. Bar graph illustrating the differential host KEGG groups ($p < 0.01$ and log ratio > 1), ordered by decreasing log ratio (DC/NC).


KEGG orthology
secretory phospholipase A2
cadherin 1, type 1, E-cadherin
major histocompatibility complex, class II
alpha-1-antitrypsin
cytosol aminopeptidase
intestinal mucin-2
protease, serine, 7 (enterokinase)
cadherin-related family member 5
dipeptidyl-peptidase 4
N-acetylated-alpha-linked acidic dipeptidase
maltase-glucoamylase
serpin B
NSAF %
NC
DC