THE UNIVERSITY OF CHICAGO

THE METAPROTEOMIC ANALYSIS OF ARCTIC SOILS WITH NOVEL BIOINFORMATIC METHODS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF THE GEOPHYSICAL SCIENCES

BY

SAMUEL MILLER

CHICAGO, ILLINOIS

DECEMBER 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I sincerely thank my wife, Alissa Miller, for her unwavering support and companionship during the last part of my Ph.D.; my parents, Jan and Richard Miller, for selflessly giving me the opportunities to pursue and complete a doctorate, and for being my cheerleaders and confidants; and my mother- and father-in-law, Mary and Gene Sherman, for their deep kindness and generosity. I am extremely grateful to the University of Chicago and the Department of the Geophysical Sciences for enabling my research. Prof. Jacob Waldbauer facilitated and backed this work, providing invaluable feedback and recommendations through the process; Prof. Michael Foote was a key guide and reviewer during the last part of the dissertation; Prof. Maureen Coleman and Prof. David Archer were a critical sounding board on my committee. Other members of the Department, past and present, also played an integral role: my friend, Dr. Gerard Olack, always dispensed insightful suggestions; Dr. Albert Colman encouraged, promoted, and supported the first part of my Ph.D. research; Mark Anderson shared his wealth of knowledge; students in the Waldbauer and Coleman labs, in particular, were generous and sound in their advice.

I dedicate this dissertation to the memory of my grandparents, Iris and Marshall Miller and Dorothy and Richard Miller, Sr., who I still miss.

ABSTRACT

Microbes control the decomposition of soil organic matter, a key biogeochemical process significant to global climate. The complex chemistry of soils and the great diversity of microbial strains with flexible metabolic capabilities have impeded the elucidation of degradation pathways from plant tissues to greenhouse gases. A mechanistic understanding of soil processes can improve models used to predict the fate of vast quantities of carbon stored in Arctic soils. Arctic warming is accelerating microbial decomposition but also increasing plant biomass, counteracting carbon loss. Floras with a significant nonvascular component are being replaced by floras dominated by larger and woodier plants. The changing vegetation may mediate the effects of warming on soil microbial activity through interactions with roots and the composition of plant detritus.

Metaproteomics is a promising approach for studying soil processes, since proteins catalyze key biogeochemical transformations. I collected soil cores from major floral ecotypes in the area of Toolik Field Station, Alaska and extracted proteins for metaproteomic analysis. To overcome impediments to the routine application of proteomics to complex samples, I developed novel bioinformatic methods to analyze protein mass spectrometry data. The standard database search method of assigning amino acid sequences to peptide mass spectra requires a tailored reference database of sequences that may be present in the proteomic dataset. Environmental metaproteomes may lack appropriate reference databases, especially in the absence of paired metagenomes. As an alternative to database search, sequences can be deduced directly from mass spectra, a computationally challenging approach known as de novo sequencing. To improve the low accuracy of de novo sequences predicted by existing algorithms, I created post-processing software called *Postnovo*, which rescores and reranks sequences from multiple input

algorithms using newly calculated metrics. I demonstrated that Postnovo improves the yield of accurate de novo sequences by about an order of magnitude and predicts the false discovery rate of the results. Postnovo extends the applicability of de novo sequencing, which is currently used with relatively simple samples such as monoclonal antibodies. Furthermore, I characterized the minimum length of environmental de novo sequences necessary for functional annotation from large reference databases.

I also employed database search methods to identify peptide sequences in my metaproteomic datasets, using Alaskan soil metagenomes and metatranscriptomes published in other studies as a reference database. To link proteins to taxa, I identified bins of metagenomic sequences representing the major bacterial groups known from 16S rRNA surveys of microbial taxonomic diversity. The challenge of utilizing the full information content of the reference nucleotide datasets – including sequence reads, unbinned contigs, and binned contigs – led me to create software called *ProteinExpress*. ProteinExpress increases the number of protein identifications and the quality of protein annotations from complex metaproteomes. Additionally, I constructed a classification system relating protein functional annotation terms from the eggNOG database to protein "Functional Groups" of biogeochemical significance.

Metaproteomic analyses revealed key processes in the soils, patterns of resource partitioning between major taxa, and changes associated with increasing plant biomass. Microbial activity in the rhizosphere appears to be distinct from activity in the bulk soil, with groups such as Rhizobiales strongly interacting with roots and other groups such as Acidobacteria dominating the degradation of plant cell wall polymers. Rhizospheric groups concentrate on the acquisition of small, soluble compounds, especially simple sugars likely exuded from roots, and most strongly express transporters for nitrogenous compounds,

potentially due to severe nutrient limitation in the proximity of roots. Acidobacteria degrade relatively labile polysaccharides, such as hemicelluloses, Actinobacteria depolymerize cellulose, and Burkholderiaceae cleave aromatics including lignin. These ecophysiological findings run counter to the expectation that major groups of heterotrophic soil bacteria are generalists without strong preferences for carbon and nutrient resources. Acidobacteria are the most active group across floral ecotypes, given their high expression of ribosomal proteins and other core functions, yet the activity of rhizospheric bacteria increases from low to high biomass floras. This suggests that further Arctic warming will be accompanied by a shift in soil microbial activity toward groups engaged in both mutualistic and competitive interactions with plants.

I. INTRODUCTION

I.A. IMPETUS FOR RESEARCH

Soils contain three times as much C in C_{org} as is found in atmospheric CO_2, and half of this pool is located in circumarctic permafrost-affected soils.[1] The mineralization of Arctic soil C in a warming climate is a positive feedback to global warming, with the potential for net release of up to ~200 Pg of C as CO_2 and CH_4 by the end of the century – nearly 20 times current annual anthropogenic CO_2 emissions.[2] Although soil microbes catalyze key biogeochemical processes, including C efflux, basic knowledge of microbial processes remains limited due to the complexity of both the microbial communities and the organic transformations they control. For example, soil microbial communities are strongly influenced by vegetation, yet the response of communities to the rapid greening of the Arctic is unknown beyond shifts in the relative abundances of poorly understood taxa.[3,4] The incorporation of simple representations of microbial activity into larger biogeochemical models of soil organic matter cycling improves the predictive accuracy of these models, so the mechanistic understanding of in situ processes has the potential to greatly increase the power of models.[5]

Proteomic methods provide a promising approach to the characterization of microbial processes in natural environments, as the detection of proteins is a more direct proxy for catalyzed reactions than the detection of genes or transcripts that only have the potential to produce an encoded protein under any given environmental conditions.[6] Since proteins catalyze specific reactions, they indicate the occurrence of reactions involving compounds that are often difficult to measure in situ. Protein-based measurements can also address basic questions concerning soil microbial ecology, such as the taxonomic distribution of the expression of pathways integral to the cycling of ubiquitous organic substrates. It is often hypothesized that

these pathways are expressed by diverse strains that quickly acquire the necessary genotypes through horizontal gene transfer,[7,8] but recent studies with stable isotopes and multi-omics methods have called that genome-based proposition into question.[9,10]

Methodological barriers have prevented the widespread adoption of metaproteomics alongside nucleotide-based methods in environmental microbiology. Outstanding issues include the assignment of accurate amino acid sequences to peptide mass spectra in complex samples[11] and the accurate taxonomic and functional annotation of peptide sequences.[12] Chapter 1 describes an algorithm that I developed to find accurate "de novo" peptide sequences generated without the need for reference data containing the sequences of interest. Chapter 2 explores the assignment of accurate taxonomic and functional annotations to de novo sequences. Chapter 3 introduces methods for the integration of metaproteomics data with other omics reference data and uses these methods to characterize Arctic soils. I recovered large amounts of information from my Arctic metaproteomic samples, including the major microbial processes occurring in the soils, the taxa performing these processes, and changes in microbial biogeochemistry with floral ecotype. This allowed me to test hypotheses regarding the major processes of organic matter degradation in Arctic soils, niche partitioning by heterotrophic microbes in soils, and the projected response of soil processes to floral growth and turnover associated with long-term warming.

I.B. PROTEOMICS BACKGROUND

Recent advances in proteomics, or the large-scale study of proteins in a sample,[13] have enabled high-resolution and high-throughput peptide sequencing and protein identification. Proteomics has proven indispensable for describing the inventory of proteins expressed in

biological systems, the post-translational modifications that regulate protein function, protein interactions in macromolecular assemblies, and statistical links between genotypes, "proteotypes" and disease states.[14] Breakthroughs in the nondestructive ionization of biological macromolecules and the sensitive and accurate mass analysis of these ions have facilitated the collection of large amounts of data by tandem mass spectrometry. The assignment of amino acid sequences to mass spectra and the association of sequences with actual proteins is a locus of methodological research that has been critical to the success of proteomics. Standard methods of sequence assignment are poorly suited to a variety of proteomic samples of interest, including the complex proteomes commonly found in natural systems such as soil,[6] water,[12] the human gut,[15] saliva,[16] snake venom,[17] pitcher plant fluid,[18] and fossils.[19–21] The *Postnovo* algorithm described in Chapter 1 was developed to overcome the sequence assignment challenge.

I.B.1. TANDEM MASS SPECTROMETRY

Most proteomics experiments use tandem mass spectrometry to measure molecular masses. The high accuracy of modern instruments permits the discrimination of molecules differing by less than one atomic mass unit (u, equivalent to the Dalton, Da) and thereby the discrimination of amino acids differing in mass. The three core components of any mass spectrometer are the ion source, the mass analyzer for filtering ions by mass-to-charge ratio (*m/z*), and the detector.[22] Tandem mass spectrometry involves multiple stages of mass analysis in quick succession on a single species of ion, with the species measured whole and subsequently selected and fragmented for further analysis. Multistage analysis (MS^n, $n \geq 2$) is critical for molecular sequence determination.

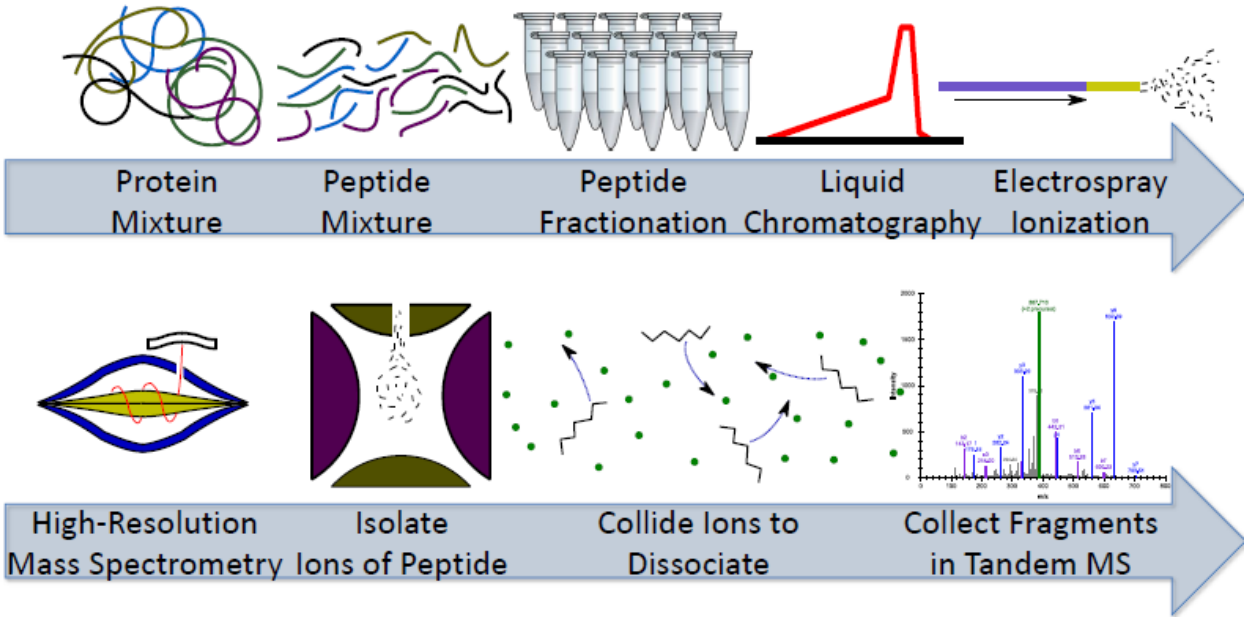

Figure I.1. Bottom-up proteomics workflow
A standard "bottom-up" proteomics workflow up to the point of spectrum sequencing. Proteins extracted from a sample are digested by a protease to peptides. (Mass spectrometry of a protein digest is "bottom-up" proteomics, whereas the analysis of intact proteins is "top-down" proteomics.) Liquid chromatography separates peptides by chemical properties. Electrospray ionization introduces peptide ions into the mass spectrometer. The high-resolution Orbitrap mass analyzer measures the *m/z* value of intact "precursor" peptides. Abundant precursors are fragmented and the peptide fragments measured, here on a lower-resolution linear ion trap (LIT) mass analyzer (the four rods of the quadrupole LIT appear in cross-section). Fragmentation spectra are produced from each selected precursor. (Figure courtesy of Prof. David Tabb.)

Common "bottom-up" proteomics experiments for the characterization of all proteins in a sample involve the digestion of proteins to shorter peptide molecules that are more tractable for high-throughput mass spectrometry (Figure I.1). Digestion is performed with another enzyme called a protease, which has the key property of predictably cleaving peptide bonds at specific amino acids. The commonly used protease, trypsin, cleaves peptide bonds on the carboxyl side of lysine and arginine. Purified and concentrated peptides are separated by high-pressure liquid chromatography (HPLC), allowing the continuous elution of simplified peptide fractions onto the mass spectrometer. Peptides are ionized and transferred to the gas phase by "soft," or

nondestructive, ionization. The principal soft ionization technique in most modern instruments is electrospray ionization,[23] the efficiency of which is enhanced by nL/min flow rates through a stretched glass capillary emitter.[24] Electrospray ionization occurs when a high voltage causes electrostatic repulsion within the liquid at the tip of the capillary, resulting in an aerosol jet of evaporating solvent droplets flowing into the lower potential vacuum inlet of the mass spectrometer.

The mass of the intact peptide ion is measured in the first stage of tandem mass spectrometry (MS1). The Fourier transform Orbitrap mass analyzer, first commercialized in 2005, is now commonly used to measure intact "precursor" peptides.[25,26] Qualities of the Orbitrap that have resulted in its widespread adoption are its resolving power (ability to separate peaks at different *m/z* values), dynamic range (linearity of the relation between ion abundance and signal), measurement speed, and independence from other stages of mass spectrometry due to ion storage and injection from a "C-trap."[27] The resolution of the Orbitrap mass analyzer is 10^{-7}-10^{-6} ppm, two to three orders of magnitude higher than the previous generation of ion trap instruments. For a peptide of mass 10,000 Da, this corresponds to a resolution of 10^{-3}-10^{-2} Da. Second stage peptide fragment analysis is often conducted with a linear-ion trap (LIT) mass analyzer.[28] The application of a resonant radio frequency potential to the LIT electrodes causes trapped ions to collide with a neutral gas. MS2 fragment peptides resulting from this collision-induced dissociation (CID) are ejected radially from the LIT and detected by a conversion dynode. Advantages of measuring fragment ions on the LIT rather than the higher-resolution Orbitrap include sensitivity (low detection limit) and acquisition speed – as precursor ions revolve around the spindle of the Orbitrap over a period of ~1 s, daughter fragment ions can be analyzed simultaneously over ~0.1 s. The "data-dependent acquisition" of fragmentation spectra

allows the most abundant ions in an Orbitrap cycle to be quickly identified and selected for fragmentation on the LIT. A single peptide precursor yields one peak in the MS1 spectrum and a number of peaks in its corresponding MS2 fragmentation spectrum. MS2 spectra generate gigabytes of raw data, presenting a challenge for computational analysis.

MS2 fragmentation spectra are critical for high-throughput sequencing by mass spectrometry, while high-resolution MS1 precursor spectra constrain sequence assignments to those that closely match the precursor mass. Ideally, the fragmentation spectrum of a peptide contains two clear peaks for each peptide bond. CID fragmentation occurs predominantly at the N-C peptide bond linking amino acid residues, resulting in two daughter peptide fragments from each side of the bond called b- and y-ions (Figure I.2). "Prefix" b-ions come from the N-terminal end of the peptide, while "suffix" y-ions come from the C-terminal end. It is extremely rare for the complete set of b- and y-ions to be represented in the fragment "peak ladder" due to the low abundance of precursors and the influence of precursor charge state and amino acid composition on cleavage site and probability.[29] When there are gaps in the peak ladder, competing sequence hypotheses can be evaluated by the probability of cleavage between different pairs of amino acids.

In an ideal peak ladder, the mass difference between each successive b- or y-ion peak corresponds to the mass of a single amino acid (Figure I.3), with the exception of the isobaric (equal mass) amino acids, leucine and isoleucine. The mass of a peak can be determined from the measured *m/z* value through the pattern of natural isotopic peaks adjacent to the main peak – when these peaks differ by an *m/z* value of 1/*z*, the fragment has a charge of *z*. In practice, peak deconvolution is often impossible due to low side-peak intensity, necessitating the determination

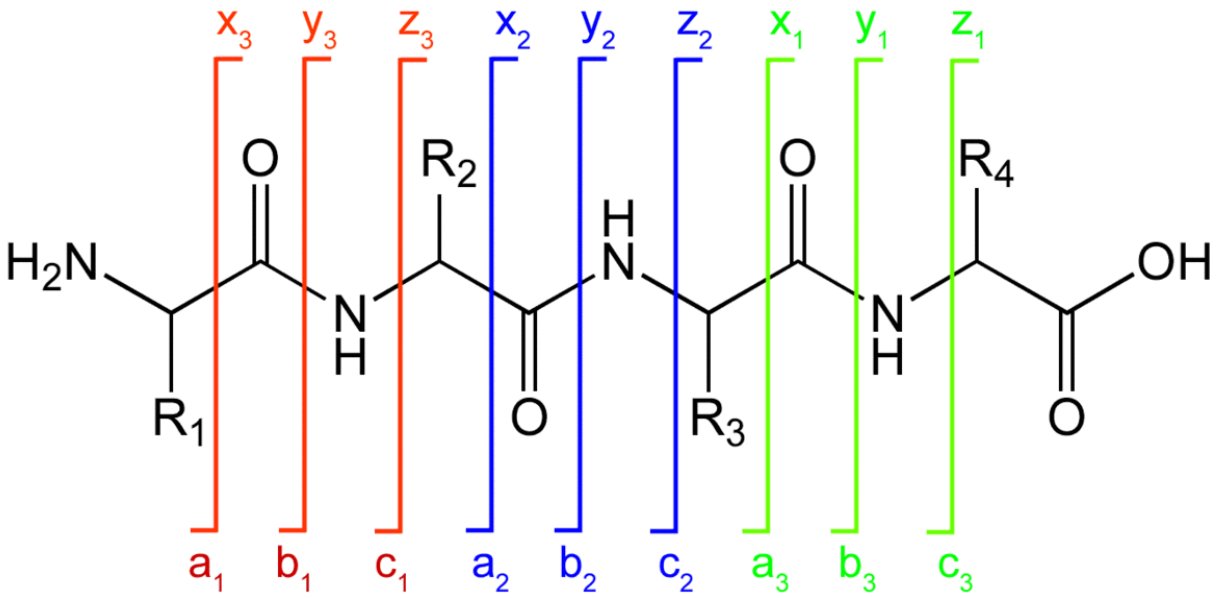

Figure I.2. Fragmentation sites on the peptide backbone
b- and y-ions are the most common fragments produced by collision-induced dissociation (CID). These ions occur at the N-C peptide bond linking amino acid residues. The numerical subscript indicates the position of the fragmentation site relative to the N-terminal (left) amino acid for "prefix" fragment ions and C-terminal (right) amino acid for "suffix" ions. (Figure from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Peptide_fragmentation.gif.)

of charge state by comparing different fragment peaks or by searching for multiple peptide sequence assignments corresponding to possible charge states (CID fragment ions typically have charges of +1, +2, or +3).[30] Given b- or y-ion evidence for each peptide bond and assuming that charge state can be deduced from the spectrum, one can manually reconstruct the sequence of a peptide from its fragmentation spectrum by subtracting consecutive peak masses and uniquely assigning the difference to amino acids. When the full set of b- and y-ions is represented, each series of ions provides redundant evidence for the sequence. Missing peaks and low resolution make the sequence deduction problem much more difficult. The absence of b- or y-ion peaks at a certain peptide bond is not necessarily fatal, as the summed mass of the two amino acids (dipeptide) straddling the bond can still be calculated from other fragment peaks. However, sets

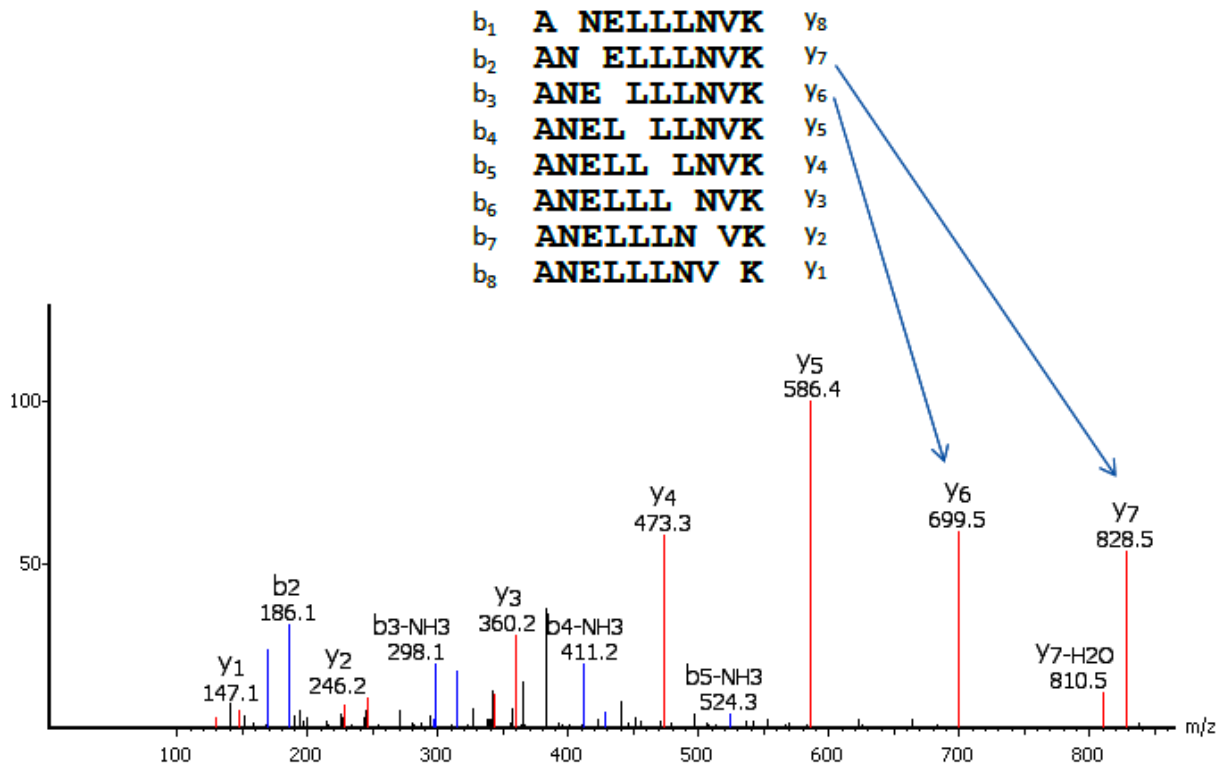

Figure I.3. Fragmentation spectrum interpretation
The mass difference between two peaks in a series (b_i, b_{i+1} or y_i, y_{i+1}) corresponds to the mass of an amino acid. The 129.0 Da mass difference between the y_6 and y_7 ion peaks in this example is about equal to the expected mass difference expected for glutamate (E), the amino acid present in y_7 but not y_6. The 129.0 Da mass difference between the complementary b_2 and $b_{3\text{-}NH3}$ peaks confirms this amino acid assignment (after correcting for the mass of the NH_3 moiety missing from the b_3 fragment). This spectrum is relatively complete, with b- or y-ion evidence for every peptide bond. (Figure courtesy of Bioinfor: http://www.bioinfor.com/wp-content/uploads/2016/11/denovo-screenshot.png.)

of amino acids can have the same mass, complicating the identification of the dipeptide. There are 12 pairs of isobaric dipeptides (e.g., AD and EG), 2 pairs of isobaric monopeptides and dipeptides (e.g., N and GG), 4 pairs of nearly isobaric dipeptides and 1 nearly isobaric monopeptide-dipeptide pair. Even when a peak ladder is complete, one could suspect that what looks like an N, for example, is actually GG due to a missing cleavage between the two glycines!

The second fundamental problem of incomplete coverage is the inference of the order of amino

acids. For example, one might calculate that alanine and cysteine must be present in a peptide sequence despite the lack of evidence for cleavage between the two amino acids, yet one would not know the order in which they occur – AC and CA are equally probable from the calculation. Subtle problems associated with the low resolution of LIT data further complicate sequence deduction. b-ions can be mistaken for y-ions, leading to erroneous amino acid assignments and inversions in the sequence.[31,32]

I.B.2. METHODS OF AUTOMATIC SEQUENCE ASSIGNMENT

The fundamental problems in the manual deduction of peptide sequences are also challenges for the automatic approaches essential to proteomics. These approaches fall in three categories: database searching, spectral libraries, and de novo sequencing. Although de novo sequencing is the most similar method to manual sequencing, this "naïve" approach has been superseded by database searching for the last two decades due to a vast increase in genomic data.

A database search compares a mass spectrum to a database of candidate peptides predicted from gene sequences. Candidate peptides are generated in silico from longer sequences using the cleavage pattern of the protease used in one's proteomic experiment (cleavage after K and R by trypsin, for example). The database search algorithm creates theoretical fragmentation spectra for each candidate peptide, filters these by the high-resolution mass of the query peptide from the experiment, and computes the cross-correlation of each theoretical spectrum with the query spectrum using the fast Fourier transform.[33] These comparisons produce a set of peptide-spectrum matches (PSMs), and a post-processing algorithm ranks PSMs by cross-correlation and other factors.[34] The false discovery rate (FDR) of a set of PSMs is often evaluated by a "target-decoy" search, in which the target protein database is reversed and used a decoy database for

finding random PSM matches (the null hypothesis).[35] Database searching is most appropriate for proteomes from well-constrained samples with genomic reference data, such as model organisms or human tissues.[11,36] Factors including the selection of the database, the parameterization of the search algorithm, and the effectiveness of the target-decoy approach for FDR estimation differ between the types of simple samples used to validate database search algorithms and complex, uncharacterized samples from natural environments or from organisms lacking a reference genome.

The identification of experimental spectra using spectral libraries is similar to database search in that PSMs are determined by spectral similarity calculations.[37] A spectral library is a large database of spectra with confident peptide assignments. Experimental spectra are compared to reference spectra with the same precursor mass. The infrequent use of spectral libraries owes to the inherent drawback that previously identified spectra are required to identify experimental spectra – an issue that is magnified in proteomes from complex, uncharacterized samples.

De novo sequencing differs greatly from both database searching and spectral libraries, as this approach does not compare experimental spectra to reference spectra. To use the terms of graph theory, de novo sequencing algorithms treat fragment peaks as nodes in a spectrum graph and possible amino acid assignments as edges connecting the nodes, a process akin to manual sequence assignment.[38] Sequencing is susceptible to all of the aforementioned pitfalls of manual interpretation stemming from missing or weak fragment peaks and insufficient resolution. De novo sequencing tools employ different scoring algorithms to find and rank sequence candidates for a spectrum. Scoring algorithms are trained with high confidence PSMs; the selection of this training dataset and the set of scoring features affects the rigor of de novo sequencing on unseen spectra. Some algorithms assign a probability score to each amino acid or sequence prediction,[39–

[41] but algorithm overfitting to specific training data reduces the generalizability of the score. The widespread use of de novo sequencing has been stymied by the difficulty of handling missing peaks in fragmentation spectra and the poor generalizability of algorithms, preventing reliable FDR estimation for de novo predictions.[42] The bases of the PepNovo+ and Novor de novo sequencing algorithms are explained below, as these algorithms are leveraged by the Postnovo algorithm introduced in Chapter 1.

I.B.2.i. DE NOVO SEQUENCING BY PEPNOVO+

PepNovo+ generates de novo sequence candidates from a spectrum graph, where each peak is a node and the nodes are connected by edges corresponding to amino acid masses.[43] The strength of each node is scored by the likelihood ratio of the probability that the peak is produced by CID fragmentation to the probability that it is caused by a random process. The probability of CID fragmentation is calculated from a probability network that relates a suite of spectral features observed in training data (972 spectra) to the occurrence of a CID fragment. Predefined network connections are highly empirical, and connection probabilities depend on the training data. As an example of a connection, the designer of PepNovo+ states that "the probability of seeing a strong b-ion in the center of the peptide, given that there is a strong y-ion, is $P_{CID}(I_b = high | I_y = high; pos(m) = 2) = 0.36$, and it drops to 0.03, if instead of the strong y-ion, a weak y-ion is detected $(I_y = low)$" [where I is intensity (on the vertical axis) and $pos(m)$ is the region of the spectrum's horizontal m/z axis].[43] Sequence candidates generated by the spectrum graph/probability network algorithm are ranked by a machine learning predictive model,[44] which is an effective general strategy for using large numbers of informative, but not decisive, discrete and continuous features in de novo sequencing as well as mainstream database search tools.[45]

This PepNovo+ sequence ranking algorithm is an implementation of the RankBoost algorithm, and it uses many of the same features as the PepNovo+ probability network model in addition to a variety of other features such as the count of b- and y-ions in the spectrum and the amino acid composition of the sequence.[46]

I.B.2.ii. DE NOVO SEQUENCING BY NOVOR

Novor relies solely on decision tree machine learning models to find and score sequence candidates (Figure I.4).[41] Decision trees are the constituent elements of random forest models, such as the random forests used in Postnovo. Compared to other supervised classification and regression machine learning models, tree-based models are very flexible and fast, although they can be overfit to the training data. A decision tree is a graph of "nodes" and "leaves." Each node, starting at the origin node, splits to two other nodes or leaves, the term for the tips of the tree. Each node represents a binary decision rule that has been learned from training data and is applied for predictive purposes on the new data being analyzed. For example, if a decision tree is used to classify a peptide sequence as accurate or inaccurate, one feature might be peptide length, and a node might split sequences into those shorter than and those of at least 6 amino acids. The (terminal) leaves of a classification tree correspond to possible classification categories – here, accurate or inaccurate. If this length ≥6 node splits to two leaves, the data that goes to one leaf will be classified as accurate, while the data ending at the other leaf will be classified as inaccurate. Node rules are often chosen by maximizing the reduction in Gini impurity, equivalent to the increase in the homogeneity of the data classification due to the split.[47] If the training sequences that encountered the length ≥6 node were split into a group of sequences all labeled "accurate" and another group all labeled "inaccurate," and the sequences

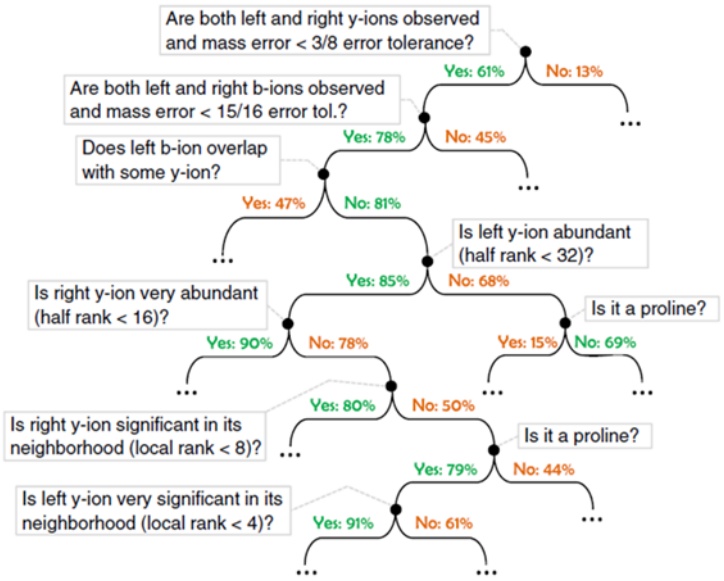

Figure I.4. Part of the Novor decision tree
Part of the decision tree used by Novor to calculate the probability of a residue in a de novo sequence (adapted from Ma, 2015).[41] The decision tree model was automatically learned from training data. Features of the training data include both discrete and continuous variables. Each node has a decision threshold from one feature, such as whether the residue is a proline or whether the right y-ion used to assign the residue is one of the 16 most abundant peaks. Residue data enters the tree from the top, first encountering the root node. If a node's decision criterion is met, the residue goes to the left, else it goes to the right. The probability on each edge represents the correctness probability of a residue after the decision split, as determined from the reserved validation dataset. The final correctness probability of a residue is assigned when a terminal leaf is reached (leaves are not shown in this part of the tree).

labeled accurate were all actually accurate, and the sequences labeled inaccurate were all actually inaccurate, then the node rule would have produced the maximum possible impurity reduction. Decision trees can be "bagged" by averaging multiple trees bootstrapped on subsets of the training data, and "bags of trees" can be turned into random forests by randomly selecting a subset of features to be considered for use at each node. For many problems, random forests are effective at optimizing classification accuracy and reducing overfitting of the model to the training data,[47] and thus Postnovo employs random forests.

Novor's first decision tree scoring function estimates the probability that a possible prefix (and suffix) mass corresponds to a real fragmentation site represented by real fragment peaks rather than noise. A prefix mass is the mass of the peptide spanning the N-terminus to the fragmentation site, and a suffix mass is the mass from the other end of the peptide. The Novor "fragmentation score" model has 72 features, an example being the relative intensity of the b-ion peak under consideration versus the most intense peak in the spectrum. In the context of the decision tree, one would expect that if this feature were used at a node, peaks that fall under a certain relative intensity threshold are less likely to be real fragmentation sites, and peaks that exceed the relative intensity threshold are more likely to be real sites. Fragmentation scores are then used as the basis of peptide predictions by a dynamic programming algorithm that is a variant of the spectrum graph approach.[48] The goal of the dynamic programming is to find the amino acid(s) that match the fragment mass, with a closer match further increasing the fragmentation score. After this process, Novor calculates a "residue score" for each amino acid prediction from another decision tree with 169 features in order to re-rank the sequence candidates for each spectrum (Figure I.4). In addition to the features factored into the fragmentation score, the residue score uses the identities of amino acid residues and the mass difference between peaks that define the single residue. Alternative amino acid assignments over small sections of the sequence are tested to boost the residue score and settle on a top de novo sequence prediction.

I.C. REFERENCES

(1) Jobbágy, E. G.; Jackson, R. B. The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. *Ecological Applications* **2000**, *10* (2), 423–436.

(2) Schuur, E. A. G.; Abbott, B. W.; Bowden, W. B.; Brovkin, V.; Camill, P.; Canadell, J. G.; Chanton, J. P.; Chapin, F. S.; Christensen, T. R.; Ciais, P.; et al. Expert Assessment of Vulnerability of Permafrost Carbon to Climate Change. *Climatic Change* **2013**, *119* (2), 359–374.

(3) Ackerman, D. E.; Griffin, D.; Hobbie, S. E.; Popham, K.; Jones, E.; Finlay, J. C. Uniform Shrub Growth Response to June Temperature across the North Slope of Alaska. *Environmental Research Letters* **2018**, *13* (4), 044013.

(4) Wallenstein, M. D.; McMahon, S.; Schimel, J. Bacterial and Fungal Community Structure in Arctic Tundra Tussock and Shrub Soils. *FEMS Microbiology Ecology* **2007**, *59* (2), 428–435.

(5) Wieder, W. R.; Grandy, A. S.; Kallenbach, C. M.; Bonan, G. B. Integrating Microbial Physiology and Physio-Chemical Principles in Soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) Model. *Biogeosciences* **2014**, *11* (14), 3899–3917.

(6) Bastida, F.; Jehmlich, N. It's All about Functionality: How Can Metaproteomics Help Us to Discuss the Attributes of Ecological Relevance in Soil? *Journal of Proteomics* **2016**, *144*, 159–161.

(7) Martiny, J. B. H.; Jones, S. E.; Lennon, J. T.; Martiny, A. C. Microbiomes in Light of Traits: A Phylogenetic Perspective. *Science* **2015**, *350* (6261), aac9323.

(8) Fierer, N. Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome. *Nature Reviews Microbiology* **2017**, *15* (10), 579–590.

(9) Pepe-Ranney, C.; Campbell, A. N.; Koechli, C. N.; Berthrong, S.; Buckley, D. H. Unearthing the Ecology of Soil Microorganisms Using a High Resolution DNA-SIP Approach to Explore Cellulose and Xylose Metabolism in Soil. *Frontiers in Microbiology* **2016**, *7*.

(10) Woodcroft, B. J.; Singleton, C. M.; Boyd, J. A.; Evans, P. N.; Emerson, J. B.; Zayed, A. A. F.; Hoelzle, R. D.; Lamberton, T. O.; McCalley, C. K.; Hodgkins, S. B.; et al. Genome-Centric View of Carbon Processing in Thawing Permafrost. *Nature* **2018**, *560* (7716), 49–54.

(11) Muth, T.; Kolmeder, C. A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; et al. Navigating through Metaproteomics Data: A Logbook of Database Searching. *Proteomics* **2015**, *15* (20), 3439–3453.

(12) May, D. H.; Timmins-Schiffman, E.; Mikan, M. P.; Harvey, H. R.; Borenstein, E.; Nunn, B. L.; Noble, W. S. An Alignment-Free "Metapeptide" Strategy for Metaproteomic Characterization of Microbiome Samples Using Shotgun Metagenomic Sequencing. *Journal of Proteome Research* **2016**, *15* (8), 2697–2705.

(13) Pandey, A.; Mann, M. Proteomics to Study Genes and Genomes. *Nature* **2000**, *405* (6788), 837–846.

(14) Aebersold, R.; Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **2016**, *537* (7620), 347–355.

(15) Gonzalez, C. G.; Zhang, L.; Elias, J. E. From Mystery to Mechanism: Can Proteomics Build Systems-Level Understanding of Our Gut Microbes? *Expert Review of Proteomics* **2017**, 1–4.

(16) Grassl, N.; Kulak, N. A.; Pichler, G.; Geyer, P. E.; Jung, J.; Schubert, S.; Sinitcyn, P.; Cox, J.; Mann, M. Ultra-Deep and Quantitative Saliva Proteome Reveals Dynamics of the Oral Microbiome. *Genome Medicine* **2016**, *8*, 44.

(17) Carregari, V. C.; Dai, J.; Verano-Braga, T.; Rocha, T.; Ponce-Soto, L. A.; Marangoni, S.; Roepstorff, P. Revealing the Functional Structure of a New PLA2 K49 from *Bothriopsis taeniata* Snake Venom Employing Automatic "de Novo" Sequencing Using CID/HCD/ETD MS/MS Analyses. *Journal of Proteomics* **2016**, *131*, 131–139.

(18) Hatano, N.; Hamada, T. Proteome Analysis of Pitcher Fluid of the Carnivorous Plant *Nepenthes alata*. *Journal of Proteome Research* **2008**, *7* (2), 809–816.

(19) Asara, J. M.; Schweitzer, M. H.; Freimark, L. M.; Phillips, M.; Cantley, L. C. Protein Sequences from *Mastodon* and *Tyrannosaurus rex* Revealed by Mass Spectrometry. *Science* **2007**, *316* (5822), 280–285.

(20) Schweitzer, M. H.; Zheng, W.; Organ, C. L.; Avci, R.; Suo, Z.; Freimark, L. M.; Lebleu, V. S.; Duncan, M. B.; Heiden, M. G. V.; Neveu, J. M.; et al. Biomolecular Characterization and Protein Sequences of the Campanian Hadrosaur *B. Canadensis*. *Science* **2009**, *324* (5927), 626–631.

(21) Pevzner, P. A.; Kim, S.; Ng, J. Comment on "Protein Sequences from *Mastodon* and *Tyrannosaurus rex* Revealed by Mass Spectrometry." *Science* **2008**, *321* (5892), 1040; author reply 1040.

(22) Bleakney, W. The Mass-Spectrograph and Its Uses. *American Journal of Physics* **1936**, *4* (1), 12–23.

(23) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64–71.

(24) Wilm, M.; Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry* **1996**, *68* (1), 1–8.

(25) Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry* **2000**, *72* (6), 1156–1162.

(26) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Analytical Chemistry* **2006**, *78* (7), 2113–2120.

(27) Zubarev, R. A.; Makarov, A. Orbitrap Mass Spectrometry. *Analytical Chemistry* **2013**, *85* (11), 5288–5296.

(28) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Müller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; et al. Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. *Molecular and Cellular Proteomics* **2012**, *11* (3), O111.013698.

(29) Kapp, E. A.; Schütz, F.; Reid, G. E.; Eddes, J. S.; Moritz, R. L.; O'Hair, R. A. J.; Speed, T. P.; Simpson, R. J. Mining a Tandem Mass Spectrometry Database To Determine the Trends and Global Factors Influencing Peptide Fragmentation. *Analytical Chemistry* **2003**, *75* (22), 6251–6264.

(30) Klammer, A. A.; Wu, C. C.; MacCoss, M. J.; Noble, W. S. Peptide Charge State Determination for Low-Resolution Tandem Mass Spectra. In *2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*; 2005; pp 175–185.

(31) Frank, A. M.; Savitski, M. M.; Nielsen, M. N.; Zubarev, R. A.; Pevzner, P. A. De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *Journal of Proteome Research* **2007**, *6* (1), 114–123.

(32) Budnik, B. A.; Nielsen, M. L.; Olsen, J. V.; Haselmann, K. F.; Hörth, P.; Haehnel, W.; Zubarev, R. A. Can Relative Cleavage Frequencies in Peptides Provide Additional Sequence Information? *International Journal of Mass Spectrometry* **2002**, *219* (1), 283–294.

(33) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society of Mass Spectrometry* **1994**, *5* (11), 976–989.

(34) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nature Methods* **2007**, *4* (11), 923–925.

(35) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nature Methods* **2007**, *4* (3), 207–214.

(36) Tanca, A.; Palomba, A.; Fraumene, C.; Pagnozzi, D.; Manghina, V.; Deligios, M.; Muth, T.; Rapp, E.; Martens, L.; Addis, M. F.; et al. The Impact of Sequence Database Choice on Metaproteomic Results in Gut Microbiota Studies. *Microbiome* **2016**, *4*, 51.

(37) Griss, J. Spectral Library Searching in Proteomics. *Proteomics* **2016**, *16* (5), 729–740.

(38) Allmer, J. Algorithms for the de Novo Sequencing of Peptides from Tandem Mass Spectra. *Expert Review of Proteomics* **2011**, *8* (5), 645–657.

(39) Jeong, K.; Kim, S.; Pevzner, P. A. UniNovo: A Universal Tool for de Novo Peptide Sequencing. *Bioinformatics* **2013**, *29* (16), 1953–1962.

(40) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **2003**, *17* (20), 2337–2342.

(41) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society of Mass Spectrometry* **2015**, *26* (11), 1885–1894.

(42) Ma, B.; Johnson, R. De Novo Sequencing and Homology Searching. *Molecular and Cellular Proteomics* **2012**, *11* (2).

(43) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **2005**, *77* (4), 964–973.

(44) Frank, A. M. A Ranking-Based Scoring Function for Peptide-Spectrum Matches. *Journal of Proteome Research* **2009**, *8* (5), 2241–2252.

(45) Kelchtermans, P.; Bittremieux, W.; De Grave, K.; Degroeve, S.; Ramon, J.; Laukens, K.; Valkenborg, D.; Barsnes, H.; Martens, L. Machine Learning Applications in Proteomics Research: How the Past Can Boost the Future. *Proteomics* **2014**, *14* (4–5), 353–366.

(46) Freund, Y.; Iyer, R.; Schapire, R. E.; Singer, Y. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research* **2003**, *4*, 933–969.

(47) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32.

(48) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. MSNovo: A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical Chemistry* **2007**, *79* (13), 4870–4878.

II. CHAPTER 1. POSTNOVO: POST-PROCESSING ENABLES ACCURATE AND FDR-CONTROLLED DE NOVO SEQUENCING

Research in this chapter was published in the Journal of Proteome Research on October 2, 2018.

Samuel E. Miller,* Adriana I. Rizzo,[†] and Jacob R. Waldbauer*

Department of the Geophysical Sciences, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, United States

Corresponding Authors:

*S.E.M. e-mail: samuelmiller@uchicago.edu. Phone: 1-952-393-5062

*J.R.W. e-mail: jwal@uchicago.edu. Phone: 1-773-702-8322

ORCID:

Samuel E. Miller: 0000-0002-2836-1401

Jacob R. Waldbauer: 0000-0002-0338-6143

Present Address:

[†]Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania 16802, United States

ABSTRACT

De novo sequencing offers an alternative to database search methods for peptide identification from mass spectra. Since it does not rely on a predetermined database of expected or potential sequences in the sample, de novo sequencing is particularly appropriate for samples

lacking a well-defined or comprehensive reference database. However, the low accuracy of many de novo sequence predictions has prevented the widespread use of the variety of sequencing tools currently available. Here, we present a new open-source tool, *Postnovo*, which post-processes de novo sequence predictions to find high-accuracy results. Postnovo uses a predictive model to re-score and re-rank candidate sequences in a manner akin to database search post-processing tools such as Percolator. Postnovo leverages the output from multiple de novo sequencing tools in its own analyses, producing many times the length of amino acid sequence information (including both full- and partial-length peptide sequences) at an equivalent false discovery rate (FDR) compared to any individual tool. We present a methodology to reliably screen the sequence predictions to a desired FDR given the Postnovo sequence score. We validate Postnovo with multiple datasets and demonstrate its ability to identify proteins that are missed by database search even in samples with paired reference databases.

II.A. INTRODUCTION

The assignment of peptide sequences to tandem mass spectra is a key step in proteomic experiments. The most widely used method of peptide identification is database search, where best matches to observed spectra are found among theoretical spectra produced from a database of peptide sequences.[1,2] An alternative to database search is de novo sequencing, which determines peptide sequences directly from mass spectra without the need for a reference database.[3–5] The generality of de novo sequencing makes it a promising approach for many types of samples. Complex environmental metaproteome samples, for example, may not have an appropriate reference database, especially if complementary metagenomic or metatranscriptomic data are not available.[6,7] Large public databases such as UniRef may contain the sequences of

interest among many others, but algorithmic search against such large databases suffers from a high rate of false positive peptide-spectrum matches (PSMs), limiting the number of statistically significant PSMs.[8] Furthermore, genetic variation in a heterogeneous population can produce protein sequences with amino acid substitutions that are not present in the database.[9,10]

De novo sequencing deduces the identity and sequence of amino acids from the pattern of peptide fragment peaks in MS/MS spectra. The method presents several challenges that have limited its use, including the difficulty of sequencing spectra with a low signal-to-noise ratio, incomplete sets of peptide fragment ions, and the lack of a reliable statistical framework for estimating the false positive rate of sequence predictions.[11] Owing to these challenges, the application of de novo sequencing has largely been confined to particular biological systems, including monoclonal antibodies[12–14] and venom peptides.[15–17] Promising improvements in de novo sequencing include the use of state of the art machine learning approaches,[18,19] complementary fragmentation methods,[20–23] isotopic labeling,[11] and refinement of existing tools through a deeper understanding of their statistical properties.[11,24,25]

Many de novo sequencing tools report peptide sequences accounting for the full mass of the parent ion, which often results in the inclusion of incorrect amino acids within an otherwise correct sequence due to weak fragment ions at certain positions. Sequencing errors are more frequent at the N-terminal end of the peptide and mostly consist of four or fewer amino acids.[24,25] Blank-Landeshammer et al. showed that de novo sequence accuracy can be improved by comparing the sequence predictions for a spectrum from multiple tools and removing conflicting N-terminal dipeptides to produce truncated sequences.[24] Here, we generalize the comparison of de novo sequence predictions in a consensus sequence algorithm that efficiently finds high-confidence subsequences from sets of de novo sequence candidates, not just top sequence

predictions. We also conduct three additional analyses to improve the accuracy of de novo sequences. First, we compare de novo sequences generated over a range of fragment mass tolerance parameterizations, showing that sequence predictions which agree over this range are more likely to be correct. Second, we cluster spectra inferred to derive from the same precursor species and compare their sequence predictions, finding that predictions shared between multiple spectra are more likely to be correct. Third, we identify subsequences within de novo sequences that can produce common isobaric errors, finding that predictions with fewer potential errors are more likely to be correct. We implement the consensus algorithm and these three additional analyses in a new open-source tool, Postnovo, which post-processes de novo sequences to boost the accuracy of predictions and assign a rigorous sequence score (Figure II.1). Postnovo fulfills a similar role in re-scoring and re-ranking de novo sequences as Percolator and PeptideProphet do with database search PSMs.[26,27]

We validate Postnovo with twelve datasets from different organisms, comprising both high- and low-resolution MS2 data, demonstrating the reliability of Postnovo's sequence score. The score can be used to screen sequences to a desired false detection rate (FDR). We demonstrate that the recall of correct sequences at any FDR significantly exceeds that of individual de novo sequencing tools. Postnovo produces a large number of high-probability sequences of sufficient length for protein analysis by homology search or the assembly of peptide sequences into full-length proteins of interest.

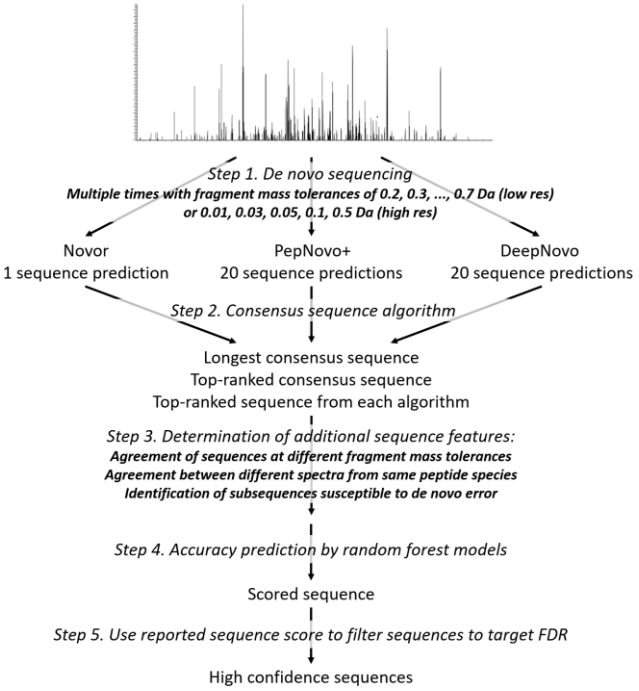

Figure II.1. Postnovo workflow
De novo sequence candidates are generated by multiple tools (Novor, PepNovo+ and DeepNovo) using multiple fragment mass tolerance settings. Postnovo finds consensus sequences from these sequences, which are further analyzed along with the top sequence candidate from each tool. Postnovo's machine learning models assign sequence candidates a score, taking into account additional information produced by Postnovo analyses, and the set of sequence predictions meeting the specified FDR cutoff is reported.

II.B. METHODS

II.B.1. PROTEOMIC DATASETS

Postnovo was validated with six low-resolution MS2 and six high-resolution MS2 proteomic datasets. Four low-resolution MS2 datasets from bacterial strains were generated in our lab: *Escherichia coli* K-12 MG1655 (197,991 spectra), *Desulfovibrio vulgaris* Hildenborough (72,458 spectra), *Rhodopseudomonas palustris* TIE-1 (62,820 spectra), and *Synechococcus sp.* WH 7803 (41,417 spectra). Two low-resolution MS2 eukaryotic datasets tested in the DeepNovo de novo sequencing study were downloaded: *Homo sapiens* testis tissue (30,051 spectra)[28] and *Drosophila melanogaster* blastoderm embryos (38,121 spectra).[29] Six

high-resolution MS2 eukaryotic, bacterial, and archaeal datasets tested in the DeepNovo study were downloaded: *Homo sapiens* macrophage vesicles[31] (114,497), *Mus musculus* fibroblast cells[32] (31,183 spectra), *Apis mellifera* tissues and hemolymph[33] (48,125 spectra), *Solanum lycopersicum* microsomal fraction[34] (94,053), *Bacillus subtilis*[35] (41,004 spectra), and *Methanosarcina mazei*[36] (50,219).

For the four low-resolution bacterial datasets, proteins were extracted by heating bacterial cell pellets to 95˚C for 20 minutes in a denaturing and reducing extraction buffer (1% SDS, 10% glycerol, 10 mM dithiothreitol, 200 mM Tris, pH 8). Cysteines were alkylated by addition of 40 mM iodoacetamide and incubation in the dark for 30 minutes. Where not otherwise specified, all solid reagents were dissolved in LC/MS-grade water (Fisher Optima). Proteins were purified by a modified eFASP (enhanced filter-aided sample preparation) protocol,[30] using Vivacon 500 concentrators (30 kDa nominal cutoff, Sartorius). Proteins were digested with MS-grade trypsin (Thermo Pierce) at 37˚C overnight, and peptides were eluted from the concentrator and dried by vacuum centrifugation. Peptide samples were reconstituted in 2% acetonitrile + 0.1% formic acid and 6 µL aliquots were injected onto a trapping column (OptiPak C18, Optimize Technologies) and separated on a capillary C18 column (Thermo Acclaim PepMap 100 Å, 2 µm particles, 50 µm I.D. × 50 cm length) using a water-acetonitrile + 0.1% formic acid gradient (2-50% acetonitrile over 180 min) at 90 nL/min using a Dionex Ultimate 3000 nanoLC system. Peptides were ionized by a nanoelectrospray source (Proxeon Nanospray Flex) fitted with a metal-coated fused silica emitter (New Objective). Mass spectra were collected on an Orbitrap Elite mass spectrometer (Thermo) operating in a data-dependent acquisition (DDA) mode, with one high-resolution (120,000 $m/\Delta m$) MS1 parent ion full scan triggering 15 MS^2 Rapid mode CID fragment ion scans of selected precursors.

Our bacterial mass spectrometry proteomics datasets have been deposited in the MassIVE Archive (https://massive.ucsd.edu) with the dataset identifier, MassIVE MSV000082266. Other datasets are available via PRIDE: low-resolution, *H. sapiens*[28] (PXD002179 File CHPP_SDS_3001), *D. melanogaster*[29] (PXD004120 File MM_BN_4a); high-resolution, *H. sapiens*[31] (PXD004424 Files 151009_exo4_2, 151009_exo4_1_3h, 151009_exo4_2_3h), *M. musculus*[32] (PXD004948 File 20160323_CoAN_CTRL1_3372), A. mellifera[33] (PXD004467 File S-1), *S. lycopersicum*[34] (PXD004947 Files 03022016_Clara_MP_Fraction_02 to 03022016_Clara_MP_Fraction_09), *B. subtilis*[35] (PXD004565 File 150710_QEp_PK_Bsub_DG_Br1), *M. mazei*[36] (PXD004325 File Mm2DLC_N_1_01).

II.B.2. ALGORITHM DESCRIPTION

Postnovo performs five principal steps to generate a set of FDR-controlled de novo peptide sequences from mass spectral data (Figure II.1): (1) run three de novo sequencing tools across multiple mass tolerance parameterizations; (2) determine top-ranked and longest consensus sequences from sequencing tool output; (3) determine additional sequence features (detailed below); (4) predict sequence accuracy (Postnovo score) using random forest models; and (5) control groupwise FDR by fitting a functional relationship between Postnovo score and local precision. Postnovo is a Python 3 application with package dependencies found in the Anaconda3 (v.5.0.1) Python distribution. In addition to the default "predict" mode, Postnovo has training and testing modes to allow the addition or validation of user training data. Postnovo is highly parallelized to exploit available CPUs.

Postnovo runs Novor[18] (v.1.1) and PepNovo+[37,38] (v.3.1) via the DeNovoGUI[39] (v.1.15.12) command line interface at each fragment mass tolerance. For low-resolution data, the

fragment mass tolerances range from 0.2-0.7 Da in steps of 0.1 Da, and for high-resolution, they are 0.01, 0.03, 0.05, 0.1, and 0.5 Da. On average, PepNovo+ sequenced 156 spectra/min/core (32 core server, Intel Xeon E5-2650 at 2.6 GHz), or 50,000 spectra in 10 minutes, and Novor was more than an order of magnitude faster. The maximum number of sequence candidates per spectrum permitted by each tool is returned (1 per spectrum for Novor and 20 per spectrum for PepNovo+). A modified version of DeepNovo[19] (v.0.0.1) was trained at low-resolution with 100,000 randomly chosen spectra from the program's default one-hour yeast dataset and at high-resolution with 50,000 randomly chosen spectra from the high-resolution yeast dataset tested in the study.[40] The modified version of DeepNovo returns the top 20 sequence candidates per spectrum as well as amino acid confidence scores for each candidate; both pieces of information are used by the program but are not normally output, so this does not significantly affect runtime. A Postnovo subcommand trained DeepNovo in parallel at each fragment mass tolerance. DeepNovo was run in a Singularity (v.2.3.2) shell using the Docker TensorFlow CPU environment. DeepNovo is significantly slower at processing high-resolution than low-resolution spectra. Training the 50,000 high-resolution yeast spectra took 483 min at 0.01 Da (20 cores, Intel Xeon E5-2680 at 2.8 GHz, 48 GB RAM). With this same computational and mass tolerance configuration, DeepNovo predicted sequences at an average rate of 22,700 spectra/hr. Postnovo refining was faster than the preceding de novo sequencing steps, completing a 50,000 sequence dataset in 25 min (32 core server, see above). Given the extensive parallelization of the de novo sequencing/Postnovo pipeline, a dataset of 50,000 spectra can be processed on a compute cluster within 4 hours or on a single server or workstation within 1 day with two Postnovo commands running DeepNovo and then Novor/PepNovo+/Postnovo.

Postnovo generates four types of features (detailed in the Results section, II.C) for each de novo sequence candidate: (1) consensus sequence information from the comparison of de novo sequence tool predictions (Section II.G.1; Figure II.5.A); (2) the agreement of sequences generated with different fragment mass tolerance parameter settings; (3) information from clusters of sequences likely to derive from the same peptide species; and (4) information on the occurrence of subsequences prone to de novo sequencing error. Postnovo determines the top sequence candidate for each spectrum from a set of candidates using machine learning classification models. Consensus sequences from each combination of tools necessarily have different sets of features from the different tools used. Therefore, Postnovo currently uses seven separate random forest models: three models for the top-one ranked sequences reported by the three individual tools, three models for the three combinations of 2-tool consensus sequences, and one model for the 3-tool consensus sequences. Each model employs features from Postnovo's four analyses (consensus information, mass tolerance agreement, precursor clustering and potential errors), in addition to features taken directly from the spectra and de novo sequencing tool output. The models report the predicted class probability of a sequence being accurate, which we term the Postnovo score. When Postnovo reports a final sequence assignment for a spectrum, it compares sequence candidates from the different models and reports the one with the highest score. Random forests are implemented in the Python 3 scikit-learn package (v.18.2),[41] with the parameterization of maximum tree depth and features optimized by grid search cross-validation of the training data.

To calculate the groupwise FDR (1-precision) of a group of de novo sequences, we fit functions that predict local precision given sequence scores from the training data. Sequences were binned by score, with the bin size being chosen to ensure a large number of sequences in

each bin and a minimal change in local precision with further divisions of the bin. For instance, Postnovo sequences were binned in score increments of 0.01 up to 0.98, the maximum value present in all training datasets. The local precision, or proportion of accurate sequences in each bin, was fit by a regression curve with a minimum R^2 of 0.995. Using this regression model, a Postnovo score cutoff can be set to yield a chosen precision (1-FDR) for a group of de novo sequences by averaging the local precisions predicted by the regression curve above a given score cutoff. Additional details regarding the Postnovo algorithm are given in Sections II.G.1-6; the Postnovo Python 3 application and an adapted version of DeepNovo are available at https://github.com/semiller10/postnovo.

II.B.3. ALGORITHM EVALUATION

"Leave-one-out" cross-validations of Postnovo were conducted for high- and low-resolution MS2 data with six proteomic datasets from different organisms for each resolution. In each of six cross-validations, a different set of five proteomic datasets was used for training and the sixth dataset for testing. The validation spectra were assigned correct sequences from the union of target-decoy search results from SEQUEST HT/Percolator[26] (implemented in Thermo Proteome Discoverer v.2.0) and MSGF+[42] (v.9949). Each set of search results was filtered to an FDR of 1% before creating the set of PSMs that agree between the two searches or were found by one search where no PSM was found for the other, so the combined FDR may be higher than 1%. For all of the datasets, cysteine carbamidomethylation was set as a static modification and methionine oxidation as a variable modification in both database search and de novo sequencing. Additionally, glutamine and asparagine deamidation were set as variable modifications for the *H. sapiens* dataset.[28] Postnovo was run in "training" mode with the five training datasets from each

cross-validation experiment, first generating the new feature set and then training the random forest models with the spectral feature matrices. The top-scoring Postnovo sequence prediction for each spectrum was compared to the corresponding database search result, with all isoleucines in the PSM replaced by leucine. The Postnovo sequence was labeled "correct" if it was a substring of the database search PSM and "incorrect" if not or if there was no corresponding database search PSM.

Postnovo results were compared to the individual results of Novor, PepNovo+, DeepNovo, and PEAKS (v.8.5) for the six datasets. Fragment mass tolerance was set to 0.5 Da for each tool, and sequences of at least 7 amino acids were retained. We used the binary classification statistics, precision and recall, to measure the ability of Postnovo to distinguish correct from incorrect sequences. Precision (1-FDR) for the purposes of validation is defined as the proportion of spectra with sequences labeled by Postnovo as correct that are "truly" correct, as determined by database search. Recall is the proportion of spectra with "truly" correct database search identifications that also have de novo sequences labeled as correct. Precision and recall can be calculated with all of the top-ranked reported predictions or with subsets filtered by confidence score threshold. Raising the score threshold above which sequences are labeled as correct filters out false positive sequence identifications, increasing the proportion of true positives (precision) among the remaining sequences. Each de novo sequencing tool reports a score that is a metric of sequence confidence and can be used as a variable threshold (Novor and DeepNovo amino acid score averaged over the sequence; PepNovo+ rank score; PEAKS average local confidence).

II.C. RESULTS AND DISCUSSION

II.C.1. POSTNOVO PERFORMANCE COMPARED TO INDIVIDUAL DE NOVO SEQUENCING TOOLS

Postnovo post-processes results from three freely-available de novo sequencing tools – Novor,[18] PepNovo+,[37,38] and DeepNovo[19], adding additional features to each sequence candidate in order to improve discrimination between accurate and inaccurate sequences. To test and validate Postnovo, we analyzed 12 datasets – six with high-resolution and six with low-resolution MS2 data – from a variety of organisms.[28,29] SEQUEST and MSGF+ database searches against the UniProt reference proteome of each organism were controlled to a 1% spectrum-level FDR and used as quasi-ground truth to label de novo sequences as correct or incorrect. First, we explored the results of running each de novo sequencing tool individually, and also compared to de novo sequencing results from the commercial software PEAKS. There are significant differences between the tools, among which is the generation of sequences always accounting for the full peptide mass by Novor, DeepNovo and PEAKS, while PepNovo+ also reports partial-length sequences. To analyze the recall of correct de novo sequences from the corresponding database search PSMs, we discard low-quality spectra that have neither de novo sequence prediction nor database search result and, for consistency with Postnovo output (see below), spectra lacking predictions longer than 7 amino acids. This leaves 394,382 spectra from the low-resolution datasets and 263,844 spectra from the high-resolution datasets. The average recall of correct sequences is 7.6% (high-resolution)/6.7% (low-resolution) for Novor, 2.3%/9.1% for PepNovo+, 7.9%/6.9% for DeepNovo, and 7.6%/6.7% for PEAKS when the top-scoring sequence candidate from each tool is considered. The higher recall of PepNovo+ is due to the use of partial-length sequences;[19] recall calculated on an amino acid basis is lower for

PepNovo+ than for the other tools (Figure II.2). A large proportion of the spectra correctly sequenced by a given tool is not correctly sequenced by the other two tools (Figure II.6), suggesting the different algorithms provide complementary output.

Precision-recall and precision-yield plots show that post-processing by Postnovo significantly improves the yield of correct sequences (Figures II.2, II.7-8). The gains in accuracy are similar for both the high- and low-resolution fragmentation spectral datasets considered. On average, Postnovo recalls 76.4% (high-resolution; σ = 16.6%) and 56.9% (low-resolution; σ = 6.7%) of truly correct sequences at a precision (1-FDR, as determined by comparison with database search results) of 50% (Table II.2). In comparison, the best-performing de novo sequencing tools recall 13.1% (high-resolution; σ = 8.3%) and 7.1% (low-resolution; σ = 1.9%) at 50% precision. At a precision of 80%, recall is 22.1% (high-resolution; σ = 25.5%) and 17.6% (low-resolution; σ = 5.5%) for Postnovo, compared to 1.7% (high-resolution; σ = 1.5%) and 1.9% (low-resolution; σ = 0.7%) for the top single tool. 80% precision (i.e., 20% sequences with at least one error) is substantially lower than the 99% (i.e., 1% FDR) to which peptide database search results are commonly controlled−a level at which de novo recall is <0.1% and the data yield impractically small − but is actually similar to the precision of current-generation DNA sequencing. However, 80% precision is actually similar to that of current-generation DNA sequencing.[55] It is clear that, despite the improvements in de novo sequencing afforded by Postnovo, peptide-spectrum matching by database search remains informatically far superior, and de novo sequencing is most applicable to biological cases inadequately served by available sequence databases. De novo tools in general show lower recall when calculated at the amino acid level (i.e., the proportion of the total length of database search results returned by de novo sequencing), due to the difficulty of accurately sequencing long peptide fragments, but Postnovo

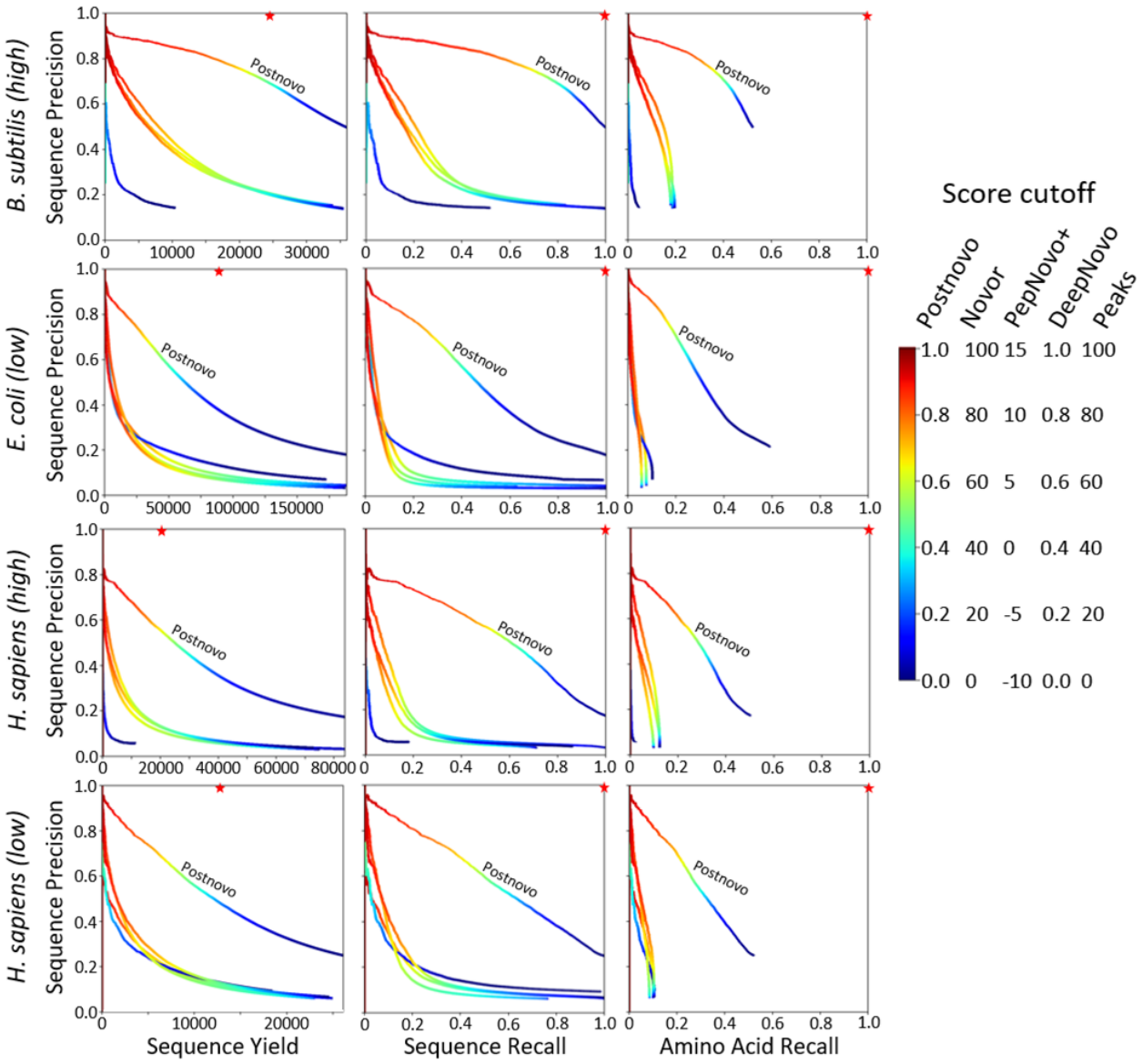

Figure II.2. Comparison of Postnovo to individual tools (datasets 1-4)
Precision-yield and precision-recall plots for de novo sequences ≥length 7 predicted from two high-resolution MS2 and two low-resolution MS2 proteomic datasets. The four unlabeled curves are the top-one ranked candidate sequence predictions of four individual algorithms. Sequence precision and yield measure the correctness of de novo sequences, whereas amino acid recall measures the number of amino acids recovered in these sequences. The variable score cutoff, represented by the color of the curve, depends on the algorithm. Postnovo assigns a score to each sequence; the scores for Novor, PepNovo+, DeepNovo, and Peaks, respectively, are the average Novor amino acid score, the rank score, the average DeepNovo amino acid score, and the Peaks average local confidence. The stars show the sequences returned by database search at a 1% FDR.

still affords substantial gains in amino acid-level performance despite sometimes sacrificing lower-confidence amino acid positions in generating partial-length sequences. In all cases, the Postnovo score has substantially greater discriminatory power than the scores output by individual de novo sequencing tools, and goes some way to closing the gap between de novo sequencing and database matching for peptide identification.

II.C.2. CONTRIBUTION OF NOVEL FEATURES TO THE POSTNOVO CLASSIFICATION MODEL

Postnovo's sequence classification model uses features derived both directly from the output of de novo sequencing tools, such as the average Novor amino acid score, and from further processing of de novo sequence predictions by Postnovo. The features added by Postnovo fall into four categories: consensus sequences, fragment mass tolerance parameterization, clustering by precursor ion, and potential subsequence errors. These additional features provide the majority of Postnovo's classification power (Figure II.3.A-C). The average importance of Postnovo's additional features in the seven random forest models is 76%. The average importance of the confidence scores used by Novor, PepNovo+, and DeepNovo to rank de novo sequence candidates is 17%, and the average importance of "other features," such as precursor mass, is 7% (Section II.G). Postnovo produces a bimodal score distribution (Figure II.3.D), with inaccurate sequences concentrated at low scores and accurate sequences concentrated at high scores.

II.C.2.i. CONSENSUS SEQUENCES AND MODEL RESULTS

Consensus sequences are shared subsequences from the de novo sequence predictions of different tools. Postnovo recovers consensus sequences from each possible combination of tools, which currently numbers four (Novor-PepNovo+, Novor-DeepNovo, PepNovo+-DeepNovo, and Novor-PepNovo+-DeepNovo). They can span the entire length of the peptide or be shorter in order to avoid lower-confidence amino acids, especially at the N-terminus, where there is often less fragment evidence.[24] We demonstrated that the use of partial-length sequences increases precision on a per sequence and per amino acid basis (Figures II.2, II.7-8), indicating that the reduction in sequence length does not artificially boost precision. Postnovo has a default sequence length threshold of 7 amino acids, mainly to reduce runtime of the consensus routine, and the user can set a different threshold as desired. Consensus information is a critical part of the success of the predictive model, with multi-algorithm consensus sequences comprising 87% of reported sequences with Postnovo scores of at least 0.5, and consensus sequences that are shorter than full peptide length comprising 73% (Table II.4). Consensus sequences originating from three tools are the most frequent type in the set of sequences with scores exceeding 0.9 (Figure II.3.D), but 2-algorithm consensus sequences are a substantial portion as well, indicating the importance of multi-algorithm consensus and additional Postnovo features in discriminating high-accuracy de novo sequences.

Postnovo finds consensus sequences by solving the longest common substring (LCS) problem for pairs of de novo sequences using a novel and efficient dynamic programming algorithm for comparing ranked lists of strings (Section II.G.1; Figure II.5). Postnovo compares not only the top-ranked sequences reported by each tool, but also the sets of sequence candidates for a spectrum in a pairwise comparison procedure. Novor only reports 1 sequence per spectrum,

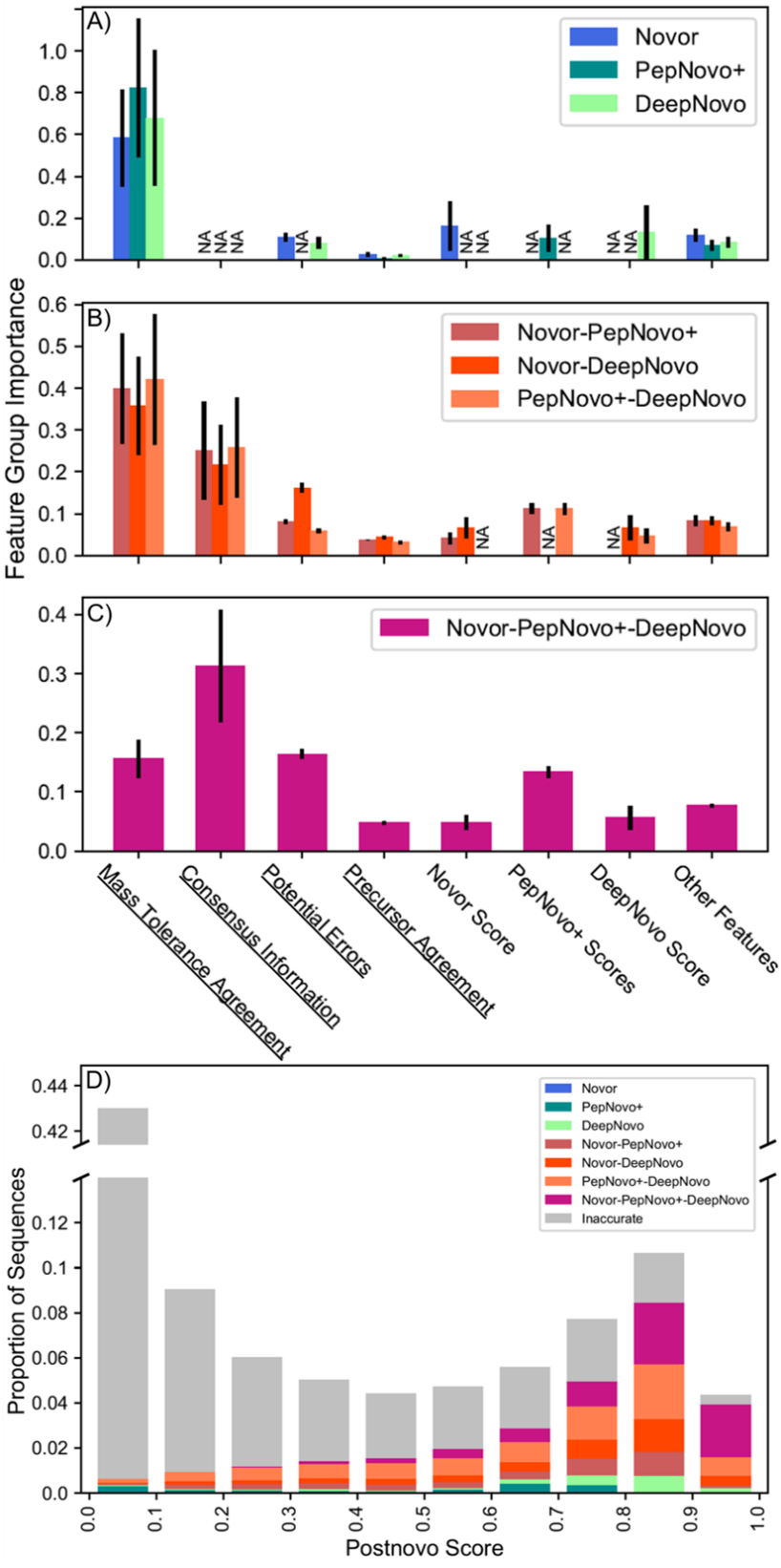

Figure II.3. Contributions to Postnovo model

(continued from previous page)
Categories of sequence candidate features determined by Postnovo (underlined) rather than those taken directly from de novo sequencing tool output account for most of the predictive power of the Postnovo machine learning classification models. Seven separate random forest models were produced from low-resolution MS2 data for (A) top-one ranked sequences from single tools, (B) consensus sequences from the output of 2 tools, and (C) consensus sequences from 3 tools. "NA" indicates features not in the specific model. The importance of each feature in the models is measured as the average decrease in Gini impurity of the full set of training data at each decision tree node that uses the feature, with error bars being the standard deviation of this metric across the decision trees of the forest.
(D) Reported Postnovo sequences across the six low-res training datasets binned by score, with the origins of accurate sequence predictions from among the seven random forest models shown in color.

but PepNovo+ and the modified version of DeepNovo used with Postnovo report 20 per spectrum, and other de novo sequencing algorithms such as Peaks can also report multiple candidates. Our algorithm reduces the number of extraneous comparisons required to find the *longest* and *top-ranked consensus sequences* from sequence lists. Given *N* lists of ranked sequences, Postnovo first finds common substrings (shared subsequences) from pairs of 2 lists. Common substrings from 2 lists must by definition contain any common substrings from 3 or more lists that include those 2 lists; likewise, common substrings from *N-1* lists must contain any common substrings from *N* lists. Therefore, Postnovo sequentially finds common substrings from 2 to *N* lists, using the common substrings found at each stage as input for comparisons in the next stage. If common substrings are not found from *N-1* lists, then comparisons of *N* lists are not required, as common substrings among *N* lists will not be found. Furthermore, at each stage, once the *longest* and *top-ranked consensus sequences* are identified, further sequence comparisons are halted (by pausing a generator function that performs pairwise sequence comparisons), as the common substrings meeting the criteria for longest and top-ranked consensus sequences have been found. In subsequent stages, if the common substrings from prior stages are not sufficient to identify these consensus sequences, then the generator functions from

the prior stages are restarted to find more common substrings that may form the basis of new common substrings in the subsequent stage (Section II.G.1).

The inclusion of lower-ranked candidates in Postnovo proves important in our cross-validation experiments. Of the sequences reported by Postnovo with a score of at least 0.5, 22% on average are consensus sequences derived from at least one lower-ranked de novo sequence candidate (Table II.5). The most important feature group in the classification accuracy of the 3-tool model is consensus information (Figure II.3.C), which encompasses features recording the fraction of single-algorithm candidate sequence length preserved in the consensus sequence and the ranks of those contributing single-algorithm candidate sequences.

II.C.2.ii. MASS TOLERANCE AGREEMENT

The parameterization of fragment mass tolerance in de novo sequencing tools defines the allowable mismatch between calculated peptide fragment masses and peaks in the observed spectrum. This window size affects the ability of the tool to distinguish between amino acids of similar masses, especially with low-resolution MS2 spectra.[20,25] We found that changing the fragment mass tolerance, from 0.2-0.7 Da in 0.1 Da increments for low-resolution MS2 data or from 0.01 through 0.03, 0.05, 0.1, and 0.5 for high-resolution data, significantly affects sequence predictions. Optimization of this parameter has also been found to increase the number of database search PSMs by up to 52%.[43] Our analysis of the overlap of de novo sequence predictions generated at each value of the fragment mass tolerance shows that accurate predictions are more likely to be shared across multiple tolerance settings than are inaccurate sequences (Figure II.9). Postnovo makes use of this observation by running the de novo sequencing tools over the parameter space and determining whether the predictions at each

parameter value agree, a binary feature included in each classification model. Fragment mass tolerance agreement features are the most important category of features in the single-tool and 2-tool consensus sequence models (Figure II.8.A-B).

II.C.2.iii. PRECURSOR CLUSTERING

The comparison of sequence predictions for spectra derived from the same molecular species (precursor ion) provides another means to evaluate the robustness of predictions.[43] Postnovo determines the agreement between de novo sequence predictions for peptide-level spectral clusters (Section II.G.2). For each sequence in the cluster, Postnovo counts the number of other sequences in which that sequence occurs, recording both the proportion of sequence matches in the cluster and the total size of the cluster for the sequence. We find that this Precursor Clustering contributes to the discriminatory power of the classification models (Figure II.3), although this effect is relatively small, possibly because relatively abundant peptides with many spectra also produce richer fragmentation spectra that are more reliably sequenced.

II.C.2.iv. POTENTIAL SEQUENCE ERRORS

It has been documented that a majority of de novo sequence errors occur in short subsequences of four or fewer amino acids and consist of isobarically substituted and misordered amino acids.[24,25] Substitutions (e.g., Q/AG and AD/EG) and inversions (e.g., GA/AG) between inferred and correct sequences are typically caused by weak fragmentation patterns or the confusion of b- and y-ions. Postnovo counts the occurrences of mono- and dipeptides with perfectly isobaric and near-isobaric substitutions (Section II.G.3), and weights the counts by amino acid-level Novor and DeepNovo scores. Postnovo also identifies di- and tripeptide

subsequences with an average Novor or DeepNovo amino acid score more than one standard deviation lower than the average confidence score of the sequence as a whole and the adjoining amino acids. The purpose of this procedure is to identify areas with missing or ambiguous fragment peaks that can lead to short inversions (e.g., AG/GA and AGD/DAG). The post hoc identification of potential sequencing errors is important in many of the Postnovo classification models (Figure II.3), and is the second-most powerful feature in the 3-algorithm consensus model.

II.D. FDR CONTROL FROM POSTNOVO SCORING

A reliable sequence confidence score allows the precision (1-FDR) of a set of de novo sequences to be determined in the absence of database search results. We tested the ability of sequence scoring metrics from Postnovo and the individual de novo sequencing tools to predict the precision of sequences binned by score from our six datasets. We found that local precision (i.e., proportion of accurate sequences within a score bin) was best described by a quadratic fit to Postnovo score and an exponential fit to the scores of the individual tools (Figures II.4, II.10-12), with Postnovo score closely approximating the probability of sequence accuracy for scores above 0.8. Regression of binned score data produced a better fit than logistic regression of unbinned sequence accuracy (0 or 1) against score, especially at high scores (Figure II.13). The empirical regression models allow a score cutoff to be found that yields a chosen precision for a set of de novo sequences. The groupwise FDR of the set is calculated from the average of local precisions predicted by the regression curve above the score cutoff.

We performed leave-one-out experiments with each of the six low-resolution test datasets to determine the generalizability of the regression models to unseen data. In each experiment, a

regression curve was fit to the grand mean of the binned proportions of accurate sequences for the other five datasets. The sequence scores of the reserved dataset were mapped to a predicted local precision, and the predicted overall precision was calculated at each score threshold. The error in this prediction is the difference versus the actual precision at each threshold (Figure II.4.B). The error in predicted precision at each threshold is generally less than 0.1 for any dataset, and the average error across all of the leave-one-out trials is close to zero. This holds for the scoring metrics from Postnovo and the three individual de novo sequencing tools (Figures II.10.B-12.B), indicating that Postnovo scoring, with its attendant increases in accuracy, is at least as reliable as the parent algorithm scores for setting experimental FDR. Considering sequences filtered by Postnovo score to a predicted precision of 0.9, the grand mean of the actual precisions from the six leave-one-out experiments is 0.914, with the most discrepant individual datasets having actual precisions of 0.972 and 0.824. Regression equations from the grand mean of all six training datasets (Section II.G.6) are used to estimate precision at each scoring threshold in the Postnovo output.

II.E. ACCURATE SEQUENCES NOT FOUND BY DATABASE SEARCH

We investigated the highest scoring Postnovo predictions from each low-resolution MS2 dataset that were called incorrect by comparison to database search results (Tables II.6-11), in order to assess the causes of variations between de novo sequencing and database search results. We performed BLAST[44] homology searches of the 30 highest scoring de novo sequences with a minimum length of 12 that did not agree with the corresponding database search result from each dataset (180 sequences total; Table II.1). These sequences were queried against both RefSeq and the appropriate reference proteome. Of these 180 nominally incorrect sequences, 96 are from

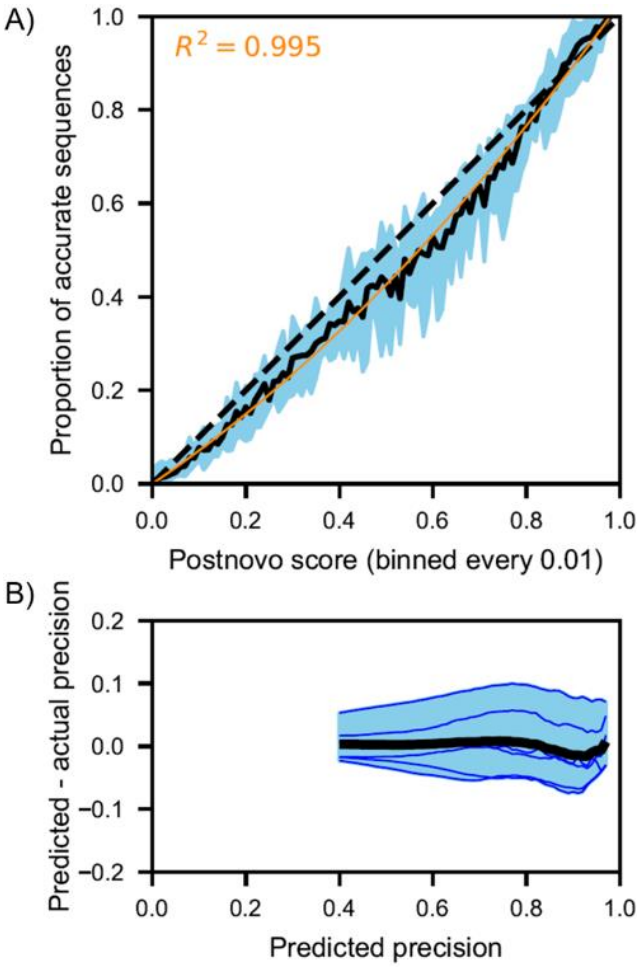

Figure II.4. Relation of Postnovo score to sequence precision
(A) The local precision, or accurate proportion, of Postnovo sequences for each score bin. The range of results for each of the six low-resolution MS2 test datasets is represented by the blue envelope, and the average of the six datasets is represented by the solid black line. The dashed black line is 1:1. A quadratic regression (orange) relates sequence score to local precision.
(B) Overall precision was predicted over the range of Postnovo score thresholds for each test dataset using leave-one-out regression models. The residual error, or difference between predicted and actual precision, is plotted as blue lines for each dataset, with the leftmost point in the line containing all Postnovo sequences (lowest score threshold) and the rightmost point containing only the highest scoring sequence. The black line is the average of the error in precision across all six datasets.

peptides isobaric with an aligned tryptic peptide from the reference proteome (the "reference peptide"), implying an error in de novo sequencing. Of the other 84 non-isobaric sequences, 38 are likely to be correct but could not be identified by database search against the reference

proteome, for reasons explored below, and 8 are identifiable keratin contaminants. The correctness of the remaining sequences could not be determined, as 27 are non-isobaric with the reference peptide, often differing by a single amino acid, and 11 do not produce strong alignments with sequences in either database.

Of the 38 sequences likely to be correct, 32 differ from the reference peptide by non-isobaric amino acid mismatches that may correspond to post-translational modifications not included in the database search settings. The most common potential modification is asparagine deamidation to aspartate, which most likely occurs in sample preparation. Two datasets also have de novo sequences with evidence of singly and doubly oxidized tryptophan.[45] Other non-isobaric mismatches can be explained by amino acid substitutions in protein variants. For instance, we found a known polymorphism of *H. sapiens* transferrin (C2 variant)[46] and a sequence from the variable region of immunoglobulin kappa light chain, which are not present in the reference proteome.

Four peptides from unexpected organisms were found among the 60 sequences investigated from the two eukaryotic datasets. Two de novo sequences from the *D. melanogaster* embryo dataset[29] are identical to proteins from *Wolbachia* species of bacterial intracellular parasites. *D. melanogaster* is known to harbor maternally-transmitted *Wolbachia* bacteria which are capable of altering the host phenotype in unexpected ways.[47–49] Two de novo sequences from the *H. sapiens* dataset[28] match proteins from *Lactococcus lactis* and *Saccharomyces cerevisiae*, respectively, with no homologous hits to proteins from other organisms in RefSeq, suggesting that the proteins originate from these or closely related organisms. The *H. sapiens* samples consist of homogenized testis tissue from cadavers; fungi of the family *Saccharomycetales* (e.g., *Candida* spp.) and lactic acid bacteria of the order *Lactobacillales* frequently reside on the skin

Type of mismatch	Count
Mismatches not caused by de novo error	
Unidentified asparagine deamidation	32
Unidentified mono- or di-oxidized tryptophan	4
Common contaminant	8
Unexpected protein from different organism	4
Unexpected sequence variant	2
De novo sequencing errors	
Isobaric substitution (4 or fewer amino acids)	54
Isobaric inversion (4 or fewer amino acids)	34
Isobaric error involving amino acids in peptide but not in partial-length Postnovo sequence	8
Mismatches of ambiguous origin	
Non-isobaric amino acid substitution	27
No homologous sequence found	12

Table II.1. Summary of 180 high-scoring Postnovo sequences in the six low-resolution MS2 datasets that do not match their respective database search PSM

of the genitals[50,51] and in seminal fluid,[52–54] respectively. Our analyses confirm that high-accuracy de novo sequencing can be used to discover protein modifications and variants, as well as unexpected sequences from uncharacterized organisms, parasites and contaminants (Table II.1).[11] Finally, correct sequences that were mistakenly called incorrect in our Postnovo cross-validation experiments had the effect of increasing the error of the precisions estimated from the Postnovo score, so Figure II.4.B provides an upper bound on the estimated error in precision.

II.F. CONCLUSIONS

We developed an open-source tool, Postnovo, that greatly increases the accuracy of de novo sequences through the syncretic use of multiple tools[18,19,37,38] and the generation of a set of novel features of candidate sequences that provide greater discriminatory power. In our cross-validation experiments with both low- and high-resolution MS2 proteomic datasets from a variety of organisms, Postnovo sequence recall exceeds that of the best-performing de novo sequencing tools, DeepNovo and Novor, by a factor of seven to fifteen at an FDR of 10% (Tables II.5-6) – a higher FDR than generally applied to database-matching results (which remain superior when an accurate, comprehensive reference database is available), but roughly equivalent to the percentage of imperfect reads in current-generation Illumina sequencing.[55] Although Postnovo includes partial-length sequences in its reported results, accuracy on a per amino acid basis is significantly higher than individual tools, since Postnovo's enhancements extend beyond trimming low-confidence amino acids from the full-length peptide sequences reported by Novor and DeepNovo. High sequence accuracy is required for many downstream uses, such as homology searches and sequence assembly, due to the short length of many peptide sequences and the increased incidence of sequencing errors in longer de novo sequences.[25]

Postnovo fills a similar role in de novo sequencing to that of Percolator[26] and PeptideProphet[27] in peptide-spectrum matching by database search – the re-ranking and re-scoring of PSMs to boost accuracy and report a confident FDR.[11,56,57] A significantly larger fraction of spectra can now be sequenced de novo at a low FDR using Postnovo than is possible with individual de novo tools. Our cross-validation experiments also discovered unexpected proteins not present in the reference databases, showing that de novo sequencing can reveal additional biological signals even in datasets that are generally well-described by a reference

database. We demonstrated that the Postnovo score assigned to each sequence prediction by our machine learning models can be reliably used to filter the set of sequences to a target FDR. Postnovo allows the user to perform validation tests with their own proteomic datasets and retrain the models as needed.

The generality of the principles developed in Postnovo will allow it to be extended to additional de novo sequencing tools. Postnovo and the de novo sequencing tools it uses are all highly parallelized and can take advantage of high-performance computation. Postnovo gives the user the option of creating a new training database for the machine learning model or updating an existing database, as well as testing Postnovo predictions against database search results. The incorporation of a post-processing methodology into de novo sequencing analyses should facilitate research on a variety of sample types that are especially challenging for peptide identification by traditional database search, including complex environmental and clinical samples, splicing and translational variants, and antibodies. The Postnovo Python 3 application and an adapted version of DeepNovo are available at https://github.com/semiller10/postnovo.

Acknowledgments

We are grateful to Maureen Coleman for discussions, comments on the manuscript and for providing the *R. palustris* and *Synechococcus* cell samples; to Alex Bradley and Wil Leavitt for the *D. vulgaris* cell samples; to Tao Pan for the *E. coli* cell samples; to Lichun Zhang for assistance with sample preparation and operation of the mass spectrometer; to Albert Colman, Gerard Olack and Mark Anderson for helpful discussions; and to three anonymous reviewers whose comments improved the manuscript. This work was supported by the Gordon and Betty Moore Foundation (awards 3305 and 3306) and the Simons Foundation (award 402971).

II.G. SUPPORTING INFORMATION

The research in this section is the basis of a supporting information file for the October 3, 2018 article in the Journal of Proteome Research.

II.G.1. CONSENSUS SEQUENCE IDENTIFICATION

Two major considerations of Postnovo's consensus sequence routine are the reduction of unnecessary pairwise comparisons to reduce runtime and the generation of consensus information that can be encoded as features in the Postnovo classification model. It is possible for a spectrum to have a 2-algorithm consensus sequence meeting the minimum length requirement but not a 3-algorithm consensus sequence. The most naïve approach to the consensus sequence problem for each combination of algorithms would be to find the LCS from every possible pairwise comparison of de novo sequence candidates. However, any consensus sequence from three parent sequences (algorithms) must also contain a consensus sequence from two parent sequences, so consensus comparisons of N (>2) algorithms should start with the set of $N-1$ algorithm consensus sequences, reducing the number of comparisons required (Figure II.5.B).

Secondly, Postnovo recovers only two types of consensus sequence, the "longest" and the "top-ranked" (Figure II.5.A). Consensus sequences are encoded in classification model features as being one or both of these. The longest consensus sequence is the longest encountered among the pairwise comparisons, and the top-ranked consensus sequence has the lowest sum of parent ranks encountered (e.g., Novor candidate #1 and PepNovo+ candidate #2 could form a consensus sequence with a rank sum of 3). Importantly, these two consensus sequences can be encountered and verified as "longest" and "top-ranked" prior to the completion of every pairwise comparison

if, in the case of the "longest," a consensus sequence spans the length of the shortest parent candidate in the sets under consideration, and, in the case of the "top-ranked," a consensus sequence has a lower rank sum than subsequent possible comparisons. The discovery of these two consensus sequences allows the set of pairwise comparisons to be truncated, further reducing runtime. This becomes more complex when considering more than 2 algorithms, as the longest and top-ranked consensus sequences for *N-1* algorithms may not be able to form any *N* algorithm consensus sequences, while unconsidered *N-1* algorithm consensus sequences may be able to do so. Therefore, instead of stopping the pairwise comparison routine, Postnovo pauses a pairwise comparison generator function and restarts it as needed to generate additional consensus sequences during higher-order algorithm comparisons.

The novel consensus sequence algorithm, an extension of the solution to the longest common substring (LCS) problem for ranked lists of strings, is described by the following pseudocode representation.

Algorithm:
Given *N* lists of ranked strings, find the longest common substrings (LCS) from each combination of lists, retaining the longest LCS and lowest-ranked LCS.

Input:
S is a list of ranked strings. $s_1, s_2, ..., s_r \in S$.
Ex. *S* = [seq 1, seq 2, ..., seq *r*].
StringLists is a dictionary mapping a tuple of string list names, *C*, to a combination of string lists. $S_1, S_2, ..., S_N \in$ *StringLists*. *N* is the number of string lists in *StringLists*.
D is a list of all combinations of string list names. $D_2, D_3, ..., D_k \in D$. $C_1, C_2, ..., C_k \in D_k$.
Ex. D_2 = [(Novor, PN+), (Novor, DeepNovo), (PN+, DeepNovo)].
C_1 = (Novor, PN+).
StringLists[C_1][1] = [Novor seq 1].
StringLists[C_1][2] = [PN+ seq 1, PN+ seq 2, ..., PN+ seq 20].
LCS_Dict is a dictionary mapping a tuple of string list names, *C*, to a list of LCS strings;
Rank_Dict is a related dictionary mapping *C* to a list of the summed ranks of each LCS's "parent" strings from which the LCS is derived.

Output:
Results is a dictionary mapping combinations of string list names to the longest LCS and lowest-ranked LCS.

Procedure GetConsensusSeqs(*StringLists*, *D*)

```
1   for k from 2 to N:
2       for C in D[k]:
3           initialize LCS_longest as None
4           initialize LCS_lowest_rank as None
5           LCS_Dict[C] = list()
6           Rank_Dict[C] = list()
7           if k == 2:
8               S1 = StringLists[C[1]]
9               S2 = StringLists[C[2]]
10              L_max = min(max_len(s for each s in S) for S in [S1, S2])
11              Generator_Dict[C] = g = LCS_comparison_generator(S1, S2)
12              for next LCS_output in g:
13                  s_x = string of rank x from S1 that was compared to s_y from S2 in this call to generator
14                  s_y = string of rank y from S1 that was compared to s_x from S2 in this call to generator
15                  LCS_x,y = LCS meeting min length criterion from this call to generator
16                  Rank_x,y = x + y
17                  LCS_Dict[C].append(LCS_x,y)
18                  Rank_Dict[C].append(Rank_x,y)
19                  if len(LCS_x,y) > LCS_longest:
20                      LCS_longest = LCS_x,y
21                  if Rank_x,y < rank(LCS_lowest_rank):
22                      LCS_lowest_rank = LCS_x,y
23                  if len(LCS_x,y) == L_max and LCS_longest == LCS_lowest_rank:
24                      if there is no possibility of a lower LCS rank sum in further comparisons:
25                          Results[C] = tuple(LCS_longest, LCS_lowest_rank)
26                          break iteration of g
27          else k > 2:
28              B = tuple(C[1], C[2], ..., C[N - 1]])
29              S1 = LCS_Dict[B]
30              Ranks1 = Rank_Dict[B]
31              S2 = StringLists[C[N]]
32              L_B,max = len(Results[B][0])
33              Rank_B,min = Results[B][1]
34              if L_B,max != None and Rank_B,min != None:
```

```
        35                                  Generator_Dict[C] = g =
LCS_comparison_generator(S1, S2) *
        36                                  for next LCS_output in g, starting with sx =
LCSlongest,Bs then sx = LCStop_rank,Bs:
        37                                      sx = string of rank x from S1 that was
compared to sy from S2 in this call to generator
        38                                      sy = string of rank y from S1 that was
compared to sx from S2 in this call to generator
        39                                      LCSx,y = LCS meeting min length criterion
from this call to generator
        40                                      Rankx,y = Ranks1[x] + y
        41                                      LCS_Dict[C].append(LCSx,y)
        42                                      Rank_Dict[C].append(Rankx,y)
        43                                      if LCSx,y != None:
        44                                          Results[C] = tuple(LCSlongest,
LCSlowest_rank)
        45                                          break iteration of g
        47                                  if x+1 < len(LCS_Dict[B]):
        48                                      if LCS_Dict[B][x+1] == None:
        49                                          Recursively generate LCSs to
find next string in S1 using paused generators in Generator_Dict[B]
        50      return Results
```

* The generator function, *LCS_comparison_generator*, finds the longest common substrings of strings from two string lists, returning the result of a pairwise LCS string comparison with each iteration. An outer loop iterates the first list of strings in ascending order of string rank, and an inner loop iterates the second list of strings also in ascending order of string rank.

II.G.2. CLUSTERING SPECTRA FROM THE SAME MOLECULAR SPECIES

A scoring metric was formulated to efficiently find clusters of spectra likely derived from the same molecular ion. The de novo sequence predictions of spectra in the cluster were compared to generate features for the Postnovo machine learning model, as greater agreement between the predictions positively correlates with the likelihood of the predictions. First, spectra were clustered by MS1 peptide mass with an error tolerance of 10 ppm. Next, a preliminary score was calculated for each member of the cluster by pairwise comparison of the amino acid composition of each sequence prediction. Sequences of like type were clustered and compared, e.g., top-ranked DeepNovo de novo sequences generated at a fragment mass tolerance of 0.2 Da

were clustered. The preliminary score assigned to the sequence was the number of amino acids shared between the sequences divided by the length of the shorter sequence in the pair. This sequence level information is derived by the de novo sequencing algorithms largely from the MS2 fragmentation spectra, so we used it as an abstraction of this spectral data for the purposes of clustering like species. The preliminary score was then modified by the difference in length between the sequences, with a larger difference reducing the score, and the difference in retention time between the spectra, with a small difference boosting the score. If the score was less than a minimum value of 0.7, then the spectrum was removed from the preliminary cluster and made available for other clusters. To test this methodology, spectra from the *H. sapiens* test dataset were controlled to a 1% spectrum-level FDR and clustered by PSM, as well as being clustered by this model. The parameters of the model were tuned to minimize the difference between the PSM and model-based clusters.

$clustering\_score = p + b - d * t$

p is the proportion of shared amino acid composition, *b* is the proximity bonus of 0.2 if spectra co-occur within 1% of the total chromatography time, *d* is the difference in de novo sequence length, and *t* is the penalty factor of 0.1 for differing sequence lengths.

II.G.3. POTENTIAL SEQUENCE ERRORS

Postnovo identifies short isobaric and near-isobaric subsequences. There are two mono-/dipeptide substitutions and 12 isobaric di-/dipeptide substitutions, not counting the reverse sequences as well: N/GG, Q/AG, AD/EG, AN/GQ, AS/GT, AV/GL, AY/FS, C(+57.02)T/M(+15.99)N, DL/EV, DQ/EN, DT/ES, LN/QV, LS/TV, and NT/QS. We consider one mono-/dipeptide substitution and four di-/dipeptide near-isobaric substitutions with a mass

difference of 0.0112 Da or less: R/GV, C(+57.02)L/SW, ER/VW, FQ/KM(+15.99), and LM(+15.99)/PY.

II.G.4. OTHER FEATURES OF POSTNOVO MODEL

A small number of other features easily obtained from the spectral and de novo sequencing data were included in the Postnovo sequence classification model. The other features include: measured mass of the precursor ion, the relative retention time of the ion over the total run time, the length of the sequence prediction, and the mass tolerances at which the sequence prediction was generated. These last mass tolerance features differ from the mass tolerance agreement features; to illustrate the difference, consider two sequences, A and B. A was generated with a mass tolerance of 0.2 Da, and B was generated at each mass tolerance from 0.3-0.7 Da. A is a substring of B, so the 0.2-0.7 Da agreement features for A all have a value of 1, whereas B is not a substring of A, so the 0.2 Da agreement feature for B has value of 0 and the 0.3-0.7 Da features have values of 1. In contrast, the simpler mass tolerance features only consider the origin of the sequences and not substring relationships, so A has a value of 1 for 0.2 Da and 0 for 0.3-0.7 Da, whereas B has a value of 0 for 0.2 Da and 1 for 0.3-0.7 Da.

II.G.5. LENGTH-ACCURACY TRADEOFF OF PARTIAL-LENGTH SEQUENCES

The highest scoring sequence candidate is reported from the set of sequence candidates for each spectrum found by the Postnovo single-algorithm and consensus sequence random forest models. Since this biases reported sequences toward shorter partial-length sequences only spanning high-probability amino acids, the user has the option of trading some accuracy for length by allowing Postnovo to report a longer sequence candidate at the expense of sequence

accuracy. We have found that the longer sequence typically includes the more accurate shorter sequence. The default tradeoff settings increase the average length of reported sequences from 8.09 to 8.74, with most of the loss among the length 7 sequences and most of the gain among the length 10-15 sequences (Figure II.14.A). The impact of this tradeoff on the score cutoff required to achieve a desired precision is relatively small. From the average of our cross-validation experiments, the score cutoff must be increased from 0.73 to 0.77 to achieve 80% precision and from 0.88 to 0.89 to achieve 90% precision (Figure II.14.B).

The amount of Postnovo score that can be traded for a sequence extension is structured on the addition of 1 amino acid to the minimum length sequence of 7 amino acids. A tradeoff of 0.07% Postnovo score per 1% length is equivalent to a tradeoff of 1% of score for an addition of 1 amino acid in 7. Increasing this value beyond 0.35%/1% results in little additional gain in sequence length (Figure II.14.A), so this is used as the default. The tradeoff mainly results in the loss of length 7 sequences and the gain of length 10-14 sequences.

II.G.6. SCORE MODELS

Sequence predictions from Postnovo, Novor, PepNovo+ and DeepNovo were binned by Postnovo score, the average Novor amino acid score over the sequence, PepNovo+ rank score, and the average DeepNovo amino acid score over the sequence, respectively. Sequences were binned by the following score intervals: 0.01 for Postnovo (over the range 0 to 0.98), 1.0 for Novor (over the range 0 to 100), 0.5 for PepNovo+ (over the range -10 to 13), and 0.01 for DeepNovo (over the range 0 to 0.98). For each of the six test datasets, the proportion of accurate sequences in the bin (local precision) was calculated. The grand mean of the local precision of

each bin was calculated across the datasets. Regression curves were fit to the score bin midpoint and local precision data. Terms were added until the R^2 value of the fit exceeded 0.995.

Postnovo

$y = (3.532*10^{-1})*x^2 + (6.745*10^{-1})*x$

x is the Postnovo score of the sequence.

Novor

$x_1 = x/100$

$y = (4.168*10^{-3})*exp((5.635*10^{0})*x_1) + (-2.850*10^{-1})*x_1^2$

x is the Novor average amino acid score of the sequence, which is scaled to x_1 to accommodate the exponential function.

PepNovo+

$x_1 = x/1000$

$y = (3.474*10^{-2})*exp((4.622*10^{2})*x_1 + (-2.466*10^{4})*x_1^2 + (6.471*10^{5})*x_1^3)$

x is the PepNovo+ rank score of the sequence, which is scaled to x_1 to accommodate the exponential function.

DeepNovo

$y = (1.525*10^{-8})*exp((3.447*10^{1})*x + (-1.655*10^{1})*x^2)$

x is the DeepNovo average amino acid score of the sequence.

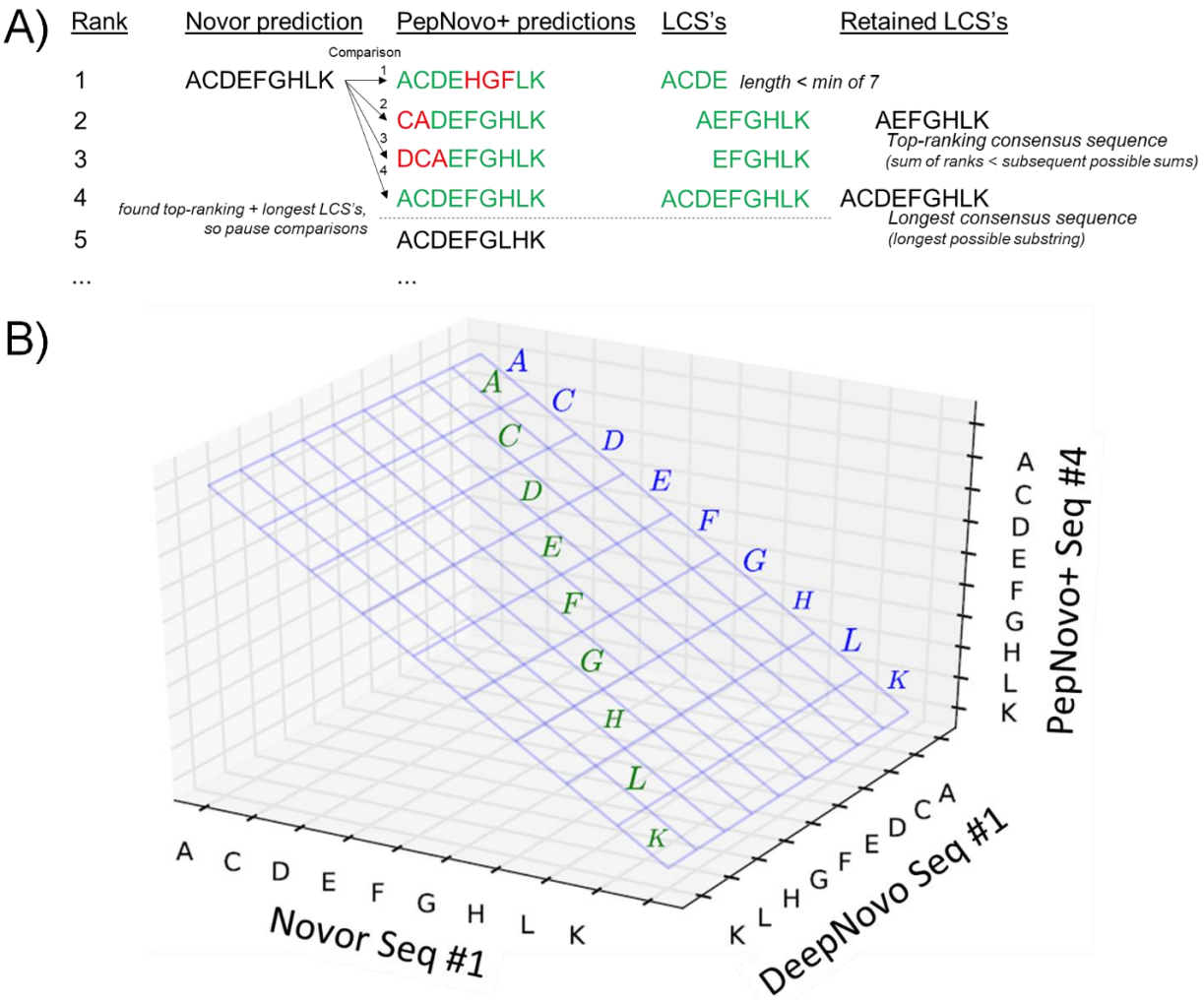

Figure II.5. Postnovo consensus sequence procedure
(A) Example of the generation of 2-algorithm consensus sequences. The reported Novor sequence candidate is compared to the 20 reported PepNovo+ sequence candidates in ascending order of rank, finding the longest common substring (LCS) from each pairwise comparison. The first comparison produces two common substrings, ACDE and LK, with ACDE being the LCS. This LCS does not meet the minimum length criterion of 7 amino acids, however. The comparison routine is halted when two target LCSs are found: the top-ranked LCS and the longest possible LCS. Here, the top-ranked LCS and has a summed rank of 3 (Novor rank #1 + PepNovo+ rank #2), and the longest LCS spans both Novor sequence #1 and PN sequence #4.
(B) Example of the generation of a 3-algorithm consensus sequence. To find Novor-PepNovo+-DeepNovo consensus sequences, build upon the 2-algorithm LCSs. This greatly simplifies the LCS task by reducing the 3-dimensional dynamic programming task to a 2-dimensional task. Here, the longest Novor-PepNovo+ LCS (blue) matches DeepNovo sequence #1, forming the 3-algorithm longest LCS (green) – we know that this is the longest LCS as a longer substring is not possible.

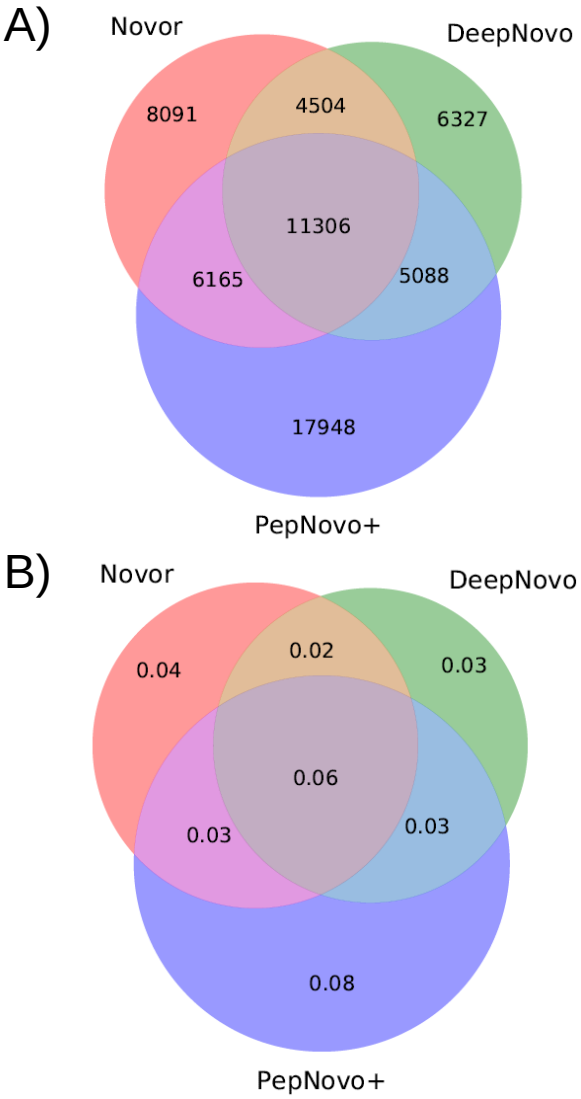

Figure II.6. Pooled de novo sequencing results from six low-resolution test datasets
(A) Counts of spectra with correct top-one ranked sequences of at least seven amino acids from each tool run individually. Substring matches against paired database search PSMs were used to measure correctness.
(B) Recall of spectra with correct top-one ranked sequences of at least seven amino acids.

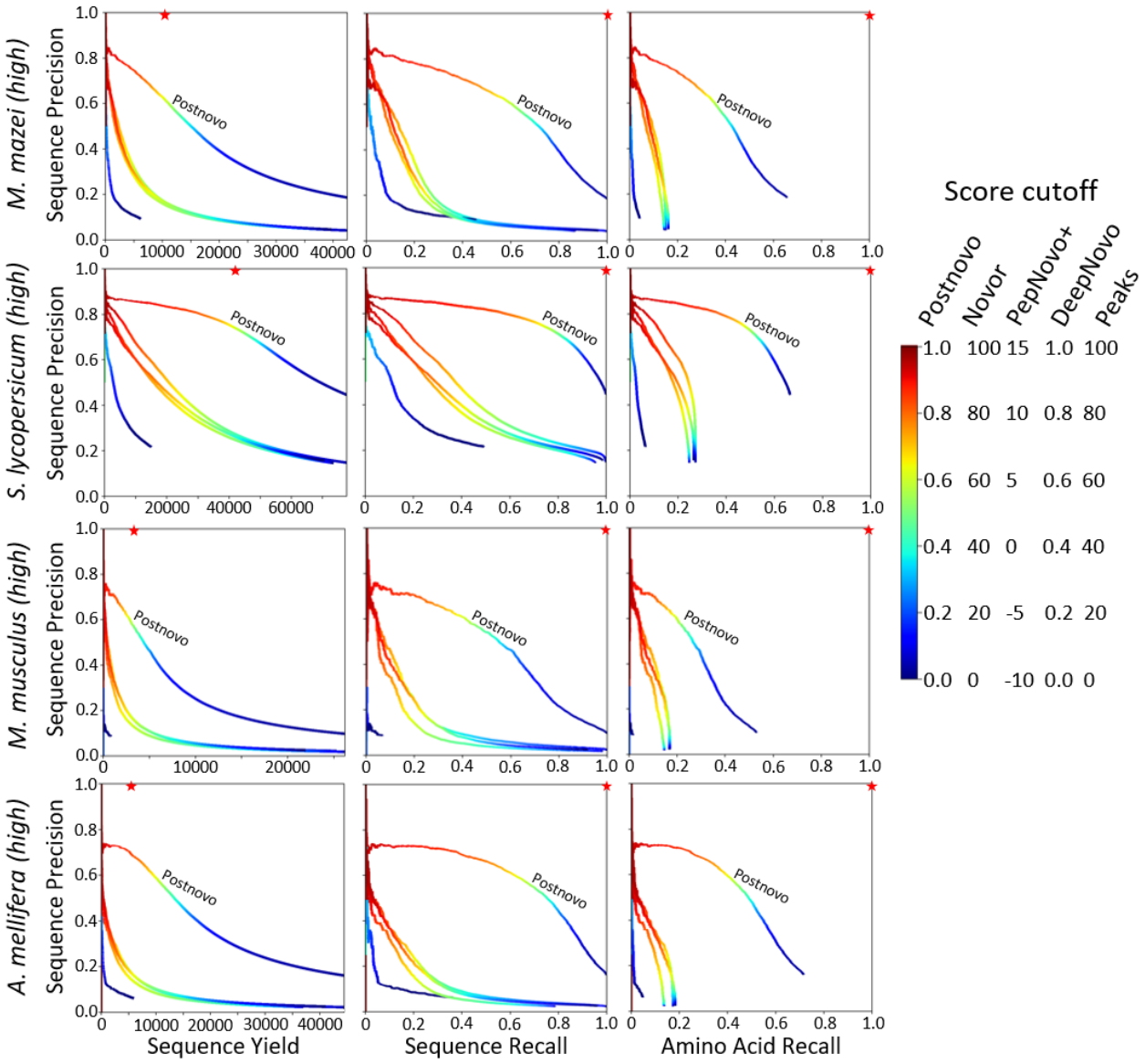

Figure II.7. Comparison of Postnovo to individual tools (datasets 5-8)
Precision-yield and precision-recall plots for de novo sequences ≥length 7 predicted from four high-resolution MS2 proteomic datasets. The four unlabeled curves are the top-one ranked candidate sequence predictions of four individual algorithms. Sequence precision and yield measure the correctness of de novo sequences, whereas amino acid recall measures the number of amino acids recovered in these sequences. The variable score cutoff, represented by the color of the curve, depends on the algorithm. Postnovo assigns a score to each sequence; the scores for Novor, PepNovo+, DeepNovo, and Peaks, respectively, are the average Novor amino acid score, the rank score, the average DeepNovo amino acid score, and the Peaks average local confidence. The stars show the sequences returned by database search at a 1% FDR.

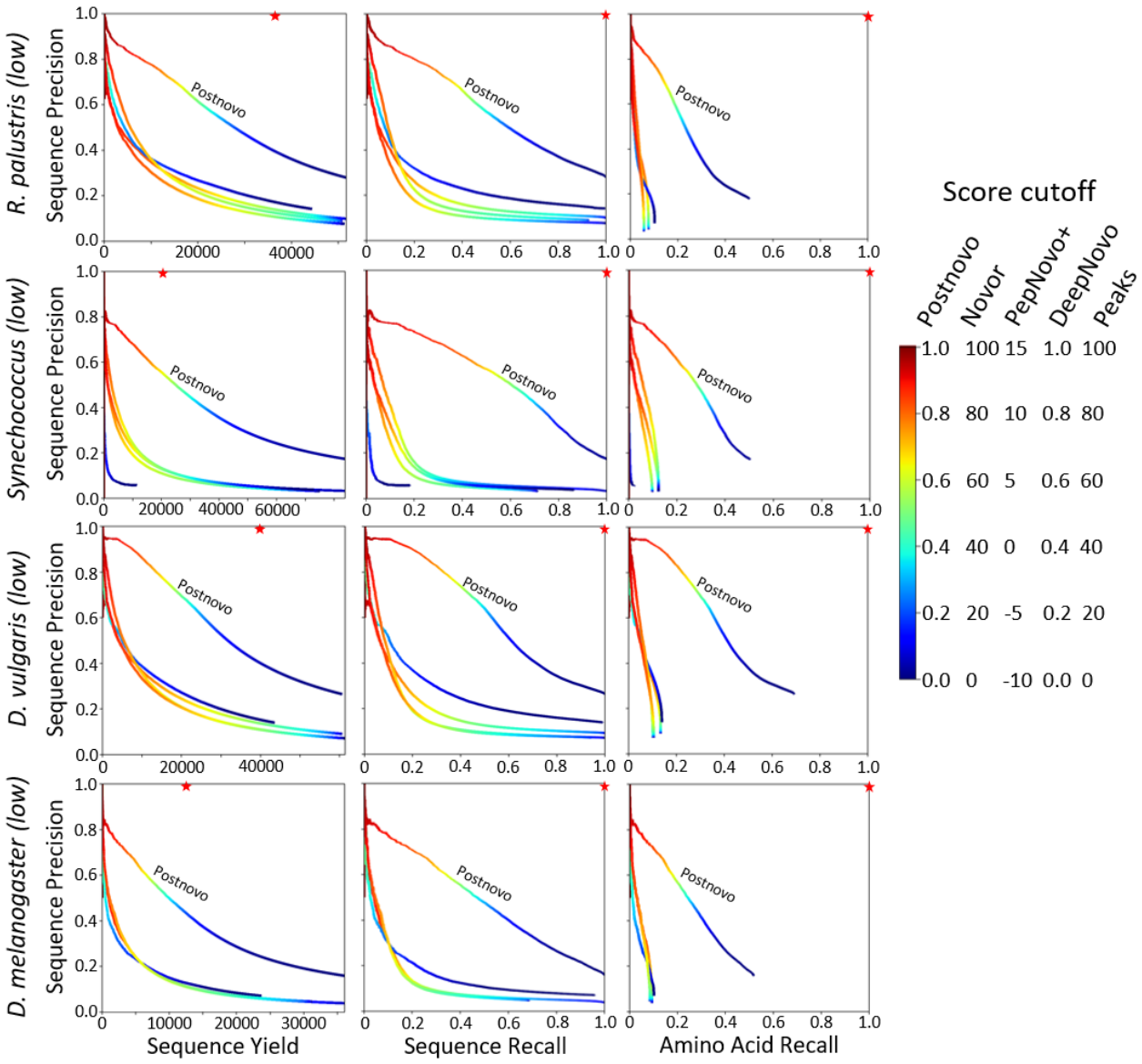

Figure II.8. Comparison of Postnovo to individual tools (datasets 9-12)
Precision-yield and precision-recall plots for de novo sequences ≥length 7 predicted from four low-resolution MS2 proteomic datasets. The four unlabeled curves are the top-one ranked candidate sequence predictions of four individual algorithms. Sequence precision and yield measure the correctness of de novo sequences, whereas amino acid recall measures the number of amino acids recovered in these sequences. The variable score cutoff, represented by the color of the curve, depends on the algorithm. Postnovo assigns a score to each sequence; the scores for Novor, PepNovo+, DeepNovo, and Peaks, respectively, are the average Novor amino acid score, the rank score, the average DeepNovo amino acid score, and the Peaks average local confidence. The stars show the sequences returned by database search at a 1% FDR.

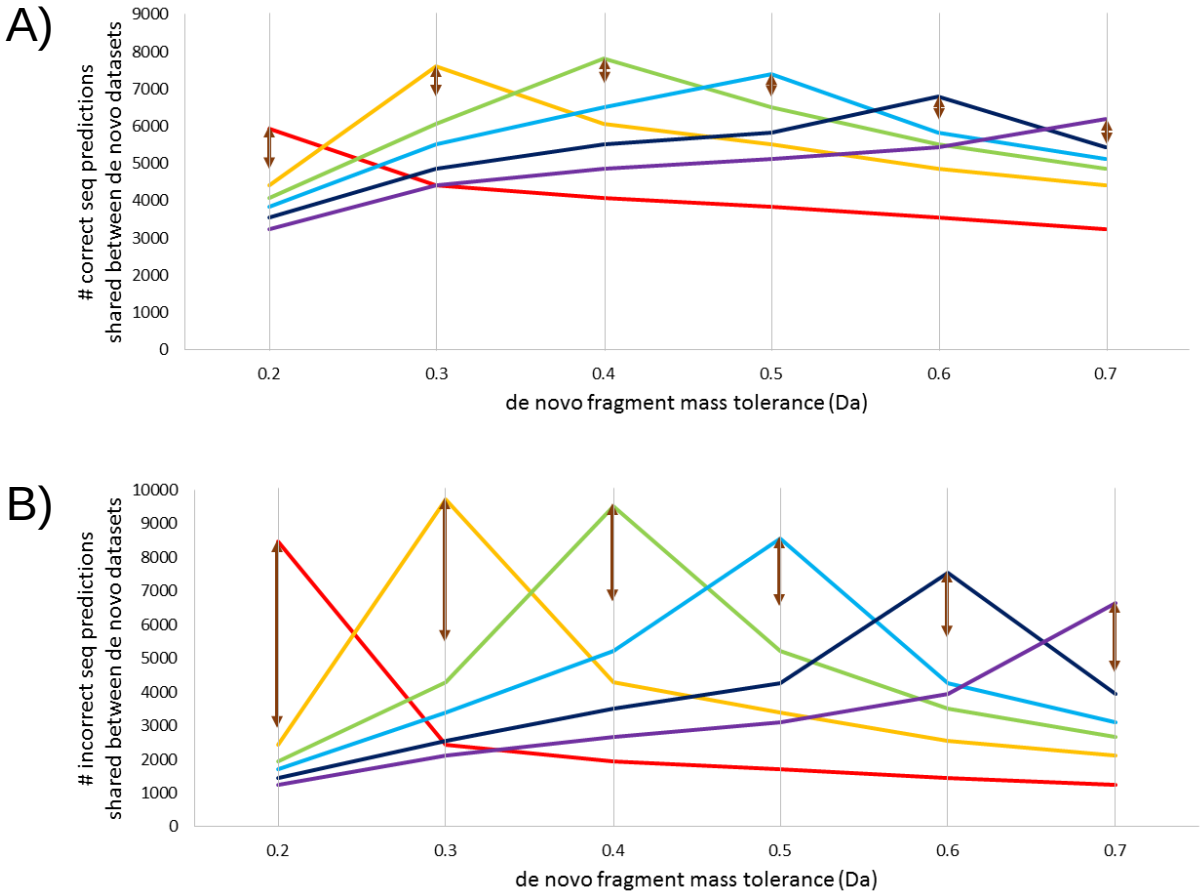

Figure II.9. Fragment mass tolerance comparison
Comparison of consensus de novo sequences (from *D. vulgaris* proteome) generated six times using different fragment mass tolerances (0.2, 0.3, ..., 0.7 Da). Each color corresponds to a tolerance, with hotter colors representing lower tolerances (0.2 is red, 0.3 orange, ..., 0.7 purple). The peak of each colored line is the number of predictions for the tolerance. For example, in (A), which shows accurate predictions, the 0.2 Da predictions are in red, and there are ~6,000 of them. The values of the colored line at other tolerances equals the number of predictions from the colored line tolerance that are shared with the other tolerances. In (A), the number of 0.2 Da predictions that are also predicted at 0.3 Da is ~4,500. The brown arrows show the number of unique predictions for a given tolerance, i.e., sequences not predicted at any other tolerance. In (A), the number of unique 0.2 Da predictions is ~1,000. Note that accurate predictions (A) are more likely to be shared between different mass tolerance predictions than inaccurate predictions (B).

(A) Accurate sequence predictions. Accuracy is defined as the prediction being found in the benchmark database search PSM for the same given spectrum or the underlying reference proteome.
(B) Inaccurate sequence predictions.

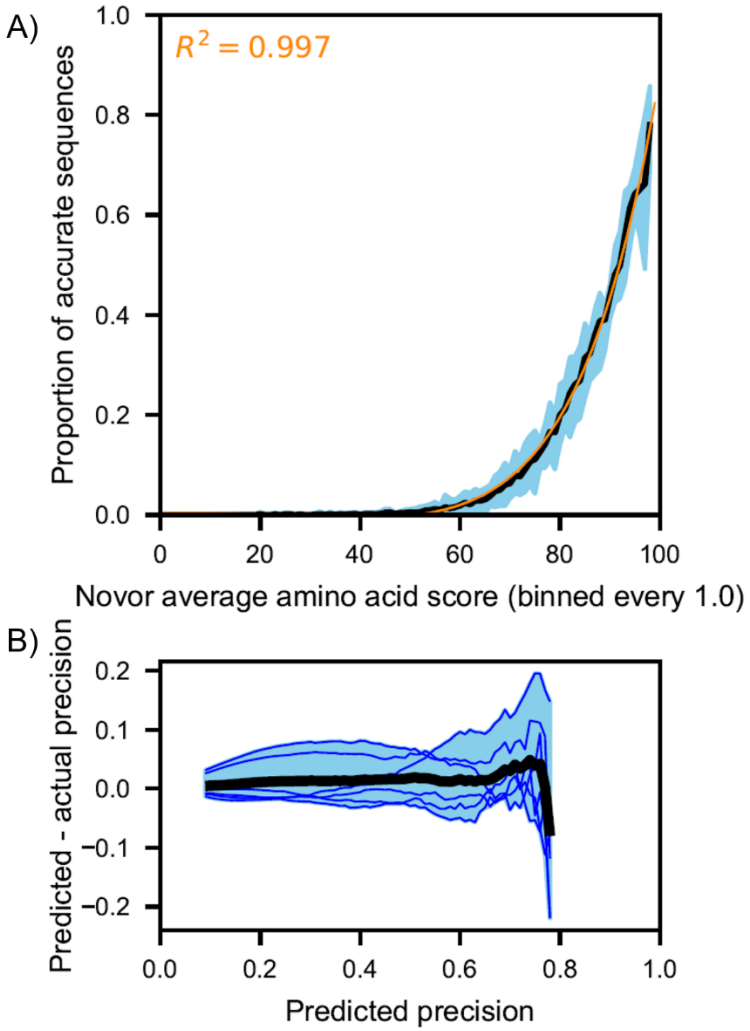

Figure II.10. Prediction of precision from Novor score
(A) The local precision, or accurate proportion, of Novor sequences is plotted for each score bin, with the range of results for each of the six test datasets represented by the blue envelope, and the average of the six datasets represented by the solid black line. A regression of the binned data (orange) relates sequence score to local precision.
(B) Overall precision was predicted over the range of Novor score thresholds for each test dataset using leave-one-out regression models. The error, or difference between predicted and actual precision, is plotted as blue lines for each dataset, with the leftmost point in the line containing all Novor sequences and the rightmost point containing only the highest scoring sequence. The black line is the average of the error in precision across all six datasets.

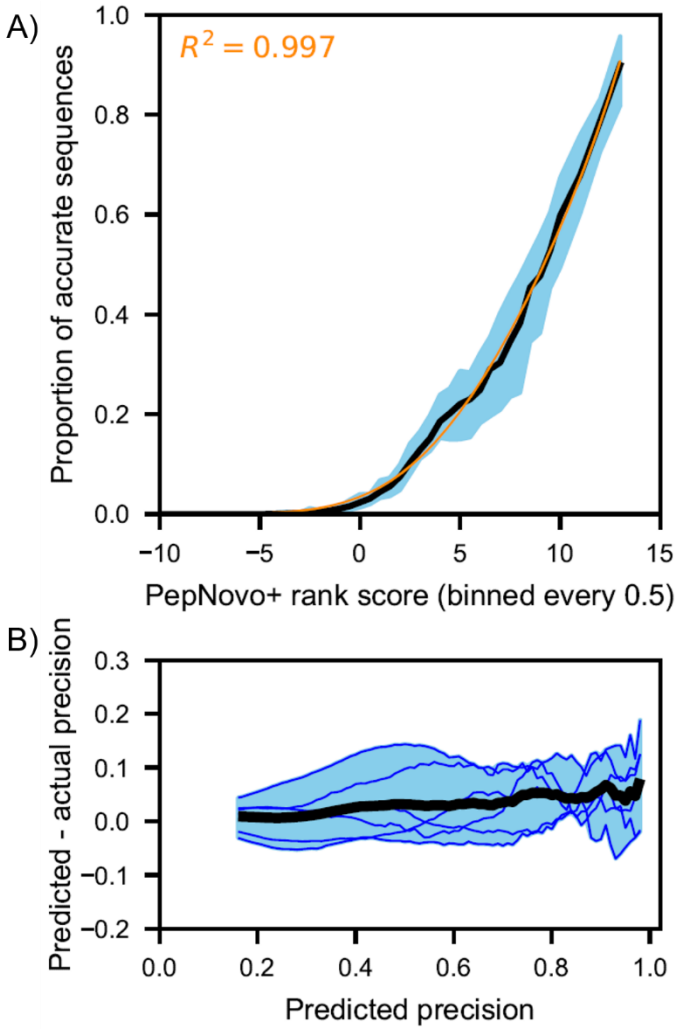

Figure II.11. Prediction of precision from PepNovo+ Score
(A) The local precision, or accurate proportion, of PepNovo+ sequences is plotted for each score bin, with the range of results for each of the six test datasets represented by the blue envelope, and the average of the six datasets represented by the solid black line. A regression of the binned data (orange) relates sequence score to local precision.
(B) Overall precision was predicted over the range of PepNovo+ score thresholds for each test dataset using leave-one-out regression models. The error, or difference between predicted and actual precision, is plotted as blue lines for each dataset, with the leftmost point in the line containing all PepNovo+ sequences and the rightmost point containing only the highest scoring sequence. The black line is the average of the error in precision across all six datasets.

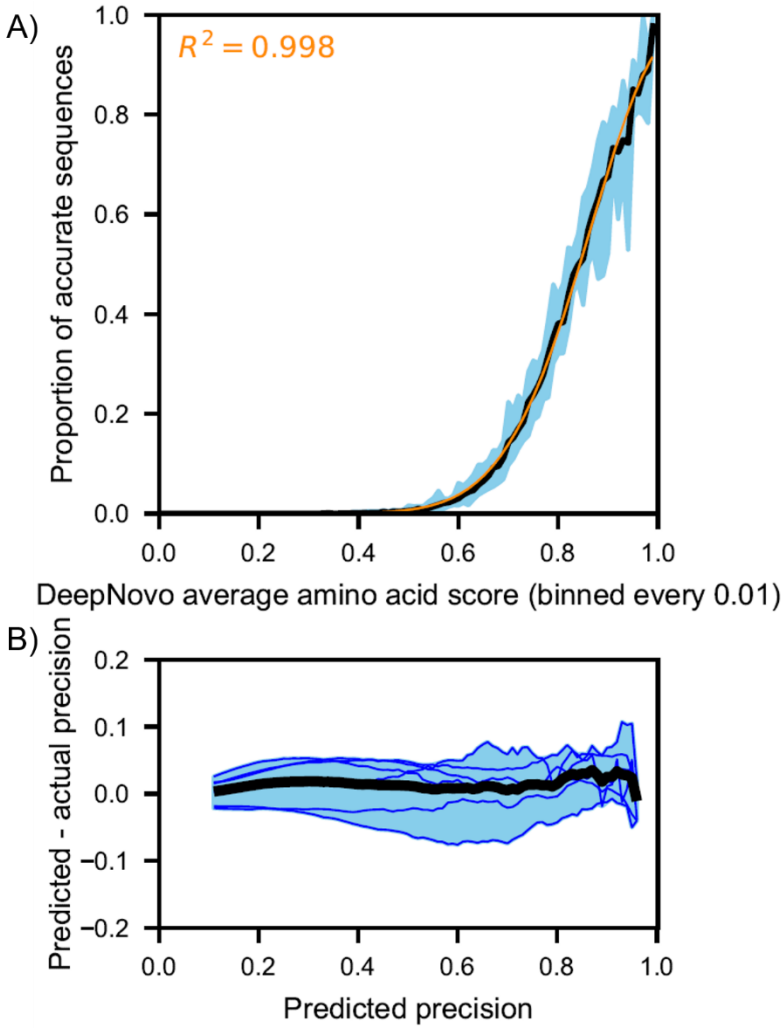

Figure II.12. Prediction of precision from DeepNovo score
(A) The local precision, or accurate proportion, of DeepNovo sequences is plotted for each score bin, with the range of results for each of the six test datasets represented by the blue envelope, and the average of the six datasets represented by the solid black line. A regression of the binned data (orange) relates sequence score to local precision.
(B) Overall precision was predicted over the range of DeepNovo score thresholds for each test dataset using leave-one-out regression models. The error, or difference between predicted and actual precision, is plotted as blue lines for each dataset, with the leftmost point in the line containing all DeepNovo sequences and the rightmost point containing only the highest scoring sequence. The black line is the average of the error in precision across all six datasets.

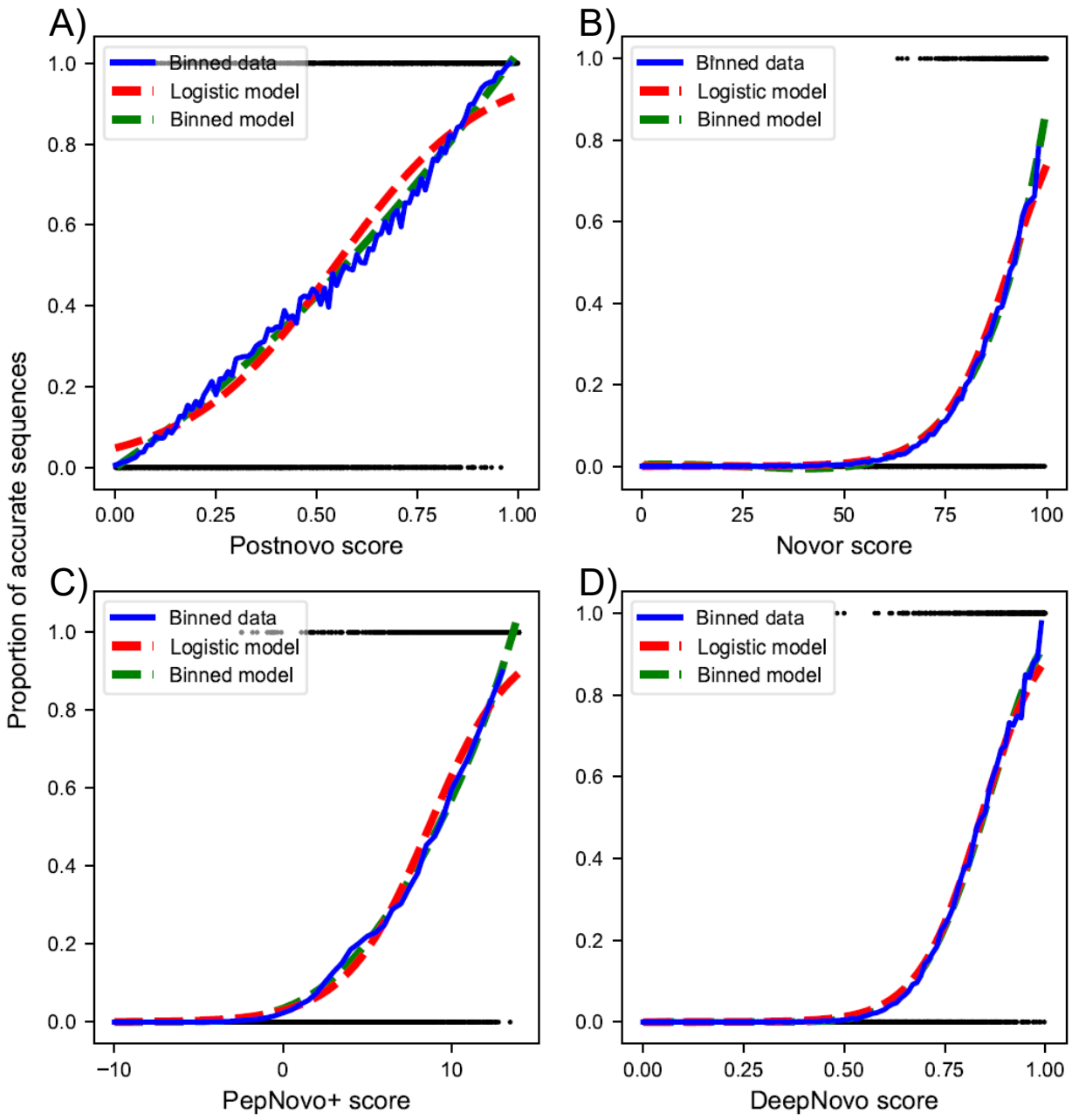

Figure II.13. Local precision model selection
Comparison of binned score regression models and logistic regression models of (A) Postnovo, (B) Novor, (C) PepNovo+, and (D) DeepNovo sequence accuracy from score metrics. The sigmoid shape of logistic curves produces a poorer fit to binned accuracy data than the curves fit to the binned data.

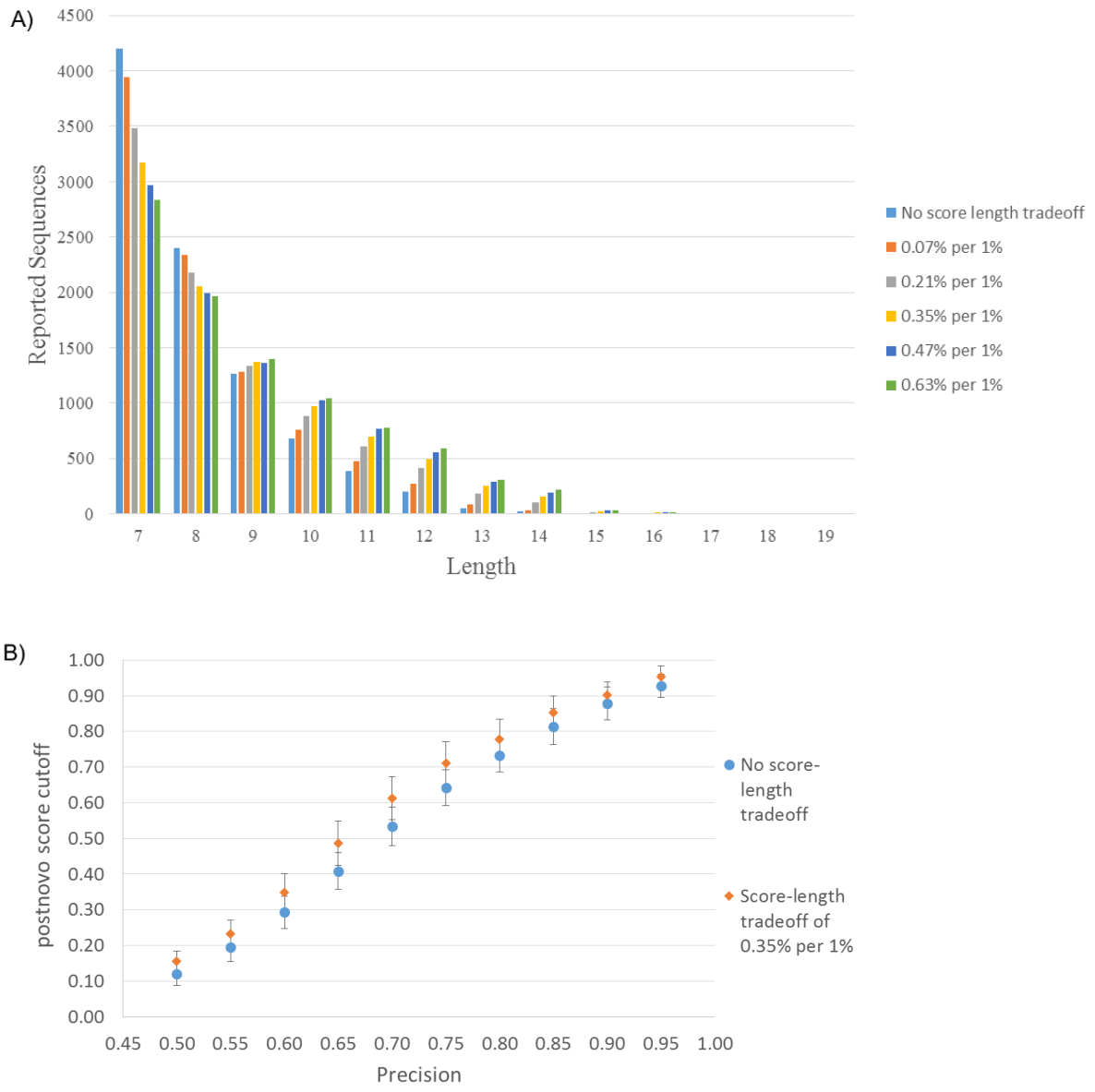

Figure II.14. Candidate sequence length control
(A) The number of *H. sapiens* Postnovo sequences reported at each length given the score-length tradeoff parameterization. By default, 0.35% of the candidate Postnovo score can be sacrificed for a 1% increase in sequence length.
(B) The use of a score-length tradeoff does not have a major effect on the classification statistics of Postnovo. This plot shows the Postnovo score cutoff required to achieve a target precision with and without the score-length tradeoff.

II.G.8. SUPPORTING TABLES

Dataset	Postnovo recall at precision = 0.5	Top single tool recall at precision = 0.5	Postnovo recall at precision = 0.8	Top single tool recall at precision = 0.8	Postnovo recall at precision = 0.9	Top single tool recall at precision = 0.9
H. sapiens	0.602	0.071 (Novor)	0.026	0.003 (Novor)	0.001	0
M. musculus	0.552	0.077 (Novor)	0.011	0.010 (Novor)	0.006	0.004 (Novor)
A. mellifera	0.753	0.030 (Novor)	0.003	0.001 (DeepNovo)	0.002	0
S. lycopersicum	0.970	0.269 (DeepNovo)	0.593	0.041 (DeepNovo)	0.007	0.002 (DeepNovo)
B. subtilis	0.987	0.206 (Novor)	0.561	0.034 (Novor)	0.068	0.001 (DeepNovo)
M. mazei	0.721	0.135 (Novor)	0.136	0.011 (Novor)	0.004	0.001 (DeepNovo)
Average (standard deviation)	0.764 (0.166)	0.131 (0.083)	0.221 (0.255)	0.017 (0.015)	0.015 (0.024)	0.001 (0.001)

Table II.2. Recall at three precisions (0.5, 0.8, and 0.9) for each high-resolution dataset

Dataset	Postnovo recall at precision = 0.5	Top single tool recall at precision = 0.5	Postnovo recall at precision = 0.8	Top single tool recall at precision = 0.8	Postnovo recall at precision = 0.9	Top single tool recall at precision = 0.9
H. sapiens	0.642	0.071 (DeepNovo)	0.215	0.019 (DeepNovo)	0.070	0.005 (DeepNovo)
D. melanogaster	0.483	0.047 (DeepNovo)	0.060	0.007 (DeepNovo)	0.003	0.003 (DeepNovo)
E. coli	0.468	0.046 (DeepNovo)	0.158	0.013 (DeepNovo)	0.024	0.002 (DeepNovo)
D. vulgaris	0.616	0.084 (DeepNovo)	0.203	0.030 (DeepNovo)	0.063	0.010 (DeepNovo)
R. palustris	0.595	0.090 (DeepNovo)	0.202	0.022 (DeepNovo)	0.038	0.006 (DeepNovo)
Synechococcus sp.	0.608	0.089 (DeepNovo)	0.218	0.022 (DeepNovo)	0.006	0.001 (PepNovo+)
Average (standard deviation)	0.569 (0.067)	0.071 (0.019)	0.176 (0.055)	0.019 (0.007)	0.034 (0.026)	0.005 (0.003)

Table II.3. Recall at three precisions (0.5, 0.8, and 0.9) for each low-resolution dataset

Dataset	Total predictions	Number that are consensus sequences	Number that are partial-length consensus sequences	Proportion that are consensus sequences	Proportion that are partial-length consensus sequences
H. sapiens	7030	5886	4958	0.837	0.705
D. melanogaster	9214	7678	6227	0.833	0.676
E. coli	41797	38354	34553	0.918	0.827
D. vulgaris	18970	15984	12626	0.843	0.666
R. palustris	29274	25686	21857	0.877	0.747
Synechococcus sp.	17372	15354	13102	0.884	0.754
Average				0.865	0.729

Table II.4. The contribution of consensus sequences to Postnovo sequence predictions with a probability ≥0.5

Dataset	Total predictions	Number that are consensus sequences with at least 1 lower-ranked sequence	Proportion that are consensus sequences with at least 1 lower-ranked sequence
H. sapiens	7030	1428	0.203
D. melanogaster	9214	1712	0.186
E. coli	41797	9990	0.239
D. vulgaris	18970	5257	0.277
R. palustris	29274	4860	0.166
Synechococcus sp.	17372	3905	0.225
Average			0.216

Table II.5. The contribution of consensus sequences derived from at least one lower-ranked de novo sequencing tool candidate sequence to Postnovo sequence predictions with a probability ≥0.5

Table II.6. *Homo sapiens* cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
17801	TGVSTGWTQLSK	[S]GLSTGWTQLSK	[S]GLSTGWTQLSK	1.4	0.002
23695	VELLDNHEDAPLR	VEILDGNHEDAPLR	VEILDGGHEDAPLR	0.052	0.004
24551	LALANALTSALR	LALADALTAALR	[AL]LANALTSALR	2.8	0.006
13495	NSGNTATLTLTR	NSGNTATLTISR	SGHTVTLTLT	2.8	1.6
9693	SSGSAVVSGDGK	SSGSAVVNGDGK	SSGSAVVSG[GSR]	1.4	0.88
11446	SVEEYANCHLAR	[P]VEEYANCHLAR	[P]VEEYANCHLAR	0.005	2.E-05
18984	GPAGPLSGAGPL	PAGPLSGAGP	GPAGPIGSAGP[I]	5.6	10
24193	ETMLYLAPTLAA	QTMLYLAPTLAA	[Q]TMIYLAPTLAA	0.021	0.001
19351	SVALTLVHLEPR	VALTLVHLEKPR	LSLVNLEPR	1.4	3.6
13340	LAEEANADLEVK	[AL]EEANADLEVK	[AL]EEANADLEVK	0.69	0.002
10210	LVDGQSHLSLSK	LVDGQSHISLSK	LVNGQSHISLSK	0.24	0.007
21318	AYLWVGTQSEAEK	AYLWVGTGASEAEK	AYLWVGTGASEAEK	0.21	7.E-04
28877	Q(+0.98)Q(+0.98)AASGLLTSLK	QEAASGLLTSLK	EEAASGLLTSLK	0.49	0.019
18427	LGLFGQDDEVTSK	LDLFGQDDDEVTS	[I]GIFGQDEDVTSK	4.8	0.023
20365	QQ(+0.98)QTVQLQSELSR	QEQTVQLQSELSR	QEQTVQLQSELSR	0.003	5.E-06
9849	QSGDSQESVTEQ	QSADSQDGVTEQ	QSGNSQESVTEQ	16	2.E-04
15638	TQSPSSLSASLGDR	TQSPSSLSASLGD	TQSPSSLSASVGDR	0.008	0.015
32901	SNDFDEYLFALE	SNDFDEYLMAIE	SNDFDEYIMAIE	0.69	0.038
20690	VFSDGADLSGVTEEA	VFSNGADLSGVTEEA	VFSNGADLSGVTEEA	1.E-04	4.E-07
13360	LSQEEHVAVAQLR	[SL]QEEHVAVAQLR	[SL]QEEHVAVAQLR	0.037	1.E-04
28218	LTVLHQDWLDGK	LTVLHQDWLNGK	VLH-DWADGK	0.004	4.5
26185	LDDMDELMAGFK	[I]DDMDELMAGFK	LDGMAELMAG	0.003	0.19
12869	VMSQQ(+0.98)LQQ(+0.98)QLHK	VMSQEIQEQLHK	VMSQEIQEQLHK	0.97	0.001
33351	LLQQLYSQLQSK	QQLYSQLQS	[II]QEIYSQIQSK	3.9	1.6
26840	SSEELSTLETLK	SSEELSTLAALK	SSELE-STLETLK	7.9	0.27
15254	SSGLVSLGVDGR	SSGLVSLGIDGR	SSGIVSLGVGDR	0.35	1.6
13589	ELESGAVSGLEK	ELETGAVTGLE	ELES-QVSGLEK	23	0.39
28181	GVDQFLTDYQLK	GVDQLLTDYQIK	GVDQLFTDYQIK	16	0.048
21810	LQLAQQ(+0.98)YCGDCK	LQLAQEYCGDCK	LQLAQEYCGDCK	0.004	6.E-06
20965	ESLHYAVAAATK	EPLHYAVAAA	ESLHSFVAAATK	23	0.78

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
17801	isobaric: TGV-SGL	1263.65	1263.65
23695	isobaric: N-GG	1781.87	1781.86
24551	isobaric: LA-AL	1212.72	1212.72
13495	unknown	1625.81	1611.79; 4413.31
9693	isobaric: DGK-GSR	1253.59	1253.60
11446	[1] non-isobaric: known S-P variant	1447.65	1390.63 (+57.02)
18984	isobaric: SG-GS	2588.28	2540.27
24193	[2] non-isobaric: includes E-Q	2098.94	2401.17
19351	unknown	1579.84	2458.49; 1811.96
13340	contaminant: keratin; isobaric: LA-AL	1300.65	1300.65
10210	isobaric: D-N(+0.98)	1282.69	1281.70 (+0.98)
21318	isobaric: Q-GA	1836.89	1836.89
28877	isobaric: Q(+0.98)Q(+0.98)-EE	1745.89	1745.88
18427	isobaric: DE-ED	1407.69	1407.69
20365	isobaric: Q(+0.98)-E	1771.91	1771.91
9849	isobaric: D-N(+0.98)	2135.96	2134.96 (+0.98)
15638	[3] non-isobaric: includes L-V	1907.90	1962.93
32901	isobaric: F-M(+15.99)	2272.12	2256.1 (+15.99)
20690	isobaric: D-N(+0.98)	1833.90	1832.92 (+0.98)
13360	isobaric: LS-SL	1478.78	1478.78
28218	[4] unexpected protein: includes D-N(+0.98)	1808.00	1807.00 (+0.98); 1039.51
26185	[5] unexpected protein	1383.60	1383.61; 2726.32
12869	isobaric: Q(+0.98)-E, Q(+0.98)-E	1468.73	1468.73
33351	isobaric: Q(+0.98)-E	2022.03	2022.01
26840	isobaric: EL-LE	1877.98	1877.98
15254	isobaric: DG-GD	1145.60	1145.60
13589	isobaric: GA-Q	1458.79	1458.79
28181	isobaric: FL-LF	1837.95	1837.95
21810	isobaric: Q(+0.98)-E	1759.79	1645.74 (+57.02, +57.02)
20965	isobaric: YA-SF	1459.78	1459.77

With this dataset, unlike the datasets from the other five organisms, asparagine and glutamine deamidation was specified as a variable modification in de novo sequencing and database search, consistent with the treatment of the data in the original study (Zhang et al., 2015).[28] This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be "incorrect" by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit's tryptic peptide, include amino acid modifications and sequence variants, such

(continued from previous page)
as polymorphisms, of a reference protein. Contaminants can be found by searching spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

Further annotations:
[1] transferrin (C2 variant), known polymorphism
[2] complement C4d
[3] immunoglobulin kappa light chain
[4] *Lactococcus lactis*: hypothetical protein
[5] *Saccharomyces cerevisiae* str. S288C: asparagine-tRNA ligase DED81

Table II.7. *Drosophila melanogaster* cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
12297	LLEVGNDGVAAG	LIEVGNNGVAAG	LIEVGNNGVAAG	3.9	0.002
25018	VDVEGVYSYLNK	VDVEGVYSYLNK	VDVEGKMYSY	0.001	0.60
11029	ALEESNYELEGK	ALEESNYELEGK	LQEDNYELE	0.003	0.30
37197	AALFLLNADAGK	[V]LFLLNADAGK	[V]LFLLNADAGK	1.4	6.E-04
21506	DLGQGVEVVLTK	DLGGGVEVILTK	DLGQVGEVVITK	11	0.21
24791	VVSGEQLQEAFR	VVSGEQLQEAFR	VSGEQLHE	0.004	0.60
6678	YSDYASNPAPGR	[SY]DYASNPAPGR	[SY]DYASNPAPGR	0.49	2.E-04
28395	AESLEATNLASNLR	ADNLEATDLASSLR	AESLAETNLASNIR	1.4	6.E-04
12602	LAAAGDLETMGAR	LAAAGDLETM	[I]AAAGDLETM[QR]	1.2	0.005
30308	EVQDLLQQYDSK	EVQDLIQQYASK	EVQDLIQQYASK	0.69	3.E-04
36437	SLLLMELLNNVAK	LLMELVNNVAK	[TV]LIMELINNVAK	0.59	0.003
19390	SDLVNQVGTLAK	DLVNQIGTLA	SDLVNNLGTIAK	7.9	0.10
20016	SEPLLDVGSPEK	SEPLLDVGSPEK	SEPLLDVCSPEK	0.008	6.E-04
25712	VACGAGVFDAVK	VACGAGAVFDAIK	VACGANYGVFHA	2.0	2.9
35779	LFGLNVELAQLK	LNVELAQLK	GLNVELSFTAVARQLK	11	1.5
12667	ALEEANNDLENK	ALEEANNDLENK	LQEANNDLAN	0.003	0.062
16770	VGVSDTALQCVSSAR	VSDTAAQCVSSAR	[GV]VSDTAPQCVSSAR	0.15	1.E-04
14374	VDNLGNNVTFER	VNNLGNNVTFER	VNNLGNNVTFER	0.011	7.E-06
22580	VTPAESALAELR	[A]TPAESALAELR	[A]TPAESALAELR	0.17	1.E-04
25594	QSGLCVSGLTLD	QSGLCVSGLTIN	QSGLCVSGLTIN	1.4	6.E-04
14585	NDGDGGLNSGYG	NNGNGGINSGYG	NNGNGGINSGYG	11	0.004
35639	PATANLLGLLAD	PATVNLLGLLAD	PATVNLLGLLAD	0.35	1.E-04
24582	YEGVDGLLEASAK	YEGVNGLLEASAK	YEGVNGLLEASAK	0.009	4.E-06
25309	VDSSDQLDELLR	VDSSDNLDELLR	[DV]SSDQLDEILR	0.17	0.004
10585	MNSLESGLSTAK	MNSLESGISTAK	MNSLESGISTA[T]	0.17	8.E-04
22410	YDLDALSTLDGK	LDALSTLDGK	YDIDASL-TLDGK	2.8	0.51
30902	VLTDGLDFLVSK	VITAGLDFLVSK	VITAGLDFLVSK	5.6	0.002
28807	TQGLQQLLAAEK	TQGIQQLLAAEK	TQGIQQLLAAEK	0.17	7.E-05
13612	LDEDTYEDFGAK	LDEDTYEDF[QK]	LDEDTYEDF[QK]	0.69	3.E-04
27787	LDEGVLVATVDNFK	[V]EEGVLVATVDNFK	[V]EEGVLVATVDNFK	0.04	2.E-06

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
12297	non-isobaric: D-N	1326.69	1325.71
25018	[1] unexpected protein	1384.69	1384.69; 2107.01
11029	contaminant: keratin	1380.64	1380.64
37197	[2] non-isobaric: AA-V (Δm = -43.00)	1202.66	1159.66
21506	isobaric: GV-VG	1499.82	1442.80
24791	[3] unexpected protein	1675.85	1675.85; 1857.88
6678	isobaric: YS-SY	1296.57	1296.57
28395	isobaric: EA-AE	2045.99	2045.99
12602	isobaric: GA-Q	1274.63	1274.63
30308	[4] non-isobaric: includes D-A	1824.86	1828.95
36437	isobaric: SL-TV	1456.83	1456.83
19390	isobaric: QV-NL	1243.68	1243.68
20016	[5] non-isobaric: G-C (Δm = 45.99)	1510.79	1556.78
25712	unknown	1419.72	1717.91; 2075.04
35779	unknown	1343.78	1212.67; 1476.79
12667	contaminant: keratin	1585.76	1585.76
16770	[6] non-isobaric L-P (Δm = -16.05); isobaric: VG-GV	1548.76	1475.70 (+57.02)
14374	non-isobaric: D-N	1376.67	1375.68
22580	[7] non-isobaric: includes V-A	1255.68	1477.79
25594	non-isobaric: D-N	1744.91	1686.91 (+57.02)
14585	non-isobaric: D-N, D-N	2522.05	2520.08
35639	[8] non-isobaric: A-V (Δm = 28.03)	1712.92	1740.95
24582	non-isobaric: D-N	1679.83	1678.84
25309	isobaric: VD-DV	1388.68	1388.68
10585	[9] non-isobaric: includes K-T	1236.60	1365.66
22410	isobaric: LS-SL	1309.64	1309.64
30902	[10] non-isobaric: D-A (Δm = -43.99)	1305.72	1261.73
28807	[11] non-isobaric	1584.85	1626.86
13612	isobaric: GA-Q	1401.60	1401.59
27787	isobaric: LD-VE	1518.79	1518.79

This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be “incorrect” by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit’s tryptic peptide, include amino acid modifications and sequence variants, such as polymorphisms, of a reference protein. Contaminants can be found by searching

(continued from previous page)
spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

Further annotations:
[1] *Wolbachia* endosymbiont of *Drosophila*: P44/Msp2 family outer membrane protein
[2] NADH dehydrogenase (ubiquinone) 75 kDa subunit, isoform A
[3] *Wolbachia* endosymbiont of *Drosophila*: chaperonin GroEL
[4] gamma-aminobutyric acid transaminase, isoform A
[5] mitochondrial ribosomal protein L9 sequence variant found in other *Drosophila* species proteomes
[6] uncharacterized protein Dmel_CG11208
[7] mitochondrial ribosomal protein S22
[8] ATP synthase, oligomycin sensitivity conferring protein
[9] uncharacterized protein Dmel_CG11790, isoform A: de novo sequence is identical to thioredoxin domain-containing protein found in other *Drosophila* species
[10] thioester-containing protein 2, isoform A
[11] vacuolar H[+] ATPase 13kD subunit, isoform A

Table II.8. *Escherichia coli* str. K-12 substr. MG1655 cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
180522	LVGVVAGGGVALLR	GVVAGGGVALLR	GVVAGGGVALIR	0.13	8.E-05
75974	ELADGVEGYLR	ELADGVEGYLR	ELADGVEGYLR	0.043	2.E-06
165328	LESLTEELAYLK	LESLTEELAYLK	LEALIEEL	0.004	1.0
121758	GSDVYWTSFTEL	SDVHWTAFTEL	GSDVYSVTSFTEL	2.8	3.E-04
103538	LAGTEVDALLGR	LAGTEVSALLGR	LAGTEVSALLGR	1.4	6.E-05
17797	LGEVGNAEHYLR	LGEVGNAEHNLR	LGEVGNAEHMLR	0.25	1.E-04
65543	ALLNADGENAWK	ALLNANGENAWK	ALLNANGENAWK	0.015	6.E-07
136255	EELDTELLNLLR	EELNTELLNLLR	EELNTELLNLLR	0.010	7.E-07
118318	GYDGDYFLVYPLK	GYDGDYFLVYPIK	GYNGDYFLVYPIK	4.E-07	1.E-07
118894	FTALTVVGDGDGR	FTALTVVGDGDGR	FTALTVVGDGNGR	0.001	4.E-07
162915	LVYDALETLAQR	VYNALETLAQR	[I]VYSALETLAQR	0.25	8.E-05
89904	ADAVEAAGVEVAK	ADAVSAAGVEVAK	ADAVTAAGVEVAK	0.30	3.E-05
44281	LNALEEVTLASK	NALEEVTLSS	[I]NALETVTIASK	16	0.015
71768	LLDVLAEQAELSK	LLDVLAEQAE	LIDVIAEKAELSK	1.2	6.E-04
120369	DVSASLYGVVGVG	ASLYGVVGVG	[DW]ASIYGVVGVG	4.8	8.E-04
147111	PVVTEEDELVGL	PVVTEENELVG[I]	PVVTEENELVG[I]	0.35	1.E-05
53110	LVADSLESQLER	LVADSIASQLER	LVADSITSQLER	4.0	7.E-04
78259	LALESVLLGDEK	ALESVLLGDQK	[I]ALESVLLGD[KE]	2.0	0.001
126833	LLLDDSLFSLPK	LFDDSLFSLPK	LLLNDTLF[LS]PK	2.0	0.25
53581	LAELPTYEEALAR	LATLPTYEEALAR	LATLPTYEEAIAR	0.038	2.E-05
116447	SFTALTVVGDGDGR	SFTALTVVGDGDGR	SFTALTVVGDGNGR	2.E-04	6.E-08
77352	ALELGPQGLLTTLK	ELGPQGLLTTL	[LA]ELGPQGLITTLK	0.18	1.E-05
66248	APDNVAQAVLEAR	APNNVAQAVIEAR	APSNVAQAVIEAR	0.10	3.E-05
117680	VDLLNQELEFLK	VDLLNQELEFLK	MLNQELE	0.011	0.25
112417	FLLANLDGFDPAK	FLLANLNGFDPAK	FLLANLNGFDPAK	0.002	1.E-07
28170	LVADSLSSQLER	LVADSISSQLER	LVADSITSQLER	0.25	8.E-05
88446	GLGTNYEEFGVR	LGANYEEFGV	GLGTNYEE	3.9	0.003
67735	LLDQAEAELVET	LDQAEAELTET	LIDQATAEIVET	2.0	0.008
96322	LGDGVVLSPAAL	[I]GNGVVLSPAAL	[I]GNGVVLSPAAL	3.9	2.E-04
94849	SDDPEVLLEALR	SDQPEVLLEAL	[DS]DPEVLLEAIR	2.0	3.E-04

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
180522	isobaric	1566.94	1566.87
75974	isobaric	1548.78	1548.78
165328	contaminant: keratin	2111.03	2095.04 (+15.99)
121758	isobaric: W-SV	2064.96	2064.96
103538	[1] non-isobaric: D-S (Δm = -28.00)	1477.78	1449.78
17797	[2] non-isobaric: Y-M (Δm = -32.02)	1528.73	1496.74
65543	non-isobaric: D-N	1484.76	1483.78
136255	non-isobaric: D-N	1642.88	1641.89
118318	non-isobaric: D-N	1927.89	1926.90
118894	non-isobaric: D-N	1653.84	1923.00
162915	[3] non-isobaric: D-S (Δm = -28.00)	1865.94	1837.94
89904	[4] non-isobaric: E-T (Δm = -27.99)	1456.74	1428.75
44281	[5] non-isobaric: E-T (Δm = -27.99)	1286.71	1258.71
71768	isobaric	1656.93	1656.93
120369	isobaric: VS-W	1796.91	1796.91
147111	non-isobaric: D-N	2336.15	2334.19
53110	[6] non-isobaric: E-T (Δm = -27.99)	1358.70	1330.71
78259	[7] isobaric: EK-KE	1285.71	1285.71
126833	isobaric: SL-LS	1359.76	1313.77
53581	[8] non-isobaric: E-T (Δm = -27.99)	1474.77	1446.77
116447	non-isobaric: D-N	1653.84	1923.00
77352	isobaric: AL-LA	1452.86	1452.86
66248	[9] non-isobaric: D-S (Δm = -28.00)	1759.84	1731.84
117680	contaminant: keratin	1459.79	1459.79
112417	non-isobaric: D-N	1419.74	1418.76
28170	[10] non-isobaric: S-T (Δm = 14.02)	1316.69	1330.71
88446	isobaric	1595.79	1595.80
67735	[11] non-isobaric: E-T (Δm = -27.99)	1528.80	1500.80
96322	non-isobaric: D-N	2029.08	2012.10
94849	isobaric: SD-DS	1355.69	1355.69

This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be “incorrect” by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit’s tryptic peptide, include amino acid modifications and sequence variants, such as polymorphisms, of a reference protein. Contaminants can be found by searching spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

(continued from previous page)
Further annotations:
[1] ATP synthase subunit beta
[2] 50S ribosomal protein L2
[3] 30S ribosomal protein S7
[4] 50S ribosomal protein L9
[5] 50S ribosomal protein L9
[6] 30S ribosomal protein S3
[7] It seems that there was a missed K-E cleavage at the C-terminal end of purine-nucleoside phosphorylase
[8] 50S ribosomal protein L10
[9] elongation factor G
[10] 30S ribosomal protein S3
[11] 30S ribosomal protein S10

Table II.9. *Desulfovibrio vulgaris* str. Hildenborough cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
60266	DGLEGGLAEVVK	NGLEGGLAEVVK	NGLEGGLAEVVK	0.17	5.E-06
60639	LTDATFEASVLK	[I]TDATFEASVLK	[I]TDATFEASVLK	0.13	4.E-06
39923	VQGLDGDLGNLR	VQGLNGDIGNLR	VQGLNGDIGNLR	0.99	3.E-05
51806	YGSVQADSEETTER	YGSVQADSEE[WER]	YGSVQADSEE[WER]	1.0	3.E-05
26800	STVTAGLAAVGK	[TS]VTAGLAAVGK	[TS]VTAGLAAVGK	11	3.E-04
36096	LTAMDATEGLVR	[TI]AMDATEGLVR	[TI]AMDATEGLVR	0.49	1.E-05
38941	LPGAMEFPLVAK	LPGAVELPLVAK	[I]PGAFEMPLVAK	8.0	0.011
46106	FVAENMGNVPAK	FVAENFGNVPAK	FVAENFGNVPAK	0.25	7.E-06
75253	LQPLPAAELAAL	LRPLPAAELAAL	LQPIPAAELAA[I]	0.48	8.E-05
18377	ALTGLGLSALAK	AMTGLALSALAK	ALTGLGL	16	0.031
58910	LLGLLSGTGAAN	GLLSGTGAAN	GLLSGTGAAN	11	3.E-04
12987	GDAVQSMQSQAR	GDAVQAMQSQAR	GDAVQAMQSQAR	0.030	9.E-07
19849	PAAAAGVQSAEK	PAAAAPGVQSAE	PAAAAGVQS[GDR]	22	0.003
71307	GVLEGLQEAEAL	VLEGLQEAEA	VIEGIQEAEAL	1.4	0.001
16201	ALNASGAETVHVAK	ALNASGATEVHVAK	ALNASGATEVHVAK	4.1	1.E-04
51043	LNATAEGDVLVPR	ATAEGDVLVAR	LNAT-EQDVIVPR	4.8	0.009
72352	LTPLDLSLVDAE	[TL]PLDLSLVDAE	[TL]PLDLSLVDAE	0.98	3.E-05
61099	YGLSPYFVTDPEK	[GY]LSPYFVTDPEK	[GY]LSPYFVTDPEK	0.013	4.E-07
75565	ENLSLVAEFGYLK	ENLSLVAEMGYIK	ENLSLVAEMGYIK	0.42	1.E-05
30865	ALGLSGGEAALR	GLSGGEAALR	ALGLSGGQ[QQR]	7.8	0.35
77730	PELNAGALAAVR	PELDAAALAAVR	PELNAGALAGLR	2.8	2.E-04
75386	LESGPWPSMVSDLK	LESGPWPSFVSDLK	LESGPWPSFVSDIK	0.001	4.E-07
22226	LGLEAASSGDLK	LEAANSGDLK	[AV]LEAASSG[LDK]	45	0.37
61418	LNMPNFDGLELLR	LNMPNMDGLELIR	LNMPNMDGIELIR	0.054	3.E-05
17380	VLGAEHPTEEAR	VLGAEHPQTEESR	VIGAEPHTEEAR	0.98	0.010
72847	NVGLVTEADFLK	VGIVTEADFLK	[L]VGIVTEADFLK	1.4	4.E-05
62184	LTADDLSEAVLAG	LTAIDELSEAVLAG	[I]TANDLSEAVLA	0.84	7.E-05
50470	GSTDYGLLQLNSR	GSTDYGILQINSR	GLLQLSTR	0.15	0.89
63595	VLDFDSLYDVPK	VIDFNSLYDVPK	VIDFNSLYDVPK	0.13	4.E-06
17037	ASANVVPSDQMK	ASSNVIPSAQMK	ASANVVPSNGAMK	7.9	0.002

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
60266	non-isobaric: D-N	1185.62	1184.64
60639	[1] non-isobaric	1563.81	1694.85
39923	non-isobaric: D-N	1255.65	1254.67
51806	non-isobaric: TT-W(+31.98)	2002.82	1970.83 (+31.98)
26800	isobaric: ST-TS	1073.61	1073.61
36096	isobaric: LT-TI	1291.64	1275.65
38941	isobaric: MEF-FEM	1287.69	1271.69
46106	isobaric: M(+15.99)-F	1752.81	1736.82
75253	non-isobaric	2446.31	2445.32
18377	unknown	1113.68	2818.34; 4726.21
58910	isobaric	1626.91	1626.91
12987	[2] non-isobaric: S-A (Δm = -15.99)	1504.69	1488.70
19849	isobaric: AEK-GDR	1553.78	1553.78
71307	isobaric	1770.89	1770.90
16201	isobaric: ET-TE	1366.72	1366.72
51043	isobaric: AEG-EQ	1691.92	1691.92
72352	isobaric: LT-TL	1555.81	1555.81
61099	isobaric: YG-GY	1514.73	1514.73
75565	isobaric: F-M(+15.99)	1757.89	1741.90
30865	isobaric: EAAL-QQQ	1584.79	1584.79
77730	isobaric: AV-GL	1618.85	1618.85
75386	isobaric: M(+15.99)-F	1916.96	1916.95
22226	isobaric: LG-AV, DL-LD	1159.61	1159.61
61418	isobaric: F-M(+15.99)	2122.00	2074.03
17380	isobaric: HP-PH	1491.77	1491.77
72847	[3] non-isobaric: N-L (Δm = -0.96)	1304.70	1303.74
62184	non-isobaric: D-N	1530.79	1529.80
50470	unknown	1752.83	3143.42; 1215.68
63595	non-isobaric: D-N	1409.71	1408.72
17037	[4] non-isobaric: D-N (Δm = +0.98); isobaric: Q-GA	1559.76	1542.78 (+15.99)

This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be “incorrect” by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit’s tryptic peptide, include amino acid modifications and sequence variants, such as polymorphisms, of a reference protein. Contaminants can be found by searching spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

(continued from previous page)
Further annotations:
[1] thioredoxin
[2] methyl-accepting chemotaxis protein
[3] CBS domain-containing protein
[4] phosphate ABC transporter substrate-binding protein

Table II.10. *Rhodopseudomonas palustris* str. TIE-1 cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
62647	YGSSDSQTLGDL	YGSTDSQTLGN	YGSSDSGATLGD	8.0	0.10
58410	LFGASGVGFYVS	LFGASGVGMYVS	LFGASGVGMYVS	0.99	4.E-04
65604	ALLTNQALSEDL	LLTNQALSQNL	AILTGGQAISEDL	7.9	1.7
52274	VSDNAGVDGLGLS	VSNNAGVDGIGLS	VSNNAGVDGIGLS	0.60	2.E-04
42612	LDCSSNLLGSATA	SSNLLGSATA	[I]DCSSGGLLGSATA	14	0.015
61335	GDGSSVAGALSDF	GNGSSVAGALSDF	GNGSSVAGALSDF	0.052	2.E-05
34704	ELVEGSDFTVAR	[LE]VEGSDFTVAR	[LE]VEGSDFTVAR	0.99	4.E-04
53331	DSYAGTSLPDLVGK	[N]SYAGTSIPDIVGK	[N]SYAGTSIPDIVGK	0.25	1.E-04
64786	PLSDFSNASFLE	PLSDFSNDEQDDASFLE	PISDFSNASMLE	0.50	0.002
66183	LTDLTMLPVLEK	LTDLTLLPVL	[I]TDLTFIPVLEK	7.9	0.018
62295	AALEGFEFDGLGDGK	AALEGFEFDGLGNGK	AALEGFEFDGLGNGK	7.E-05	3.E-08
40795	VSTLDGDANVPFYK	VSTLNGDANVPFYK	VSTINGDANVPFYK	2.E-04	1.E-06
53606	LEGTDGLALGPLLK	LEGTDGVAIGPMLK	LEGTDVAAIGPLLK	0.72	0.006
42562	ALAGSGAYNSPAWA	LAGSGAYNSPAWA	LAGSGAYNSPAWA	7.E-04	3.E-07
51824	TELDNNLEQLSSYK	TEIDNNIEQMSSYK	ELEAELEQLSRY	0.031	0.30
47115	ATLQGTGLGVASLK	ATLQSAGLGVASLK	ATLQSAGLGVASLK	0.25	1.E-04
43311	STLDGDANVPFYK	STLNGDANVPFYK	STINGDANVPFYK	0.002	1.E-05
56523	TPTLNLDFADVR	LNLDFADVR	[EGLI]NLDFADVR	4.0	0.018
61449	LVVLGGALEVSDK	LVVLNGALEVTD	[I]VVLGGIAEVSDK	6.8	0.062
60569	GSSDSQTLGDLL	GSSDS-TLGDLL	GSSDSGATLGDIL	16	0.42
25828	EGTELGTATLTESGK	EGTEIGTATLTESSK	EGTEIGTATLTEGSK	0.010	4.E-06
43138	DCSSNLLGSATANR	DCSSGGLLGSATANR	DCSSGGLLGSATANR	0.18	7.E-05
41939	NLTVEDGLMTAR	NITVENGIMTAR	NITVENGIMTAR	4.0	0.002
64726	SSLSLLEAAELAK	SLSLLEAAELVK	SSLTVLEAAELAK	0.86	5.E-04
23955	LTGEGENVAGFAK	LTGEGENVAFDGMFNDDGFA	[I]TGEGEGGVAGFAK	6.8	0.011
46539	NDDGLVAAAVLS	NDNGLVAAAVLS	NDNGIVAAAVLS	0.24	0.002
54671	SSLQATDALLATT	SSLQATADLLATT	SSLQATADLLATT	4.8	0.002
33496	LMNSNASADLTGK	LMNSGGASADLTGK	LMNSGGASADLTGK	1.2	5.E-04
54283	VADDDAAGALYR	VADAIEGGRFDDAAGALYR	VADDDAA	2.0	1.2
53769	LALEGTDGLALGPLLK	ALEGTDGVAIGPMLK	[I]ALEGTDVAAIGPLLK	0.13	0.001

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
62647	isobaric: Q-GA	2337.17	2337.17
58410	isobaric: F-M(+15.99)	2399.11	2383.12 (+15.99)
65604	isobaric: N-GG	2272.20	2256.20 (+15.99)
52274	[1]non-isobaric: includes D-N	1834.91	1784.94
42612	isobaric: N-GG	1933.92	1876.89 (+57.02)
61335	non-isobaric: D-N	2014.99	1998.01 (+15.99)
34704	isobaric: EL-LE	1321.65	1321.65
53331	[2]non-isobaric: includes D-N	2044.02	9195.33
64786	isobaric: F-M(+15.99)	2060.99	2044.99 (+15.99)
66183	isobaric: M(+15.99)-F	1633.90	1617.91 (+15.99)
62295	non-isobaric: D-N	1736.86	1735.88
40795	non-isobaric: D-N	1795.90	1794.92
53606	isobaric: GL-VA	1579.92	1579.92
42562	isobaric	2176.05	2176.05
51824	contaminant: keratin	1995.96	1995.96
47115	isobaric: GT-SA	1626.93	1626.93
43311	non-isobaric: D-N	1795.90	1794.92
56523	isobaric: TPT-EGL	1360.70	1360.70
61449	isobaric: AL-IA	1298.75	1298.74
60569	isobaric: Q-GA	2337.17	2337.17
25828	isobaric: SG-GS	1677.81	1677.81
43138	isobaric: N-GG	1933.92	1876.89 (+57.02)
41939	non-isobaric: D-N	1667.78	1666.80
64726	isobaric: SL-TV	1886.03	1886.02
23955	isobaric: N-GG	1291.64	1291.64
46539	non-isobaric: D-N	1670.86	1669.87
54671	isobaric: DA-AD	2017.04	2017.04
33496	isobaric: N-GG	1634.79	1634.79
54283	unknown	1754.79	2094.01; 1302.65
53769	isobaric: GL-VA	1579.92	1579.92

This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be “incorrect” by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit’s tryptic peptide, include amino acid modifications and sequence variants, such as polymorphisms, of a reference protein. Contaminants can be found by searching spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

(continued from previous page)
Further annotations:
[1] CoA transferase subunit A
[2] porin

Table II.11. *Synechococcus sp.* WH7803 cross-validation

Scan	De novo query sequence	Top RefSeq hit	Top reference proteome hit	E-value of top RefSeq hit	E-value of top reference proteome hit
31096	VSDSLLNADAEAR	VSNSIINADAEAR	VSNSIINADAEAR	3.4	6.E-05
27369	SNSLLDADAEAR	SNSVLDANAEAR	SNSIINADAEAR	2.8	6.E-04
22478	YTEYWDGDEPAR	EFWDGDEPAR	YTEYSVDGDEPAR	0.35	2.E-05
36663	EAVAETGEELLEK	EAVAETSEELLEK	EAVAETDEALIEK	0.013	2.E-04
13769	LESTASAPDLAR	LESSMSAPDLAR	LESTASPDALAR	11	0.081
21220	GVNLGAGTVGGLGK	DLGAGTVGGLG	GVNLGATGVGGIGK	12	0.002
10343	SVESATESTTTR	SVESATESTTAR	ESATESTTTR	0.34	4.E-05
27274	VSNSLLDADAEAR	ISNSLLDEDAEA	VSNSIINADAEAR	2.4	6.E-05
20092	VTLVSESEGLDK	VTLVSESEGLNK	VTLVSESEGLNK	0.13	2.E-06
30287	LNGNDSALQLLR	LNGNDSALELL	LNGNDGTLQLLR	1.4	3.E-05
35401	LTVGFDLAPLGLK	LTVGMDLAPLGLK	LTVGMDLAPLGLK	0.11	2.E-06
12839	AQEGSTASNLLK	AQEGSTGTNLLK	AQEGSTGTNLLK	4.0	1.E-05
17693	ALEESNYELEGK	ALEESNYELEGK	LEESDIERLE	0.003	2.8
13455	FLADSDGDSGPR	FLADSDGDEAPR	FLADSDGDS[PGR]	3.9	3.E-04
27070	SNLQQSLSDAEQR	SNLQQSISDAEQR	LEQSLSDAQ	0.009	0.056
34703	AGGGLASDLVSR	AGGNLASELVSR	GDLSSDLVTR	16	0.16
37578	LPLAVALGLALK	LPLAVALGLAL	MPLALALGLAL	0.49	0.002
20646	WVSGGAVAMTTK	WVSGGAVAM	WVSGGAV[WTTK]	3.9	0.028
20479	TTTNVLQGSLHR	TTNVLQGSLQ	VLAPQGSLH	16	2.0
24989	LLAADAESLVAR	LAADAESLVAR	[II]AANAESIVAR	0.34	0.010
19986	AQGSAMDSPASLR	GSAMDSNASLR	AQGS--WDSPASLR	19	0.012
13479	ALEGEAMPSEAK	AVEGEAMPSE	LEVEALPSE	11	0.12
33388	SSSPENPDLAASMA	SSSPENPDLAAS	SSSPENPDLAAS	0.031	6.E-07
12365	LGETNTQADGQK	LGETNTRADGQ	GETNTQANG[GAK]	2.8	0.002
11547	ALEGEAESMEAK	ALEVEAEAMEAK	AIEGEAEW-EAK	2.0	0.018
35750	SSAVSPVSLALL	AVSPVSLALL	SSAVSPDALALL	5.6	0.002
18005	AAANPDGLVALAK	AANPDVLVALAK	AAANPDAVVAIAK	1.7	5.E-04
24052	SLSVSSLKPLGDR	SLAISSLKPLGDR	SLSVSTVKPLGDR	0.15	1.E-05
24696	FALKPTSLSDEVR	ALKPTSLTDEVR	FALKPSTISDEVR	0.10	2.E-05
26788	PVNSQLCMVGLK	NSALCMVGLK	PVNSQ--SWMVGLK	11	0.014

(continued from previous page)

Scan	Explanation for database search discrepancy (red annotations in caption)	De novo mass	Mass of tryptic peptide in protein hit
31096	non-isobaric: D-N	1559.78	1558.80
27369	non-isobaric: D-N	1559.78	1558.80
22478	isobaric: W-SV	1757.77	1757.77
36663	isobaric: GEE-DEA	1940.93	1924.93 (+15.99)
13769	isobaric: APD-PDA	1229.63	1229.63
21220	isobaric: GT-TG	1198.67	1198.67
10343	[1] unknown	1569.71	1539.70; 1081.49
27274	non-isobaric: D-N	1559.78	1558.80
20092	non-isobaric: D-N	1275.65	1274.67
30287	isobaric: SA-GT	1511.81	1511.81
35401	isobaric: F-M(+15.99)	1584.88	1568.88 (+15.99)
12839	isobaric: AS-GT	1217.63	1217.63
17693	contaminant: keratin	1380.64	1380.64
13455	isobaric: GP-PG	1235.54	1235.54
27070	contaminant: keratin	1715.85	1715.84
34703	unknown	1644.81	1326.69; 1691.84
37578	unknown	1177.78	6467.56; 1508.91
20646	non-isobaric: AM-W(+15.99)	1206.61	1190.61 (+15.99)
20479	unknown	1509.84	8674.45; 1891.98
24989	non-isobaric: D-N	1227.68	1226.70
19986	non-isobaric: AM-W(+15.99)	1637.74	1621.75 (+15.99)
13479	unknown	1247.57	2915.32; 2624.33
33388	isobaric	2117.01	2117.03
12365	[2] non-isobaric: includes D-N; isobaric: Q-GA	1812.82	1797.86
11547	non-isobaric: SM(+15.99)-W(+31.98)	1263.57	1231.57 (+31.98)
35750	isobaric: VS-DA	1813.00	1812.99
18005	isobaric: GL-AV	1209.67	1209.67
24052	isobaric: SL-TV	1598.90	1598.90
24696	isobaric: TS-ST	1461.78	1461.78
26788	isobaric: LC(+57.02)-SW	1575.77	1559.78 (+15.99)

This table considers the 30 top-scoring de novo sequences of at least 12 amino acids that were determined to be “incorrect” by comparison to database search results. These sequences were searched by BLAST against RefSeq (updated Feb. 1, 2018) and the reference proteome used in the database search. In some cases, the top RefSeq and reference proteome hits were the same. The top hits are shown here, and the better hit, as determined by alignment E-value, peptide mass difference, and sequence similarity, is highlighted. The sequence differences are shown in red, with those shown in square brackets not included in the reported sequence alignment. An explanation of why each sequence did not match the database search is provided – those in green indicate a Postnovo sequence that is likely to be correct due to a full explanation of the discrepancy. Isobaric discrepancies, in which the mass of the precursor peptide spectrum mass agrees with that of the corresponding tryptic peptide in the best BLAST hit, are due to de novo sequencing errors. Non-isobaric discrepancies, in which the precursor peptide spectrum mass does not agree with the best hit’s tryptic peptide, include amino acid modifications and sequence variants, such as polymorphisms, of a reference protein. Contaminants can be found by searching spectra against common contaminant protein sequences. Unexpected proteins are proteins from other organisms that are not represented in the reference proteome.

(continued from previous page)
Further annotations:
[1] The de novo peptide mass does not match the mass of the tryptic peptides from either the RefSeq (*Synechococcus sp.* WH7805) or reference proteome (WH7803) hit. The two hits are orthologous proteins (NADPH-dependent assimilatory sulfite reductase hemoprotein subunit), and the de novo sequence is found in the WH7805 hit but not the WH7803 hit. The partial length de novo sequence is SVESATESTTAR, and the tryptic peptides from the proteins are N-terminal: MSQSSVESATESTTAR from WH7805 and MESATESTTTR from WH7803. The matching SV amino acids between the de novo sequence and WH7805, which are missing in WH7803, are also found in orthologous proteins from other *Synechococcus* strains.
[2] C-phycoerythrin class 1 subunit beta

II.H. REFERENCES

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society of Mass Spectrometry* **1994**, *5* (11), 976–989.

(2) Verheggen, K.; Raeder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and Evolution of Database Search Engines – A Central Component of Mass Spectrometry Based Proteomic Workflows. *Mass Spectrometry Reviews* **2017**, 1-15, https://doi.org/10.1002/mas.21543.

(3) Medzihradszky, K. F.; Chalkley, R. J. Lessons in de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Mass Spectrometry Reviews* **2015**, *34* (1), 43–63.

(4) Taylor, J. A.; Johnson, R. S. Sequence Database Searches via de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **1997**, *11* (9), 1067–1075.

(5) Bartels, C. Fast Algorithm for Peptide Sequencing by Mass Spectroscopy. *Biological Mass Spectrometry* **1990**, *19* (6), 363–368.

(6) Tanca, A.; Palomba, A.; Fraumene, C.; Pagnozzi, D.; Manghina, V.; Deligios, M.; Muth, T.; Rapp, E.; Martens, L.; Addis, M. F.; et al. The Impact of Sequence Database Choice on Metaproteomic Results in Gut Microbiota Studies. *Microbiome* **2016**, *4*, 51.

(7) Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and Perspectives of Metaproteomic Data Analysis. *Journal of Biotechnology* **2017**, *261* (Supplement C), 24–36.

(8) Muth, T.; Kolmeder, C. A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; et al. Navigating through Metaproteomics Data: A Logbook of Database Searching. *Proteomics* **2015**, *15* (20), 3439–3453.

(9) Wilmes, P.; Andersson, A. F.; Lefsrud, M. G.; Wexler, M.; Shah, M.; Zhang, B.; Hettich, R. L.; Bond, P. L.; VerBerkmoes, N. C.; Banfield, J. F. Community Proteogenomics Highlights Microbial Strain-Variant Protein Expression within Activated Sludge Performing Enhanced Biological Phosphorus Removal. *ISME Journal* **2008**, *2* (8), 853–864.

(10) Denef, V. J.; Kalnejais, L. H.; Mueller, R. S.; Wilmes, P.; Baker, B. J.; Thomas, B. C.; VerBerkmoes, N. C.; Hettich, R. L.; Banfield, J. F. Proteogenomic Basis for Ecological Divergence of Closely Related Bacteria in Natural Acidophilic Microbial Communities. *Proceedings of the National Academy of Science U. S. A.* **2010**, *107* (6), 2383–2390.

(11) Devabhaktuni, A.; Elias, J. E. Application of de Novo Sequencing to Large-Scale Complex Proteomics Data Sets. *Journal of Proteome Research* **2016**, *15* (3), 732–742.

(12) Bandeira, N.; Pham, V.; Pevzner, P.; Arnott, D.; Lill, J. R. Automated de Novo Protein Sequencing of Monoclonal Antibodies. *Nature Biotechnology* **2008**, *26* (12), 1336–1338.

(13) Tran, N. H.; Rahman, M. Z.; He, L.; Xin, L.; Shan, B.; Li, M. Complete *De Novo* Assembly of Monoclonal Antibody Sequences. *Scientific Reports* **2016**, *6*, 31730.

(14) Guthals, A.; Gan, Y.; Murray, L.; Chen, Y.; Stinson, J.; Nakamura, G.; Lill, J. R.; Sandoval, W.; Bandeira, N. De Novo MS/MS Sequencing of Native Human Antibodies. *Journal of Proteome Research* **2017**, *16* (1), 45–54.

(15) Bringans, S.; Eriksen, S.; Kendrick, T.; Gopalakrishnakone, P.; Livk, A.; Lock, R.; Lipscombe, R. Proteomic Analysis of the Venom of *Heterometrus longimanus* (Asian Black Scorpion). *Proteomics* **2008**, *8* (5), 1081–1096.

(16) Brinkman, D. L.; Aziz, A.; Loukas, A.; Potriquet, J.; Seymour, J.; Mulvenna, J. Venom Proteome of the Box Jellyfish *Chironex fleckeri*. *PLoS One* **2012**, *7* (12), e47866.

(17) Carregari, V. C.; Dai, J.; Verano-Braga, T.; Rocha, T.; Ponce-Soto, L. A.; Marangoni, S.; Roepstorff, P. Revealing the Functional Structure of a New PLA2 K49 from *Bothriopsis taeniata* Snake Venom Employing Automatic "de Novo" Sequencing Using CID/HCD/ETD MS/MS Analyses. *Journal of Proteomics* **2016**, *131*, 131–139.

(18) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society of Mass Spectrometry* **2015**, *26* (11), 1885–1894.

(19) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. De Novo Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Science U. S. A.* **2017**, *114* (31), 8247–8252.

(20) Chi, H.; Sun, R.-X.; Yang, B.; Song, C.-Q.; Wang, L.-H.; Liu, C.; Fu, Y.; Yuan, Z.-F.; Wang, H.-P.; He, S.-M.; et al. PNovo: De Novo Peptide Sequencing and Identification Using HCD Spectra. *Journal of Proteome Research* **2010**, *9* (5), 2713–2724.

(21) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R.-X.; Liu, J.; Zeng, W.-F.; Song, C.-Q.; He, S.-M.; et al. PNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *Journal of Proteome Research* **2013**, *12* (2), 615–625.

(22) Robotham, S. A.; Horton, A. P.; Cannon, J. R.; Cotham, V. C.; Marcotte, E. M.; Brodbelt, J. S. UVnovo: A de Novo Sequencing Algorithm Using Single Series of Fragment Ions via Chromophore Tagging and 351 nm Ultraviolet Photodissociation Mass Spectrometry. *Analytical Chemistry* **2016**, *88* (7), 3990–3997.

(23) Horton, A. P.; Robotham, S. A.; Cannon, J. R.; Holden, D. D.; Marcotte, E. M.; Brodbelt, J. S. Comprehensive de Novo Peptide Sequencing from MS/MS Pairs Generated through Complementary Collision Induced Dissociation and 351 nm Ultraviolet Photodissociation. *Analytical Chemistry* **2017**, *89* (6), 3747–3753.

(24) Blank-Landeshammer, B.; Kollipara, L.; Biß, K.; Pfenninger, M.; Malchow, S.; Shuvaev, K.; Zahedi, R. P.; Sickmann, A. Combining De Novo Peptide Sequencing Algorithms, A Synergistic Approach to Boost Both Identifications and Confidence in Bottom-up Proteomics. *Journal of Proteome Research* **2017**, *16* (9), 3209–3218.

(25) Muth, T.; Renard, B. Y. Evaluating de Novo Sequencing in Proteomics: Already an Accurate Alternative to Database-Driven Peptide Identification? *Briefings in Bioinformatics* **2017**, https://doi.org/10.1093/bib/bbx033.

(26) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nature Methods* **2007**, *4* (11), 923–925.

(27) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* **2002**, *74* (20), 5383–5392.

(28) Zhang, Y.; Li, Q.; Wu, F.; Zhou, R.; Qi, Y.; Su, N.; Chen, L.; Xu, S.; Jiang, T.; Zhang, C.; et al. Tissue-Based Proteogenomics Reveals That Human Testis Endows Plentiful Missing Proteins. *Journal of Proteome Research* **2015**, *14* (9), 3583–3594.

(29) Hampoelz, B.; Mackmull, M.-T.; Machado, P.; Ronchi, P.; Bui, K. H.; Schieber, N.; Santarella-Mellwig, R.; Necakov, A.; Andrés-Pons, A.; Philippe, J. M.; et al. Pre-Assembled Nuclear Pores Insert into the Nuclear Envelope during Early Development. *Cell* **2016**, *166* (3), 664–678.

(30) Erde, J.; Loo, R. R. O.; Loo, J. A. Enhanced FASP (EFASP) to Increase Proteome Coverage and Sample Recovery for Quantitative Proteomic Experiments. *Journal of Proteome Research* **2014**, *13* (4), 1885–1895.

(31) Cypryk, W.; Lorey, M.; Puustinen, A.; Nyman, T. A.; Matikainen, S. Proteomic and Bioinformatic Characterization of Extracellular Vesicles Released from Human Macrophages upon Influenza A Virus Infection. *Journal of Proteome Research* **2017**, *16* (1), 217–227.

(32) Nevo, N.; Thomas, L.; Chhuon, C.; Andrzejewska, Z.; Lipecka, J.; Guillonneau, F.; Bailleux, A.; Edelman, A.; Antignac, C.; Guerrera, I. C. Impact of Cystinosin Glycosylation on Protein Stability by Differential Dynamic Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). *Molecular and Cellular Proteomics* **2017**, *16* (3), 457–468.

(33) Hu, H.; Bienefeld, K.; Wegener, J.; Zautke, F.; Hao, Y.; Feng, M.; Han, B.; Fang, Y.; Wubie, A. J.; Li, J. Proteome Analysis of the Hemolymph, Mushroom Body, and Antenna Provides Novel Insight into Honeybee Resistance against *Varroa* Infestation. *Journal of Proteome Research* **2016**, *15* (8), 2841–2854.

(34) Mata, C.I.; Fabre, B.; Hertog, M.L.; Parsons, H.T.; Deery, M.J.; Lilley, K.S.; Nicolaï, B.M. In-depth Characterization of the Tomato Fruit Pericarp Proteome. *Proteomics* **2017**, *17* (1-2), 1600406.

(35) Reuß, D. R.; Altenbuchner, J.; Mäder, U.; Rath, H.; Ischebeck, T.; Sappa, P. K.; Thürmer, A.; Guérin, C.; Nicolas, P.; Steil, L.; et al. Large-Scale Reduction of the *Bacillus subtilis* Genome: Consequences for the Transcriptional Network, Resource Allocation, and Metabolism. *Genome Research* **2017**, *27* (2), 289–299.

(36) Cassidy, L.; Prasse, D.; Linke, D.; Schmitz, R. A.; Tholey, A. Combination of Bottom-up 2D-LC-MS and Semi-Top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon *Methanosarcina mazei*. *Journal of Proteome Research* **2016**, *15* (10), 3773–3783.

(37) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **2005**, *77* (4), 964–973.

(38) Frank, A. M. A Ranking-Based Scoring Function for Peptide-Spectrum Matches. *Journal of Proteome Research* **2009**, *8* (5), 2241–2252.

(39) Muth, T.; Weilnböck, L.; Rapp, E.; Huber, C. G.; Martens, L.; Vaudel, M.; Barsnes, H. DeNovoGUI: An Open Source Graphical User Interface for de Novo Sequencing of Tandem Mass Spectra. *Journal of Proteome Research* **2014**, *13* (2), 1143–1146.

(40) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Molecular and Cellular Proteomics* **2014**, *13* (1), 339–347.

(41) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (Oct), 2825–2830.

(42) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nature Communications* **2014**, *5*, 5277.

(43) May, D. H.; Tamura, K.; Noble, W. S. Param-Medic: A Tool for Improving MS/MS Database Search Yield by Optimizing Parameter Settings. *Journal of Proteome Research* **2017**, *16* (4), 1817–1824.

(44) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **1990**, *215* (3), 403–410.

(45) Perdivara, I.; Deterding, L. J.; Przybylski, M.; Tomer, K. B. Mass Spectrometric Identification of Oxidative Modifications of Tryptophan Residues in Proteins: Chemical Artifact or Post-Translational Modification? *Journal of the American Society of Mass Spectrometry* **2010**, *21* (7), 1114–1117.

(46) Namekata, K.; Oyama, F.; Imagawa, M.; Ihara, Y. Human Transferrin (Tf): A Single Mutation at Codon 570 Determines Tf C1 or Tf C2 Variant. *Human Genetics* **1997**, *100* (3–4), 457–458.

(47) Min, K.-T.; Benzer, S. *Wolbachia*, Normally a Symbiont of *Drosophila*, Can Be Virulent, Causing Degeneration and Early Death. *Proceedings of the National Academy of Science U. S. A.* **1997**, *94* (20), 10792–10796.

(48) Clark, M. E.; Anderson, C. L.; Cande, J.; Karr, T. L. Widespread Prevalence of *Wolbachia* in Laboratory Stocks and the Implications for *Drosophila* Research. *Genetics* **2005**, *170* (4), 1667–1675.

(49) Glaser, R. L.; Meola, M. A. The Native *Wolbachia* Endosymbionts of *Drosophila melanogaster* and *Culex quinquefasciatus* Increase Host Resistance to West Nile Virus Infection. *PLoS One* **2010**, *5* (8), e11977.

(50) Huffnagle, G. B.; Noverr, M. C. The Emerging World of the Fungal Microbiome. *Trends in Microbiology* **2013**, *21* (7), 334–341.

(51) Blaschke-Hellmessen, R. Standorte Für *Candida* Aus Medizinisch-Hygienischer Sicht: Habitats for *Candida* in Medical and Hygienic Respects. *Mycoses* **1999**, *42* (S1), 22–29.

(52) Weng, S.-L.; Chiu, C.-M.; Lin, F.-M.; Huang, W.-C.; Liang, C.; Yang, T.; Yang, T.-L.; Liu, C.-Y.; Wu, W.-Y.; Chang, Y.-A.; et al. Bacterial Communities in Semen from Men of Infertile Couples: Metagenomic Sequencing Reveals Relationships of Seminal Microbiota to Semen Quality. *PLoS One* **2014**, *9* (10).

(53) Kiessling, A. A.; Desmarais, B. M.; Yin, H.-Z.; Loverde, J.; Eyre, R. C. Detection and Identification of Bacterial DNA in Semen. *Fertility and Sterility* **2008**, *90* (5), 1744–1756.

(54) Akutsu, T.; Motani, H.; Watanabe, K.; Iwase, H.; Sakurada, K. Detection of Bacterial 16S Ribosomal RNA Genes for Forensic Identification of Vaginal Fluid. *Legal Medicine (Tokyo, Japan)* **2012**, *14* (3), 160–162.

(55) Manley, L. J.; Ma, D.; Levine, S. S. Monitoring Error Rates in Illumina Sequencing. *Journal of Biomolecular Techniques* **2016**, *27* (4), 125–128.

(56) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nature Methods* **2007**, *4* (3), 207–214.

(57) Tschager, T.; Rösch, S.; Gillet, L.; Widmayer, P. A Better Scoring Model for de Novo Peptide Sequencing: The Symmetric Difference between Explained and Measured Masses. *Algorithms in Molecular Biology* **2017**, *12*, 12.

III. CHAPTER 2. CONSIDERATIONS IN THE ANALYSIS OF DE NOVO PEPTIDE SEQUENCES

III.A. INTRODUCTION

The application of proteomics to environmental samples requires the development of methods for the taxonomic and functional characterization of peptide sequences identified from mass spectra. The assignment of sequences to spectra is itself a significant challenge in complex samples. In well-defined samples, such as a pure culture of *E. coli*, possible sequence assignments to spectra are selected from a database of proteins that could be present in the sample, such as the translated *E. coli* genome. Database search becomes less effective without an appropriate reference database.[1] As described in Chapter 3, I used 28 metagenomic and metatranscriptomic datasets from Alaskan soils that have been published in other studies as a reference database for peptide identification from my Alaskan metaproteomes.[2,3] The adequacy of this reference database may be facilitated by greater sequence homogeneity in colder than warmer soils – some metagenome-assembled genomes from the reference datasets have remarkably been identified at Alaskan sites separated by hundreds of kilometers.[3] Other environmental samples, such as those from warmer soils, may have a greater need for the alternative method of de novo sequencing in the absence of an adequate reference database. Post-processing of de novo sequences by *Postnovo*, introduced in Chapter 1, improves the accuracy of de novo sequencing to make it a viable alternative to database search.

The taxonomic and functional annotation of de novo sequences requires homology search against sequence databases using programs such as BLAST, presenting a number of challenges.[4] De novo sequence predictions contain degenerate leucine/isoleucine residues, unlike peptide-

spectrum matches returned by traditional search methods. Left unresolved, this can be a major source of uncertainty when blasting short proteomic peptide sequences. De novo sequences returned by Postnovo and other algorithms[5] can be partial-length sequences covering the higher confidence amino acids in the spectrum. The length of the query sequence can prevent the recovery of a statistically significant homologous hit. False positive errors can be controlled in BLAST hits by discarding hits failing to meet an E-value threshold. The E-value is the expectation value, or the number of equally strong hits to a database of the given size that is expected by chance alone.[6] The probability of a random hit can also be reduced by nested searches against large and small databases. For example, if a peptide sequence is blasted against the full RefSeq database and only hits bacterial sequences, then it is appropriate to blast the sequence against prokaryotic RefSeq in order to lower the E-values of the hits.[7]

Here, I describe necessary considerations in homology searches of short de novo peptide sequences and have implemented solutions to some of the problems raised in an extension of the Postnovo post-processing software.

III.B. HOMOLOGOUS SEQUENCE IDENTIFICATION

There are practical solutions to certain inherent difficulties raised by de novo sequences. To address L/I ambiguity in sequences, I BLAST every sequence permutation of L/I residues (e.g., the sequences PEPTIDE and PEPTLDE) and subsequently compare the hits (BLAST+ v.2.6.0, option blastp-short). The precomputed databases required by BLAST do not permit the use of a degenerate amino acid for both L and I, and the MS BLAST tool which addresses this issue is a web-based tool that only accepts a single query at a time.[8] I retain the maximum set of 500 BLAST hits reported per query sequence, and hits to Postnovo sequences with L/I

permutations are merged under a single query (e.g., the hits to PEPTIDE and PEPTLDE are merged back together into a single list of 1,000 hits). The option of employing multiple databases in the BLAST search is provided by my extension to Postnovo in order to allow recovery of hits with higher statistical significance from a smaller database nested within a larger database. Hits with an E-value ≤ 0.10 are retained as "strong" hits, or probable sequence homologs. Although this purely arbitrary threshold is quite high for traditional BLAST searches of long, protein-length sequences, it technically means that 0.1 false positive hits are found per query. Further screening based on comparison of the annotations of the 500 hits boosts confidence in the homologous sequences that are ultimately returned, as explained below.

Many queries do not have any strong hits due to the sequence's shortness and/or significant divergence from the closest database sequence. In my study of Arctic soil metaproteomes, the minimum sequence length is set at 9 residues, as it is difficult to extract meaningful information from shorter query sequences blasted against RefSeq, a large and inclusive database.[9] The relationship between query sequence length and E-value is explored in Figure III.1. One thousand subsequences were randomly drawn from RefSeq release 83 (the database as of July 17, 2017) at each of a range of lengths from 7 to 15 residues and blasted back against RefSeq.[i] Each line relates sequence length to the average E-value of the top hit for sequences of that length. The different colored lines show how E-value changes with database size. By definition, E-value is directly proportional to database size.[6] Sequences of length 11 lie on the cusp of statistical significance over a range of database sizes ($0.01 < \text{E-value} < 1$).

[i] Sequences of length 6 are often reported by database search and de novo sequencing, but these are often computationally intractable to BLAST against large databases due to the number of alignments that must be performed. Sequences longer than 15 residues are also found, but the bulk of sequences are shorter than this, especially in partial-length de novo peptide sequences.

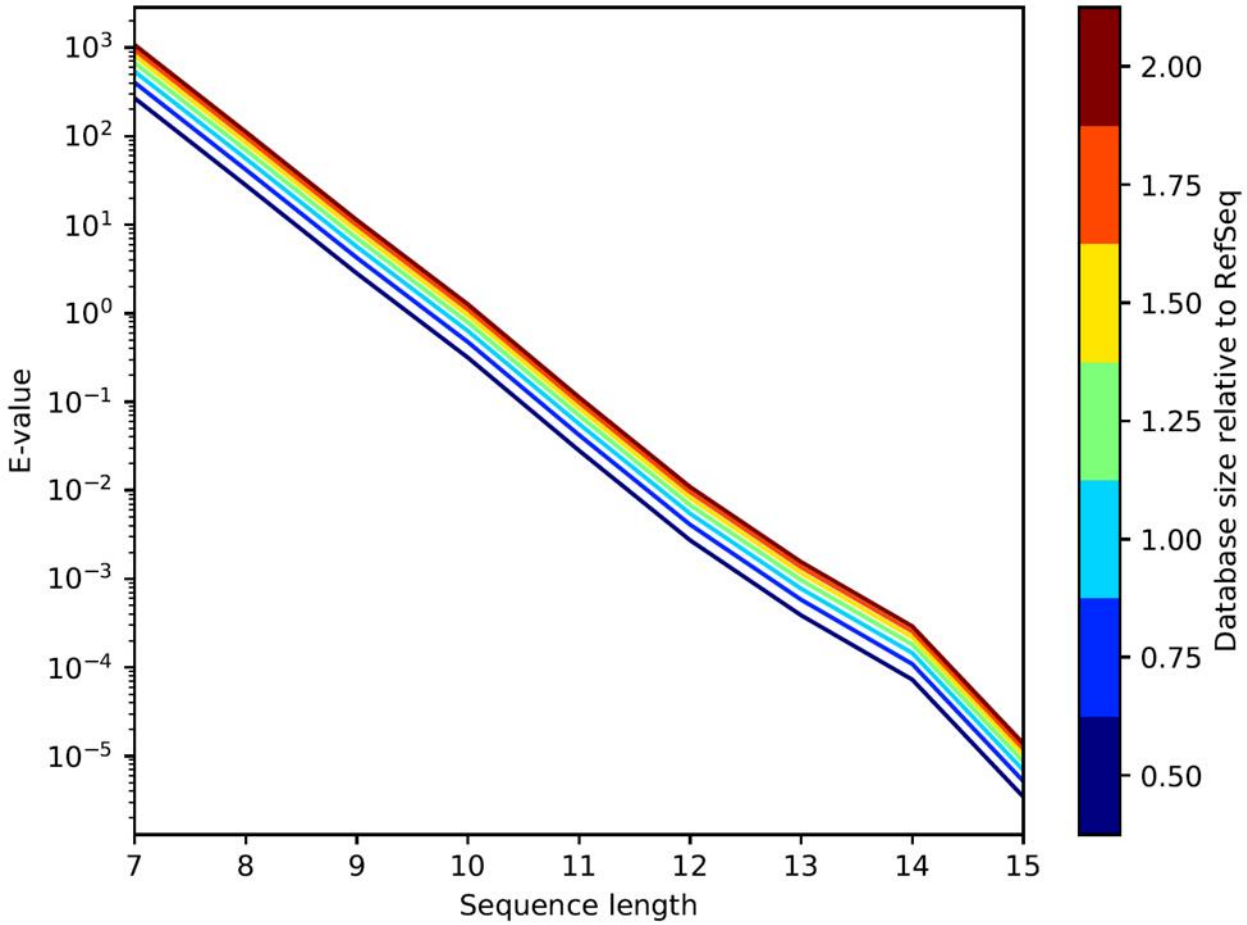

Figure III.1. Effect of sequence length on BLAST homolog recovery
One thousand sequences were randomly drawn from RefSeq release 83 (the database as of July 17, 2017) at each of lengths 7-15 and blasted back against RefSeq. Colored lines show how E-value changes with database size, relative to RefSeq. Many metaproteomic sequences lie in this range of lengths, so it is important to distinguish hits that are truly homologous from those that occur by chance alone.

Figure III.1 only addresses sequences that yield identical matches to homologous sequences, yet it is possible that many metaproteomic sequences from environmental samples diverge from the closest database sequence. To understand the representation of Arctic soil sequences in RefSeq, protein-coding sequences from an Arctic soil metagenome (NCBI SRA ERR1017187)[3] were randomly sampled over a range of sequence lengths and blasted against RefSeq (Figure III.2). One thousand subsequences were taken from MEGAHIT-assembled

contigs at each length, and the top-scoring BLAST hit(s) to each was retained. The number of amino acids in the query sequence not identical to the hit was calculated. Then the proportions of the 1,000 query sequences at each length with 1, 2, ..., *N* amino acid differences were calculated (the proportions in each length group sum to 1). Figure III.1 showed that many shorter query sequences have high E-values and thus non-homologous identical hits, so it is unsurprising that a large proportion of the shorter soil sequences yield identical hits. However, the small proportion of longer sequences with identical hits is not expected by chance alone, demonstrating that the vast majority of proteins from the Alaskan soil sample are not represented by identical sequences in RefSeq.

The question remains as to how many of the non-identical top hits to these longer soil sequences are non-homologous, arising by chance alone. To address this, the random subsequences that were drawn from RefSeq were mutated *in silico* using amino acid substitution probabilities from the PAM30 substitution matrix, the default matrix employed in blastp-short searches.[ii] The mutated sequences were blasted against RefSeq, returning 500 hits per sequence.

[ii] The placement of simulated mutations in a sequence, *s*, is determined from the PAM30 substitution matrix and M.B. Dayhoff's corresponding amino acid frequency data.[10] The relative frequencies of amino acid pairs are first calculated from the substitution matrix:

$$\boldsymbol{S^n} = \lambda \log_2 \frac{\boldsymbol{Q^n}}{\boldsymbol{P}}$$

$\boldsymbol{S^n}$ is the substitution matrix representing the log odds of amino acid pairs in a sequence alignment occurring by evolutionary processes versus a random process based on overall amino acid frequencies. $\boldsymbol{Q^n}$ is the matrix of expected pairing frequencies in homologous sequences. The superscript $\boldsymbol{n}$ represents the evolutionary divergence from the time 1 observations, with the PAM1 matrix normalized to 1% sequence change (99% average identity among aligned sequences in the PAM dataset). PAM30 is extrapolated from the $\boldsymbol{n}$ = 1 data over 30 units of time, resulting in 74% average identity (0.99^{30}), which is appropriate enough for the range of mutation frequencies in Figure III.3 (56-92% identity). $\boldsymbol{P}$ is the matrix of joint amino acid frequencies given their background frequencies in the sequence data – the null model for an alignment. $\boldsymbol{\lambda}$ is a scaling factor chosen for convenience. The extrapolation involved in PAM$^{\boldsymbol{n}}$ matrices is inappropriate for low-similarity, highly divergent homologs, but the simulated homologs of Figure III.3 are similar enough to permit the use of the PAM dataset.[11]

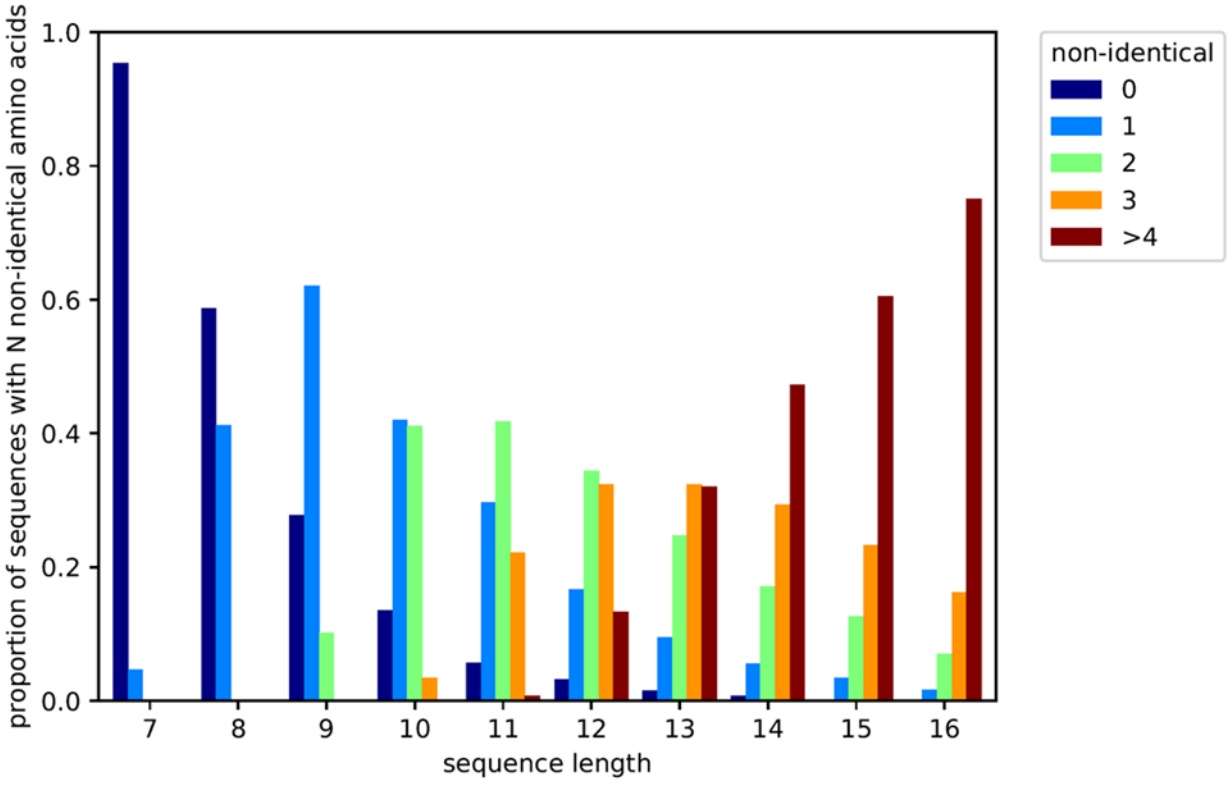

Figure III.2. Divergence of Arctic soil sequences from RefSeq sequences
One thousand sequences were randomly drawn from translated Arctic soil metagenomic contigs at each of lengths 7-16 and blasted against RefSeq. The number of non-identical amino acids in a query sequence was calculated from the top hit and is represented by each color. The bars in each sequence length group sum to 1. Sequences shorter than 11 often hit non-homologous sequences by chance alone, resulting in an inflated number of identical amino acids. The longer sequences show the divergence of Arctic soil proteins from homologous proteins in RefSeq, with very few of these peptide-length sequences present in RefSeq.

(continued from previous page)

To introduce *m* mutations in sequence *s*, *m* amino acids are randomly chosen with weight w_i equal to the probability *U* of amino acid *i* mutating to any non-identical amino acid *j* relative to the overall probability of mutation for the amino acids in $\boldsymbol{s}$.

$$U_i = \frac{\left(\sum_j \boldsymbol{Q}[i,j]\right) - \boldsymbol{Q}[i,i]}{\sum_j \boldsymbol{Q}[i,j]}$$

$$w[i] = \frac{U_i}{\sum_s \boldsymbol{U}}$$

Once the amino acids to mutate in *s* are chosen, the amino acids to which they mutate are chosen. The random selection of mutations is again weighted by the pairing frequencies in $\boldsymbol{Q}$.

The position of the unmutated "parent" sequence among the hits was found by searching for the unique parent sequence accession. The rank of the parent sequence in the BLAST results was used to study the recovery of homologous sequences among non-homologous sequences hit by chance. As seen in Figure III.3, unmutated sequences drawn from the database and blasted back against it do not always recover the query accession in position 1 – if they did, the boxes would be one dimensional lines flush with the x-axis. The reason for this is that other organisms can have identical sequences in conserved proteins with different accessions. The distribution of hit positions for unmutated sequences serves as a baseline for the analysis of mutated sequences. For example, sequences of length 13 with one mutation yield approximately the same distribution of hit positions as the parent sequences, and the parent sequences are known to have a negligible rate of high E-value null hits, as seen in Figure III.1. Therefore, length 13 peptide sequences can easily be matched to closest homologous sequences in RefSeq that diverge by one amino acid. As the number of mutations in a sequence increases, the parent sequence falls in BLAST table position and may cease to be found among the top 500 hits that are reported. The proportion of parent sequences not found in the results is recorded above each box in Figure III.3. For example, 2.5% of length 13 sequences randomly drawn from RefSeq (zero mutations) fail to recover their own accessions anywhere in the results. Again, this is due to identical sequences occurring in conserved proteins shared by many organisms and so is not an issue in analyzing the validity of homologous search results.

Despite the frequent separation by multiple non-identical amino acids of Arctic soil sequences from their closest homolog in RefSeq (Figure III.2), sequences with this amount of divergence can typically be recovered near the top of the BLAST table. Consider length 11 sequences, 96% of which diverge by at least one non-identical amino acid from their top hit in

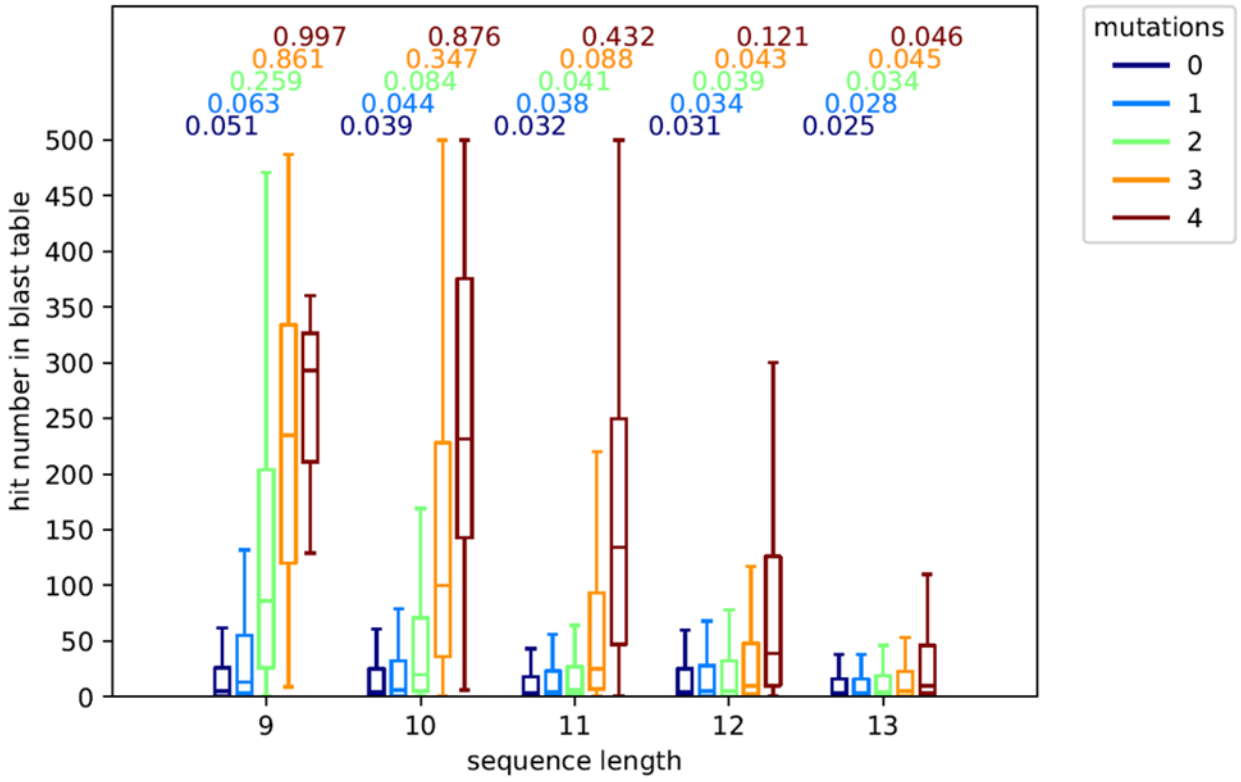

Figure III.3. Effect of sequence divergence on BLAST homolog recovery
One thousand sequences were randomly drawn from RefSeq at each of lengths 9-13. Simulated mutations were introduced into each sequence, resulting in five sets of 1,000 sequences with different numbers of mutations at each length. These sequences were blasted against RefSeq, returning the maximum of 500 hits per sequence. Each list was searched for the query accession, indicating the identical sequence was not due to chance alone. The distributions of these hit ranks are plotted, and the proportion of query sequences that did not return the selfsame hit in the BLAST results (position > 500) is recorded above each box. For example, 51 out of 1,000 length 9 sequences with zero mutations were not found in the BLAST results and are not represented by the box. For longer sequences that do not have any confounding null hits to the database, the failure to find the query accession is due to the large number of identical homologous sequences among many organisms. The plot explores the recovery of divergent homologs in RefSeq. Sequences with the number of amino acid differences expected in Arctic soils (see Figure III.2) recover homologous sequences at or near the top of the BLAST results, and null hits to unrelated proteins are not a significant problem.

RefSeq (Figure III.2) and have a low rate of null hit complications associated with shorter sequences (Figure III.1). 99% of length 11 soil sequences with three or fewer non-identical amino acids can be closely matched to homologous sequences in RefSeq (Figure III.3). The

increasing number of non-identical amino acids in longer soil sequences does not pose a problem, as the higher score of longer alignments retains homologous sequences near the top of the results.

III.C. TAXONOMIC AND FUNCTIONAL SCREENING AND ANNOTATION

Although sequences of length 9 and 10 have relatively high E-values, de novo sequences of this length were still analyzed due to additional filters used to winnow homologous from null hits. First, taxonomic identity is used to screen BLAST results that lack any "strong" hits meeting the E-value threshold of 0.1. A profile of resolved taxonomic groups in the metaproteome is constructed from sequences with strong hits that can be resolved by lowest common taxonomic rank to at least the family level. Lowest common ranks that appear multiple times in the set of strong hits are added to the profile. As an example of screening with the taxonomic profile, consider a length 9 Arctic soil sequence which yields a cluster of hits to a protein expressed by members of *Rhizobiaceae*. The hits do not meet the E-value threshold of 0.1, and another clearly spurious hit to *Macropodidae* (kangaroo) is observed in the results. The taxonomic profile was constructed from sequences with strong hits to *Rhizobiaceae* but not *Macropodidae*, so this sequence's hits to *Rhizobiaceae* but not *Macropodidae* are retained. Taxonomic considerations also provide the basis for the utilization of results from BLAST searches against nested databases. Multiple BLAST searches can be useful in complex communities with significant representation of certain taxonomic groups. In some soils, bacterial proteins may be quite abundant but not overwhelmingly dominant, so it is useful to blast peptide sequence predictions against the full RefSeq database as well as the prokaryotic RefSeq database, as the latter is smaller and thus yields more statistically significant hits with lower E-

values.[7] BLAST results against the broader database are parsed first, and if a query sequence's hits all fall within the taxonomic ambit of the narrower database, such as a set of hits to *Bacteria*, then the BLAST hits from the narrower database are used instead.

Both the "strong" hits and taxonomically selected "weak" hits are passed through a functional screen in order to further amplify true homologs. The retained hits at this stage are passed to eggNOG-mapper functional annotation software for comparison to eggNOG superkingdom-level databases.[12] The eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database was developed to group orthologous genes, or related genes which did not derive from gene duplication (paralogs).[13,14] Consistent with the hypothesis that orthologs are more likely to retain the ancestral function than paralogs, sequence mapping to eggNOG results in more accurate functional annotation than broader homology-based mapping. As an example of functional screening, consider a peptide sequence with two "strong" BLAST hits that barely meet the E-value threshold of 0.1 but have different eggNOG functional annotations – the sequence is excluded from the Pipeline's reported results, as the hits do not have a consistent annotation. The functional consistency screen, like the taxonomic profile screen before, can improve confidence in weak BLAST hits that are otherwise difficult to interpret. For example, if a peptide sequence has two weak BLAST hits (E-values of 0.5 and 1) that have passed the taxonomic filter, then the sequence is retained if the two subject sequences have identical functional annotations.

III.D. DISCUSSION

Postnovo has been extended to handle the annotation of de novo sequences. This Python 3 application is available at https://github.com/semiller10/postnovo. Annotation results are

reported in tabular format, with each peptide, represented by one or more spectra, assigned functional terms by eggNOG-mapper and a taxonomic lineage annotation to the lowest common rank of screened BLAST hits.

We tested the annotation of de novo sequences using 13 soil metaproteomes from the area of Toolik Field Station, Alaska (see Chapter 3 for analysis of these samples and more, and for methods of protein extraction, mass spectrometry, and metagenome utilization). Hundreds to thousands of unique peptides were identified from each metaproteome. Database search was conducted with 13 soil metagenomes from three permafrost areas in Alaska.[3,15,16] No metagenomes were paired to the Toolik metaproteomic soil samples, and the 10 Illumina HiSeq metagenomes from a site 250 miles distant provided a majority of the best peptide sequence predictions. De novo sequencing in conjunction with post-processing by Postnovo succeeded at identifying a correct sequence for 65% of the unique peptides that were also identified by database search.[5,17] Many of the de novo sequences did not cover the full length of the peptide and were ultimately superseded by database search PSMs as the reported sequence, yet an average of 18% of the unique reported sequences were de novo.

Our metaproteomic analyses leveraged recent developments in de novo sequencing, metaproteomic database search,[18] and functional annotation[12] to perform high-throughput metaproteomic sequence identification and peptide annotation. The analysis of short and often partial peptide sequences is enabled by rigorous screening criteria, including sequence length relative to database properties and the consistency of taxonomic and functional annotations among putative homologs. The use of BLAST as a homology search tool was validated with soil peptide sequences, which are divergent from homologous RefSeq sequences, but not so divergent as to prevent the identification of these homologs among BLAST hits. The growth of

RefSeq and other large databases may adversely affect the statistical significance of environmental metaproteomic BLAST results given the relatively short length of many sequences identified by de novo sequencing and database search. The creation of tailored reference databases akin to prokaryotic RefSeq[7] but for specific environments like soil could improve the E-values of homologous hits. This may be counterproductive, however, if the exclusion of certain classes of organisms, such as the numerous closely-related strains of pathogenic bacteria in RefSeq, results in the removal of homologs to organisms from the environment of interest. Alternatively, the production of single-cell or metagenome-assembled genomes from diverse environments may lead to higher-scoring alignments that compensate for increasing database size.

Parts of the MetaProteomeAnalyzer software would be worth emulating in further upgrades of Postnovo.[19] Useful features include interfaces with programs for the graphical interpretation of results and the creation of a database of annotation terms from the results that can be used for filtering by user-defined keywords. The soil metaproteomic results show that even in the absence of metagenomic data from the same sample, a pool of metagenomic sequences from similar samples is sufficient to yield many protein identifications from relatively complex samples. The tandem application of de novo sequencing and database search has many potential applications in complex and uncharacterized samples, including the human gut microbiome and protein variants caused by cancer mutations.[20]

III.E. REFERENCES

(1) Muth, T.; Kolmeder, C. A.; Salojärvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; et al. Navigating through Metaproteomics Data: A Logbook of Database Searching. *Proteomics* **2015**, *15* (20), 3439–3453.

(2) Ward, C. P.; Nalven, S. G.; Crump, B. C.; Kling, G. W.; Cory, R. M. Photochemical Alteration of Organic Carbon Draining Permafrost Soils Shifts Microbial Metabolic Pathways and Stimulates Respiration. *Nature Communications* **2017**, *8* (1), 772.

(3) Johnston, E. R.; Rodriguez-R, L. M.; Luo, C.; Yuan, M. M.; Wu, L.; He, Z.; Schuur, E. A. G.; Luo, Y.; Tiedje, J. M.; Zhou, J.; et al. Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Frontiers in Microbiology* **2016**, *7*.

(4) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. **1990**, *215* (3), 403–410.

(5) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **2005**, *77* (4), 964–973.

(6) Karlin, S.; Altschul, S. F. Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proceedings of the National Academy of Science U. S. A.* **1990**, *87* (6), 2264–2268.

(7) Tatusova, T.; Ciufo, S.; Federhen, S.; Fedorov, B.; McVeigh, R.; O'Neill, K.; Tolstoy, I.; Zaslavsky, L. Update on RefSeq Microbial Genomes Resources. *Nucleic Acids Research* **2015**, *43* (Database issue), D599–D605.

(8) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. Charting the Proteomes of Organisms with Unsequenced Genomes by MALDI-Quadrupole Time-of-Flight Mass Spectrometry and BLAST Homology Searching. *Analytical Chemistry* **2001**, *73* (9), 1917–1926.

(9) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Research* **2016**, *44* (D1), D733-745.

(10) Dayhoff, M. O.; Schwartz, R. M. Chapter 22: A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*; 1978.

(11) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Science U. S. A.* **1992**, *89* (22), 10915–10919.

(12) Huerta-Cepas, J.; Forslund, K.; Coelho, L. P.; Szklarczyk, D.; Jensen, L. J.; von Mering, C.; Bork, P. Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper. *Molecular Biology and Evolution* **2017**, *34* (8), 2115–2122.

(13) Powell, S.; Forslund, K.; Szklarczyk, D.; Trachana, K.; Roth, A.; Huerta-Cepas, J.; Gabaldón, T.; Rattei, T.; Creevey, C.; Kuhn, M.; et al. EggNOG v4.0: Nested Orthology Inference across 3686 Organisms. *Nucleic Acids Research* **2014**, *42* (D1), D231–D239.

(14) Altenhoff, A. M.; Boeckmann, B.; Capella-Gutierrez, S.; Dalquen, D. A.; DeLuca, T.; Forslund, K.; Huerta-Cepas, J.; Linard, B.; Pereira, C.; Pryszcz, L. P.; et al. Standardized Benchmarking in the Quest for Orthologs. *Nature Methods* **2016**, *13* (5), 425–430.

(15) Lipson, D. A.; Haggerty, J. M.; Srinivas, A.; Raab, T. K.; Sathe, S.; Dinsdale, E. A. Metagenomic Insights into Anaerobic Metabolism along an Arctic Peat Soil Profile. *PLoS One* **2013**, *8* (5).

(16) Fierer, N.; Leff, J. W.; Adams, B. J.; Nielsen, U. N.; Bates, S. T.; Lauber, C. L.; Owens, S.; Gilbert, J. A.; Wall, D. H.; Caporaso, J. G. Cross-Biome Metagenomic Analyses of Soil Microbial Communities and Their Functional Attributes. *Proceedings of the National Academy of Science U. S. A.* **2012**, *109* (52), 21390–21395.

(17) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society of Mass Spectrometry* **2015**, *26* (11), 1885–1894.

(18) Tang, H.; Li, S.; Ye, Y. A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. *PLoS Computational Biology* **2016**, *12* (12), e1005224.

(19) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *Journal of Proteome Research* **2015**, *14* (3), 1557–1565.

(20) Alfaro, J. A.; Ignatchenko, A.; Ignatchenko, V.; Sinha, A.; Boutros, P. C.; Kislinger, T. Detecting Protein Variants by Mass Spectrometry: A Comprehensive Study in Cancer Cell-Lines. *Genome Medicine* **2017**, *9*, 62.

IV. CHAPTER 3. THE METAPROTEOMIC ANALYSIS OF ARCTIC SOILS

IV.A. INTRODUCTION

Arctic soils contain large quantities of organic carbon susceptible to mineralization in a warming climate. The upper 3 m of these soils hold 1,035±150 Pg C_{org},[1,2] or 1.4 times the amount of C in the atmosphere and 50% of the global stock of soil C_{org},[3] despite Arctic soils spanning only 15% of soil surface area. The quality of C_{org} is higher than in lower latitude soils, as low temperatures and high soil moisture limit decomposer activity, while cryoturbation rapidly advects organic matter and fresh plant materials into deeper, colder layers.[4,5] Soil incubation experiments up to 12 years in length indicate that a significant fraction of C_{org} is converted into CO_2 under aerobic, unsaturated conditions at 4-5°C.[6,7] Climate models with basic representations of permafrost properties project that the Arctic region will lose 37-174 Pg C by 2100 given the current trajectory of greenhouse gas emissions (RCP 8.5); the average across models, 92±17 Pg, is the equivalent of 9 years of current anthropogenic emissions.[8,9] The permafrost carbon-climate feedback has been modeled as a mechanism for the Paleocene-Eocene Thermal Maximum and ensuing periodic hyperthermals due to the vast extent of unglaciated land that existed at both poles and the sensitivity of the poles to strong Milankovich orbital forcing that coincided with the hyperthermals.[10] A number of relatively unconstrained factors have the potential to magnify soil C losses with warming, including the formation of thermokarst, or permafrost collapse features,[11,12] the higher frequency of deep-burning fires,[13] the effects of augmented plant growth and the turnover of floral ecotypes,[14–16] and early winter soil respiration during top-down freezing to the permafrost table.[17]

Gains in plant biomass due to rising temperatures and CO_2 fertilization counteract soil C losses in net ecosystem exchange (NEE). Satellite observations since 1982 indicate that tundra plant biomass has increased by ~20%,[18] equivalent to ~0.4 Pg C, a small amount relative to the large C pools in the soil and lower-latitude boreal forest biomass (~53.9 Pg).[19] Plant growth is projected to increase substantially in the coming decades, counteracting soil C losses in NEE until 2100, but doing little to offset much larger losses after that point.[20] The greening of the Arctic is evident in all types of vascular plants, with community plant height projected to increase by 20-60% by 2100 given the current rate of change.[21] The ingrowth of taller floral ecotypes better adapted to warmer climates, especially woody deciduous shrubs, is responsible for part of this change, and comes at the expense of mosses, lichens and prostrate herbaceous shrubs.[22,23] In situ summer greenhouse warming manipulations conducted since 1989 at Toolik Field Station on the North Slope of Alaska have produced large increases in plant biomass and woody dominance at the expense of nonvascular plants, with vascular biomass increasing by 77% and non-vascular biomass decreasing by 73% in the first 14 years of the experiment.[24] There was no statistically significant change in organic C in the surface organic layer of the soil but measurable loss in the deeper mineral layer. A reduction in microbial biomass and food web complexity at the surface and greater inputs of leaf litter are hypothesized to have caused a gain in organic layer C after an initial period of loss,[25,26] while the deeper penetration of roots and soil insulation from winter cold due to snow trapping by taller plants may have stimulated C respiration by dormant microbes in the mineral layer. The results of this experiment illustrate the complex interplay between warming air temperatures, plant responses, microbial responses, and soil biogeochemistry.

A greater understanding of the microbial communities that catalyze soil organic matter transformations is needed in mechanistic models of soil biogeochemistry. The structural complexity and occlusion of soil organic matter (SOM) precludes the in situ measurement of most fluxes between plant and microbial products, increasing the predictive value of relationships between specific microbial taxa and biogeochemical processes.[27] Although many genes associated with SOM metabolism are distributed among diverse genomes, potentially indicating an absence of phylogenetic niche partitioning, functional potential does not necessarily translate into ecological importance.[28,29] Taxa possessing the same genetic pathway but exhibiting different growth rates, resuscitation rates from dormancy, or substrate use efficiencies can respond divergently to the substrate given a set of soil conditions.[30,31] Surveys of 16S rRNA diversity have revealed that the relative abundances of major phylogenetic clades strongly correlate with certain general soil properties, especially pH, O_2 availability, climatic factors, and plant productivity.[32–34] In contrast, the distribution of genomic potential for the utilization of substrates and electron acceptors generally appears to be much shallower,[35–38] with a notable exception being the deep phylogenetic conservation of methanogenesis in the Archaea, as this pathway is intertwined with other core cellular pathways. Traits encoded by small, modular operons are more easily transferred between strains by horizontal gene transfer,[39] although recombination is often favored among related bacterial strains,[40] perhaps constraining the spread of shallow traits. Recent work has revealed patterns of deeper phylogenetic conservation in the expression of seemingly shallowly distributed substrate usage traits.[28,30,41,42] DNA-stable isotope probing (DNA-SIP) of soil microcosms amended with isotopically labeled xylose and cellulose found that specific clades acquired the label from each substrate, and in the case of xylose, the label was transferred between clades over 7 days, suggesting interactions

between bacteria inhabiting separate trophic levels.[30] Rather than amending a natural microbial community with a specific substrate, as in DNA-SIP, exometabolomic studies have amended specific isolates with natural mixtures of exudate metabolites to measure substrate preferences of each isolate.[41] Not only are rhizosphere isolates selective in the assimilation of root exudate compounds, but they specialize in substrates produced at different plant developmental stages.[42]

Metaproteomic methods have the potential to characterize in situ protein expression and therefore the intra- and extracellular biogeochemical processes occurring under a given set of natural conditions.[43] I developed a novel computational pipeline called *ProteinExpress* to overcome certain limitations in existing metaproteomic data analysis methodologies and applied it to soil samples from the area of Toolik Field Station on the North Slope of Alaska. Samples were taken from three major floral ecotypes representative of the trajectory of Arctic greening – intertussock, tussock, and woody shrub – with the intertussock ecotype dominated by nonvascular plants and the tussock and woody shrub ecotypes dominated by vascular plants.[44] ProteinExpress processes metagenomic datasets from the environment of interest to create a peptide reference database for sequence identification from mass spectra. Due to the high genomic microdiversity of soil microbes, the use of full sets of metagenomic reads and contigs from multiple datasets increases the number of peptide-spectrum matches (PSMs). This in turn raises the issue that a peptide can be found in multiple proteins from different genomes; different protein sequences linked to a spectrum must be screened to ensure they have identical functional annotations. Determination of the taxonomic origin of proteins is complicated by the nature of the database search, as the database contains a number of sequences that cannot be confidently assigned a phylogenetic identity due to insufficient sequence length or a lack of related genomes in public sequence databases. ProteinExpress addresses this problem by aligning PSM-bearing

protein-coding sequences to taxonomic bins of metagenomic contig sequences identified in the metagenomic datasets. The strength of the alignment multiplied by the overall abundance of the protein function provides a useful metric for the functional activity of a taxon, which I term bin fidelity. Bin fidelity patterns and changes in protein abundance across the floral environments elucidate ecophysiological traits of poorly understood but ubiquitous groups of soil bacteria and how the microbiome shifts with Arctic greening.

IV.B. METHODS

IV.B.1. SAMPLES

Cores of permafrost-affected soil were collected from two areas near Toolik Field Station (TFS) on the North Slope of Alaska, USA (Table IV.1). To extract cores, a 2.5 inch diameter serrated push-corer was pushed to the permafrost table or water table, and the core was extracted with the assistance of suction force from the corer's handled plunger. Cores were wrapped in aluminum foil sterilized with isopropyl alcohol wipes, carried back to TFS on blue ice, and placed in a -80°C freezer within 3 hours of sampling. Cores were transported from TFS to the University of Chicago on dry ice in an insulated box, and again stored at -80°C.

Sampling was conducted in early August at the end of the lower Arctic growing season. Soil cores were taken from three of the common vegetation types of Arctic North American moist acidic tundra (MAT): tussock, intertussock, and shrub.[45,46] The sedge, *Eriophorum vaginatum,* forms hemispherical tussocks ~0.5 m in radius, with a dense structure of slender roots extending through the organic soil to the base of the active layer.[15,47] Tussock sedges are spaced ~0.5 m apart and surrounded by diverse intertussock vegetation, including mosses, lichens, and prostrate herbaceous plants such as *Rubus chamaemorus* (cloudberry).[48] Intertussock

Sample ID	Vegetation Type	Soil Horizon	Northing (UTM, Zone 6 N)	Easting (UTM, Zone 6 N)	Site	Active Layer Depth (cm)	Depth to Water Table (cm)	Sampling Date (in 2014)
1	Intertussock	Organic	7612385.993	406199.426	Imnavait	40	23	Aug 11
2	Intertussock	Organic	7612385.993	406199.426	Imnavait	40	23	Aug 11
3	Intertussock	Organic	7612296.508	405983.235	Imnavait	64	23	Aug 11
4	Intertussock	Organic	7612296.508	405983.235	Imnavait	64	23	Aug 11
5	Intertussock	Organic	7614070.44	393755.304	Toolik	61	NA	Aug 12
6	Intertussock	Mineral	7614070.44	393755.304	Toolik	61	NA	Aug 12
7	Tussock	Organic	7612386.411	406199.526	Imnavait	53	NA	Aug 11
8	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
9	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
10	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
11	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
12	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
13	Tussock	Organic	7612345.005	406061.461	Imnavait	45	NA	Aug 11
14	Tussock	Organic	7614069.165	393754.96	Toolik	60	NA	Aug 12
15	Tussock	Mineral	7614069.165	393754.96	Toolik	60	NA	Aug 12
16	Shrub	Organic	7614490.628	393616.694	Toolik	45	NA	Aug 8
17	Shrub	Mineral	7614490.628	393616.694	Toolik	45	NA	Aug 8
18	Shrub	Organic	7612425.334	406153.236	Imnavait	58	35	Aug 11

Table IV.1. Metaproteomic field sample information

plants have a low profile compared to tussocks and are shallow rooting, with the moss carpet transitioning into the organic soil. Woody shrubs ~1 m in height are prevalent around water tracks, or near-surface runoff channels. *Betula nana* (dwarf birch) and *Salix pulchra* (diamondleaf willow) are main members of the shrub ecotype, with mosses, lichens and herbaceous plants carpeting the soil. Woody shrubs have a higher density of fine roots distributed at a shallower depth in the organic layer than tussock sedges.[15,49]

Samples from the three vegetation types were collected from two areas with MAT soils: the north-south facing hillslopes just south of Toolik Lake and the west facing hillslope of the Imnavait Creek valley. The sites lie 10 km apart but are situated on glacial outwash of different ages.[50,51] The Toolik site is on Itkillik I deposits of ~55 ka, and the Imnavait site is on Sagavanirktok deposits of ~125 ka. Although surface development affects soil pH through the erosion of buffering minerals, accumulation of organic matter, and growth of *Sphagnum* moss,

the Toolik and Imnavait organic soils have average pH values of 4.34 and 4.80, respectively, in contrast to nearby glacial deposits of Itkillik II age (~11.5 ka), which have circumneutral pH.[52,53]

IV.B.2. PROTEIN EXTRACTION

Discs ~3 cm in thickness were cut from frozen cores in a 4°C cold room. Discs were shattered with a hammer, with the discs enclosed in a sterile bag surrounded by another bag containing chips of dry ice. ~5 g of soil fragments free of visible roots were placed in a 50 mL Falcon tube. 40 mL of boiling detergent solution (4% (w/w) SDS, 100 mM Tris-HCl, 100 mM NaCl solution) was poured into the tube which was placed in boiling water for 5 min.[54] The mixture was shaken for 15 min at 99°C and 750 rpm. It was then horn sonicated at high power for 4 min (10 s on/off cycles) and shaken for 4 min at 99°C and 750 rpm; this step was repeated. To separate larger minerals and organic solids from the protein-bearing solution, the tube was centrifuged for 10 min at 2,000×g and the supernatant transferred to a 50 mL Teflon centrifuge tube prerinsed with acetone/methanol (1:1 v/v).

Dithiothreitol was added to a concentration of 24 mM to reduce disulfide bonds. Proteins were precipitated overnight at -20°C in 4 solvent volume: 1 sample volume of chilled acetone/methanol (1:1 v/v). The mixture was centrifuged at 7,000×g for 45 min and the supernatant discarded, leaving a light brown pellet containing proteins. The pellet was disaggregated by vortexing in 10 mL of chilled acetone/methanol (1:1 v/v). After washing, the pellet was centrifuged at 7,000×g for 15 min and the supernatant discarded. The pellet was dried in the fume hood for 6 hours until only a thin veneer of solvent remained on the pellet, and it was then dissolved in 1 mL of urea denaturing buffer (8 M urea, 100 mM Tris-HCl, pH 8 in LC/MS H_2O) by repeated horn sonication and vortexing over a period of 2 hours, achieving a dark brown

color in the solution. Free cysteines were alkylated by 20 mM 2-iodoacetamide in the dark for 15 min. The solution was diluted to 2 mL with dilution buffer (100 mM Tris-HCl, 20 mM $CaCl_2$).

Proteins were digested and peptides recovered following a procedure similar to eFASP (enhanced Filter-Aided Sample Preparation).[55] 125 µL of the solution containing the dissolved pellet was mixed with 200 µL of exchange buffer (8 M urea, 0.2% (w/v) deoxycholate, 100 mM ammonium bicarbonate in LC/MS H_2O) and dispensed into a passivated 30 kDa nominal cutoff concentrator (Sartorius Vivacon 500). The solution was spun down at 14,000×g, collecting proteins from the dissolved pellet on the concentrator; filtrate was discarded. The concentrator was washed 3 times with 200 µL of exchange buffer and then washed 2 times by 200 µL of digestion buffer (0.2% deoxycholate, 100 mM ammonium bicarbonate in LC/MS H_2O). After buffer exchange, the concentrator was transferred to a passivated collection tube. Proteins were digested with 2.5 µg Trypsin (1 µg Thermo Pierce MS-grade Trypsin / 50 µL sample) in 100 µL digestion buffer added to the concentrator and let stand overnight at 37°C. To collect peptides, 50 µL of peptide recovery buffer (100 mM ammonium bicarbonate in LC/MS H_2O) was twice spun through the concentrator at 14,000×g for 10 min. Finally, residual organic contaminants were removed by phase transfer to ethyl acetate. 200 µL of ethyl acetate was added to the filtrate and transferred to a LoBind Eppendorf tube. 2.5 µL of trifluroacetic acid was added, followed by vortexing, 10 s in an ultrasonic bath, and centrifugation at 16,000×g for 10 min. The upper organic layer was discarded, and the phase transfer was repeated. Residual ethyl acetate was evaporated off the aqueous layer for 5 min at 60°C. The peptide solution was frozen at -80°C and dried by centrifugal evaporation.

Peptides were reconstituted in 30 µL of 2% acetonitrile and 0.1% formic acid (v/v) in LC/MS H_2O. A 6 µL aliquot was injected onto a trapping column (OptiPak C18, Optimize

Technologies). Analytes were then separated on a capillary C18 column (Thermo Acclaim PepMap 100 Å, 2 µm particles, 50 µm I.D. × 50 cm length) by a Dionex Ultimate 3000 nanoLC system using a water-acetonitrile + 0.1% formic acid gradient (2-50% acetonitrile over 180 min) at 90 nL/min. Peptides were ionized by a nanoelectrospray source (Proxeon Nanospray Flex) fitted with a metal-coated fused silica emitter (New Objective). Mass spectra were collected on an Orbitrap Elite mass spectrometer (Thermo) operating in a data-dependent acquisition mode, with one high-resolution (120,000 $m/\Delta m$) MS1 parent ion full scan triggering 15 MS^2 Rapid mode CID fragment ion scans of selected precursors.

IV.B.3. DATA ANALYSIS

I developed software in Python 3 called ProteinExpress which uses a combination of existing bioinformatic tools and new methods. The following data analysis steps were implemented by ProteinExpress, as summarized in Figures IV.1 and IV.2. ProteinExpress interfaces with modules of Postnovo to allow spectrum identification by both database search and de novo sequencing. ProteinExpress source code is freely available at https://github.com/semiller10/protein-express.

IV.B.3.i. NUCLEOTIDE DATA

Peptide sequences were assigned to mass spectra by searching a sequence database generated from Alaskan soil metagenomes and metatranscriptomes (Table IV.2). The relatively low genetic diversity of Arctic soils in comparison to warmer soils facilitates mass spectral database search against nucleotide datasets from unpaired soil samples.[56] Databases were constructed from 12 metagenomes and 6 metatranscriptomes from Imnavait Creek,[57] one of the

Nucleotide dataset (ENA run accession)	Study (Reference)	Type of sequence data
ERR1017187	58	DNA
ERR1019366	58	DNA
ERR1022687	58	DNA
ERR1022692	58	DNA
ERR1034454	58	DNA
ERR1035437	58	DNA
ERR1035438	58	DNA
ERR1035441	58	DNA
ERR1039457	58	DNA
ERR1039458	58	DNA
SRR5208451	59	RNA
SRR5208454	59	RNA
SRR5208455	59	RNA
SRR5208541	59	RNA
SRR5208544	59	RNA
SRR5208545	59	RNA
SRR5450431	59	DNA
SRR5450432	59	DNA
SRR5450434	59	DNA
SRR5450438	59	DNA
SRR5450631	59	DNA
SRR5450755	59	DNA
SRR5471030	59	DNA
SRR5471031	59	DNA
SRR5471032	59	DNA
SRR5471221	59	DNA
SRR5476649	59	DNA
SRR5476651	59	DNA

Table IV.2. Metagenomic and metatranscriptomic sample information

two areas sampled for metaproteomes, as well as 10 metagenomes from the CiPEHR site near Fairbanks, 200 km south of TFS.[56] All 28 of these paired-end 2×150 bp datasets were generated on the Illumina HiSeq 2500 sequencing platform. Reads were trimmed with the SolexaQA++ dynamictrim subcommand at a 1% nucleotide error probability cutoff.[58] Full and partial genes were called and translated with Prodigal.[59] Gene sequences from each dataset were substring dereplicated with CD-Hit,[60] forming 28 databases of unassembled sequences. Reads were also assembled into contigs using MEGAHIT,[61] producing 28 databases of longer sequences.

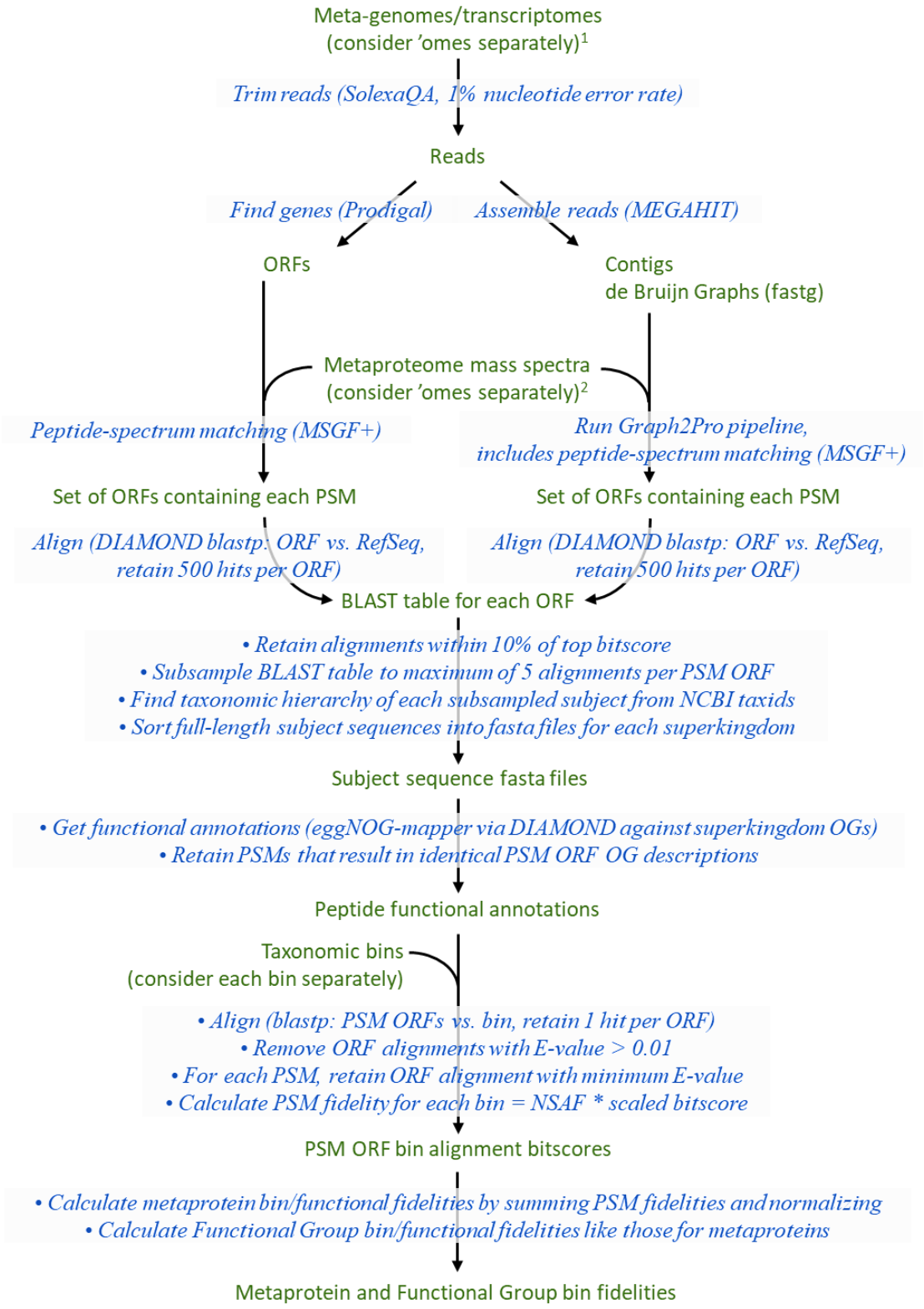

Figure IV.1. Flowchart of metaproteomic data analysis in ProteinExpress

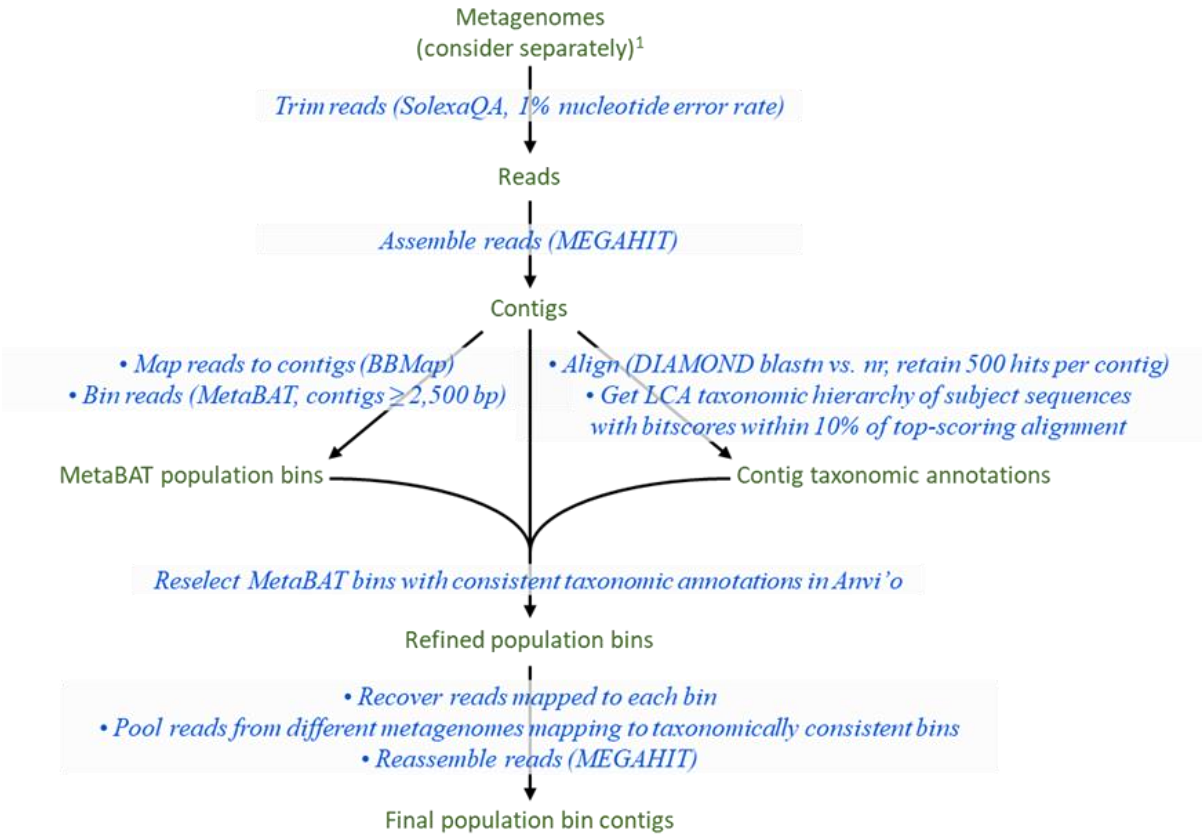

Figure IV.2. Flowchart of bin construction from metagenomes

To study the taxonomic origin of proteins, taxonomic bins of metagenomic contig sequences were generated in a two-stage process (Figure IV.2). Bins were identified from the read abundance, tetranucleotide frequency, and taxonomic affiliation of contigs. For each dataset, reads were first mapped to contigs and sorted using BBMap.[62] Contigs longer than 2,500 bp were binned with MetaBAT 2 on the basis of read abundance and tetranucleotide frequency.[63] These contigs were aligned to the nr nucleotide database using DIAMOND,[64] retaining 500 hits per query. A parsimonious taxonomic assignment at the lowest common rank was determined from subject sequences within 10% of the bitscore (a measure of alignment strength) of the query's top hit. Anvi'o was used to visualize bin data and select contigs with consistent

taxonomic ranks.[65] In the second stage of bin selection, reads mapping to taxonomically consistent contigs were pooled across metagenomic datasets and reassembled, producing one or more new bins from each taxonomic group. CheckM was used to estimate bin completeness and contamination using lineage-specific single-copy marker genes.[66] Bins with completeness ≥ 90% were retained and merged into a single bin for the taxonomic group. Highly complete bins often had ≥ 10% contamination (multiple occurrences of the same marker genes) due to the high genomic heterogeneity of soil bacterial communities.[67]

IV.B.3.ii. PEPTIDE DATA

Spectra from each metaproteomic dataset were searched using MSGF+ against the 28 unassembled and 28 assembled nucleotide datasets (Table IV.3).[68] Since peptides are often shorter than reads, each spectrum could return a peptide-spectrum match (PSM) from multiple database sequences. High-confidence PSMs were selected by finding the 1% false detection rate (FDR) cutoff from each target-decoy search. The full or partial gene sequences in which PSMs are located are called PSM ORFs (open reading frames); again, each PSM is associated with one or more high-confidence PSM ORF.

Database search against assembled nucleotide datasets was improved with the Graph2Pro pipeline.[69] Graph2Pro uses metaproteomic data in conjunction with the assembler's de Bruijn assembly graph to increase the number of PSMs. The graph includes both reported contigs and ambiguous contigs that represent multiple paths connecting reads. The identification of a high-confidence PSM in an ambiguous contig validates the existence of that sequence. Inclusion of these ambiguous contigs expands the size of the database search space, allowing matches beyond the ends of contigs normally reported by the assembler.

Sample ID	MS/MS spectra	Peptide-spectrum matches (≤1% FDR)	Spectra with screened functional annotations	Unique screened peptide sequences	Spectra/ PSM	Spectra/ unique peptide
1	38245	15878	4920	2099	7.8	18.2
2	39121	16235	4814	2223	8.1	17.6
3	36688	10485	3216	2081	11.4	17.6
4	36214	10064	3066	2050	11.8	17.7
5	32087	6052	1967	1202	16.3	26.7
6	30950	3142	1079	372	28.7	83.2
7	38778	4898	1481	727	26.2	53.3
8	39321	15055	5003	3071	7.9	12.8
9	38814	13583	4341	2640	8.9	14.7
10	39415	13915	4283	2624	9.2	15.0
11	39269	13879	4207	2543	9.3	15.4
12	39583	12978	3853	2305	10.3	17.2
13	40495	14184	4417	2572	9.2	15.7
14	33192	8621	2828	1456	11.7	22.8
15	33588	5462	2013	893	16.7	37.6
16	33378	8445	2801	1670	11.9	20.0
17	33596	6674	1807	728	18.6	46.1
18	36997	10919	3424	2292	10.8	16.1

Table IV.3. Metaproteomic sample analysis information

Upon searching spectra against unassembled and assembled nucleotide databases, PSMs were associated with sets of PSM ORFs. These were then searched against the NCBI Reference Sequence Database (RefSeq, release 83) using BLASTp to identify homologous proteins.[70,71] 500 hits were returned per ORF query, and subject sequences within 10% of the bitscore of the top hit were retained. If more than 5 subject sequences were selected by bitscore, then 5 were evenly sampled in descending order of bitscore. Selected subject sequences were sorted by NCBI Taxonomy ID into fasta files for each superkingdom and searched by eggNOG-mapper (via DIAMOND) against superkingdom-level OGs (Orthologous Groups of proteins).[72] eggNOG-mapper reports functional annotation terms from multiple protein classification systems, including Gene Family names, GO terms, KEGG Orthology (KO) IDs, and a functional description string inferred from the best-matching OG. At this point in ProteinExpress, each PSM is associated with a set of PSM ORFs, and each PSM ORF is associated with a set of functional annotations. PSMs found in a wide range of PSM ORFs or in short PSM ORFs may

not contain enough sequence information to identify a unique function for the PSM. Therefore, PSMs were filtered to those returning an identical set of OG functional description strings. Four functional annotation systems reported by eggNOG-mapper were maintained for each PSM (GO terms, KO IDs, and unique pairs of Gene Family and OG description), not only to compare results using each system, but also because some eggNOG-mapper queries do not receive annotations from every system (only the OG description is returned for every query). PSMs sharing a functional annotation are called metaproteins.[73] I assigned 2,659 unique Gene Family + OG description annotations to 141 Functional Groups of significance for biogeochemistry and cellular biology (Appendix: Table VI.3). For example, "CITA" + "Citrate synthase" was assigned to the "TCA Cycle" Functional Group.

Some Functional Groups, such as "TCA Cycle," are equivalent to biochemical pathways. Others, such as "ATP Synthase," "Transposase," "Ribosome," and "Ribose Transport," are comprised of protein functions which are not necessarily related through a common biochemical pathway. The Gene Family and OG description annotations from eggNOG-mapper were cross-referenced when needed to the UniProt KnowledgeBase, the InterPro database, the KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology database, and the CAZY (carbohydrate-active enzymes) database of glycoside hydrolase families to determine likely protein functions. For example, sequences annotated as "Hydrolase, family 38" were assigned to the Functional Group, "Mannose Cleavage," because CAZY records that enzymes assigned to Family 38 on the basis of sequence similarity are only known to be active on alpha-linked mannose residues.

The relative expression levels of peptides, metaproteins, and Functional Groups were calculated with the NSAF (normalized spectral abundance factor) metric.[74] Metaprotein and

Functional Group NSAF values can be calculated by summing the NSAF values of peptides in the metaprotein and Functional Group.

$$NSAF_N = \frac{S_N/L_N}{\sum_{i=1}^{n}(S_i/L_i)} \quad (1)$$

N is the peptide index; S_N is the number of spectra matched to the peptide; L_N is the average length of the protein subject sequences returned from the BLASTp searches of the PSM ORFs (see above); *n* is the total number of peptides identified in the dataset. The spectrum count is normalized to protein length since longer proteins generate more tryptic peptides and are more likely to be observed. I found that the standard deviation of subject sequence lengths from sets of PSM ORFs was 2.0%, on average.

I developed a metric called "fidelity" to estimate metaprotein expression levels of different taxa in a sample. Fidelity takes into account both the relatedness of a metaprotein to a particular bin of contig sequences (representing the likeliness of expression by organisms of the bin) and the level of metaprotein expression irrespective of bin (NSAF).

$$Fidelity_{N,T} = NSAF_N \times \frac{A_{N,T} - A_{N,min}}{A_{N,max} - A_{N,min}} \quad (2)$$

N is the peptide index; *T* is the taxonomic bin index; $A_{N,T}$ is the bitscore of the top-scoring alignment between the peptide's PSM ORFs and the taxonomic bin of contig sequences; $A_{N,min}$ and $A_{N,max}$ are the minimum and maximum bitscores, respectively, from the alignments against each taxonomic bin. $Fidelity_{N,T}$ lies on the range [0, 1].

$$Bin\ Fidelity_{N,T} = \frac{Fidelity_{N,T}}{Fidelity_{N,max}} \quad (3)$$

$Fidelity_{N,max}$ is the maximum fidelity for peptide *N* across the set of bins. $Bin\ Fidelity_{N,T}$ lies on the range [0, 1], with at least one value across the set of taxa having the value of 1.

$$Peptide\ Fidelity_{N,T} = \frac{Fidelity_{N,T}}{Fidelity_{max,T}} \quad (4)$$

$Fidelity_{max,T}$ is the maximum fidelity for taxon T across the set of peptides. If $Fidelity_{max,T}$ has a value of 0, it is instead set to 1 for the purposes of calculating peptide fidelity. $Peptide\ Fidelity_{N,T}$ lies on the range [0, 1].

To calculate fidelity, first find high-confidence BLASTp alignments between the set of PSM ORFs associated with a peptide's functionally-screened PSMs and each taxon's bin of contig sequences. For each PSM, take the ORF with the strongest alignment to the bin, only considering alignments with E-values ≤ 0.01 (expectation of random alignment to bin contigs ≤ 1%). This procedure returns the strongest association between each PSM and each bin, with the strength of association measured by bitscore, which is used in the calculation of fidelity.

I designed the fidelity metric to remove the dependence of the bitscore sequence similarity metric on sequence length and to be effective at determining the taxonomic origin of both conserved and divergent protein sequences. More conserved sequences have a narrower range of alignment bitscores across the bins. The PSM bitscore is normalized by subtracting the minimum bitscore in the range and dividing by the difference between maximum and minimum, resulting in normalized bitscores ranging from 0 to 1 regardless of alignment length and sequence conservation across bins. PSMs that do not have a statistically significant alignment to the bin (E-value > 0.01), as mentioned above, are given a normalized bitscore of 0. The normalized bitscore is then multiplied by the NSAF value of the peptide in the dataset, producing values representing peptide abundance × taxonomic similarity. These values are divided by the maximum value for the peptide across bins, yielding bin fidelity as defined in Equation 3, or by the maximum value for the bin across peptides, yielding peptide fidelity as defined in Equation

4. These different normalizations are useful for different purposes. Normalization across taxa is used to explore patterns of functional partitioning between taxa, whereas normalization across functions is used to explore changes in expression within individual taxa.

For metaproteins and Functional Groups, the abundance × similarity values of the peptides in the metaprotein or Functional Group are summed before dividing by the maximum fidelity across bins or metaproteins/Functional Groups.

$$Fidelity_{P,T} = \sum_{i}^{n}(NSAF_i \times \frac{A_{i,T} - A_{i,min}}{A_{i,max} - A_{i,min}}) \quad (5)$$

P is the metaprotein (or Functional Group) index; *T* is the taxonomic bin index; *n* is the total number of peptides in the metaprotein (or Functional Group). $Fidelity_{P,T}$ lies on the range [0, 1].

$$Bin\ Fidelity_{P,T} = \frac{Fidelity_{P,T}}{Fidelity_{P,max}} \quad (6)$$

$Fidelity_{P,max}$ is the maximum fidelity for metaprotein (or Functional Group) *P* across the set of bins. $Bin\ Fidelity_{P,T}$ lies on the range [0, 1], with at least one value across the set of taxa having the value of 1.

$$Metaprotein\ Fidelity_{N,T} = \frac{Fidelity_{P,T}}{Fidelity_{max,T}} \quad (7)$$

$Fidelity_{max,T}$ is the maximum fidelity for taxon *T* across the set of peptides. If $Fidelity_{max,T}$ has a value of 0, it is instead set to 1 for the purposes of calculating metaprotein (or Functional Group) fidelity. $Metaprotein\ Fidelity_{N,T}$ lies of the range [0, 1].

IV.C. RESULTS

IV.C.1. COMPARISON OF ENVIRONMENTS USING PROTEIN EXPRESSION PROFILES

IV.C.1.i. COMPARISON OF OVERALL PROTEIN EXPRESSION LEVELS

The most abundant cellular functions of microbial communities are identified at high levels in the metaproteomic datasets (Figures IV.3-IV.6). The following metaprotein Functional Groups have average NSAF values exceeding 1% across samples (only organic soil datasets, accounting for all but 3 mineral soil datasets, are considered for now): ribosomal proteins, ranging from 2.4-11.1% of identified spectra; cold shock proteins, 0.8-10.1%; ATP synthase, 0.9-7.1%; DNA supercoiling (including gyrase and topoisomerase), 1.6-9.2%; chromatin packaging (including histones), 0.1-6.9%; Gro chaperones, 0.9-6.7%; outer membrane porins, 0.8-3.7%; pili/fimbriae, 0.3-3.3%; and peroxide resistance proteins (including superoxide dismutase and catalase), 0.2-2.8%. Functional Groups with average NSAF values from 0.1-1.0% account for other integral microbial community functions, including DNA synthesis, replication, and transposition; RNA polymerase; tRNA ligases and other proteins involved in translation; central C metabolism pathways; amino acid synthesis and ammonia metabolism (glutamine synthetase); phosphate assimilation; flagella and chemotaxis.

Four Functional Groups with a majority of non-zero NSAF values in tussock and intertussock samples were found to have statistically significant differences between environments (shrub is ignored due to the fewer number of metaproteomic datasets than tussock and intertussock): ribose transport (Welch's t-test; $p = 0.00021$), xylose+arabinose transport ($p = 0.019$), sugar alcohol transport ($p = 0.014$), and succinoglycan (EPS) synthesis ($p = 0.014$). These Functional Groups are more highly represented in tussock than intertussock samples, and are most strongly expressed by Rhizobiales, as is explored in Section IV.C.2.ii. Alkanesulfonate

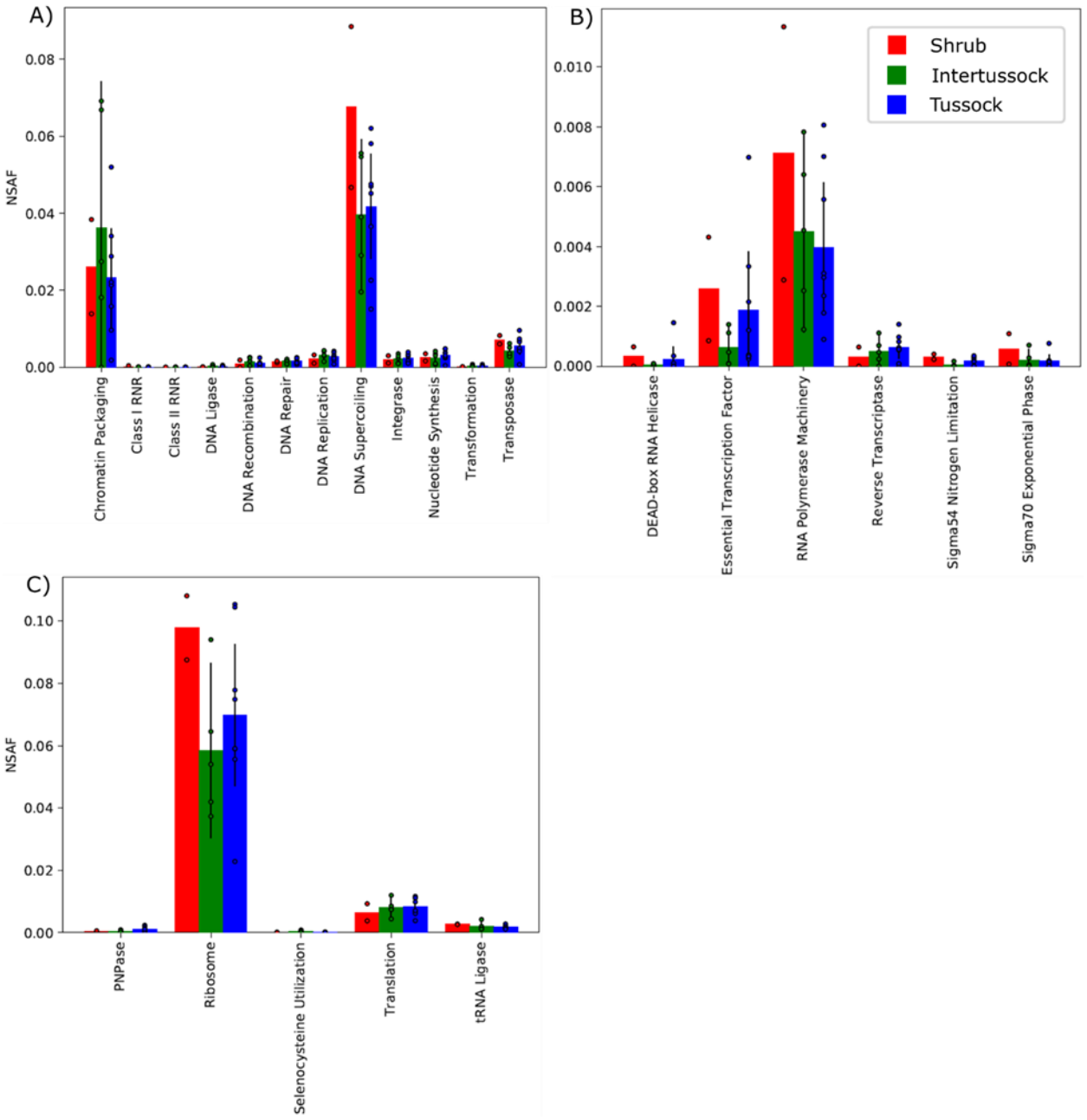

Figure IV.3. Expression levels (NSAF) of Functional Groups associated with A) DNA, B) RNA, and C) translation

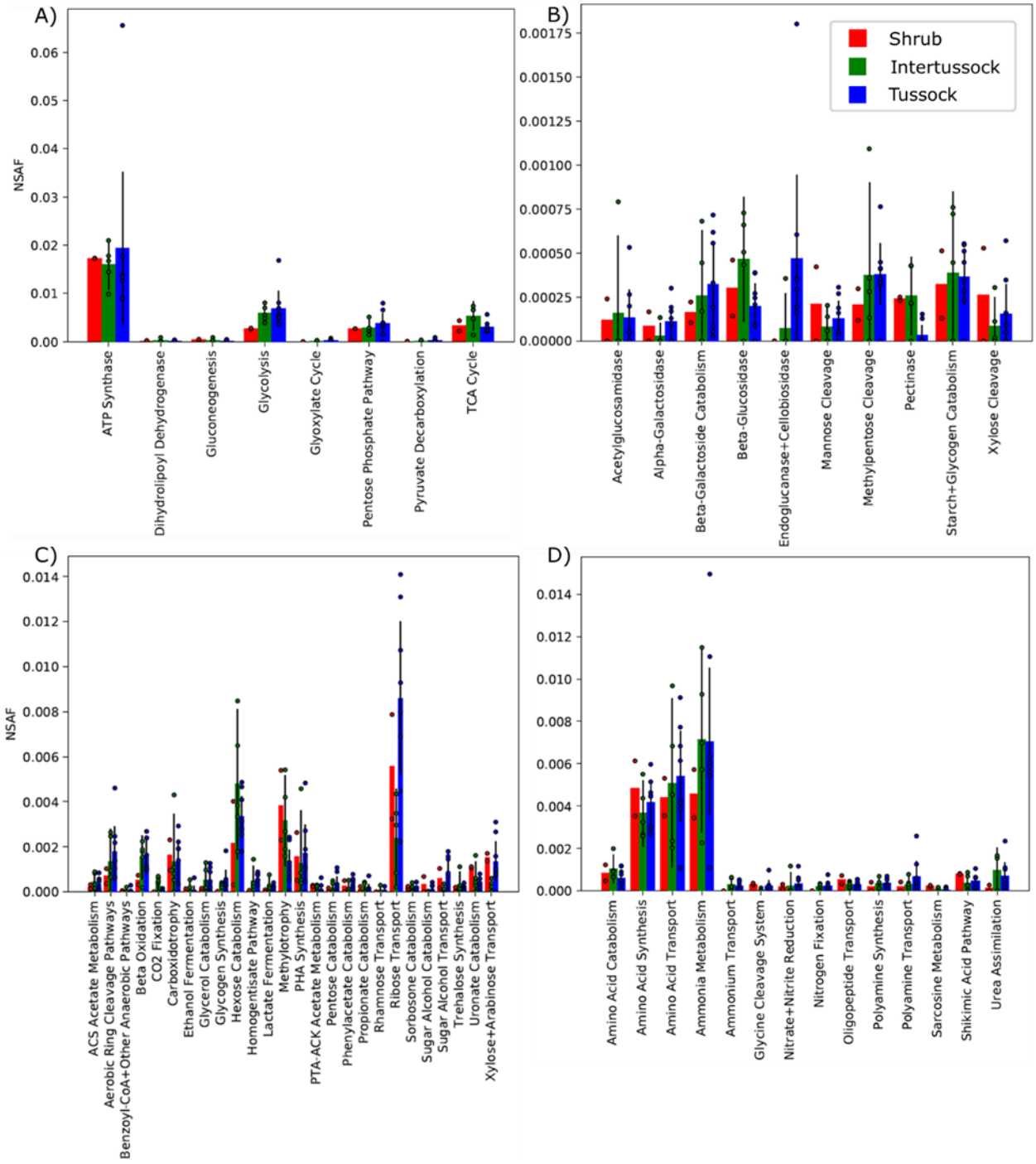

Figure IV.4. Expression levels (NSAF) of Functional Groups associated with A) central carbon metabolism and energy conservation, B) polysaccharide degradation, C) carbon metabolism, and D) nitrogen

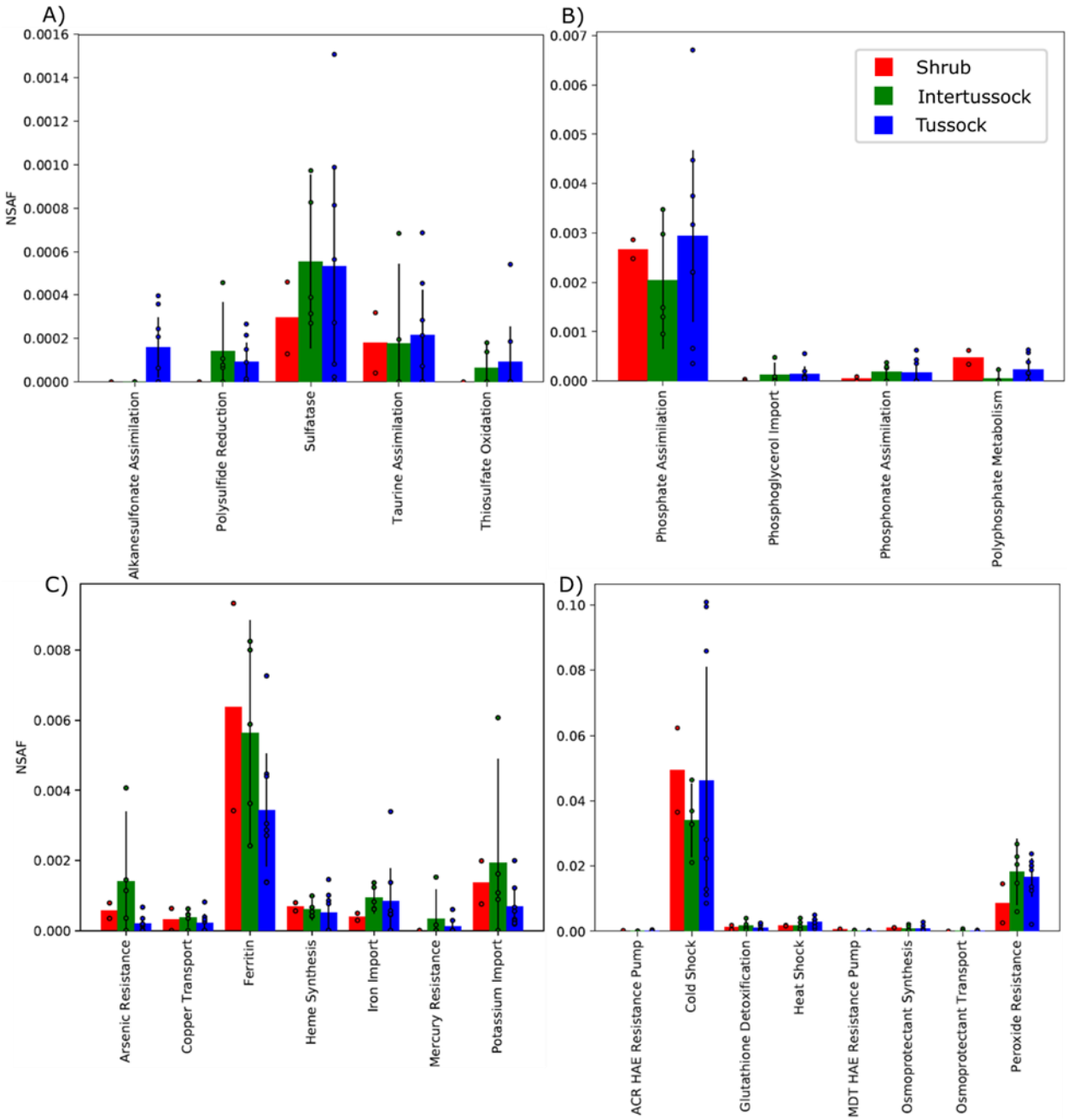

Figure IV.5. Expression levels (NSAF) of Functional Groups associated with A) sulfur, B) phosphorus, C) trace elements, and D) stress

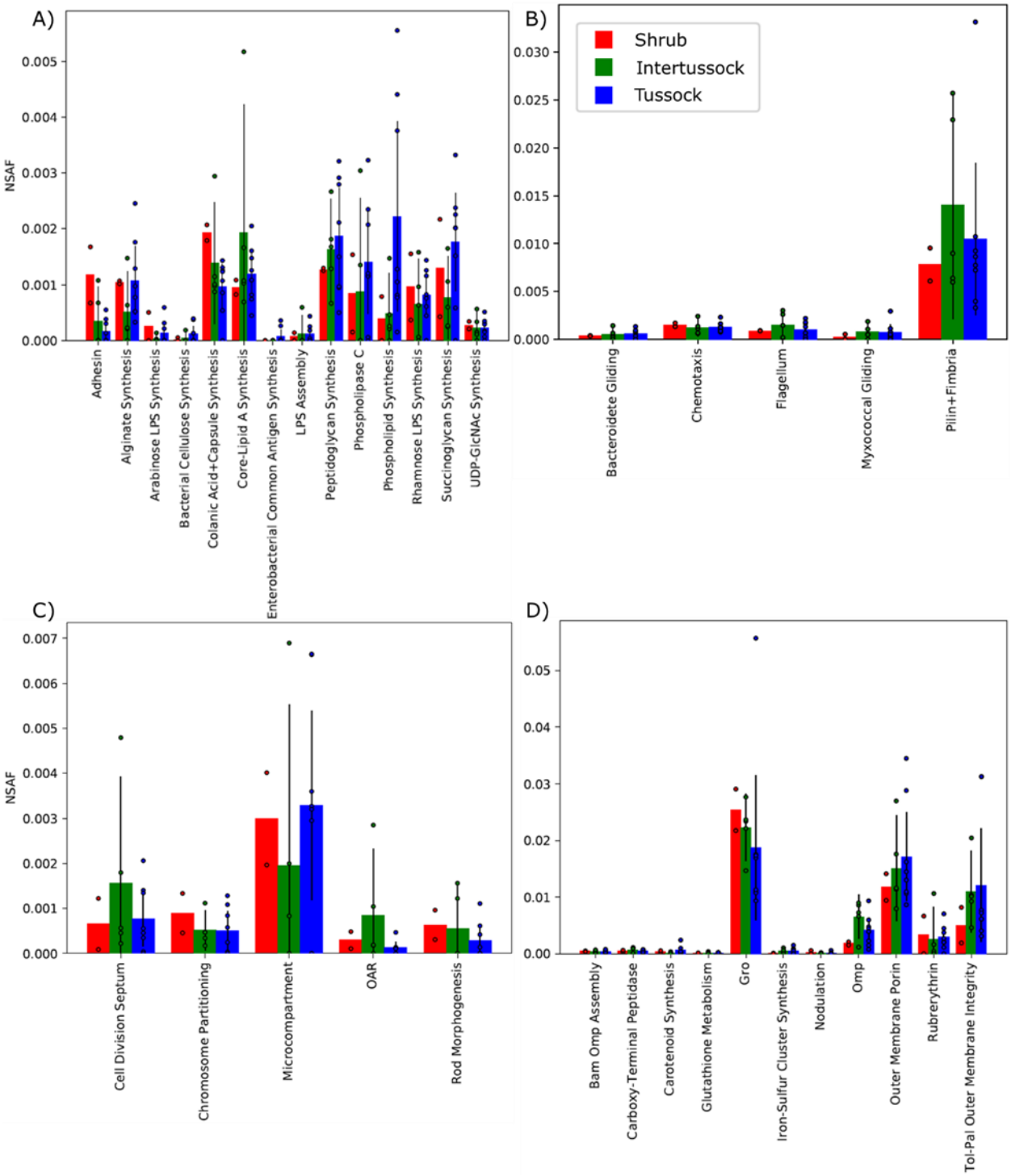

Figure IV.6. Expression levels (NSAF) of Functional Groups associated with A) membrane and wall synthesis, B) movement, C) cell division and structure, and D) other functions

assimilation is detected in tussock but not intertussock samples, and displays a statistically significant difference between the two environments ($p = 0.032$).

Organic soil datasets from different vegetation types form distinct clusters of NSAF data in linear discriminant analysis (LDA; Figure IV.7). Additionally, the three mineral soil datasets often cluster together separately from organic clusters. LDA identifies new axes in a multidimensional dataset (here, sample rows and metaprotein columns) which maximize separation between samples assigned to predefined classes (here, vegetation type). Since a large number of metaproteins with high covariance were considered, the dimensionality of the data was first reduced by principal component analysis (PCA), retaining principal components (PCs) accounting for 90% of variance and using these as input for LDA. Different functional annotation systems for defining metaproteins were compared. GO terms are the most numerous annotations, with many PSMs assigned multiple terms, and describe a variety of protein traits, including cellular localization, at finer or coarser detail. GO terms produce the cleanest separation of the organic and mineral datasets, although the clusters are looser than with metaproteins defined by other annotation systems (Figure IV.7A). KEGG IDs are unique annotations defining proteins with specific, orthologous functions; LDA produces clusters similar to the GO clusters, except that one tussock organic dataset clusters with the mineral datasets (Figure IV.7B). Gene Families and eggNOG orthologous group (OG) descriptions – assigned by eggNOG-mapper to the largest number of PSMs of any annotation system – were used in combination to define metaproteins. This system separates clusters, including a mineral cluster, but one tussock and one intertussock organic dataset also fall with the mineral datasets (Figure IV.7C). Lastly, broader Functional Groups based on the previous annotation system allow datasets to separate cleanly by environment regardless of organic or mineral status (Figure

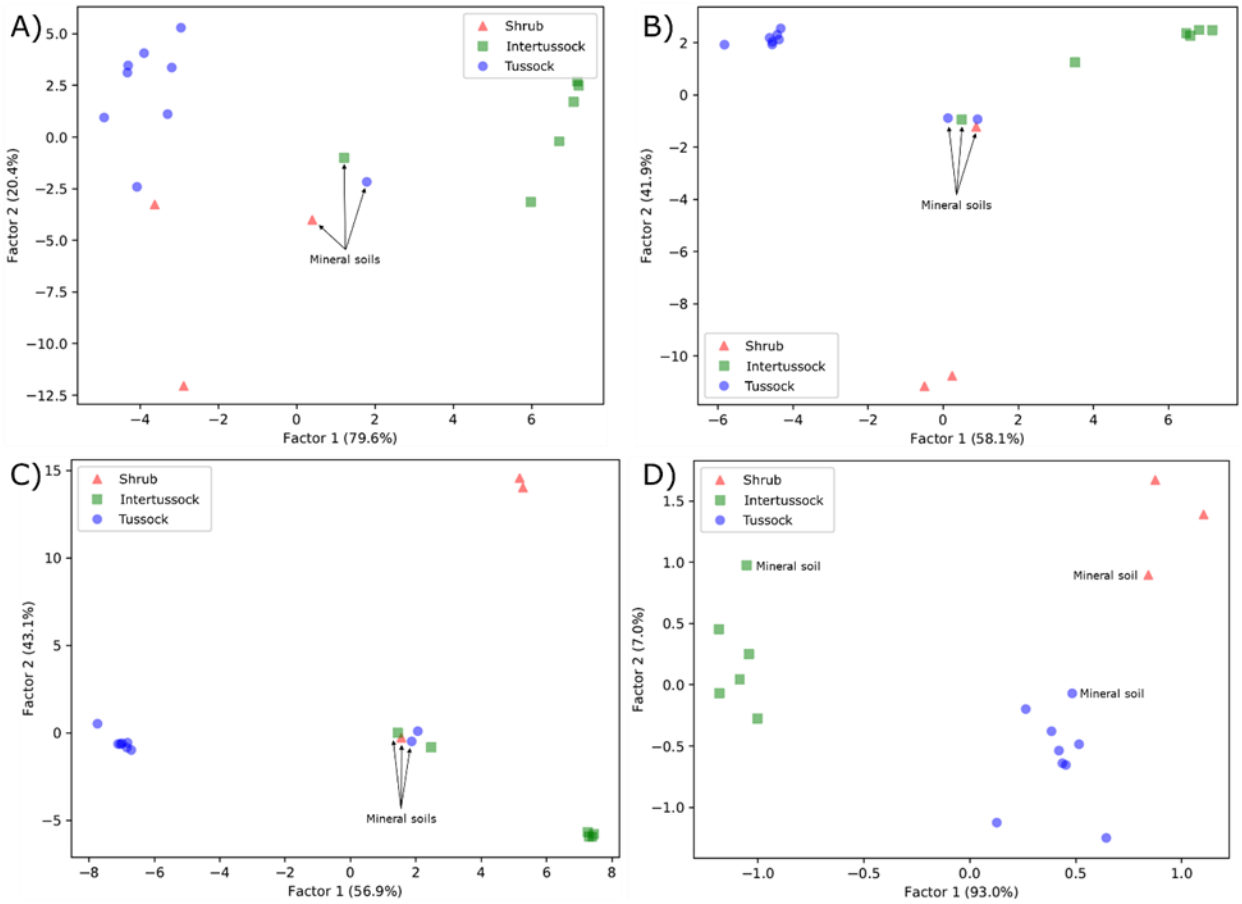

Figure IV.7. Linear discriminant analyses of NSAF data from metaproteins defined by multiple functional annotation systems: A) GO terms, B) KEGG IDs, C) Gene Families + eggNOG OG descriptions, and D) Functional Groups of Gene Family + OG metaproteins. Each point is a different sample, with both organic and mineral soils considered.

IV.7D).

The intermediate position of a distinct mineral cluster between the organic clusters suggests that a set of core functions is present in the mineral soils but functions associated with specific vegetation types are largely absent. Metaprotein NSAF values for the mineral soils confirm that a number of functions related to organic matter decomposition were not detected. Mineral soils lie below the rooting zone adjacent to permafrost, so they are frozen for a longer period of time, preventing the accumulation of plant-derived organics. The separation of the

mineral and three organic clusters is consistent with strong control of microbial community function by the plant products available for consumption.

The latent Functional Groups that contribute most to discrimination between environments by the dominant first discriminant function (axis) are among those that change the most between environments. Sugar transporters have strong positive coefficients, indicating higher expression in tussock and shrub than intertussock soils. Again, these proteins have statistically significant differences in expression between tussock and intertussock datasets. Rhizobiales dominate the expression of sugar transporters along with others that have highly positive coefficients (see Section IV.C.2), including cold shock proteins, proteins required for polyphosphate metabolism, and proteins required for biosynthesis of the extracellular polymeric substances (EPS), succinoglycan and alginate. Rhizobiales and their unique functional profile, which seems to involve interactions with plants (see Section IV.C.2), are more prevalent in soils with greater floral biomass. Proteins involved in the assimilation of inorganic and small ($\leq$3 C) compounds have more negative coefficients, indicating higher expression in intertussock than tussock and shrub soils. These Functional Groups include CO_2 assimilation, acetate metabolism via acetyl-CoA synthetase, and ammonium transport. Although these Functional Groups have much lower overall expression levels than sugar transporters, they are more strongly expressed in lower biomass intertussock soils. Increasing Arctic plant biomass and the displacement of nonvascular by vascular plants may therefore alter microbial biogeochemistry by increasing the dependence of the microbial community on organic substrates derived from plants.

IV.C.1.ii. MULTIVARIATE ANALYSIS OF PROTEIN EXPRESSION BY TAXA IN DIFFERENT ENVIRONMENTS

Bin fidelity measures the strength of metaprotein expression by taxa, or bins of nucleotide sequence data. This metric takes into account both metaprotein abundance and the likelihood of metaprotein expression by a taxon. With the exception of three unique peptide-spectrum matches (PSMs) from Archaea, all annotated PSMs are bacterial in origin, reflecting the preponderance of bacterial rather than plant, fungal or archaeal protein biomass in the soils. Fungi are better adapted to well-oxygenated litter at the surface than the moist soil environment. Pathways involved in the final stages of lignin degradation are represented in the metaproteomic datasets, but the initial steps typically catalyzed by fungi and, to a lesser degree, bacterial actinomycetes are not detected, suggesting that lignin breakdown by oxygenases occurs more at the surface, as in other biomes. The near absence of methanogenic Archaea in the datasets is in accordance with very low levels of identified anaerobic metabolisms, indicating that some amount of O_2 is present in porewaters. A lack of archaeal methanogenesis is also consistent with the non-detection of proteins required for methane oxidation (e.g., methane monooxygenase), despite the identification of methylotrophic pathways throughout the datasets.

Multivariate analyses of bin fidelity data were used to understand relationships between taxonomic functional profiles as well as the magnitude of changes in profiles between taxa versus environments. LDA was conducted with 1) data points for each taxon defined by a vector of bin fidelities for each Functional Group (Figure IV.8A), and 2) data points for each Functional Group defined by a vector of bin fidelities for each taxon (Figure IV.8B). These complementary analyses reveal that vegetation types cluster by the functional profiles of taxa but not by the taxonomic profiles of functions. In other words, taxa differ between environments in terms of the

levels of functions they express to a much greater extent than functions differ between environments in terms of the taxa in which they are expressed. This accords with the occurrence of the same taxa across all of the vegetation types and differentiated functional profiles between taxa that seem tailored to different plant inputs.

PCA was used to project taxon data points from a space of dimensionality equal to the number of metaproteins onto a small number of principal component (PC) axes in the directions of maximum variance (Figures IV.9 and IV.10). Considering Functional Groups, the first two PCs account for 37.1% of variance in the data (Figure IV.9A), and the next two account for a further 17.0% (Figure IV.9B). Some of the separation between environments that is made clear by LDA is also apparent along PCs 1 and 3. There is more separation between taxa when metaproteins are defined by GO terms (Figure IV.10). To quantify the difference between environments along PCs 1 and 2, MANOVA was applied to the Functional Group and GO bin fidelities. The non-significant result for Functional Groups (Wilks's lambda $F = 0.41$, $p = 0.67$) and marginally significant result for GO terms (Wilks's lambda $F = 3.8$, $p = 0.032$) corroborates the relatively weak effect of environment on metaprotein expression by taxa. Differences in expression profiles between environments do not greatly exceed the differences between taxa that exist within environments, reflecting the relative stability of the partitioning of functions between taxa.

The projection on PCs 1 and 2 in Figure IV.9A shows that Rhizobiales (α-proteobacteria), β-proteobacteria, and γ-proteobacteria lie in the lower left corner; actinobacterial groups lie in the lower right, with Class Actinobacteria occupying a more central position; Acidobacteria lie at the top, with Myxococcales and Bacteroidetes intermediate between Acidobacteria and Actinobacteria. Implications of this arrangement of points in terms of the

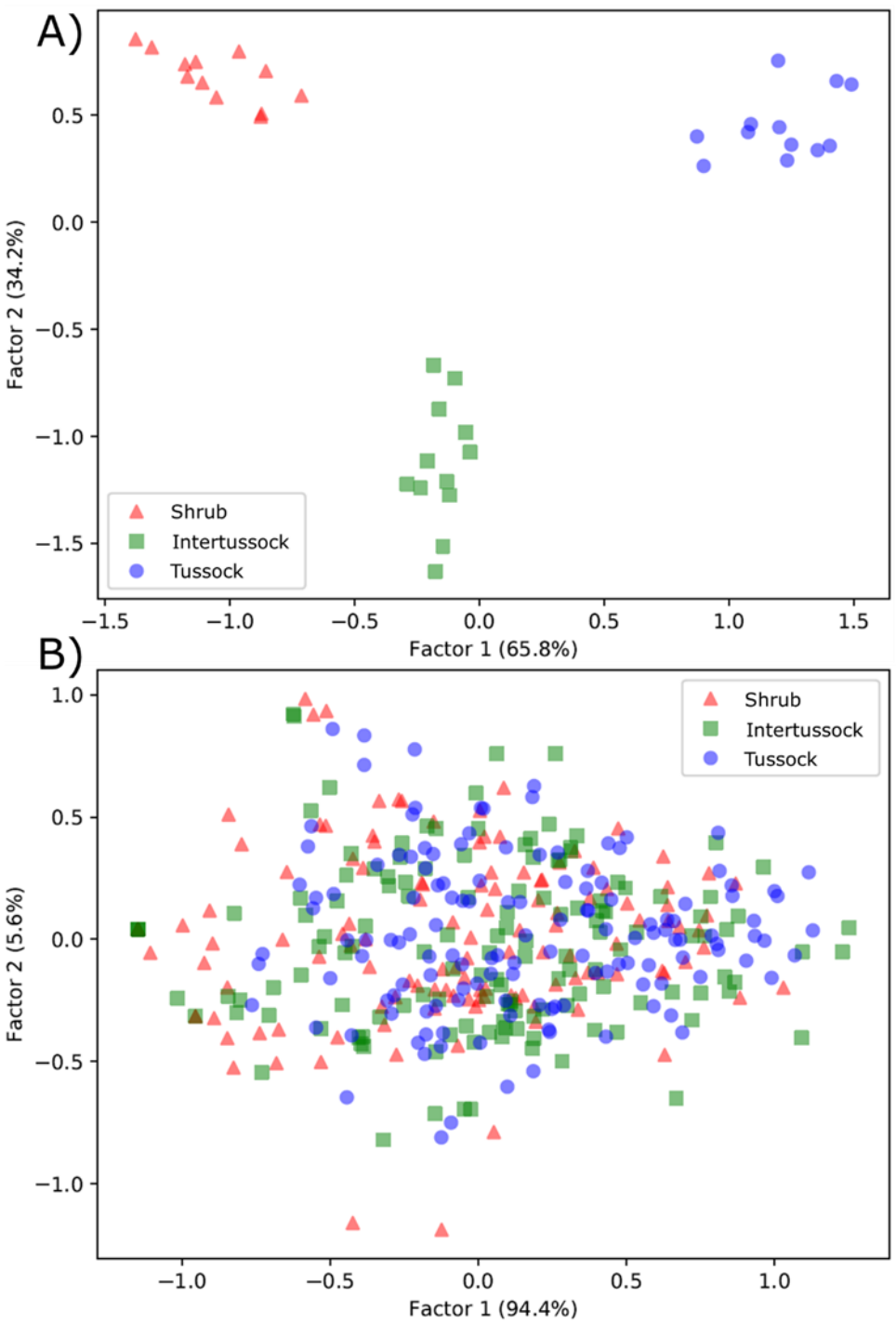

Figure IV.8. Linear discriminant analyses of bin fidelities with A) each data point representing a bin and B) each data point representing a Functional Group.

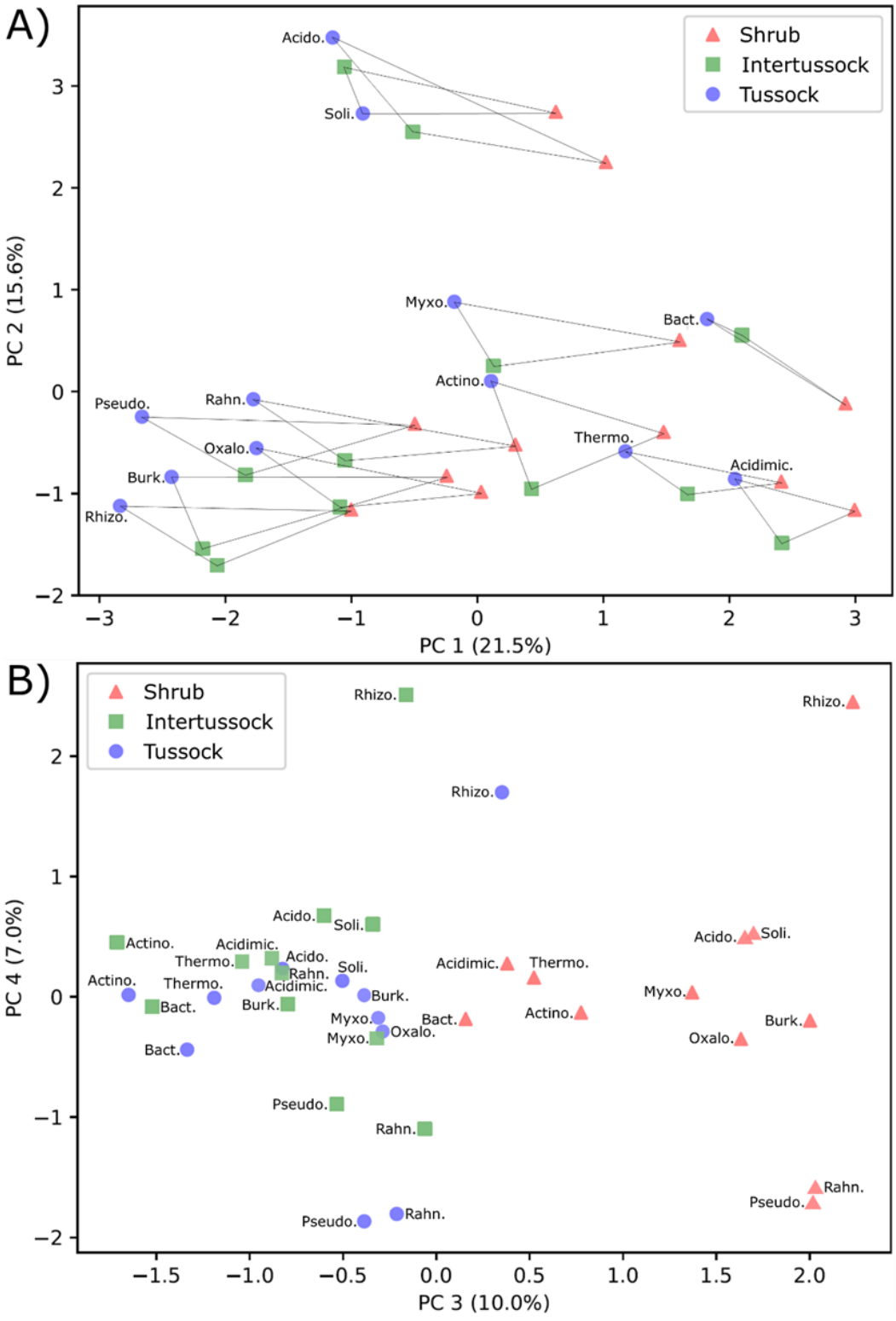

Figure IV.9. Principal component analyses of bin fidelities for Functional Groups, showing A) principal components 1 and 2 and B) principal components 3 and 4. The percentage in parentheses is the proportion of variance explained by the PC.

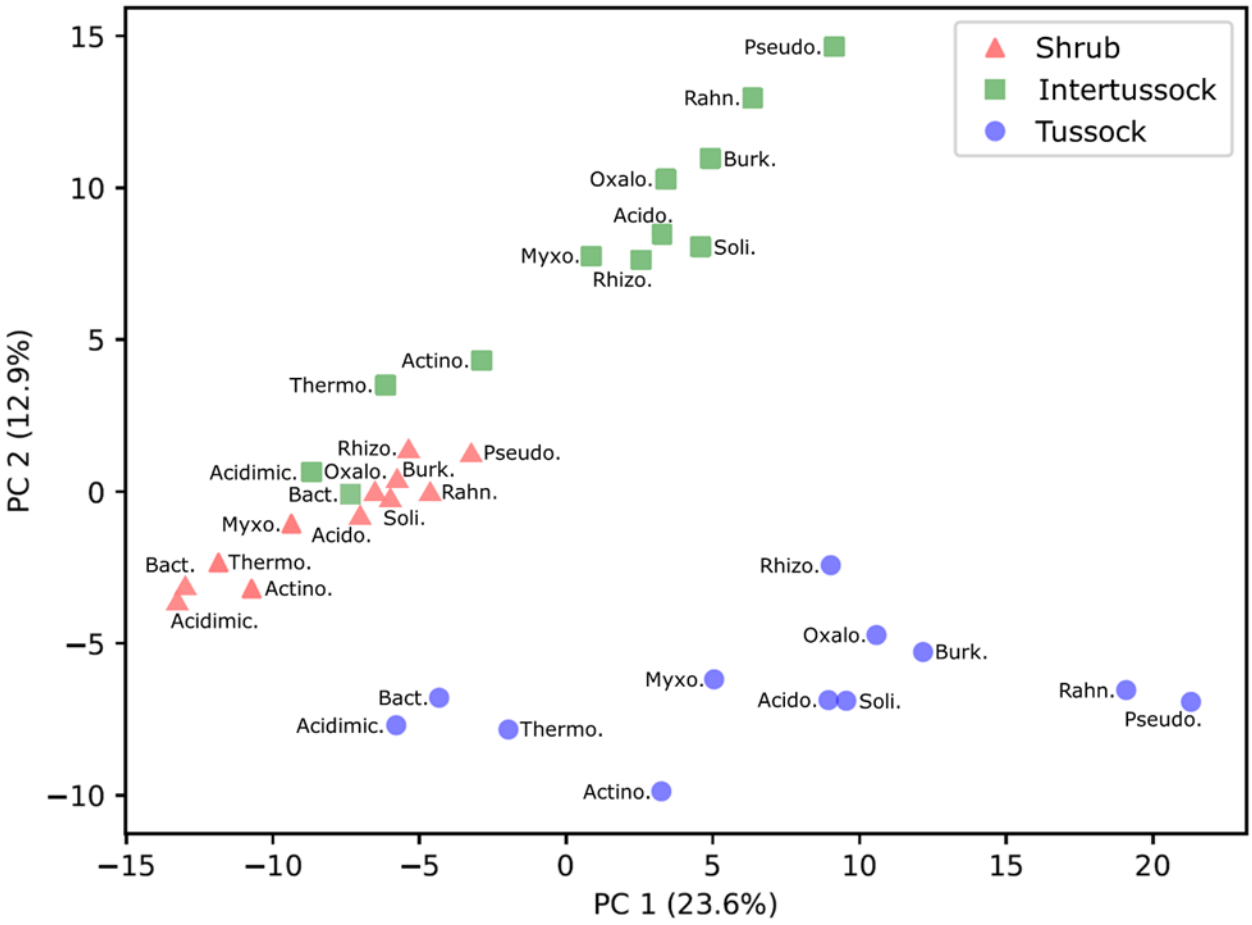

Figure IV.10. Principal component analyses of bin fidelities for metaproteins defined by GO terms. The percentage in parentheses is the proportion of variance explained by the PC.

similarities of taxonomic functional profiles are confirmed by the detailed investigation in Section IV.C.2. For instance, the central position of Class Actinobacteria reflects its moderate expression of some functions that are more strongly expressed by either Proteobacteria or Acidobacteria. The positions of taxa along PCs 1 and 2 shift in a predictable manner with changes in environment, with the two exceptions being Solibacteres (Acidobacteria) for tussock and intertussock samples, and Bacteroidetes. This indicates that the functional response of taxa to changes in environment is predictable to some extent.

The contribution of latent functional variables to the bin fidelity linear discriminant functions in Figure IV.8.A (loadings shown in Appendix: Tables VI.1-2) as well as differences in *functional* fidelity values between environments (Section IV.C: Figures IV.14, IV.18, IV.22 and IV.26) provide insight into the drivers of functional differences between environments. As defined in Section IV.B.3.ii, functional fidelity indicates the relative expression of functions within a given taxon, as opposed to bin fidelity's measurement of the partitioning of a given function among taxa. Factor 2, or the second discriminant function in the bin fidelity LDA, cleanly separates low (intertussock) from high (tussock/shrub) floral biomass environments.

Some of the most abundant Functional Groups, which are also indicative of overall cellular activity, contribute strongly to differences between environments. The Ribosome, DNA Supercoiling, and Chromosome Packaging contribute positively to Factor 2 and have higher functional fidelities in higher biomass floras (Section IV.C.2.i: Figure IV.14). Proteobacteria and Actinobacteria but not other groups, including the most active group, Acidobacteria, have higher Ribosome functional fidelities in higher biomass floras, supporting the possibility of greater rhizospheric proteobacterial activity in more heavily rooted soils. Sugar transporters (e.g., Ribose Transport) are significantly more abundant (Section IV.C.1.i) and also have higher Factor 2 loadings and functional fidelities in higher biomass floras (Section IV.C.2.ii: Figure IV.18), with functional fidelity values increasing relatively uniformly across taxa, maintaining predominant expression among Proteobacteria. These changes in sugar transporters are consistent with the greater importance of root interactions in the microbial functional profile, as the transporters may be used to acquire root exudates (Section IV.C.2.ii). Ammonia Metabolism is an abundant Functional Group with a *lower* Factor 2 loading and *lower* functional fidelities in higher biomass floras (Section IV.C.2.iii: Figure IV.22), supporting the hypothesis developed in Section

IV.C.2.iii that rhizospheric bacteria are competing with plants for scarce N; glutamine synthetase, the predominant metaprotein in this Ammonia Metabolism Functional Group, is required for the biosynthesis of key nitrogenous compounds. The change in the TCA Cycle between environments is more convoluted in terms of fidelity metrics but nonetheless makes sense in light of the suggested shift toward proteobacterial activity in more heavily rooted soils. This TCA Cycle Functional Group contributes positively to Factor 2 but decreases strongly in the non-Proteobacteria and less so in the Proteobacteria in terms of functional fidelity (Section IV.C.2.i: Figure IV.14). Decreasing functional fidelity indicates that TCA cycle expression decreases relative to that of other proteins in cells, but the positive loading on Factor 2 indicates that within the community, TCA cycle expression becomes more concentrated in Proteobacteria than non-Proteobacteria.

IV.C.2. COMPARISON OF THE FUNCTIONAL PROFILES OF TAXA

IV.C.2.i. OVERALL PATTERNS AND CELLULAR ACTIVITY

The bin fidelity data reveal strong functional niche differentiation between the 12 major bacterial taxa identified, as illustrated in complementary ways by Figures IV.11-IV.26. For each function, bin fidelities were normalized to the maximum value, so at least one taxon has the maximum value of 1 while the other taxa range from 0 to 1. Overarching patterns in the bin fidelity data were explored using *k*-means clustering by Euclidean distance, with $k = 3$ clusters corresponding to a breakpoint in the reduction of the sum of squared errors with the addition of clusters (Table IV.4). The three clusters highlight functional differences relevant to the interlinked soil C and N cycles and the ecophysiology of abundant but poorly characterized microbial groups – especially Acidobacteria.[75–80]

Number of clusters	Sum of squared errors
1	119.5
2	92.51
3	74.89
4	68.37
5	62.34
6	57.42
7	53.67
8	50.11
9	47.42
10	45.04

Table IV.4. *k*-means clustering of bin fidelity vectors

The largest cluster (Cluster 2) comprises Functional Groups generally expressed at moderate or high levels by both Acidobacteria and Proteobacteria. ("Proteobacteria" is used in this paper to indicate α-, β-, and γ-proteobacterial taxa but not the δ-proteobacterial order, Myxococcales, which displays distinct functional patterns and is only mentioned explicitly.) Cluster 2 contains most of the Functional Groups responsible for the essential cellular functions of DNA replication and repair, transcription, translation, and cell division (Figure IV.12). Acidobacteria most strongly express these Functional Groups, indicating that by a combination of cell biomass and per cell activity, Acidobacteria are the most active bacterial group. This is broadly consistent with the relatively high abundance of Acidobacteria in many of the 16S rRNA gene libraries from the Toolik area.[81–83] Clusters 1 and 3 generally contain Functional Groups that are dominated by Proteobacteria and Acidobacteria, respectively. Cluster 1 includes many functions related to the transport of N, including the high-abundance (by NSAF) amino acid transporters, suggesting a critical role of Proteobacteria in N throughput (Figure IV.20). Regarding the C cycle, Cluster 1 includes many monosaccharide transporters and pathways for

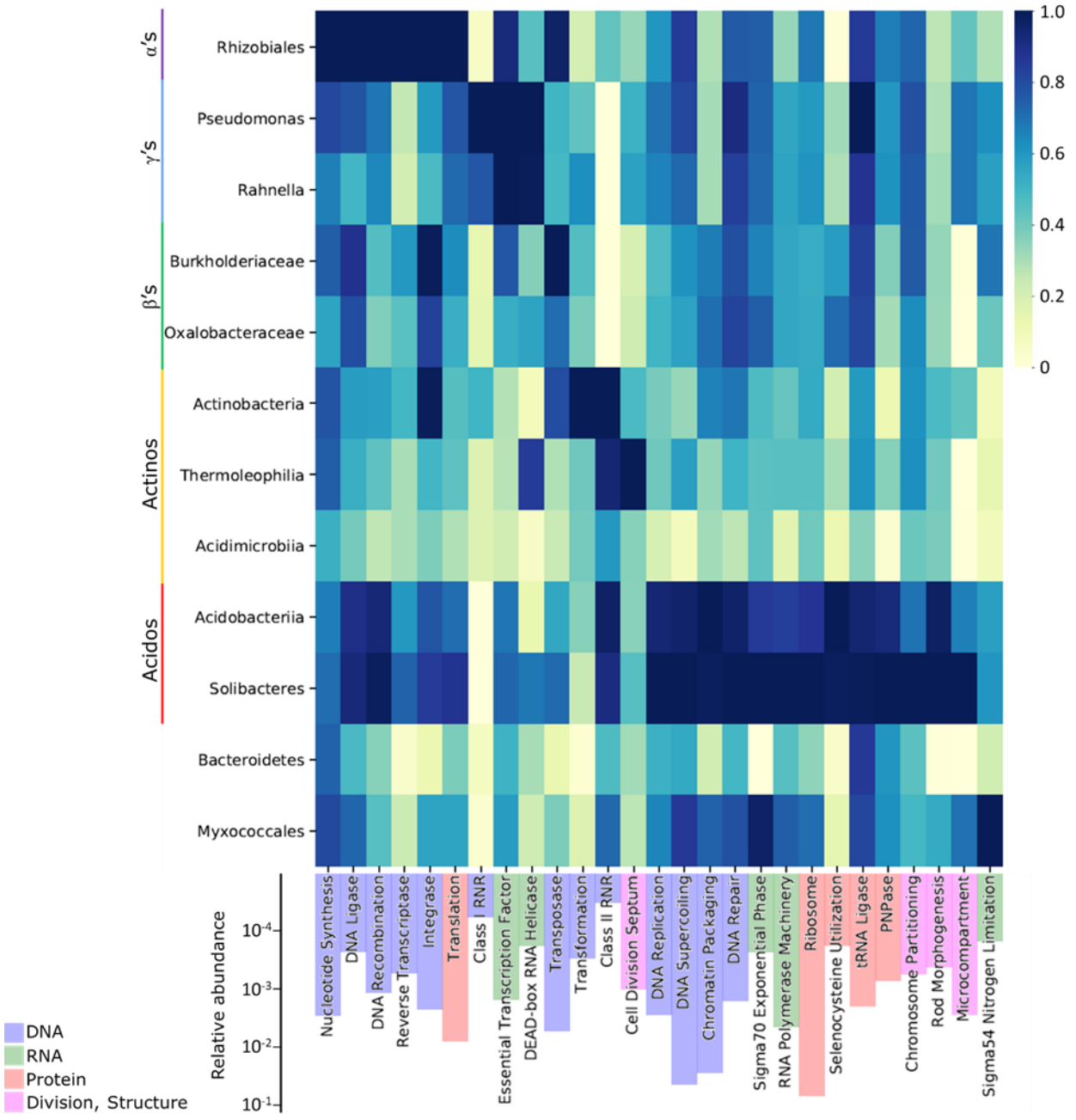

Figure IV.11. Cell growth-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples. Columns are ordered by the taxonomic bin with the maximum expression (darkest blue cell) and then by Functional Group category (bar color).

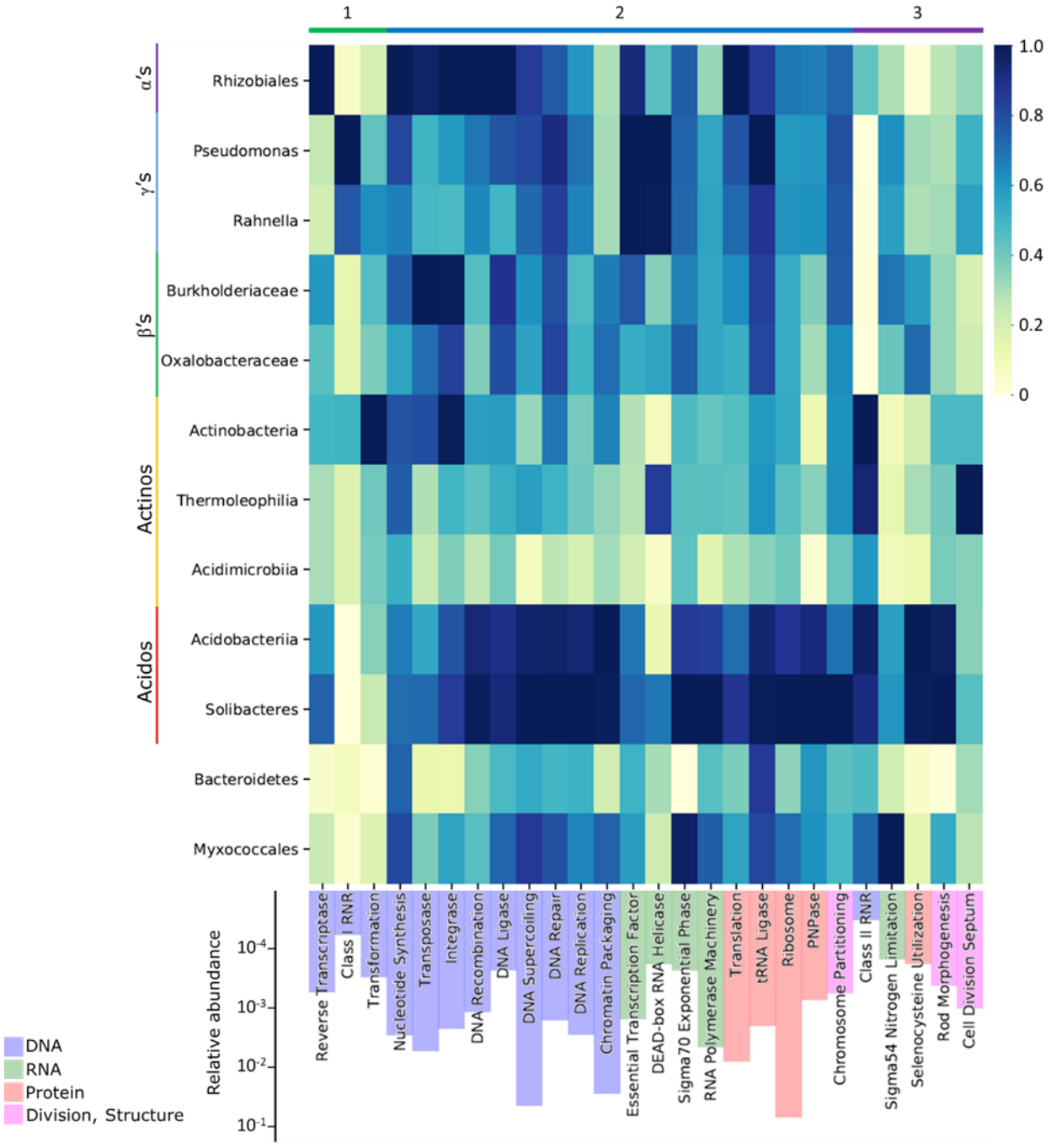

Figure IV.12. Cell growth-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples, with k-means cluster assignments at top. Columns are ordered by cluster assignment and then by Functional Group category (bar color).

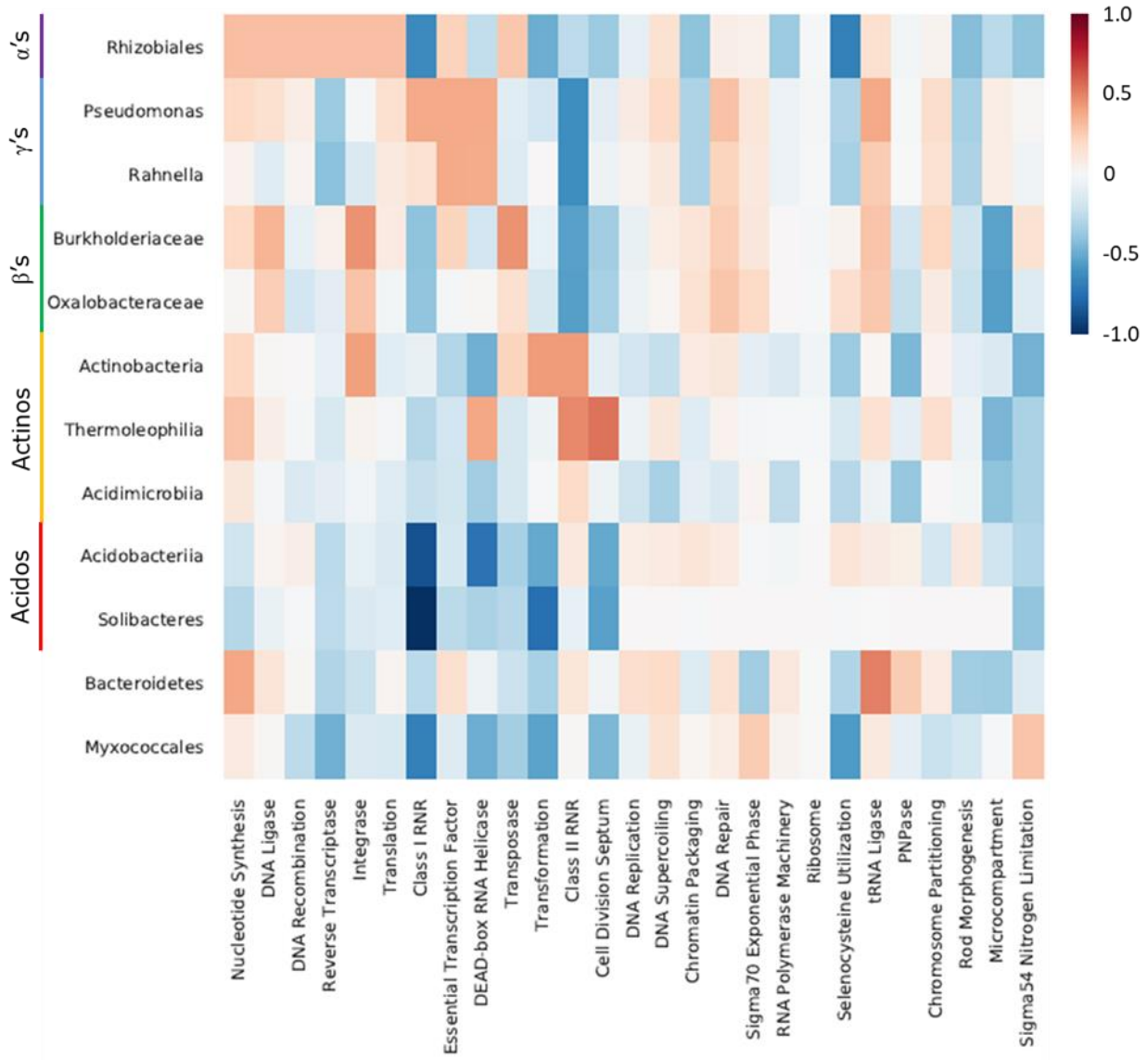

Figure IV.13. Cell growth-related Functional Group bin fidelities normalized to the maximum value in the column, averaged over all organic soil samples, with the Ribosome values then subtracted. This indicates the relative levels of Functional Group expression by taxa compared to a baseline of Ribosome expression. For instance, Nucleotide Synthesis is not dominated as heavily by the Acidobacteria as is Ribosome expression (Nucleotide Synthesis proteins are more evenly expressed across groups), so the values for Acidobacteria are negative (blue) while those for the other taxa are positive (red). Columns are in the same order as Figure IV.11.

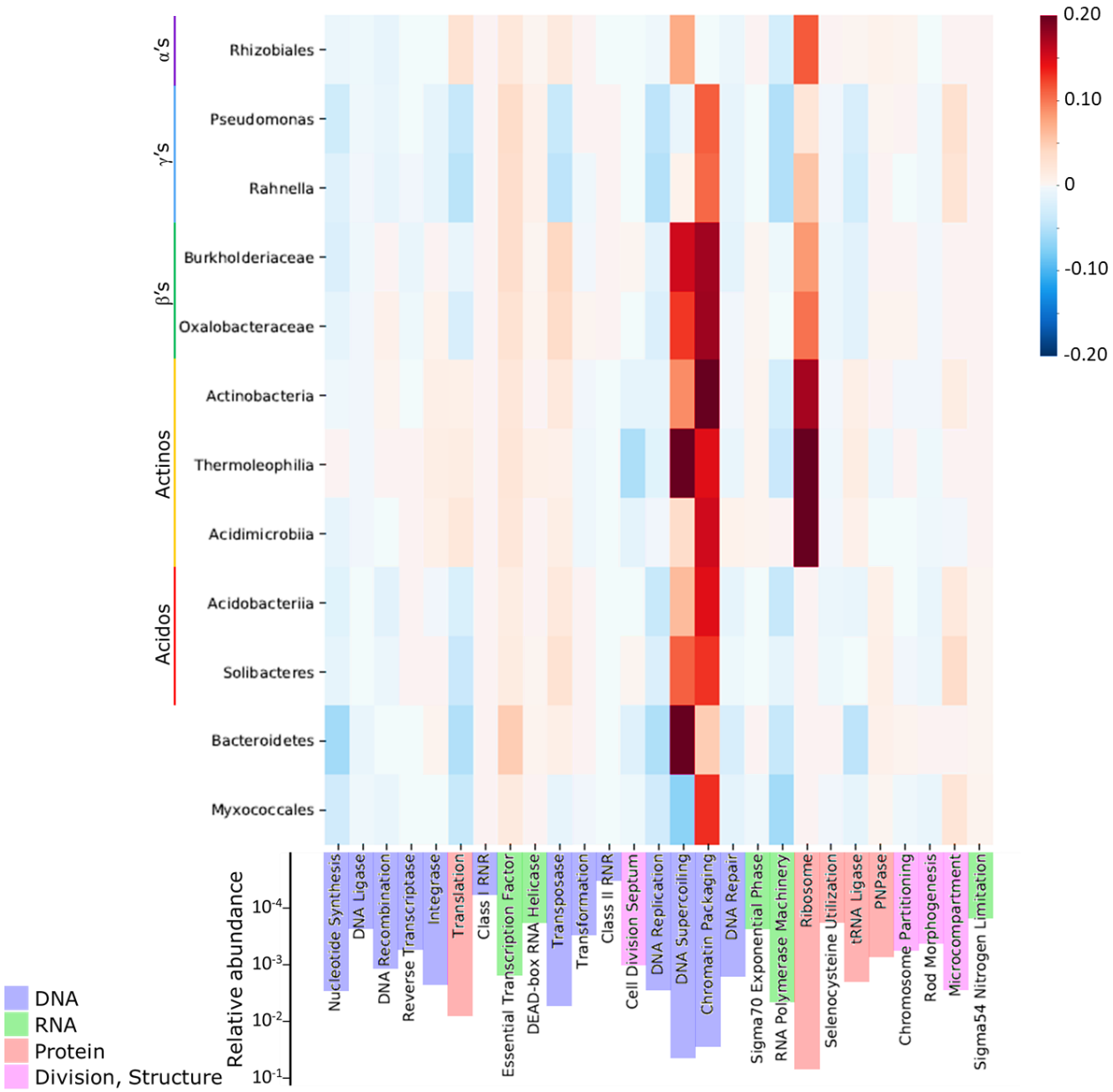

Figure IV.14. Difference in cell growth-related Functional Group functional fidelities between tussock/shrub (high plant biomass) and intertussock (low plant biomass) organic soil samples. The values represent the relative change in average functional fidelity from intertussock to tussock/shrub samples. Columns are in the same order as Figure IV.11.

the utilization of gases, such as CO, and ≤3 C solutes, such as methanol. In contrast, Cluster 3 contains most of the pathways required for the degradation of the abundant plant polysaccharides, cellulose, hemicellulose, and pectin (Figure IV.16). The division of processes involving polysaccharides and monosaccharides between Acidobacteria and Proteobacteria is explored further in Section IV.C.2.ii. Ribosomal proteins, the most abundant group of proteins identified in the datasets, are used to benchmark the expression of other functions in Figures IV.13, IV.17, IV.21, and IV.25 (Section IV.C), as the production of ribosomes reflects a combination of cellular abundance and growth rate. Taxonomic expression profiles with marked differences to the ribosome suggest that the function is not expressed in proportion to the overall activity of the taxa. This comparison highlights differences between Cluster 2, representative of overall activity, and Clusters 1 and 3, containing functions that skew strongly toward particular taxa.

Bin fidelity data also reveal finer patterns of niche partitioning, such as the monopolization of abundant EPS production pathways by Rhizobiales, suggesting that this group is a prodigious producer of biofilms – an ecophysiological trait with biogeochemical implications (Section IV.C.2.iv). The 12 identified bacterial taxa are therefore summarized in the remainder of this section, with bin names emboldened. Acidobacterial contigs fall in two bins most closely related to the isolates *Koribacter versatilis* (**Acidobacteriia**) and *Solibacter usitatus* (**Solibacteres**), with both having similar functional profiles. **Bacteroidetes** (with a majority of contigs affiliated with *Chitinophaga*) and **Myxococcales** (δ-proteobacteria) share a number of acidobacterial characteristics, especially regarding C substrate preferences. Three groups from Phylum Actinobacteria were identified. **Class Actinobacteria** overlap certain traits of Acidobacteria and others of Proteobacteria; **Thermoleophilia** and **Acidimicrobiia** appear to be

the most inactive groups in the community. **Rhizobiales** (an order of Class α-proteobacteria), β-proteobacteria (**Burkholderiaceae** and **Oxalobacteraceae**), and γ-proteobacteria (***Pseudomonas*** and ***Rahnella***) have expression profiles that cluster together yet display certain distinctions. The main functions of biogeochemical significance in the soil microbial community are explored in the following sections and summarized in Section IV.D: Figure IV.27.

IV.C.2.ii. CARBON METABOLISM AND ENERGY CONSERVATION

The bin fidelities of Functional Groups involved in C metabolism are explored in Figures IV.15-IV.18. Acidobacteria play a central role in C biogeochemistry in these datasets, as the group exhibits the highest expression of most core C metabolism pathways. Acidobacteria, Class Actinobacteria, Bacteroidetes, and Myxococcales dominate the depolymerization of polysaccharides into monosaccharides by extracellular enzymes within the soils. Actinobacteria most strongly express endoglucanase and cellobiosidase, enzymes required for the debranching and cleavage of oligosaccharides from cellulose, as well as accessory enzymes to the glycolytic pathway required for the catabolism of hexoses beside glucose, such as fructose and galactose. Actinobacteria also have relatively high average expression levels of the following enzymes essential for organic matter degradation: β-glucosidase, required for the cleavage of terminal glucose monomers from oligosaccharides; enzymes such as debranching enzyme, α-amylases and phosphorylases involved in the degradation of starch and glycogen, which are weaker glucose homopolymers than cellulose; enzymes required for the cleavage of pentose monomers from heteropolysaccharides such as hemicellulose; and enzymes cleaving N-acetylglucosamine (GlcNAc), a monomeric unit of microbial cell walls and chitin.

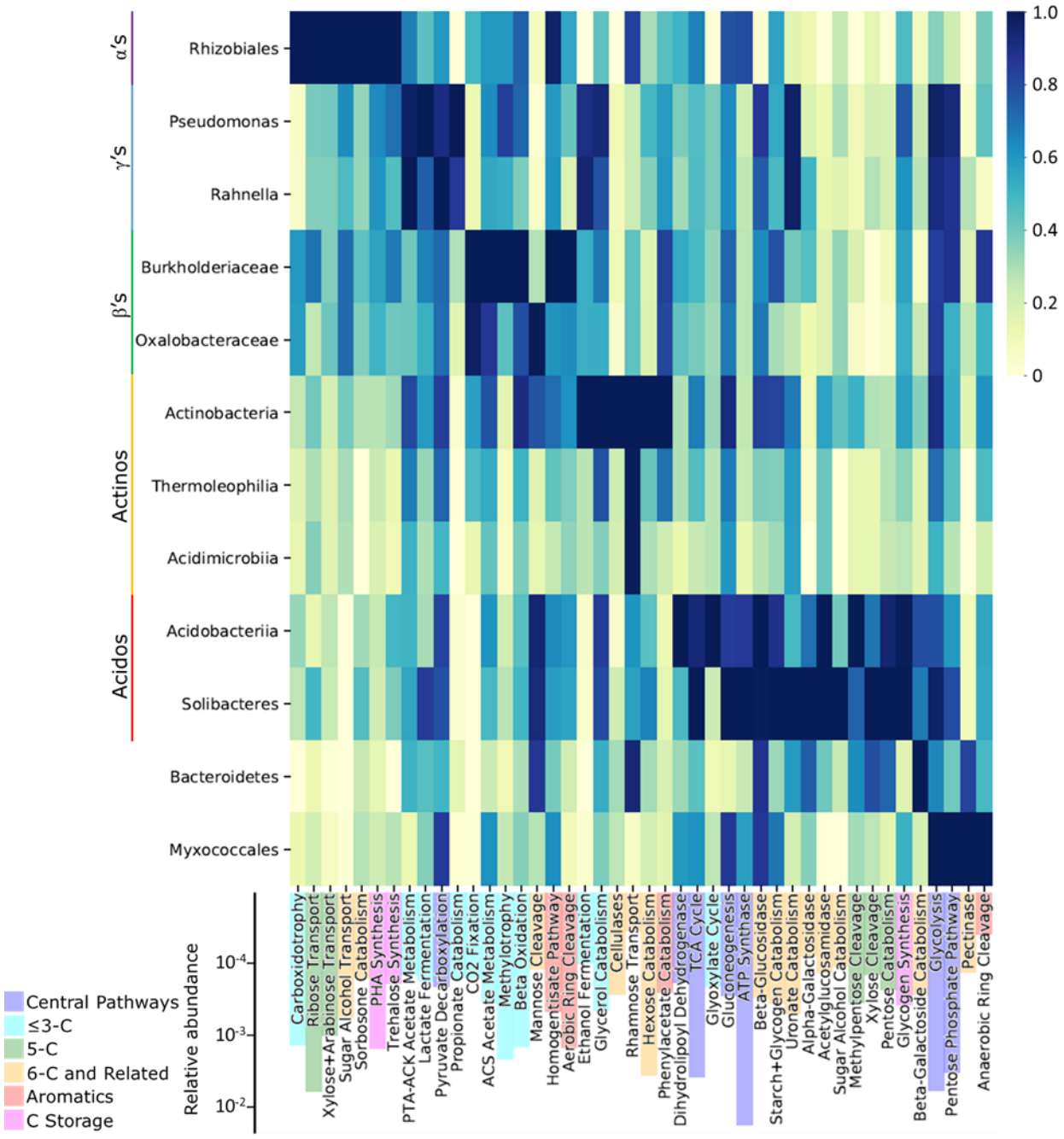

Figure IV.15. Carbon-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples. Columns are ordered by the taxonomic bin with the maximum expression (darkest blue cell) and then by Functional Group category (bar color).

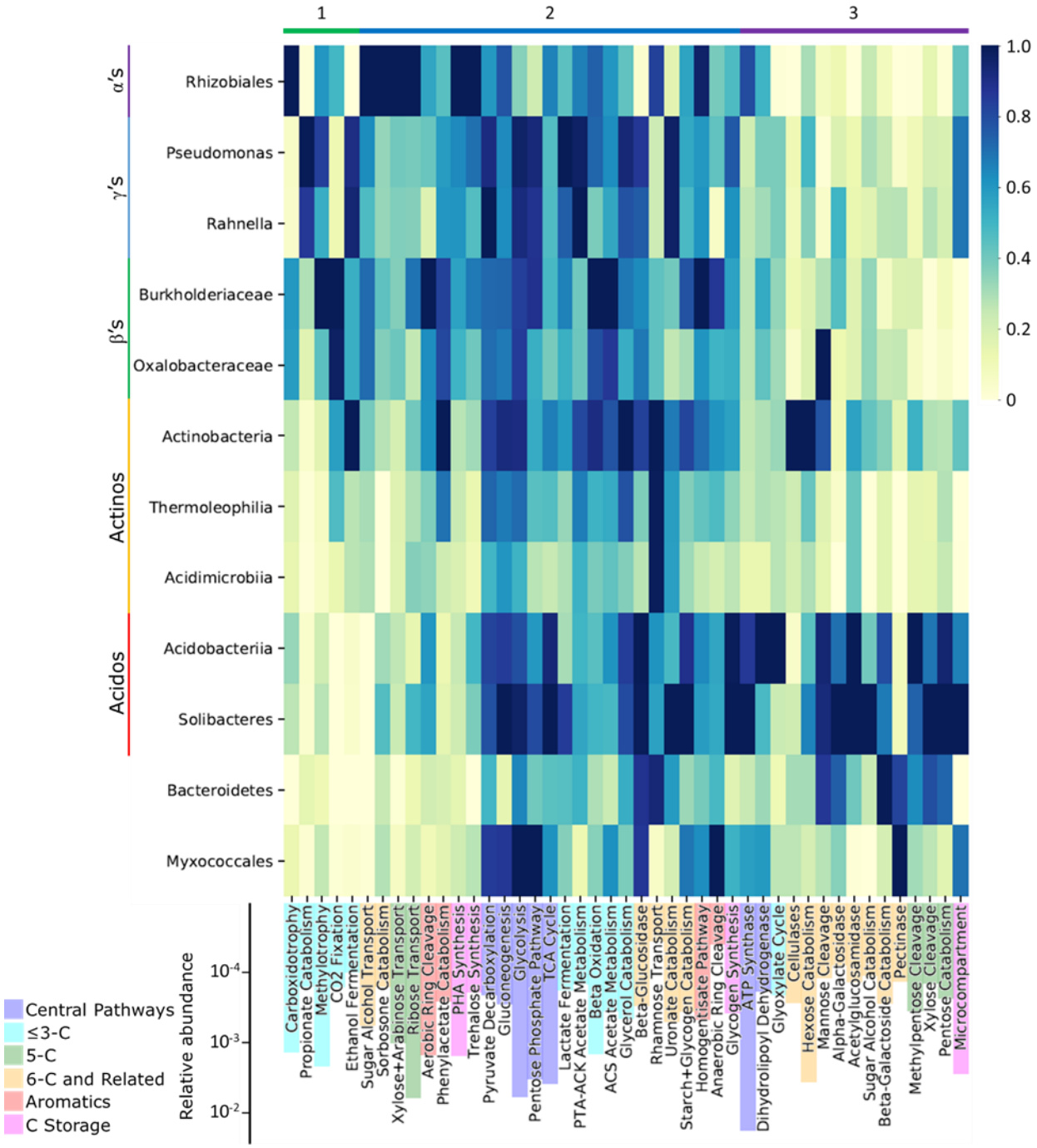

Figure IV.16. Carbon-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples, with k-means cluster assignments at top. Columns are ordered by cluster assignment and then by Functional Group category (bar color).

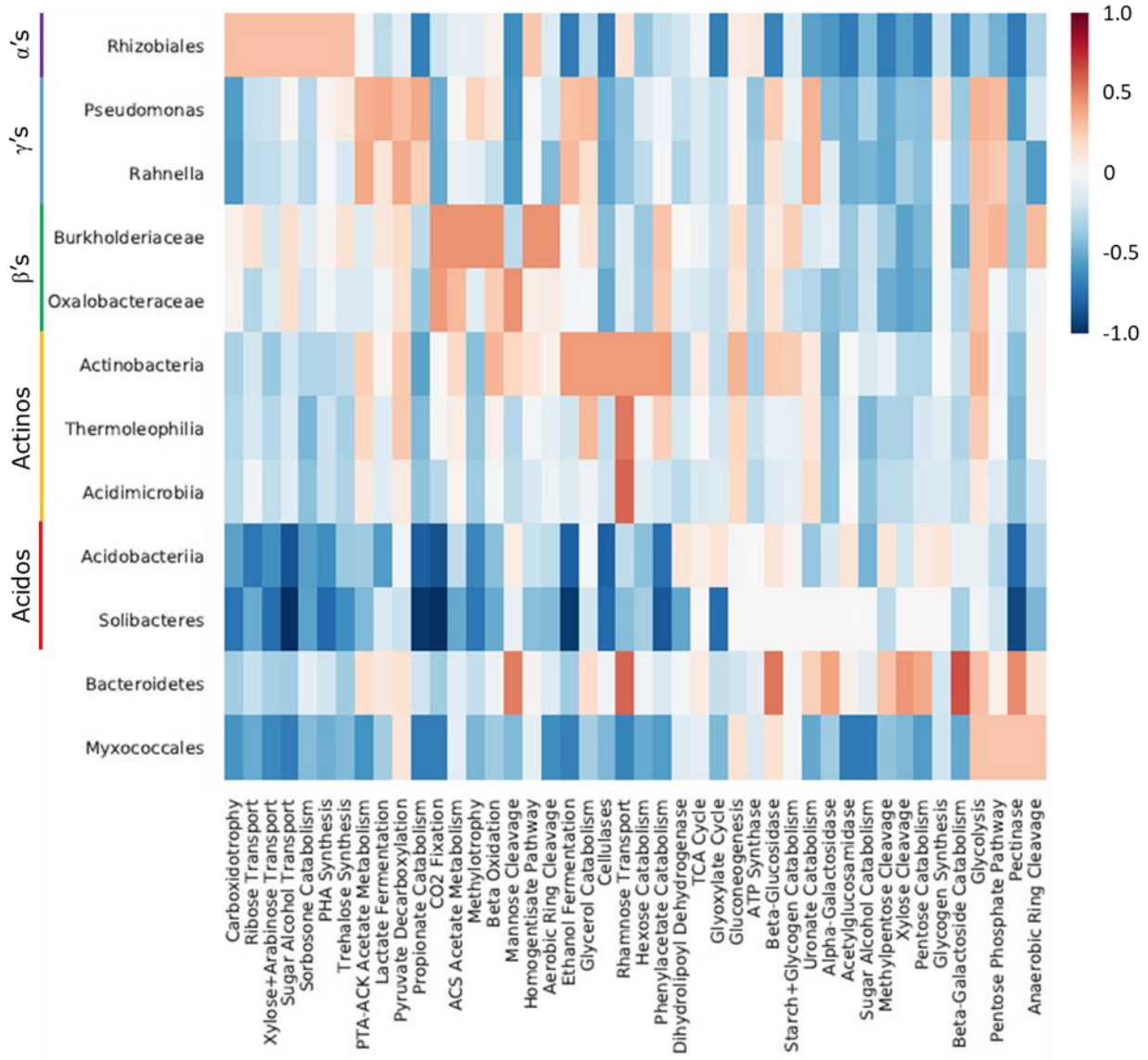

Figure IV.17. Carbon-related Functional Group bin fidelities normalized to the maximum value in the column, averaged over all organic soil samples, with the Ribosome values then subtracted. This indicates the relative levels of Functional Group expression by taxa compared to a baseline of Ribosome expression. For instance, Carboxidotrophy is dominated by Rhizobiales, so the values for most taxa decrease versus the Ribosome (blue) while values for Rhizobiales increase (red). Columns are in the same order as Figure IV.15.

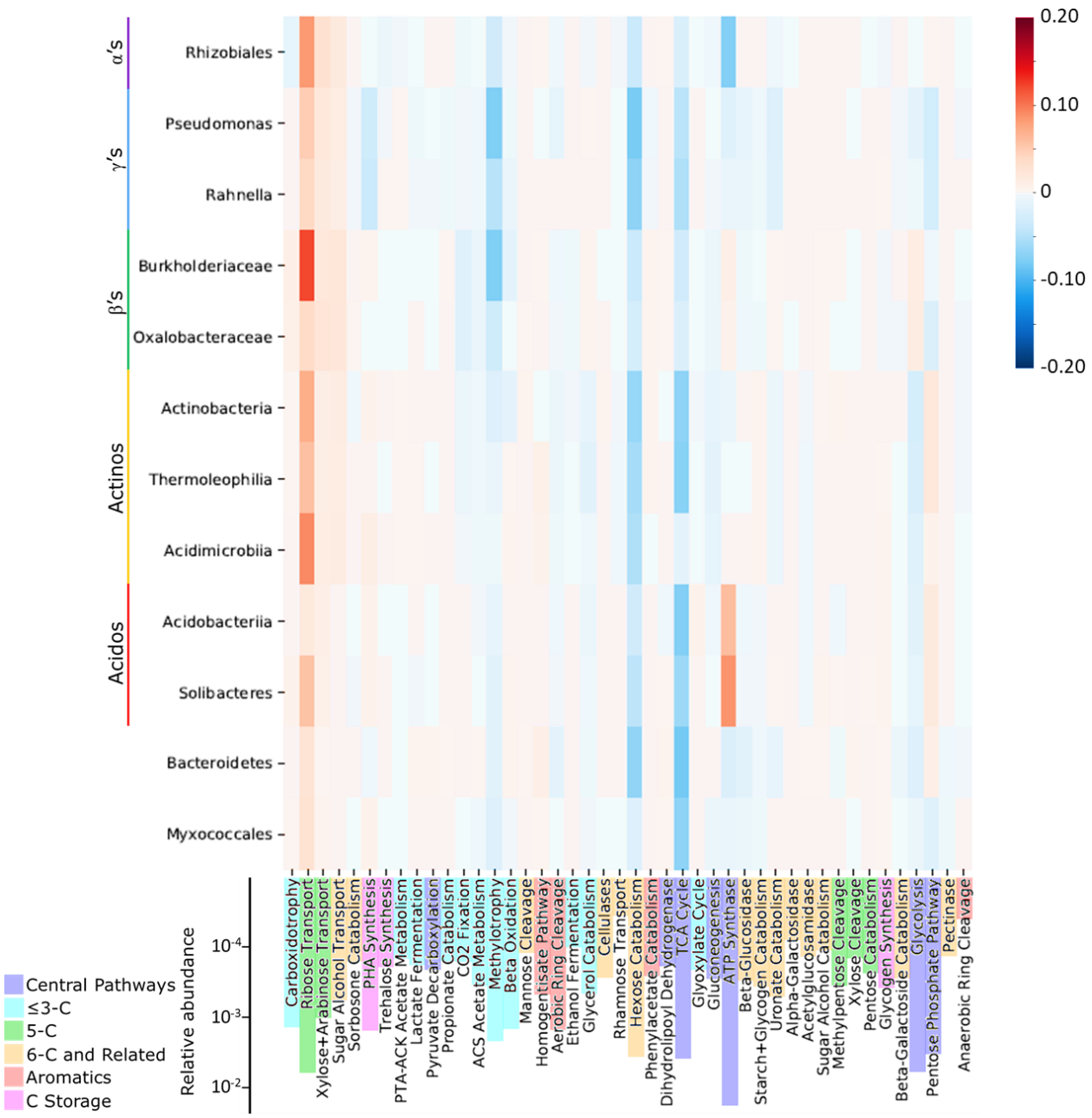

Figure IV.18. Difference in carbon-related Functional Group functional fidelities between tussock/shrub (high plant biomass) and intertussock (low plant biomass) organic soil samples. The values represent the relative change in average functional fidelity from intertussock to tussock/shrub samples. Columns are in the same order as Figure IV.15.

Acidobacteria appear to dominate hemicellulose degradation and play a major role in the degradation of starch/glycogen and GlcNAc. Hemicellulose and pectin form a matrix around cellulose microfibrils in plant cell walls. Hemicellulose constitutes 20-30% of plant dry weight.[84] The depolymerization of hemicelluloses involves enzymes acting on a variety of bonds, including β-1,4-linked D-xylose, β-1,4-linked D-glucose, and β-1,4-linked D-mannose.[85] Acidobacteria also dominate the expression of enzymes required for the metabolism of diverse pentoses yielded from polymer breakdown, such as xylose and arabinose, which, unlike ribose, do not directly enter the pentose phosphate pathway. Bacteroidetes and Myxococcales express moderate levels of enzymes required for cellulose and hemicellulose decomposition. These taxa are distinguished from Actinobacteria and Acidobacteria by their capacity to degrade the anionic heteropolysaccharide, pectin, which has a characteristic galacturonic acid backbone which crosslinks cellulose and hemicellulose in the plant cell wall.[84]

Rhizobiales (α-), β-, and γ-proteobacteria express negligible levels of polysaccharide-depolymerizing enzymes, yet are heavily invested in the uptake of monosaccharides. Rhizobiales have the lowest fidelities for polysaccharide depolymerization and relatively low fidelities for glycolysis and the catabolism of diverse hexoses and pentoses. However, Rhizobiales dominate sugar transport functions, followed by β-proteobacteria, while the groups with high levels of polysaccharide degradation and sugar consumption have low transporter levels that are comparable to the relatively inactive actinobacterial groups, Thermoleophilia and Acidimicrobiia. There are multiple possible explanations for the seemingly counterintuitive patterns of sugar transporter expression. Cells excreting extracellular enzymes to degrade polysaccharides are likely in close proximity to the reaction product, so they may maintain a lower level of transporters that can capture most but not all of the diffusing product. In the case

of a biofilm growing directly on the substrate, the product is likely to quickly encounter the extracellular matrix, reducing the required density of outer membrane transporters. Rhizobiales sugar scavengers might live on the periphery of acidobacterial biofilms and invest in transporters rather than glycosidases.

Alternatively, Rhizobiales may import sugars generated by plant roots rather than saprotrophic bacteria. Certain groups of Proteobacteria and Actinobacteria are frequently found in the rhizosphere, or the soil immediately surrounding and strongly influenced by roots.[86,87] Bacteria dependent on root interactions have less of a need to produce enzymes for organic matter decomposition, but still require a standing stock of transporters to assimilate exudates. To prevent diffusive escape of organic exudates into porewaters, the density of transporters may need to be higher in bacteria on or near the root surface than in symbiotic bacteria encapsulated in legume root nodules. No plants in the study areas are known to form root nodules, although the main shrub species have ectomycorrhizae and some prostrate herbaceous plants associate with ericoid mycorrhizal fungi.[88] Multiple lines of evidence in the metaproteomic data substantiate plant interactions with Proteobacteria (Rhizobiales in particular) and Actinobacteria. First, these taxa have the highest overall expression of Nod factors, which initiate the formation of nodules in legumes and are also known to modulate interactions between mutualistic bacteria that do not form nodules and plants (Figures IV.23-IV.25).[89] Furthermore, an enzyme in the bacterial pathway for the biosynthesis of indole-3-acetic acid (IAA), a key plant hormone, was identified and found to be closely related to sequences in the Rhizobiales bin. Second, as discussed in Section IV.C.1, sugar transporters and other metaproteins linked to Proteobacteria have higher overall (NSAF) levels in the tussock than intertussock soils – soils with higher plant biomass. The rhizosphere has higher concentrations of bioavailable, soluble C sources than the

bulk soil.[90] Third, N_2 fixation and other modes of N acquisition are most strongly expressed by Rhizobiales followed by other Proteobacteria (Section IV.C.2.iii). The greater expression of proteins for N acquisition by these putative rhizospheric taxa has at least two potential explanations, including competition with plants for scarce N resources and a higher inherent demand for N as "copiotrophic" taxa.[91]

Rhizobiales, β-, and γ-proteobacteria most strongly express pathways required for the utilization of small soluble molecules and gases. Rhizobiales dominate CO catabolism, followed by β-proteobacteria, and Rhizobiales, β-proteobacteria, and Class Actinobacteria have the highest levels of enzymes required for CO_2 fixation. Methylotrophy pathways are most expressed by Burkholderiaceae followed by *Pseudomonas* and then a number of other groups. The non-detection of methane monooxygenase in conjunction with the near absence of Archaea, the prevalence of aerobic metabolisms, and the low level of the water table late in the growing season indicates that methane is not an important part of soil C cycling at the time of sampling, and also suggests that methylotrophs are consuming methanol.[92] The ACS and PTA-AckA pathways of acetate metabolism are highly expressed by Proteobacteria and Class Actinobacteria, and β-oxidation of fatty acids to acetyl-CoA (activated acetate) is dominated by β-proteobacteria, followed by Rhizobiales, γ-proteobacteria, and Class Actinobacteria. Rhizobiales have the highest production of outer membrane porins for small solutes (Class 2 outer membrane proteins), in agreement with their reliance on small substrates and disproportionate expression of transporters (Figure IV.17). In contrast, Rhizobiales exhibit very low production of other outer membrane proteins involved in the transport of larger compounds such as aromatics. In conjuction with the high levels of sugar transporters discussed before, Rhizobiales, β-, and γ-proteobacteria appear to catabolize soluble C sources to a greater extent

than Acidobacteria, Bacteroidetes, and Myxococcales, which invest more in enzymes required for the degradation of insoluble organic matter.

Aromatic ring cleavage pathways integral to lignin degradation are largely expressed by Burkholderiaceae, Class Actinobacteria, and Rhizobiales. Burkholderiaceae dominate the expression of pathways involved in the aerobic degradation of heterogeneous aromatic compounds derived from lignin, such as the protocatechuate-4,5 cleavage pathway. Burkholderiaceae and Rhizobiales most strongly express the homogentisate pathway, whereas Actinobacteria and Burkholderiaceae have the highest expression of the phenylacetate pathway. Both of these pathways degrade phenylalanine and an array of related aromatic compounds. Anaerobic pathways for the degradation of aromatic compounds have significantly lower overall expression levels than aerobic pathways and are expressed most by Myxococcales and Burkholderiaceae. Regarding other evidence of anaerobic activity in the soils, ethanol and lactate fermentation pathways are found at very low levels relative to the TCA cycle and are expressed the most by γ-proteobacteria, with lower bin fidelities in a range of other groups.

Finally, C storage molecules are important for organic C sequestration in soils and for buffering bacteria against perturbations in the environment. Rhizobiales followed by other Proteobacteria most strongly express pathways for the biosynthesis of polyhydroxyalkanoate compounds, including polyhydroxybutyrate granules, demonstrated to be critical for the survival of rhizobia through long periods of starvation.[93] Trehalose has a very similar taxonomic profile, whereas Acidobacteria expresses the highest level of glycogen biosynthetic pathways, followed by Proteobacteria beside Rhizobiales.

IV.C.2.iii. NUTRIENTS AND TRACE ELEMENTS

Bin fidelity data indicate that Proteobacteria, and particularly Rhizobiales, dominate N uptake (Figures IV.19-IV.22). The N-related Functional Group with the highest overall relative abundance is ammonia metabolism, which largely consists of the metaprotein, glutamine synthetase. This key enzyme in cellular metabolism incorporates ammonia into organic molecules. The ammonia can originate from amino acid catabolism, so high levels of this metaprotein reinforce its importance but do not indicate that ammonium is a significant source of N. Amino acid transporters are highly expressed, and ammonium transport, nitrate/nitrite reduction, and N_2 fixation are quite low in comparison – lower even than urea assimilation, oligopeptide transport, and polyamine transport. This could indicate that amino acids and other organic N sources are highly important in the tundra N cycle, consistent with previous measurements.[94] Proteobacteria also have the highest expression of polyamine biosynthesis pathways, which may be used for intracellular functions or to regulate plant activity.[95] Amino acid biosynthesis pathways, including the separately grouped shikimic acid pathway, are still highest in Proteobacteria but are more evenly expressed by other taxa as well.

There are at least two possible explanations for the dominant expression of N transporters by Proteobacteria and the relative evenness of glutamine synthetase and amino acid biosynthesis pathways. Rhizospheric Proteobacteria may be in direct competition for N with roots and therefore produce more transporters to increase the uptake of this limiting nutrient. The depletion of N in the rhizosphere occurs as root hairs grow through soil over the course of days, taking up N within 1-5 cm of the tip.[96] Competition rather than symbiosis seems likely, as N transfer between bacteria and plants is only known to occur via N_2 fixation,[97] and plants in the area of Toolik are not known to form root nodules, which are typically the site of symbiotic N_2 fixation.

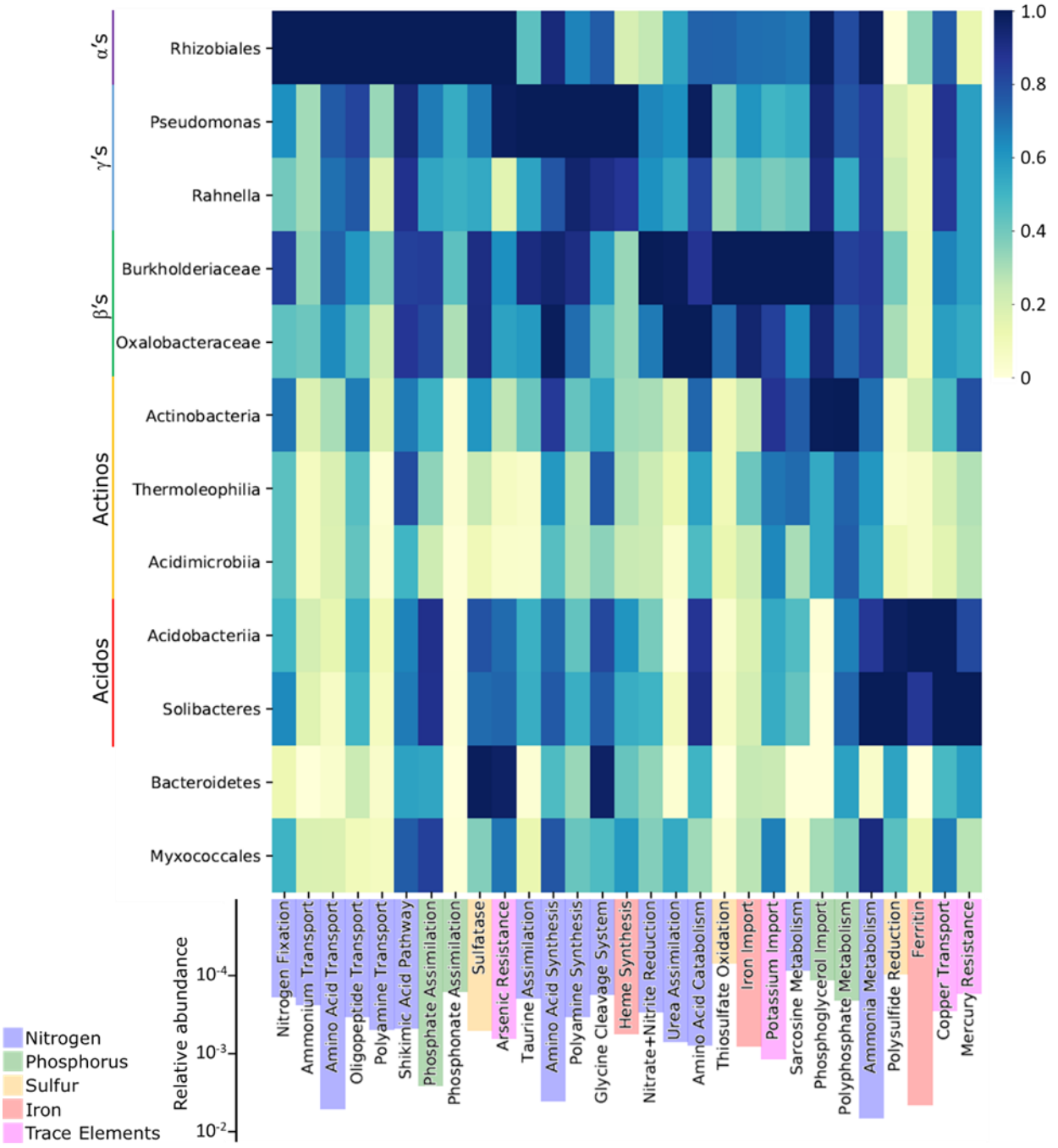

Figure IV.19. Nutrient-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples. Columns are ordered by the taxonomic bin with the maximum expression (darkest blue cell) and then by Functional Group category (bar color).

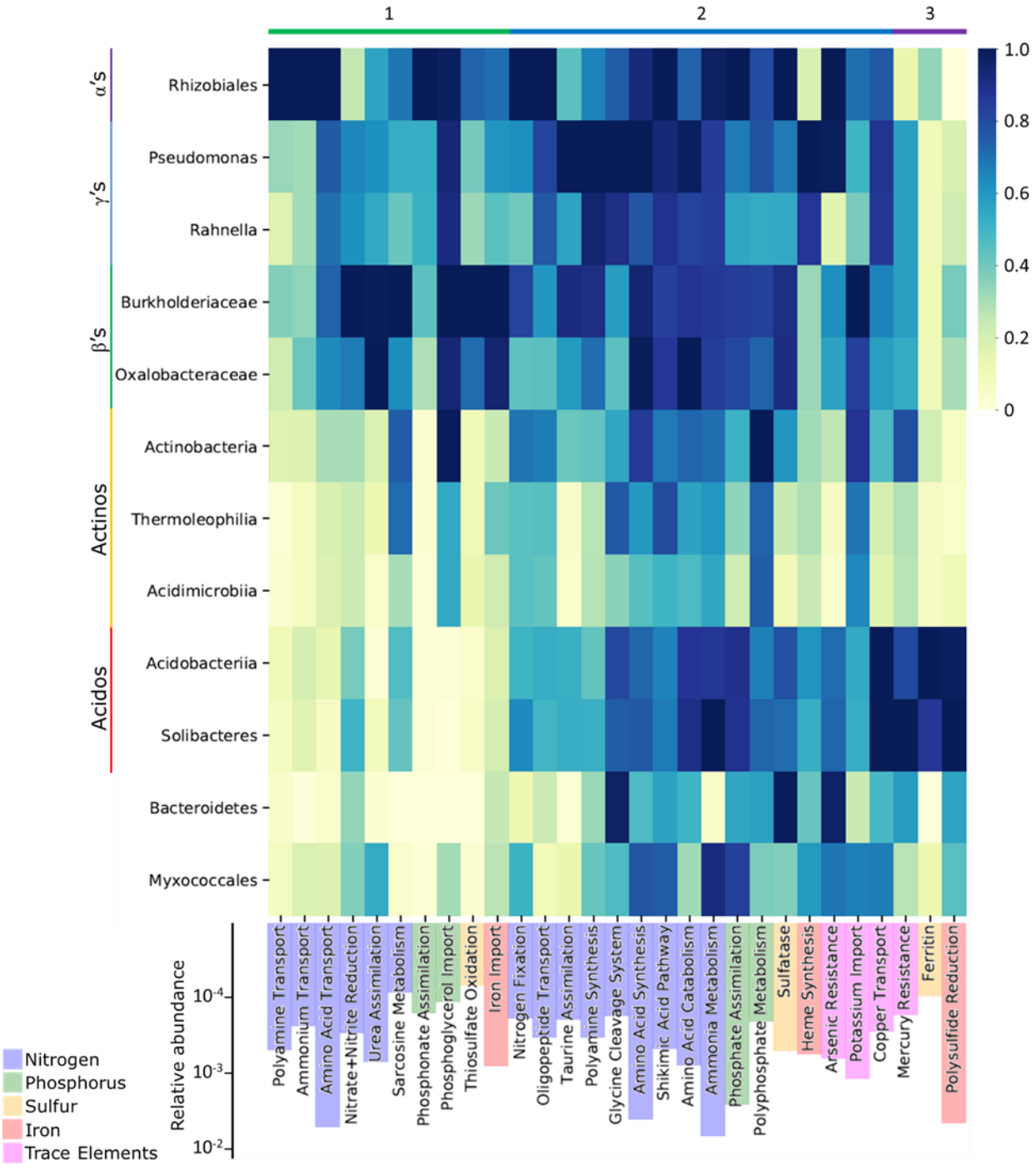

Figure IV.20. Nutrient-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples, with k-means cluster assignments at top. Columns are ordered by cluster assignment and then by Functional Group category (bar color).

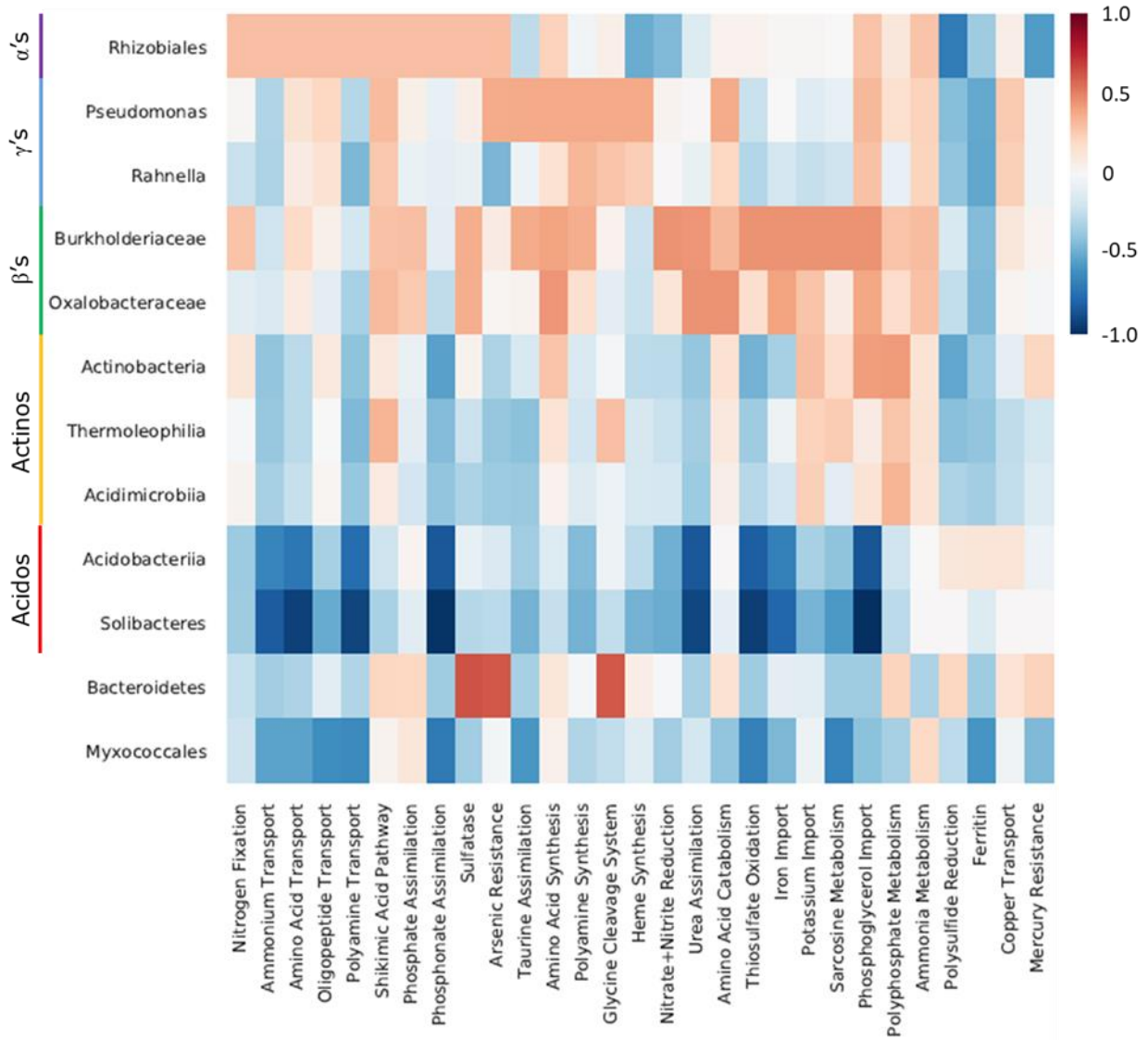

Figure IV.21. Nutrient-related Functional Group bin fidelities normalized to the maximum value in the column, averaged over all organic soil samples, with the Ribosome values then subtracted. This indicates the relative levels of Functional Group expression by taxa compared to a baseline of Ribosome expression. For instance, proteins involved in Nitrogen Fixation are disproportionately expressed by Rhizobiales, Burkholderiales, and Class Actinobacteria (values are positive and red) versus most other taxa (values are negative and blue) when compared to the Ribosome expression profile. Columns are in the same order as Figure IV.19.

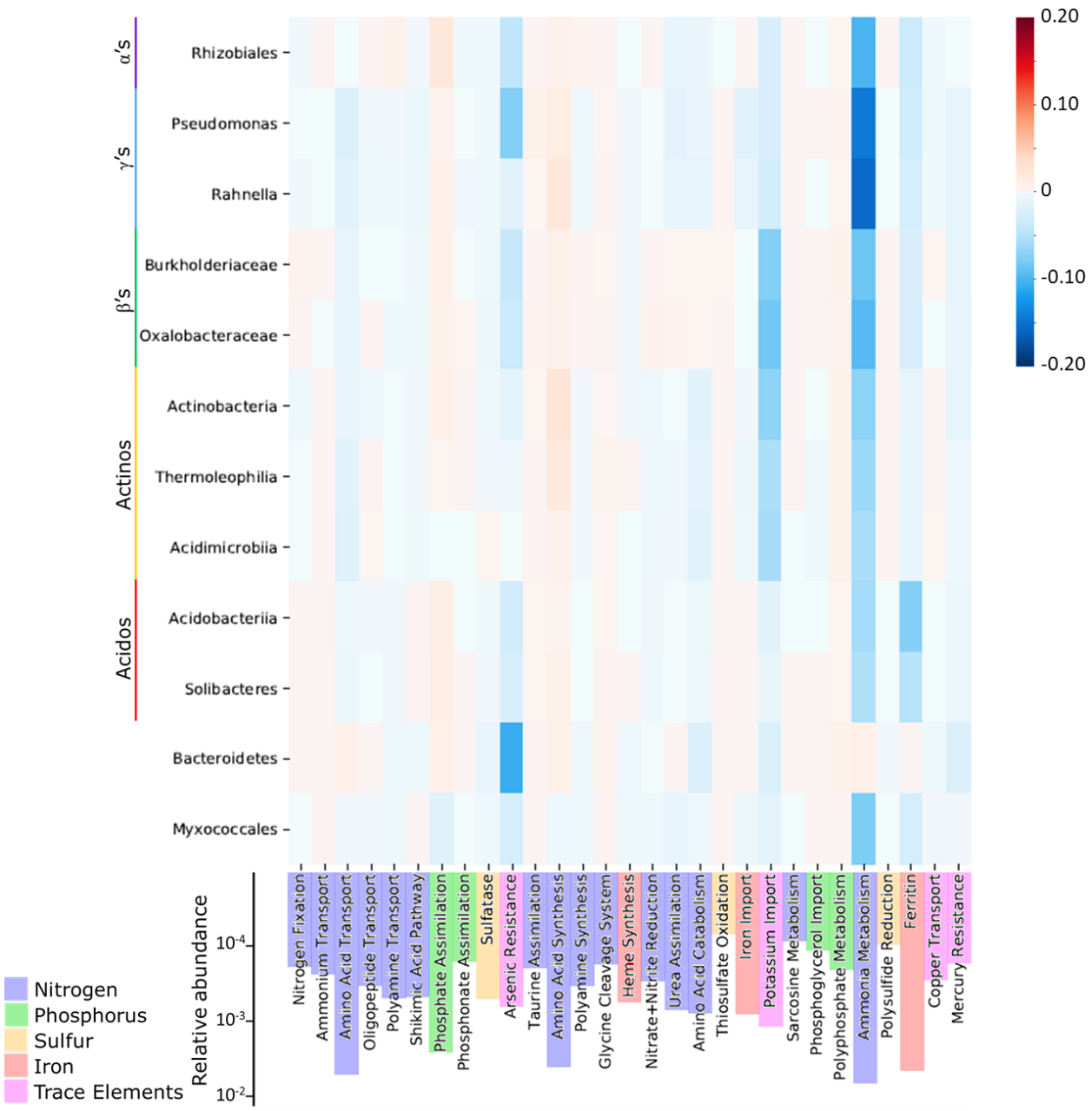

Figure IV.22. Difference in nutrient-related Functional Group functional fidelities between tussock/shrub (high plant biomass) and intertussock (low plant biomass) organic soil samples. The values represent the relative change in average functional fidelity from intertussock to tussock/shrub samples. Columns are in the same order as Figure IV.19.

Bacteria with N_2-fixing genetic potential in Alaskan taiga soils are mainly Rhizobiales and groups of Betaproteobacteria, including Burkholderiaceae,[98] which is concordant with the metaproteomic data showing that Rhizobiales and Burkholderiaceae have the highest bin fidelities for proteins involved in N_2 fixation. However, these proteins are found at a relatively low abundance, and cryptic mutualism via diazotrophy would not explain the high demand for other N sources in diazotrophic groups. Another line of evidence for N limitation in Rhizobiales is the dominance of polyhydroxyalkanoate production by this group, a type of C storage molecule produced under nutrient limited conditions.[99]

Alternatively, Proteobacteria in soils have been hypothesized to be boom-bust copiotrophs or ruderal, "weedy" taxa (in analogy to J.P. Grime's ecological model of plants), with a high inherent demand for limiting nutrients due to selection for fast growth. Copiotrophs or ruderal taxa contrast with slow-growing oligotrophs or stress tolerators, which have specific substrate requirements – the latter ecological categorization has been posited for Acidobacteria.[32,100] High levels of N transporters in Proteobacteria are consistent with this hypothesis, although other aspects of the bin fidelity profiles seem inconsistent with it. Copiotrophs are thought to immediately use substrates for biosynthesis or respiration, yet Rhizobiales dominate the expression of C storage compounds. The low level of ribosomes in Proteobacteria versus Acidobacteria, despite the roughly equal abundances of proteobacterial and acidobacterial groups in 16S rRNA gene libraries from Toolik soils,[81–83] also runs counter to the expectation that copiotrophs are poised for rapid growth. Therefore, competition with plants seems the more likely explanation for the profile of N uptake proteins in Proteobacteria.

Although N is often found to limit primary productivity in tundra fertilization studies, some soils are limited by P.[101] The phosphate assimilation Functional Group, which includes

phosphatases and phosphate transporters/porins, has relatively even expression across taxa. Polyphosphate metabolism and phosphonate assimilation have much lower overall relative abundances (NSAF) and bin fidelities skewed toward Rhizobiales. Microbial effects on rhizosphere P dynamics remain ambiguous, with some studies finding that plant-growth promoting bacteria increase P availability.[97]

Throughout the tundra, abundant rust-colored mats of oxidized Fe and iridescent Mn sheens coat plant stems and float on slow-moving waters, with bacteria catalyzing redox reactions at the oxic-anoxic interface.[102] Metal oxidation and reduction are not apparent in the metaproteomic datasets, although these metabolisms may manifest at other times of the year, such as spring thaw, when the water table is closer to the surface within the organic layer. Fe oxidizers and reducers in floating Fe mats from the Toolik area are largely from families of Betaproteobacteria and Deltaproteobacteria, respectively,[102] that are not represented highly enough among the metagenomic datasets to form identifiable bins. The Fe cycle in oxygenated tundra soils therefore primarily involves the assimilation of Fe into biomass rather than the use of Fe for energy conservation. γ-proteobacteria most strongly express the heme biosynthesis pathway, and β-proteobacteria followed by the other Proteobacteria have the highest levels of proteins involved in Fe import, with siderophores for Fe^{3+} acquisition being the most important. Acidobacteria dominate the expression of the Fe storage protein, ferritin, which is relatively abundant in these soils. This protein is ubiquitous across all domains of life, storing Fe^{3+} and releasing Fe^{2+}. Purposes of controlled Fe storage and release include protection against Fe overload, Fe deficiency, and oxidative stress.[103] I hypothesize that Acidobacteria may slowly accumulate Fe^{3+} in aerobic soils during the summer – given their relatively low expression of Fe

transporters – and anaerobically respire it during spring thaw or at other times when soil water content is high and O_2 is scarce.

IV.C.2.iv. CELL ENVELOPE AND MOVEMENT

The analysis of bin fidelity data elucidates in situ bacterial phenotypes, from C and nutrient preferences to the production and composition of extracellular polysaccharides and mechanisms of motility (Figures IV.23-IV.26). Reassuringly, bacteroidete and myxococcal gliding proteins have fidelities equal to 1 (the maximum) for the respective taxa, whereas other taxa have negligible fidelities about equal to 0. Additionally, Gram-positive Actinobacteria (with only a single membrane) have low bin fidelities for the production of outer membrane components but higher fidelities for phospholipids and cell wall components, traits shared by all bacteria. These data support the utility of the alignment bitscore as a measure of relatedness in the calculation of the fidelity metric.

Rhizobiales dominate the expression of proteins involved in succinoglycan production, and both Rhizobiales and γ-proteobacteria appear to be the major producers of alginate. Succinoglycan and alginate are known as components of biofilms formed on plants by nodulating rhizobia and pathogenic *Pseudomonas*, respectively.[104] Acidobacteria strongly express capsule production pathways, although production is relatively even across Gram-negative taxa, including Rhizobiales. The relatively high levels of EPS biosynthesis by Rhizobiales suggest that they are the predominant producers of biofilms in the soils, consistent with the potential existence of significant interactions between this group and plant roots. Biofilms provide benefits to both microbe and plant and play an instrumental role in the initiation of symbioses.[104]

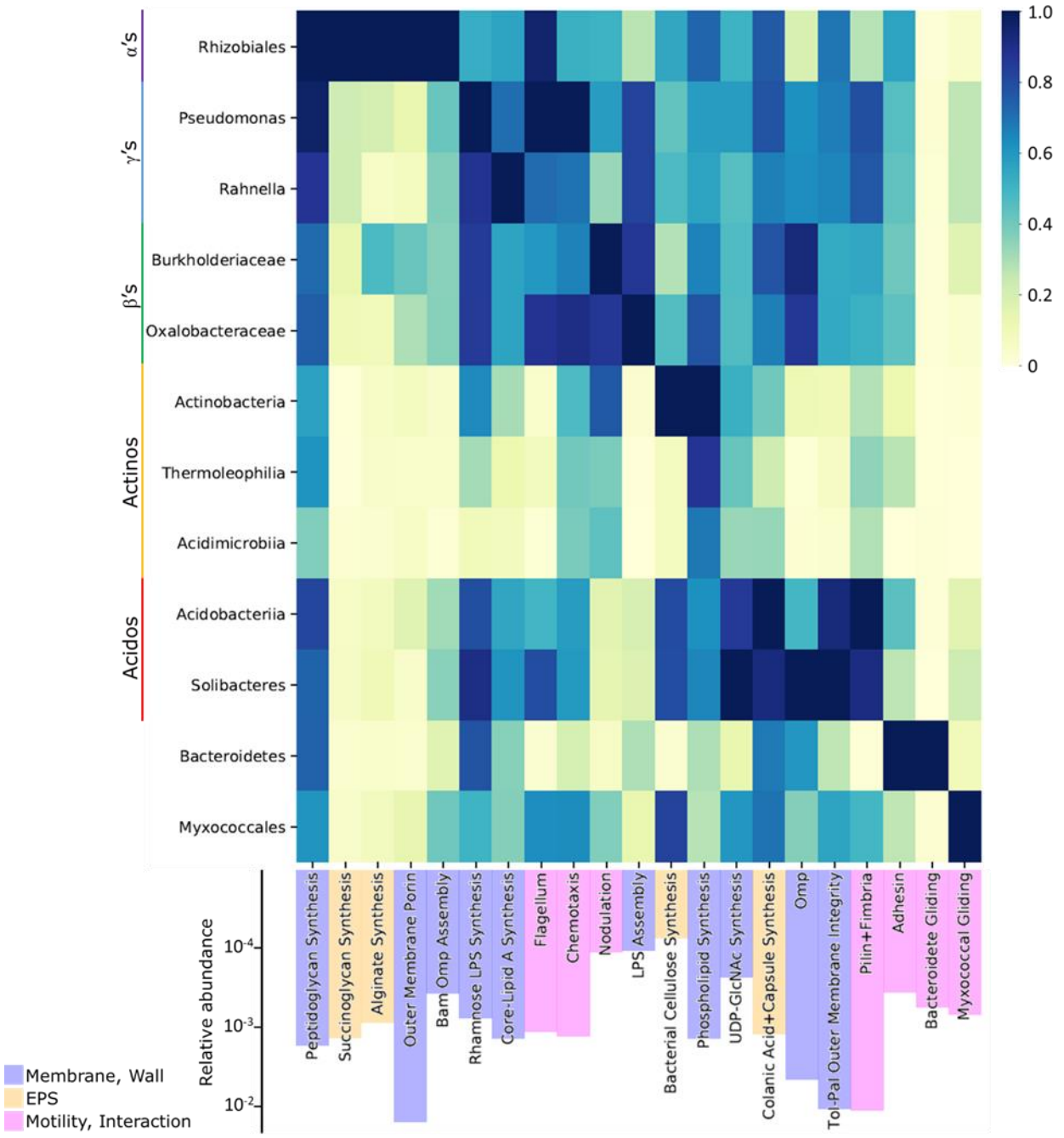

Figure IV.23. Cell envelope-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples. Columns are ordered by the taxonomic bin with the maximum expression (darkest blue cell) and then by Functional Group category (bar color).

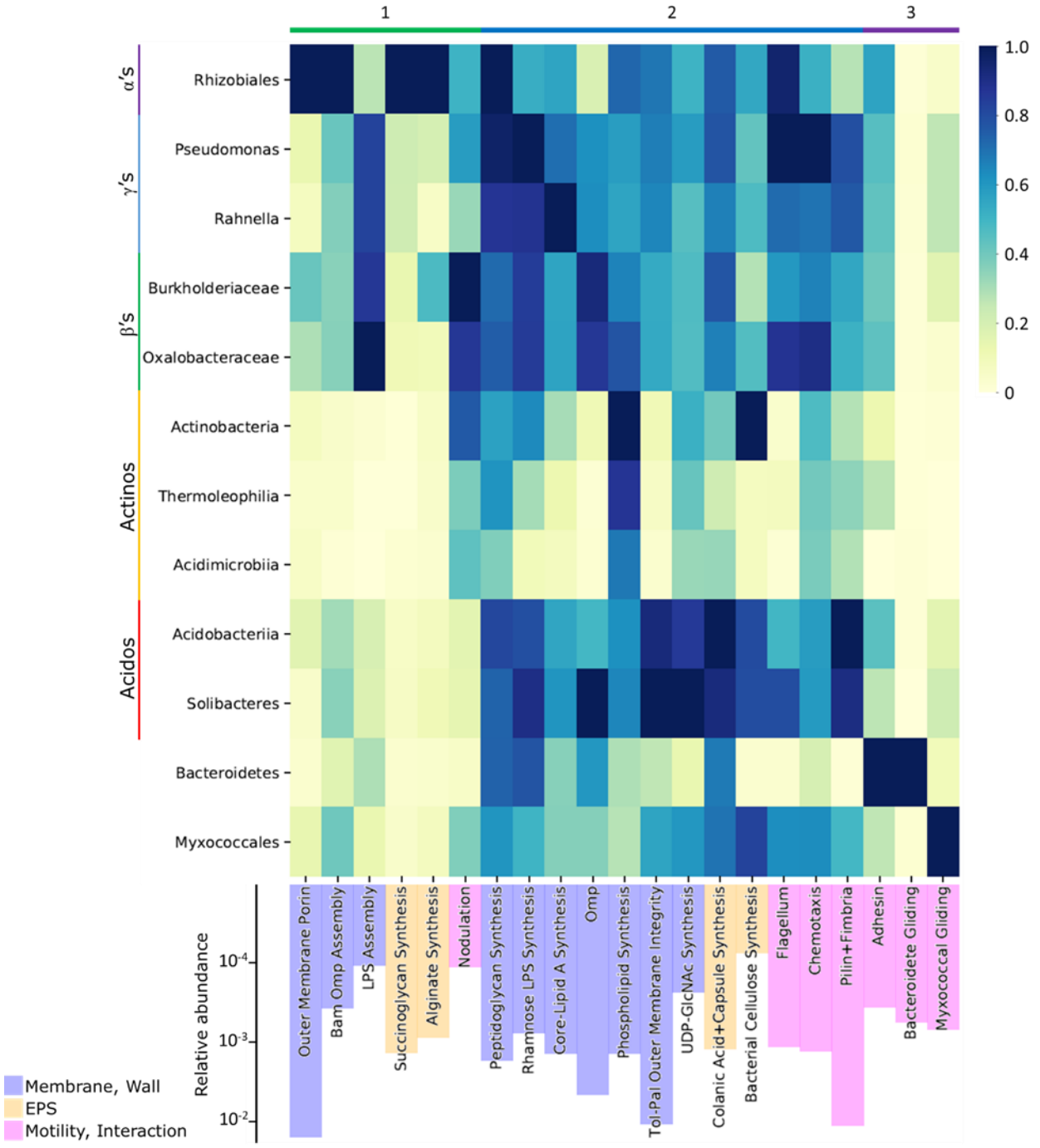

Figure IV.24. Cell envelope-related Functional Group bin fidelities normalized to the maximum value in the column (heatmap) and spectral relative abundances (NSAF; bars), averaged over all organic soil samples, with k-means cluster assignments at top. Columns are ordered by cluster assignment and then by Functional Group category (bar color).

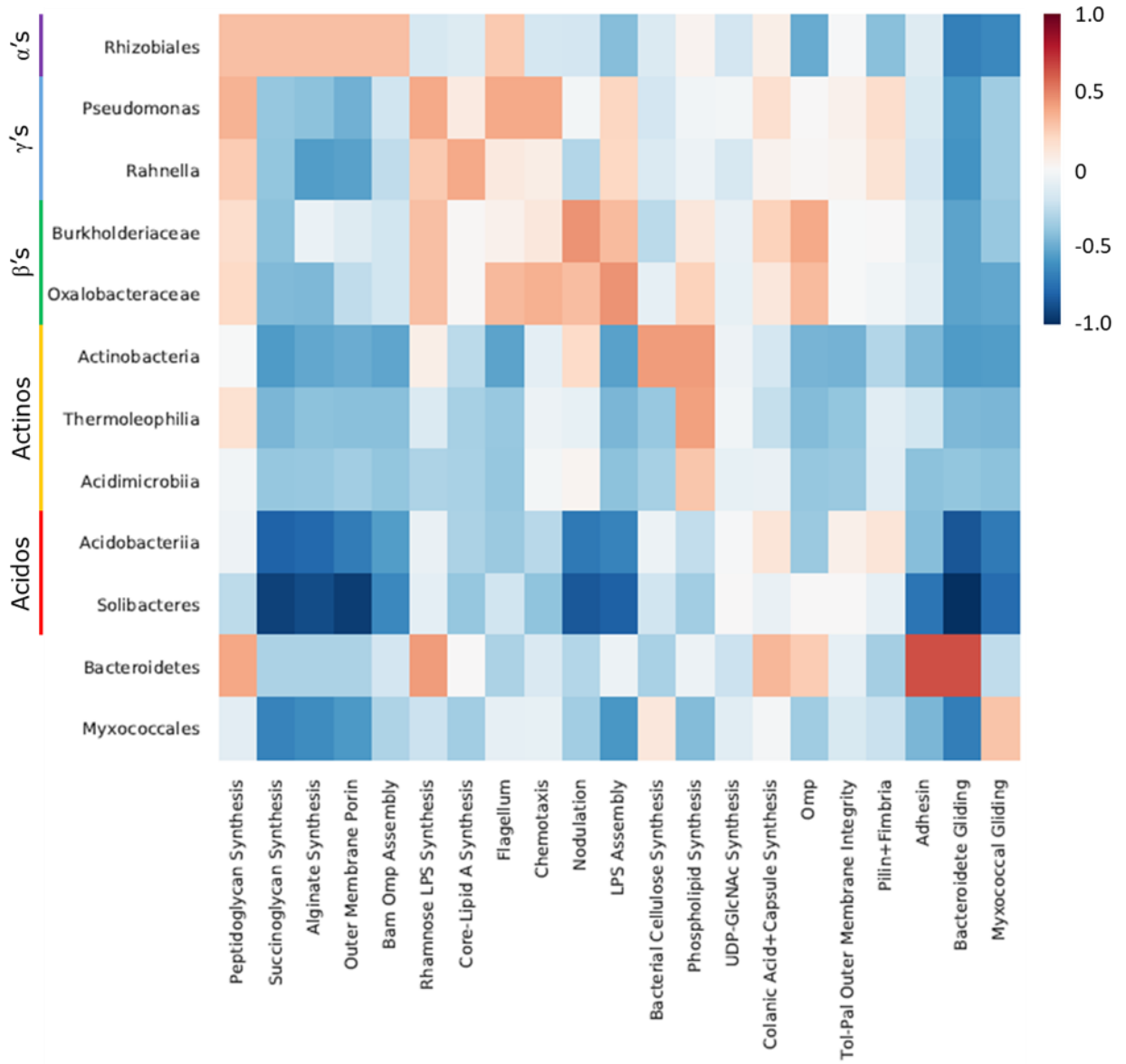

Figure IV.25. Cell envelope-related Functional Group bin fidelities normalized to the maximum value in the column, averaged over all organic soil samples, with the Ribosome values then subtracted. This indicates the relative levels of Functional Group expression by taxa compared to a baseline of Ribosome expression. For instance, the expression of Peptidoglycan Synthesis proteins is more even between taxa than Ribosome expression, so values are positive (red) for most taxa. Columns are in the same order as Figure IV.23.

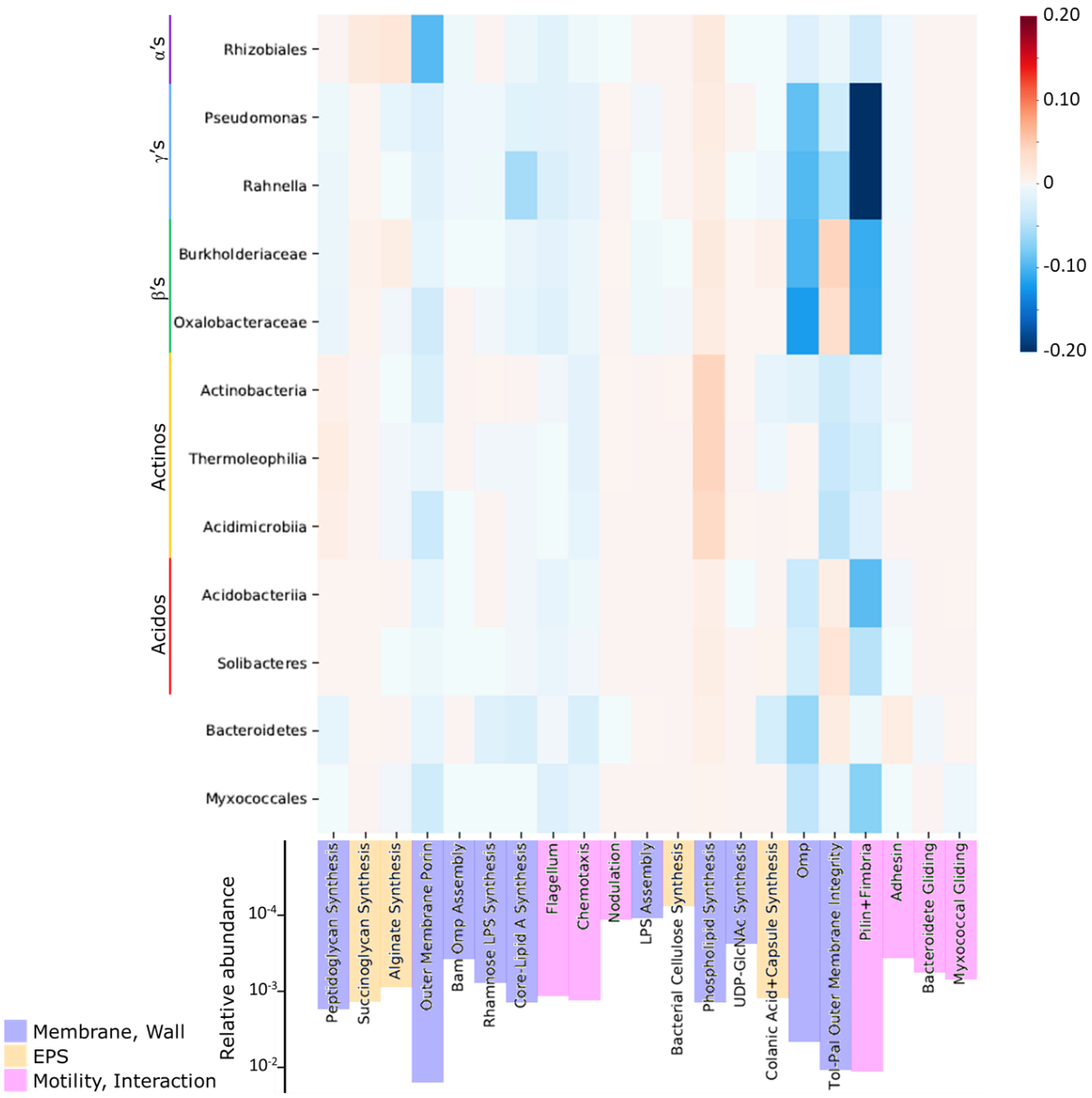

Figure IV.26. Difference in cell envelope-related Functional Group functional fidelities between tussock/shrub (high plant biomass) and intertussock (low plant biomass) organic soil samples. The values represent the relative change in average functional fidelity from intertussock to tussock/shrub samples. Columns are in the same order as Figure IV.23.

Certain Functional Groups related to cell structure have high relative abundances in the soils. Bacterial microcompartments are similar in many respects to eukaryotic organelles, serving to segregate specific cellular functions (e.g., CO_2 fixation in the carboxysome). Microcompartments form from a shell of proteins and are used to contain pathways that require high reactant concentrations or produce volatile metabolites that need to be sequestered from the rest of the cell. Acidobacteria dominate the expression of microcompartment proteins (Figure IV.11), and although their purpose is unclear, they are probably not carboxysomes, as Acidobacteria appear not to express proteins required for CO_2 fixation (Figure IV.15). Acidobacteria also dominate the expression of rod morphogenesis proteins. Cell shape relates to ecophysiology, as the ratio of cell surface area to volume affects the rate of substrate uptake, with putatively copiotrophic taxa more likely to grow in a spherical shape than putatively oligotrophic taxa such as Acidobacteria.[32]

Proteins involved in the structure and operation of flagella and pili/fimbriae have relatively even taxonomic expression patterns, except for the actinobacterial groups, which often form hyphae, and Bacteroidetes. Bacteroidetes and Myxococcales depend on gliding motility, with some level of pilin production possible in the Myxococcales as well. The Chemotaxis Functional Group is most highly expressed by *Pseudomonas*, yet has a relatively even expression profile across taxa, suggesting the ubiquity of environmentally regulated movement.

IV.D. DISCUSSION AND CONCLUSION

Microbial activity in soils controls the fluxes of vast quantities of C and other elements derived from plant detritus and minerals. The balance of photosynthesis and respiration stabilizes atmospheric CO_2 and soil C_{org} storage, but environmental perturbations have the capacity to

disrupt this equilibrium.[105] Rapid Arctic warming is increasing the activity of both plants and microbes, yet the mineralization of C_{org} stored in permafrost-affected soil (half the global stock of soil C_{org}) has the potential to greatly exceed any C gains in low stature Arctic vegetation.[106] Predicting the fate of Arctic soil C may depend on a greater understanding of the microbial processes controlling the transformations of myriad soil organic molecules that are difficult to measure in situ,[107] and how the soil microbiome interacts with the rapidly changing vegetation.[81] Direct investigation of the intra- and extracellular proteins catalyzing biogeochemical cycles has the potential to reveal key decomposition pathways and the largely uncharacterized ecophysiology of microbial taxa.[27,108]

I elucidated in situ microbial activity in Arctic soils using novel methods that I developed for the analysis of metaproteomic data. The novel software pipeline called ProteinExpress leverages large search databases of meta-genomic/transcriptomic reads and contigs and sequence assembly graphs[69] to boost peptide sequence and protein identifications. For comparison, a recent methodological study of organic prairie soils using a similar instrumental setup with higher mass resolution (1D LC-MS/MS on a Q-Exactive HF mass spectrometer), a reference database of short- and long-read hybrid metagenomic assemblies, and the same MSGF+ database search tool[68] resulted in an average of 29.4 spectra/PSM and 34.1 spectra/unique peptide.[109] In my organic soil datasets, I found average ratios of 11.4 spectra/PSM and 20.1 spectra/unique peptide, a higher yield of sequence information from the spectral data.

ProteinExpress retains the suite of protein-coding sequences that can be matched to a single spectrum and screens these results for high-confidence metaprotein functional assignments. GO, KEGG, and COG terms and eggNOG gene families and descriptions are generated by eggNOG-mapper[72] and supplemented by assignment to a database of "Functional

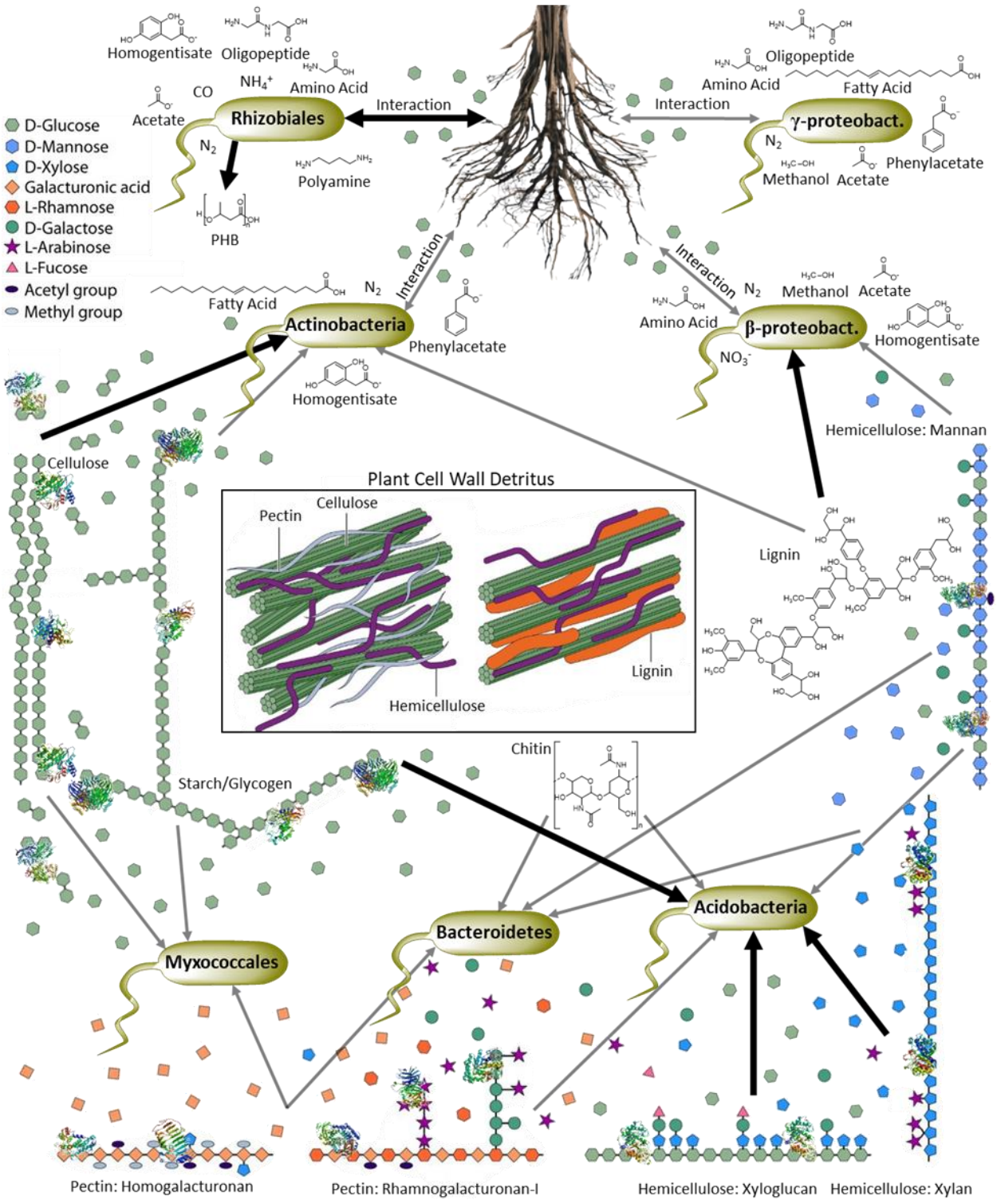

Figure IV.27. Summary of resource partitioning in moist acidic tundra soils

Groups" that I developed based on eggNOG gene families and descriptions. eggNOG-mapper produces high annotation coverage using those two systems, but not other systems such as KEGG, for the vast majority of queries. Functional Groups facilitate the interpretation of metaproteins involved in common cellular and biogeochemical processes, and I created a database of 141 Functional Groups such as "ATP Synthase," "Amino Acid Transport," and "Cellulases" from 2,659 eggNOG terms in the datasets (Appendix: Table VI.3).

A final methodological improvement introduced in ProteinExpress is a metric called fidelity that serves as a proxy for the relative expression of functions by taxa. Bin fidelity multiplies the overall abundance of a metaprotein (measured here by NSAF) by the scaled alignment score of the metaprotein sequence to bins of taxonomic reference sequences, representing relatedness of the expressed sequence to each taxon. From the 20 Alaskan soil metagenomes considered in this study,[56,57] I identified 12 bins of assembled sequences covering most of the abundant groups found in 16S rRNA gene surveys of Toolik-area soils.[81–83] Some bins correspond to genera (*Pseudomonas*, *Rahnella*), while others are rather more coarse (e.g., Bacteroidetes, Class Actinobacteria). Finer taxonomic resolution of the exceptionally microdiverse strains found in soils[67] can be achieved by longer sequence assemblies, which are possible using paired short- and long-read sequence data.[110]

This study revealed how the functional profile of the soil microbiome changes with the flora, with the lower biomass, nonvascular plant-dominated intertussock ecotype better adapted to cooler climates and the higher biomass tussock and shrub ecotypes better adapted to warmer climates.[24] Organic soils from the three ecotypes cleanly separate by metaprotein relative abundance, whereas mineral soils lie between the three organic clusters in decompositions of the multivariate dataset, suggesting that a set of shared microbial functions in the mineral soils is

expanded to process a greater diversity of substrates in the plant colonized organic soils. A major difference between environments with lower and higher plant biomass is the greater representation of proteins strongly associated with Rhizobiales, including sugar transporters and enzymes for the biosynthesis of succinoglycan EPS. The functional profile of Rhizobiales strongly suggests that they form biofilms around roots of the non-legumous flora – a possibility that could be investigated by fluorescent DNA tagging[111] – and consume small organic exudates such as simple sugars. Rhizobiales and other proteobacterial groups also appear to be more active in soils with higher biomass floras, judging by their greater investment in ribosomal proteins, for example.

Rhizobiales also dominate the expression of N transporters, although proteins involved in intracellular N usage, such as the highly abundant protein glutamine synthetase, are more evenly expressed across taxa. N is a common limiting nutrient in tundra soils;[101] the overall abundances of proteins involved in the N cycle suggest that organic rather than inorganic N is the major currency of N consumption in the microbiome. This is consistent with N addition studies in N-limited environments that demonstrate rapid immobilization of added organic and inorganic N in microbial biomass and the rapid turnover of the pool of free amino acids in soil porewaters (estimated at 20 day^{-1} in a study of permafrost-affected soils).[96,112] The likely localization of Rhizobiales to the rhizosphere suggests that this group competes for N with plant roots and therefore requires a high level of transporters to meet nutritional requirements. Although microbes outcompete plants in N addition experiments, plants accumulate N over time,[113] which is likely the reason why extractable soil N in Arctic soils falls from measurable concentrations after spring thaw to undetectable concentrations by the middle of the growing season.[94] The competition hypothesis seems more likely than an alternative that Rhizobiales are “copiotrophs”

with an inherently high demand for substrates and nutrients,[114] given that Rhizobiales also exhibit features of "oligotrophs," such as the high expression of C storage compounds (PHAs).

The same groups of bacteria are found in all of the floral ecotypes and their protein expression profiles change. However, the functional specialization of taxa does not change as much as the overall activity of taxa, with the same taxa largely dominating the same functions in each environment. Functional profiles of the taxa reveal a substantial amount about their ecophysiology and the biogeochemical transformations occurring in the soils. Regarding C biogeochemistry, the clearest divide exists between taxa specializing in small, soluble compounds and those specializing in the degradation of insoluble polymers. Rhizobiales and other Proteobacteria are in the former category, strongly overlapping with the cluster of functions involved in small molecule catabolism, while Acidobacteria are in the latter, strongly overlapping with the cluster of polysaccharide degradation functions. Acidobacteria are the most active group in all of the soils, as shown by their high expression of ribosomes, RNA polymerase, and other core proteins required for growth and replication, although Proteobacteria are more active in higher biomass floras. Specific components of soil organic matter appear to be degraded by distinct taxa, with Acidobacteria dominating the degradation of hemicelluloses, such as xylan, and other relatively labile polysaccharides. Actinobacteria specialize in cellulose, Myxococcales in pectin, and Burkholderiaceae in lignin. A very recent, deep multi-omics study of a permafrost thaw transect in Sweden provides a useful reference point for my Alaskan observations.[115] Metagenome-assembled genomes (MAGs) of Acidobacteria reconstructed from 214 metagenomic samples generally contained cellulases, β-glucosidases, and xylanases. The strong expression of these genes was confirmed by complementary metatranscriptomic and metaproteomic data (with peptide mass spectra searched against the 1,529 MAGs recovered from

the metagenomes). In contrast to my samples, Acidobacteria rather than Actinobacteria seemed to have the highest expression of cellulases despite many actinobacterial MAGs encoding cellulases and not xylanases. The unexpectedly low contribution of Actinobacteria to cellulose degradation also contrasts with metatranscriptomic samples from Svalbard peat soils.[116] The preference of Acidobacteria for hemicellulose but not cellulose that is observed in my datasets is supported by patterns of substrate utilization in the pure culture isolates of the group, with a majority able to grow on xylan and glucans, but only one isolate able to grow on crystalline cellulose and another able to grow on carboxymethylcellulose (these two isolates are unrelated to the Acidobacteria present in my Alaskan metagenomic reference datasets).[76]

This metaproteomic study reveals deep phylogenetic resource partitioning of organic compounds in soils, a result which runs counter to the theory from metagenomics and comparative genomics that these metabolic transformations are shallow traits scattered by horizontal gene transfer among strains from a variety of clades.[39] In a 2017 Nature review on the state of soil microbiome research, N. Fierer wrote, "[F]or many 'broad' processes – including the processes that drive soil carbon dynamics, or those that contribute to nitrogen mineralization and/or immobilization – it is far more difficult to link microbial community data to process rates. This is because there are numerous individual processes and taxa associated with the metabolism of the thousands of organic compounds found in soil. This complexity makes it very difficult to predict soil function. If, for example, we want to know the fate of labile carbon compounds in soil (which is important in soil carbon models), information about what taxa are present in a given soil sample is unlikely to be useful."[100] Again, I found that regardless of floral environment, Acidobacteria are the most active group, specializing in the degradation of relatively labile polysaccharides – a potentially widespread substrate preference supported by

recent cultivation and multi-omics work. This suggests that measurements of hemicellulose degradation rates at different temperatures by acidobacterial isolates related to those in tundra soils should be a high priority for the elaboration of terrestrial biogeochemical models. The more accurate representation of microbial physiology in a soil C cycle model has already been shown to improve predictions of decomposition rates.[117] Likewise, the higher expression of functions strongly linked to Rhizobiales in more vegetated soils may augment C sequestration through plant growth-promoting interactions, soil aggregate stabilization by EPS production, and the production of polyhydroxyalkanoates, key C storage molecules in soils.[99] Metaproteomic methods developed in ProteinExpress should next be applied to soils from a variety of environments with complementary third-generation metagenomic reference data, as well as to Arctic soils collected over a period of time.

IV.E. REFERENCES

(1) Hugelius, G.; Strauss, J.; Zubrzycki, S.; Harden, J. W.; Schuur, E. A. G.; Ping, C.-L.; Schirrmeister, L.; Grosse, G.; Michaelson, G. J.; Koven, C. D.; et al. Estimated Stocks of Circumpolar Permafrost Carbon with Quantified Uncertainty Ranges and Identified Data Gaps. *Biogeosciences* **2014**, *11* (23), 6573–6593.

(2) Hugelius, G.; Bockheim, J. G.; Camill, P.; Eberling, B.; Grosse, G.; Harden, J. W.; Johnson, K.; Jorgenson, T.; Koven, C.; Kuhry, P.; et al. A New Data Set for Estimating Organic Carbon Storage to 3 m Depth in Soils of the Northern Circumpolar Permafrost Region. *Earth System Science Data* **2013**, *5*, 393–402.

(3) Jobbágy, E. G.; Jackson, R. B. The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. *Ecological Applications* **2000**, *10* (2), 423–436.

(4) Zimov, S. A.; Davydov, S. P.; Zimova, G. M.; Davydova, A. I.; Schuur, E. A. G.; Dutta, K.; Chapin, F. S. Permafrost Carbon: Stock and Decomposability of a Globally Significant Carbon Pool. *Geophysical Research Letters* **2006**, *33* (20).

(5) Dutta, K.; Schuur, E. A. G.; Neff, J. C.; Zimov, S. A. Potential Carbon Release from Permafrost Soils of Northeastern Siberia. *Global Change Biology* **2006**, *12* (12), 2336–2351.

(6) Elberling, B.; Michelsen, A.; Schädel, C.; Schuur, E. A. G.; Christiansen, H. H.; Berg, L.; Tamstorf, M. P.; Sigsgaard, C. Long-Term CO_2 Production Following Permafrost Thaw. *Nature Climate Change* **2013**, *3* (10), 890–894.

(7) Schädel, C.; Schuur, E. A. G.; Bracho, R.; Elberling, B.; Knoblauch, C.; Lee, H.; Luo, Y.; Shaver, G. R.; Turetsky, M. R. Circumpolar Assessment of Permafrost C Quality and Its Vulnerability over Time Using Long-Term Incubation Data. *Global Change Biology* **2014**, *20* (2), 641–652.

(8) Schuur, E. A. G.; McGuire, A. D.; Schädel, C.; Grosse, G.; Harden, J. W.; Hayes, D. J.; Hugelius, G.; Koven, C. D.; Kuhry, P.; Lawrence, D. M.; et al. Climate Change and the Permafrost Carbon Feedback. *Nature* **2015**, *520* (7546), 171–179.

(9) Koven, C. D.; Ringeval, B.; Friedlingstein, P.; Ciais, P.; Cadule, P.; Khvorostyanov, D.; Krinner, G.; Tarnocai, C. Permafrost Carbon-Climate Feedbacks Accelerate Global Warming. *Proceedings of the National Academy of Science U. S. A.* **2011**, *108* (36), 14769–14774.

(10) DeConto, R. M.; Galeotti, S.; Pagani, M.; Tracy, D.; Schaefer, K.; Zhang, T.; Pollard, D.; Beerling, D. J. Past Extreme Warming Events Linked to Massive Carbon Release from Thawing Permafrost. *Nature* **2012**, *484* (7392), 87–91.

(11) Mauritz, M.; Bracho, R.; Celis, G.; Hutchings, J.; Natali, S. M.; Pegoraro, E.; Salmon, V. G.; Schädel, C.; Webb, E. E.; Schuur, E. A. G. Nonlinear CO2 Flux Response to 7 Years of Experimentally Induced Permafrost Thaw. *Global Change Biology* **2017**, *23* (9), 3646–3666.

(12) Anthony, K. W.; Deimling, T. S. von; Nitze, I.; Frolking, S.; Emond, A.; Daanen, R.; Anthony, P.; Lindgren, P.; Jones, B.; Grosse, G. 21st-Century Modeled Permafrost Carbon Emissions Accelerated by Abrupt Thaw beneath Lakes. *Nature Communications* **2018**, *9* (1), 3262.

(13) Turetsky, M. R.; Kane, E. S.; Harden, J. W.; Ottmar, R. D.; Manies, K. L.; Hoy, E.; Kasischke, E. S. Recent Acceleration of Biomass Burning and Carbon Losses in Alaskan Forests and Peatlands. *Nature Geoscience* **2011**, *4* (1), 27–31.

(14) Hartley, I. P.; Garnett, M. H.; Sommerkorn, M.; Hopkins, D. W.; Fletcher, B. J.; Sloan, V. L.; Phoenix, G. K.; Wookey, P. A. A Potential Loss of Carbon Associated with Greater Plant Growth in the European Arctic. *Nature Climate Change* **2012**, *2* (12), 875–879.

(15) Iversen, C. M.; Sloan, V. L.; Sullivan, P. F.; Euskirchen, E. S.; McGuire, A. D.; Norby, R. J.; Walker, A. P.; Warren, J. M.; Wullschleger, S. D. The Unseen Iceberg: Plant Roots in Arctic Tundra. *New Phytologist* **2015**, *205* (1), 34–58.

(16) Zhu, Q.; Iversen, C. M.; Riley, W. J.; Slette, I. J.; Stel, H. M. V. Root Traits Explain Observed Tundra Vegetation Nitrogen Uptake Patterns: Implications for Trait-Based Land Models. *Journal of Geophysical Research: Biogeosciences* **2016**, *121* (12), 3101–3112.

(17) Commane, R.; Lindaas, J.; Benmergui, J.; Luus, K. A.; Chang, R. Y.-W.; Daube, B. C.; Euskirchen, E. S.; Henderson, J. M.; Karion, A.; Miller, J. B.; et al. Carbon Dioxide Sources from Alaska Driven by Increasing Early Winter Respiration from Arctic Tundra. *Proceedings of the National Academy of Science U. S. A.* **2017**, 201618567.

(18) Epstein, H. E.; Raynolds, M. K.; Walker, D. A.; Bhatt, U. S.; Tucker, C. J.; Pinzon, J. E. Dynamics of Aboveground Phytomass of the Circumpolar Arctic Tundra during the Past Three Decades. *Environmental Research Letters* **2012**, *7* (1), 015506.

(19) Pan, Y.; Birdsey, R. A.; Fang, J.; Houghton, R.; Kauppi, P. E.; Kurz, W. A.; Phillips, O. L.; Shvidenko, A.; Lewis, S. L.; Canadell, J. G.; et al. A Large and Persistent Carbon Sink in the World's Forests. *Science* **2011**, 1201609.

(20) McGuire, A. D.; Lawrence, D. M.; Koven, C.; Clein, J. S.; Burke, E.; Chen, G.; Jafarov, E.; MacDougall, A. H.; Marchenko, S.; Nicolsky, D.; et al. Dependence of the Evolution of Carbon Dynamics in the Northern Permafrost Region on the Trajectory of Climate Change. *Proceedings of the National Academy of Science U. S. A.* **2018**, *115* (15), 3882–3887.

(21) Bjorkman, A. D.; Myers-Smith, I. H.; Elmendorf, S. C.; Normand, S.; Rüger, N.; Beck, P. S. A.; Blach-Overgaard, A.; Blok, D.; Cornelissen, J. H. C.; Forbes, B. C.; et al. Plant Functional Trait Change across a Warming Tundra Biome. *Nature* **2018**, *562* (7725), 57–62.

(22) Fraser, R. H.; Lantz, T. C.; Olthof, I.; Kokelj, S. V.; Sims, R. A. Warming-Induced Shrub Expansion and Lichen Decline in the Western Canadian Arctic. *Ecosystems* **2014**, *17* (7), 1151–1168.

(23) Myers-Smith, I. H.; Forbes, B. C.; Wilmking, M.; Hallinger, M.; Lantz, T.; Blok, D.; Tape, K. D.; Macias-Fauria, M.; Sass-Klaassen, U.; Lévesque, E.; et al. Shrub Expansion in Tundra Ecosystems: Dynamics, Impacts and Research Priorities. *Environmental Research Letters* **2011**, *6* (4), 045509.

(24) Sistla, S. A.; Moore, J. C.; Simpson, R. T.; Gough, L.; Shaver, G. R.; Schimel, J. P. Long-Term Warming Restructures Arctic Tundra without Changing Net Soil Carbon Storage. *Nature* **2013**, *497* (7451), 615–618.

(25) Hobbie, S. E.; Chapin, F. S. The Response of Tundra Plant Biomass, Aboveground Production, Nitrogen, and CO2 Flux to Experimental Warming. *Ecology* **1998**, *79* (5), 1526–1544.

(26) Rustad, L.; Campbell, J.; Marion, G.; Norby, R.; Mitchell, M.; Hartley, A.; Cornelissen, J.; Gurevitch, J.; GCTE-NEWS. A Meta-Analysis of the Response of Soil Respiration, Net Nitrogen Mineralization, and Aboveground Plant Growth to Experimental Ecosystem Warming. *Oecologia* **2001**, *126* (4), 543–562.

(27) Jansson, J. K.; Hofmockel, K. S. The Soil Microbiome—from Metagenomics to Metaphenomics. *Current Opinion in Microbiology* **2018**, *43*, 162–168.

(28) Bryson, S.; Li, Z.; Chavez, F.; Weber, P. K.; Pett-Ridge, J.; Hettich, R. L.; Pan, C.; Mayali, X.; Mueller, R. S. Phylogenetically Conserved Resource Partitioning in the Coastal Microbial Loop. *ISME Journal* **2017**, *11* (12), 2781–2792.

(29) Bryson, S.; Li, Z.; Pett-Ridge, J.; Hettich, R. L.; Mayali, X.; Pan, C.; Mueller, R. S. Proteomic Stable Isotope Probing Reveals Taxonomically Distinct Patterns in Amino Acid Assimilation by Coastal Marine Bacterioplankton. *mSystems* **2016**, *1* (2).

(30) Pepe-Ranney, C.; Campbell, A. N.; Koechli, C. N.; Berthrong, S.; Buckley, D. H. Unearthing the Ecology of Soil Microorganisms Using a High Resolution DNA-SIP Approach to Explore Cellulose and Xylose Metabolism in Soil. *Frontiers in Microbiology* **2016**, *7*.

(31) Kieft, B.; Li, Z.; Bryson, S.; Crump, B. C.; Hettich, R.; Pan, C.; Mayali, X.; Mueller, R. S. Microbial Community Structure–Function Relationships in Yaquina Bay Estuary Reveal Spatially Distinct Carbon and Nitrogen Cycling Capacities. *Frontiers in Microbiology* **2018**, *9*.

(32) Fierer, N.; Bradford, M. A.; Jackson, R. B. Toward an Ecological Classification of Soil Bacteria. *Ecology* **2007**, *88* (6), 1354–1364.

(33) Placella, S. A.; Brodie, E. L.; Firestone, M. K. Rainfall-Induced Carbon Dioxide Pulses Result from Sequential Resuscitation of Phylogenetically Clustered Microbial Groups. *Proceedings of the National Academy of Science U. S. A.* **2012**, *109* (27), 10931–10936.

(34) Delgado-Baquerizo, M.; Oliverio, A. M.; Brewer, T. E.; Benavent-González, A.; Eldridge, D. J.; Bardgett, R. D.; Maestre, F. T.; Singh, B. K.; Fierer, N. A Global Atlas of the Dominant Bacteria Found in Soil. *Science* **2018**, *359* (6373), 320–325.

(35) Berlemont, R.; Martiny, A. C. Genomic Potential for Polysaccharide Deconstruction in Bacteria. *Applied and Environmental Microbiology* **2015**, *81* (4), 1513–1519.

(36) Berlemont, R.; Martiny, A. C. Glycoside Hydrolases across Environmental Microbial Communities. *PLoS Computational Biology* **2016**, *12* (12), e1005300.

(37) Zimmerman, A. E.; Martiny, A. C.; Allison, S. D. Microdiversity of Extracellular Enzyme Genes among Sequenced Prokaryotic Genomes. *ISME Journal* **2013**, *7* (6), 1187–1199.

(38) Martiny, A. C.; Treseder, K.; Pusch, G. Phylogenetic Conservatism of Functional Traits in Microorganisms. *ISME Journal* **2013**, *7* (4), 830–838.

(39) Martiny, J. B. H.; Jones, S. E.; Lennon, J. T.; Martiny, A. C. Microbiomes in Light of Traits: A Phylogenetic Perspective. *Science* **2015**, *350* (6261), aac9323.

(40) Doroghazi, J. R.; Buckley, D. H. Widespread Homologous Recombination within and between *Streptomyces* Species. *ISME Journal* **2010**, *4* (9), 1136–1143.

(41) Baran, R.; Brodie, E. L.; Mayberry-Lewis, J.; Hummel, E.; da Rocha, U. N.; Chakraborty, R.; Bowen, B. P.; Karaoz, U.; Cadillo-Quiroz, H.; Garcia-Pichel, F.; et al. Exometabolite Niche Partitioning among Sympatric Soil Bacteria. *Nature Communications* **2015**, *6*, 8289.

(42) Zhalnina, K.; Louie, K. B.; Hao, Z.; Mansoori, N.; da Rocha, U. N.; Shi, S.; Cho, H.; Karaoz, U.; Loqué, D.; Bowen, B. P.; et al. Dynamic Root Exudate Chemistry and Microbial Substrate Preferences Drive Patterns in Rhizosphere Microbial Community Assembly. *Nature Microbiology* **2018**, *3* (4), 470–480.

(43) Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and Perspectives of Metaproteomic Data Analysis. *Journal of Biotechnology* **2017**, *261* (Supplement C), 24–36.

(44) *Alaska's Changing Arctic: Ecological Consequences for Tundra, Streams, and Lakes*; Hobbie, J., Kling, G., Eds.; Long-Term Ecological Network Series; Oxford University Press, 2014.

(45) Walker, M. D.; Walker, D. A.; Auerbach, N. A. Plant Communities of a Tussock Tundra Landscape in the Brooks Range Foothills, Alaska. *Journal of Vegetation Science* **1994**, *5* (6), 843–866.

(46) Walker, D. A.; Raynolds, M. K.; Daniëls, F. J. A.; Einarsson, E.; Elvebakk, A.; Gould, W. A.; Katenin, A. E.; Kholod, S. S.; Markon, C. J.; Melnikov, E. S.; et al. The Circumpolar Arctic Vegetation Map. *Journal of Vegetation Science* **2005**, *16* (3), 267–282.

(47) Shaver, G. R.; Cutler, J. C. The Vertical Distribution of Live Vascular Phytomass in Cottongrass Tussock Tundra. *Arctic and Alpine Research* **1979**, *11* (3), 335–342.

(48) Shaver, G. R.; Fetcher, N.; Chapin, F. S. Growth and Flowering in Eriophorum Vaginatum: Annual and Latitudinal Variation. *Ecology* **1986**, *67* (6), 1524–1535.

(49) Wang, P.; Mommer, L.; van Ruijven, J.; Berendse, F.; Maximov, T. C.; Heijmans, M. M. P. D. Seasonal Changes and Vertical Distribution of Root Standing Biomass of Graminoids and Shrubs at a Siberian Tundra Site. *Plant and Soil* **2016**, *407* (1), 55–65.

(50) Hamilton, T. D. Late Cenozoic Glaciation of the Central Brooks Range. In *Glaciation in Alaska: The Geologic Record*; Hamilton, T. D., Reed, K. M., Thorson, R. M., Eds.; Alaska Geological Society: Anchorage, Alaska, 1986; pp 9–49.

(51) Walker, D. A.; Hamilton, T. D.; Maier, H. A.; Munger, C. A.; Raynolds, M. K. Glacial History and Long-Term Ecology in the Toolik Lake Region. In *Alaska's Changing Arctic: Ecological Consequences for Tundra, Streams, and Lakes*; Long-Term Ecological Network Series; Oxford University Press, 2014.

(52) Mercado-Díaz, J. A.; Gould, W. A.; González, G. Soil Nutrients, Landscape Age, and *Sphagno-Eriophoretum Vaginati* Plant Communities in Arctic Moist-Acidic Tundra Landscapes. *Open Journal of Soil Science* **2014**, *04*, 375.

(53) Bockheim, J. G.; Walker, D. A.; Everett, L. R.; Nelson, F. E.; Shiklomanov, N. I. Soils and Cryoturbation in Moist Nonacidic and Acidic Tundra in the Kuparuk River Basin, Arctic Alaska, U.S.A. *Arctic and Alpine Research* **1998**, *30* (2), 166–174.

(54) Chourey, K.; Jansson, J.; VerBerkmoes, N.; Shah, M.; Chavarria, K. L.; Tom, L. M.; Brodie, E. L.; Hettich, R. L. Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. *Journal of Proteome Research* **2010**, *9* (12), 6615–6622.

(55) Erde, J.; Loo, R. R. O.; Loo, J. A. Enhanced FASP (EFASP) to Increase Proteome Coverage and Sample Recovery for Quantitative Proteomic Experiments. *Journal of Proteome Research* **2014**, *13* (4), 1885–1895.

(56) Johnston, E. R.; Rodriguez-R, L. M.; Luo, C.; Yuan, M. M.; Wu, L.; He, Z.; Schuur, E. A. G.; Luo, Y.; Tiedje, J. M.; Zhou, J.; et al. Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Frontiers in Microbiology* **2016**, *7*.

(57) Ward, C. P.; Nalven, S. G.; Crump, B. C.; Kling, G. W.; Cory, R. M. Photochemical Alteration of Organic Carbon Draining Permafrost Soils Shifts Microbial Metabolic Pathways and Stimulates Respiration. *Nature Communications* **2017**, *8* (1), 772.

(58) Cox, M. P.; Peterson, D. A.; Biggs, P. J. SolexaQA: At-a-Glance Quality Assessment of Illumina Second-Generation Sequencing Data. *BMC Bioinformatics* **2010**, *11* (1), 485.

(59) Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **2010**, *11*, 119.

(60) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28* (23), 3150–3152.

(61) Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph. *Bioinformatics* **2015**, *31* (10), 1674–1676.

(62) Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. LBNL Report #: LBNL-7065E **2014**.

(63) Kang, D. D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities. *PeerJ* **2015**, *3*, e1165.

(64) Buchfink, B.; Xie, C.; Huson, D. H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nature Methods* **2015**, *12* (1), 59–60.

(65) Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data. *PeerJ* **2015**, *3*, e1319.

(66) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Research* **2015**, *25* (7), 1043–1055.

(67) Delmont, T. O.; Eren, A. M.; Maccario, L.; Prestat, E.; Esen, Ö. C.; Pelletier, E.; Le Paslier, D.; Simonet, P.; Vogel, T. M. Reconstructing Rare Soil Microbial Genomes Using in Situ Enrichments and Metagenomics. *Frontiers in Microbiology* **2015**, *6*.

(68) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nature Communications* **2014**, *5*, 5277.

(69) Tang, H.; Li, S.; Ye, Y. A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. *PLoS Computational Biology* **2016**, *12* (12), e1005224.

(70) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **1990**, *215* (3), 403–410.

(71) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Research* **2016**, *44* (D1), D733-745.

(72) Huerta-Cepas, J.; Forslund, K.; Coelho, L. P.; Szklarczyk, D.; Jensen, L. J.; von Mering, C.; Bork, P. Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper. *Molecular Biology and Evolution* **2017**, *34* (8), 2115–2122.

(73) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehtevä, M.; Reichl, U.; Martens, L.; Rapp, E. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *Journal of Proteome Research* **2015**, *14* (3), 1557–1565.

(74) Paoletti, A. C.; Parmely, T. J.; Tomomori-Sato, C.; Sato, S.; Zhu, D.; Conaway, R. C.; Conaway, J. W.; Florens, L.; Washburn, M. P. Quantitative Proteomic Analysis of Distinct Mammalian Mediator Complexes Using Normalized Spectral Abundance Factors. *Proceedings of the National Academy of Science U. S. A.* **2006**, *103* (50), 18928–18933.

(75) Eichorst, S. A.; Trojan, D.; Roux, S.; Herbold, C.; Rattei, T.; Woebken, D. Genomic Insights into the Acidobacteria Reveal Strategies for Their Success in Terrestrial Environments. *Environmental Microbiology* **2018**, *20* (3), 1041–1063.

(76) Kielak, A. M.; Barreto, C. C.; Kowalchuk, G. A.; Veen, V.; A, J.; Kuramae, E. E. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Frontiers in Microbiology* **2016**, *7*.

(77) Pascual, J.; Wüst, P. K.; Geppert, A.; Foesel, B. U.; Huber, K. J.; Overmann, J. Novel Isolates Double the Number of Chemotrophic Species and Allow the First Description of Higher Taxa in Acidobacteria Subdivision 4. *Systematic and Applied Microbiology* **2015**, *38* (8), 534–544.

(78) Rawat, S. R.; Männistö, M. K.; Bromberg, Y.; Häggblom, M. M. Comparative Genomic and Physiological Analysis Provides Insights into the Role of Acidobacteria in Organic Carbon Utilization in Arctic Tundra Soils. *FEMS Microbiology Ecology* **2012**, *82* (2), 341–355.

(79) Ward, N. L.; Challacombe, J. F.; Janssen, P. H.; Henrissat, B.; Coutinho, P. M.; Wu, M.; Xie, G.; Haft, D. H.; Sait, M.; Badger, J.; et al. Three Genomes from the Phylum Acidobacteria Provide Insight into the Lifestyles of These Microorganisms in Soils. *Applied and Environmental Microbiology* **2009**, *75* (7), 2046–2056.

(80) Pankratov, T. A.; Serkebaeva, Y. M.; Kulichevskaya, I. S.; Liesack, W.; Dedysh, S. N. Substrate-Induced Growth and Isolation of Acidobacteria from Acidic *Sphagnum* Peat. *ISME Journal* **2008**, *2* (5), 551–560.

(81) Wallenstein, M. D.; McMahon, S.; Schimel, J. Bacterial and Fungal Community Structure in Arctic Tundra Tussock and Shrub Soils. *FEMS Microbiology Ecology* **2007**, *59* (2), 428–435.

(82) Deslippe, J. R.; Hartmann, M.; Simard, S. W.; Mohn, W. W. Long-Term Warming Alters the Composition of Arctic Soil Microbial Communities. *FEMS Microbiology Ecology* **2012**, *82* (2), 303–315.

(83) Koyama, A.; Wallenstein, M. D.; Simpson, R. T.; Moore, J. C. Soil Bacterial Community Composition Altered by Increased Nutrient Availability in Arctic Tundra Soils. *Frontiers in Microbiology* **2014**, *5*.

(84) Sjostrom, E. *Wood Chemistry: Fundamentals and Applications*; Gulf Professional Publishing, 1993.

(85) Harris, P. J.; Stone, B. A. Chemistry and Molecular Organization of Plant Cell Walls. In *Biomass Recalcitrance*; Wiley-Blackwell, 2009; pp 61–93.

(86) Yeoh, Y. K.; Dennis, P. G.; Paungfoo-Lonhienne, C.; Weber, L.; Brackin, R.; Ragan, M. A.; Schmidt, S.; Hugenholtz, P. Evolutionary Conservation of a Core Root Microbiome across Plant Phyla along a Tropical Soil Chronosequence. *Nature Communications* **2017**, *8* (1), 215.

(87) Garrido-Oter, R.; Nakano, R. T.; Dombrowski, N.; Ma, K.-W.; McHardy, A. C.; Schulze-Lefert, P. Modular Traits of the Rhizobiales Root Microbiota and Their Evolutionary Relationship with Symbiotic Rhizobia. *Cell Host and Microbe* **2018**, *24* (1), 155-167.e5.

(88) Gardes, M.; Dahlberg, A. Mycorrhizal Diversity in Arctic and Alpine Tundra: An Open Question. *New Phytologist* **1996**, *133* (1), 147–157.

(89) Steenhoudt, O.; Vanderleyden, J. Azospirillum, a Free-Living Nitrogen-Fixing Bacterium Closely Associated with Grasses: Genetic, Biochemical and Ecological Aspects. *FEMS Microbiology Reviews* **2000**, *24* (4), 487–506.

(90) Jones, D. L.; Nguyen, C.; Finlay, R. D. Carbon Flow in the Rhizosphere: Carbon Trading at the Soil–Root Interface. *Plant and Soil* **2009**, *321* (1), 5–33.

(91) Hinsinger, P.; Bengough, A. G.; Vetterlein, D.; Young, I. M. Rhizosphere: Biophysics, Biogeochemistry and Ecological Relevance. *Plant and Soil* **2009**, *321* (1), 117–152.

(92) Christensen, T. R. Methane Emission from Arctic Tundra. *Biogeochemistry* **1993**, *21* (2), 117–139.

(93) Ratcliff, W. C.; Kadam, S. V.; Denison, R. F. Poly-3-Hydroxybutyrate (PHB) Supports Survival and Reproduction in Starving Rhizobia. *FEMS Microbiology Ecology* **2008**, *65* (3), 391–399.

(94) Weintraub, M. N.; Schimel, J. P. The Seasonal Dynamics of Amino Acids and Other Nutrients in Alaskan Arctic Tundra Soils. *Biogeochemistry* **2005**, *73* (2), 359–380.

(95) Fujihara, S. Biogenic Amines in Rhizobia and Legume Root Nodules. *Microbes and Environments* **2009**, *24* (1), 1–13.

(96) Kuzyakov, Y.; Xu, X. Competition between Roots and Microorganisms for Nitrogen: Mechanisms and Ecological Relevance. *New Phytologist* **2013**, *198* (3), 656–669.

(97) Richardson, A. E.; Barea, J.-M.; McNeill, A. M.; Prigent-Combaret, C. Acquisition of Phosphorus and Nitrogen in the Rhizosphere and Plant Growth Promotion by Microorganisms. *Plant and Soil* **2009**, *321* (1–2), 305–339.

(98) Penton, C. R.; Yang, C.; Wu, L.; Wang, Q.; Zhang, J.; Liu, F.; Qin, Y.; Deng, Y.; Hemme, C. L.; Zheng, T.; et al. NifH-Harboring Bacterial Community Composition across an Alaskan Permafrost Thaw Gradient. *Frontiers in Microbiology* **2016**, *7*.

(99) Anderson, A. J.; Dawes, E. A. Occurrence, Metabolism, Metabolic Role, and Industrial Uses of Bacterial Polyhydroxyalkanoates. *Microbiology Reviews* **1990**, *54* (4), 450–472.

(100) Fierer, N. Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome. *Nature Reviews Microbiology* **2017**, *15* (10), 579–590.

(101) Shaver, G. R.; Chapin, F. S. Long–Term Responses to Factorial, NPK Fertilizer Treatment by Alaskan Wet and Moist Tundra Sedge Species. *Ecography* **1995**, *18* (3), 259–275.

(102) Emerson, D.; Scott, J. J.; Benes, J.; Bowden, W. B. Microbial Iron Oxidation in the Arctic Tundra and Its Implications for Biogeochemical Cycling. *Applied and Environmental Microbiology* **2015**, *81* (23), 8066–8075.

(103) Smith, J. L. The Physiological Role of Ferritin-Like Compounds in Bacteria. *Critical Reviews in Microbiology* **2004**, *30* (3), 173–185.

(104) Danhorn, T.; Fuqua, C. Biofilm Formation by Plant-Associated Bacteria. *Annual Review of Microbiology* **2007**, *61* (1), 401–422.

(105) Raich, J. W.; Schlesinger, W. H. The Global Carbon Dioxide Flux in Soil Respiration and Its Relationship to Vegetation and Climate. *Tellus B* **1992**, *44* (2), 81–99.

(106) Schuur, E. A. G.; Abbott, B. W.; Bowden, W. B.; Brovkin, V.; Camill, P.; Canadell, J. G.; Chanton, J. P.; Chapin, F. S.; Christensen, T. R.; Ciais, P.; et al. Expert Assessment of Vulnerability of Permafrost Carbon to Climate Change. *Climatic Change* **2013**, *119* (2), 359–374.

(107) Luo, Y.; Ahlström, A.; Allison, S. D.; Batjes, N. H.; Brovkin, V.; Carvalhais, N.; Chappell, A.; Ciais, P.; Davidson, E. A.; Finzi, A.; et al. Toward More Realistic Projections of Soil Carbon Dynamics by Earth System Models. *Global Biogeochemical Cycles* **2016**, *30* (1), 40–56.

(108) Keiblinger, K. M.; Fuchs, S.; Zechmeister-Boltenstern, S.; Riedel, K. Soil and Leaf Litter Metaproteomics – A Brief Guideline from Sampling to Understanding. *FEMS Microbiology Ecology* **2016**, *92* (11).

(109) Callister, S. J.; Fillmore, T. L.; Nicora, C. D.; Shaw, J. B.; Purvine, S. O.; Orton, D. J.; White, R. A.; Moore, R. J.; Burnet, M. C.; Nakayasu, E. S.; et al. Addressing the Challenge of Soil Metaproteome Complexity by Improving Metaproteome Depth of Coverage through Two-Dimensional Liquid Chromatography. *Soil Biology and Biochemistry* **2018**, *125*, 290–299.

(110) White, R. A.; Bottos, E. M.; Chowdhury, T. R.; Zucker, J. D.; Brislawn, C. J.; Nicora, C. D.; Fansler, S. J.; Glaesemann, K. R.; Glass, K.; Jansson, J. K. Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. *mSystems* **2016**, *1* (3), e00045-16.

(111) Cardinale, M. Scanning a Microhabitat: Plant-Microbe Interactions Revealed by Confocal Laser Microscopy. *Frontiers in Microbiology* **2014**, *5*.

(112) Jones, D. L.; Kielland, K. Soil Amino Acid Turnover Dominates the Nitrogen Flux in Permafrost-Dominated Taiga Forest Soils. *Soil Biology and Biochemistry* **2002**, *34* (2), 209–219.

(113) Näsholm, T.; Kielland, K.; Ganeteg, U. Uptake of Organic Nitrogen by Plants. *New Phytologist* **2009**, *182* (1), 31–48.

(114) Fierer, N.; Bradford, M. A.; Jackson, R. B. Toward an Ecological Classification of Soil Bacteria. *Ecology* **2007**, *88* (6), 1354–1364.

(115) Woodcroft, B. J.; Singleton, C. M.; Boyd, J. A.; Evans, P. N.; Emerson, J. B.; Zayed, A. A. F.; Hoelzle, R. D.; Lamberton, T. O.; McCalley, C. K.; Hodgkins, S. B.; et al. Genome-Centric View of Carbon Processing in Thawing Permafrost. *Nature* **2018**, *560* (7716), 49–54.

(116) Tveit, A. T.; Urich, T.; Svenning, M. M. Metatranscriptomic Analysis of Arctic Peat Soil Microbiota. *Applied and Environmental Microbiology* **2014**, *80* (18), 5761–5772.

(117) Wieder, W. R.; Grandy, A. S.; Kallenbach, C. M.; Bonan, G. B. Integrating Microbial Physiology and Physio-Chemical Principles in Soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) Model. *Biogeosciences* **2014**, *11* (14), 3899–3917.

V. CONCLUSION

Novel metaproteomic methods were used to elucidate the microbial biogeochemistry of Arctic soils, revealing key processes in the soils, patterns of resource partitioning between major taxa, and changes associated with increasing plant biomass across the warming Arctic. Microbial activity in the rhizosphere is distinct from that centered on the breakdown of soil organic matter. Rhizospheric groups specialize in the acquisition of small, soluble carbon compounds exuded by plant roots and the acquisition of scarce N to alleviate nutrient limitation. These groups are more active and express higher levels of their associated functional profile in soils with higher floral biomass, pointing to a greater role of plant interactions in soil microbial activity as the Arctic warms. The polysaccharide-degrading, highly active Acidobacteria concentrate on certain components of plant detritus while different groups degrade other components. These results can both guide bacterial cultivation and inform biogeochemical models of soil processes.

New data analysis methods were critical to this work, with the *ProteinExpress* pipeline handling the identification and annotation of protein sequences from the Arctic samples. De novo sequencing was also developed into a viable alternative to traditional methods of peptide identification in the absence of an appropriate reference dataset, as can be the case with complex metaproteomes. Post-processing of de novo sequences by the novel *Postnovo* algorithm increases the accuracy of sequence predictions by about an order of magnitude. Postnovo has the potential to be used with de novo sequencing just as mainstream post-processing methods are regularly used with traditional database search.[1] Postnovo and ProteinExpress can be applied to any environmental sample. The metaproteomic investigation of microbial activity in soils holds great promise for the elucidation of biogeochemistry and ecophysiology, and for bridging the gap between the two.

REFERENCE

(1) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nature Methods* **2007,** *4*, 923–925.

VI. APPENDIX

Table VI.1. Contributions of Functional Group latent variables to the first discriminant function of the bin fidelity LDA (shown in Figure IV.8.A)

Functional Group	Factor 1 Loading
Lactate Fermentation	0.241
Polyphosphate Metabolism	0.221
Alkanesulfonate Assimilation	0.187
Nitrogen Fixation	0.167
DNA Ligase	0.145
Homogentisate Pathway	0.141
Polysulfide Reduction	0.137
Glycerol Catabolism	0.134
Shikimic Acid Pathway	0.123
Trehalose Synthesis	0.121
Iron Import	0.121
Taurine Assimilation	0.115
Bacterial Cellulose Synthesis	0.114
Chromosome Partitioning	0.114
Beta Oxidation	0.109
CO2 Fixation	0.105
tRNA Ligase	0.105
Thiosulfate Oxidation	0.099
Mercury Resistance	0.098
Methylpentose Cleavage	0.097
Dihydrolipoyl Dehydrogenase	0.093
Pyruvate Decarboxylation	0.091
Ribose Transport	0.086
RNA Polymerase Machinery	0.086
Glycogen Synthesis	0.084
Glycine Cleavage System	0.077
Beta-Galactoside Catabolism	0.075
Ammonium Transport	0.074
Bam Omp Assembly	0.074
Ribosome	0.073
Nucleotide Synthesis	0.073
Arsenic Resistance	0.071
TCA Cycle	0.061
Ammonia Metabolism	0.058
ACS Acetate Metabolism	0.057
DNA Recombination	0.050
UDP-GlcNAc Synthesis	0.050
Rhamnose Transport	0.048
Phosphonate Assimilation	0.046
Phosphoglycerol Import	0.046
DEAD-box RNA Helicase	0.045
Pentose Catabolism	0.045
LPS Assembly	0.038
Alpha-Galactosidase	0.038
Glyoxylate Cycle	0.035
Heme Synthesis	0.034
Polyamine Synthesis	0.031
Sugar Alcohol Catabolism	0.031
DNA Supercoiling	0.029
Class II RNR	0.028
Phenylacetate Catabolism	0.018
Core-Lipid A Synthesis	0.018
Chemotaxis	0.018
Starch+Glycogen Catabolism	0.010
Myxococcal Gliding	0.006
Sugar Alcohol Transport	0.006

(continued from previous page)

Cell Division Septum	0.005
Bacteroidete Gliding	0.004
Peptidoglycan Synthesis	0.004
Nodulation	0.003
Carboxidotrophy	0.001
Rod Morphogenesis	0.000
Ferritin	-0.001
Transformation	-0.003
PNPase	-0.003
Xylose+Arabinose Transport	-0.005
Class I RNR	-0.010
Translation	-0.010
Omp	-0.011
Cellulases	-0.011
Succinoglycan Synthesis	-0.013
Oligopeptide Transport	-0.016
Rhamnose LPS Synthesis	-0.017
Chromatin Packaging	-0.021
Amino Acid Synthesis	-0.022
Pentose Phosphate Pathway	-0.022
Pilin+Fimbria	-0.025
Integrase	-0.026
Alginate Synthesis	-0.027
DNA Repair	-0.027
Phosphate Assimilation	-0.028
Outer Membrane Porin	-0.028
Sigma70 Exponential Phase	-0.033
Carotenoid Synthesis	-0.036
Uronate Catabolism	-0.036
DNA Replication	-0.037
Amino Acid Catabolism	-0.039
Sorbosone Catabolism	-0.040
Gluconeogenesis	-0.041
Flagellum	-0.044
Sigma54 Nitrogen Limitation	-0.044
Beta-Glucosidase	-0.050
Amino Acid Transport	-0.052
Colanic Acid+Capsule Synthesis	-0.055
Tol-Pal Outer Membrane Integrity	-0.056
Xylose Cleavage	-0.057
Glycolysis	-0.059
Essential Transcription Factor	-0.059
Sarcosine Metabolism	-0.061
Sulfatase	-0.063
Microcompartment	-0.065
Ethanol Fermentation	-0.068
Mannose Cleavage	-0.071
Copper Transport	-0.071
Acetylglucosamidase	-0.071
Anaerobic Ring Cleavage	-0.072
ATP Synthase	-0.077
Propionate Catabolism	-0.078
Adhesin	-0.079
Aerobic Ring Cleavage	-0.082
Reverse Transcriptase	-0.088
Selenocysteine Utilization	-0.089
Transposase	-0.089
Methylotrophy	-0.090
Phospholipid Synthesis	-0.092
Potassium Import	-0.123
Nitrate+Nitrite Reduction	-0.126
Polyamine Transport	-0.128
Pectinase	-0.131
Urea Assimilation	-0.132

(continued from previous page)

PHA Synthesis	-0.135
Hexose Catabolism	-0.135
PTA-ACK Acetate Metabolism	-0.156

Table VI.2. Contributions of Functional Group latent variables to the second discriminant function of the bin fidelity LDA (shown in Figure IV.8.A)

Functional Group	Factor 2 Loading
Dihydrolipoyl Dehydrogenase	0.264
Core-Lipid A Synthesis	0.219
Taurine Assimilation	0.217
Carotenoid Synthesis	0.138
Sugar Alcohol Catabolism	0.132
Bacterial Cellulose Synthesis	0.128
Cell Division Septum	0.127
Alkanesulfonate Assimilation	0.123
Chromosome Partitioning	0.106
Phospholipid Synthesis	0.095
Ribosome	0.094
Glycine Cleavage System	0.092
TCA Cycle	0.089
Ribose Transport	0.087
Trehalose Synthesis	0.086
ACS Acetate Metabolism	0.079
Sigma70 Exponential Phase	0.077
Xylose+Arabinose Transport	0.075
Lactate Fermentation	0.075
Phosphate Assimilation	0.065
Carboxidotrophy	0.063
Starch+Glycogen Catabolism	0.062
Beta Oxidation	0.060
Bam Omp Assembly	0.057
Sigma54 Nitrogen Limitation	0.056
UDP-GlcNAc Synthesis	0.056
Methylpentose Cleavage	0.055
Sugar Alcohol Transport	0.054
Phosphonate Assimilation	0.051
DNA Replication	0.046
Bacteroidete Gliding	0.044
DNA Recombination	0.043
Class I RNR	0.041
Propionate Catabolism	0.039
Rhamnose LPS Synthesis	0.038
Polyphosphate Metabolism	0.037
Succinoglycan Synthesis	0.036
PHA Synthesis	0.034
Mannose Cleavage	0.030
Uronate Catabolism	0.029
Sulfatase	0.020
Xylose Cleavage	0.020
Outer Membrane Porin	0.020
Tol-Pal Outer Membrane Integrity	0.018
LPS Assembly	0.016
Nodulation	0.016
Beta-Galactoside Catabolism	0.014
Rod Morphogenesis	0.012
Chromatin Packaging	0.012
Pentose Phosphate Pathway	0.011
Thiosulfate Oxidation	0.007
Pyruvate Decarboxylation	0.005
DNA Repair	0.003
Amino Acid Synthesis	0.002
Methylotrophy	0.001
CO2 Fixation	-0.002
Amino Acid Transport	-0.004
Translation	-0.005
Potassium Import	-0.006

(continued from previous page)

Polyamine Synthesis	-0.006
Pectinase	-0.006
Homogentisate Pathway	-0.006
Glycolysis	-0.007
Essential Transcription Factor	-0.011
Phenylacetate Catabolism	-0.013
Transformation	-0.015
Ethanol Fermentation	-0.015
Nitrate+Nitrite Reduction	-0.017
Oligopeptide Transport	-0.017
DNA Supercoiling	-0.017
Shikimic Acid Pathway	-0.018
Pentose Catabolism	-0.019
Myxococcal Gliding	-0.022
ATP Synthase	-0.024
Transposase	-0.025
Selenocysteine Utilization	-0.025
Alpha-Galactosidase	-0.027
Colanic Acid+Capsule Synthesis	-0.029
Alginate Synthesis	-0.032
Class II RNR	-0.032
Polysulfide Reduction	-0.032
Gluconeogenesis	-0.032
Microcompartment	-0.033
tRNA Ligase	-0.035
DNA Ligase	-0.036
Phosphoglycerol Import	-0.038
Reverse Transcriptase	-0.038
Urea Assimilation	-0.040
Aerobic Ring Cleavage	-0.043
RNA Polymerase Machinery	-0.049
Ferritin	-0.050
Acetylglucosamidase	-0.051
Beta-Glucosidase	-0.052
Arsenic Resistance	-0.055
Copper Transport	-0.056
Glycerol Catabolism	-0.057
PTA-ACK Acetate Metabolism	-0.059
Ammonium Transport	-0.065
Sarcosine Metabolism	-0.067
Nitrogen Fixation	-0.070
Pilin+Fimbria	-0.071
Omp	-0.071
Integrase	-0.071
Iron Import	-0.080
DEAD-box RNA Helicase	-0.081
Heme Synthesis	-0.086
Ammonia Metabolism	-0.089
Sorbosone Catabolism	-0.091
Rhamnose Transport	-0.096
Peptidoglycan Synthesis	-0.103
PNPase	-0.110
Mercury Resistance	-0.114
Nucleotide Synthesis	-0.121
Hexose Catabolism	-0.124
Glycogen Synthesis	-0.126
Cellulases	-0.130
Polyamine Transport	-0.131
Flagellum	-0.136
Adhesin	-0.144
Amino Acid Catabolism	-0.149
Chemotaxis	-0.153
Glyoxylate Cycle	-0.181
Anaerobic Ring Cleavage	-0.221

Table VI.3. Functional Group definitions from unique pairs of eggNOG predicted Gene Family name and eggNOG HMM model annotation

Functional Group	Gene Family	eggNOG HMM Model Annotation
Glycolysis	PPGK	ROK family
Glycolysis	PPGK	Polyphosphate glucokinase
Glycolysis	GLK	Glucokinase (EC 2.7.1.2)
Glycolysis	PGI	Phosphohexose isomerase
Glycolysis	PGI	Glucose-6-phosphate isomerase
Glycolysis	PFKA	Phosphohexokinase
Glycolysis	PFKA	Phosphofructokinase
Glycolysis	PFKA	Ec 2.7.1.11
Glycolysis	PFK	K00850 6-phosphofructokinase 1 EC 2.7.1.11
Glycolysis	PFK	Ec 2.7.1.11
Glycolysis	PFKB	PfkB domain protein
Glycolysis	FBAB	Fructose-bisphosphate aldolase
Glycolysis	FBAB	Fructose-bisphosphate aldolase (EC 4.1.2.13)
Glycolysis	FBAB	DeoC
Glycolysis	FBA	Fructose-bisphosphate aldolase class-II
Glycolysis	FBA	Fructose-bisphosphate aldolase
Glycolysis	FBA	Fructose-bisphosphate aldolase, class II
Glycolysis	FBAA	Fructose-bisphosphate aldolase
Glycolysis	TPIA	Triose-phosphate isomerase
Glycolysis	TPIA	DoxX family
Glycolysis	GAPA	Catalyzes the NAD-dependent conversion of D-erythrose 4- phosphate to 4-phosphoerythronate (By similarity)
Glycolysis	GAPA	Glyceraldehyde-3-phosphate dehydrogenase
Glycolysis	GAPA	Glyceraldehyde-3-phosphate dehydrogenase, type I
Glycolysis	GAP	Catalyzes the NAD-dependent conversion of D-erythrose 4- phosphate to 4-phosphoerythronate (By similarity)
Glycolysis	GAP	Glyceraldehyde-3-phosphate dehydrogenase
Glycolysis	GAP	Glyceraldehyde-3-phosphate dehydrogenase, type I
Glycolysis	PGK	Phosphoglycerate kinase
Glycolysis	APGM	Phosphoglycerate mutase
Glycolysis	GPMA	Catalyzes the interconversion of 2-phosphoglycerate and 3-phosphoglycerate (By similarity)
Glycolysis	GPMI	Catalyzes the interconversion of 2-phosphoglycerate and 3-phosphoglycerate (By similarity)
Glycolysis	ENO	Catalyzes the reversible conversion of 2- phosphoglycerate into phosphoenolpyruvate. It is essential for the degradation of carbohydrates via glycolysis (By similarity)
Glycolysis	PYK	Pyruvate kinase
Glycolysis	PYKA	Pyruvate kinase
Glycolysis	PYKA	Pyruvate kinase (EC 2.7.1.40)
Glycolysis	PYKF4	Pyruvate kinase
Glycolysis		Glyceraldehyde-3-phosphate dehydrogenase
Glycolysis		Phosphoglycerate kinase
Glycolysis		Phosphoglycerate mutase
Pentose Phosphate Pathway	ZWF	Glucose-6-phosphate 1-dehydrogenase
Pentose Phosphate Pathway	ZWF1	Glucose-6-phosphate 1-dehydrogenase
Pentose Phosphate Pathway	ZWF2	Glucose-6-phosphate 1-dehydrogenase
Pentose Phosphate Pathway	PGL	Inherit from bactNOG: 6-phosphogluconolactonase (EC 3.1.1.31)
Pentose Phosphate Pathway	PGL	6-phosphogluconolactonase EC 3.1.1.31
Pentose Phosphate Pathway	PGL	Catalyzes the hydrolysis of 6-phosphogluconolactone to 6-phosphogluconate (By similarity)
Pentose Phosphate Pathway	PGL	K01057 6-phosphogluconolactonase EC 3.1.1.31
Pentose Phosphate Pathway	GND	Catalyzes the oxidative decarboxylation of 6- phosphogluconate to ribulose 5-phosphate and CO(2), with concomitant reduction of NADP to NADPH (By similarity)
Pentose Phosphate Pathway	GND	6-phosphogluconate dehydrogenase (Decarboxylating)
Pentose Phosphate Pathway	GND	6-phosphogluconate dehydrogenase
Pentose Phosphate Pathway	RPIA	Ribose 5-phosphate isomerase A (phosphoriboisomerase A)
Pentose Phosphate Pathway	RPIA	Phosphoriboisomerase A
Pentose Phosphate Pathway	TKT	Transketolase (EC 2.2.1.1)
Pentose Phosphate Pathway	TKT	Transketolase
Pentose Phosphate Pathway	TKTA	Transketolase (EC 2.2.1.1)
Pentose Phosphate Pathway	TKTA	Transketolase

(continued from previous page)

Pentose Phosphate Pathway	TAL	Transaldolase is important for the balance of metabolites in the pentose-phosphate pathway (By similarity)
Pentose Phosphate Pathway	FSA,TAL	Transaldolase is important for the balance of metabolites in the pentose-phosphate pathway (By similarity)
Pentose Phosphate Pathway		Glucose-6-phosphate 1-dehydrogenase
Pentose Phosphate Pathway		6-phosphogluconolactonase EC 3.1.1.31
Pentose Phosphate Pathway		6-phosphogluconate dehydrogenase
Pentose Phosphate Pathway		NAD binding domain of 6-phosphogluconate dehydrogenase
Pentose Phosphate Pathway		Transketolase
Pentose Phosphate Pathway		Transketolase, thiamine diphosphate binding domain
Pentose Phosphate Pathway		Transaldolase
Pyruvate Decarboxylation	ACEE	Component of the pyruvate dehydrogenase (PDH) complex, that catalyzes the overall conversion of pyruvate to acetyl-CoA and CO(2) (By similarity)
Pyruvate Decarboxylation	ACEE	Pyruvate dehydrogenase E1 component
Pyruvate Decarboxylation	PDHA	Pyruvate dehydrogenase (Acetyl-transferring) E1 component, alpha subunit
Pyruvate Decarboxylation	PDHA	Pyruvate dehydrogenase
Pyruvate Decarboxylation	PDHB	Pyruvate dehydrogenase subunit beta
Pyruvate Decarboxylation	ACEF	Pyruvate dehydrogenase complex
Pyruvate Decarboxylation	ACEF	Dihydrolipoamide acetyltransferase
Pyruvate Decarboxylation	ACEF	Catalytic domain of components of various dehydrogenase complexes
Dihydrolipoyl Dehydrogenase	LPDA	Dihydrolipoyl dehydrogenase
Dihydrolipoyl Dehydrogenase	LPDA	Dihydrolipoamide dehydrogenase
Dihydrolipoyl Dehydrogenase	LPDA	Mercuric reductase
Dihydrolipoyl Dehydrogenase	LPD	Dihydrolipoyl dehydrogenase
Dihydrolipoyl Dehydrogenase	LPD	(dihydrolipoamide dehydrogenase) (EC 1.8.1.4)
Dihydrolipoyl Dehydrogenase	LPDG	Dihydrolipoyl dehydrogenase
Dihydrolipoyl Dehydrogenase	LPDG	Dihydrolipoamide dehydrogenase
TCA Cycle	CITA	Transporter
TCA Cycle	CITA	Signal transduction histidine kinase regulating citrate malate metabolism
TCA Cycle	CITA	Citrate synthase
TCA Cycle	GLTA	Pyridine nucleotide-disulfide oxidoreductase
TCA Cycle	GLTA	Glutamate synthase
TCA Cycle	GLTA	Citrate synthase
TCA Cycle	CITZ	2-methylcitrate synthase citrate synthase II
TCA Cycle	CITZ	Citrate synthase
TCA Cycle	KORA	2-oxoglutarate ferredoxin oxidoreductase subunit alpha
TCA Cycle	KORA	2-oxoacid acceptor oxidoreductase, alpha subunit
TCA Cycle	KORA	Pyruvate ferredoxin/flavodoxin oxidoreductase
TCA Cycle	KORA	Ferredoxin oxidoreductase
TCA Cycle	KORA	Ferredoxin
TCA Cycle	ACNA	Aconitate hydratase
TCA Cycle	ACNA	Aconitate hydratase 1
TCA Cycle	SUCA	2-oxoglutarate dehydrogenase, E1 subunit
TCA Cycle	SUCA	2-oxoglutarate dehydrogenase, E1
TCA Cycle	SUCA	2-oxoglutarate dehydrogenase e1 component
TCA Cycle	SUCB	Dihydrolipoamide
TCA Cycle	SUCB	2-oxoglutarate dehydrogenase E2 component
TCA Cycle	SUCB	2-oxoglutarate dehydrogenase E2 component, dihydrolipoamide succinyltransferase
TCA Cycle	SUCB	Of components of various dehydrogenase complexes
TCA Cycle	SUCB	Dihydrolipoamide succinyltransferase
TCA Cycle	SUCC	Succinyl-CoA synthetase subunit beta
TCA Cycle	SUCD	Succinyl-CoA ligase ADP-forming subunit alpha
TCA Cycle	SDHA	Succinate dehydrogenase or fumarate reductase, flavoprotein subunit
TCA Cycle	SDHA	Succinate dehydrogenase (Flavoprotein subunit)
TCA Cycle	SDHA	Succinate dehydrogenase
TCA Cycle	SDHA	Succinate dehydrogenase, flavoprotein subunit
TCA Cycle	SDHA1	Succinate dehydrogenase (Flavoprotein subunit)

(continued from previous page)

TCA Cycle	SDHB	Succinate dehydrogenase
TCA Cycle	FUMC	Fumarate hydratase class II
TCA Cycle	FUMA	Fumarate
TCA Cycle	MQO	Malate dehydrogenase quinone
TCA Cycle	MQO	Malate dehydrogenase (quinone)
TCA Cycle	MDH	Catalyzes the reversible oxidation of malate to oxaloacetate (By similarity)
TCA Cycle	MDH	L-Lactate dehydrogenase
TCA Cycle	MDH	Malate dehydrogenase
TCA Cycle		Citrate synthase
TCA Cycle		Isocitrate dehydrogenase (NADp)
Glyoxylate Cycle	ACEA	Isocitrate lyase
Glyoxylate Cycle	ACEB	Malate synthase (EC 2.3.3.9)
Glyoxylate Cycle	ACEB	Malate synthase
Glyoxylate Cycle	GLCB	Malate synthase g
Glyoxylate Cycle	GLCB	Malate synthase
Glyoxylate Cycle		Malate synthase
Gluconeogenesis	GLPX	Fructose-1,6-bisphosphatase
Gluconeogenesis	GLPX	Bacterial fructose-1,6-bisphosphatase, glpX-encoded
Gluconeogenesis	PPSA	Catalyzes the phosphorylation of pyruvate to phosphoenolpyruvate (By similarity)
Gluconeogenesis	PPSA	Phosphoenolpyruvate synthase
Gluconeogenesis	PCKA	Phosphoenolpyruvate carboxylase
Gluconeogenesis	PCKA	Phosphoenolpyruvate Carboxylase
Gluconeogenesis	PCKA	Phosphoenolpyruvate carboxykinase
Gluconeogenesis	PCKG	Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP), the rate-limiting step in the metabolic pathway that produces glucose from lactate and other precursors derived from the citric acid cycle (By similarity)
Gluconeogenesis	MAEB	Malic enzyme
Gluconeogenesis	MAEB	Malic protein NAD-binding
Gluconeogenesis	DME	Malic enzyme
Gluconeogenesis	SFCA	Malic_M
Gluconeogenesis	MLEA	Malic enzyme, NAD binding domain
Gluconeogenesis		Phosphoenolpyruvate carboxykinase
ATP Synthase	ATPA	Produces ATP from ADP in the presence of a proton gradient across the membrane. The V-type alpha chain is a catalytic subunit (By similarity)
ATP Synthase	ATPA	Produces ATP from ADP in the presence of a proton gradient across the membrane. The alpha chain is a regulatory subunit (By similarity)
ATP Synthase	ATPD	Produces ATP from ADP in the presence of a proton gradient across the membrane. The catalytic sites are hosted primarily by the beta subunits (By similarity)
ATP Synthase	ATPE	F(1)F(0) ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F(1) containing the extramembraneous catalytic core and F(0) containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F(1) is coupled via a rotary mechanism of the central stalk subunits to proton translocation (By similarity)
ATP Synthase	ATPE	Subunit C
ATP Synthase	ATPE	ATP synthase, F0 subunit c
ATP Synthase	ATPE	ATP synthase F0, C subunit
ATP Synthase	ATPE	ATP synthase, subunit C
ATP Synthase	ATPE	ATP synthase
ATP Synthase	ATPE	Atp synthase
ATP Synthase	ATPH	F(1)F(0) ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F(1) containing the extramembraneous catalytic core and F(0) containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F(1) is coupled via a rotary mechanism of the central stalk subunits to proton translocation (By similarity)
ATP Synthase	ATPH	ATP synthase delta (OSCP) subunit
ATP Synthase	ATPA,ATPA1	Produces ATP from ADP in the presence of a proton gradient across the membrane. The alpha chain is a regulatory subunit (By similarity)
ATP Synthase	ATPG	Produces ATP from ADP in the presence of a proton gradient across the membrane. The gamma chain is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex (By similarity)
ATP Synthase	ATPC	F0F1 ATP synthase subunit epsilon (EC 3.6.3.14)
ATP Synthase	ATPC	Produces ATP from ADP in the presence of a proton gradient across the membrane (By similarity)
ATP Synthase	ATPD2	Produces ATP from ADP in the presence of a proton gradient across the membrane. The catalytic sites are hosted primarily by the beta subunits (By similarity)

(continued from previous page)

ATP Synthase	ATPB	Produces ATP from ADP in the presence of a proton gradient across the membrane. The archaeal beta chain is a regulatory subunit
ATP Synthase	ATPB	It plays a direct role in the translocation of protons across the membrane (By similarity)
ATP Synthase	ATPF2	F(1)F(0) ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F(1) containing the extramembraneous catalytic core and F(0) containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F(1) is coupled via a rotary mechanism of the central stalk subunits to proton translocation (By similarity)
ATP Synthase	ATPF	F(1)F(0) ATP synthase produces ATP from ADP in the presence of a proton or sodium gradient. F-type ATPases consist of two structural domains, F(1) containing the extramembraneous catalytic core and F(0) containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F(1) is coupled via a rotary mechanism of the central stalk subunits to proton translocation (By similarity)
ATP Synthase	ATPF	Component of the F(0) channel, it forms part of the peripheral stalk, linking F(1) to F(0) (By similarity)
ATP Synthase		H transporting two-sector ATPase subunit C
ATP Synthase		ATP synthase subunit C
ATP Synthase		ATP synthase F0 C subunit
ATP Synthase		H -transporting two-sector ATPase subunit C
ATP Synthase		ATP synthase, Delta/Epsilon chain, beta-sandwich domain
Ethanol Fermentation	PORC	Pyruvate ketoisovalerate oxidoreductase, gamma subunit
Ethanol Fermentation	PFLA	Pyruvate formate-lyase 1-activating enzyme
Ethanol Fermentation	PFLB	Formate acetyltransferase
Lactate Fermentation	DLD	FAD linked oxidase domain protein
Lactate Fermentation	DLD	FAD linked oxidases, C-terminal domain
Lactate Fermentation	DLD	D-lactate dehydrogenase
Lactate Fermentation	DLD2	FAD linked oxidase
Lactate Fermentation	DLD2	FAD linked oxidase domain protein
Lactate Fermentation	LDHA2	D-isomer specific 2-hydroxyacid dehydrogenase
Lactate Fermentation	LDH	L-lactate dehydrogenase
Lactate Fermentation	LDH	Dehydrogenase
Lactate Fermentation	LDH	Leucine dehydrogenase
Lactate Fermentation	LLDD	L-lactate dehydrogenase
Lactate Fermentation	LCTP	L-lactate permease
Lactate Fermentation		D-lactate dehydrogenase (cytochrome)
Lactate Fermentation		Lactate utilization protein B C
CO2 Fixation	CBBL	RuBisCO catalyzes two reactions the carboxylation of D- ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate. Both reactions occur simultaneously and in competition at the same active site (By similarity)
CO2 Fixation	CBBL	RuBisCO catalyzes two reactions the carboxylation of D- ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate in the photorespiration process. Both reactions occur simultaneously and in competition at the same active site (By similarity)
CO2 Fixation	PRKB	Phosphoribulokinase (EC 2.7.1.19)
CO2 Fixation	CBBM	RuBisCO catalyzes two reactions the carboxylation of D- ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate. Both reactions occur simultaneously and in competition at the same active site
CO2 Fixation	CBBM	RuBisCO catalyzes two reactions the carboxylation of D- ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate. Both reactions occur simultaneously and in competition at the same active site (By similarity)
CO2 Fixation	CBBX	CbxX CfqX family protein
CO2 Fixation	CBBX	CbbX protein
CO2 Fixation		RuBisCO catalyzes two reactions the carboxylation of D- ribulose 1,5-bisphosphate, the primary event in carbon dioxide fixation, as well as the oxidative fragmentation of the pentose substrate in the photorespiration process. Both reactions occur simultaneously and in competition at the same active site (By similarity)
Carboxidotrophy	COOXM	Carbon monoxide dehydrogenase, medium
Carboxidotrophy	COXM	Molybdopterin dehydrogenase FAD-binding
Carboxidotrophy	COXM	Carbon monoxide dehydrogenase, medium
Carboxidotrophy	COXM	Dehydrogenase
Carboxidotrophy	COXM	CO dehydrogenase flavoprotein C-terminal domain
Carboxidotrophy	COXG	Carbon monoxide dehydrogenase
Carboxidotrophy	COXL	Dehydrogenase
Carboxidotrophy	COXL	Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding
Methylotrophy	GFA	Catalyzes the condensation of formaldehyde and glutathione to S-hydroxymethylglutathione (By similarity)

(continued from previous page)

Methylotrophy	FGHA	S-Formylglutathione hydrolase
Methylotrophy	FGHA	S-formylglutathione hydrolase
Methylotrophy	FAE	Formaldehyde-activating enzyme
Methylotrophy	FDHA	Dehydrogenase
Methylotrophy	FDHA	Formate dehydrogenase Alpha subunit
Methylotrophy	FDH	Formate dehydrogenase
Methylotrophy	FDH	Aldo Keto reductase
Methylotrophy	FDHF	Oxidoreductase alpha (molybdopterin) subunit
Methylotrophy	FDHF	Molydopterin dinucleotide binding domain
Methylotrophy	FDHF	Oxidoreductase, alpha molybdopterin subunit
Methylotrophy	FDHD	Necessary for formate dehydrogenase activity (By similarity)
Methylotrophy	FDHC	Formate dehydrogenase
Methylotrophy	FOLD	Catalyzes the oxidation of 5,10- methylenetetrahydrofolate to 5,10-methenyltetrahydrofolate and then the hydrolysis of 5,10-methenyltetrahydrofolate to 10- formyltetrahydrofolate (By similarity)
Methylotrophy	METF	Methylenetetrahydrofolate reductase
Methylotrophy	METF-2	Methylenetetrahydrofolate reductase
Methylotrophy	FHS	Formyltetrahydrofolate synthetase
Methylotrophy	MXAF	PQQ-dependent dehydrogenase, methanol ethanol family
Methylotrophy	MXAF	Dehydrogenase
Methylotrophy	SGAA	Cys/Met metabolism PLP-dependent enzyme
Methylotrophy	SGAA	Class V aminotransferase
Methylotrophy		Glutathione-dependent formaldehyde-activating, GFA
Methylotrophy		Glutathione-dependent formaldehyde-activating GFA
Methylotrophy		Methylenetetrahydrofolate reductase
Beta Oxidation	ACD	Acyl-CoA dehydrogenase
Beta Oxidation	ACD,MMGC	Acyl-CoA dehydrogenase
Beta Oxidation	ACDA	Acyl-CoA dehydrogenase
Beta Oxidation	ACDA	Dehydrogenase
Beta Oxidation	ACDB	Acyl-CoA dehydrogenase
Beta Oxidation	ALKK	AMP-dependent synthetase and ligase
Beta Oxidation	ALKK	Amp-dependent synthetase and ligase
Beta Oxidation	ALKK	AMP-binding enzyme
Beta Oxidation	BMUL_0578	Acyl-CoA dehydrogenase
Beta Oxidation	BMUL_0578	Acyl-Coa dehydrogenase
Beta Oxidation	FADA	Catalyzes the final step of fatty acid oxidation in which acetyl-CoA is released and the CoA ester of a fatty acid two carbons shorter is formed (By similarity)
Beta Oxidation	FADA	Acetyl-CoA acetyltransferase
Beta Oxidation	FADA	Acetyl-coa acetyltransferase
Beta Oxidation	FADB	3-hydroxyacyl-CoA dehydrogenase
Beta Oxidation	FADB	Catalyzes the formation of a hydroxyacyl-CoA by addition of water on enoyl-CoA. Also exhibits 3-hydroxyacyl-CoA epimerase and 3-hydroxyacyl-CoA dehydrogenase activities (By similarity)
Beta Oxidation	FADB	3-hydroxyacyl-coa dehydrogenase
Beta Oxidation	FADB	Oxidation complex subunit alpha
Beta Oxidation	FADB	3-hydroxyacyl-COA dehydrogenase
Beta Oxidation	FADB2X, HADH2	Short-chain dehydrogenase
Beta Oxidation	FADD	Amp-dependent synthetase and ligase
Beta Oxidation	FADD	AMP-dependent synthetase and ligase
Beta Oxidation	FADD4	Amp-dependent synthetase and ligase
Beta Oxidation	FADD5	Amp-dependent synthetase and ligase
Beta Oxidation	FADD6	AmP-dependent synthetase and ligase
Beta Oxidation	FADD19	Long-chain fatty acid-CoA ligase activity
Beta Oxidation	FADD22	AMP-binding enzyme
Beta Oxidation	FADD35	Amp-dependent synthetase and ligase
Beta Oxidation	FADD35	AMP-dependent synthetase and ligase
Beta Oxidation	FADE	Acyl-Coa dehydrogenase
Beta Oxidation	FADE	Acyl-CoA dehydrogenase
Beta Oxidation	FADE	Dehydrogenase
Beta Oxidation	FADE1_1	Acyl-Coa dehydrogenase
Beta Oxidation	FADE5	Acyl-Coa dehydrogenase
Beta Oxidation	FADE9	Acyl-CoA dehydrogenase

(continued from previous page)

Beta Oxidation	FADE10	Acyl-Coa dehydrogenase
Beta Oxidation	FADE12_3	Acyl-Coa dehydrogenase
Beta Oxidation	FADE13	Acyl-Coa dehydrogenase
Beta Oxidation	FADE15,FADE5	Acyl-Coa dehydrogenase
Beta Oxidation	FADE26	Acyl-CoA dehydrogenase
Beta Oxidation	FADE30	Acyl-Coa dehydrogenase
Beta Oxidation	FADF	Iron-sulfur cluster-binding protein
Beta Oxidation	FADF	Cysteine-rich domain
Beta Oxidation	FADH	NADH flavin oxidoreductase, NADH oxidase
Beta Oxidation	FADH	2,4-dienoyl-coA reductase
Beta Oxidation	FADL	Long-chain fatty acid transport protein
Beta Oxidation	FADL	Membrane protein involved in aromatic hydrocarbon degradation
Beta Oxidation	HADH2	Short-chain dehydrogenase
Beta Oxidation	HADH2	Short-chain dehydrogenase reductase sdr
Beta Oxidation	HADH2	Dehydrogenase
Beta Oxidation		3-hydroxyacyl-COA dehydrogenase
Beta Oxidation		Acyl-CoA dehydrogenase
Beta Oxidation		Acyl-Coa dehydrogenase
Beta Oxidation		Acyl-CoA dehydrogenase domain protein
Beta Oxidation		Acyl-CoA dehydrogenase, N-terminal domain
Beta Oxidation		Acyl-CoA dehydrogenase-related protein
Beta Oxidation		Enoyl-CoA hydratase
Beta Oxidation		Long-chain-fatty-acid--CoA ligase
PTA-ACK Acetate Metabolism	ACKA	Catalyzes the formation of acetyl phosphate from acetate and ATP. Can also catalyze the reverse reaction (By similarity)
PTA-ACK Acetate Metabolism	PTA	Phosphate acetyltransferase EC
PTA-ACK Acetate Metabolism	PTA	Involved in acetate metabolism (By similarity)
ACS Acetate Metabolism	ACSA	Acetate CoA ligase
ACS Acetate Metabolism	ACSA	Acetoacetyl-CoA synthase
ACS Acetate Metabolism	ACSA	Pfam:DUF3448
ACS Acetate Metabolism	ACSA	Acetoacetyl-CoA synthetase
ACS Acetate Metabolism	ACSA	Catalyzes the conversion of acetate into acetyl-CoA (AcCoA), an essential intermediate at the junction of anabolic and catabolic pathways. AcsA undergoes a two-step reaction. In the first half reaction, AcsA combines acetate with ATP to form acetyl-adenylate (AcAMP) intermediate. In the second half reaction, it can then transfer the acetyl group from AcAMP to the sulfhydryl group of CoA, forming the product AcCoA (By similarity)
ACS Acetate Metabolism	ACSA_1	Synthetase
ACS Acetate Metabolism	ACTP	SSS sodium solute transporter superfamily
ACS Acetate Metabolism	ACTP	P-type atpase
ACS Acetate Metabolism	ACTP	Solute symporter
Propionate Catabolism	PRPB	Methylisocitrate lyase
Propionate Catabolism	PRPC	Citrate synthase
Propionate Catabolism	PRPD	2-methylcitrate dehydratase
Propionate Catabolism	ACNB	Aconitate hydratase 2
Propionate Catabolism		MMGE PRPD family protein
Glycerol Catabolism	GLPF	MIP family channel protein
Glycerol Catabolism	GLPF	Major intrinsic protein
Glycerol Catabolism	GLPK	Key enzyme in the regulation of glycerol uptake and metabolism (By similarity)
Glycerol Catabolism	GYLR	Glycerol operon regulatory protein
Glycerol Catabolism	GLPD	Fad dependent oxidoreductase
Glycerol Catabolism	GLPD	Glycerol-3-phosphate dehydrogenase
Glycerol Catabolism	GLPC	Dehydrogenase subunit c
Glycerol Catabolism	GLPC	Ferredoxin
Glycerol Catabolism	GLPC	Dehydrogenase
Glycerol Catabolism	GLPQ	Diester phosphodiesterase
Glycerol Catabolism	GLPQ	Glycerophosphoryl diester phosphodiesterase
Glycerol Catabolism	GLPCD	FAD linked oxidase domain protein
Glycerol Catabolism	GLPQ2	Glycerophosphoryl diester phosphodiesterase
Glycerol Catabolism		NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus
Sorbosone Catabolism	SNDH	Dehydrogenase

(continued from previous page)

Sorbosone Catabolism	SNDH	L-sorbosone dehydrogenase
Sorbosone Catabolism		L-sorbosone dehydrogenase
Sugar Alcohol Transport	SMOE	Extracellular solute-binding protein, family 1
Sugar Alcohol Transport	SMOE	Extracellular solute-binding protein
Sugar Alcohol Transport	SMOM	TRAP dicarboxylate transporter-DctP subunit
Sugar Alcohol Transport	SMOM	Extracellular solute-binding protein, family 7
Sugar Alcohol Catabolism	YEIQ	Mannitol dehydrogenase
Sugar Alcohol Catabolism	IOLE	Xylose isomerase domain-containing protein TIM barrel
Sugar Alcohol Catabolism	IOLE	Catabolism protein
Sugar Alcohol Catabolism	IOLE	IolE protein
Sugar Alcohol Catabolism	IOLC	PfkB domain protein
Sugar Alcohol Catabolism		Myo-inosose-2 dehydratase
Sugar Alcohol Catabolism		Myo-inosose-2 dehydratase (EC 4.2.1.44)
Ribose Transport	RBSA1	ABC transporter
Ribose Transport	RBSA1	ABC transPORTER
Ribose Transport	RBSA	ABC transporter
Ribose Transport	RBSA	Abc transporter
Ribose Transport	RBSA	ABC transPORTER
Ribose Transport	RBSA	Part of the ABC transporter complex RbsABCD involved in ribose import. Responsible for energy coupling to the transport system (By similarity)
Ribose Transport	RBSB	D-ribose transporter subunit RbsB
Ribose Transport	RBSB	Periplasmic binding protein LacI transcriptional regulator
Ribose Transport	RBSB	ABC transporter substrate-binding protein
Ribose Transport	RBSB	Ribose ABC transporter
Ribose Transport	RBSB2	ABC transporter substrate-binding protein
Ribose Transport	RBSB1	Substrate binding component of ABC transporter
Ribose Transport	RBSB1	ABC transporter
Ribose Transport	RBSB1	(ABC) transporter
Ribose Transport	RBSB11	Periplasmic binding protein LacI transcriptional regulator
Ribose Transport	RBSB9	Periplasmic binding protein LacI transcriptional regulator
Ribose Transport	RBSB,YTFQ	Periplasmic binding protein LacI transcriptional regulator
Ribose Transport	RBSC	ABC transporter (permease)
Ribose Transport	RBSC	ABC transporter
Ribose Transport	RBSB10	Periplasmic binding proteins and sugar binding domain of LacI family
Ribose Transport	RBSB13	(ABC) transporter
Ribose Transport	RBSB5	(ABC) transporter
Ribose Transport		Sugar (D-ribose) ABC transporter (Periplasmic
Ribose Transport		Ribose binding protein of ABC transporter
Ribose Transport		Part of the ABC transporter complex RbsABCD involved in ribose import. Responsible for energy coupling to the transport system (By similarity)
Xylose+Arabinose Transport	XYLF	ABC transporter periplasmic
Xylose+Arabinose Transport	XYLF	ABC transporter
Xylose+Arabinose Transport	XYLG	Xylose transporter ATP-binding subunit
Xylose+Arabinose Transport	XYLG	Part of the ABC transporter complex XylFGH involved in xylose import. Responsible for energy coupling to the transport system
Xylose+Arabinose Transport	ARAF	Periplasmic binding protein LacI transcriptional regulator
Xylose+Arabinose Transport	ARAF	L-arabinose-binding periplasmic protein
Xylose+Arabinose Transport	ARAG	L-arabinose transporter ATP-binding protein
Xylose+Arabinose Transport	ARAG	Part of the ABC transporter complex RbsABCD involved in ribose import. Responsible for energy coupling to the transport system (By similarity)
Pentose Catabolism	XYLA	Xylose isomerase
Pentose Catabolism	XYLB	Xylulokinase (EC 2.7.1.17)
Pentose Catabolism	ARAA	Catalyzes the conversion of L-arabinose to L-ribulose (By similarity)
Pentose Catabolism	ARAA	L-arabinose isomerase
Pentose Catabolism	ARAB	Ribulokinase
Pentose Catabolism	ARAB	Ec 2.7.1.16
Pentose Catabolism	ARAB	K00853 L-ribulokinase EC 2.7.1.16
Pentose Catabolism	FUCO	Lactaldehyde reductase
Pentose Catabolism	Y0877	L-fucose isomerase, C-terminal domain
Pentose Catabolism	YAGF	Dehydratase family
Rhamnose Transport	RHAP	Monosaccharide-transporting ATPase (EC 3.6.3.17)
Rhamnose Transport	RHAT	L-rhamnose-proton symport
Hexose Catabolism	FRK	Fructokinase

(continued from previous page)

Hexose Catabolism	FRK	PfkB domain protein
Hexose Catabolism	SCRK	Fructokinase
Hexose Catabolism	GALM	Converts alpha-aldose to the beta-anomer. It is active on D-glucose, L-arabinose, D-xylose, D-galactose, maltose and lactose (By similarity)
Hexose Catabolism	GALT	Galactose-1-phosphate uridyl transferase, C-terminal domain
Hexose Catabolism	GALT	Galactose-1-phosphate uridylyltransferase
Hexose Catabolism	GALK	Galactokinase (EC 2.7.1.6)
Hexose Catabolism	GALK	Galactokinase
Hexose Catabolism	LACC	K00917 tagatose 6-phosphate kinase EC 2.7.1.144
Hexose Catabolism	LACD	Tagatose-bisphosphate aldolase
Hexose Catabolism	RHAD	Rhamnulose-1-phosphate aldolase alcohol dehydrogenase
Hexose Catabolism	RHAD	Catalyzes the reversible cleavage of L-rhamnulose-1- phosphate to dihydroxyacetone phosphate (DHAP) and L-lactaldehyde (By similarity)
Hexose Catabolism	RHAI	Xylose isomerase domain-containing protein
Hexose Catabolism	RHAM	Domain of unknown function (DUF718)
Hexose Catabolism	RHAS	Rhamnose ABC transporter, periplasmic rhamnose-binding protein
Hexose Catabolism	RHAS	ABC, transporter
Hexose Catabolism	RHAS	Rhamnose ABC transporter periplasmic rhamnose-binding protein
Hexose Catabolism	RHMD	Mandelate racemase muconate lactonizing
Uronate Catabolism	GNTR	Transcriptional regulator, GntR family
Uronate Catabolism	GNTR	LacI family transcriptional regulator
Uronate Catabolism	GNTR	Transcriptional regulator
Uronate Catabolism	GNTR	GntR family transcriptional regulator
Uronate Catabolism	GNTR	Transcriptional regulator, LacI family
Uronate Catabolism	GNTP	GntP family permease
Uronate Catabolism	GNTT	GntP family permease
Uronate Catabolism	GNTT	Gluconate transporter
Uronate Catabolism	UXAB	Altronate oxidoreductase
Uronate Catabolism	UXAC	Glucuronate isomerase
Uronate Catabolism	UXAC	Uronic isomerase
Uronate Catabolism	UIDA	Glycosyl hydrolases family 2, TIM barrel domain
Uronate Catabolism	KDUD	2-deoxy-D-gluconate 3-dehydrogenase
Uronate Catabolism	KDUD	Short-chain dehydrogenase reductase sdr
Uronate Catabolism	KDUI	Catalyzes the isomerization of 5-dehydro-4-deoxy-D- glucuronate to 3-deoxy-D-glycero-2,5-hexodiulosonate (By similarity)
Uronate Catabolism	KDGK	PfkB family carbohydrate kinase
Uronate Catabolism	KDGK	PfkB domain protein
Uronate Catabolism	KDGK,KGUK	PfkB domain protein
Uronate Catabolism	KDGD	5-keto-4-deoxy-glucarate dehydratase
Uronate Catabolism	KDGF	Cupin 2 conserved barrel domain protein
Uronate Catabolism		Glycosyl hydrolase family 67 N-terminus
Phenylacetate Catabolism	PAAK	AMP-binding enzyme
Phenylacetate Catabolism	PAAK	Phenylacetate--CoA ligase (EC 6.2.1.30)
Phenylacetate Catabolism	PAAK	Phenylacetate-CoA ligase
Phenylacetate Catabolism	PAAA	Phenylacetate-CoA oxygenase, PaaG subunit
Phenylacetate Catabolism	PAAB	Enoyl-CoA hydratase
Phenylacetate Catabolism	PAAB	Inherit from proNOG: Prolyl 4-hydroxylase, alpha subunit
Phenylacetate Catabolism	PAAB,PAAG	Enoyl-CoA hydratase
Phenylacetate Catabolism	PAAE	Phenylacetate-CoA oxygenase reductase, PaaK subunit
Phenylacetate Catabolism	PAAG	EnoyL-CoA hydratase
Phenylacetate Catabolism	PAAG	Enoyl-CoA hydratase
Phenylacetate Catabolism	PAAF	EnoyL-CoA hydratase
Phenylacetate Catabolism	PAAH	3-hydroxyacyl-coa dehydrogenase
Phenylacetate Catabolism	PAAH	3-hydroxyacyl-CoA dehydrogenase (EC
Phenylacetate Catabolism	PAAN	Aldehyde dehydrogenase
Phenylacetate Catabolism	PAAN	Phenylacetic acid degradation protein
Phenylacetate Catabolism	PAAX	Transcriptional regulator, PaaX family
Phenylacetate Catabolism	PAAX	Phenylacetic acid degradation operon negative regulatory protein
Phenylacetate Catabolism	PAAM	Outer membrane porin
Phenylacetate Catabolism		Phenylacetate-CoA ligase
Aerobic Ring Cleavage	BEND	Dehydrogenase
Aerobic Ring Cleavage	BEND	1,6-dihydroxycyclohexa-2,4-diene-1-carboxylate dehydrogenase

(continued from previous page)

Aerobic Ring Cleavage	BENR	Transcriptional regulator
Aerobic Ring Cleavage	PCAB	3-carboxy-cis-cis-muconate cycloisomerase
Aerobic Ring Cleavage	PCAB	3-carboxy-cis,cis-muconate cycloisomerase
Aerobic Ring Cleavage	PCAD	Carboxymuconolactone decarboxylase family
Aerobic Ring Cleavage	PCAD	3-oxoadipate enol-lactonase
Aerobic Ring Cleavage	PCAF	Catalyzes the final step of fatty acid oxidation in which acetyl-CoA is released and the CoA ester of a fatty acid two carbons shorter is formed (By similarity)
Aerobic Ring Cleavage	PCAF	Acetyl-CoA acetyltransferase
Aerobic Ring Cleavage	PCAF	Beta-ketoadipyl CoA thiolase
Aerobic Ring Cleavage	PCAF	Thiolase
Aerobic Ring Cleavage	PCAH	Protocatechuate 3,4-dioxygenase subunit beta
Aerobic Ring Cleavage	PCAK	Major facilitator Superfamily
Aerobic Ring Cleavage	PCAK	Major Facilitator superfamily
Aerobic Ring Cleavage	PCAK	Transporter
Aerobic Ring Cleavage	PCAK	Major Facilitator Superfamily protein
Aerobic Ring Cleavage	PCAR	IclR family transcriptional regulator
Aerobic Ring Cleavage	PCAR	Transcriptional regulator
Aerobic Ring Cleavage	LIGB	Protocatechuate 4,5-dioxygenase
Aerobic Ring Cleavage	LIGB	Catalyzes the formation of phosphodiester linkages between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction (By similarity)
Aerobic Ring Cleavage	LIGE	Glutathione S-transferase
Aerobic Ring Cleavage	LIGJ	4-Oxalomesaconate hydratase
Aerobic Ring Cleavage	LIGR	LysR family Transcriptional regulator
Aerobic Ring Cleavage	OCAR_5219	Carboxymuconolactone decarboxylase
Aerobic Ring Cleavage	CATB	Mandelate racemase muconate lactonizing
Aerobic Ring Cleavage	CATE	Glyoxalase bleomycin resistance protein dioxygenase
Aerobic Ring Cleavage	MANR	Mandelate racemase
Aerobic Ring Cleavage	XYLE	2,3-dioxygenase
Aerobic Ring Cleavage	BMUL_4012,XYLH	4-oxalocrotonate tautomerase
Aerobic Ring Cleavage	PHTD	4,5-dihydroxyphthalate decarboxylase
Aerobic Ring Cleavage	PCPA	12-dioxygenase
Aerobic Ring Cleavage	PCPB	Monooxygenase, FAD-binding
Aerobic Ring Cleavage	BOXC	Benzoyl-CoA-dihydrodiol lyase
Aerobic Ring Cleavage	POXD	Methane phenol toluene hydroxylase
Aerobic Ring Cleavage	HCAG	Feruloyl esterase
Aerobic Ring Cleavage		Aromatic-ring-hydroxylating dioxygenase beta subunit
Aerobic Ring Cleavage		4,5-dihydroxyphthalate decarboxylase
Aerobic Ring Cleavage		Dienelactone hydrolase
Aerobic Ring Cleavage		4-oxalocrotonate tautomerase
Aerobic Ring Cleavage		Feruloyl-CoA synthase
Aerobic Ring Cleavage		Feruloyl esterase
Aerobic Ring Cleavage		Benzoate-CoA ligase
Aerobic Ring Cleavage		3-carboxy-cis,cis-muconate cycloisomerase
Aerobic Ring Cleavage		Beta-ketoadipyl CoA thiolase
Aerobic Ring Cleavage		Protocatechuate 4,5-dioxygenase
Aerobic Ring Cleavage		Protocatechuate 4,5-dioxygenase subunit alpha
Aerobic Ring Cleavage		Mandelate racemase muconate lactonizing
Aerobic Ring Cleavage		Mandelate racemase muconate lactonizing protein
Aerobic Ring Cleavage		Mandelate racemase
Aerobic Ring Cleavage		Carboxymuconolactone decarboxylase
Aerobic Ring Cleavage		Carboxymuconolactone decarboxylase family
Homogentisate Pathway	HMGA	Homogentisate 1,2-dioxygenase
Homogentisate Pathway	HMGA	Homogentisate 1,2-dioxygenase (EC 1.13.11.5)
Homogentisate Pathway	HPCE	Fumarylacetoacetate hydrolase family protein
Homogentisate Pathway	HPCH	Aldolase
Homogentisate Pathway	HPAE	Aldehyde dehydrogenase
Homogentisate Pathway	HPAH	2-oxo-hepta-3-ene-1,7-dioic acid hydratase
Homogentisate Pathway	HPAI	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase EC
Homogentisate Pathway	HPAI	Aldolase
Homogentisate Pathway	HPAF	Leucine-rich repeat-containing protein
Homogentisate Pathway	YCGM	Fumarylacetoacetate (FAA) hydrolase family

(continued from previous page)

Homogentisate Pathway	YCGM	Fumarylacetoacetate (faa) hydrolase
Homogentisate Pathway	YCGM	5-carboxymethyl-2-hydroxymuconate Delta-isomerase (EC 5.3.3.10)
Homogentisate Pathway	MAIA	Maleylacetoacetate isomerase
Homogentisate Pathway		Fumarylacetoacetate hydrolase (fumarylacetoacetase)
Homogentisate Pathway		Fumarylacetoacetate hydrolase
Homogentisate Pathway		Fumarylacetoacetate (FAA) hydrolase family
Anaerobic Ring Cleavage	BADD	2-hydroxyglutaryl-CoA dehydratase subunit D
Anaerobic Ring Cleavage	BADG	Benzoyl-CoA reductase subunit D
Anaerobic Ring Cleavage	HGDA	Benzoyl-CoA reductase subunit
Anaerobic Ring Cleavage	BAMV	PAS PAC sensor signal transduction histidine kinase
Anaerobic Ring Cleavage	BAMW	Sigma54 specific transcriptional regulator Fis family
Anaerobic Ring Cleavage	RDC	Amidohydrolase 2
PHA Synthesis	PHBB	Acetoacetyl-CoA reductase
PHA Synthesis	PHBB	NAD dependent epimerase/dehydratase family
PHA Synthesis	PHAB2	Acetoacetyl-CoA reductase
PHA Synthesis	PHAZ	Polyhydroxyalkanoate depolymerase, intracellular
PHA Synthesis	PHAZ	Depolymerase
PHA Synthesis	PHAZ	Poly(3-hydroxyalkanoate) depolymerase
PHA Synthesis	PHBA	AcetyL-CoA acetyltransferase
PHA Synthesis	PHBA	Acetyl-coa acetyltransferase
PHA Synthesis	ATOB,PHBA	Acetyl-coa acetyltransferase
PHA Synthesis	PHBC	Poly-beta-hydroxybutyrate polymerase domain protein
PHA Synthesis	PHAR	Synthesis repressor, PhaR
PHA Synthesis	PHAB	Enoyl-CoA hydratase
PHA Synthesis	PHAB	KR domain
PHA Synthesis	PHAI	Polyhydroxyalkanoate granule-associated protein
PHA Synthesis	PHAI	Poly granule associated
PHA Synthesis	BKTB	AcetyL-CoA acetyltransferase
PHA Synthesis	BKTB	Acetyl-CoA acetyltransferase
PHA Synthesis		Esterase PHB depolymerase
PHA Synthesis		Dehydrogenase (EC 1.1.1.30)
Trehalose Synthesis	OTSA	Catalyzes the transfer of glucose from UDP-glucose to glucose-6-phosphate to form alpha,alpha-1,1 trehalose-6-phosphate. Acts with retention of the anomeric configuration of the UDP-sugar donor
Trehalose Synthesis	OTSA	Alpha,alpha-trehalose-phosphate synthase
Trehalose Synthesis	OTSA	Alpha,alpha-trehalose-phosphate synthase (EC 2.4.1.15)
Trehalose Synthesis	OTSA	Alpha-alpha-trehalose-phosphate synthase
Trehalose Synthesis	OTSB	Ec 3.1.3.12
Trehalose Synthesis	TREY	Malto-oligosyltrehalose synthase
Trehalose Synthesis	TREY	Maltooligosyl trehalose synthase
Trehalose Synthesis	TREY	Alpha amylase, catalytic domain
Trehalose Synthesis	TREZ	Maltooligosyl trehalose trehalohydrolase
Trehalose Synthesis	TRES	Trehalose synthase
Trehalose Synthesis		Bifunctional 4-alpha-glucanotransferase malto-oligosyltrehalose synthase
Glycogen Synthesis	GLGB	Alpha amylase, C-terminal all-beta domain
Glycogen Synthesis	GLGB	Catalyzes the formation of the alpha-1,6-glucosidic linkages in glycogen by scission of a 1,4-alpha-linked oligosaccharide from growing alpha-1,4-glucan chains and the subsequent attachment of the oligosaccharide to the alpha-1,6 position (By similarity)
Glycogen Synthesis	GLGE	Alpha amylase, catalytic domain
Glycogen Synthesis	GLGE	Maltosyltransferase that uses maltose 1-phosphate (M1P) as the sugar donor to elongate linear or branched alpha-(1- 4)- glucans. Is involved in a branched alpha-glucan biosynthetic pathway from trehalose, together with TreS, Mak and GlgB (By similarity)
Glycogen Synthesis	GLGC	Glucose-1-phosphate adenylyltransferase
Glycogen Synthesis	GLGC	Catalyzes the synthesis of ADP-glucose, a sugar donor used in elongation reactions on alpha-glucans (By similarity)
Glycogen Synthesis	GLGX	Glycogen debranching enzyme
Glycogen Synthesis	GLGX	Glycogen debranching enzyme GlgX
Glycogen Synthesis	GLGA	Synthesizes alpha-1,4-glucan chains using ADP-glucose (By similarity)
Glycogen Synthesis	GLGA	Glycogen synthase
Glycogen Synthesis	GLGP	Phosphorylase is an important allosteric enzyme in carbohydrate metabolism. Enzymes from different sources differ in their regulatory mechanisms and in their natural substrates. However, all known phosphorylases share catalytic and structural properties (By similarity)
Xylose Cleavage	BXLA	Glycoside hydrolase family 3 domain protein
Xylose Cleavage	XYL31A	Hydrolase, family 31

(continued from previous page)

Xylose Cleavage	XYL31A	Glycoside hydrolase family 31
Xylose Cleavage		Acetyl xylan esterase
Xylose Cleavage		Acetyl xylan esterase (AXE1)
Xylose Cleavage	XYNA	Endo-1,4-beta-xylanase (EC 3.2.1.8)
Xylose Cleavage	XYNA	Glyco_10
Methylpentose Cleavage		Alpha-L-arabinofuranosidase B, catalytic
Methylpentose Cleavage		Alpha-L-AF_C
Methylpentose Cleavage		Alpha-L-arabinofuranosidase
Methylpentose Cleavage		Alpha-L-arabinofuranosidase domain protein
Methylpentose Cleavage		Alpha-N-arabinofuranosidase (EC 3.2.1.55)
Methylpentose Cleavage	FUCA1	Alpha-L-fucosidase EC 3.2.1.51
Methylpentose Cleavage	ALFA	Alpha-L-fucosidase EC 3.2.1.51
Methylpentose Cleavage	ALFA	Glycoside hydrolase family 29 (Alpha-L-fucosidase)
Methylpentose Cleavage		Alpha_L_fucos
Methylpentose Cleavage		Glycoside hydrolase family 29 (Alpha-L-fucosidase)
Methylpentose Cleavage		Alpha-L-rhamnosidase
Methylpentose Cleavage		Alpha-L-rhamnosidase N-terminal domain
Methylpentose Cleavage		Inherit from bactNOG: alpha-L-rhamnosidase
Alpha-Galactosidase	MELA	Alpha-galactosidase
Alpha-Galactosidase	MELA	Alpha-galactosidase (EC 3.2.1.22)
Alpha-Galactosidase	AGA	Alpha-galactosidase (EC 3.2.1.22)
Alpha-Galactosidase	AGAB3	Alpha-galactosidase (EC 3.2.1.22)
Alpha-Galactosidase		Glycoside hydrolase 97
Alpha-Galactosidase		Alpha-galactosidase
Mannose Cleavage	SP_2145	Alpha-1,2-mannosidase
Mannose Cleavage	AMS1	Hydrolase, family 38
Mannose Cleavage		Alpha-1,2-mannosidase
Mannose Cleavage		Hydrolase, family 38
Mannose Cleavage		Glycosyl hydrolases family 38 C-terminal domain
Mannose Cleavage		Glycosyl hydrolase family 92
Mannose Cleavage	BMNA	Beta-mannosidase EC 3.2.1.25
Mannose Cleavage	SSCG_01475	Hydrolase, family 26
Mannose Cleavage		Mannan endo-1,4-beta-mannosidase
Beta-Galactoside Catabolism	GALD	Beta-galactosidase
Beta-Galactoside Catabolism	LACZ	Beta-galactosidase
Beta-Galactoside Catabolism	BGAA	Hydrolase family 2, sugar binding
Beta-Galactoside Catabolism	BGAA	Hydrolase, family
Beta-Galactoside Catabolism	BGAT	Beta-galactosidase
Beta-Galactoside Catabolism		Beta-galactosidase
Beta-Galactoside Catabolism		Beta-galactosidase trimerisation domain
Beta-Galactoside Catabolism		Glycosyl hydrolases family 2
Beta-Galactoside Catabolism		Hydrolase family 2, sugar binding
Beta-Galactoside Catabolism		Glycoside hydrolase family 2 sugar binding
Beta-Galactoside Catabolism		Beta-galactosidase EC 3.2.1.23
Beta-Galactoside Catabolism		Glycosyl hydrolase family 2, sugar binding domain protein
Beta-Galactoside Catabolism		Hydrolase, family 2
Beta-Galactoside Catabolism		Glycoside hydrolase family 2, sugar binding
Beta-Galactoside Catabolism		Glycoside hydrolase family 2 sugar binding protein
Beta-Galactoside Catabolism		Inherit from bctoNOG: Beta-galactosidase I
Beta-Glucosidase	BGLB	Glycosyl hydrolase family 3 C-terminal domain
Beta-Glucosidase	BGLB	K05349 beta-glucosidase EC 3.2.1.21
Beta-Glucosidase	BGLB	Glycoside hydrolase, family 3 domain protein
Beta-Glucosidase	BGLB	Beta-glucosidase
Beta-Glucosidase	BGLB	Fn3_like
Beta-Glucosidase	BGLB	Glycoside hydrolase family 3 domain protein
Beta-Glucosidase	BGLX2	Glycoside hydrolase family 3 domain protein
Beta-Glucosidase	BGLX2	Glycoside hydrolase, family 3 domain protein
Beta-Glucosidase	BGLX	K05349 beta-glucosidase EC 3.2.1.21
Beta-Glucosidase	BGLX	Glycoside hydrolase, family 3 domain protein
Beta-Glucosidase	BGXA	Glycoside hydrolase, family 3 domain protein
Beta-Glucosidase	ENC_03470	Beta-glucosidase EC 3.2.1.21
Beta-Glucosidase	BLGA	K05350 beta-glucosidase EC 3.2.1.21

(continued from previous page)

Beta-Glucosidase	BLGA	Ec 3.2.1.21
Beta-Glucosidase	ID880	Inherit from COG: Glycoside hydrolase Family 5
Beta-Glucosidase		Ec 3.2.1.21
Beta-Glucosidase		Glycoside hydrolase family 3 domain protein
Beta-Glucosidase		K05349 beta-glucosidase EC 3.2.1.21
Beta-Glucosidase		Hydrolase, family 3
Beta-Glucosidase		Glycosyl hydrolase family 3 C-terminal domain
Cellulases	CELD	Glycoside hydrolase family 3 domain protein
Cellulases	EGL2	Glycosyl hydrolase family 9
Cellulases		Endoglucanase (EC 3.2.1.4)
Cellulases		Glycoside hydrolase family 5
Cellulases		Cellulase (glycosyl hydrolase family 5)
Cellulases		1,4-beta-cellobiosidase
Cellulases		Glycosyl hydrolase family 9
Cellulases		Cellulose-binding protein
Starch+Glycogen Catabolism	AGLA,AGLA2	Alpha amylase, catalytic region
Starch+Glycogen Catabolism	AGLA	Alpha amylase, catalytic region
Starch+Glycogen Catabolism	AGLA	Trehalose-6-phosphate hydrolase
Starch+Glycogen Catabolism	MALS	Alpha amylase catalytic region
Starch+Glycogen Catabolism	MALL	Alpha amylase, catalytic region
Starch+Glycogen Catabolism	MALL	Trehalose-6-phosphate hydrolase (EC 3.2.1.93)
Starch+Glycogen Catabolism	MALP	Phosphorylase is an important allosteric enzyme in carbohydrate metabolism. Enzymes from different sources differ in their regulatory mechanisms and in their natural substrates. However, all known phosphorylases share catalytic and structural properties (By similarity)
Starch+Glycogen Catabolism	MALZ	Alpha-glucosidase
Starch+Glycogen Catabolism	MALZ	Glycoside hydrolase family 31
Starch+Glycogen Catabolism	MALQ	K00705 4-alpha-glucanotransferase EC 2.4.1.25
Starch+Glycogen Catabolism	MALQ	4-alpha-glucanotransferase (EC 2.4.1.25)
Starch+Glycogen Catabolism	MALQ	4-alpha-glucanotransferase
Starch+Glycogen Catabolism	CGA	Glucan 14-alpha-glucosidase
Starch+Glycogen Catabolism	CGA	Glucan 1,4-alpha-glucosidase (EC 3.2.1.3)
Starch+Glycogen Catabolism	SUSB	Alpha-glucosidase
Starch+Glycogen Catabolism	SUSB	Glycoside hydrolase 97
Starch+Glycogen Catabolism		Glycoside hydrolase 15-related
Starch+Glycogen Catabolism		4-alpha-glucanotransferase (EC 2.4.1.25)
Starch+Glycogen Catabolism		Alpha amylase, catalytic region
Starch+Glycogen Catabolism		Alpha amylase
Starch+Glycogen Catabolism		Alpha amylase, catalytic
Starch+Glycogen Catabolism		AmylO-alpha-16-glucosidase
Starch+Glycogen Catabolism		Amylo-alpha-1,6-glucosidase
Pectinase		Glycosyl hydrolases family 28
Pectinase		Glycoside hydrolase family 28
Pectinase		Glycoside hydrolase, family 28

(continued from previous page)

Pectinase		Pectinesterase (EC 3.1.1.11)
Pectinase		Pectate lyase
Acetylglucosamidase	CHIA	Chitinase (EC 3.2.1.14)
Acetylglucosamidase	CHIA	Chitinase EC 3.2.1.14
Acetylglucosamidase	NAHA	Beta-N-acetylhexosaminidase
Acetylglucosamidase	NAHA	Ec 3.2.1.52
Acetylglucosamidase	NAH	Glycosyl hydrolase family 20, catalytic domain
Acetylglucosamidase	NAH	K12373 hexosaminidase EC 3.2.1.52
Acetylglucosamidase		Hydrolase family 20, catalytic
Acetylglucosamidase		Glycosyl hydrolase family 20, catalytic domain
Nitrogen Fixation	NIFH	The key enzymatic reactions in nitrogen fixation are catalyzed by the nitrogenase complex, which has 2 components the iron protein and the molybdenum-iron protein (By similarity)
Nitrogen Fixation	NIFJ	Oxidoreductase required for the transfer of electrons from pyruvate to flavodoxin (By similarity)
Nitrogen Fixation	NIFU	Nitrogen-fixing NifU domain protein
Nitrogen Fixation	NIFB	Cofactor biosynthesis protein NifB
Nitrogen Fixation	FIXJ	Two component transcriptional regulator, LuxR family
Nitrogen Fixation	FIXK	Transcriptional regulator, Crp Fnr family
Nitrogen Fixation	FIXL	Signal transduction histidine kinase
Nitrogen Fixation	FIXL	Histidine kinase
Nitrate+Nitrite Reduction	NASA	Catalytic subunit of the periplasmic nitrate reductase (NAP). Only expressed at high levels during aerobic growth. NapAB complex receives electrons from the membrane-anchored tetraheme protein NapC, thus allowing electron flow between membrane and periplasm. Essential function for nitrate assimilation and may have a role in anaerobic metabolism (By similarity)
Nitrate+Nitrite Reduction	NASA	Nitrate reductase
Nitrate+Nitrite Reduction	NARG	Nitrate reductase, alpha subunit
Nitrate+Nitrite Reduction	NARI	Respiratory nitrate reductase
Nitrate+Nitrite Reduction	NARX	Nitrate nitrite sensor protein
Nitrate+Nitrite Reduction	NARX	Histidine Kinase
Nitrate+Nitrite Reduction	NARL	Regulator
Nitrate+Nitrite Reduction	NIRA	Sulfite reductase
Nitrate+Nitrite Reduction	NIRB	Nitrite reductase NADPH large subunit
Nitrate+Nitrite Reduction	NIRB	BFD-like [2Fe-2S] binding domain
Nitrate+Nitrite Reduction	NIRB	Nitrite reductase (NAD(P)H) large subunit
Nitrate+Nitrite Reduction	NIRV	Nitrate reductase
Nitrate+Nitrite Reduction	NIRJ	Pyrroloquinoline quinone biosynthesis protein E
Nitrate+Nitrite Reduction	NIRJ	Radical SAM
Nitrate+Nitrite Reduction	NIRM	Cytochrome C, class I
Nitrate+Nitrite Reduction	NIRD	Asnc family transcriptional regulator
Ammonium Transport	AMTB	Ammonium Transporter
Ammonium Transport	AMTB	Ammonium transporter
Ammonium Transport		Ammonium transporter
Ammonia Metabolism	GLNA	Glutamine synthetase
Ammonia Metabolism	GLNA	Glutamine synthetase catalytic region
Ammonia Metabolism	GLNA3	Glutamine synthetase
Ammonia Metabolism	GLNA,GLNN	Glutamine synthetase
Ammonia Metabolism	GLNB	Nitrogen regulatory protein P-II
Ammonia Metabolism	GLNB	Nitrogen regulatory protein PII
Ammonia Metabolism	GLNB	Nitrogen regulatory protein pii
Ammonia Metabolism	GLNB,GLNK	Nitrogen regulatory protein PII
Ammonia Metabolism	GLNK	Nitrogen regulatory protein P-II
Ammonia Metabolism	GLNK	Nitrogen regulatory protein PII
Ammonia Metabolism	GLNK	Nitrogen regulatory protein pii
Ammonia Metabolism	GLNII	Glutamine synthetase
Ammonia Metabolism	GLNII	Glutamine synthetase, beta-Grasp domain
Ammonia Metabolism	GLNN	Glutamine synthetase
Ammonia Metabolism	GDH	Short-chain dehydrogenase reductase SDR
Ammonia Metabolism	GDH	Short-chain dehydrogenase reductase sdr
Ammonia Metabolism	GDH	Dehydrogenase
Ammonia Metabolism	GDHA	Glu/Leu/Phe/Val dehydrogenase, dimerisation domain
Ammonia Metabolism	GDHA	Glutamate dehydrogenase
Ammonia Metabolism	GDHB	Dehydrogenase
Urea Assimilation	URTA	Urea ABC transporter, urea binding protein

(continued from previous page)

Urea Assimilation	URTA	ABC transporter
Urea Assimilation	URTA	Extracellular ligand-binding receptor
Urea Assimilation	URTA	Inherit from bactNOG: ABC, transporter
Urea Assimilation	URTA	Branched-chain amino acid ABC transporter
Urea Assimilation	URTA	(ABC) transporter
Urea Assimilation	URTB2	ABC transporter permease
Urea Assimilation	UREC	Urea amidohydrolase subunit alpha
Urea Assimilation	UREG	Facilitates the functional incorporation of the urease nickel metallocenter. This process requires GTP hydrolysis, probably effectuated by UreG (By similarity)
Urea Assimilation	UREE	Involved in urease metallocenter assembly. Binds nickel. Probably functions as a nickel donor during metallocenter assembly (By similarity)
Urea Assimilation	UREB	Urea amidohydrolase subunit beta
Amino Acid Transport	TCYA	Extracellular solute-binding protein, family 3
Amino Acid Transport	AAPJ	Acid-binding periplasmic protein
Amino Acid Transport	AAPJ	Amino acid ABC transporter substrate-binding protein
Amino Acid Transport	AAPJ	Glutamate glutamine aspartate asparagine ABC transporter, periplasmic substrate-binding protein
Amino Acid Transport	GLTI	Extracellular solute-binding protein
Amino Acid Transport	GLTI	ABC transporter
Amino Acid Transport	BZTA	Glutamate glutamine aspartate asparagine ABC transporter, periplasmic substrate-binding protein
Amino Acid Transport	GLTK	ABC transporter
Amino Acid Transport	FLIY	Cystine transporter subunit
Amino Acid Transport	ARTQ	Transporter permease
Amino Acid Transport	GLNH	Glutamine ABC transporter periplasmic protein
Amino Acid Transport	PUTP	Symporter
Amino Acid Transport	METN	Part of the ABC transporter complex MetNIQ involved in methionine import. Responsible for energy coupling to the transport system (By similarity)
Amino Acid Transport	YHDZ	Abc transporter atp-binding protein
Amino Acid Transport	YHDZ	ABC transporter
Amino Acid Transport	LIVK	Extracellular ligand-binding receptor
Amino Acid Transport	LIVK	(ABC) transporter
Amino Acid Transport	LIVK2	Extracellular ligand-binding receptor
Amino Acid Transport	LIVK2	ABC transporter substrate-binding protein
Amino Acid Transport	BMUL_4261,LIVK2	Extracellular ligand-binding receptor
Amino Acid Transport	LIVJ	Extracellular ligand-binding receptor
Amino Acid Transport	LIVF	Branched-chain amino acid ABC transporter (ATP-binding protein)
Amino Acid Transport	LIVF	ABC transporter
Amino Acid Transport	LIVF	Abc transporter
Amino Acid Transport	LIVH	Branched-chain amino acid ABC transporter (Permease protein)
Amino Acid Transport	LIVH	ABC transporter permease protein
Amino Acid Transport	GLTL	ABC transporter
Amino Acid Transport	KYNA	Catalyzes the oxidative cleavage of the L-tryptophan (L- Trp) pyrrole ring (By similarity)
Amino Acid Transport	YAAJ	Amino acid carrier protein
Amino Acid Transport	SSTT	Involved in the import of serine and threonine into the cell, with the concomitant import of sodium (symport system) (By similarity)
Amino Acid Transport	YDAO	Amino acid
Amino Acid Transport		Polar amino acid uptake family ABC transporter periplasmic substrate-binding protein
Amino Acid Transport		Amino acid ABC transporter
Amino Acid Transport		Amino acid permease-associated region
Amino Acid Transport		ABC transporter periplasmic branched chain amino acid binding protein
Amino Acid Transport		Branched-chain amino acid ABC transporter, periplasmic substrate-binding protein
Amino Acid Transport		Branched-chain amino acid ABC transporter (ATP-binding protein)
Amino Acid Transport		Branched-chain amino acid ABC transporter, periplasmic substrate-binding
Amino Acid Transport		Branched-chain amino acid transporter substrate-binding protein
Amino Acid Transport		High-affinity branched-chain amino acid transport system permease protein
Amino Acid Transport		Inherit from NOG: ABC-type amino acid transport signal transduction systems, periplasmic component domain
Amino Acid Catabolism	TDCB	Threonine dehydratase
Amino Acid Catabolism	SDAA	L-serine dehydratase 1
Amino Acid Catabolism	TNAA	Tryptophanase EC 4.1.99.1
Amino Acid Catabolism	TNAA	L-tryptophan indole-lyase
Amino Acid Catabolism	TNAA	Beta-eliminating lyase
Amino Acid Catabolism	ADI	Decarboxylase
Amino Acid Catabolism	MCCA	CarbamoyL-phosphate synthase l chain ATP-binding

(continued from previous page)

Amino Acid Catabolism	MCCA	Carboxylase, alpha
Amino Acid Catabolism	MCCB	Carboxylase
Amino Acid Catabolism	IVD	Isovaleryl-CoA dehydrogenase
Amino Acid Catabolism	IVD	Dehydrogenase
Amino Acid Catabolism	MMSB	NADP oxidoreductase coenzyme F420-dependent
Amino Acid Catabolism	MMSB	3-hydroxyisobutyrate dehydrogenase (EC 1.1.1.31)
Amino Acid Catabolism	MMSB	3-hydroxyisobutyrate dehydrogenase
Amino Acid Catabolism	MMSB	Dehydrogenase
Amino Acid Catabolism	IBD	Dehydrogenase
Amino Acid Catabolism	PHHA	Phenylalanine 4-monooxygenase
Amino Acid Catabolism	PHHA	Phenylalanine-4-hydroxylase
Amino Acid Catabolism	FAHA	Hydrolase
Amino Acid Catabolism	FAHA	Fumarylacetoacetase EC 3.7.1.2
Amino Acid Catabolism	FAHA	Fumarylacetoacetate (FAA) hydrolase family
Amino Acid Catabolism	FAHA	Fumarylacetoacetase
Amino Acid Catabolism	GABD	Dehydrogenase
Amino Acid Catabolism	GABT	4-aminobutyrate aminotransferase
Amino Acid Catabolism	HUTU	Urocanate hydratase (EC 4.2.1.49)
Amino Acid Catabolism	HUTU	Urocanate hydratase
Amino Acid Catabolism	ANSA	L-asparaginase
Amino Acid Catabolism	ANSA	Asparaginase
Amino Acid Catabolism	ANSA	L-asparaginase (EC 3.5.1.1)
Amino Acid Catabolism	KAMA	Lysine 2,3-aminomutase
Amino Acid Catabolism	MEGL	Methionine gamma-lyase
Amino Acid Catabolism	ASTA	Arginine n-succinyltransferase
Amino Acid Catabolism	ASTE	Transforms N(2)-succinylglutamate into succinate and glutamate (By similarity)
Amino Acid Catabolism	PDH	Proline dehydrogenase
Amino Acid Catabolism	HUTF	N-formimino-l-glutamate deiminase
Amino Acid Catabolism	HUTI	Amidohydrolase family
Amino Acid Catabolism	HUTI	Imidazolonepropionase (EC 3.5.2.7)
Amino Acid Catabolism	HUTI	Imidazolone-5-propionate hydrolase
Amino Acid Catabolism	HCNB	FaD-dependent pyridine nucleotide-disulfide oxidoreductase
Amino Acid Catabolism	HCNB	Bfd domain protein (2fe-2s)-binding domain protein
Amino Acid Catabolism	KAMD	D-lysine 5,6-aminomutase subunit alpha
Amino Acid Catabolism		N-formylglutamate amidohydrolase
Amino Acid Catabolism		3-Hydroxyisobutyrate dehydrogenase
Amino Acid Catabolism		4-aminobutyrate aminotransferase
Amino Acid Catabolism		Beta-eliminating lyase
Oligopeptide Transport	APPF	ABC transporter, ATP-binding protein
Oligopeptide Transport	APPF	Oligopeptide dipeptide abc transporter, atpase subunit
Oligopeptide Transport	APPB	Binding-protein-dependent transport systems inner membrane component
Oligopeptide Transport	APPD	ABC transporter
Oligopeptide Transport	APPA	Bacterial extracellular solute-binding proteins, family 5 Middle
Oligopeptide Transport	APPA	Peptide opine nickel uptake family ABC transporter periplasmic substrate-binding protein
Oligopeptide Transport	OPPA	Family 5
Oligopeptide Transport	OPPA	Oligopeptide ABC transporter system, substrate-binding protein
Oligopeptide Transport	OPPA	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein
Oligopeptide Transport	OPPC	Binding-protein-dependent transport systems inner membrane component
Oligopeptide Transport	OPPC	Permease protein
Oligopeptide Transport	OPPD	ABC, transporter
Oligopeptide Transport	OPPD	(ABC) transporter
Oligopeptide Transport	OPPF	(ABC) transporter
Oligopeptide Transport	DPPD	ABC transporter
Oligopeptide Transport	DPPD	Abc transporter
Oligopeptide Transport	DPPD	(ABC) transporter
Oligopeptide Transport	SCLAV_4611	Peptide transport system secreted peptide binding protein
Oligopeptide Transport		Oligopeptide dipeptide ABC transporter, periplasmic substrate-binding protein
Oligopeptide Transport		Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein
Oligopeptide Transport		Peptide opine nickel uptake family ABC transporter periplasmic substrate-binding protein
Oligopeptide Transport		Peptide transport
Polyamine Transport	POTA	Part of the ABC transporter complex PotABCD involved in spermidine putrescine import. Responsible for energy coupling to the transport system (By similarity)
Polyamine Transport	POTA	ABC transporter

(continued from previous page)

Polyamine Transport	POTH	Binding-protein-dependent transport systems inner membrane component
Polyamine Transport	POTC	Binding-protein-dependent transport systems inner membrane component
Polyamine Transport	POTE5	Amino acid
Polyamine Transport		Part of the ABC transporter complex PotABCD involved in spermidine putrescine import. Responsible for energy coupling to the transport system (By similarity)
Polyamine Transport		ABC spermidine putrescine transporter, periplasmic binding protein
Polyamine Transport		Spermidine putrescine-binding periplasmic protein
Polyamine Transport		POT family
Amino Acid Synthesis	CYSE	Serine o-acetyltransferase
Amino Acid Synthesis	CYSE	Serine acetyltransferase
Amino Acid Synthesis	CYSK	Cysteine synthase
Amino Acid Synthesis	CYSK	Cysteine synthase A
Amino Acid Synthesis	CYSM	Cysteine synthase
Amino Acid Synthesis	GLTB	Glutamate synthase
Amino Acid Synthesis	METH	Methionine synthase
Amino Acid Synthesis	ILVE	Branched-chain-amino-acid aminotransferase
Amino Acid Synthesis	ILVE	Amino acid aminotransferase
Amino Acid Synthesis	ILVE	BrancheD-chain amino acid aminotransferase
Amino Acid Synthesis	ILVE	Branched-chain amino acid aminotransferase
Amino Acid Synthesis	ASPC	Aspartate aminotransferase
Amino Acid Synthesis	ASPC	Aromatic amino acid aminotransferase
Amino Acid Synthesis	ASPC	Cys/Met metabolism PLP-dependent enzyme
Amino Acid Synthesis	ASPC	Aminotransferase class I and II
Amino Acid Synthesis	ASPC	Aminotransferase
Amino Acid Synthesis	ASNB	Asparagine synthase
Amino Acid Synthesis	ASNB	Asparagine synthetase
Amino Acid Synthesis	PROC	Pyrroline-5-carboxylate reductase
Amino Acid Synthesis	PUTA	Bifunctional proline dehydrogenase pyrroline-5-carboxylate dehydrogenase
Amino Acid Synthesis	PUTA	Proline dehydrogenase, pyrroline-5-carboxylate dehydrogenase
Amino Acid Synthesis	PUTA	Proline dehydrogenase
Amino Acid Synthesis	ARGE	Peptidase M20
Amino Acid Synthesis	ARGE	Acetylornithine deacetylase
Amino Acid Synthesis	ARGE	Peptidase
Amino Acid Synthesis	DAP2	Peptidase s9 prolyl oligopeptidase active site domain protein
Amino Acid Synthesis	DAPB	Catalyzes the conversion of 4-hydroxy- tetrahydrodipicolinate (HTPA) to tetrahydrodipicolinate (By similarity)
Amino Acid Synthesis	SERC	Catalyzes the reversible conversion of 3- phosphohydroxypyruvate to phosphoserine and of 3-hydroxy-2-oxo-4- phosphonooxybutanoate to phosphohydroxythreonine (By similarity)
Amino Acid Synthesis	GLYA	Catalyzes the reversible interconversion of serine and glycine with tetrahydrofolate (THF) serving as the one-carbon carrier. This reaction serves as the major source of one-carbon groups required for the biosynthesis of purines, thymidylate, methionine, and other important biomolecules. Also exhibits THF-independent aldolase activity toward beta-hydroxyamino acids, producing glycine and aldehydes, via a retro-aldol mechanism (By similarity)
Amino Acid Synthesis	ILVG	Acetolactate synthase
Amino Acid Synthesis	ILVG	Thiamine pyrophosphate protein
Amino Acid Synthesis	ILVG	Thiamine pyrophosphate
Amino Acid Synthesis	MET17	O-acetylhomoserine O-acetylserine sulfhydrylase
Amino Acid Synthesis	MET17	O-acetylhomoserine
Amino Acid Synthesis	TRPA	The alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3- phosphate (By similarity)
Amino Acid Synthesis	ARGG	Argininosuccinate synthase
Amino Acid Synthesis	ARGG	Citrulline--aspartate ligase
Amino Acid Synthesis	HISF	IGPS catalyzes the conversion of PRFAR and glutamine to IGP, AICAR and glutamate. The HisF subunit catalyzes the cyclization activity that produces IGP and AICAR from PRFAR using the ammonia provided by the HisH subunit (By similarity)
Amino Acid Synthesis	PHEA	Prephenate dehydratase
Amino Acid Synthesis	PHEA	Prephenate dehydratase (EC 4.2.1.51)
Amino Acid Synthesis	PHEA	Chorismate mutase
Amino Acid Synthesis	HISC	Imidazole acetol-phosphate transaminase
Amino Acid Synthesis	OPLAH	5-oxoprolinase (ATP-hydrolyzing)
Amino Acid Synthesis	OPLAH	5-oxoprolinase (EC 3.5.2.9)
Amino Acid Synthesis	DAPE	Peptidase dimerisation domain
Amino Acid Synthesis	DAPE	Peptidase
Amino Acid Synthesis	DAPE	Peptidase, M20

(continued from previous page)

Amino Acid Synthesis	DAPE	Catalyzes the hydrolysis of N-succinyl-L,L- diaminopimelic acid (SDAP), forming succinate and LL-2,6- diaminoheptanedioate (DAP), an intermediate involved in the bacterial biosynthesis of lysine and meso-diaminopimelic acid, an essential component of bacterial cell walls (By similarity)
Amino Acid Synthesis	AATA	Aspartate aminotransferase
Amino Acid Synthesis	HISD	Catalyzes the sequential NAD-dependent oxidations of L- histidinol to L-histidinaldehyde and then to L-histidine (By similarity)
Amino Acid Synthesis	PROB	Catalyzes the transfer of a phosphate group to glutamate to form glutamate 5-phosphate which rapidly cyclizes to 5- oxoproline (By similarity)
Amino Acid Synthesis	ILVD4	Dihydroxy-acid dehydratase
Amino Acid Synthesis	TYRA	Prephenate dehydrogenase
Amino Acid Synthesis	ILVI	Acetolactate synthase
Amino Acid Synthesis	ILVI	Acetolactate synthase large subunit
Amino Acid Synthesis	LTAE	Aldolase
Amino Acid Synthesis	ASNB2	Asparagine synthetase
Amino Acid Synthesis	LEUA2	2-isopropylmalate synthase homocitrate synthase family protein
Amino Acid Synthesis	YFDZ	Aminotransferase
Amino Acid Synthesis	ILVA	Threonine dehydratase
Amino Acid Synthesis	ILVD3	Dihydroxy-acid dehydratase
Amino Acid Synthesis	ARGH	Arginosuccinase
Amino Acid Synthesis	CYSE	Serine o-acetyltransferase
Amino Acid Synthesis	CYSE	Serine acetyltransferase
Amino Acid Synthesis	DAPA2	Dihydrodipicolinate
Amino Acid Synthesis	DAPD	N-succinyltransferase (EC 2.3.1.117)
Amino Acid Synthesis	TRPF	N-(5'-phosphoribosyl)anthranilate isomerase
Amino Acid Synthesis	TRPF	N-(5'phosphoribosyl)anthranilate isomerase
Amino Acid Synthesis	LYS1	Saccharopine dehydrogenase
Amino Acid Synthesis	BMUL_3672	Saccharopine dehydrogenase
Amino Acid Synthesis	HISG	Catalyzes the condensation of ATP and 5-phosphoribose 1- diphosphate to form N'-(5'-phosphoribosyl)-ATP (PR-ATP). Has a crucial role in the pathway because the rate of histidine biosynthesis seems to be controlled primarily by regulation of HisG enzymatic activity (By similarity)
Amino Acid Synthesis	HISH	IGPS catalyzes the conversion of PRFAR and glutamine to IGP, AICAR and glutamate. The HisH subunit provides the glutamine amidotransferase activity that produces the ammonia necessary to HisF for the synthesis of IGP and AICAR (By similarity)
Amino Acid Synthesis	ARGH2	Lyase
Amino Acid Synthesis	ASNB3	Asparagine synthase
Amino Acid Synthesis	HISZ	Required for the first step of histidine biosynthesis. May allow the feedback regulation of ATP phosphoribosyltransferase activity by histidine (By similarity)
Amino Acid Synthesis	HISZ	ATP phosphoribosyltransferase, regulatory subunit
Amino Acid Synthesis	PHEC	Dehydratase (EC
Amino Acid Synthesis	HISB	Imidazoleglycerol-phosphate dehydratase
Amino Acid Synthesis	TRPD	Anthranilate phosphoribosyltransferase
Amino Acid Synthesis	THRC	Threonine synthase
Amino Acid Synthesis	THRC	Catalyzes the gamma-elimination of phosphate from L- phosphohomoserine and the beta-addition of water to produce L- threonine (By similarity)
Amino Acid Synthesis	HISN	Histidinol-phosphate phosphatase
Amino Acid Synthesis	META	Homoserine O-transsuccinylase
Amino Acid Synthesis	METC	Cystathionine beta-lyase
Amino Acid Synthesis	TYRB	Aromatic amino acid aminotransferase
Amino Acid Synthesis	AVTA	Valine-pyruvate transaminase
Amino Acid Synthesis	DAPF	Catalyzes the stereoinversion of LL-2,6- diaminoheptanedioate (L,L-DAP) to meso-diaminoheptanedioate (meso- DAP), a precursor of L-lysine and an essential component of the bacterial peptidoglycan (By similarity)
Amino Acid Synthesis	ALSS	Acetolactate synthase
Amino Acid Synthesis	ASDA	Aspartate aminotransferase
Amino Acid Synthesis	ASNA2	K01444 N4-(beta-N-acetylglucosaminyl)-L-asparaginase EC 3.5.1.26
Amino Acid Synthesis	ILVN	Synthase small subunit
Amino Acid Synthesis	PHEB	Chorismate mutase type II
Amino Acid Synthesis	SERA1	Dehydrogenase
Amino Acid Synthesis	YBDL	Aminotransferase
Amino Acid Synthesis	PABB	Para-aminobenzoate synthase
Amino Acid Synthesis	PABB	Anthranilate synthase component I, N terminal region
Amino Acid Synthesis	PABC	4-amino-4-deoxychorismate lyase
Amino Acid Synthesis	PABC	Aminotransferase

(continued from previous page)

Amino Acid Synthesis	PABC	Aminotransferase class IV
Amino Acid Synthesis	TRPG	Synthase component II
Amino Acid Synthesis		Methionine biosynthesis protein MetW
Amino Acid Synthesis		Cystathionine gamma-synthase
Amino Acid Synthesis		Methionine synthase
Amino Acid Synthesis		Threonine synthase
Amino Acid Synthesis		Threonine dehydratase
Amino Acid Synthesis		Asparagine synthetase
Amino Acid Synthesis		Acetolactate synthase
Amino Acid Synthesis		Saccharopine dehydrogenase
Amino Acid Synthesis		Asparagine synthase
Amino Acid Synthesis		Homoserine dehydrogenase
Amino Acid Synthesis		Methionine synthase (EC 2.1.1.13)
Amino Acid Synthesis		Ornithine cyclodeaminase
Amino Acid Synthesis		Dihydroxyacid dehydratase (EC 4.2.1.9)
Amino Acid Synthesis		Proline racemase
Amino Acid Synthesis		Methionine synthase, vitamin-B12 independent
Amino Acid Synthesis		Prephenate dehydrogenase
Amino Acid Synthesis		Dihydrodipicolinate synthase
Shikimic Acid Pathway	AROG	Phospho-2-dehydro-3-deoxyheptonate aldolase
Shikimic Acid Pathway	AROG	Stereospecific condensation of phosphoenolpyruvate (PEP) and D-erythrose-4-phosphate (E4P) giving rise to 3-deoxy-D- arabino-heptulosonate-7-phosphate (DAHP) (By similarity)
Shikimic Acid Pathway	AROG-1	Phospho-2-dehydro-3-deoxyheptonate aldolase
Shikimic Acid Pathway	AROA	3-phosphoshikimate 1-carboxyvinyltransferase
Shikimic Acid Pathway	AROA	EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)
Shikimic Acid Pathway	AROA	5-enolpyruvylshikimate-3-phosphate synthase
Shikimic Acid Pathway	AROK	Catalyzes the specific phosphorylation of the 3-hydroxyl group of shikimic acid using ATP as a cosubstrate (By similarity)
Shikimic Acid Pathway	AROB	3-dehydroquinate synthase
Shikimic Acid Pathway	AROF	3-deoxy-7-phosphoheptulonate synthase
Shikimic Acid Pathway	AROF	Ec 2.5.1.54
Shikimic Acid Pathway	AROF	Phospho-2-dehydro-3-deoxyheptonate aldolase
Shikimic Acid Pathway	AROC	Chorismate synthase
Shikimic Acid Pathway	AROC	5-enolpyruvylshikimate-3-phosphate phospholyase
Shikimic Acid Pathway	AROQ	Catalyzes a trans-dehydration via an enolate intermediate (By similarity)
Polyamine Synthesis	AGUA	Agmatine deiminase
Polyamine Synthesis	AGUA	Porphyromonas-type peptidyl-arginine deiminase
Polyamine Synthesis	AGUA	Glycosyl hydrolase family 67 C-terminus
Polyamine Synthesis	AGUA	Deiminase
Polyamine Synthesis	AGUB	Nitrilase cyanide hydratase and apolipoprotein N-acyltransferase
Polyamine Synthesis	AGUB	N-carbamoylputrescine amidase
Polyamine Synthesis	AGUB	Carbon-nitrogen hydrolase
Polyamine Synthesis	AGUR	TetR family transcriptional regulator
Polyamine Synthesis	DAT	Diaminobutyrate--2-oxoglutarate aminotransferase
Polyamine Synthesis	DAT	Aminotransferase
Polyamine Synthesis	DDC	Decarboxylase
Polyamine Synthesis	DBDD	Decarboxylase
Polyamine Synthesis	RHBB	Decarboxylase
Polyamine Synthesis	SPEA	Catalyzes the biosynthesis of agmatine from arginine (By similarity)
Polyamine Synthesis	SPEA	Orn DAP Arg decarboxylase 2
Polyamine Synthesis	SPEE	Catalyzes the production of spermidine from putrescine and decarboxylated S-adenosylmethionine (dcSAM), which serves as an aminopropyl donor (By similarity)
Polyamine Synthesis	SPEF	Decarboxylase
Polyamine Synthesis	SPEF	Ornithine decarboxylase
Glycine Cleavage System	GCVA	Transcriptional regulator
Glycine Cleavage System	GCVA	Transcriptional Regulator LysR family
Glycine Cleavage System	GCVP	The glycine cleavage system catalyzes the degradation of glycine. The P protein binds the alpha-amino group of glycine through its pyridoxal phosphate cofactor
Glycine Cleavage System	GCVPA	The glycine cleavage system catalyzes the degradation of glycine. The P protein binds the alpha-amino group of glycine through its pyridoxal phosphate cofactor
Glycine Cleavage System	GCVPB	The glycine cleavage system catalyzes the degradation of glycine. The P protein binds the alpha-amino group of glycine through its pyridoxal phosphate cofactor
Glycine Cleavage System	GCVT	The glycine cleavage system catalyzes the degradation of glycine (By similarity)
Sarcosine Metabolism	SARDH	Fad dependent oxidoreductase

(continued from previous page)

Sarcosine Metabolism	SARDH	Aminomethyl transferase
Sarcosine Metabolism	SOXA	Sarcosine oxidase alpha subunit
Sarcosine Metabolism	SOXA	Sarcosine oxidase (alpha subunit)
Sarcosine Metabolism	SOXA2	Sarcosine oxidase alpha subunit
Sarcosine Metabolism	SOXD	Cytochrome C, class I
Sarcosine Metabolism	SOXD	Sarcosine oxidase delta subunit
Sarcosine Metabolism	HYUB	N-methylhydantoinase B
Sarcosine Metabolism		N-methylhydantoinase B
Phosphate Assimilation	OPRO	Phosphate-Selective Porin O and P
Phosphate Assimilation	OPRO	Polyphosphate-selective porin O
Phosphate Assimilation	OPRO	Phosphate-selective porin O and P
Phosphate Assimilation	YKAA	Phosphate transport regulator
Phosphate Assimilation	PSTS	Phosphate-binding protein
Phosphate Assimilation	PSTS	Phosphate binding protein
Phosphate Assimilation	PSTS	Phosphate ABC transporter substrate-binding protein
Phosphate Assimilation	PSTS	Part of the ABC transporter complex PstSACB involved in phosphate import (By similarity)
Phosphate Assimilation	PSTB	Part of the ABC transporter complex PstSACB involved in phosphate import. Responsible for energy coupling to the transport system (By similarity)
Phosphate Assimilation	PSTC	Phosphate ABC transporter, permease
Phosphate Assimilation	PSTC	Phosphate ABC transporter, permease protein
Phosphate Assimilation	PHOU	Plays a role in the regulation of phosphate uptake (By similarity)
Phosphate Assimilation	PHOU	Plays a role in the regulation of phosphate uptake. Encoded together with proteins of the phosphate-specific transport (Pst) system in the polycistronic pstSCAB-phoU operon (By similarity)
Phosphate Assimilation	PHOU	Part of the phosphate (Pho) regulon, which plays a key role in phosphate homeostasis. Encoded together with proteins of the phosphate-specific transport (Pst) system in the polycistronic pstSCAB-phoU operon. PhoU is essential for the repression of the Pho regulon at high phosphate conditions. In this role, it may bind, possibly as a chaperone, to PhoR, PhoB or a PhoR-PhoB complex to promote dephosphorylation of phospho-PhoB, or inhibit formation of the PhoR-PhoB transitory complex (By similarity)
Phosphate Assimilation	PHOU	Plays a role in the regulation of phosphate uptake
Phosphate Assimilation	PHOU	Plays a role in the regulation of phosphate uptake. In this role, it may bind, possibly as a chaperone, to PhoR, PhoP or a PhoR-PhoP complex to promote dephosphorylation of phospho-PhoP, or inhibit formation of the PhoR-PhoP transitory complex (By similarity)
Phosphate Assimilation	PHOD	Alkaline phosphatase
Phosphate Assimilation	PHOB	Two component transcriptional regulator
Phosphate Assimilation	PHOB	Two component transcriptional regulator (Winged helix family
Phosphate Assimilation	PHOB	Phosphate regulon transcriptional regulatory protein PhoB
Phosphate Assimilation	PHOB	Two component transcriptional regulator, winged helix family
Phosphate Assimilation	PHOH	PhoH family
Phosphate Assimilation	PHOP	Regulator
Phosphate Assimilation	PHOP2	Two component transcriptional regulator (Winged helix family
Phosphate Assimilation	PHOR	Signal transduction histidine kinase
Phosphate Assimilation	PHOR	Integral membrane sensor signal transduction histidine kinase
Phosphate Assimilation	PHOR	Histidine kinase
Phosphate Assimilation	PHOR	Phosphate regulon sensor
Phosphate Assimilation	PITA	Phosphate transporter family
Phosphate Assimilation	PITA	Phosphate transporter
Phosphate Assimilation	PIT	Phosphate transporter
Phosphate Assimilation		Part of the ABC transporter complex PstSACB involved in phosphate import. Responsible for energy coupling to the transport system (By similarity)
Phosphate Assimilation		ABC-type phosphate transport system, periplasmic component
Phosphate Assimilation		Phosphate ABC transporter substrate-binding protein
Phosphate Assimilation		Inherit from COG: phosphate abc transporter
Phosphate Assimilation		PhoD-like phosphatase
Phosphate Assimilation		Phosphate-selective porin O and P
Phosphate Assimilation		Phosphate-Selective porin O and P
Phosphate Assimilation		Inherit from bactNOG: Phosphate-Selective Porin O and P
Phosphate Assimilation		Na Pi-cotransporter
Phosphonate Assimilation	PHNK	Phosphonate C-P lyase system protein PhnK
Phosphonate Assimilation	PHNL	Phosphonate C-P lyase system protein PhnL
Phosphonate Assimilation	PHND	Phosphonate ABC transporter, periplasmic
Phosphonate Assimilation	PHND	Phosphonate ABC transporter, periplasmic phosphonate-binding protein
Phosphonate Assimilation	PHNN	Phosphonate metabolism protein 1,5-bisphosphokinase (PRPP-forming) PhnN
Phosphonate Assimilation	PHNJ	Phosphonate metabolism

(continued from previous page)

Phosphonate Assimilation	PHNI	Phosphonate metabolism
Phosphonate Assimilation	PHNI	Phosphonate metabolism protein
Phosphonate Assimilation	PHNC	ABC, transporter
Phosphonate Assimilation	PHNM2	Alkylphosphonate utilization protein PhnM
Phosphonate Assimilation	PHNA	Phosphonoacetate hydrolase
Polyphosphate Metabolism	PPK	Catalyzes the reversible transfer of the terminal phosphate of ATP to form a long-chain polyphosphate (polyP) (By similarity)
Polyphosphate Metabolism	PPK1	Catalyzes the reversible transfer of the terminal phosphate of ATP to form a long-chain polyphosphate (polyP) (By similarity)
Polyphosphate Metabolism	PAP	Polyphosphate
Polyphosphate Metabolism	PAP	Polyphosphate kinase 2
Polyphosphate Metabolism	PPK2	Polyphosphate AMP phosphotransferase
Polyphosphate Metabolism	PPK2	Polyphosphate kinase 2
Polyphosphate Metabolism	PPK2	Polyphosphate nucleotide phosphotransferase, ppk2 family
Polyphosphate Metabolism		Exopolyphosphatase-related protein
Phosphoglycerol Import	UGPA	Sn-glycerol-3-phosphate transport system, permease protein
Phosphoglycerol Import	UGPA	UTP-glucose-1-phosphate uridylyltransferase
Phosphoglycerol Import	UGPA	Binding-protein-dependent transport systems inner membrane component
Phosphoglycerol Import	UGPB	Extracellular solute-binding protein, family 1
Phosphoglycerol Import	UGPB	Extracellular solute-binding protein
Phosphoglycerol Import	UGPB	Glycerol-3-phosphate transporter periplasmic binding protein
Phosphoglycerol Import	UGPC	ABC transporter
Phosphoglycerol Import	UGPC	(ABC) transporter
Thiosulfate Oxidation	SOXB	Sarcosine oxidase beta subunit
Thiosulfate Oxidation	SOXB	Sulfur oxidation protein
Thiosulfate Oxidation	SOXB	Sulfur oxidation B protein
Thiosulfate Oxidation	SOXB	5'-Nucleotidase domain protein
Thiosulfate Oxidation	SOXC	The exact function is not known. Can catalyze the reduction of a variety of substrates like dimethyl sulfoxide, trimethylamine N-oxide, phenylmethyl sulfoxide and L-methionine sulfoxide. Cannot reduce cyclic N-oxides. Shows no activity as sulfite oxidase (By similarity)
Thiosulfate Oxidation		Thiosulfate reductase cytochrome B subunit (Membrane anchoring protein)
Polysulfide Reduction	NFRD	Polysulfide reductase NrfD
Polysulfide Reduction	NFRD	Polysulphide reductase NrfD
Polysulfide Reduction	NRFC	Iron-sulfur binding
Polysulfide Reduction	NRFC	Fe-S-cluster-containing hydrogenase
Polysulfide Reduction	NRFC	Molybdopterin oxidoreductase, iron-sulfur binding subunit
Polysulfide Reduction	TTRA	Molybdopterin oxidoreductase
Polysulfide Reduction	TTRA	Molydopterin dinucleotide-binding region
Polysulfide Reduction		Polysulphide reductase NrfD
Alkanesulfonate Assimilation	SSUA3	ABC transporter substrate-binding protein
Alkanesulfonate Assimilation		Alkanesulfonate monooxygenase
Taurine Assimilation	TAUB	Abc transporter atp-binding protein
Taurine Assimilation	TAUB	ABC transporter, (ATP-binding protein)
Taurine Assimilation	TAUB	ABC transporter
Taurine Assimilation	TAUB	(ABC) transporter
Taurine Assimilation	TAUA	Taurine ABC transporter, periplasmic binding protein
Taurine Assimilation	TAUA	Taurine ABC transporter, periplasmic
Taurine Assimilation	TAUA	ABC transporter substrate-binding protein
Taurine Assimilation	TAUA2	Solute-binding periplasmic protein of ABC
Taurine Assimilation	BMUL_1604	SyrP protein
Taurine Assimilation	BMUL_1604	Taurine catabolism dioxygenase TauD, TfdA family
Sulfatase	SULFATASE	K01130 arylsulfatase EC 3.1.6.1
Sulfatase	SULFATASE	Arylsulfatase (EC 3.1.6.1)
Sulfatase	ASLA	Sulfatase
Sulfatase	ATSA	Arylsulfatase (EC 3.1.6.1)
Sulfatase	EGTB	Sulfatase-modifying factor enzyme 1
Sulfatase	EGTB	Sulfatase modifying factor
Sulfatase		K01130 arylsulfatase EC 3.1.6.1

(continued from previous page)

Sulfatase		Sulfatase
Sulfatase		Arylsulfatase
Sulfatase		Arylsulfatase (EC 3.1.6.1)
Sulfatase		Sulfatase-modifying factor protein
Sulfatase		Sulfatase-modifying factor enzyme 1
Iron Import	FEOB	Ferrous iron transport protein B
Iron Import	EFEU	Iron permease
Iron Import	FIEF	Cation diffusion facilitator family transporter
Iron Import	HBPA	Extracellular solute-binding protein
Iron Import	HMUP	Hemin uptake protein
Iron Import	FEPA	Outer membrane receptor FepA
Iron Import	CIRA	Receptor
Iron Import	FECA	Receptor
Iron Import	FECR	Anti-FecI sigma factor, FecR
Iron Import	FCUA	Receptor
Iron Import	FCUA	TonB-dependent siderophore receptor
Iron Import	IROD	Esterase
Iron Import	FPVA	Receptor
Iron Import	FBPA	Extracellular solute-binding protein, family 1
Iron Import	FBPA	Iron ABC transporter substrate binding protein
Iron Import	FBPC	Part of the ABC transporter complex FbpABC involved in Fe(3) ions import. Responsible for energy coupling to the transport system (By similarity)
Iron Import	FBPC	ABC transporter
Iron Import	FUR	Uptake regulator, Fur family
Iron Import	FUR	Ferric uptake
Iron Import	FHUE	Receptor
Iron Import	FHUC	ABC transporter
Iron Import	FHUC	Abc transporter
Iron Import	FHUA	Receptor
Iron Import	FHUA	TonB-dependent siderophore receptor
Iron Import	ENTF	Amino acid adenylation domain protein
Iron Import	BASH	Thioesterase
Iron Import	PIUB	Membrane
Iron Import	PIUB	Component of the sulfite reductase complex that catalyzes the 6-electron reduction of sulfite to sulfide. This is one of several activities required for the biosynthesis of L- cysteine from sulfate. The flavoprotein component catalyzes the electron flow from NADPH - FAD - FMN to the hemoprotein component (By similarity)
Iron Import		Ferric enterobactin esterase-related protein Fes
Iron Import		Ferric uptake regulator, Fur family
Iron Import		Tonb-dependent siderophore receptor
Iron Import		Iron siderophore sensor protein
Iron Import		FecR protein
Ferritin	BFR	Iron-storage protein
Ferritin	BFR	Bacterioferritin
Ferritin	BFR3	Ferritin dps family protein
Ferritin	BFR3	Ferritin-like domain
Ferritin	DPS	Ferritin dps family protein
Ferritin	DPS	Ferritin-like domain
Ferritin	DPS	DNA protection during starvation protein
Ferritin		Ferritin Dps family protein
Heme Synthesis	HEMA	5-aminolevulinate synthase
Heme Synthesis	HEMA	Catalyzes the NADPH-dependent reduction of glutamyl- tRNA(Glu) to glutamate 1-semialdehyde (GSA) (By similarity)
Heme Synthesis	HEME	Catalyzes the decarboxylation of four acetate groups of uroporphyrinogen-III to yield coproporphyrinogen-III (By similarity)
Heme Synthesis	HEMB	Delta-aminolevulinic acid dehydratase
Heme Synthesis	HEMN	Coproporphyrinogen III oxidase
Heme Synthesis	HEMN	Oxygen-independent coproporphyrinogen III oxidase
Heme Synthesis	HEMN	Coproporphyrinogen iii oxidase
Heme Synthesis	HEMH	Catalyzes the ferrous insertion into protoporphyrin IX (By similarity)
Heme Synthesis	HEMC	Tetrapolymerization of the monopyrrole PBG into the hydroxymethylbilane pre-uroporphyrinogen in several discrete steps (By similarity)
Heme Synthesis	HEMD	Uroporphyrinogen-III synthase
Heme Synthesis	HEMD	Synthase

(continued from previous page)

Heme Synthesis	HEMG	Protoporphyrinogen oxidase
Heme Synthesis	HEML	Glutamate-1-semialdehyde aminotransferase
Heme Synthesis	HEML	Glutamate-1-semialdehyde 2,1-aminomutase
Heme Synthesis	HEML	Aminotransferase class-III
Heme Synthesis	HEMF	Key enzyme in heme biosynthesis. Catalyzes the oxidative decarboxylation of propionic acid side chains of rings A and B of coproporphyrinogen III (By similarity)
Copper Transport	PACS	Heavy metal translocating p-type ATPase
Copper Transport	PACS	P-type atpase
Copper Transport	PACS	P-type ATPase
Copper Transport	CUSA	Heavy metal efflux pump, CzcA
Copper Transport	CUSA	AcrB/AcrD/AcrF family
Copper Transport	CUSA	Copper silver resistance-related transport membrane protein
Copper Transport	CUSB	Efflux transporter, rnd family, mfp subunit
Copper Transport	CUSB	RND family efflux transporter, MFP subunit
Copper Transport	COPB	P-type atpase
Copper Transport	COPB	Outer membrane efflux protein
Copper Transport	YCNJ	Copper resistance protein CopC
Copper Transport	SCO	Electron transport protein SCO1 SenC
Copper Transport	YEBZ	Copper resistance protein D
Copper Transport		Electron transport protein SCO1 SenC
Potassium Import	KDPB	One of the components of the high-affinity ATP-driven potassium transport (or KDP) system, which catalyzes the hydrolysis of ATP coupled with the exchange of hydrogen and potassium ions (By similarity)
Potassium Import	KDPA	One of the components of the high-affinity ATP-driven potassium transport (or KDP) system, which catalyzes the hydrolysis of ATP coupled with the exchange of hydrogen and potassium ions (By similarity)
Potassium Import		Transport of potassium into the cell (By similarity)
Potassium Import	KUP	Transport of potassium into the cell (By similarity)
Potassium Import	KUP	Transport of potassium into the cell
Potassium Import	TRKH	Low-affinity potassium transport system. Interacts with Trk system potassium uptake protein TrkA (By similarity)
Mercury Resistance	MERP	Ion binding protein
Mercury Resistance	MERR	Transcriptional regulator
Mercury Resistance	MERR	Transcriptional regulator, merr family
Mercury Resistance	MERR1	Merr family transcriptional regulator
Arsenic Resistance	ARSA	Arsenite-activated ATPase (ArsA)
Arsenic Resistance	ARSA	K01130 arylsulfatase EC 3.1.6.1
Arsenic Resistance	ARSA	Arsenical pump-driving ATPase
Arsenic Resistance	YFFB	Arsenate reductase
Arsenic Resistance	ARSR	Transcriptional regulator, arsR family
Arsenic Resistance		Arsenical pump membrane protein
Chromatin Packaging	ACUC	Histone deacetylase
Chromatin Packaging	APHA	Histone deacetylase superfamily
Chromatin Packaging	APHA	Histone deacetylase superfamily protein
Chromatin Packaging	HDA	Histone deacetylase domain
Chromatin Packaging		Core component of nucleosome. Nucleosomes wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template. Histones thereby play a central role in transcription regulation, DNA repair, DNA replication and chromosomal stability. DNA accessibility is regulated via a complex set of post-translational modifications of histones, also called histone code, and nucleosome remodeling
Chromatin Packaging		Swib mdm2 domain-containing protein
Chromatin Packaging		SWIB MDM2 domain
Chromatin Packaging		Histone H2A
Chromatin Packaging		Core component of nucleosome. Nucleosomes wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template. Histones thereby play a central role in transcription regulation, DNA repair, DNA replication and chromosomal stability. DNA accessibility is regulated via a complex set of post-translational modifications of histones, also called histone code, and nucleosome remodeling (By similarity)
Chromatin Packaging		Histone deacetylase
DNA Supercoiling	HUP	Histone family protein DNA-binding
DNA Supercoiling	HUP	DNA-binding protein
DNA Supercoiling	HUPB	DNA-binding protein
DNA Supercoiling	HUPN	DNA-binding protein

(continued from previous page)

DNA Supercoiling	HUP,HUPB	DNA-binding protein
DNA Supercoiling	HUP,HUPA	DNA-binding protein
DNA Supercoiling	HUPA	DNA-binding protein
DNA Supercoiling	GYRA	DNA gyrase negatively supercoils closed circular double- stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings (By similarity)
DNA Supercoiling	GYRB	DNA gyrase negatively supercoils closed circular double- stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings (By similarity)
DNA Supercoiling	GYRA2	DNA topoisomerase IV subunit A
DNA Supercoiling	TOPA	Releases the supercoiling and torsional tension of DNA, which is introduced during the DNA replication and transcription, by transiently cleaving and rejoining one strand of the DNA duplex. Introduces a single-strand break via transesterification at a target site in duplex DNA. The scissile phosphodiester is attacked by the catalytic tyrosine of the enzyme, resulting in the formation of a DNA-(5'-phosphotyrosyl)-enzyme intermediate and the expulsion of a 3'-OH DNA strand. The free DNA strand then undergoes passage around the unbroken strand, thus removing DNA supercoils. Finally, in the religation step, the DNA 3'-OH attacks the covalent intermediate to expel the active-site tyrosine and restore the DNA phosphodiester backbone (By similarity)
DNA Supercoiling	TOPB	DNA topoisomerase
DNA Supercoiling	TOPB	ATP-dependent DNA helicase RecQ
DNA Supercoiling	TOPB	DNA topoisomerase iii
DNA Supercoiling	TOPI	DNA topoisomerase
DNA Replication	DNAE	DNA polymerase III, subunit alpha
DNA Replication	DNAE	DNA polymerase III, alpha subunit
DNA Replication	DNAE	DNA polymerase III alpha subunit
DNA Replication	DNAE	DNA polymerase III subunit alpha
DNA Replication	DNAE2	DNA polymerase III (alpha subunit)
DNA Replication	DNAE2	DNA polymerase involved in damage-induced mutagenesis and translesion synthesis (TLS). It is not the major replicative DNA polymerase (By similarity)
DNA Replication	DNAE2	Bacterial DNA polymerase III alpha subunit
DNA Replication	DNAE2	DNA polymerase III subunit alpha (EC 2.7.7.7)
DNA Replication	DNAN	DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. This DNA polymerase also exhibits 3' to 5' exonuclease activity. The beta chain is required for initiation of replication once it is clamped onto DNA, it slides freely (bidirectional and ATP-independent) along duplex DNA (By similarity)
DNA Replication	DNAG	DNA primase is the polymerase that synthesizes small RNA primers for the Okazaki fragments on both template strands at replication forks during chromosomal DNA synthesis (By similarity)
DNA Replication	DNAB	Replicative DNA helicase
DNA Replication	DNAB	Replicative dna helicase
DNA Replication	DNAX	DNA polymerase iii subunits gamma and tau
DNA Replication	DNAX	Dna polymerase iii subunits gamma and tau
DNA Replication	DNAX	DNA polymerase III subunits gamma and tau
DNA Replication	DNAX	DNA polymerase III, subunits gamma
DNA Replication	DNAX	DNA polymerase III, subunits gamma and tau
DNA Replication	DNAA	It binds specifically double-stranded DNA at a 9 bp consensus (dnaA box) 5'-TTATC CA A CA A-3'. DnaA binds to ATP and to acidic phospholipids (By similarity)
DNA Replication	DNAQ	Helicase
DNA Replication	DNAQ	EXOIII
DNA Replication	DNAQ	DNA polymerase III, epsilon subunit
DNA Replication	DNAQ	DNA polymerase III subunit epsilon
DNA Replication	DNAK	Acts as a chaperone (By similarity)
DNA Replication	DNAJ	ATP binding to DnaK triggers the release of the substrate protein, thus completing the reaction cycle. Several rounds of ATP-dependent interactions between DnaJ, DnaK and GrpE are required for fully efficient folding. Also involved, together with DnaK and GrpE, in the DNA replication of plasmids through activation of initiation proteins (By similarity)
DNA Replication	DARO_0690	RNA-directed DNA polymerase
DNA Replication	HALSA_1121	RNA-directed DNA polymerase
DNA Replication	HDEF_1223	RNA-directed DNA polymerase
DNA Replication	UMUC	DNA-directed DNA polymerase
DNA Replication	ECA3407	DnaG primase-like protein

(continued from previous page)

DNA Replication	DINB	DNA polymerase
DNA Replication	DINB	Poorly processive error-prone DNA polymerase involved in untargeted mutagenesis. Copies undamaged DNA at stalled replication forks which arise in vivo from mismatched or misaligned primer ends. These misaligned primers can be extended by polIV. Exhibits no 3-5 exonuclease (proofreading) activity. May be involved in translesional synthesis in conjunction with the beta clamp from polIII (By similarity)
DNA Replication	DINB	Poorly processive, error-prone DNA polymerase involved in untargeted mutagenesis. Copies undamaged DNA at stalled replication forks, which arise in vivo from mismatched or misaligned primer ends. These misaligned primers can be extended by PolIV. Exhibits no 3'-5' exonuclease (proofreading) activity. May be involved in translesional synthesis, in conjunction with the beta clamp from PolIII (By similarity)
DNA Replication	DING	Helicase
DNA Replication	POLA	Dna polymerase I
DNA Replication	POLA	DNA polymerase I
DNA Replication	POLA	DNA polymerase i
DNA Replication	POLC	Required for replicative DNA synthesis. This DNA polymerase also exhibits 3' to 5' exonuclease activity (By similarity)
DNA Replication	POLC	DNA polymerase III, epsilon subunit
DNA Replication	POLC	Possesses two activities a DNA synthesis (polymerase) and an exonucleolytic activity that degrades single stranded DNA in the 3'- to 5'-direction. Has a template-primer preference which is characteristic of a replicative DNA polymerase (By similarity)
DNA Replication	POLX	PHP domain protein
DNA Replication	POLB	DNA polymerase
DNA Replication	POLB	PHP domain protein
DNA Replication	PRIA	Primosomal protein N'
DNA Replication	PRIA	Primosomal protein n'
DNA Replication	PRIA	Primosomal protein N''
DNA Replication	PRIB	Primosomal replication protein
DNA Replication	HOLA	DNA polymerase III, delta' subunit
DNA Replication	HOLA	DNA polymerase III (Delta subunit)
DNA Replication	HOLB	DNA polymerase III delta prime subunit
DNA Replication	HOLB	DNA polymerase III subunit delta'
DNA Replication	HOLC	Dna polymerase III (Chi subunit)
DNA Replication	HOLC	Dna polymerase iii, chi subunit
DNA Replication	REP	UvrD Rep helicase
DNA Replication	REP	Helicase
DNA Replication	SEQA	Negative regulator of replication initiation, which contributes to regulation of DNA replication and ensures that replication initiation occurs exactly once per chromosome per cell cycle. Binds to pairs of hemimethylated GATC sequences in the oriC region, thus preventing assembly of replication proteins and re- initiation at newly replicated origins. Repression is relieved when the region becomes fully methylated (By similarity)
DNA Replication		DNA primase small subunit
DNA Replication		Bifunctional DNA primase polymerase
DNA Ligase	LIGA	DNA ligase that catalyzes the formation of phosphodiester linkages between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction. It is essential for DNA replication and repair of damaged DNA (By similarity)
DNA Ligase	LIGD	DNA ligase
DNA Ligase	LIGD	ATP-dependent DNA ligase
DNA Ligase	LIGD	ATP dependent DNA ligase C terminal region
DNA Ligase	LIG	ATP-dependent DNA ligase I
DNA Ligase	LIG	DNA ligase
DNA Ligase		DNA ligase
DNA Repair	UVRA	The UvrABC repair system catalyzes the recognition and processing of DNA lesions. UvrA is an ATPase and a DNA-binding protein. A damage recognition complex composed of 2 UvrA and 2 UvrB subunits scans DNA for abnormalities. When the presence of a lesion has been verified by UvrB, the UvrA molecules dissociate (By similarity)
DNA Repair	UVRA	Excinuclease ABC subunit A
DNA Repair	UVRC	The UvrABC repair system catalyzes the recognition and processing of DNA lesions. UvrC both incises the 5' and 3' sides of the lesion. The N-terminal half is responsible for the 3' incision and the C-terminal half is responsible for the 5' incision (By similarity)
DNA Repair	UVRA2	The UvrABC repair system catalyzes the recognition and processing of DNA lesions. UvrA is an ATPase and a DNA-binding protein. A damage recognition complex composed of 2 UvrA and 2 UvrB subunits scans DNA for abnormalities. When the presence of a lesion has been verified by UvrB, the UvrA molecules dissociate (By similarity)
DNA Repair	UVRD	ATP-dependent DNA helicase pcra
DNA Repair	UVRD	Atp-dependent dna helicase
DNA Repair	UVRD	Helicase

(continued from previous page)

DNA Repair	UVRD	ATP-dependent DNA helicase
DNA Repair	UVRD	DNA helicase
DNA Repair	UVRB	Damaged site, the DNA wraps around one UvrB monomer. DNA wrap is dependent on ATP binding by UvrB and probably causes local melting of the DNA helix, facilitating insertion of UvrB beta-hairpin between the DNA strands. Then UvrB probes one DNA strand for the presence of a lesion. If a lesion is found the UvrA subunits dissociate and the UvrB-DNA preincision complex is formed. This complex is subsequently bound by UvrC and the second UvrB is released. If no lesion is found, the DNA wraps around the other UvrB subunit that will check the other stand for damage (By similarity)
DNA Repair	UVRD2	Helicase
DNA Repair	PHRB	Deoxyribodipyrimidine photolyase
DNA Repair	PHRB	Deoxyribodipyrimidine photo-lyase
DNA Repair	RECA	Can catalyze the hydrolysis of ATP in the presence of single-stranded DNA, the ATP-dependent uptake of single-stranded DNA by duplex DNA, and the ATP-dependent hybridization of homologous single-stranded DNAs. It interacts with LexA causing its activation and leading to its autocatalytic cleavage (By similarity)
DNA Repair	RECQ	ATP-dependent DNA helicase RecQ
DNA Repair	RECQ	ATP-dependent DNA helicase, RecQ family
DNA Repair	RECQ	Atp-dependent dna helicase
DNA Repair	RECN	May be involved in recombinational repair of damaged DNA (By similarity)
DNA Repair	RECG	DEAD/DEAH box helicase
DNA Repair	RECG	ATP-dependent DNA helicase RecG
DNA Repair	RECF	It is required for DNA replication and normal SOS inducibility. RecF binds preferentially to single-stranded, linear DNA. It also seems to bind ATP (By similarity)
DNA Repair	RECF2	SMC domain protein
DNA Repair	NTH	Endonuclease III
DNA Repair	NTH	Hhh-gpd family
DNA Repair	MUTS	That it carries out the mismatch recognition step. This protein has a weak ATPase activity (By similarity)
DNA Repair	MUTL	This protein is involved in the repair of mismatches in DNA. It is required for dam-dependent methyl-directed DNA mismatch repair. May act as a molecular matchmaker , a protein that promotes the formation of a stable complex between two or more DNA-binding proteins in an ATP-dependent manner without itself being part of a final effector complex (By similarity)
DNA Repair	MUTM	Involved in base excision repair of DNA damaged by oxidation or by mutagenic agents. Acts as DNA glycosylase that recognizes and removes damaged bases. Has a preference for oxidized purines, such as 7,8-dihydro-8-oxoguanine (8-oxoG). Has AP (apurinic apyrimidinic) lyase activity and introduces nicks in the DNA strand. Cleaves the DNA backbone by beta-delta elimination to generate a single-strand break at the site of the removed base with both 3'- and 5'-phosphates (By similarity)
DNA Repair	MUTM	Formamidopyrimidine-DNA glycosylase N-terminal domain
DNA Repair	MUTM2	Glycosylase
DNA Repair	MUTM2	DNA-(apurinic or apyrimidinic site) lyase formamidopyrimidine-DNA glycosylase
DNA Repair	MUTY	A g-specific adenine glycosylase
DNA Repair	MUTY	HhH-GPD family
DNA Repair	MUTY	A G-specific adenine glycosylase
DNA Repair	MUTT	Mutator MutT protein
DNA Repair	MUTS2	DNA mismatch repair protein MutS
DNA Repair	MUTS2	MutS2 protein
DNA Repair	UDGA	Phage SPO1 DNA polymerase-related protein
DNA Repair	UDGB	Uracil-DNA glycosylase superfamily
DNA Repair	ALKB	2OG-Fe(II) oxygenase
DNA Repair	ALKB	Alkylated DNA repair protein
DNA Repair	ALKB	2og-fe(ii) oxygenase
DNA Repair	ALKA	Glycosylase II
DNA Repair	ALKA	DNA-3-methyladenine glycosylase
DNA Repair	ALKA	Transcriptional regulator
DNA Repair	ALKA	DNA-3-methyladenine glycosylase II transcriptional regulator Ada DNA-O6-methylguanine--protein-cysteine S-methyltransferase
DNA Repair	ALKA	HhH-GPD superfamily base excision DNA repair protein
DNA Repair	MAG	3-methyladenine DNA glycosylase
DNA Repair	MUG	G U mismatch-specific DNA glycosylase
DNA Repair	VSR	DNA mismatch endonuclease
DNA Recombination	RUVA	The RuvA-RuvB complex in the presence of ATP renatures cruciform structure in supercoiled DNA with palindromic sequence, indicating that it may promote strand exchange reactions in homologous recombination. RuvAB is a helicase that mediates the Holliday junction migration by localized denaturation and reannealing. RuvA stimulates, in the presence of DNA, the weak ATPase activity of RuvB (By similarity)

(continued from previous page)

DNA Recombination	RUVB	The RuvA-RuvB complex in the presence of ATP renatures cruciform structure in supercoiled DNA with palindromic sequence, indicating that it may promote strand exchange reactions in homologous recombination. RuvAB is a helicase that mediates the Holliday junction migration by localized denaturation and reannealing (By similarity)
DNA Recombination	RUVC	Nuclease that resolves Holliday junction intermediates in genetic recombination. Cleaves the cruciform structure in supercoiled DNA by nicking to strands with the same polarity at sites symmetrically opposed at the junction in the homologous arms and leaves a 5'-terminal phosphate and a 3'-terminal hydroxyl group (By similarity)
DNA Recombination	RUVX	Could be a nuclease that resolves Holliday junction intermediates in genetic recombination (By similarity)
DNA Recombination	XSEA	Bidirectionally degrades single-stranded DNA into large acid-insoluble oligonucleotides, which are then degraded further into small acid-soluble oligonucleotides (By similarity)
DNA Recombination	XSEB	Bidirectionally degrades single-stranded DNA into large acid-insoluble oligonucleotides, which are then degraded further into small acid-soluble oligonucleotides (By similarity)
DNA Recombination	RECD	Helicase, RecD TraA family
DNA Recombination	RECD	Helicase RecD TraA
DNA Recombination	RECD	Exodeoxyribonuclease v alpha
DNA Recombination	RECJ	Single-stranded-DNA-specific exonuclease (RecJ)
DNA Recombination	RECJ	Single-stranded-DNA-specific exonuclease RecJ
DNA Recombination	RECJ	Exonuclease RecJ
DNA Recombination	RECB	Uvrd rep
DNA Recombination	RECB	Exodeoxyribonuclease V beta subunit
DNA Recombination	RECR	May play a role in DNA repair. It seems to be involved in an RecBC-independent recombinational process of DNA repair. It may act with RecF and RecO (By similarity)
DNA Recombination	RARA	AAA ATPase central domain protein
DNA Recombination	RARA	Recombination factor protein RarA
DNA Recombination	ADDA	Uvrd rep helicase
DNA Recombination	ADDA	Helicase
DNA Recombination	ADDB	Double-strand break repair protein Addb
DNA Recombination	ADDB	The heterodimer acts as both an ATP-dependent DNA helicase and an ATP-dependent, dual-direction single-stranded exonuclease. Recognizes the chi site generating a DNA molecule suitable for the initiation of homologous recombination
DNA Recombination	ADDB	Exonuclease-like protein
DNA Recombination	SXCC_02867	Resolvase
DNA Recombination	YBCK	Resolvase domain protein
DNA Recombination	OCAR_4954	Resolvase
DNA Recombination	TNPR	Plasmid pRiA4b ORF-3 family protein
DNA Recombination	TNPR	Resolvase domain-containing protein
DNA Recombination	BMUL_2472	Resolvase
DNA Recombination	TNPX	Recombinase
DNA Recombination	RADC	DNA repair protein radc
DNA Recombination		Inherit from bactNOG: recb family
Transformation	DPNA	SNF2 family N-terminal domain
Transformation	DPNA	Helicase
Transformation	DPNA	DEXDc
Transformation	COME	Competence protein
Transformation	COMF	Competence protein
Transformation	COMEC	DNA internalization-related competence protein ComEC Rec2
Transformation	COMEC	Competence protein
Transformation	COMEC	ComEC Rec2-related protein
Transformation	DPRA	DNA protecting protein DprA
Transformation	TRAA	Transfer relaxase TraA
Transformation	TRAA	TrwC relaxase
Transformation	TRAI	TrwC protein
Transformation	TRAI	Relaxase mobilization nuclease family protein
Transformation	TRAI	Relaxase/Mobilisation nuclease domain
Transformation	TRAW	Type-F conjugative transfer system protein TraW
Transformation	TRAD	Conjugative transfer protein TraD
Transformation	TRAD	Inherit from proNOG: TRANSFER protein
Transformation	TRAD	Pfam:TraG
Transformation	TRAD	Type IV secretion system protein VirD4

(continued from previous page)

Transformation	TRAG	Conjugal transfer coupling protein TraG
Transformation	TRAG	TraG domain-containing protein
Transformation	TRAG	Conjugative transfer protein TraG
Transformation	TRBG	Transfer protein, trbG
Transformation	VIRB4	Conjugal transfer ATPase
Transformation	VIRB4	Conjugal transfer ATPase TrbE
Transformation	VIRB6	TrbL VirB6 plasmid conjugal transfer protein
Transformation		Conjugative relaxase domain protein
Transformation		Conjugative transfer protein
Transposase	METTU_1963	Transposase
Transposase	TOLA_1058	Transposase
Transposase	TNP	Transposase
Transposase	TNP	Pfam:Transposase_25
Transposase	SSAG_00936	Pfam:Transposase_36
Transposase	SSAG_00936	Transposase
Transposase	LBL_2628	Transposase IS116/IS110/IS902 family
Transposase	LBL_2628	Transposase IS116 IS110 IS902 family protein
Transposase	LBL_2628	Transposase IS116 IS110 IS902
Transposase	LBL_2628	Transposase
Transposase	RV1313C	Transposase, IS204 IS1001 IS1096 IS1165 family protein
Transposase	RV1313C	Transposase IS204 IS1001 IS1096 IS1165 family protein
Transposase	RV1313C	Pfam:Transposase_12
Transposase	TNP3508A	Transposase
Transposase	TNP3508A	Transposase mutator type
Transposase	AM1_0223	Transposase
Transposase	AM1_0223	Transposase, IS4 family protein
Transposase	PARC	DNA topoisomerase IV, subunit A
Transposase	PARC	DNA topoisomerase
Transposase	PARC	DNA topoisomerase (EC 5.99.1.3)
Transposase	ACID_3180	Transposase IS116 IS110 IS902
Transposase	MMC1_0442	Transposase, is66
Transposase	MMC1_0442	Transposase (IS66
Transposase	GLOV_0006	Transposase
Transposase	AJS_0041	Transposase
Transposase	KT99_11013	DDE_Tnp_IS1595
Transposase	KT99_11013	Transposase
Transposase	LBYS_0348	Transposase (IS4 family protein)
Transposase	LBYS_0348	Is4 family
Transposase	MNOD_0308	Transposase
Transposase	SCE2281	Transposase
Transposase	NOCA_1024	Transposase IS116 IS110 IS902 family protein
Transposase	CLIM_0806	Transposase
Transposase	NHAM_0512	Transposase
Transposase	ILYOP_0070	Transposase

(continued from previous page)

Transposase	SPB_1147	Transposase
Transposase	BRADO0294	Transposase TnpC protein
Transposase	BRADO0294	Transposase
Transposase	STROP_2021	Inherit from bactNOG: Transposase
Transposase	INSG	Transposase, IS4 family protein
Transposase	ACID_2273	Transposase
Transposase	KRAC_1754	Transposase, IS4 family protein
Transposase	KRAC_1754	Transposase IS4 family
Transposase	AMBT_05390	Transposase
Transposase	CYCMA_3505	Pfam:Transposase_11
Transposase	CYCMA_3505	Transposase
Transposase	MICAU_1851	Transposase
Transposase	AJS_1107	Transposase
Transposase	BMA1265	Transposase
Transposase	CAUL_0340	Transposase
Transposase	HDEF_0251	Transposase
Transposase	KSE_01030T	Transposase
Transposase	MYPE60	Transposase
Transposase	RPIC_0280	Transposase, IS4
Transposase	SVI_2501	Transposase
Transposase	BT_0485	Transposase
Transposase	BT_0485	Transposase is116 is110 is902 family
Transposase	CLP_0001	Transposase
Transposase	OA238_1743	Transposase
Transposase	TTHA0234	Transposase
Transposase	BIND_0328	Transposase
Transposase	RB5370	Transposase
Transposase	RB5370,SSAG_00936	Transposase
Transposase	TNPA	Transposase
Transposase	TPY_0546	Transposase, Mutator family
Transposase	TPY_0546	Transposase
Transposase	BT_2352	Transposase IS66
Transposase	BT_2352	Pfam:Transposase_25
Transposase	HALHY_0339	Transposase
Transposase	HALHY_0339	Transposase is4
Transposase	PNAP_0427	Transposase
Transposase	SLG_18410	Transposase
Transposase	SLG_18410	Pfam:Transposase_25
Transposase	YAFF	Pfam:Transposase_11
Transposase	ASA_1780	IS630 family transposase

(continued from previous page)

Transposase	AVIN_13510	Inherit from proNOG: transposase
Transposase	CKL_0494,NMUL_A1443	Transposase
Transposase	DMR_01520	Transposase for insertion sequence element
Transposase	HIPMA_0060	Pfam:Transposase_17
Transposase	INSQ	DNA (cytosine-5-)-methyltransferase (EC 2.1.1.37)
Transposase	INSQ	Transposase
Transposase	KRAC_2650	Transposase
Transposase	KRAC_2650	Transposase for insertion sequence
Transposase	MPOP_0102	Transposase (IS4 family protein)
Transposase	MPOP_0102	Transposase IS4 Family Protein
Transposase	RC1_0998	Transposase
Transposase	S7335_436	Transposase
Transposase	SWOO_1064	Transposase IS116 IS110 IS902 family protein
Transposase	SWOO_1064	Transposase
Transposase	DALK_1299	Inherit from bactNOG: Transposase-like protein
Transposase	GK0887	Transposase
Transposase	SHEL_27520	Transposase
Transposase	ACIFE_0396	Transposase
Transposase	AFLV_1427	Transposase
Transposase	GBRO_0815	Transposase
Transposase	HTH_0473	Transposase, IS605 OrfB family
Transposase	PBPRA1820	Pfam:Transposase_25
Transposase	PSYC_0537	Transposase
Transposase	RPE_0533	Transposase
Transposase	VIA_002750	Transposase
Transposase	ALL0363	Transposase
Transposase	BT_1821	Transposase
Transposase	GALF_0323	Transposase (IS4
Transposase	GALF_0323	Transposase, IS4 family protein
Transposase	GURA_2394	Transposase
Transposase	MAE_21040	Transposase
Transposase	MPE_A0854	Transposase, IS4 family protein
Transposase	NAMU_0221	Transposase
Transposase	SLIN_0170	Transposase
Transposase	YFAD	Transposase
Transposase	BCOA_0505	Transposase, IS605 OrfB family

(continued from previous page)

Transposase	BPR_I0156	Transposase
Transposase	CALKR_0444	Transposase
Transposase	JNB_05235	Transposase
Transposase	KRAC_1846	Transposase
Transposase	MAQU_3187	Integrase catalytic subunit
Transposase	MAQU_3187	Transposase
Transposase	MMAR_1396	Transposase for insertion sequence ISMyma02
Transposase	MNOD_2993	Transposase, is4-like protein
Transposase	TNPB	Integrase catalytic subunit
Transposase	TNPB	Transposase
Transposase	ALL0016	Transposase and inactivated derivatives-like
Transposase	RAHAQ_0099	Transposase
Transposase	REIS_0002	Transposase
Transposase	RF_0379	Transposase
Transposase	RPIC_1476	Transposase Tn3 family protein
Transposase	CVAR_0201	Transposase, IS4 family protein
Transposase	DESPR_0301	Transposase
Transposase	DRET_1561	Transposase IS116 IS110 IS902 family protein
Transposase	GDIA_2359	Transposase
Transposase	GURA_0561	Transposase, IS204 IS1001 IS1096 IS1165 family protein
Transposase	GURA_1179	Transposase, IS4
Transposase	GURA_1179	Transposase (IS4,)
Transposase	INSL1,PP_1865	Transposase
Transposase	MICAU_1880	Transposase IS116 IS110 IS902 family protein
Transposase	NMUL_A0047	IS4 family transposase
Transposase	NWI_0954	Transposase
Transposase	R15	Transposase
Transposase	RPE_0249	Transposase
Transposase	YDCC	Transposase
Transposase	YDCC	Transposase IS4 family
Transposase	AFLV_1426	IS630 family transposase
Transposase	CCEL_1484	Transposase
Transposase	CYCMA_3596	Transposase
Transposase	LFERR_0267	Transposase
Transposase	MSC_0172	Transposase
Transposase	NPUN_F2104	Transposase IS4 family

(continued from previous page)

Transposase	PDEN_2092	Transposase
Transposase	PREMU_2024	Transposase, IS204 IS1001 IS1096 IS1165 family protein
Transposase	RPDX1_0336	Transposase IS116 IS110 IS902 family protein
Transposase	YAFM	Inherit from proNOG: transposase
Transposase	BMA1016	ISBma1, transposase
Transposase	MC7420_546	Rhodopirellula transposase family protein
Transposase		K07480 insertion element IS1 protein InsB
Transposase		Transposase
Transposase		Transposase (IS4 family protein)
Transposase		Pfam:Transposase_11
Transposase		Transposase Tn3 family protein
Transposase		Transposase (IS66
Transposase		Transposase domain (DUF772)
Transposase		Inherit from bactNOG: Transposase
Transposase		Transposase, is4 family protein
Transposase		Inherit from bactNOG: transposase IS605 OrfB family
Transposase		Transposase, IS605 OrfB family
Integrase	AFE_0507	Integrase
Integrase	BL0239	Integrase
Integrase	DACE_1327	Integrase, catalytic region
Integrase	MVAN_1091	Integrase core domain
Integrase	NAMU_1215	Integrase
Integrase	SSMG_01709	Integrase catalytic subunit
Integrase	BP2214	Integrase, catalytic region
Integrase	KOLE_1136	Integrase core domain
Integrase	LFERR_0326	Integrase
Integrase	MVAN_0479	Integrase
Integrase	CPAP_0279	Integrase
Integrase	REIS_0088	Integrase catalytic
Integrase	BMUL_0495	Integrase family
Integrase	DSUI_1507	Integrase catalytic subunit
Integrase	SULAZ_0974	Integrase, catalytic region
Integrase	VAPAR_0892	Integrase catalytic subunit
Integrase	M446_0582	Integrase catalytic subunit
Integrase	OA238_1294	Integrase
Integrase	AVA_1330	Integrase catalytic subunit
Integrase	AZL_011680	Integrase catalytic subunit
Integrase	SDEN_1475	Integrase catalytic subunit
Integrase	HRM2_33600,TTHE_0044	Integrase catalytic subunit
Integrase	INTD	Integrase

(continued from previous page)

Integrase	DDES_2232	Phage integrase
Integrase	INTIA	Integrase
Integrase	CKL_0494	Integrase catalytic subunit
Integrase	CKL_0494	Integrase catalytic
Integrase	INTB	Integrase
Integrase	MSMEG_1857	Integrase core domain
Integrase	TMAR_0884	Integrase core domain
Integrase	TTHE_0044	Integrase catalytic subunit
Integrase	TNPS	Phage integrase
Integrase	TNPS	Site-specific recombinase, phage integrase family
Integrase	BL00575	Site-specific recombinase, phage integrase
Integrase	CKL_0494	Integrase catalytic subunit
Integrase	CKL_0494	Integrase catalytic
Integrase	INTB	Integrase
Integrase	MSMEG_1857	Integrase core domain
Integrase	TMAR_0884	Integrase core domain
Integrase	TTHE_0044	Integrase catalytic subunit
Integrase	AVA_1330	Integrase catalytic subunit
Integrase	AZL_011680	Integrase catalytic subunit
Integrase	SDEN_1475	Integrase catalytic subunit
Integrase	DDES_2232	Phage integrase
Integrase	INTIA	Integrase
Integrase	LVIS_1721	Integrase catalytic subunit
Integrase	MAHAU_0136	Integrase catalytic subunit
Integrase	METTU_0272	Integrase
Integrase	NAMU_1237	Integrase
Integrase	KRAC_10383	Integrase catalytic subunit
Integrase	KRAC_10383	Transposase
Integrase	INSI	Integrase catalytic subunit
Integrase	INSI	Integrase, catalytic region
Integrase	MXAN_2168	Integrase catalytic subunit
Integrase	YAGA	Integrase catalytic subunit
Integrase	YAGA	Integrase catalytic
Integrase	MLL5956	Integrase
Integrase	OCAR_6151	Integrase
Integrase	MLL5958	Integrase
Integrase	INT	Phage integrase family protein
Integrase	INT	Integrase family
Integrase	BMUL_2282	Phage integrase family protein
Integrase	BMUL_2282	Integrase

(continued from previous page)

Integrase	ACID_0719	Integrase catalytic subunit
Integrase	ACID_0719	Inherit from bactNOG: Integrase catalytic subunit
Integrase	ISTA	Integrase catalytic subunit
Integrase	ISTA	Transposase
Integrase	INTT	Inherit from proNOG: Integrase
Integrase		Inherit from bactNOG: Integrase
Integrase		Inherit from proNOG: Integrase
Integrase		Inherit from bactNOG: Integrase, catalytic region
Integrase		Phage integrase family protein
Integrase		Phage integrase
Integrase		Phage integrase family
Integrase		Integrase core domain
Integrase		Integrase core domain containing protein
Integrase		Integrase family
Integrase		Integrase, catalytic region
Nucleotide Synthesis	GUAA	Catalyzes the synthesis of GMP from XMP (By similarity)
Nucleotide Synthesis	GUAB	Catalyzes the conversion of inosine 5'-phosphate (IMP) to xanthosine 5'-phosphate (XMP), the first committed and rate- limiting step in the de novo synthesis of guanine nucleotides, and therefore plays an important role in the regulation of cell growth (By similarity)
Nucleotide Synthesis	PYRH	Catalyzes the reversible phosphorylation of UMP to UDP (By similarity)
Nucleotide Synthesis	PYRC	Dihydroorotase EC 3.5.2.3
Nucleotide Synthesis	PYRC	Dihydroorotase, multifunctional complex type
Nucleotide Synthesis	PYRC	Dihydroorotase
Nucleotide Synthesis	PYRC	Dihydropyrimidinase
Nucleotide Synthesis	PYRC	Dihydro-orotase (EC 3.5.2.3)
Nucleotide Synthesis	PYRG	Catalyzes the ATP-dependent amination of UTP to CTP with either L-glutamine or ammonia as the source of nitrogen (By similarity)
Nucleotide Synthesis	PURA	Plays an important role in the de novo pathway of purine nucleotide biosynthesis
Nucleotide Synthesis	PURA	Plays an important role in the de novo pathway of purine nucleotide biosynthesis (By similarity)
Nucleotide Synthesis	PURH	Phosphoribosylaminoimidazolecarboxamide formyltransferase IMP cyclohydrolase
Nucleotide Synthesis	PURH	AICARFT/IMPCHase bienzyme
Nucleotide Synthesis	PURH	Bifunctional purine biosynthesis protein PurH
Nucleotide Synthesis	PURL	Formylglycinamide ribotide synthetase
Nucleotide Synthesis	PURL	Phosphoribosylformylglycinamidine synthase
Nucleotide Synthesis	PURL	Phosphoribosylformylglycinamidine synthase ii
Nucleotide Synthesis	PURL	Phosphoribosylformylglycinamidine synthase (EC 6.3.5.3)
Nucleotide Synthesis	PURL	Phosphoribosylformylglycinamidine synthase II
Nucleotide Synthesis	PURB	Adenylosuccinate lyase
Nucleotide Synthesis	PURM	Phosphoribosylformylglycinamidine cyclo-ligase
Nucleotide Synthesis	PURM	Phosphoribosylaminoimidazole synthetase
Nucleotide Synthesis	PURD	Phosphoribosylglycinamide synthetase
Nucleotide Synthesis	PURK	Phosphoribosylaminoimidazole carboxylase atpase subunit
Nucleotide Synthesis	PURK	Phosphoribosylaminoimidazole carboxylase ATPase subunit
Nucleotide Synthesis	PYRB	Aspartate carbamoyltransferase
Nucleotide Synthesis	PYRB	Aspartate transcarbamylase
Nucleotide Synthesis	PYRE	Catalyzes the transfer of a ribosyl phosphate group from 5-phosphoribose 1-diphosphate to orotate, leading to the formation of orotidine monophosphate (OMP) (By similarity)
Nucleotide Synthesis	PURE	Catalyzes the conversion of N5-carboxyaminoimidazole ribonucleotide (N5-CAIR) to 4-carboxy-5-aminoimidazole ribonucleotide (CAIR) (By similarity)
Nucleotide Synthesis	PYRD	Catalyzes the conversion of dihydroorotate to orotate (By similarity)
Nucleotide Synthesis	PYRD	Catalyzes the conversion of dihydroorotate to orotate with quinone as electron acceptor (By similarity)
Nucleotide Synthesis	PYRD	Catalyzes the conversion of dihydroorotate to orotate
Nucleotide Synthesis	PURC	SAICAR synthetase
Nucleotide Synthesis	PURF	Glutamine amidotransferases class-II
Nucleotide Synthesis	PURF	Amidophosphoribosyltransferase (EC 2.4.2.14)
Nucleotide Synthesis	PURF	Glutamine phosphoribosylpyrophosphate amidotransferase
Nucleotide Synthesis	GUAC	Catalyzes the conversion of inosine 5'-phosphate (IMP) to xanthosine 5'-phosphate (XMP), the first committed and rate- limiting step in the de novo synthesis of guanine nucleotides, and therefore plays an important role in the regulation of cell growth (By similarity)
Nucleotide Synthesis	PYRF	Orotidine 5'-phosphate decarboxylase
Nucleotide Synthesis	PYRF	Catalyzes the decarboxylation of orotidine 5'- monophosphate (OMP) to uridine 5'-monophosphate (UMP) (By similarity)

(continued from previous page)

Nucleotide Synthesis	PURS	Phosphoribosylformylglycinamidine synthase, purS
Nucleotide Synthesis	PURQ	Phosphoribosylformylglycinamidine synthase I
Nucleotide Synthesis	GUAD	Guanine deaminase
Nucleotide Synthesis	GUAD	Deaminase
Nucleotide Synthesis	SPOT	In eubacteria ppGpp (guanosine 3'-diphosphate 5-' diphosphate) is a mediator of the stringent response that coordinates a variety of cellular activities in response to changes in nutritional abundance (By similarity)
Nucleotide Synthesis	NDK	Nucleoside diphosphate kinase
Nucleotide Synthesis	NDK	Major role in the synthesis of nucleoside triphosphates other than ATP. The ATP gamma phosphate is transferred to the NDP beta phosphate via a ping-pong mechanism, using a phosphorylated active-site intermediate (By similarity)
Nucleotide Synthesis	ADK	Catalyzes the reversible transfer of the terminal phosphate group between ATP and AMP. Plays an important role in cellular energy homeostasis and in adenine nucleotide metabolism (By similarity)
Nucleotide Synthesis	PRS	Phosphoribosyl pyrophosphate synthase
Nucleotide Synthesis	TMK	Phosphorylation of dTMP to form dTDP in both de novo and salvage pathways of dTTP synthesis (By similarity)
Nucleotide Synthesis	CMK	Cytidylate kinase
Nucleotide Synthesis	CMK	Cytidine monophosphate kinase
Nucleotide Synthesis	GMK	Essential for recycling GMP and indirectly, cGMP (By similarity)
Nucleotide Synthesis	DGK	Deoxynucleoside kinase
Nucleotide Synthesis		Dihydroorotase
Nucleotide Synthesis		Dihydroorotase (EC 3.5.2.3)
Class I RNR	NRDA	Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides (By similarity)
Class I RNR	NRDB	Reductase, subunit beta
Class I RNR	NRDB	Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides (By similarity)
Class I RNR	NRDR	Negatively regulates transcription of bacterial ribonucleotide reductase nrd genes and operons by binding to NrdR- boxes (By similarity)
Class II RNR	NRDJ	Class II vitamin B12-dependent ribonucleotide reductase
Class II RNR	NRDJ	Reductase
Class II RNR	NRDJ	Ribonucleoside-diphosphate reductase
Class II RNR	NRDJ	Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides (By similarity)
Reverse Transcriptase	HAUR_0135	RNA-directed DNA polymerase
Reverse Transcriptase	DARO_0690	RNA-directed DNA polymerase
Reverse Transcriptase	PPHA_0416	Reverse transcriptase
Reverse Transcriptase	PPHA_0416	RNA-directed DNA polymerase
Reverse Transcriptase	BCELL_1613	RNA-directed DNA polymerase
Reverse Transcriptase	SCE0729	RNA-directed DNA polymerase (Reverse transcriptase)
Reverse Transcriptase	HALSA_1121	RNA-directed DNA polymerase
Reverse Transcriptase	HDEF_1223	RNA-directed DNA polymerase
Reverse Transcriptase	ARAD_7889	RNA-directed DNA polymerase
Reverse Transcriptase	NTHER_0975	RNA-directed DNA polymerase
Reverse Transcriptase	HALSA_1121,STH1346	RNA-directed DNA polymerase
Reverse Transcriptase	PSPTO_2165	RNA-directed DNA polymerase
Reverse Transcriptase		RNA-directed DNA polymerase
Reverse Transcriptase		Reverse transcriptase
RNA Polymerase Machinery	RPOA	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)
RNA Polymerase Machinery	RPOB	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)

(continued from previous page)

RNA Polymerase Machinery	RPOB,RPOC	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)
RNA Polymerase Machinery	RPOC	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)
RNA Polymerase Machinery	RPOZ	Promotes RNA polymerase assembly. Latches the N- and C- terminal regions of the beta' subunit thereby facilitating its interaction with the beta and alpha subunits (By similarity)
RNA Polymerase Machinery	MFD	Transcription-repair coupling factor
RNA Polymerase Machinery	MFD	TRCF domain
RNA Polymerase Machinery		RNA polymerase
DEAD-box RNA Helicase	CSHA	DEAD DEAH box helicase
DEAD-box RNA Helicase	RHLB	ATP-dependent RNA helicase
DEAD-box RNA Helicase	RHLE2	Dead deah box helicase domain protein
DEAD-box RNA Helicase		DEAD DEAH box helicase
DEAD-box RNA Helicase		ATP-dependent RNA helicase, DEAD box family
Essential Transcription Factor	RHO	Facilitates transcription termination by a mechanism that involves Rho binding to the nascent RNA, activation of Rho's RNA-dependent ATPase activity, and release of the mRNA from the DNA template (By similarity)
Essential Transcription Factor	GREA	Necessary for efficient RNA polymerase transcription elongation past template-encoded arresting sites. The arresting sites in DNA have the property of trapping a certain fraction of elongating RNA polymerases that pass through, resulting in locked ternary complexes. Cleavage of the nascent transcript by cleavage factors such as GreA or GreB allows the resumption of elongation from the new 3'terminus. GreA releases sequences of 2 to 3 nucleotides (By similarity)
Essential Transcription Factor	NUSA	NusA antitermination factor
Essential Transcription Factor	NUSA	Transcription elongation factor NusA
Essential Transcription Factor	NUSA	Factor nusa
Essential Transcription Factor	NUSG	Influences transcription termination and antitermination. Acts as a component of the transcription complex, and interacts with the termination factor rho and RNA polymerase (By similarity)
Essential Transcription Factor	NUSG	Transcription termination antitermination protein nusG
Essential Transcription Factor	NUSG	Participates in transcription elongation, termination and antitermination (By similarity)
Essential Transcription Factor	NUSB	Involved in the transcription termination process (By similarity)
Sigma70 Exponential Phase	RPOD	Sigma factors are initiation factors that promote the attachment of RNA polymerase to specific initiation sites and are then released (By similarity)
Sigma54 Nitrogen Limitation	RPON	Sigma factors are initiation factors that promote the attachment of RNA polymerase to specific initiation sites and are then released (By similarity)
Sigma54 Nitrogen Limitation	RPON	RNA polymerase
Sigma54 Nitrogen Limitation		Two component, sigma54 specific, transcriptional regulator, Fis family
Ribosome	RPLL	Seems to be the binding site for several of the factors involved in protein synthesis and appears to be essential for accurate translation (By similarity)
Ribosome	RPSG	One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the head domain of the 30S subunit. Is located at the subunit interface close to the decoding center, probably blocks exit of the E-site tRNA (By similarity)
Ribosome	RPSC	Binds the lower part of the 30S subunit head. Binds mRNA in the 70S ribosome, positioning it for translation (By similarity)
Ribosome	RPSD	One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the body of the 30S subunit (By similarity)
Ribosome	RPSJ	Involved in the binding of tRNA to the ribosomes (By similarity)
Ribosome	RPLE	This is 1 of the proteins that binds and probably mediates the attachment of the 5S RNA into the large ribosomal subunit, where it forms part of the central protuberance. In the 70S ribosome it contacts protein S13 of the 30S subunit (bridge B1b), connecting the 2 subunits
Ribosome	RPLB	One of the primary rRNA binding proteins. Required for association of the 30S and 50S subunits to form the 70S ribosome, for tRNA binding and peptide bond formation. It has been suggested to have peptidyltransferase activity
Ribosome	RPLF	This protein binds to the 23S rRNA, and is important in its secondary structure. It is located near the subunit interface in the base of the L7 L12 stalk, and near the tRNA binding site of the peptidyltransferase center (By similarity)
Ribosome	RPSB	30S ribosomal protein S2

(continued from previous page)

Ribosome	RPSE	Located at the back of the 30S subunit body where it stabilizes the conformation of the head with respect to the body (By similarity)
Ribosome	RPLS	This protein is located at the 30S-50S ribosomal subunit interface and may play a role in the structure and function of the aminoacyl-tRNA binding site (By similarity)
Ribosome	RPSM	Located at the top of the head of the 30S subunit, it contacts several helices of the 16S rRNA. In the 70S ribosome it contacts the 23S rRNA (bridge B1a) and protein L5 of the 50S subunit (bridge B1b), connecting the 2 subunits
Ribosome	RPSI	30S ribosomal protein S9
Ribosome	RPSH	One of the primary rRNA binding proteins, it binds directly to 16S rRNA central domain where it helps coordinate assembly of the platform of the 30S subunit (By similarity)
Ribosome	RPLA	Binds directly to 23S rRNA. The L1 stalk is quite mobile in the ribosome, and is involved in E site tRNA release (By similarity)
Ribosome	RPLO	Binds to the 23S rRNA (By similarity)
Ribosome	RPLO	50S ribosomal protein L15
Ribosome	RPSO	One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it helps nucleate assembly of the platform of the 30S subunit by binding and bridging several RNA helices of the 16S rRNA (By similarity)
Ribosome	RPSO	Forms an intersubunit bridge (bridge B4) with the 23S rRNA of the 50S subunit in the ribosome (By similarity)
Ribosome	RPSO	30S ribosomal protein S15
Ribosome	RPLM	This protein is one of the early assembly proteins of the 50S ribosomal subunit, although it is not seen to bind rRNA by itself. It is important during the early stages of 50S assembly (By similarity)
Ribosome	RPLT	Binds directly to 23S ribosomal RNA and is necessary for the in vitro assembly process of the 50S ribosomal subunit. It is not involved in the protein synthesizing functions of that subunit (By similarity)
Ribosome	RPLJ	50S ribosomal protein L10
Ribosome	RPSP	30S ribosomal protein S16
Ribosome	RPSP	30s ribosomal protein s16
Ribosome	RPSP	30s ribosomal protein S16
Ribosome	RPSA	Thus facilitating recognition of the initiation point. It is needed to translate mRNA with a short Shine-Dalgarno (SD) purine-rich sequence (By similarity)
Ribosome	RPSA	RNA binding S1 domain protein
Ribosome	RPLN	Binds to 23S rRNA. Forms part of two intersubunit bridges in the 70S ribosome (By similarity)
Ribosome	RPLR	This is one of the proteins that binds and probably mediates the attachment of the 5S RNA into the large ribosomal subunit, where it forms part of the central protuberance (By similarity)
Ribosome	RPSS	Protein S19 forms a complex with S13 that binds strongly to the 16S ribosomal RNA (By similarity)
Ribosome	RPLY	This is one of the proteins that binds to the 5S RNA in the ribosome where it forms part of the central protuberance (By similarity)
Ribosome	RPLI	Binds to the 23S rRNA (By similarity)
Ribosome	RPLI	50S ribosomal protein L9
Ribosome	RPSR	Binds as a heterodimer with protein S6 to the central domain of the 16S rRNA, where it helps stabilize the platform of the 30S subunit (By similarity)
Ribosome	RPLD	50S ribosomal protein L4
Ribosome	RPLD	One of the primary rRNA binding proteins, this protein initially binds near the 5'-end of the 23S rRNA. It is important during the early stages of 50S assembly. It makes multiple contacts with different domains of the 23S rRNA in the assembled 50S subunit and ribosome (By similarity)
Ribosome	RPLV	Its binding is stimulated by other ribosomal proteins, e.g. L4, L17, and L20. It is important during the early stages of 50S assembly. It makes multiple contacts with different domains of the 23S rRNA in the assembled 50S subunit and ribosome (By similarity)
Ribosome	RPLV	The globular domain of the protein is located near the polypeptide exit tunnel on the outside of the subunit, while an extended beta-hairpin is found that lines the wall of the exit tunnel in the center of the 70S ribosome (By similarity)
Ribosome	RPSK	Located on the platform of the 30S subunit, it bridges several disparate RNA helices of the 16S rRNA. Forms part of the Shine-Dalgarno cleft in the 70S ribosome (By similarity)
Ribosome	RPSL	Interacts with and stabilizes bases of the 16S rRNA that are involved in tRNA selection in the A site and with the mRNA backbone. Located at the interface of the 30S and 50S subunits, it traverses the body of the 30S subunit contacting proteins on the other side and probably holding the rRNA structure together. The combined cluster of proteins S8, S12 and S17 appears to hold together the shoulder and platform of the 30S subunit (By similarity)
Ribosome	RPMI	50S ribosomal protein l35
Ribosome	RPMI	50S ribosomal protein L35
Ribosome	RPSQ	One of the primary rRNA binding proteins, it binds specifically to the 5'-end of 16S ribosomal
Ribosome	RPLP	Binds 23S rRNA and is also seen to make contacts with the A and possibly P site tRNAs (By similarity)
Ribosome	RPLW	One of the early assembly proteins it binds 23S rRNA. One of the proteins that surrounds the polypeptide exit tunnel on the outside of the ribosome. Forms the main docking site for trigger factor binding to the ribosome (By similarity)

(continued from previous page)

Ribosome	RPLK	This protein binds directly to 23S ribosomal RNA (By similarity)
Ribosome	RPLQ	50S ribosomal protein l17
Ribosome	RPLQ	50S ribosomal protein L17
Ribosome	RPLC	One of the primary rRNA binding proteins, it binds directly near the 3'-end of the 23S rRNA, where it nucleates assembly of the 50S subunit (By similarity)
Ribosome	RPST	Binds directly to 16S ribosomal RNA (By similarity)
Ribosome	RPLX	One of the proteins that surrounds the polypeptide exit tunnel on the outside of the subunit (By similarity)
Ribosome	RPSN	Binds 16S rRNA, required for the assembly of 30S particles and may also be responsible for determining the conformation of the 16S rRNA at the A site (By similarity)
Ribosome	RPSU	30S ribosomal protein S21
Ribosome	RSMA	Specifically dimethylates two adjacent adenosines (A1518 and A1519) in the loop of a conserved hairpin near the 3'-end of 16S rRNA in the 30S particle. May play a critical role in biogenesis of 30S subunits (By similarity)
Ribosome	RSMD	Methyltransferase
Ribosome	RPMF	50S ribosomal protein L32
Ribosome	RPMF	50s ribosomal protein l32
Ribosome	RPMA	50S ribosomal protein L27
Ribosome	RPMC	50s ribosomal protein l29
Ribosome	RPMC	50S ribosomal protein L29
Ribosome	RPSF	Binds together with S18 to 16S ribosomal RNA (By similarity)
Ribosome	RPMJ	50S ribosomal protein L36
Ribosome	RPL2	One of the primary rRNA binding proteins. Required for association of the 30S and 50S subunits to form the 70S ribosome, for tRNA binding and peptide bond formation. It has been suggested to have peptidyltransferase activity
Ribosome	RPMB	50S ribosomal protein L28
Ribosome	RPMB	50S ribosomal protein l28
Ribosome	RPMH	50s ribosomal protein l34
Ribosome	RPS2	30S ribosomal protein S2
Ribosome	RPMD	50S ribosomal protein L30
Ribosome	RPME2	50s ribosomal protein l31
Ribosome	RPS19E	May be involved in maturation of the 30S ribosomal subunit (By similarity)
Ribosome		Ribosomal protein S4/S9 N-terminal domain
Ribosome		Ribosomal protein S20
Ribosome		Ribosomal protein S5
Ribosome		Ribosomal protein S18
Ribosome		40S ribosomal protein
Ribosome		Ribosomal protein L22p/L17e
Ribosome		Ribosomal protein L23, component of cytosolic 80S ribosome and 60S large subunit
Ribosome		Ribosomal protein
Ribosome		SSU ribosomal protein S30P
Ribosome		Mitochondrial 37S ribosomal protein SWS2
Ribosome		Ribosomal protein
tRNA Ligase	ALAS	Catalyzes the attachment of alanine to tRNA(Ala) in a two-step reaction alanine is first activated by ATP to form Ala- AMP and then transferred to the acceptor end of tRNA(Ala). Also edits incorrectly charged Ser-tRNA(Ala) and Gly-tRNA(Ala) via its editing domain (By similarity)
tRNA Ligase	ILES	Amino acids such as valine, to avoid such errors it has two additional distinct tRNA(Ile)-dependent editing activities. One activity is designated as 'pretransfer' editing and involves the hydrolysis of activated Val-AMP. The other activity is designated 'posttransfer' editing and involves deacylation of mischarged Val-tRNA(Ile) (By similarity)
tRNA Ligase	LEUS	Leucyl-tRNA synthetase
tRNA Ligase	VALS	Amino acids such as threonine, to avoid such errors, it has a posttransfer editing activity that hydrolyzes mischarged Thr-tRNA(Val) in a tRNA-dependent manner (By similarity)
tRNA Ligase	THRS	Threonyl-tRNA synthetase
tRNA Ligase	THRS	ThreonyL-tRNA synthetase
tRNA Ligase	PHET	Phenylalanyl-tRNA synthetase, beta subunit
tRNA Ligase	PHET	Phenylalanyl-tRNA synthetase beta subunit
tRNA Ligase	PHET	Phenylalanyl-tRNA synthetase (beta subunit)
tRNA Ligase	PHET	Phenylalanyl-tRNA synthetase subunit beta
tRNA Ligase	LYSS	LysyL-tRNA synthetase
tRNA Ligase	LYSS	Lysyl-tRNA synthetase
tRNA Ligase	ASPS	Aspartyl-tRNA synthetase

(continued from previous page)

tRNA Ligase	SERS	Catalyzes the attachment of serine to tRNA(Ser). Is also able to aminoacylate tRNA(Sec) with serine, to form the misacylated tRNA L-seryl-tRNA(Sec), which will be further converted into selenocysteinyl-tRNA(Sec) (By similarity)
tRNA Ligase	GLYS	Glycyl-tRNA synthetase beta subunit
tRNA Ligase	GLYS	Glycyl-tRNA synthetase (EC 6.1.1.14)
tRNA Ligase	GLYS	Glycyl-tRNA synthetase subunit beta
tRNA Ligase	GLNS	Glutaminyl-tRNA synthetase
tRNA Ligase	GATA	Amidase (EC 3.5.1.4
tRNA Ligase	GATA	K02433 aspartyl-tRNA(Asn) glutamyl-tRNA (Gln) amidotransferase subunit A EC 6.3.5.6 6.3.5.7
tRNA Ligase	GATA	PTS system galactitol-specific transporter subunit IIA
tRNA Ligase	GATA	Allows the formation of correctly charged Gln-tRNA(Gln) through the transamidation of misacylated Glu-tRNA(Gln) in organisms which lack glutaminyl-tRNA synthetase. The reaction takes place in the presence of glutamine and ATP through an activated gamma-phospho-Glu-tRNA(Gln) (By similarity)
tRNA Ligase	GATA	K01426 amidase EC 3.5.1.4
tRNA Ligase	GATA	Amidase
tRNA Ligase	GATA	Amidase EC 3.5.1.4
tRNA Ligase	GATA1	Amidotransferase subunit A
tRNA Ligase	GATB	Allows the formation of correctly charged Asn-tRNA(Asn) or Gln-tRNA(Gln) through the transamidation of misacylated Asp- tRNA(Asn) or Glu-tRNA(Gln) in organisms which lack either or both of asparaginyl-tRNA or glutaminyl-tRNA synthetases. The reaction takes place in the presence of glutamine and ATP through an activated phospho-Asp-tRNA(Asn) or phospho-Glu-tRNA(Gln) (By similarity)
tRNA Ligase	GATB	The phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate active -transport system, catalyzes the phosphorylation of incoming sugar substrates concomitant with their translocation across the cell membrane. This system is involved in galactitol transport
tRNA Ligase	PROS	Catalyzes the attachment of proline to tRNA(Pro) in a two-step reaction proline is first activated by ATP to form Pro- AMP and then transferred to the acceptor end of tRNA(Pro) (By similarity)
tRNA Ligase	PROS	Catalyzes the attachment of proline to tRNA(Pro) in a two-step reaction proline is first activated by ATP to form Pro- AMP and then transferred to the acceptor end of tRNA(Pro). As ProRS can inadvertently accommodate and process non-cognate amino acids such as alanine and cysteine, to avoid such errors it has two additional distinct editing activities against alanine. One activity is designated as 'pretransfer' editing and involves the tRNA(Pro)-independent hydrolysis of activated Ala-AMP. The other activity is designated 'posttransfer' editing and involves deacylation of mischarged Ala-tRNA(Pro). The misacylated Cys- tRNA(Pro) is not edited by ProRS (By similarity)
tRNA Ligase	CYSS	Cysteinyl-tRNA synthetase
tRNA Ligase	AMIDASE	K02433 aspartyl-tRNA(Asn) glutamyl-tRNA (Gln) amidotransferase subunit A EC 6.3.5.6 6.3.5.7
tRNA Ligase	AMIDASE	Allows the formation of correctly charged Gln-tRNA(Gln) through the transamidation of misacylated Glu-tRNA(Gln) in organisms which lack glutaminyl-tRNA synthetase. The reaction takes place in the presence of glutamine and ATP through an activated gamma-phospho-Glu-tRNA(Gln) (By similarity)
tRNA Ligase	TYRS	Catalyzes the attachment of tyrosine to tRNA(Tyr) in a two-step reaction tyrosine is first activated by ATP to form Tyr- AMP and then transferred to the acceptor end of tRNA(Tyr) (By similarity)
tRNA Ligase	ARGS	ArginyL-tRNA synthetase
tRNA Ligase	ARGS	Arginine--tRNA ligase
tRNA Ligase	ARGS	Arginyl-tRNA synthetase
tRNA Ligase	TRPS	Tryptophanyl-tRNA synthetase
tRNA Ligase	ASNS	TRNA synthetases class II (D, K and N)
tRNA Ligase	ASNS	Asparaginyl-tRNA synthetase
tRNA Ligase	PHES	Phenylalanyl-tRNA synthetase alpha subunit
tRNA Ligase	PHES	Phenylalanyl-tRNA synthetase subunit alpha
tRNA Ligase	HISS	Histidyl-tRNA synthetase
tRNA Ligase	GLUQ	Catalyzes the tRNA-independent activation of glutamate in presence of ATP and the subsequent transfer of glutamate onto a tRNA(Asp). Glutamate is transferred on the 2-amino-5-(4,5- dihydroxy-2-cyclopenten-1-yl) moiety of the queuosine in the wobble position of the QUC anticodon (By similarity)
tRNA Ligase	GLTX	Catalyzes the attachment of glutamate to tRNA(Glu) in a two-step reaction glutamate is first activated by ATP to form Glu-AMP and then transferred to the acceptor end of tRNA(Glu) (By similarity)
tRNA Ligase	GLYQ	Glycyl-tRNA synthetase, alpha subunit
tRNA Ligase	GLYQS	Catalyzes the attachment of glycine to tRNA(Gly) (By similarity)
tRNA Ligase		K02433 aspartyl-tRNA(Asn) glutamyl-tRNA (Gln) amidotransferase subunit A EC 6.3.5.6 6.3.5.7
Translation	TSF	Associates with the EF-Tu.GDP complex and induces the exchange of GDP to GTP. It remains bound to the aminoacyl-tRNA.EF- Tu.GTP complex up to the GTP hydrolysis stage on the ribosome (By similarity)
Translation	INFB	One of the essential components for the initiation of protein synthesis. Protects formylmethionyl-tRNA from spontaneous hydrolysis and promotes its binding to the 30S ribosomal subunits. Also involved in the hydrolysis of GTP during the formation of the 70S ribosomal complex (By similarity)

(continued from previous page)

Translation	INFA	However, it seems to stimulate more or less all the activities of the other two initiation factors, IF-2 and IF-3 (By similarity)
Translation	METG	Is required not only for elongation of protein synthesis but also for the initiation of all mRNA translation through initiator tRNA(fMet) aminoacylation (By similarity)
Translation	INFC	IF-3 binds to the 30S ribosomal subunit and shifts the equilibrum between 70S ribosomes and their 50S and 30S subunits in favor of the free subunits, thus enhancing the availability of 30S subunits on which protein synthesis initiation begins (By similarity)
Translation	MIAA	Catalyzes the transfer of a dimethylallyl group onto the adenine at position 37 in tRNAs that read codons beginning with uridine, leading to the formation of N6-(dimethylallyl)adenosine (i(6)A) (By similarity)
Translation	FRR	Responsible for the release of ribosomes from messenger RNA at the termination of protein biosynthesis. May increase the efficiency of translation by recycling ribosomes from one round of translation to another (By similarity)
Translation	RNE	Ribonuclease, Rne Rng family
Translation	RNE	Ribonuclease
Translation	RNE	Ribonuclease E
Translation	PRFA	Peptide chain release factor 1 directs the termination of translation in response to the peptide chain termination codons UAG and UAA (By similarity)
Translation	MAP	Removes the N-terminal methionine from nascent proteins (By similarity)
Translation	FUSA2	Elongation factor g
Translation	FUSA2	Elongation factor G
Translation	FUSA2	EFG_IV
Translation	FUSA2	Translation elongation
Translation	FMT	Possible lysine decarboxylase
Translation	FMT	Modifies the free amino group of the aminoacyl moiety of methionyl-tRNA(fMet). The formyl group appears to play a dual role in the initiator identity of N-formylmethionyl-tRNA by (I) promoting its recognition by IF2 and (II) impairing its binding to EFTu-GTP (By similarity)
Translation	TGT	Exchanges the guanine residue with 7-aminomethyl-7- deazaguanine in tRNAs with GU(N) anticodons (tRNA-Asp, -Asn, -His and -Tyr). After this exchange, a cyclopentendiol moiety is attached to the 7-aminomethyl group of 7-deazaguanine, resulting in the hypermodified nucleoside queuosine (Q) (7-(((4,5-cis- dihydroxy-2-cyclopenten-1-yl)amino)methyl)-7-deazaguanosine) (By similarity)
Translation	TRUA	Formation of pseudouridine at positions 38, 39 and 40 in the anticodon stem and loop of transfer RNAs (By similarity)
Translation	DEF	Removes the formyl group from the N-terminal Met of newly synthesized proteins. Requires at least a dipeptide for an efficient rate of reaction. N-terminal L-methionine is a prerequisite for activity but the enzyme has broad specificity at other positions (By similarity)
Translation	MIAB	Catalyzes the methylthiolation of N6- (dimethylallyl)adenosine (i(6)A), leading to the formation of 2-methylthio-N6-(dimethylallyl)adenosine (ms(2)i(6)A) at position 37 in tRNAs that read codons beginning with uridine (By similarity)
Translation	EFP	Involved in peptide bond synthesis. Stimulates efficient translation and peptide-bond synthesis on native or reconstituted 70S ribosomes in vitro. Probably functions indirectly by altering the affinity of the ribosome for aminoacyl-tRNA, thus increasing their reactivity as acceptors for peptidyl transferase (By similarity)
Translation	EFP	Involved in peptide bond synthesis. Alleviates ribosome stalling that occurs when 3 or more consecutive Pro residues or the sequence PPG is present in a protein, possibly by augmenting the peptidyl transferase activity of the ribosome. Modification of Lys-34 is required for alleviation (By similarity)
Translation	MNMA	Catalyzes the 2-thiolation of uridine at the wobble position (U34) of tRNA, leading to the formation of s(2)U34 (By similarity)
Translation	TRMD	Specifically methylates guanosine-37 in various tRNAs (By similarity)
Translation	YCHF	GTP-dependent nucleic acid-binding protein engD
Translation	YCHF	GTP-binding protein YchF
Translation	MTAB	MiaB-like tRNA modifying enzyme
Translation	DTD	Hydrolyzes D-tyrosyl-tRNA(Tyr) into D-tyrosine and free tRNA(Tyr). Could be a defense mechanism against a harmful effect of D-tyrosine (By similarity)
Translation	RPH	Phosphorolytic exoribonuclease that removes nucleotide residues following the -CCA terminus of tRNA and adds nucleotides to the ends of RNA molecules by using nucleoside diphosphates as substrates (By similarity)
Translation	SELB	Selenocysteine-specific translation elongation factor
Translation	LAST	RNA methyltransferase, TrmH family, group 1
Translation	LAST	Methyltransferase
Translation	SUN	Fmu (Sun) domain protein
Translation	SUN	Nol1 Nop2 Sun family protein
Translation	PTH	The natural substrate for this enzyme may be peptidyl- tRNAs which drop off the ribosome during protein synthesis (By similarity)
Translation	TRMB	Catalyzes the formation of N(7)-methylguanine at position 46 (m7G46) in tRNA (By similarity)

(continued from previous page)

Translation	BMUL_0927	Endoribonuclease L-PSP
Translation	BMUL_3462	N-acetyltransferase
Translation	CCA	Catalyzes the addition and repair of the essential 3'- terminal CCA sequence in tRNAs without using a nucleic acid template. Adds these three nucleotides in the order of C, C, and A to the tRNA nucleotide-73, using CTP and ATP as substrates and producing inorganic pyrophosphate. Also shows phosphatase, 2'- nucleotidase and 2',3'-cyclic phosphodiesterase activities. These phosphohydrolase activities are probably involved in the repair of the tRNA 3'-CCA terminus degraded by intracellular RNases (By similarity)
Translation	CCA	Polynucleotide adenylyltransferase
Translation	CCA	Polynucleotide adenylyltransferase metal dependent phosphohydrolase
Translation	HSLR	RNA-binding S4
Translation	HSLR	RNA-binding S4 domain-containing protein
Translation	PRMA	Methylates ribosomal protein L11 (By similarity)
Translation	PRMC	Methylates the class 1 translation termination release factors RF1 PrfA and RF2 PrfB on the glutamine residue of the universally conserved GGQ motif (By similarity)
Translation	RIMO	Catalyzes the methylthiolation of an aspartic acid residue of ribosomal protein S12 (By similarity)
Translation	RLUA	Pseudouridine synthase
Translation	RPPH	Accelerates the degradation of transcripts by removing pyrophosphate from the 5'-end of triphosphorylated RNA, leading to a more labile monophosphorylated state that can stimulate subsequent ribonuclease cleavage (By similarity)
Translation	SCLAV_0086	Acetyltransferase
Translation	SCLAV_0086	Acetyltransferase (GNAT) family
Translation	TEF1	This protein promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis (By similarity)
Translation	TRUB	Responsible for synthesis of pseudouridine from uracil- 55 in the psi GC loop of transfer RNAs (By similarity)
Translation	TUF3	This protein promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis (By similarity)
Translation	YJGF	Endoribonuclease L-PSP
Translation	DUSB	Catalyzes the synthesis of dihydrouridine, a modified base found in the D-loop of most tRNAs (By similarity)
Translation	MTNA	Catalyzes the interconversion of methylthioribose-1- phosphate (MTR-1-P) into methylthioribulose-1-phosphate (MTRu-1-P) (By similarity)
Translation	RLME	Specifically methylates the uridine in position 2552 of 23S rRNA at the 2'-O position of the ribose in the fully assembled 50S ribosomal subunit (By similarity)
Translation	RLMF	Specifically methylates the adenine in position 1618 of 23S rRNA (By similarity)
Translation	RLMH	Specifically methylates the pseudouridine at position 1915 (m3Psi1915) in 23S rRNA (By similarity)
Translation	RLUD	RNA pseudouridylate synthase
Translation	RLUD	Pseudouridine synthase
Translation	DUSC	Catalyzes the synthesis of dihydrouridine, a modified base found in the D-loop of most tRNAs (By similarity)
Translation	PRFB	Peptide chain release factor 2 directs the termination of translation in response to the peptide chain termination codons UGA and UAA (By similarity)
Translation	MSHC	Catalyzes the ATP-dependent condensation of GlcN-Ins and L-cysteine to form L-Cys-GlcN-Ins (By similarity)
Translation	SMTA	Methyltransferase small
Translation	SMTA	Methyltransferase
Translation	TRUC	Pseudouridine synthase
Translation	YCIH	Translation initiation factor SUI1
Translation	RRMJ	Hemolysin A
Translation	DUSA	Catalyzes the synthesis of dihydrouridine, a modified base found in the D-loop of most tRNAs (By similarity)
Translation	PMRIA	Modifies the free amino group of the aminoacyl moiety of methionyl-tRNA(fMet). The formyl group appears to play a dual role in the initiator identity of N-formylmethionyl-tRNA by (I) promoting its recognition by IF2 and (II) impairing its binding to EFTu-GTP (By similarity)
Translation	QUEA	Transfers and isomerizes the ribose moiety from AdoMet to the 7-aminomethyl group of 7-deazaguanine (preQ1-tRNA) to give epoxyqueuosine (oQ-tRNA) (By similarity)
Translation	RLMD	Catalyzes the formation of 5-methyl-uridine at position 1939 (m5U1939) in 23S rRNA (By similarity)
Translation	CAFA	Ribonuclease, Rne Rng family
Translation	DUS	Catalyzes the synthesis of dihydrouridine, a modified base found in the D-loop of most tRNAs (By similarity)

(continued from previous page)

Translation	PCNB	PolyA polymerase
Translation	RNT	Responsible for the end-turnover of tRNA specifically removes the terminal AMP residue from uncharged tRNA (tRNA-C-C-A). Also appears to be involved in tRNA biosynthesis (By similarity)
Translation	TRMA	Catalyzes the formation of 5-methyl-uridine at position 54 (m5U54) in all tRNAs (By similarity)
Translation	TRML	Could methylate the ribose at the nucleotide 34 wobble position in tRNA (By similarity)
Translation	FUSA	Catalyzes the GTP-dependent ribosomal translocation step during translation elongation. During this step, the ribosome changes from the pre-translocational (PRE) to the post- translocational (POST) state as the newly formed A-site-bound peptidyl-tRNA and P-site-bound deacylated tRNA move to the P and E sites, respectively. Catalyzes the coordinated movement of the two tRNA molecules, the mRNA and conformational changes in the ribosome
Translation	FUSA	Catalyzes the GTP-dependent ribosomal translocation step during translation elongation. During this step, the ribosome changes from the pre-translocational (PRE) to the post- translocational (POST) state as the newly formed A-site-bound peptidyl-tRNA and P-site-bound deacylated tRNA move to the P and E sites, respectively. Catalyzes the coordinated movement of the two tRNA molecules, the mRNA and conformational changes in the ribosome (By similarity)
Translation	RSMH	Specifically methylates the N4 position of cytidine in position 1402 (C1402) of 16S rRNA (By similarity)
Translation	MNMG	NAD-binding protein involved in the addition of a carboxymethylaminomethyl (cmnm) group at the wobble position (U34) of certain tRNAs, forming tRNA-cmnm(5)s(2)U34 (By similarity)
Translation	TILS	Ligates lysine onto the cytidine present at position 34 of the AUA codon-specific tRNA(Ile) that contains the anticodon CAU, in an ATP-dependent manner. Cytidine is converted to lysidine, thus changing the amino acid specificity of the tRNA from methionine to isoleucine (By similarity)
Translation	TTCA	Required for the thiolation of cytidine in position 32 of tRNA, to form 2-thiocytidine (s(2)C32) (By similarity)
Translation	DER	GTPase that plays an essential role in the late steps of ribosome biogenesis (By similarity)
Translation	RIMP	Required for maturation of 30S ribosomal subunits (By similarity)
Translation	MNME	Exhibits a very high intrinsic GTPase hydrolysis rate. Involved in the addition of a carboxymethylaminomethyl (cmnm) group at the wobble position (U34) of certain tRNAs, forming tRNA- cmnm(5)s(2)U34 (By similarity)
Translation	RHLE	DEAD DEAH box helicase domain protein
Translation	RHLE	DEAD DEAH box helicase
Translation	RHLE	Dead deah box helicase domain protein
Translation	RHLE	Helicase
Translation	RHLE	ATP-dependent RNA helicase
Translation	DEAD	Dead deah box helicase domain protein
Translation	DEAD	DEAD DEAH box helicase domain protein
Translation	DEAD	Dead deah box
Translation	DEAD	ATP-dependent RNA helicase
Translation	HRPA	ATP-dependent helicase HrpA
Translation	HRPA	ATP-dependent helicase hrpA
Translation	HRPA	ATP-dependent Helicase
Translation	HRPB	ATP-dependent helicase HrpB
Translation	SRMB	ATP-dependent RNA helicase
Translation		S-adenosylmethionine-dependent methyltransferase
Translation		Associated with ribosomes but is not required for canonical ribosome function and has extra-ribosomal functions. Component of the GAIT (gamma interferon-activated inhibitor of translation) complex which mediates interferon-gamma-induced transcript-selective translation inhibition in inflammation processes. Upon interferon-gamma activation and subsequent phosphorylation dissociates from the ribosome and assembles into the GAIT complex which binds to stem loop-containing GAIT elements in the 3'-UTR of diverse inflammatory mRNAs (such as ceruplasmin) and suppresses their translation. In the GAIT complex interacts with m7G cap-bound eIF4G at or near the eIF3-binding site and blocks the recruitment of the 43S ribosomal complex
Translation		Inherit from arCOG: mRNA 3-end processing factor
Translation		Catalyzes the last two steps in the biosynthesis of 5- methylaminomethyl-2-thiouridine (mnm(5)s(2)U) at the wobble position (U34) in tRNA. Catalyzes the FAD-dependent demodification of cmnm(5)s(2)U34 to nm(5)s(2)U34, followed by the transfer of a methyl group from S-adenosyl-L-methionine to nm(5)s(2)U34, to form mnm(5)s(2)U34 (By similarity)
Selenocysteine Utilization	SELB	Selenocysteine-specific translation elongation factor
Selenocysteine Utilization	SELU	TRNA 2-selenouridine synthase
Selenocysteine Utilization	SELU	Catalyzes the transfer of selenium from selenophosphate for conversion of 2-thiouridine to 2-selenouridine at the wobble position in tRNA (By similarity)
Selenocysteine Utilization	SELA	Converts seryl-tRNA(Sec) to selenocysteinyl-tRNA(Sec) required for selenoprotein biosynthesis (By similarity)
Selenocysteine Utilization	SELD	Selenophosphate synthase
Selenocysteine Utilization	SELD	AIR synthase related protein, N-terminal domain

(continued from previous page)

Selenocysteine Utilization	SELD	Synthesizes selenophosphate from selenide and ATP (By similarity)
Selenocysteine Utilization		Selenocysteine synthase (Seryl-tRNASer selenium transferase)
PNPase	PNP	Involved in mRNA degradation. Hydrolyzes single-stranded polyribonucleotides processively in the 3'- to 5'-direction (By similarity)
Phospholipid Synthesis	CFA	Cyclopropane-fatty-acyl-phospholipid synthase
Phospholipid Synthesis	FABA	Necessary for the introduction of cis unsaturation into fatty acids. Catalyzes the dehydration of (3R)-3-hydroxydecanoyl- ACP to E-(2)-decenoyl-ACP and then its isomerization to Z-(3)- decenoyl-ACP. Can catalyze the dehydratase reaction for beta- hydroxyacyl-ACPs with saturated chain lengths up to 16 0, being most active on intermediate chain length (By similarity)
Phospholipid Synthesis	FABB	Catalyzes the condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP (By similarity)
Phospholipid Synthesis	FABD	Malonyl CoA-acyl carrier protein transacylase
Phospholipid Synthesis	FABG	3-oxoacyl-(Acyl-carrier-protein) reductase
Phospholipid Synthesis	FABG	Reductase
Phospholipid Synthesis	FABG	3-oxoacyl-acyl-carrier-protein reductase
Phospholipid Synthesis	FABG1	Reductase
Phospholipid Synthesis	FABG-1	Short-chain dehydrogenase reductase SDR
Phospholipid Synthesis	FABG2	Short-chain dehydrogenase reductase SDR
Phospholipid Synthesis	FABG2	Reductase
Phospholipid Synthesis	FABG4	Short chain dehydrogenase
Phospholipid Synthesis	FABF	Catalyzes the condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP (By similarity)
Phospholipid Synthesis	FABF	Synthase
Phospholipid Synthesis	FABF2	3-oxoacyl-(Acyl carrier protein) synthase II
Phospholipid Synthesis	FABF2	Synthase ii
Phospholipid Synthesis	FABH	Catalyzes the condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP. Catalyzes the first condensation reaction which initiates fatty acid synthesis and may therefore play a role in governing the total rate of fatty acid production. Possesses both acetoacetyl-ACP synthase and acetyl transacylase activities. Its substrate specificity determines the biosynthesis of branched- chain and or straight-chain of fatty acids (By similarity)
Phospholipid Synthesis	FABH	Synthase
Phospholipid Synthesis	FABI	Enoyl- acyl-carrier-protein reductase NADH
Phospholipid Synthesis	PFAA	Synthase
Phospholipid Synthesis	TGS1	Acyltransferase WS DGAT MGAT
Phospholipid Synthesis	DGKA	Diacylglycerol kinase
Phospholipid Synthesis	MT3314	Diacylglycerol kinase, catalytic region
Phospholipid Synthesis		Diacylglycerol kinase
Phospholipid Synthesis		Monogalactosyldiacylglycerol synthase
UDP-GlcNAc Synthesis	GLMS	Catalyzes the first step in hexosamine metabolism, converting fructose-6P into glucosamine-6P using glutamine as a nitrogen source (By similarity)
UDP-GlcNAc Synthesis	GLMS2	Glutamine-fructose-6-phosphate transaminase
UDP-GlcNAc Synthesis	GLMS2	Glutamine--fructose-6-phosphate transaminase (isomerizing)
UDP-GlcNAc Synthesis	GLMM	Catalyzes the conversion of glucosamine-6-phosphate to glucosamine-1-phosphate (By similarity)
UDP-GlcNAc Synthesis	GLMU	Catalyzes the last two sequential reactions in the de novo biosynthetic pathway for UDP-N-acetylglucosamine (UDP- GlcNAc). The C-terminal domain catalyzes the transfer of acetyl group from acetyl coenzyme A to glucosamine-1-phosphate (GlcN-1-P) to produce N-acetylglucosamine-1-phosphate (GlcNAc-1-P), which is converted into UDP-GlcNAc by the transfer of uridine 5-monophosphate (from uridine 5-triphosphate), a reaction catalyzed by the N-terminal domain (By similarity)
Peptidoglycan Synthesis	MURA	Cell wall formation. Adds enolpyruvyl to UDP-N- acetylglucosamine (By similarity)
Peptidoglycan Synthesis	MURB	ATP-dependent carboxylate-amine ligase (By similarity)
Peptidoglycan Synthesis	MURB	Pfam:DUF404
Peptidoglycan Synthesis	MURB	Cell wall formation (By similarity)
Peptidoglycan Synthesis	MURC	Cell wall formation (By similarity)
Peptidoglycan Synthesis	MURI	Provides the (R)-glutamate required for cell wall biosynthesis (By similarity)
Peptidoglycan Synthesis	MURD	Cell wall formation. Catalyzes the addition of glutamate to the nucleotide precursor UDP-N-acetylmuramoyl-L-alanine (UMA) (By similarity)
Peptidoglycan Synthesis	MURE	Catalyzes the addition of an amino acid to the nucleotide precursor UDP-N-acetylmuramoyl-L-alanyl-D-glutamate (UMAG) in the biosynthesis of bacterial cell-wall peptidoglycan (By similarity)
Peptidoglycan Synthesis	MURE	Catalyzes the addition of meso-diaminopimelic acid to the nucleotide precursor UDP-N-acetylmuramoyl-L-alanyl-D-glutamate (UMAG) in the biosynthesis of bacterial cell-wall peptidoglycan (By similarity)
Peptidoglycan Synthesis	MURE	Acid to the nucleotide precursor UDP-N-acetylmuramoyl-L-alanyl-D-glutamate (UMAG) in the biosynthesis of bacterial cell-wall peptidoglycan (By similarity)

(continued from previous page)

Peptidoglycan Synthesis	MURF	Involved in cell wall formation. Catalyzes the final step in the synthesis of UDP-N-acetylmuramoyl-pentapeptide, the precursor of murein (By similarity)
Peptidoglycan Synthesis	DDL	Cell wall formation (By similarity)
Peptidoglycan Synthesis	MURG	Cell wall formation. Catalyzes the transfer of a GlcNAc subunit on undecaprenyl-pyrophosphoryl-MurNAc-pentapeptide (lipid intermediate I) to form undecaprenyl-pyrophosphoryl-MurNAc-(pentapeptide)GlcNAc (lipid intermediate II) (By similarity)
Peptidoglycan Synthesis	MTGA	Monofunctional biosynthetic peptidoglycan transglycosylase
Peptidoglycan Synthesis	PBP1A	Penicillin-binding protein
Peptidoglycan Synthesis	PBP2B	Penicillin-binding protein
Peptidoglycan Synthesis	MRCA	Penicillin-binding protein 1A
Peptidoglycan Synthesis	MRCA	Peptidoglycan glycosyltransferase (EC 2.4.1.129)
Peptidoglycan Synthesis	MRCB	Penicillin-binding protein 1B
Peptidoglycan Synthesis	MRCB	Penicillin-binding protein 1A
Peptidoglycan Synthesis	MRDA	Penicillin-binding protein 2
Peptidoglycan Synthesis	MRDA	Penicillin-binding protein
Peptidoglycan Synthesis	YCBB	Dolichyl-phosphate beta-D-mannosyltransferase (EC 2.4.1.83)
Peptidoglycan Synthesis	YCBB	Peptidoglycan binding domain-containing protein
Peptidoglycan Synthesis	YCBB	ErfK YbiS YcfS YnhG family protein
Peptidoglycan Synthesis	YBIS	ErfK YbiS YcfS YnhG
Peptidoglycan Synthesis	YBIS	ErfK YbiS YcfS YnhG family protein
Peptidoglycan Synthesis	DACA	D-alanyl-d-alanine carboxypeptidase
Peptidoglycan Synthesis	DACA	Carboxypeptidase
Peptidoglycan Synthesis	DACB	D-alanyl-D-alanine carboxypeptidase
Peptidoglycan Synthesis	DACC	Ec 3.4.16.4
Peptidoglycan Synthesis	DACF	Carboxypeptidase
Peptidoglycan Synthesis	PRC	C-terminal domain of tail specific protease (DUF3340)
Peptidoglycan Synthesis	PRC	Carboxyl-terminal protease (EC 3.4.21.102)
Peptidoglycan Synthesis	PRC	Carboxyl-terminal protease
Peptidoglycan Synthesis	MLTB	Lytic Murein transglycosylase
Peptidoglycan Synthesis	MLTB	Lytic murein transglycosylase
Peptidoglycan Synthesis	MLTC	Transglycosylase
Peptidoglycan Synthesis	MLTC	Murein-degrading enzyme. May play a role in recycling of muropeptides during cell elongation and or cell division
Peptidoglycan Synthesis	MLTD2	Lytic transglycosylase
Peptidoglycan Synthesis	PONA	Peptidoglycan glycosyltransferase
Peptidoglycan Synthesis	PONA	Penicillin-binding protein 1A
Peptidoglycan Synthesis		Lytic transglycosylase
Peptidoglycan Synthesis		NLP P60 protein
Peptidoglycan Synthesis		PBPb
Core-Lipid A Synthesis	LAPB	ABC transporter
Core-Lipid A Synthesis	LPSA	Lipopolysaccharide A protein
Core-Lipid A Synthesis	LPXA	Involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell (By similarity)
Core-Lipid A Synthesis	LPXB	Condensation of UDP-2,3-diacylglucosamine and 2,3- diacylglucosamine-1-phosphate to form lipid A disaccharide, a precursor of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell (By similarity)
Core-Lipid A Synthesis	LPXC	Involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell (By similarity)
Core-Lipid A Synthesis	LPXD	Catalyzes the N-acylation of UDP-3-O-acylglucosamine using 3-hydroxyacyl-ACP as the acyl donor. Is involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell (By similarity)
Core-Lipid A Synthesis	LPXF	Phosphoesterase, PA-phosphatase related
Core-Lipid A Synthesis	KDSA	Phospho-2-dehydro-3-deoxyoctonate aldolase
Core-Lipid A Synthesis	KDSB	Activates KDO (a required 8-carbon sugar) for incorporation into bacterial lipopolysaccharide in Gram-negative bacteria (By similarity)
Core-Lipid A Synthesis	KDSC	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase, YrbI family
Core-Lipid A Synthesis	KDSC	3-deoxy-d-manno-octulosonate 8-phosphate phosphatase
Core-Lipid A Synthesis	KDSC	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase
Core-Lipid A Synthesis	KDSD	Arabinose 5-phosphate isomerase
Core-Lipid A Synthesis	KDSD	KpsF GutQ family protein
Core-Lipid A Synthesis	KDTA	3-Deoxy-D-manno-octulosonic-acid transferase (kdotransferase)
Core-Lipid A Synthesis	KDTA	3-Deoxy-D-manno-octulosonic-acid transferase
Core-Lipid A Synthesis	KDTA	Transferase
Core-Lipid A Synthesis	WAAG	Glycosyl transferase group 1

(continued from previous page)

Core-Lipid A Synthesis	GALU	UTP-glucose-1-phosphate uridylyltransferase
Core-Lipid A Synthesis	GALU	Utp--glucose-1-phosphate uridylyltransferase
Core-Lipid A Synthesis	GMHA	Catalyzes the isomerization of sedoheptulose 7-phosphate in D-glycero-D-manno-heptose 7-phosphate (By similarity)
Core-Lipid A Synthesis	GMHB	D,D-heptose 1,7-bisphosphate phosphatase
Core-Lipid A Synthesis	GMHB	D,d-heptose 1,7-bisphosphate phosphatase
Core-Lipid A Synthesis	HLDD	Catalyzes the interconversion between ADP-D-glycero- beta-D-manno-heptose and ADP-L-glycero-beta-D-manno-heptose via an epimerization at carbon 6 of the heptose (By similarity)
Core-Lipid A Synthesis	HLDE	Cytidylyltransferase
Core-Lipid A Synthesis	HLDE	Bifunctional protein
Core-Lipid A Synthesis	HLDE	Catalyzes the ADP transfer to D-glycero-D-manno-heptose 1-phosphate, yielding ADP-D,D-heptose (By similarity)
Core-Lipid A Synthesis	RFAC	Lipopolysaccharide heptosyltransferase i
Core-Lipid A Synthesis	RFAC2	Glycosyl transferase, family 9
Core-Lipid A Synthesis	RFAD	NAD-dependent epimerase dehydratase
Core-Lipid A Synthesis	RFAD	Nad-dependent epimerase dehydratase
Core-Lipid A Synthesis	RFAF	Glycosyltransferase family 9 (heptosyltransferase)
Core-Lipid A Synthesis	RFAF	Heptosyltransferase II
Core-Lipid A Synthesis	RFAI	UDP-glucose
Core-Lipid A Synthesis	RFAL	O-antigen
Core-Lipid A Synthesis	HTRB	Lipid A biosynthesis lauroyl
Core-Lipid A Synthesis	HTRB	Lipid a biosynthesis
Core-Lipid A Synthesis	HTRB	Lipid A biosynthesis acyltransferase
Core-Lipid A Synthesis	MSBA	ABC transporter
Core-Lipid A Synthesis	MSBA	ABC transporter related
Core-Lipid A Synthesis	MSBA	Involved in lipid A export and possibly also in glycerophospholipid export and for biogenesis of the outer membrane. Transmembrane domains (TMD) form a pore in the inner membrane and the ATP-binding domain (NBD) is responsible for energy generation (By similarity)
Core-Lipid A Synthesis	MSBA	ABC, transporter
Rhamnose LPS Synthesis	RFBA	Inherit from bctoNOG: Transferase
Rhamnose LPS Synthesis	RFBA	Catalyzes the formation of dTDP-glucose, from dTTP and glucose 1-phosphate, as well as its pyrophosphorolysis (By similarity)
Rhamnose LPS Synthesis	RFBB	Polysaccharide biosynthesis protein
Rhamnose LPS Synthesis	RFBB	ABC transporter
Rhamnose LPS Synthesis	RFBB	NAD-dependent epimerase dehydratase
Rhamnose LPS Synthesis	RFBB	DTDP-glucose 4-6-dehydratase
Rhamnose LPS Synthesis	RFBB	Dtdp-glucose 4,6-dehydratase
Rhamnose LPS Synthesis	RFBB	Epimerase dehydratase
Rhamnose LPS Synthesis	RFBC	DTDP-4-dehydrorhamnose 3,5-epimerase
Rhamnose LPS Synthesis	RFBD	DTDP-4-dehydrorhamnose reductase
Rhamnose LPS Synthesis	RFBD	Dtdp-4-dehydrorhamnose reductase
Rhamnose LPS Synthesis	RFBH	DegT DnrJ EryC1 StrS aminotransferase
Rhamnose LPS Synthesis	RFBH	DegT/DnrJ/EryC1/StrS aminotransferase family
Rhamnose LPS Synthesis	RFBH	DegT DnrJ EryC1 StrS
Rhamnose LPS Synthesis	RFBP	Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase (EC 2.7.8.6)
Rhamnose LPS Synthesis	RFBP	Transferase
Rhamnose LPS Synthesis	RFBF	Glucose-1-phosphate cytidylyltransferase
Rhamnose LPS Synthesis	RFBF	Nucleotidyl transferase
Rhamnose LPS Synthesis	RFBU	Glycosyl transferase group 1
Rhamnose LPS Synthesis	RFBE	DegT DnrJ EryC1 StrS aminotransferase
Rhamnose LPS Synthesis	RFBE	Nad-dependent epimerase dehydratase
Rhamnose LPS Synthesis	RMLC	DTDP-4-dehydrorhamnose 3,5-epimerase
Rhamnose LPS Synthesis	WBBL	Family 2
Rhamnose LPS Synthesis	WBBL	Glycosyl transferase
Rhamnose LPS Synthesis	WBBL	Glycosyl transferase family 2
Rhamnose LPS Synthesis	WBBL	Glycosyl transferase family
Rhamnose LPS Synthesis	WBBL	Glycosyl transferase, family 2
Arabinose LPS Synthesis	ARNA	Nad-dependent epimerase dehydratase
Arabinose LPS Synthesis	ARNB	DegT DnrJ EryC1 StrS aminotransferase
Arabinose LPS Synthesis	ARNB	Catalyzes the conversion of UDP-4-keto-arabinose (UDP- Ara4O) to UDP-4-amino-4-deoxy-L-arabinose (UDP-L-Ara4N). The modified arabinose is attached to lipid A and is required for resistance to polymyxin and cationic antimicrobial peptides (By similarity)
Arabinose LPS Synthesis	ARNC	Glycosyl transferase family 2

(continued from previous page)

Arabinose LPS Synthesis	ARNC	Catalyzes the transfer of 4-deoxy-4-formamido-L- arabinose from UDP to undecaprenyl phosphate. The modified arabinose is attached to lipid A and is required for resistance to polymyxin and cationic antimicrobial peptides (By similarity)
Arabinose LPS Synthesis	ARNT	Catalyzes the transfer of the L-Ara4N moiety of the glycolipid undecaprenyl phosphate-alpha-L-Ara4N to lipid A. The modified arabinose is attached to lipid A and is required for resistance to polymyxin and cationic antimicrobial peptides (By similarity)
Arabinose LPS Synthesis	ARNT	Glycosyl transferase family 39
Enterobacterial Common Antigen Synthesis	WECC	Dehydrogenase
Enterobacterial Common Antigen Synthesis	WECC	Nucleotide sugar dehydrogenase
Enterobacterial Common Antigen Synthesis	WECA	Undecaprenyl-Phosphate
LPS Assembly	LPTB	Abc transporter atp-binding protein
LPS Assembly	LPTB	ABC, transporter
LPS Assembly	LPTB	ABC transporter
LPS Assembly	LPTD	Organic solvent tolerance protein
LPS Assembly	LPTD	Involved in the assembly of LPS in the outer leaflet of the outer membrane. Determines N-hexane tolerance and is involved in outer membrane permeability. Essential for envelope biogenesis (By similarity)
LPS Assembly	LPTE	Rare lipoprotein B
LPS Assembly	YHBN	Lipopolysaccharide transport periplasmic protein LptA
Phospholipase C	PLCC	Phospholipase C
Phospholipase C	PLCC	Acid phosphatase
Phospholipase C	PLCC	Phosphoesterase family
Phospholipase C		K01114 phospholipase C EC 3.1.4.3
Phospholipase C		Inherit from COG: phospholipase C
Phospholipase C		Phospholipase C
Phospholipase C		Inherit from bactNOG: Phosphatidylinositol-specific phospholipase C
Alginate Synthesis	LADS	Histidine kinase
Alginate Synthesis	ALG8	Alginate biosynthesis protein Alg8
Alginate Synthesis	ALGB	Two component, sigma54 specific, transcriptional regulator, Fis family
Alginate Synthesis	ALGC	Phosphomannomutase
Alginate Synthesis	ALGG	Bifunctional protein that converts poly(beta-D- mannuronate) to alpha-L-guluronate and that is also part of a periplasmic protein complex that serves as a scaffold that leads the newly formed alginate polymer through the periplasmic space to the outer membrane secretin AlgE
Alginate Synthesis	ALGI	Membrane bound O-acyl transferase, MBOAT family protein
Alginate Synthesis	ALGI	Membrane bOund o-acyl transferase mboat family protein
Alginate Synthesis	ALGI	Membrane bound O-acyl transferase, MBOAT
Alginate Synthesis	ALGX	Alginate biosynthesis protein AlgX
Alginate Synthesis	MUCD	Protease
Alginate Synthesis	MUCD	Protease, Do
Alginate Synthesis	MUCR	Transcriptional regulator
Alginate Synthesis	MUCS	Transcriptional regulator, MarR family
Succinoglycan Synthesis	CHVI	Two component transcriptional regulator, winged helix family
Succinoglycan Synthesis	EXOO	Glycosyl transferase, family 2
Succinoglycan Synthesis	EXOI	Succinoglycan biosynthesis protein
Succinoglycan Synthesis	EXOI	Nuclease (SNase domain protein)
Succinoglycan Synthesis	PSSA	CDP-diacylglycerol--serine O-phosphatidyltransferase
Succinoglycan Synthesis	PSSA	Phosphatidylserine synthase
Succinoglycan Synthesis	PSSN	Export protein
Succinoglycan Synthesis	PSSP	Capsular exopolysaccharide family
Succinoglycan Synthesis	PSSP	Exopolysaccharide
Bacterial Cellulose Synthesis	BSCB	Cellulose synthase regulator protein
Bacterial Cellulose Synthesis	BCSZ	Cellulase (EC 3.2.1.4)
Bacterial Cellulose Synthesis	BCSA	Cellulose synthase, catalytic subunit
Bacterial Cellulose Synthesis	BCSA	Glycosyl transferase family 2
Bacterial Cellulose Synthesis	BCSA	Synthase

(continued from previous page)

Bacterial Cellulose Synthesis	BCSC	Cellulose synthase
Colanic Acid+Capsule Synthesis	GMD	GDP-mannose 4,6-dehydratase
Colanic Acid+Capsule Synthesis	GMD	GDPmannose 4,6-dehydratase
Colanic Acid+Capsule Synthesis	FCL	NAD-dependent epimerase dehydratase
Colanic Acid+Capsule Synthesis	FCL	Nad-dependent epimerase dehydratase
Colanic Acid+Capsule Synthesis	WCAG7	Inherit from bactNOG: Ceramide Glucosyltransferase
Colanic Acid+Capsule Synthesis	WCAG8	Nad-dependent epimerase dehydratase
Colanic Acid+Capsule Synthesis	WCAJ	Bacterial sugar transferase
Colanic Acid+Capsule Synthesis	WCAJ	Undecaprenyl-phosphate glucose phosphotransferase
Colanic Acid+Capsule Synthesis	WCAJ	Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase
Colanic Acid+Capsule Synthesis	WCAJ	Sugar transferase
Colanic Acid+Capsule Synthesis	WCAJ	Transferase
Colanic Acid+Capsule Synthesis	WZA	Polysaccharide biosynthesis/export protein
Colanic Acid+Capsule Synthesis	WZA	Polysaccharide biosynthesis export protein
Colanic Acid+Capsule Synthesis	WZA	Polysaccharide export protein
Colanic Acid+Capsule Synthesis	WZC	Tyrosine-protein kinase
Colanic Acid+Capsule Synthesis	DJLA	Regulatory DnaK co-chaperone. Direct interaction between DnaK and DjlA is needed for the induction of the wcaABCDE operon, involved in the synthesis of a colanic acid polysaccharide capsule, possibly through activation of the RcsB RcsC phosphotransfer signaling pathway. The colanic acid capsule may help the bacterium survive conditions outside the host (By similarity)
Colanic Acid+Capsule Synthesis	CAPD	Polysaccharide biosynthesis protein
Colanic Acid+Capsule Synthesis	CAPM	Glycosyl transferase
Colanic Acid+Capsule Synthesis	GUMC	Chain length determinant protein
Colanic Acid+Capsule Synthesis	GUMC	Capsular exopolysaccharide family protein
Colanic Acid+Capsule Synthesis	GUMB	Polysaccharide export protein
Colanic Acid+Capsule Synthesis	WBPL	Glycosyl transferase, family 4
Colanic Acid+Capsule Synthesis	WBPL	Undecaprenyl-Phosphate
Colanic Acid+Capsule Synthesis	WBPP	NAD-dependent epimerase dehydratase
Colanic Acid+Capsule Synthesis	WBPP	UDP-glucose 4-epimerase (EC 5.1.3.2)
Colanic Acid+Capsule Synthesis	KPSD	Polysaccharide export protein
Colanic Acid+Capsule Synthesis	NEUB	Synthase
Colanic Acid+Capsule Synthesis	MANA	Mannose-6-phosphate isomerase
Colanic Acid+Capsule Synthesis	MANA	Mannose-6-phosphate isomerase, class I
Colanic Acid+Capsule Synthesis	MANA	Mannan endo-1,4-beta-mannosidase

(continued from previous page)

Colanic Acid+Capsule Synthesis	MANB	Phosphoglucomutase phosphomannomutase
Colanic Acid+Capsule Synthesis	MANB	Phosphomannomutase
Colanic Acid+Capsule Synthesis	MANB	Cellulase (glycosyl hydrolase family 5)
Colanic Acid+Capsule Synthesis	MANC	Mannose-1-phosphate guanylyltransferase
Colanic Acid+Capsule Synthesis	MANC	Mannose-1-phosphate guanylyltransferase, mannose-6-phosphate isomerase
Colanic Acid+Capsule Synthesis	MANC	Mannose-1-phosphate guanylyltransferase mannose-6-phosphate isomerase
Colanic Acid+Capsule Synthesis	MANC	Nucleotidyl transferase
Colanic Acid+Capsule Synthesis	MANC	Cupin 2, conserved barrel domain protein
Colanic Acid+Capsule Synthesis		Capsular exopolysaccharide family
Colanic Acid+Capsule Synthesis		Capsular exopolysaccharide family protein
Colanic Acid+Capsule Synthesis		Capsular polysaccharide biosynthesis protein
Colanic Acid+Capsule Synthesis		Capsular polysaccharide biosynthesis
Adhesin	ICAA	Glycosyl hydrolases family 18
Adhesin		Beta-Ig-H3 fasciclin
Adhesin		Fasciclin
Cell Division Septum	FTSZ	Essential cell division protein that forms a contractile ring structure (Z ring) at the future cell division site. The regulation of the ring assembly controls the timing and the location of cell division. One of the functions of the FtsZ ring is to recruit other cell division proteins to the septum to produce a new cell wall between the dividing cells. Binds GTP and shows GTPase activity (By similarity)
Cell Division Septum	FTSA	This protein may be involved in anomalous filament growth. May be a component of the septum (By similarity)
Cell Division Septum	FTSK	Cell division protein FtsK
Cell Division Septum	FTSK	Essential cell division protein that coordinates cell division and chromosome segregation. The N-terminus is involved in assembly of the cell-division machinery. The C-terminus functions as a DNA motor that moves dsDNA in an ATP-dependent manner towards the dif recombination site, which is located within the replication terminus region. Translocation stops specifically at Xer-dif sites, where FtsK interacts with the Xer recombinase, allowing activation of chromosome unlinking by recombination. FtsK orienting polar sequences (KOPS) guide the direction of DNA translocation. FtsK can remove proteins from DNA as it translocates, but translocation stops specifically at XerCD-dif site, thereby preventing removal of XerC and XerD from dif
Cell Division Septum	FTSE	Cell division ATP-binding protein FtsE
Cell Division Septum	FTSE	Cell division atp-binding protein ftse
Cell Division Septum	FTSE	Cell division ATP-binding protein
Cell Division Septum	FTSX	Cell division protein
Cell Division Septum	FTSX	Part of the ABC transporter FtsEX involved in cellular division (By similarity)
Cell Division Septum	FTSX	Cell division protein FtsX
Cell Division Septum	FTSN	Cell division protein
Cell Division Septum	FTSI	Peptidoglycan glycosyltransferase
Cell Division Septum	FTSI	Penicillin-binding protein
Cell Division Septum	FTSI	Peptidoglycan synthetase ftsI
Cell Division Septum	FTSQ	Essential cell division protein (By similarity)
Cell Division Septum	ZAPC	Contributes to the efficiency of the cell division process by stabilizing the polymeric form of the cell division protein FtsZ. Acts by promoting interactions between FtsZ protofilaments and suppressing the GTPase activity of FtsZ (By similarity)
Cell Division Septum	ZAPA	Activator of cell division through the inhibition of FtsZ GTPase activity, therefore promoting FtsZ assembly into bundles of protofilaments necessary for the formation of the division Z ring. It is recruited early at mid-cell but it is not essential for cell division (By similarity)
Cell Division Septum	ENVC	Peptidase
Cell Division Septum	AMIA	Cell wall hydrolase autolysin
Cell Division Septum	AMIA	N-acetylmuramoyl-L-alanine amidase
Cell Division Septum	AMIB	N-acetylmuramoyl-L-alanine amidase
Cell Division Septum	AMIC	N-acetylmuramoyl-l-alanine amidase (EC 3.5.1.28)
Cell Division Septum	AMIC	N-acetylmuramoyl-L-alanine amidase

(continued from previous page)

Cell Division Septum	MIND	Site-determining protein
Cell Division Septum	ZIPA	Cell division protein ZipA
Cell Division Septum		Cell division protein FtsL
Cell Division Septum		Activator of cell division through the inhibition of FtsZ GTPase activity, therefore promoting FtsZ assembly into bundles of protofilaments necessary for the formation of the division Z ring. It is recruited early at mid-cell but it is not essential for cell division (By similarity)
Cell Division Septum		Cell division protein FtsK
Cell Division Septum		Inherit from COG: Divisome component that associates with the complex late in its assembly, after the Z-ring is formed, and is dependent on DivIC and PBP2B for its recruitment to the divisome. Together with EzrA, is a key component of the system that regulates PBP1 localization during cell cycle progression. Its main role could be the removal of PBP1 from the cell pole after pole maturation is completed. Also contributes to the recruitment of PBP1 to the division complex. Not essential for septum formation (By similarity)
Chromosome Partitioning	SMC	Chromosome segregation protein SMC
Chromosome Partitioning	SMC	Required for chromosome condensation and partitioning (By similarity)
Chromosome Partitioning	MRP	Involved in chromosome partitioning
Chromosome Partitioning	MRP	ATPase-like, ParA MinD
Chromosome Partitioning	XERC	Site-specific tyrosine recombinase, which acts by catalyzing the cutting and rejoining of the recombining DNA molecules. The XerC-XerD complex is essential to convert dimers of the bacterial chromosome into monomers to permit their segregation at cell division. It also contributes to the segregational stability of plasmids (By similarity)
Chromosome Partitioning	XERC	Integrase
Chromosome Partitioning	XERD	Site-specific tyrosine recombinase, which acts by catalyzing the cutting and rejoining of the recombining DNA molecules. The XerC-XerD complex is essential to convert dimers of the bacterial chromosome into monomers to permit their segregation at cell division. It also contributes to the segregational stability of plasmids (By similarity)
Chromosome Partitioning	PARE	DNA topoisomerase type IIA subunit B region 2 domain protein
Chromosome Partitioning	PARE	DNA topoisomerase IV, subunit B
Chromosome Partitioning	PARE	DNA topoisomerase type iia subunit b region 2 domain protein
Chromosome Partitioning	PARE	Dna topoisomerase iv (Subunit b)
Chromosome Partitioning	PARC	DNA topoisomerase IV, subunit A
Chromosome Partitioning	PARC	DNA topoisomerase
Chromosome Partitioning	PARC	DNA topoisomerase (EC 5.99.1.3)
Chromosome Partitioning	PARB	ParB-like partition protein
Chromosome Partitioning	PARB	Chromosome segregation DNA-binding protein
Chromosome Partitioning	PARB	Partitioning protein
Rod Morphogenesis	MREB	Cell shape determining protein, MreB Mrl family
Rod Morphogenesis	MREB	Rod shape-determining protein mreB
Rod Morphogenesis	MREB	Rod shape-determining protein mreb
Rod Morphogenesis	MREB	MreB Mrl family cell shape determining protein
Rod Morphogenesis	MREB	Rod shape-determining protein MreB
Rod Morphogenesis	MREC	Rod shape-determining protein MreC
Rod Morphogenesis	MRED	Rod shape-determining protein
Rod Morphogenesis	RODZ	Cytoskeletal protein that is involved in cell-shape control through regulation of the length of the long axis (By similarity)
Rod Morphogenesis	WAG31	DivIVA family
Microcompartment	PDUA	Microcompartments protein
Microcompartment	CCHA	Microcompartments protein
Microcompartment	CCMK	Major carboxysome shell protein
Microcompartment	CCMK	Microcompartments protein
Microcompartment		Microcompartments protein
OAR	OAR	TonB-dependent Receptor Plug Domain
OAR	OAR	Oar protein
OAR	OAR	TonB-dependent receptor
Flagellum	FLIC	Flagellin
Flagellum	FLIC	Flagellin domain protein
Flagellum	FLHF	Flagellar biosynthesis regulator FlhF
Flagellum	FLGC	Flagellar basal-body rod protein FlgC
Flagellum	FLGC	Flagellar basal-body rod protein (FlgC)
Flagellum	FLIF	The M ring may be actively involved in energy transduction (By similarity)
Flagellum	FLGH	Assembles around the rod to form the L-ring and probably protects the motor basal body from shearing forces during rotation (By similarity)
Flagellum	FLGK	Flagellar hook-associated protein flgk
Flagellum	FLGK	Flagellar hook-associated protein

(continued from previous page)

Flagellum	FLGE	Flagellar hook protein flgE
Flagellum	FLGE	Hook-basal body protein
Flagellum	FLGE	Flagellar basal body protein FlaE
Flagellum	FLGE	Flagellar hook protein, FlgE
Flagellum	FLHB	Flagellar biosynthetic protein flhB
Flagellum	FLIP	Flagellar biosynthetic protein FliP
Flagellum	FLGI	Assembles around the rod to form the L-ring and probably protects the motor basal body from shearing forces during rotation (By similarity)
Flagellum	FLIS	Flagellar protein FliS
Flagellum	FLGA	Flagellar basal body P-ring biosynthesis protein FlgA
Flagellum	FLGA	Flagella basal body P-ring formation protein
Flagellum	FLHA	Flagellar biosynthesis protein (FlhA)
Flagellum	FLHA	Flagellar biosynthesis protein, FlhA
Flagellum	FLAG	Flagellar protein
Flagellum	FLIJ	Flagellar export
Flagellum	FLIN	Flagellar motor switch protein
Flagellum	FLGL	Flagellar hook-associated protein flgL
Flagellum	FLII	Type iii secretion
Flagellum	FLII	Flagellum-specific ATP synthase
Flagellum	FLII	ATPase, FliI
Flagellum	FLII	ATP synthase alpha/beta family, nucleotide-binding domain
Flagellum	HRCN	Flagellum-specific ATP synthase
Flagellum	OCAR_5373	Flagellin
Flagellum	TLL0138	Flagellar biosynthesis protein FlhF
Flagellum		Flagellar hook-associated
Flagellum		Flagellar hook-associated protein
Flagellum		Bacterial flagellin N-terminal helical region
Flagellum		Flagellar hook-length control protein
Flagellum		Regulatory protein FlaEY
Pilin+Fimbria	FIMA	Family of unknown function (DUF1028)
Pilin+Fimbria	FIMB	Type 1 fimbriae regulatory protein
Pilin+Fimbria	FIMR	Two component transcriptional regulator, LuxR family
Pilin+Fimbria	FIMV	Domain-containing protein
Pilin+Fimbria	FIMV	Pilus assembly protein
Pilin+Fimbria	FIMV	Domain protein
Pilin+Fimbria	HRPA	ATP-dependent helicase HrpA
Pilin+Fimbria	HRPA	ATP-dependent helicase hrpA
Pilin+Fimbria	HRPA	ATP-dependent Helicase
Pilin+Fimbria	HRPB	ATP-dependent helicase HrpB
Pilin+Fimbria	HRPX	Signal transduction histidine kinase
Pilin+Fimbria	HRPY	Two component transcriptional regulator luxr family
Pilin+Fimbria	PILT	Pfam:GSPII_E
Pilin+Fimbria	PILT	Twitching motility protein
Pilin+Fimbria	UPTC	Twitching motility protein
Pilin+Fimbria	PILP	Pilus assembly protein
Pilin+Fimbria	PILR	Two component, sigma54 specific, transcriptional regulator, Fis family
Pilin+Fimbria	PILR	(Type IV) pilus
Pilin+Fimbria	PILA	Pilin (bacterial filament)
Pilin+Fimbria	PILA	Fimbrial protein (Pilin)
Pilin+Fimbria	PILA	Type IV fimbrial pilin protein
Pilin+Fimbria	PILA	Fimbrial protein
Pilin+Fimbria	PPDD	Type IV pilin
Pilin+Fimbria	PPDD	Fimbrial protein
Pilin+Fimbria	PILQ	(type IV) pilus
Pilin+Fimbria	PILQ	(Type IV) pilus
Pilin+Fimbria	PILQ	Type IV pilus secretin PilQ
Pilin+Fimbria	PILB	Type II secretion system protein E
Pilin+Fimbria	PILB	Pathway protein e
Pilin+Fimbria	PILE	Pilus assembly protein
Pilin+Fimbria	PILE	Fimbrial protein
Pilin+Fimbria	PILO	Pilus assembly protein, PilO
Pilin+Fimbria	PILO	Pilin accessory protein (PilO)
Pilin+Fimbria	PILO	Pilus assembly protein pilo

(continued from previous page)

Pilin+Fimbria	PILO	Assembly protein PilO
Pilin+Fimbria	PILA,PPDD	Fimbrial protein
Pilin+Fimbria	PILM	Type IV pilus assembly protein PilM
Pilin+Fimbria	PILN	Fimbrial assembly
Pilin+Fimbria	PILN	(Type IV) pilus
Pilin+Fimbria	PILN	Fimbrial assembly family protein
Pilin+Fimbria	PILC	Type ii secretion system
Pilin+Fimbria	PILC	General secretion pathway protein f
Pilin+Fimbria	PILC	Type II secretion system
Pilin+Fimbria	PILC	Type II secretion system protein
Pilin+Fimbria	PILC	Type IV pilin biogenesis protein
Pilin+Fimbria	PILW	Type IV pilus assembly protein PilW
Pilin+Fimbria		Response regulator receiver modulated PilZ sensor protein
Pilin+Fimbria		Inherit from COG: Pilus assembly protein tip-associated adhesin
Pilin+Fimbria		Inherit from bactNOG: Pilin, type IV
Myxococcal Gliding	MGLA	ADP-ribosylation factor family
Myxococcal Gliding	MGLA	Gliding motility protein MglA
Myxococcal Gliding	MGLB	Roadblock LC7 family protein
Myxococcal Gliding	AGLR	MotA TolQ exbB proton channel
Myxococcal Gliding	AGLS	Adventurous gliding motility protein
Myxococcal Gliding	AGMK	Repeat protein
Myxococcal Gliding		Adventurous gliding protein T
Bacteroidete Gliding	GLDN	Gliding motility associated protein GldN
Bacteroidete Gliding	GLDL	Gliding motility-associated protein GldL
Bacteroidete Gliding	GLDJ	Sulphatase-modifying factor protein
Bacteroidete Gliding	GLDM	Gliding motility-associated protein GldM
Bacteroidete Gliding	GLDC	Gliding motility-associated protein GldC
Bacteroidete Gliding	GLDG	#NAME?
Bacteroidete Gliding	GLDG	ABC transporter substrate-binding component GldG
Bacteroidete Gliding	GLDK	Sulphatase-modifying factor protein
Bacteroidete Gliding	SPRA	Inherit from bctoNOG: Gliding motility-related protein
Chemotaxis	MCPU	Methyl-accepting chemotaxis sensory transducer
Chemotaxis	MCPU	Methyl-accepting chemotaxis
Chemotaxis	MCP64H-2	Methyl-accepting chemotaxis sensory transducer
Chemotaxis	CHER	MCP methyltransferase, CheR-type
Chemotaxis	CHER	Methylation of the membrane-bound methyl-accepting chemotaxis proteins (MCP) to form gamma-glutamyl methyl ester residues in MCP (By similarity)
Chemotaxis	WSPC	Methyl-transferase
Chemotaxis	WSPC	Methyltransferase
Chemotaxis	WSPC	MCP methyltransferase, CheR-type with Tpr repeats
Chemotaxis	TLPA	Methyl-accepting chemotaxis
Chemotaxis	CHER2	Methylation of the membrane-bound methyl-accepting chemotaxis proteins (MCP) to form gamma-glutamyl methyl ester residues in MCP (By similarity)
Chemotaxis	CHEW1	Chemotaxis protein CheW
Chemotaxis	CHEW40H-1	CheW protein
Chemotaxis	WSPD	Chemotaxis protein CheW
Chemotaxis	WSPD	CheW
Chemotaxis	HEMAT	Methyl-accepting chemotaxis
Chemotaxis	PILI	Chew protein
Chemotaxis	CTPL	Methyl-accepting chemotaxis
Chemotaxis	MCP-4	Methyl-accepting chemotaxis sensory transducer
Chemotaxis	AER	Methyl-accepting chemotaxis sensory transducer with Pas Pac sensor
Chemotaxis	CHEY1	Response regulator
Chemotaxis	CHED	Probably deamidates glutamine residues to glutamate on methyl-accepting chemotaxis receptors (MCPs), playing an important role in chemotaxis (By similarity)
Chemotaxis	CHEB	Catalyzes the demethylation of specific methylglutamate residues introduced into the chemoreceptors (methyl-accepting chemotaxis proteins) by CheR (By similarity)
Chemotaxis	CHEA	CheA Signal Transduction Histidine Kinase
Chemotaxis	CHEA	CheA signal transduction histidine kinase
Chemotaxis	CHEA	Histidine kinase
Chemotaxis	CHEA	Chea signal transduction histidine kinase

(continued from previous page)

Chemotaxis	CHEV	Response regulator receiver modulated CheW protein
Chemotaxis	CHEY	Response regulator receiver protein
Chemotaxis	CHEY	Response regulator
Chemotaxis	CHEBR	Catalyzes the demethylation of specific methylglutamate residues introduced into the chemoreceptors (methyl-accepting chemotaxis proteins) by CheR (By similarity)
Chemotaxis	CHEY40H-2	Response regulator
Chemotaxis	PILJ	Methyl-accepting chemotaxis
Chemotaxis	WSPR	Response regulator
Chemotaxis		MCP methyltransferase methylesterase, CheR CheB with PAS PAC sensor
Chemotaxis		Methyl-accepting chemotaxis
Chemotaxis		MCP methyltransferase, CheR-type
Chemotaxis		Chemotaxis sensory transducer
Chemotaxis		Chemotaxis
Chemotaxis		Chemotaxis protein
Chemotaxis		Methyl-accepting chemotaxis protein (MCP) signalling domain
Chemotaxis		Methyl-accepting chemotaxis sensory transducer
Chemotaxis		Catalyzes the demethylation of specific methylglutamate residues introduced into the chemoreceptors (methyl-accepting chemotaxis proteins) by CheR (By similarity)
Chemotaxis		Signal transduction histidine kinase with CheB and CheR activity
Cold Shock	CSPE	Cold shock protein
Cold Shock	CSPE	Cold-shock DNA-binding domain protein
Cold Shock	CSPA	Cold shock protein
Cold Shock	CSPA	Cold-shock DNA-binding protein family
Cold Shock	CSPA,CSPE	Cold shock protein
Cold Shock	CSPB	Cold-shock protein
Cold Shock	CSPB	Cold-shock DNA-binding
Cold Shock	CSPB	'Cold-shock' DNA-binding domain
Cold Shock	CSPC	Cold-shock DNA-binding protein family
Cold Shock	CSPA,CSPC,CSPE	Cold shock protein
Cold Shock	CSPA3	DNA-binding domain protein
Cold Shock	CSPD	Cold-shock DNA-binding domain protein
Cold Shock		Cold shock protein ScoF
Cold Shock		Cold-shock DNA-binding domain protein
Cold Shock		Cold shock protein
Cold Shock		Cold-shock DNA-binding
Heat Shock	HRCA	Negative regulator of class I heat shock genes (grpE- dnaK-dnaJ and groELS operons). Prevents heat-shock induction of these operons (By similarity)
Heat Shock	HTPG	Heat shock protein Hsp90
Heat Shock	HTPG	Molecular chaperone. Has ATPase activity (By similarity)
Heat Shock	GRPE	Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins, in association with DnaK and GrpE. It is the nucleotide exchange factor for DnaK and may function as a thermosensor. Unfolded proteins bind initially to DnaJ
Heat Shock	IBPA	HeAt shock protein
Heat Shock	IBPA	Heat shock protein, Hsp20
Heat Shock	IBPA	Heat shock protein
Heat Shock	MMC1_1348	Heat shock protein (HSP20
Heat Shock	HSP20	Heat shock protein (HSP20
Heat Shock	HSP20	Heat shock protein, Hsp20
Heat Shock	HSP20	Heat shock protein
Heat Shock	HSP20	Heat shock protein Hsp20
Heat Shock	BMUL_2287	HeAt shock protein
Heat Shock	BMUL_2287	Heat shock protein
Heat Shock	IBPA1	Heat Shock Protein
Heat Shock	HSP	Heat shock protein
Heat Shock	OCAR_5761	Heat shock protein
Heat Shock	CLPB	ATP-dependent chaperone ClpB
Heat Shock	CLPB	K03695 ATP-dependent Clp protease ATP-binding subunit ClpB

(continued from previous page)

Heat Shock	CLPB	ATP-dependent CLP protease ATP-binding subunit
Heat Shock	CLPB	Part of a stress-induced multi-chaperone system, it is involved in the recovery of the cell from heat-induced damage, in cooperation with DnaK, DnaJ and GrpE. Acts before DnaK, in the processing of protein aggregates. Protein binding stimulates the ATPase activity
Heat Shock	CLPB	ATP-dependent chaperone
Heat Shock	CLPB	ATP-dependent chaperone clpb
Heat Shock		Heat shock
Heat Shock		Heat shock cognate 70 kDa
Heat Shock		Heat shock protein Hsp20
Heat Shock		Response to heat
Osmoprotectant Transport	PROW	Binding-protein-dependent transport systems inner membrane component
Osmoprotectant Transport	PROV	ABC transporter
Osmoprotectant Transport	PROX	ABC transporter
Osmoprotectant Transport	PROP26	Major Facilitator
Osmoprotectant Transport	PROP16	Transporter
Osmoprotectant Transport	PROP	)transporter
Osmoprotectant Transport	PROP11	MembrAne
Osmoprotectant Synthesis	BETB	Dehydrogenase
Osmoprotectant Synthesis	EHUB	Ectoine hydroxyectoine ABC transporter solute-binding protein
Osmoprotectant Synthesis	GBSA	Aldehyde dehydrogenase
Osmoprotectant Synthesis	OPUCB	ABC-type glycine betaine transport, periplasmic subunit
Osmoprotectant Synthesis	OPUB	Substrate-binding region of ABC-type glycine betaine transport system
Osmoprotectant Synthesis	YEHZ	Glycine Betaine
Osmoprotectant Synthesis	BETT	Choline carnitine betaine transporter
Osmoprotectant Synthesis	BETA	Glucose-methanol-choline oxidoreductase
Osmoprotectant Synthesis	BETA	GMC oxidoreductase
Osmoprotectant Synthesis	BETA	Choline dehydrogenase
Osmoprotectant Synthesis	BETA	Can catalyze the oxidation of choline to betaine aldehyde and betaine aldehyde to glycine betaine (By similarity)
Osmoprotectant Synthesis		Choline ABC transporter periplasmic binding protein
Osmoprotectant Synthesis	MDOD	Glucan biosynthesis protein D
Osmoprotectant Synthesis	MDOD	Glucan biosynthesis protein
Osmoprotectant Synthesis	MDOH	Involved in the biosynthesis of osmoregulated periplasmic glucans (OPGs) (By similarity)
Osmoprotectant Synthesis	MDOB	Sulfatase
Peroxide Resistance	AHPC	Peroxiredoxin
Peroxide Resistance	AHPC	Alkyl hydroperoxide reductase
Peroxide Resistance	AHPC	Alkyl hydroperoxide reductase Thiol specific antioxidant Mal allergen
Peroxide Resistance	AHPC	Redoxin domain protein
Peroxide Resistance	AHPD	Antioxidant protein with alkyl hydroperoxidase activity. Required for the reduction of the AhpC active site cysteine residues and for the regeneration of the AhpC enzyme activity (By similarity)
Peroxide Resistance	AHPF	Alkyl hydroperoxide reductase
Peroxide Resistance	SODA	Manganese and iron superoxide dismutase
Peroxide Resistance	SODA	Destroys radicals which are normally produced within the cells and which are toxic to biological systems (By similarity)
Peroxide Resistance	SODA	Iron/manganese superoxide dismutases, C-terminal domain
Peroxide Resistance	SODB	Destroys radicals which are normally produced within the cells and which are toxic to biological systems (By similarity)
Peroxide Resistance	SODA,SODB	Destroys radicals which are normally produced within the cells and which are toxic to biological systems (By similarity)
Peroxide Resistance	SODC	Superoxide dismutase copper zinc binding protein
Peroxide Resistance	SODC	Destroys radicals which are normally produced within the cells and which are toxic to biological systems (By similarity)
Peroxide Resistance	SODC	Superoxide dismutase copper zinc binding
Peroxide Resistance	SODN	Superoxide dismutase
Peroxide Resistance	KATE	Catalase EC 1.11.1.6
Peroxide Resistance	KATE	Catalase (EC 1.11.1.6)
Peroxide Resistance	KATE	Catalase (EC 1.11.1.6
Peroxide Resistance	KATE	Catalase
Peroxide Resistance	KATA	Catalase (EC 1.11.1.6)
Peroxide Resistance	KATA	Catalase (EC 1.11.1.6
Peroxide Resistance	KATA	Catalase
Peroxide Resistance	OCAR_5492	Antioxidant protein with alkyl hydroperoxidase activity. Required for the reduction of the AhpC active site cysteine residues and for the regeneration of the AhpC enzyme activity (By similarity)

(continued from previous page)

Peroxide Resistance	VEIS_1594	Antioxidant protein with alkyl hydroperoxidase activity. Required for the reduction of the AhpC active site cysteine residues and for the regeneration of the AhpC enzyme activity (By similarity)
Peroxide Resistance	OCAR_6549	Antioxidant protein with alkyl hydroperoxidase activity. Required for the reduction of the AhpC active site cysteine residues and for the regeneration of the AhpC enzyme activity (By similarity)
Peroxide Resistance		Di-haem cytochrome c peroxidase
Peroxide Resistance		Catalase domain protein
Peroxide Resistance		Catalase
Peroxide Resistance		Cytochrome c peroxidase
Peroxide Resistance		Antioxidant protein with alkyl hydroperoxidase activity. Required for the reduction of the AhpC active site cysteine residues and for the regeneration of the AhpC enzyme activity (By similarity)
Glutathione Detoxification	BMUL_3027	Glutathione S-transferase
Glutathione Detoxification	GSTA	Glutathione S-transferase
Glutathione Detoxification	OCAR_5403	Glutathione S-Transferase
Glutathione Detoxification	BTUE	Glutathione peroxidase
Glutathione Detoxification	GPO	Glutathione peroxidase
Glutathione Detoxification	GST3	Glutathione S-transferase
Glutathione Detoxification	YGHU	Glutathione S-transferase
Glutathione Detoxification	YGHU	Glutathione S-Transferase
Glutathione Detoxification	YIBF	Glutathione S-transferase
Glutathione Detoxification	BMUL_3027	Glutathione S-transferase
Glutathione Detoxification	GSTN	Glutathione S-transferase
Glutathione Detoxification	GOR	Glutathione reductase
Glutathione Detoxification	GRXB	Glutaredoxin 2
Glutathione Detoxification	YFCF	Glutathione S-transferase
Glutathione Detoxification	YLIJ	Glutathione S-transferase
Glutathione Detoxification	LIGE	Glutathione S-transferase
Glutathione Detoxification	GST	Glutathione S-transferase
Glutathione Detoxification	GST	Glutathione S-Transferase
Glutathione Detoxification	GRXD	Glutaredoxin
Glutathione Detoxification	GRXC	Glutaredoxin
Glutathione Detoxification	GRLA	Glutaredoxin
Glutathione Detoxification	GSP	Glutathionylspermidine synthase
Glutathione Detoxification		Glutathione S-transferase, C-terminal domain
Glutathione Detoxification		Glutathione S-transferase, C-terminal domain
Glutathione Detoxification		Lactoylglutathione lyase
Glutathione Detoxification		Glutathione S-transferase
ACR HAE Resistance Pump	ACRA	RND Family Efflux Transporter MFP Subunit
ACR HAE Resistance Pump	ACRB	Transporter, hydrophobe amphiphile efflux-1 (HAE1) family
ACR HAE Resistance Pump	ACRB	Acriflavin resistance protein
ACR HAE Resistance Pump	ACRB	Resistance protein
ACR HAE Resistance Pump	ACRB3	Acriflavin resistance protein
ACR HAE Resistance Pump	ACR	Acriflavin resistance protein
ACR HAE Resistance Pump	ACRE	RND family efflux transporter, MFP subunit
MDT HAE Resistance Pump	MDTA	RND family efflux transporter, MFP subunit
MDT HAE Resistance Pump	MDTA	Efflux transporter, RND family, MFP subunit
MDT HAE Resistance Pump	MDTB	Acriflavin resistance protein
MDT HAE Resistance Pump	MDTB	Resistance protein
MDT HAE Resistance Pump	MDTC	Acriflavin resistance protein
MDT HAE Resistance Pump	MDTC	Resistance protein
Gro	GROL	Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions (By similarity)
Gro	GROS	Binds to Cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter (By similarity)
Gro	GROES2	Binds to cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter (By similarity)
Iron-Sulfur Cluster Synthesis	ISCS	Cysteine desulfurase
Iron-Sulfur Cluster Synthesis	ISCS	Catalyzes the removal of elemental sulfur from cysteine to produce alanine (By similarity)
Iron-Sulfur Cluster Synthesis	ISCS	Aminotransferase class-V

(continued from previous page)

Iron-Sulfur Cluster Synthesis	ISCU	Scaffold protein
Iron-Sulfur Cluster Synthesis	HSCA	Chaperone involved in the maturation of iron-sulfur cluster-containing proteins. Has a low intrinsic ATPase activity which is markedly stimulated by HscB (By similarity)
Iron-Sulfur Cluster Synthesis	HSCA	Heat shock protein 70
Iron-Sulfur Cluster Synthesis	SUFB	FeS assembly protein SufB
Iron-Sulfur Cluster Synthesis	SUFB	FeS assembly protein sufB
Iron-Sulfur Cluster Synthesis	SUFB	Cysteine desulfurase activator complex subunit SufB
Iron-Sulfur Cluster Synthesis	SUFB	SufBD protein
Iron-Sulfur Cluster Synthesis	SUFC	FeS assembly ATPase sufC
Iron-Sulfur Cluster Synthesis	SUFC	FeS assembly ATPase SufC
Iron-Sulfur Cluster Synthesis	SUFD	FeS assembly protein SufD
Tol-Pal Outer Membrane Integrity	PAL	Peptidoglycan-associated lipoprotein
Tol-Pal Outer Membrane Integrity	PAL	OmpA family
Tol-Pal Outer Membrane Integrity	PAL	OmpA MotB domain-containing protein
Tol-Pal Outer Membrane Integrity	TOLA	TolA protein
Tol-Pal Outer Membrane Integrity	TOLA	Cell envelope integrity inner membrane protein TolA
Tol-Pal Outer Membrane Integrity	TOLB	Involved in the TonB-independent uptake of proteins (By similarity)
Tol-Pal Outer Membrane Integrity	TOLC	Type I secretion outer membrane protein, TolC
Tol-Pal Outer Membrane Integrity	TOLC	Outer membrane protein tolc
Tol-Pal Outer Membrane Integrity	TOLC	RND efflux system, outer membrane lipoprotein
Tol-Pal Outer Membrane Integrity	TOLQ	Mota tolq exbb proton channel
Tol-Pal Outer Membrane Integrity	TOLQ	MotA TolQ ExbB proton channel
Tol-Pal Outer Membrane Integrity	YBGC	Thioesterase
Omp	SKP	Outer membrane chaperone Skp (OmpH)
Omp	SKP	Molecular chaperone that interacts specifically with outer membrane proteins, thus maintaining the solubility of early folding intermediates during passage through the periplasm (By similarity)
Omp	OMPA	OmpA-like transmembrane domain
Omp	OMPA	Ompa motb domain protein
Omp	OMPA	OmpA MotB domain protein
Omp	OMPA	Outer membrane protein a
Omp	OMPA	OmpA MotB domain-containing protein
Omp	OMPX	Outer membrane protein x
Omp	OMPW	Outer membrane protein W
Omp	OMPC	MembrAne
Omp	OMPC,OMPF,PHOE	MembrAne
Omp		Outer membrane chaperone Skp
Omp		Outer membrane chaperone Skp (OmpH)
Omp		OmpA MotB family outer membrane protein
Bam Omp Assembly	BAMA	Outer membrane protein assembly
Bam Omp Assembly	BAMA	Part of the outer membrane protein assembly complex, which is involved in assembly and insertion of beta-barrel proteins into the outer membrane (By similarity)

(continued from previous page)

Bam Omp Assembly	BAMA	Part of the outer membrane protein assembly complex, which is involved in assembly and insertion of beta-barrel proteins into the outer membrane
Bam Omp Assembly	BAMA	Outer membrane protein assembly complex, YaeT protein
Bam Omp Assembly	BAMB	Part of the outer membrane protein assembly complex, which is involved in assembly and insertion of beta-barrel proteins into the outer membrane (By similarity)
Bam Omp Assembly	BAMB	Enzyme repeat domain protein
Bam Omp Assembly	BAMB	PQQ enzyme repeat family protein
Bam Omp Assembly	BAMD	Part of the outer membrane protein assembly complex, which is involved in assembly and insertion of beta-barrel proteins into the outer membrane
Bam Omp Assembly	BAMD	Part of the outer membrane protein assembly complex, which is involved in assembly and insertion of beta-barrel proteins into the outer membrane (By similarity)
Bam Omp Assembly	YAET	Outer membrane protein assembly complex, yaet protein
Bam Omp Assembly	YAET	Outer membrane protein assembly complex, YaeT protein
Bam Omp Assembly	NLPB	(Lipo)protein
Bam Omp Assembly	NLPB	(LipO)protein
Outer Membrane Porin	BMUL_4600	Porin Gram-negative type
Outer Membrane Porin	BMUL_4600	Porin, Gram-negative type
Outer Membrane Porin		Omp2b porin
Outer Membrane Porin		Outer membrane porin
Outer Membrane Porin		Inherit from NOG: Porin Gram-negative type
Outer Membrane Porin		Porin Gram-negative type
Carboxy-Terminal Peptidase	CTP	Carboxyl-terminal protease
Carboxy-Terminal Peptidase	CTPA	Protease
Carboxy-Terminal Peptidase	CTPA	Peptidase family S41
Carboxy-Terminal Peptidase	TRI1	Peptidase, S41
Carboxy-Terminal Peptidase		Peptidase S41
Carboxy-Terminal Peptidase		Peptidase family S41
Carboxy-Terminal Peptidase		Peptidase family S41, nonpeptidase-like protein
Carboxy-Terminal Peptidase		Peptidase, S41
Carboxy-Terminal Peptidase		Carboxy-terminal processing protease
Rubrerythrin	RBR	Rubrerythrin
Glutathione Metabolism	GGT	Gamma-glutamyltranspeptidase
Glutathione Metabolism	GGT	Gamma-glutamyltransferase (EC 2.3.2.2)
Glutathione Metabolism	GGT	Gamma-glutamyltranspeptidase EC 2.3.2.2
Glutathione Metabolism	GGT	Gamma-glutamyltranspeptidase (EC 2.3.2.2)
Glutathione Metabolism	GGT	K00681 gamma-glutamyltranspeptidase EC 2.3.2.2
Glutathione Metabolism	GGT1	Gamma-glutamyltranspeptidase EC 2.3.2.2
Glutathione Metabolism	GGT3	Gamma-glutamyltranspeptidase EC 2.3.2.2
Glutathione Metabolism	GSIA	ABC, transporter
Glutathione Metabolism	GSIA	(ABC) transporter
Carotenoid Synthesis	HOPE	Squalene synthase HpnC
Carotenoid Synthesis	HOPE	Synthase
Carotenoid Synthesis	HOPE	Squalene/phytoene synthase
Carotenoid Synthesis	CRTB	Phytoene synthase
Carotenoid Synthesis	CRTI	Phytoene desaturase
Carotenoid Synthesis	CRTI	Phytoene
Carotenoid Synthesis	CRTI	Phytoene dehydrogenase
Carotenoid Synthesis	CRTB	Phytoene synthase
Carotenoid Synthesis	CRTO	FAD dependent oxidoreductase
Carotenoid Synthesis	CRTD	Methoxyneurosporene dehydrogenase
Carotenoid Synthesis	CRTD	Flavin containing amine oxidoreductase
Carotenoid Synthesis	CRTQ	Flavin containing amine oxidoreductase
Carotenoid Synthesis		Phytoene synthase
Nodulation	NODD	Lysr family transcriptional regulator
Nodulation	NODD	LysR family transcriptional regulator
Nodulation	NODI	ABC transporter
Nodulation	NOLO	K00612 carbamoyltransferase EC 2.1.3
Nodulation	NOLG	Acriflavin resistance protein
Nodulation	NOLF	Efflux transporter, rnd family, mfp subunit
Nodulation	NOLF	Efflux transporter RND family MFP subunit