SYSTEMS BIOLOGY

Functional analysis of natural microbial consortia using community proteomics

Nathan C. VerBerkmoes, Vincent J. Denef‡, Robert L. Hettich* and Jillian F. Banfield‡*

Abstract | We know very little about the metabolic functioning and evolutionary dynamics of microbial communities. Recent advances in comprehensive, sequencing-based methods, however, are laying a molecular foundation for new insights into how microbial communities shape the Earth's biosphere. Here we explore the convergence of microbial ecology, genomics, biological mass spectrometry and informatics that form the new field of microbial community proteogenomics. We discuss the first applications of proteogenomics and its potential for studying the physiology, ecology and evolution of microbial populations and communities.

Consortia
A coexisting group of microbial populations.

Metaproteomics
The term metaproteomics is preferred for more partial, gene-centric approaches to community analysis.

Community proteomics
Application of proteomics beyond single isolate studies aimed at comprehensive system analysis.

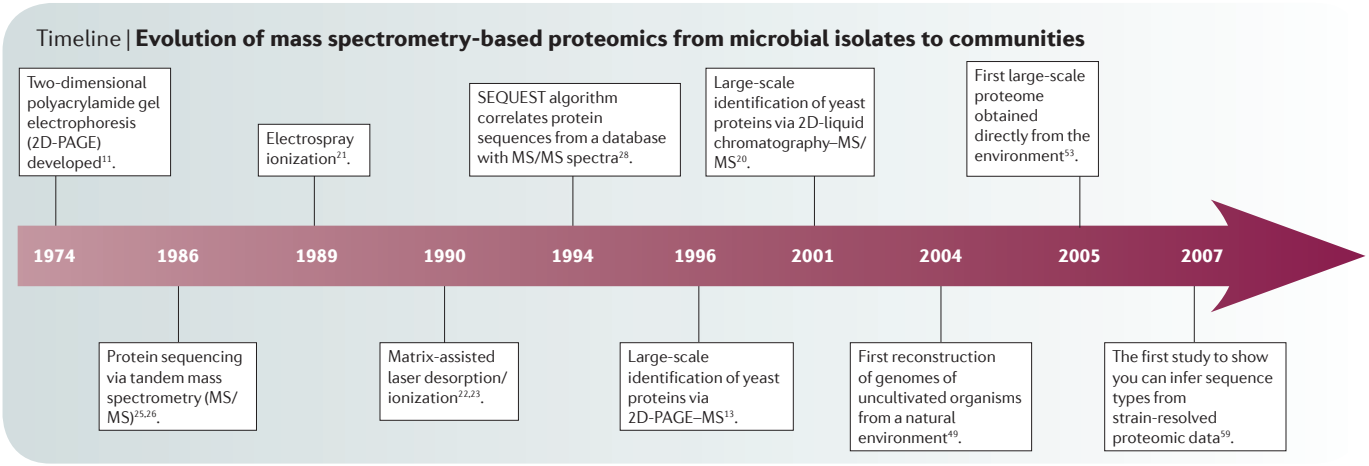
**Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA.*
‡University of California, Berkeley, California 94720, USA.
Correspondence to N.C.V.
e-mail:
verberkmoesn@ornl.gov
doi:10.1038/nrmicro2080

The ubiquity of microorganisms in Earth's near-surface environments has spurred a worldwide scientific effort to classify and understand microbial communities at the molecular level. Systems of interest include the human body[1], soils and the plant rhizosphere[2], the ocean[3,4], communities of biotechnological interest[5,6] and extreme environments, such as acid mine drainage (AMD)[7] and hydrothermal systems[8]. Because many microorganisms (estimates vary between 80–99%) cannot yet be cultured, and given that isolated strains might behave differently in culture than in their natural environments, there has been considerable interest in developing cultivation-independent methods to study microbial communities. Molecular fingerprinting techniques (for example, based on ribosomal RNA gene sequences) that were established in the 1980s to characterize community membership[9,10] have led to methods that focus on functional gene content and expression levels in microbial isolates and communities.

High-throughput molecular biology techniques, such as genome sequencing, have been key to the current renaissance in microbial ecology and physiology. The ability to sequence and annotate whole genomes of organisms, ranging from viruses to mammals, is well established — hundreds of microbial genome sequences have been completed and hundreds more are being characterized at a rapid pace (see the Integrated Microbial Genomes with Microbiome Samples website). This sequence information has enabled the analysis of the protein complements (proteomes) of organisms and consortia.

Here we describe the adaptation of isolate-based proteomics methods to study genomically uncharacterized or incompletely characterized natural consortia and review early progress. The approach simultaneously yields information about which members are active in a community, its detailed genetic composition and the biochemical pathways and mechanisms that are necessary for a community's survival.

The emerging fields of 'omics' technologies have spawned many new terms and acronyms. The terms metaproteomics, microbial community proteomics and microbial community proteogenomics are sometimes interchangeably used for different types of experiments and results. We would like to suggest some clarifications to these terms. In our view, metaproteomics should be used to classify experimental systems and results that are sufficiently complex that the genes and proteins that are identified from these communities cannot be binned into species or organism types. The metaproteomics approach is comparable to gene-centric environmental genomics, or metagenomics, approaches. This might be most obvious for complex systems, such as the soil or human microbiomes. Although genes and proteins can clearly be identified from such systems to moderate depths with current technologies, it is virtually impossible to confidently assign a large percentage of these identifications to a specific species or organism type. Thus, the experimental results provide a global view of the metabolic activity of the community, but cannot



properly delineate which members of a community are performing which metabolic functions. By contrast, genomics and proteomics of less diverse communities, such as the AMD system, can enable comprehensive analysis of defined populations.

The terms community proteomics and community genomics should be used to describe experiments in which most identified proteins can be mapped back onto the genomic context of individual species in the community, so the metabolic function of the members of the community can be ascertained. Community proteogenomics is a natural extension of community proteomics, going beyond the simple cataloguing of expressed proteins. Proteogenomics methods attempt to infer genomic information using candidate peptide sequences from genomic databases. Initially defined as a method to improve gene annotation, we here extend the term to include assessment of which strain variants or species are present in a sample and how the genomic makeup of a population changes over time. This aspect is crucial to a valid interpretation of community proteomic data, and as such we argue there can be no community proteomics without making these genomic inferences.

Microbial proteomics

Proteomics analyses comprise a suite of methodologies. In addition to protein cataloguing (determining the proteins present in a cell under given growth condition), proteomics can be used to evaluate how different growth states (as a function of time points or environmental settings) affect protein expression and determine protein localization, and to discover post-translational modifications (PTMs) and infer protein–protein interactions, amino-acid sequences and genotypes.

Proteomics is developing rapidly and cannot be defined by a single measurement platform, experimental approach or application. The first proteomics experiments were made possible by the invention of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)[11,12] over 30 years ago, but were limited by a lack of methods to identify the proteins that were contained in the observed spots. Simultaneously, developments in a seemingly completely unrelated field of mass spectrometry (MS) in the 1980s and 1990s ushered in the development of a new approach —MS-based proteomics (TIMELINE).

Microbial proteomics currently makes use of both gel-based (one dimensional and 2D)[13] and gel-independent liquid chromatography (LC)-based separations, each relying on MS-based peptide identification. In gel-based approaches the intact proteins are separated before an in-gel enzymatic digestion is performed to generate proteolytic peptides, whereas typical LC–tandem MS (MS/MS) methods are not carried out on intact proteins. Instead proteolytic peptides that have been separated from the complex sample are analysed — these approaches are commonly referred to as 'shotgun' or 'bottom-up' proteomics[14–17]. There are also LC–MS/MS approaches based on intact protein interrogation (termed 'top-down'); these in general do not provide the same depth of proteome coverage but can provide unique information about the molecular form of proteins, especially post-translational modifications[18]. The typical shotgun proteomics experiment involves three stages: sample preparation, LC–MS analyses and proteome informatics. Appropriate applications at all three stages are crucial for obtaining an overall quality proteome sampling, and these steps have been reviewed in detail[14–17].

Sample preparation generally begins with cellular lysis via sonication, French press or bead beating. To achieve comprehensive protein measurement, whole proteomes can be physically fractionated into cytoplasmic, membrane and extracellular samples through centrifugation techniques (if adequate biomass is available). Proteins are then denatured and reduced (to facilitate their effective proteolysis) and digested into peptides, usually with a highly specific protease (such as trypsin), although non-specific proteases are used for certain applications.

To reduce the complexity of the peptide mixture that enters the mass spectrometer simultaneously, a 2D LC-based separation is generally performed by combining strong cation exchange (based on charge) and reverse phase (based on hydrophobicity) columns[19,20]. After LC separation, charged peptides are transferred into the gas phase for MS measurements, either by electrospray ionization[21] for solution phase

Community proteogenomics
A dynamic interplay between community genomics and community proteomics, where the genomic makeup of the populations is inferred from the proteomics data, allowing for evolutionary analyses as well as for a valid interpretation of proteomics data from genomically uncharacterized samples.

Mass spectrometry-based proteomics
The application of mass spectrometry to proteome measurements.

Tandem mass spectrometry
The isolation, activation and fragmentation of peptides in mass spectrometers to obtain primary sequence information about the peptides.

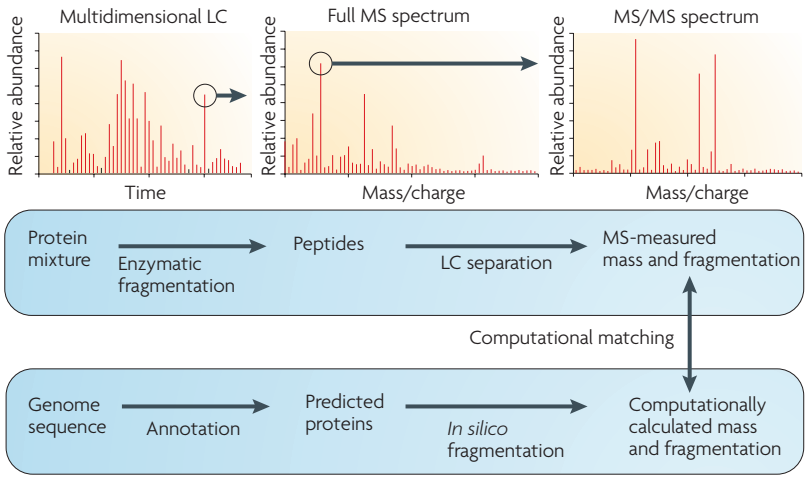


Figure 1 | **Liquid chromatography–mass spectrometry-based proteomics.** A protein mixture is enzymatically digested (for example, using trypsin) and the resulting peptide mixture is separated by multidimensional liquid chromatography (LC). Peptides are then assessed using mass spectrometry (MS) and then isolated and fragmented to obtain sequence information, by tandem MS (MS/MS). Spectra that are generated by MS are computationally matched to theoretical spectra generated from an *in silico* digest of a protein-sequence database, which itself is derived from the annotation of genomic sequence data.

samples, or deposited onto surfaces and mixed with a solid matrix, for analyses via matrix-assisted laser desorption/ionization (MALDI)[22,23]. Peptide and protein identification using MS methods became possible with the development of rapid scanning mass spectrometers capable of MS/MS measurements (also known as collision induced dissociation (CID)[24–27] on LC timescales.

Key to peptide and protein identification is the comparison of the measured intact masses and fragmentation patterns of the peptides with predicted intact masses and fragmentation patterns, which are generated *in silico* from genomic sequence data (FIG. 1). Comparisons make use of database search algorithms such as SEQUEST[28], MASCOT[29] and X-tandem[30]. Because LC–MS/MS-based identification of the peptides is so intimately tied to computational predictions, an entire field of proteome bioinformatics has emerged to tackle the flood of data that is created in typical proteomics studies.

The analysis of whole proteomes and protein complexes by MS can provide useful qualitative information, but the importance of accurate quantitative data for comparative analysis between temporally or spatially resolved samples cannot be overstated. Due to non-uniform ionization of peptides and/or increased loss of some peptides types (for example, hydrophobic peptides adhere to surfaces), direct quantification of shotgun proteomic data is challenging[31]. This can partially be addressed by stable isotope labelling, which enables quantification of the abundance of labelled protein in one growth condition relative to the abundance of unlabelled protein from another condition (for example, *Escherichia coli* grown in high salt versus low salt solutions).

Label-free methods represent an alternative approach for proteome quantification. These methods exploit intrinsic MS measurement metrics from unlabelled samples, such as peak intensities/areas of peptides[32], spectral counts[33] and normalized spectral abundance factors[34], to quantify peptides and proteins. They have grown in popularity due to their simplicity, low cost and applicability on any sample. Much effort is directed towards developing better tools and statistics for label-free methods[35–37].

The rapid advancements of multidimensional liquid-based separations coupled to rapid scanning MS/MS has enabled in-depth, accurate and quantitative studies of whole microbial proteomes[20,38–41]. The new generation of fast-scanning mass spectrometers[42,43] can identify several thousand proteins from a single growth state of a microbial isolate in 1–2 days[44–48]. With isolates grown under different metabolic conditions, it is possible to quantitatively compare thousands of proteins from the different conditions[44–48]. By analysing a microbial isolate in multiple growth states, between 50–90% of the predicted proteome can be identified.

Proteomics of microbial communities

Initial attempts to characterize the proteins that are expressed in microbial communities were hampered both by the limited resolution of 2D gel electrophoresis and by the lack of genomic information upon which to base peptide and protein identification (TABLE 1). The availability of the appropriate genomic sequence data was therefore key to the transition from isolate to community proteomics. In 2004, the genomes of co-existing members of an AMD biofilm community[49] and extensive genomic sampling of microorganisms in the Sargasso Sea[50] were published. A myriad of natural communities have now been sequenced. The resulting vast dataset is growing through the addition of genome sequences from individual microbial cells that, in some cases, were isolated from other more abundant community members[51].

An AMD biofilm was the first ecosystem for which genomic sequence information was available for the dominant species in the population (BOX 1a, top line). These chemoautotrophically-based biofilms grow underground in the Richmond Mine, California. As a consequence of their independence from other ecosystems, they represent a natural model system for community ecological studies. An intrinsic feature of the AMD biofilms is that they are dominated by only a handful of taxa. Genomes of the five dominant organisms were reconstructed using a modest sequencing allocation (initially 76 megabases[49]). Despite using a low energy resource (aerobic iron oxidation), microbial community growth is so prolific and abundant that discrete biofilms that grow at the air–solution interface can be sampled. These features made AMD biofilms an ideal system for development of cultivation-independent methods for the study of natural microbial communities and an obvious choice for the first community proteomics experiments.

Proteome bioinformatics
A subdiscipline in proteomics that is concerned with all methods of data analyses, validation, comparisons, statistics, dissemination and archival.

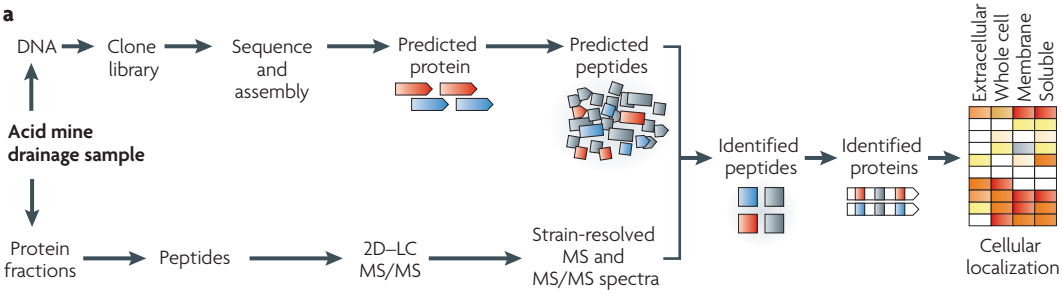
Table 1 | **Overview of meta- and community proteomic studies**

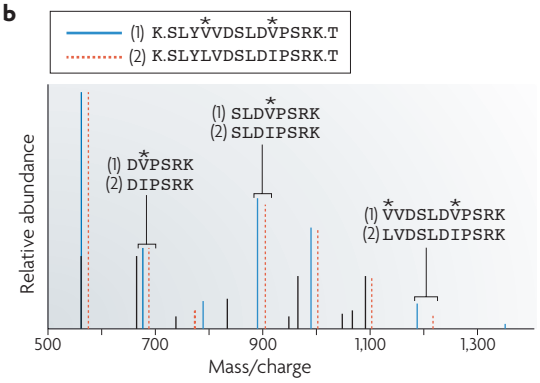
Microbiome	Number of peptides/ proteins identified*	Protein/peptide separation method	MS platform	Peptide identification method	Ref.
Ocean	184/NA‡	2D-PAGE, 2D nano-LC	LCQ, MS/MS	Spectral matching, de novo	76
Acid mine drainage	6,188§/2,033 (2p) 6,931§/5,090 (1p)	2D nano-LC	LTQ MS/MS	Spectral matching	52
Lake and soil	NA/513 (1p)‖	2D nano-LC	Q-ToF, MS/MS	Spectral matching	65
Estuary	7/3 (2p)	2D-PAGE + LC	Q-ToF, MS/MS	De novo	77
Ocean	3/1 (2p)¶	1D-PAGE	MALDI-ToF, MS	Spectral matching	62
Riftia symbionts	NA/220 (2p)¶	2D-PAGE, 1D-PAGE + 2D nano-LC	MALDI-ToF MS, Q-ToF MS/MS	Spectral matching	78
Infant gastrointestinal tract	11/1 (1p)	2D-PAGE	MALDI-ToF MS	De novo	79
Acid mine drainage	8,137§/3,234 (2p)	2D nano-LC	LTQ MS/MS	Spectral matching	59
Waste water treatment reactor	NA/109 (2p)¶	2D-PAGE	MALDI-ToF, MS/MS	Spectral matching, de novo	80
Contaminated soil/groundwater	NA/59 (1p)¶	1D + 2D-PAGE + LC	MS/MS	Spectral matching	66
Sludge	NA/46 (2p)¶	2D-PAGE	MALDI-ToF MS, Q-ToF MS/MS	Spectral matching	56
Sludge	4,472#/2,378 (2p)	2D nano-LC	LTQ MS/MS, Orbitrap, MS/MS	Spectral matching	57
Sludge EPS	50/10 (1p)¶	1D-PAGE + LC	4000Qtrap, MS/MS	Spectral matching	81
Ocean	6,533/1,042 (1p-2p)	2D nano-LC	LTQ MS/MS	Spectral matching	63
Acid mine drainage	NA/2,752** (2p)	2D nano-LC	Orbitrap, MS/MS	Spectral matching	60
Gut	NA/2,214 (2p)	2D nano-LC	Orbitrap, MS/MS	Spectral matching	64

*1p/2p filter indicates the need for at least one or two peptides to be identified to deem the corresponding protein identified. ‡Data was presented as matching membrane proteins or matching enzymes, with only specific matches given for four proteins. §Average number of peptides per MS/MS run. The complete experiment leading to the identification of the number of unique proteins included 13 MS/MS runs. ‖30% of the identified proteins were identified based on only one peptide. Number of proteins is cumulative over seven soil samples and one lake sample and might be partially redundant. ¶For MASCOT-based searches, other filters in addition to number of peptides were applied. #Average number of peptides per MS/MS run. The complete experiment leading to the identification of the number of unique proteins included four MS/MS runs. LTQ data are presented. **Average number of proteins identified in 27 samples, for each of which 3 MS/MS runs were performed. 2D-PAGE, two-dimensional polyacrylamide gel electrophoresis; EPS, extracellular polymeric substances; LC, liquid chromatography; LCQ, quadrupole ion trap; LTQ, linear ion trap; MALDI-ToF, matrix-assisted laser desorption/ionization-time of flight; MS, mass spectrometry; MS/MS, tandem MS; Q-ToF, quadrupole-time of flight.

The first large-scale proteomics experiment, aimed at obtaining a system-level snapshot of the protein-abundance levels, targeted an AMD biofilm for which no community genomic data were available[52]. Instead, peptide and protein identification (BOX 1a, bottom line) relied on the genomic sequence for a biofilm from the same system with similar microbial membership[49]. Despite differences between the sequences of predicted proteins in the dataset and those in the actual sample, it was possible to match shotgun MS/MS spectra to peptides and thus identify over 2,000 proteins from the 5 abundant populations (TABLE 1). Fifty percent of the proteome of *Leptospirillum* group II, the most abundant organism in the sample, was identified. The experiment confirmed the existence of 570 hypothetical proteins, suggested distinct functions for different organisms in the community and revealed genomic regions that show heterogeneous protein expression (for example, associated with integrated plasmid and phage). It was possible to directly map information about *Leptospirillum* group II protein localization (for example, based on detection of proteins in the extracellular or membrane fraction) and inferred abundance back into the genomic context (BOX 1). By combining abundance, localization and genomic context information, it was possible to formulate hypotheses regarding the function of some proteins of unknown function. Some highly abundant proteins of unknown function were targeted for detailed biochemical studies. Subsequent analysis revealed that two of these are cytochromes that are involved in iron oxidation (the

Box 1 | **Strain-resolved proteomics from environmental samples**





Although general genomics databases can be used, proteogenomics methods optimally rely on the generation of genomics and proteomics data from the same sample (see panel **a**). DNA is extracted from biological samples, fragmented, cloned and sequenced, and the resulting sequencing reads are assembled and/or binned. After gene annotation, the protein-sequence database is constructed and an *in silico* trypsin digest is performed on the predicted proteins, resulting in a peptide database (top). From the same or similar biological samples, total community protein is extracted and then digested using trypsin. Peptide separation by two-dimensional (2D) nano-liquid chromatography (LC) and tandem mass spectrometry (MS/MS) is performed (see FIG. 1) (bottom). The spectra are matched to peptides in the database, and after filtering a list of identified peptides is obtained. Based on their unique occurrence in one protein in the whole database, certain peptides (unique peptides, coloured red and blue) can be unambiguously tracked back to their corresponding proteins and thus permit reliable protein identification. Non-unique peptides (grey) cannot be used to uniquely identify a protein, but these data are used in the calculation of protein coverage and abundance measures. The identified proteins are placed back into the genomic context of the organisms they are derived from to allow for the biological mining of the data. In case the protein mixture was fractionated (into extracellular, soluble and membrane fractions) after the initial extraction, comparison of the fractionation data can provide information about protein localization.

The power conveyed by MS due to its high mass accuracy is demonstrated by overlaying two spectra from closely related peptides (see panel **b**) by highlighting the major peaks and by linking them to their corresponding amino-acid sequence (blue for one peptide-variant spectrum and red for another). In this example, a shift of around 14 daltons due to the substitution of valine by isoleucine and a second shift of an additional 14 daltons due to the substitution of valine by leucine are observed. Due to the high mass accuracy of the parent ion measurement and the MS/MS measurements, one amino-acid change between two orthologous peptides can easily be discerned, thus allowing strain-resolved proteomics. Asterisks represent amino acids that undergo substitution.

function that underpins the role of bacteria and archaea in driving AMD production). Both cytochromes were directly purified from natural AMD biofilms. Cyt_{572} was shown to be an unusual membrane-bound cytochrome[53] and Cyt_{579} was shown to be a soluble extracellular protein that transfers electrons that are derived from Fe(II) oxidation[54].

Currently, more than 30 proteomic datasets have been generated from samples taken in this AMD system. Together with species-abundance measurements made by FISH (fluorescent *in situ* hybridization) of ribosomal RNA probes, these data allow us to determine the relationship between species abundance and proteomics identification efficiency (FIG. 2). Saturation of the number of proteins identified, as well as the average protein coverage (percent of protein covered by identified peptides), emerges once the target organism constitutes more than 30–40% of the community. Below this threshold, both coverage and the number of proteins identified gradually decreases. However, even when an organism represents around 1% of a community, more than 100 of its proteins can be identified. Results from this model system can be used to constrain expected outcomes for other proteogenomic studies.

Proteomics has also been applied to study wastewater sludge microbial communities that are used for enhanced biological phosphorus removal (EPBR). As for AMD biofilms, EBPR samples are characterized by uneven species-abundance patterns; specifically, the community is dominated by organisms that are closely related to *Accumulibacter phosphatis*. Reference genomes for *A. phosphatis* in sludge samples from the United States and Australia [55] were used to identify MS/MS spectra for peptides from aerobic and anaerobic activated sludge communities from a UK bioreactor. Using 2D-PAGE, 46 proteins were identified from 111 excised spots[56]. Subsequent re-analysis using LC–MS/MS and methods developed for analysis of AMD biofilm proteomes

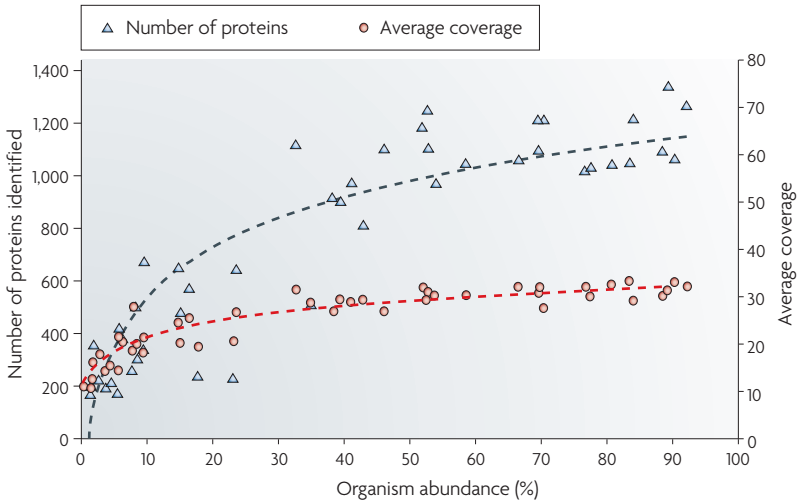


Figure 2 | **Relationship between species abundance, protein identification levels and protein coverage.** The plots were drawn based on the *Leptospirillum* group II and III data from 25 proteomics datasets from the acid mine drainage project[60]. Organism abundance levels were determined by FISH (fluorescent *in situ* hybridization). Increase in organism abundance has more effect on the number of proteins identified than on the average coverage. The left axis indicates the number of proteins (blue triangles) and the right axis refers to the average coverage (red circles). Please note the different scales for the two sides.

identified over 2,300 proteins[57], making extensive analysis of the metabolic pathways central to EBPR possible (TABLE 1).

Strain-resolved proteogenomics

Variation in the makeup of microbial communities over space and time, either due to changes in relative abundance of strain variants in populations or migration, presents a key challenge for proteomics studies. Because peptide identification depends on an exact match between the predicted and observed masses of peptides and their fragmentation products (FIG. 1; BOX 1), even a single amino-acid substitution (except for isoleucine–leucine and a small subset of other substitutions in experiments with lower mass accuracy measurements) will prevent peptide identification. For more abundant and larger proteins, this problem is less severe because protein identification relies typically on the identification of only two peptides (that is, failure to detect a few of the peptides due to amino-acid substitutions will not confound protein identification). The more diverged the protein is from the reference sequence, the more severe the problem[58]. Computational models predicted loss of half of the identifiable proteins when the organism in the sample diverges 8–23% from the available sequence data, and experimental data showed this to occur at 10% average amino-acid divergence (BOX 2).Thus, shotgun proteomics is capable of cross-strain identifications, but avoids most cross-species false positives. However, strain variation not only prevents peptide and potentially protein identification, it also confounds estimates of protein abundances that are based on spectral count per protein.

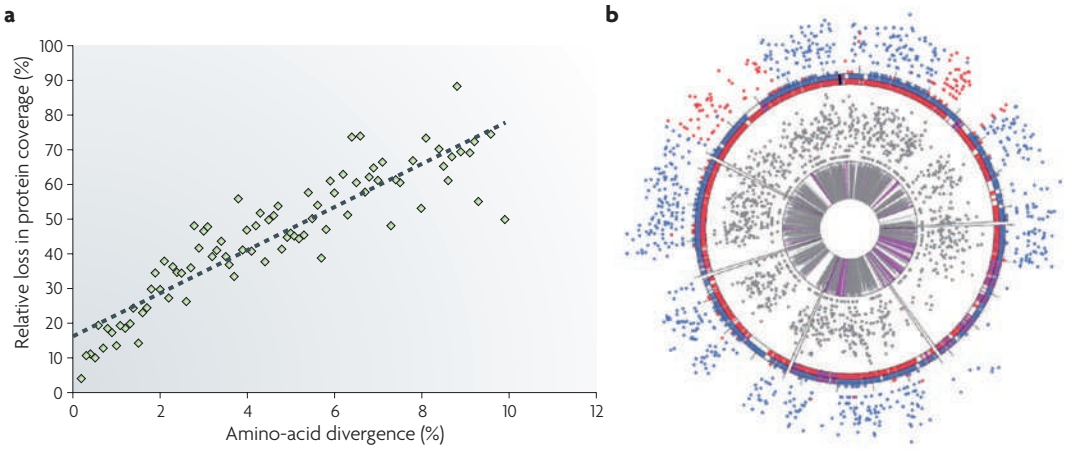
The sensitivity of MS measurements to amino-acid substitutions facilitates an important avenue for microbial (and other) community studies: strain typing, peptide by peptide. This approach was applied to the same AMD biofilm studied by Ram *et al.*[52] when a second genomic dataset from a third AMD biofilm community became available. The new dataset yielded a new genome for *Leptospirillum* group II. Each peptide for each predicted protein in the two *Leptospirillum* group II genomes was classified as either unique to one strain or shared by both strains. By mapping the peptides identified back into a genomic context it was possible to genotype the *Leptospirillum* group II type, peptide by peptide and protein by protein, in the proteomically characterized (but not genomically sequenced) sample (a process called proteomics inferred genome typing (PIGT)). This tight and reciprocal coupling between genomics and proteomics is recognized in the term proteogenomics. The first application of strain-resolved proteogenomics to a microbial community resulted in the identification of a new *Leptospirillum* group II genotype that represents a recombinant hybrid of the genomically characterized types[59] (BOX 2). The PIGT method was applied to a group of 28 community proteomics datasets to reveal several new recombinant *Leptospirillum* group II genotypes in the AMD system. The findings indicate environment-based selection for the dominant *Leptospirillum* group II type in each sample, suggesting that these bacteria use recombination as a strategy for fine-scale adaptation[60].

Strain-resolved proteogenomics analyses were also applied to the EBPR sludge samples, following methods used by Lo *et al.*[59] and making use of reference genomes from the two different previously reported EBPR samples[55]. Interestingly, substantial differences in protein abundance were found among *A. phosphatis* enzyme variants that are involved in both core-metabolism and EBPR-specific pathways. These differences in inferred activity may reflect partitioning of strain variants in microniches in the sludge[57].

Proteomics in more complex ecosystems

Both EBPR sludge and AMD biofilms are far less complex (due to dominance by one or a few populations) than other systems to which proteomics might be applied (for example water, soil sediments, the ocean and the human microbiome). The large number of unique taxa, each producing unique protein products at highly variable abundance levels (dynamic range) create numerous technical challenges, some of which are similar to those faced in proteomics studies of multicellular organisms. Complete detection of the majority of unique protein products in a sample may be impossible, even in systems of only moderate complexity[61] This raises several questions. How realistic is it to try to achieve deep proteomic sampling of complex communities? Can measurements of thousands of proteins using current technologies provide useful information if the sample contains hundreds of thousands or millions of unique proteins? Will it ever be possible to confidently determine which proteins the measured peptides come from and, more importantly, from which organisms they derive? How can we quantitatively compare the abundances of proteins among complex environmental samples, when protein abundances, organism abundances and genotypes differ? Will it be possible to reveal the dynamics of metabolic networks

Box 2 | **Implications of sequence divergence to proteogenomics**



Microbial life is extraordinarily diverse and sequence variation is an inherent characteristic of most natural species populations. For proteomics applications, differences in amino-acid sequences between closely related organisms in a natural community can be looked at from two perspectives.

On one hand, amino-acid substitutions decrease the efficiency of mass spectrometry (MS)-based protein identification, as most MS-based proteomics is based on identical matches between the measured peptides and the database entries. Computational models predicted loss of half of the identifiable proteins when the organism in the sample diverges 8–23% from the available sequence data. Experimental data indicates loss of half the identifiable proteins at 90% amino-acid identity, due to a ~6% reduction of the coverage of proteins by identified peptides per 1% amino-acid divergence (see panel **a**). Panel **a** constructed from data in REF. 58.

On the other hand, the influence of sequence variation on proteomics can be used to discriminate subtly different protein sequences from closely related organisms. This allows for the resolution of the behaviour of closely related organisms when they co-occur in the same environment (V.J.D., unpublished observations). Strain-resolved proteomics holds promise for unravelling the ecological significance of sequence divergence between strains of one species. This method has also been used for genotyping and has revealed that recombination occurs between closely related sequence types[59,60] (see panel **b**). Panel **b** provides an example of a proteomics-inferred genome typing (PIGT) dataset. Unique peptide counts — peptides with a sequence that is unique in the whole search database — allow researchers to discriminate between closely related variants of a particular protein. Protein sequences from *Leptospirillum* group II strain 1 (blue) and *Leptospirillum* group II strain 2 (red) detected in the sample are plotted in the outside scatter plot. These surround the comparative genomic representation of the two *Leptospirillum* group II genomes — white space indicates that no orthologue is present in one of the two types, red and blue indicate that a gene is present at that locus and purple indicates that the orthologues are 100% identical. Grey represents proteins identified from non-unique peptides that thus cannot be distinguished between the two strains. We observed regions where only strain 1 type proteins were identified, alternating with 10–100 kb regions where only strain 2 type proteins were found. This indicated that the strain present in the analysed sample was a recombinant variant of both strains. Reproduced with permission from REF. 60 (2009) Blackwell Publishing.

and truly begin to understand how these communities function at the molecular level? We feel that high complexity generates problems that cannot be adequately overcome using current technologies. As we will discuss below, the field is moving quickly and we anticipate that the technical challenges can be identified and overcome so that proteomics-based functional analyses of microbial communities will become a reality.

While several metaproteomics projects have attempted to characterize proteins in environmental samples representing complex communities, few such projects have identified more than a handful of proteins (TABLE 1). Generally, 2D LC–MS/MS-based methods are better for community or metaproteomics studies, although SDS-PAGE-based separation is sometimes preferred[62]. A study of the dominant populations (SAR11, *Prochlorococcus* and *Synechococcus*) in an ocean sample uncovered a high bias in expression of periplasmic substrate-binding proteins in SAR11. This was inferred to represent a means to maximize nutrient acquisition in the highly nutrient-depleted environment[63]. The availability of extensive metagenomic datasets from ocean samples also allowed for the strain-resolved analysis of periplasmic phosphate-binding proteins (PtsS), which revealed differences in protein abundance between different PtsS subclades. This aspect of the study is an example of the importance of the dynamic interplay between proteomics and genomics, and highlights the need for new analysis methods that resolve the proteomics signal from closely related protein variants on a community-wide scale.

One interesting, emerging research area for microbial community proteomics studies is the Human Microbiome Project, which is focused on a characterization of the suite of microbial species that are intimately tethered to human hosts. For example, the human gut contains a dense, complex and diverse microbial community (termed the gut microbiome) that is critical for both health and disease. Substantial efforts have been

directed at large-scale DNA sequencing to characterize the human microbiome. While this metagenome information will provide details about the repertoire of genes that are present, it provides no direct information about which genes are expressed or functioning. To investigate whether proteomics would even be possible in the complex human gut microbiome, this whole community proteomics approach was used to study the microbial consortium in faecal samples from matched human twins[64]. This approach was successful for deep proteome measurements of thousands of proteins from the microbial membership. Several unknown proteins represented previously undiscovered microbial pathways, revealing a novel and complex interaction between the human host and the associated microorganisms. While these results are preliminary and must be extended for more comprehensive evaluation of normal versus diseased states, these initial results suggest that proteomics approaches are viable for human microbiome studies.

As highlighted above, biodiversity (species richness and abundance) will be a key factor in defining challenges for studies of complex systems. Are there a few dominant organisms, or are all organisms present at comparable abundance levels? Are reference isolate genomic sequences or relevant metagenomics datasets available? Analysis of the dominant proteomes and of genomically-defined members of extremely complex samples is straightforward with current techniques; the same analysis for the low-level members remains very challenging. Although cell-enrichment techniques can enable sampling of proteomes of low-abundance members, potential artefacts associated with proteome changes during processes such as cell sorting or filtration cannot be ignored. In the longer term, the availability of parallel metagenomes is anticipated to be less of a problem as extensive sequencing efforts that are now underway will provide reference isolate genome sequences for under sampled branches of the tree of life (see The Tree of Life).

Another important challenge in environmental proteomics is the need for efficient and non-biased extraction of proteins from complex environmental matrices. Initial studies of AMD biofilm communities compared lysis by sonication with other extraction methods that made use of different pH buffers. Biases appear unavoidable, but can sometimes be advantageous. For example, the proteomes of low-abundance community members can sometimes be preferentially sampled because of biased lysis.

Additional challenges for quantitative protein extraction can arise due to the sample matrix, specifically for soil and sediment samples. Most initial proteomics studies in soil have found that efficient cell lysis and protein extraction is problematic[65,66]. Mineral assemblages act as mixed-bed chromatography columns that can irreversibly bind peptides and proteins. Consequently, only a handful of proteins have been identified or quantified from soil[65,66] and methodological challenges have limited biological insights into soil systems. Due to the extraordinarily high levels of species richness and problems with protein recovery, soils are perhaps the grand technical challenge for the emerging field of microbial community proteomics.

Another major technical challenge for community proteogenomics is the dynamic range (see also REF. 61) — that is, the range of abundances (low to high) of proteins that can be detected simultaneously. While dynamic range can be problematic for microbial isolate studies, it is significantly more so for mixtures of organisms due to the uneven abundances of different organisms (for example, if the dominant population contributes 10% of the community, whereas lowly abundant organisms only contribute 0.1%). Fortunately, the past 5 years have seen the dynamic ranges for standard proteomics of LC–MS systems increase by 1–2 orders of magnitude. This advance has been achieved through the refinement of chromatographic methods and by coupling better peptide separations to new generation MS instruments that can retain large quantities of ions in the system, thus generating better sensitivity[42,43]. The integration of high-throughput MS/MS methodologies with high-performance mass spectrometers that are capable of high-resolution (peak widths less than 0.01 daltons) and high-mass accuracies (to a few millidaltons) holds great promise for future microbial community proteomics measurements.

An important consideration that impacts the outcome of proteomics experiments is quality of genome annotation. Missing protein predictions or predictions of hypothetical proteins in the wrong reading frame will preclude protein identification. However, the existence of a protein database in which all hypothetical proteins are predicted in all reading frames makes it possible to use proteomics identification to validate gene predictions[52,67–70]. Incorrect gene starts and failure to predict signal peptides can also affect protein identification, especially for low-abundance proteins for which few peptides will be identified. Refinement of gene predictions and confirmation of signal peptide predictions also represents important feedback from proteomics and genomics studies.

The use of fragmentary metagenomics datasets will be restricted if sequences cannot be accurately assigned to the correct organism. Although the advent of new DNA sequencing technologies generating vast quantities of DNA sequence will enable proteomics studies of numerous systems, high-throughput accurate binning methods will be essential for data interpretation. Developments in sequence-signature analysis hold great promise for resolving this challenge[71,72], and proteomics may also help. For example, statistical analysis of proteomics data from communities with highly skewed membership may assist in resolving which genome fragments belongs to which organisms.

As the field advances from protein identification/validation and qualitative analysis to tackle detailed biological questions, accurate quantification of proteins across a wide dynamic range will be important. Quantification methods described above and developed for microbial isolates can be applied to consortia, but an additional consideration is community membership. For example, changes in the concentration of species in a time-series experiments will alter protein abundance levels. Consequently, monitoring community makeup (for example, using FISH is essential so that overall activity levels and cell abundance levels can be differentiated.

Binning methods
Methods used in metagenomics to group sequencing reads and assembled sequence contigs by the organism that they come from.

De novo sequencing or sequence tagging
Attempting to obtain full or partial sequences directly from tandem mass spectra without the use of a genome or proteome database.

Correct peptide identification lies at the heart of effective proteomics methods. Typical 2D–LC–MS/MS experiments generate tens of thousands of tandem MS spectra per day, a fraction of which are dominated by noise from the instrument or chemicals in the solvent gradient. Search algorithms produce identifications for every MS/MS spectrum, so filtering methods must be used to reduce the number of false positive peptide identifications. The careful balance of false positives and false negatives is essential but complex, and there are many computational techniques for solving this problem[73,74]. False positive identification rates depend highly on the accuracy of the measurement of the masses of the primary peptides and fragment ions. Instruments with high mass accuracies that are sufficient to discriminate all amino acids for which a mass difference exists will be essential to limit false discovery rates of peptides and proteins in studies of complex microbial communities.

As noted above, the move from genomically defined isolates to incompletely defined complex environmental communities is complicated by false positive peptide identifications that occur when spectra are assigned to completely different peptides that happen to have high scores. This problem can be solved if peptides can be identified from the spectra themselves (without a reference sequence). The advent of high mass accuracy MS systems makes *de novo* sequencing or sequence tagging possible[75]. *De novo* sequencing methods are designed to determine peptide sequences directly from the MS/MS spectra, whereas sequence tagging approaches produce a short tag that can be matched against a database. Both techniques are still in development, and most methods use these approaches in combination. Availability of alternative methods for fragmenting peptides in the mass spectrometer will provide additional mass constraints for peptide-sequence determination and are thus likely to be important for the successful deployment of *de novo* approaches.

A final grand challenge for proteomics studies of complex microbial communities will be the development of methods to analyse small quantities of sample. This will enable characterization of finely spatially resolved environmental samples and analysis of systems with limited biomass. Current techniques require hundreds of milligrams of wet cellular material for standard lyses and fractionation and 1–10 mg of wet cellular material for small-scale lysis without fractionation. To decrease sample size requirements by an order of magnitude may require a shift away from, for example, lyses in microfuge tubes and chromatography in low-flow columns, towards cell lysis, protein digestion and peptide separation on microchip-based platforms directly connected to MS systems. Although this is not likely to occur in the next few years, significant research in this area holds promise for commercial chip-based platforms in the foreseeable future.

Conclusions

Although the field of community proteomics is in its infancy, it is becoming clear that the technology developed for microbial isolates can be extended into these more complicated systems. Current technology can identify proteins of populations that comprise at least 1% of the community and for which closely related genomic sequences are available, but future technological improvements should make community proteomics applicable to less abundant populations. This will allow the extension of current studies of community functional partitioning, resource competition and strain-resolved adaptation. Protein-centric metaproteomics approaches will also become important tools to study microbial ecosystems. However, we need to focus on community proteogenomics, where simultaneous inferences regarding the dynamics of the genotype and identified proteins are made. Only by dynamically integrating genomics and proteomics data can we make meaningful quantitative comparisons of community proteomics data between different samples and ecosystems.

1. Turnbaugh, P. J. *et al*. The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
2. Kent, A. D. & Triplett, E. W. Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu. Rev. Microbiol.* **56**, 211–236 (2002).
3. DeLong, E. F. Marine microbial diversity: the tip of the iceberg. *Trends Biotechnol.* **15**, 203–207 (1997).
4. DeLong, E. F. Microbial community genomics in the ocean. *Nature Rev. Microbiol.* **3**, 459–469 (2005).
5. Rittmann, B. E. *et al.* A vista for microbial ecology and environmental biotechnology. *Environ. Sci. Technol.* **40**, 1096–1103 (2006).
6. Daims, H., Taylor, M. W. & Wagner, M. Wastewater treatment: a model system for microbial ecology. *Trends Biotechnol.* **24**, 483–489 (2006).
7. Baker, B. J. & Banfield, J. F. Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* **44**, 139–152 (2003).
8. Huber, J. A. *et al.* Microbial population structures in the deep marine biosphere. *Science* **318**, 97–100 (2007).
9. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337–365 (1986).
10. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
An overview of microbial diversity on earth.
11. O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021 (1974).
12. Klose, J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243 (1975).
13. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA* **93**, 14440–14445 (1996).
14. Peng, J. & Gygi, S. P. Proteomics: the move to mixtures. *J. Mass Spec.* **36**, 1083–1091 (2001).
An excellent review of proteomics methodologies for analysing complex protein mixtures.
15. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473 (2001).
16. Liu, H., Lin. D. & Yates, J. R. Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotech.* **32**, 898–902 (2002).
17. VerBerkmoes, N. C., Connelly, H. M., Pan, C. & Hettich, R. L. Mass spectrometric approaches for characterizing bacterial proteomes. *Expert Rev. Proteomics* **1**, 433–447 (2004).
18. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nature Methods* **4**, 817–821 (2007).
19. McCormack, A. L. *et al.* Direct analysis and identification of proteins in mixtures by LC–MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776 (1997).
20. Washburn, M. P., Wolters, D. A. & Yates III, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotech.* **19**, 242–247 (2001).
The first large-scale identification of a microbial proteome via shotgun proteomics or 2D–LC–MS/MS.
21. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
The first description of electrospray ionization. Fenn was awarded the Nobel Prize for this discovery.
22. Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* **63**, 1193A–1203A (1991).
23. Nakanishi, T., Okamoto, N., Tanaka, K. & Shimizu, A. Laser-desorption time-of-flight mass-spectrometric analysis of transferrin precipitated with antiserum — a unique simple method to identify molecular-weight variants. *Biol. Mass Spec.* **23**, 230–233 (1994).
24. Hunt, D. F., Buko, A. M., Ballard, J. M., Shabanowitz, J. & Giordani, A. B. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed. Mass Spectrom.* **8**, 397–408 (1981).

25. Hunt, D. F., Yates, J. R. 3rd, Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl Acad. Sci. USA* **83**, 6233–6237 (1986).
26. Biemann, K. Mass spectrometric methods for protein sequencing. *Anal. Chem.* **58**, 1288A–1300A (1986).
27. Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biomed. Environ. Mass Spectrom.* **16**, 99–111 (1988).
28. Eng, J. K., McCormack, A. L. & Yates III, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
29. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
30. Craig R. & Beavis R. C. TANDEM: matching proteins with mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
31. Ong, S. E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nature Chem. Methods.* **1**, 252–262 (2005).
A review of quantitative proteomics methodology.
32. Old, W. M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell Proteomics* **4**, 1487–1502 (2005).
33. Liu, H., Sadygov, R. G. & Yates, J. R. 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
34. Florens, L. *et al.* Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**, 303–311 (2006).
35. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics* **4**, 1265–1272 (2005).
36. Zhang, B. *et al.* Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* **5**, 2909–2918 (2006).
37. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotech.* **25**, 117–124 (2007).
38. Lipton, M. S. *et al.* Global analysis of *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* **99**, 11049–11054 (2002).
39. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multi-dimensional chromatography coupled with tandem mass spectrometry (LC/LC–MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
40. Corbin, R. W. *et al.* Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA* **100**, 9232–9237 (2003).
41. VerBerkmoes, N. C., *et al.* Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states. *J. Proteome Res.* **5**, 287–298 (2006).
42. Schwartz, J. C., Senko, M. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **13**, 659–669 (2002).
43. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443 (2005).
44. Brown, S. D. *et al.* Molecular dynamics of the *Shewanella oneidensis* response to chromate stress. *Mol. Cell Proteomics* **5**, 1054–1071 (2006).
45. Chourey, K. *et al.* Global molecular and morphological effects of 24-h chromium exposure on *Shewanella oneidensis* MR-1. *App. Environ. Microbiol.* **72**, 6331–6344 (2006).
46. Callister, S. J. *et al.* Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from *Rhodobacter sphaeroides* 2.4.1. Aerobic and photosynthetic cell cultures. *J. Proteome Res.* **5**, 1940–1947 (2006).
47. Thompson, M. R. *et al.* Dosage-dependent proteome response of *Shewanella oneidensis* MR-1 to acute chromate challenge. *J. Proteome Res.* **6**, 1745–1757 (2007).
48. Luo, Q. *et al.* Proteome analysis of *Desulfovibrio desulfuricans* G20 mutants using the accurate mass and time (AMT) tag approach. *J. Proteome Res.* **8**, 3042–3053 (2007).
49. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
First reconstruction of complete microbial genomes from the natural environment.
50. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
51. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104**, 11889–11894 (2007).
52. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920 (2005).
First large-scale measurement of proteomes from a microbial community in the natural environment.
53. Jeans, C. *et al.* Cytochrome 572 is a conspicuous membrane protein with iron oxidation activity purified directly from a natural acidophilic microbial community. *ISME J.* **2**, 542–550 (2008).
54. Singer, S. W. *et al.* Characterization of cytochrome 579, an unusual cytochrome isolated from an iron oxidizing microbial community. *Appl. Environ. Microbiol.* **74**, 4454–4462 (2008).
55. García, M. H. *et al.* Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotech.* **24**, 1263–1269 (2006).
56. Wilmes, P., Wexler, M. & Bond, P. L. Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS ONE* **3**, e1778 (2008).
57. Wilmes, P. *et al.* Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* **2**, 542–550 (2008).
58. Denef, V. J., Shah, M. B., VerBerkmoes, N. C., Hettich, R. L. & Banfield, J. F. Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *J. Proteome Res.* **6**, 3152–3161 (2007).
59. Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537–541 (2007).
Evidence of recombination of large genomic regions obtained via strain-resolved proteomics. Developed the new approach of proteomics-inferred genome typing, more extensively applied in reference 60.
60. Denef, V. J. *et al.* Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environ. Microbiol.* **11**, 313–325 (2009).
61. Wilmes, P. & Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **14**, 92–97 (2006).
62. Giovannoni S. J. *et al.* Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**, 82–85 (2005).
63. Sowell S. M. *et al.* Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* **3**, 93–105 (2009).
64. VerBerkmoes. N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **3**, 179–189 (2009).
65. Schulze, W. Z. *et al.* A proteomics fingerprint of dissolved organic carbon and of soil particles. *Oecologia* **142**, 335–343 (2005).
66. Benndorf, D., Balcke, G. U., Harms, H. & von Bergen, M. Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J.* **1**, 224–234 (2007).
67. Jaffe J. D., Berg H. C. & Church G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**, 59–77 (2004).
68. Savidor, A. *et al.* Expressed peptide tags: an additional layer of data for genome annotation. *J. Proteome Res.* **5**, 3048–3058 (2006).
69. Fermin, D. *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **7**, R35 (2006).
70. Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S. & Smith, R. D. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct. Genomic Proteomic* **7**, 50–62 (2008).
71. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163 (2004).
72. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S. & Ikemura, T. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.* **12**, 281–290 (2005).
73. Elias, J. E. & Gygi, S. P. Target–decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).
74. Tabb, D. L. What's driving false discovery rates? *J. Proteome Res.* **7**, 45–46 (2008).
75. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **4**, 787–797 (2007).
Review of proteome informatics methods, from database searching to *de novo* sequencing.
76. Powell, M. J., Sutton, J. N., Del Castillo, C. E. & Timperman, A. T. Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. *Marine Chem.* **95**, 183–198 (2005).
77. Kan, J., Hanson, T., Ginter, J., Wang, K. & Chen, F. Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems* **1**, 7 (2005).
78. Markert, S. *et al.* Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* **315**, 247–250 (2007).
79. Klaassens, E. S., de Vos, W. M. & Vaughan, E. E. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl. Environ. Microbiol.* **73**, 1388–1392 (2007).
80. Lacerda, C. M. R., Choe, L. H. & Reardon, K. F. Metaproteomic analysis of a bacterial community response to cadmium exposure. *J. Proteome Res.* **6**, 1145–1152 (2007).
81. Park, C., Novak, J. T., Helm, R. F., Ahn, Y.-O. & Esen, A. Evaluation of the extracellular proteins in full-scale activated sludges. *Water Research* **42**, 3879–3889 (2008).

Acknowledgements

Funding was provided by the United States Department of Energy: Genomics: Genomes-to-Life Program, the National Science Foundation Biocomplexity Program and the NASA Astrobiology Institute. B. R. Maggard is thanked for secretarial assistance in the preparation of this manuscript. Oak Ridge National Laboratory is managed by University of Tennessee–Battelle LLC for the Department of Energy under contract DOE-AC05-00OR22725.

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Escherichia coli | *Accumulibacter phosphatis*

FURTHER INFORMATION

Organic & Biological Mass Spectrometry Group: http://www.ornl.gov/sci/csd/Research_areas/obms_group.html
Jillian Banfield's homepage: http://eps.berkeley.edu/development/view_person.php?uid=185017&page=22
DOE Genomics:GTL: http://quicksilver.espm.berkeley.edu
Integrated Microbial Genomes with Microbiome Samples: http://img.jgi.doe.gov/cgi-bin/m/main.cgi
The Human Microbiome Project: http://nihroadmap.nih.gov/hmp
The Tree of Life: http://www.tigr.org/tol

ALL LINKS ARE ACTIVE IN THE ONLINE PDF