Next-generation approaches to the microbial ecology of food fermentations

Nicholas A. Bokulich1,2,3 & David A. Mills1,2,3*
1Department of Viticulture and Enology, 2Department of Food Science and Technology, 3Foods for Health Institute, University of California, Davis, CA 95616, USA

Food fermentations have enhanced human health since the dawn of time and remain a prevalent means of food processing and preservation. Due to their cultural and nutritional importance, many of these foods have been studied in detail using molecular tools, leading to enhancements in quality and safety. Furthermore, recent advances in high-throughput sequencing technology are revolutionizing the study of food microbial ecology, deepening insight into complex fermentation systems. This review provides insight into novel applications of select molecular techniques, particularly next-generation sequencing technology, for analysis of microbial communities in fermented foods. We present a guideline for integrated molecular analysis of food microbial ecology and a starting point for implementing next-generation analysis of food systems. [BMB Reports 2012; 45(7): 377-389]

INTRODUCTION

Fermented foods have constituted a significant proportion of the human diet globally since prehistory. These foods benefit consumers through enhanced nutritional content, digestibility, microbial stability, and in some cases detoxification (1-5), likely representing mankind’s oldest means of food preservation. In addition, these foods often serve as vehicles for beneficial microorganisms that play a beneficial role in human health, as well as prebiotic substances, which promote the growth and health-modulating activities of beneficial microbes in the human body (6). These foods are woven into the cultural fabric of the people who create them, and microbial fermentation remains a prominent means of food processing, enrichment, and preservation to the current day. Given the continued importance of fermented foods to the global population, much effort has been put towards describing the microbial communities responsible for different food fermentations to better improve and manipulate their production and safety. Molecular methods, relying on detection of nucleic acid sequences, have replaced traditional, culture-based detection methods for microbial community profiling due to their greater speed and accuracy—as most culture media bias for the growth of specific organisms, leading to distorted perspectives of microbial communities, even in simple microbial systems, such as food fermentations (7, 8). Additionally, the stressful conditions of some food systems, particularly alcoholic fermentations, can induce a viable-but-not-culturable state in microorganisms, preventing culture-based detection (9, 10).

This review will discuss current advances in molecular technology for studying food fermentations worldwide and present strategies for continued studies of global fermentations. This will not be a comprehensive review of all molecular profiling techniques, but will contrast a few leading technologies and, in particular, focus on the cutting-edge DNA sequencing technologies that are revolutionizing the study of food microbial ecology (Table 1).

MOLECULAR TOOLS

Targeted tools

A targeted profiling method is any technique that detects a taxonomically defined group of microbes, e.g., all bacteria, a specific species, or a specific strain. A multitude of targeted techniques exist, but the two discussed below are the principal methods employed in modern microbial ecology, particularly of foods. The principal issue of these and all targeted methods is that they can only quantify populations targeted by the selected primer or probe, so cannot distinguish subpopulations (without using additional probes) and cannot detect non-target populations. Therefore, these are not comprehensive profiling techniques (i.e., they do not characterize a complete microbial ecosystem), but are used to track functionally important members of a given ecosystem.

Fluorescence in situ hybridization (FISH)

Fluorescence in situ hybridization (FISH) is a culture-independent, PCR-free technique for directly visualizing microbial cells in a sample. This method utilizes fluorescence-labeled oligonucleotide probes targeting specific DNA sequences, usually re-
Table 1. Comparison of molecular tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Target size (bp)</th>
<th>Sensitivity</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>FISH</td>
<td>15-30</td>
<td>$10^2$-$10^6$</td>
<td>PCR-free, direct visualization in situ</td>
<td>Targeted technique</td>
</tr>
<tr>
<td>QPCR</td>
<td>100-150</td>
<td>$10^1$-$10^3$</td>
<td>Sensitive, quantitative</td>
<td>Targeted technique</td>
</tr>
<tr>
<td>DGGE</td>
<td>100-300</td>
<td>$10^2$-$10^3$</td>
<td>Single-nucleotide fragment separation</td>
<td>Technically difficult</td>
</tr>
<tr>
<td>TRFLP</td>
<td>1,000+</td>
<td>$10^2$</td>
<td>High-throughput, pseudoquantitative</td>
<td>Lower taxonomic resolution</td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAIhx</td>
<td>35-150</td>
<td>$10^1$</td>
<td>Superior sequence coverage, lowest</td>
<td>Shorter fragment length, theoretically</td>
</tr>
<tr>
<td>HiSeq2000</td>
<td>35-150</td>
<td>$10^1$</td>
<td>sequencing cost</td>
<td>Lower taxonomic resolution</td>
</tr>
<tr>
<td>MiSeq</td>
<td>35-150</td>
<td>$10^1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>454 Life Sciences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLX Titanium</td>
<td>300-400</td>
<td>$10^1$</td>
<td>Longer fragment length, theoretically higher</td>
<td>Low coverage, expensive</td>
</tr>
<tr>
<td>FLX+</td>
<td>500-600</td>
<td>$10^1$</td>
<td>taxonomic resolution</td>
<td></td>
</tr>
</tbody>
</table>

*aMaximum target size in bp. bSensitivity as limit of detection (cells ml$^{-1}$) (FISH, QPCR, DGGE, TRFLP) or reads per run (Illumina, 454). All values are current as of the time of writing.

New generation approaches to the microbial ecology of food fermentations
Nicholas A. Bokulich and David A. Mills

Quantitative PCR (QPCR)
Quantitative PCR (QPCR, also known as real-time PCR) has become an extremely popular technique for food microbiology testing and microbial ecology studies. In this review we will focus on novel applications of QPCR for the study of food microbiology and how advances in bioinformatics are increasing the utility of this and other molecular techniques. For a more thorough review of general methodology and applications of QPCR to food microbiology and food testing, the reader should consult the recent review of Postollec et al. (21).

QPCR incorporates the same principles of DNA replication used by PCR, but amplification is tracked in real-time using a fluorescent reporter molecule. A baseline threshold is determined, at which sample fluorescence can be distinguished from background noise. This is used to determine the threshold cycle (C) for each sample, at which fluorescence crosses the baseline threshold. A standard curve is constructed plotting C against gene copy number or cell count, and used to quantify unknown samples based on C. The result is an elegantly simple, rapid, and accurate means of tracking target cell or gene counts in samples.

QPCR has seen widespread use for quantification of microbial populations in foods (21) and can be used to profile complex, mixed-culture fermentations. Beyond cell quantification, the detection of functional genes presents a novel use of this technique with interesting applications to food microbiology. For example, this technique has been used to detect biogenic amine-producing strains of lactic acid bacteria (LAB) in wines (22, 23), ciders (24), and cheeses (25, 26) by directly targeting the genes responsible for biogenic amine production. Ibarburu and coworkers (27) have developed an assay to quantify exopolysaccharide (EPS)-producing LAB (which cause slime-production, a defect in wines and other beverages) by targeting the glycosyltransferase gene controlling EPS synthesis. QPCR of sulfite-reductase genes in Saccharomyces cerevisiae (28) can be used to track sulfide

tation, e.g., cheese, vegetable, meat, or seafood fermentations.
production, a major off-flavor concern in wine fermentations. QPCR has also been used to detect beer-spoilage LAB by targeting the horA gene, which promotes resistance to hop antimicrobial iso-α-acids (29). This general approach detects populations based not on taxonomic distinction, but instead by specific genes with functions relevant to food properties. Another novel application of QPCR is the use of strain-specific primers/probes (30). Provided a strain-specific marker can be identified, this enables high-throughput, sensitive profiling of strain-level successions in food fermentations (30), which cannot be characterized by most culture-independent molecular profiling tools. Strain-specific QPCR is likely to increase as the advent of next generation sequencing (NGS) technologies has enabled researchers worldwide to rapidly generate draft genome sequences for any target microbial strain.

**Community profiling tools**

Culture-independent, molecular techniques for microbial profiling employ universal primers targeting a taxonomically defined group (usually kingdom-level phyla) to amplify heterogeneous sequences of DNA from mixed microbial samples before separating and taxonomically classifying the sequences, giving a perspective of the complete community structure of the target population. The common issue with these techniques is that they can only provide a qualitative or pseudoquantitative (i.e., relative abundance) assessment of a microbial community, so must be paired with quantitative, targeted molecular techniques to enumerate cells (i.e., absolute abundance). While a large number of profiling methods exist, this review will focus on the three most popular methods for microbial community profiling used today.

**Denaturing gradient gel electrophoresis (DGGE)**

Denaturing gradient gel electrophoresis (DGGE) has received a great deal of attention from food microbiologists in the past 20 years since it was first adopted for studying bacterial communities (31). DGGE and its close relative, temperature gradient gel electrophoresis (TGGE), function by separating short fragments of DNA along a chemical or temperature gradient, respectively. DNA is propelled through a polyacrylamide gel matrix by an electric current until it reaches the point along the gradient at which it becomes destabilized and partially denatures (held together by a high-stability GC-motif, or "GC tail"), becoming immobilized in the gel. The most abundant of these fragments may then be visualized on the gel under UV light, physically extracted, reamplified, and sequenced to qualitatively determine the most dominant organisms in a given sample.

In spite of its widespread popularity, DGGE is hindered by several issues that make it unsuitable for large-scale studies of microbial ecology. First of all, DGGE is a time-consuming and technically challenging method, as the denaturing gradient gels are difficult to set and run properly. DGGE only serves to separate DNA fragments, and a representative band from each position on a gel must be extracted and sequenced to identify the closest match. However, only a small number of samples may be run on a single gel, and so representative bands must be sequenced from each, making analysis of large sample sets time-consuming, redundant, and expensive. Band co-migration (32, 33) and rRNA gene multi-copy heterogeneity (34) further complicate this process, respectively hindering band purity and increasing redundancy. In addition, as a PCR-based method, DGGE is subject to PCR-induced error, including amplification bias (35) and chimera formation (36), distorting community structure. Another problem with this method is that it is non-quantitative, so can detect the presence of the dominant species present but cannot reliably determine relative abundance, though others dispute this point, suggesting integration of band size and intensity against a densitometric curve (37) may be an accurate indicator of abundance.

That said, DGGE was the tool that pioneered the field of molecular ecology of foods. DGGE was first implemented to study bacterial communities in food by Ampe et al. (8), who used it to describe the special distribution of bacteria in pozol, Mexican fermented maize dough. Cocolin and coworkers (38) introduced DGGE characterization of fungi, profiling yeast succession during wine fermentations. Prokaryotic 16S rRNA genes may contain insufficient heterogeneity to reliably identify some bacterial species using the short amplicons necessary for DGGE separation, and Renouf and coworkers (39) applied DGGE of bacterial spoB gene amplicons as a means of differentiating wine LAB with high 16S rRNA gene homology. DGGE has seen widespread use in wine and food microbial community studies beyond enumeration here. For more information on the application of DGGE to study food, readers are referred to the review of Ercolini (40).

**Terminal restriction fragment length polymorphism (TRFLP)**

Terminal restriction fragment length polymorphism (TRFLP) is another extremely popular technique for studying microbial ecology, but has received much less attention by food microbiologists. This method differentiates microbial populations based on terminal restriction fragment (TRF) size. Mixed DNA samples are amplified with fluorescence-labeled universal primers, digested with select restriction enzymes, and separated by capillary electrophoresis with molecular standards for estimating fragment size. Only the labeled 5’ terminal fragments are detected and compared to a database to determine which bacterial populations (grouped as operational taxonomic units, OTUs) are represented by these fingerprint-like markers. Fluorescence intensity is integrated to pseudoquantitatively determine OTU abundance relative to total fluorescence.

Like DGGE, TRFLP is prone to the same amplification-related biases common to all PCR-based techniques. In addition, as electrophoretic mobility is a function of molecular weight, charge, and sequence structure, both sequence variability and fluorescence labels may alter the electrophoretic migration of TRFs, resulting in a reproducible disparity between expected and observed TRF sizes (41, 42). This problem may be ameliorated by...
creating an empirical TRFLP database for the microbial species expected in a sample, but this process is time-consuming and can make critical assumptions about the microbial constituents of a given system. Residual polymerase activity (43) and single-stranded “pseudo-TRFs” (44) can also create artificial OTUs, though both of these errors can be corrected by proper procedural modification. The largest issue with TRFLP, however, is that sequence homology among related microbial taxa limits the taxonomic depth achieved using this technique, so a single TRF is identified to the common taxon encompassing all possible database matches. The shallow taxonomic depth often achieved by universal primers may be increased to species-level by using clade-specific primers, for example specific for LAB (45).

TRFLP has several advantages that suit it for microbial community profiling compared to DGGE. Most importantly, TRFLP is a high-throughput, automatable process, unlike DGGE, so can compare much larger sample sets rapidly and efficiently. Thus, TRF profiles may be integrated and used for diversity comparisons between samples on a massive scale. A common approach is to use principle component analysis (PCA) or a similar approach to cluster samples based on TRF abundance and visualize these differences in 2-dimensional space (46). Similarly, TRFLP is commonly used to calculate diversity statistics such as Jaccard distance (47) (though caution is advised, as TRFLP can overestimate species diversity due to the above-mentioned errors (48)) and ANOVA or other multivariate statistics can compare significance among clusters or treatment categories (46). Additionally, TRFLP is a pseudoquantitative method that calculates relative species abundance based on sensitive capillary electrophoresis separation and fluorescence detection, whereas DGGE is not, based on polyacrylamide gel separation and UV detection. For interested readers, others have written excellent reviews detailing TRFLP methodology (42) and analysis (49) beyond the scope of this article.

TRFLP was first developed to study mixed bacterial communities by Liu and coworkers (50). Since then, it has become a popular technique for microbial profiling (49), but has been little-used for studying food systems. Rademaker and coworkers (51, 52) first applied TRFLP to study cheese and yogurt fermentation communities. Others have applied reverse-transcription TRFLP to analyze viable bacterial communities during cheese fermentations (53). Specific TRFLP assays have been developed for species-level differentiation of LAB (45) and yeasts (54) in fermented foods and beverages, and used to profile microbial successions during beer (55) and wine fermentations (56). TRFLP has also been used to analyze the bacterial composition of commercial probiotic products (57). This technique shows much promise as a low-tech, high-throughput method for studying microbial ecology, but has been largely overlooked by food microbiologists. With the advent of next-generation sequencing tools, TRFLP remains a viable complementary method for exploring species-level dynamics.

Next-generation sequencing (NGS)
Next-generation sequencing (NGS) collectively describes several technologies that achieve massively parallel sequencing of heterogeneous DNA fragments. For the purposes of microbial community surveillance, these fragments consist of short segments amplified using universal PCR primers targeting known marker genes, principally prokaryotic 16S rRNA and fungal ITS genes. Fragments are physically partitioned and base incorporation in the growing DNA strand is detected using novel technologies unique to each system. The only two NGS systems currently used for microbial community profiling (to our knowledge) are 454 Life Sciences pyrosequencing (58) and Illumina (formerly Solexa) sequencing platforms (59). These systems deliver a functionally identical product, raw DNA sequence, but with varying error rates, sequence lengths, yields, and costs. Overall, pyrosequencing can sequence longer reads (up to 600 bp) than Illumina systems (150 bp), but Illumina provides magnitudes-greater sequence coverage per run (∼109 reads on the HiSeq2000) than pyrosequencing (∼107 reads on the 454 Life Sciences Titanium FLX+), enabling greater multiplexing capacity and thus dramatically lower cost per sample (60). Longer sequencing reads theoretically provide greater taxonomic information, but reads as short as 100 bp can adequately describe microbial community changes and sequence coverage is more important for fully capturing microbial diversity (61).

Pyrosequencing, as the first commercially available NGS technique, has seen the most extensive use in food microbiology. It has been used to study the bacterial diversity of a number of fermented foods, including narezushi (fermented fish and rice) (62, 63), nukadoko (fermented rice bran) (64), Chinese liquor fermentations (65), Danish raw milk cheeses (66), Polish Oscypek cheese (66), and pearl millet fermentations (67). It has been remarkably well-utilized for studying Korean fermented foods, including bacteria and archaea of fermented seafood (68), bacterial and fungal communities of Makgeolli (rice beer) (69), meju (70) and doenjang (soybean pastes) (71), kochujang (fermented red pepper condiment) (72), and kimchi (73). It has also been used to study the viral communities inhabiting a range of Korean fermented food products (74). The greater sequence coverage and lower cost of Illumina sequencing has prompted its growing application in the study of microbial ecology in favor of pyrosequencing, and food microbiology is no exception. We have compared multiple hypervariable 16S rDNA domains for optimal taxonomic coverage of wine and food fermentations using short, 150-bp reads currently available on Illumina sequencing platforms, determining V4 as the more informative region (56). We have used this method to study botrytized wine fermentations (56) and the brewhouse-resident microbiota responsible for the fermentation of American coolship ale, a beer fermented without any inoculation (53).

The benefits of NGS over first-generation profiling methods, such as DGGE and TRFLP, are numerous. All NGS systems sequence thousands to billions of raw reads in a single run, translating to numerous observations for robust, saturated diversity
analyses. As detection occurs at the molecular level using high-resolution optics or a semiconductor chip, as opposed to the bulk UV/fluorescence excitation of DGGE and TRFLP, it is much more sensitive to low-abundance OTUs. Like TRFLP, the readout is pseudounquantitative (as read count per OTU relative to total reads), but it is sequence-based, so can be used for phylogeny-based diversity comparisons, such as UniFrac (75). Count-based diversity metrics, such as Bray-Curtis dissimilarity, calculate sample similarity as a function of shared OTUs, but do not consider phylogenetic distance, which often governs similarity in metabolic function and ecological role (75). Therefore, NGS enables comparison of communities by phylogenetic similarity and by extension functionality, in addition to count-based metrics. In food systems, this can give deeper meaning to observed microbial successions, demonstrating the impact of conditions (e.g., temperature, pH) or treatments not only on microbial diversity counts, but also on functional processes.

NGS is a growing technology and along with its many advantages come new challenges for analysis. The foremost issue is computational analysis, power, and storage. The many thousand to billion reads retrieved by NGS platforms come at the cost of many gigabytes-worth of hard drive space and enough computational power to handle these files. These requirements grow in parallel with increasing read coverage, so that the Illumina HiSeq2000 (~10 billion reads) requires magnitudes-greater space and memory than the 454 Life Sciences Titanium FLX pyrosequencer (~100,000 reads). Several open-source programs exist for processing short-amplicon NGS data, such as QIIME (76) and Mothur (77). These programs each package several features for the many steps required to translate raw reads into microbial community characterization, including high-throughput read filtration/demultiplexing, OTU selection, taxonomic assignment, sequence alignment, and downstream analysis. For a thorough discussion of the various programs available, their features, and the bioinformatic processes involved in interpreting microbial community patterns from NGS data, the reader is referred to the reviews of Kuczynski et al. (60) and Zaneveld et al. (78).

**Database quality**

A critical step for translating NGS reads into meaningful biological results is taxonomic assignment of sequences against a reference database. The resulting information is only as good as its reference database, so researchers must carefully select a database that meets the needs of their project. A database should be both comprehensive and high-quality. A comprehensive database should contain enough representative sequences to reliably describe all target taxa in thorough detail. If a given species is not represented within the database, sequences derived from that species would receive an incorrect assignment or remain unclassified. Therefore, a good database should have reasonable coverage of species of all taxa of interest.

Database quality is judged by the individual sequences contained therein. Many public sequence databases contain incorrectly classified or unclassified sequences, chimera, and sequences that are either too short or contain too many ambiguous base calls to be considered taxonomically useful (79-81). Such sequences can misconstrue NGS sequence assignment and/or limit depth of taxonomic classification, so users must carefully select databases that have been properly curated and filtered to include only high-quality, validated sequences. Fortunately, a number of options exist. For Prokaryotes, the Greengenes database (81, 82), Silva reference database (83), and Ribosomal database project (84) all provide curated, high-quality 16S rDNA sequence databases. For fungi, the situation is much less straightforward, as public sequence databases tend to contain very high levels of poor-quality and misannotated sequences (79, 80), while curated databases lack coverage or cover only specific subsections of fungi. UNITE (80, 85) offers a hand-curated ITS database of primarily ectomycorrhizal fungi (not associated with food) as well as a non-curated compilation of all ITS sequences from the International Nucleotide Sequence Database (INSD). Silva reference database (83) also contains a number of fungal large-subunit (LSU) rDNA sequences, but the list of curated sequences is far from complete. Different reference databases yield different taxonomic assignments as a function of completeness and quality (81, 86) so care must be taken to select an appropriate database for NGS studies, underlining the need for a comprehensive, carefully annotated fungal reference database (80).

**Combined tactics**

No molecular method is a stand-alone tool for describing microbial communities. Targeted methods can only detect and enumerate taxonomically defined targets, so cannot provide a comprehensive view of a community. Profiling methods, conversely, differentiate their target taxa, but cannot absolutely quantify these populations. Different molecular targets can also render dramatically different views of the same community due to primer bias (35), primer coverage (see general considerations below), and the degree of intra-taxon homology within that region (87). Even the cutting-edge profiling technology, NGS, cannot reliably differentiate most species, so cannot be considered fully comprehensive if species-level dynamics play a significant role in the studied system (as is usually the case for food fermentations).

Therefore, a combined approach is the best current strategy to study microbial ecology, utilizing multiple molecular tools to infer microbial patterns and pair these with quantifiable functional attributes impacting food fermentations. The combination of tools used will depend upon the questions asked and the system studied, but Fig. 1 presents a conceptual guideline for a combined approach to fermentation profiling. The numbered paragraphs below correspond to the steps outlined in Fig. 1.

1. Initially, a profiling tool should be used to give a broad-spectrum view of the microbial communities present. NGS would provide the most comprehensive, informative readout, but universal TRFLP or DGGE may be more appropriate given the system, sample size, and purpose. Microbial profiles may be used to calculate diversity statistics within and between samples.
Next-generation approaches to the microbial ecology of food fermentations
Nicholas A. Bokulich and David A. Mills

and cluster samples based on similarity. Optionally, a low-cost profiling tool such as TRFLP or DGGE may be used to cluster samples based on profile similarity. Then representative samples may be selected from each cluster for NGS, giving a more comprehensive, phylogenetically robust view of the microbial community (55).

2. A clade-specific profiling tool may be used to increase taxonomic depth for the most dominant and important groups. For example, LAB-TRFLP (45) may be used to profile LAB dynamics at species-level. Alternatively, a separate target region may be analyzed by the same profiling technique in step 1, confirming and/or complementing initial observations.

3. Profiling data should be complemented by QPCR or an equivalent quantitative, targeted method to quantify total bacterial/fungal load and/or track the most abundant or significant taxa detected by profiling tools.

4. If in situ visualization of microbial dispersion and/or physical interactions is an important aspect (e.g., in solid-phase fermentations), FISH may be used to observe spatial distribution of primary populations identified by profiling methods and QPCR. Representative samples may be selected by clustering based on community profiles (see item 1).

By pairing profiling and targeted tools to a specific purpose, a complete, reliable picture of whole-community dynamics may be formed, providing deeper insight into community function and process variables. In addition, further downstream steps (beyond the scope of this review) may be integrated with microbial data to form a system-wide perspective. The possibilities are virtually endless, but some examples given in Fig. 1 are:

5. Complement community-wide profiling with a strain profiling technique to observe strain-level dynamics during fermentation. This may in turn inform further targeted profiling (QPCR, FISH) using strain-specific primers/probes.

6. Complement microbial profiling with relevant physiochemical analyses (e.g., primary substrates, metabolites, volatile compounds).

7. Quantify functional gene presence/expression with QPCR to classify/compare communities by specific metabolic function. Alternatively, representative samples selected based on microbial profiles (as described in step 1) may be profiled using metatranscriptomic NGS to compare comprehensive expression profiles between different communities. Metatranscriptomics is another novel application of NGS technology beyond the scope of this discussion, and interested readers are referred to the review of Simon and Daniel (88).

8. Statistically integrate microbial, physiochemical, and func-

Fig. 1. General guideline for combined molecular approach to community profiling. General steps include: 1) broad-spectrum community profiling; 2) clade-specific profiling; 3) quantification of major taxa; 4) in situ visualization; 5) strain profiling; 6) chemical profiling; 7) functional gene quantification; and 8) integrate physiochemical and microbial profiles to statistically correlate systems-level phenomena.
tional genetic profiles to determine how sample categories (e.g., treatment variables) impact microbiota and hence composition, and/or correlate physiochemical/functional genetic parameters with specific groups/strains of microbes.

GENERAL CONSIDERATIONS FOR MOLECULAR TOOLS

Sampling
Fermented food systems are almost always heterogeneous mixtures, often combining multiple solid, liquid, and gas phases. Even in beverage fermentations, tanks are usually not mixed, and microbial and abiotic components can often stratify or associate with phase interfaces (e.g., microbial films on solid surfaces, pellicle formation at points of air contact). If testing a commercial product in situ, these strata cannot be mixed, as this may disrupt the quality of the product itself. Careful strategies must be employed to ensure that a reasonably representative sample is taken while respecting the needs of the producer. Moreover, sample homogenization may become an issue for reliable sampling of solid and mixed-phase fermentations, wherein bias may be introduced by inadequate homogenization. Subsequent storage may also alter microbial communities, so appropriate handling and prompt processing are critical (89).

DNA extraction
Sample preparation and DNA extraction present further sources of sampling bias. Inadequate DNA sampling methods can alter community representations via incomplete lysis of particularly durable cells, and brief cell homogenization and/or enzymatic degradation steps can enhance extraction efficiency (90). Additionally, a number of food components, including lipids and proteins, may inhibit PCR amplification, and must be removed during extraction (91).

Viable cells
Food fermentations represent active, changing environments, and accurate assessment of the most dominant populations at a single time may require specific detection of viable cells. Reverse-transcription methods have been used frequently for viable cell detection with a large number of profiling methods (e.g., TRFLP) (53). As RNA exhibits lower stability than DNA, this technique can successfully capture viable cell activity through reverse-transcription of RNA transcripts, using cDNA instead as an amplification template for any PCR-based profiling method. However, RNA expression depends upon the physiological status of the cell, (92) and reverse-transcription techniques may not be sufficient for profiling communities of metabolically active cells during stationary phase. Another approach is to use the DNA-binding dyes propidium monoazide (PMA) or ethidium monoazide (EMA) prior to DNA extraction. These compounds covalently bind DNA, preventing PCR amplification, but cannot permeate the membrane of intact cells, so selectively bind DNA in nonviable cells (92). PMA has been successfully used to profile bacterial populations in environmental water samples using pyrosequencing (93) and yeast communities of wine fermentations using QPCR (94).

Molecular targets
Judicious DNA target selection is critical to the success of any molecular profiling technique. An optimal region will be chosen based on the limitations of the technique used, the degree of taxonomic discrimination required, and the system studied. Furthermore, primers or probes targeting this region must possess a reasonable balance of specificity and coverage for their intended purpose, a feature that has as much to do with the design of the probe as it has for the topology of the target site. Detailed maps of the prokaryotic 16S rRNA gene and the fungal rRNA operon, the most common molecular targets for taxonomic differentiation, are shown in Fig. 2.

Technological limitations include sequence length, copy number, and other requirements unique to the method. For example, DGGE can only accommodate very short sequences to reliably destabilize and separate heterogeneous sequences, and most NGS platforms can only sequence fragments in the range of 100-600 bp, so a molecular target must be chosen wherein adequate sequence heterogeneity exists to differentiate target organisms. For QPCR, sequences should be between 100-150 bp to efficiently amplify and render accurate quantification, so primer targets must be carefully chosen to amplify all target organisms with a high degree of specificity, usually by targeting a highly variable region (e.g., for quantification of a specific genus, species, or strain) or conserved region (e.g., for total bacteria).

Molecular target copy number influences both absolutely quantitative methods (e.g., QPCR) and pseudoquantitative methods (e.g., TRFLP, NGS), as copy number, especially of rRNA operon genes, can vary widely among taxa, skewing quantitative estimates. This is particularly true of the fungal rRNA operon (including ITS), which exhibits large copy number variation among species, and is

---

**Fig. 2.** Map of prokaryotic 16S rRNA gene (A) and the fungal rRNA operon (B). Scale is in bp.
not firmly quantified in most species (95). To circumvent this is-

issue, many researchers have chosen alternative, single-copy target

sites, such as the \( \text{rpoB} \) gene in bacteria (39), but the lack of repre-
sentative sequence coverage of such genes poses a challenge for

widespread implementation (see below).

**Taxonomic discrimination**

Community profiling methods carry the additional need for a tar-

get site that contains sufficient heterogeneity to differentiate taxa

at the desired level within size constraints. For example, an opti-
nal NGS target region should contain as much heterogeneity as

possible within 100-600 bp. Not all phyla will be equally hetero-
geneous at a given site, however, so the “optimal” target region

is dependent upon the sample type and the species contained

therein (86). We have found that the V4 domain of prokaryotic

16S rDNA is a suitable region for short-amplicon NGS of food

and wine fermentations, as it can reliably discriminate most LAB
to genus-level in a 150 bp read (56). This is consistent with the

findings of others, who bioinformatically determined the V4 do-

main to be appropriate for a range of environments (61), but not

for all (86). Select short, hypervariable segments of the 16S rRNA

gene contain a surprisingly high degree of information, rendering

phylogenetic assessments comparable to those of full-length 16S

sequences (61).

Fungi (and other eukarya) present a similar situation. Most

yeasts contain high interspecific variability in the D1/D2 loop of

LSU rDNA (96), yet this site cannot reliably discriminate species

belonging to certain clades (97). The ITS region is generally hy-
pervariable across all fungi, and thus has been adopted as the

“universal barcode” for fungal identification (97), but the ITS1

and ITS2 subdomains display varying levels of intraspecific and

intragenic heterogeneity within different clades (87), and these

regions can deliver dramatically different views of community di-

versity using short-amplicon NGS reads (98).

**Coverage and specificity**

Coverage describes how well a primer or probe matches all

members of a given taxon along the target binding site. For ex-

ample, a universal bacterial primer should ideally be homolo-
gous to all bacteria at a single, designated site, otherwise it will
give an inaccurate estimate of bacterial quantity (QPCR) or com-
munity structure (profiling methods). A perfect scenario is highly
unlikely, as 1) sequence heterogeneity along most universal
genes dictates that at least some bacteria will have mismatches,
increasing the potential for nonbinding or preferential amplifica-
tion of targets displaying full homology and 2) public collections
of sequence data on all microbial genes are incomplete and do
not comprehend even all currently described microbial species,
increasing the probability that current “universal” primers (based
on sequence alignments of currently described species) do not
capture this potential diversity. This is particularly pronounced
for fungi, with ITS sequences in the INSD representing <1% of
known fungal species (98). For prokaryotes, a number of degen-
erate, high-coverage primers targeting hypervariable domains

across the 16S rRNA gene have been developed and bio-

informatically validated (86). For fungi, primer choice is less

straightforward, as different established “universal” primers ex-
thibit bias toward specific clades (99). Most fungal primers target-
ing the rRNA operon were designed 10-20 years ago based on
contemporary sequence data, and a need exists for a modern re-

vision of these primers based on current data.

Specificity describes the degree to which a primer or probe ex-
clusively binds members of the target taxon. For example, a uni-
versal bacterial primer should not also bind archaeal or eukary-
otic sequences and a primer for *Lactobacillus helveticus* should
not also bind *Lactobacillus casei*. This is particularly important
for QPCR, as mispriming will result in overestimation of the tar-
gent, but also impacts profiling tools, as it can limit detection of
target taxa. For example, many prokaryotic (100) and fungal
(101) rDNA primers contain high homology to chloroplast and
mitochondrial sequences. As almost all foods are plant- or ani-
mal-based, the majority of signal (e.g., DGGE, TRFLP, FISH) or
sequence reads (NGS) may consist of sequences from the source
material, dampening detection of the microbiota present, even if
it can be unambiguously distinguished from microbial OTUs.

Primers that do not bind plant/plastid DNA have been developed
for both fungi (101) and bacteria (100), but these primers have
not been validated against a comprehensive database and their
increased specificity may risk compromising the universal cover-
age of these kingdoms. Thus, all primers used for molecular tools
should be computationally validated to ensure adequate cover-
age of and specificity for target taxa. This must be balanced
against the threat of eukaryotic DNA contamination on a sam-
ples-dependent basis when choosing the right target site and
probe.

**Selection**

All factors considered, it is difficult to distinguish any one site as
the optimal target for most molecular methods. For studying pro-
karyotic communities, the rRNA operon is historically the most
common target, particularly the small-subunit 16S rRNA gene, as
it contains highly conserved regions (ideal for universal primer
sites) interpolated by highly variable regions (ideal for taxonomic
discrimination). This gene remains the most prevalent target, as it
is also the best-represented prokaryotic gene contained in all
public sequence databases. For over 20 years this gene has pro-
ven its worth for phylogenetic comparison and identification of
prokaryotes, but some properties of this gene make it less favor-
able for specific applications. Copy number variation can skew
accurate quantification, making single-copy genes more appro-
priate for QPCR of some targets. In addition, some closely re-
lated species cannot be discriminated by short 16S rDNA ampli-
cons, including some LAB (39). Thus, some protein-encoding
genes have gained favor, such as *rpoB* (39). Such targets are
poorly characterized across all bacterial phyla, however, so their
degree of conservation is not well established. This increases the
risk that a chosen primer/probe sequence will lack suitable spe-
ficity for or coverage of their intended target taxon. As
whole-genome and environmental metagenomic sequencing data continue to increase, the distribution and conservation of alternative gene sequences will become more apparent, possibly leading to the identification of a superior molecular target for microbial diversity studies.

For fungi, the situation is more complicated, and attention has fluctuated between the small-subunit 18S rRNA gene (following the lead of prokaryotic systematics) to the large subunit rRNA gene and to the ITS (102). The ITS has finally been chosen as the “universal barcode” for fungi, as it exhibits the best discrimination of most fungal species (97), but is less heterogeneous among certain fungal clades (87) and, as it is a non-coding, hypervariable region, lacks the phylogenetically metric nature of conserved rRNA genes for comparing highly divergent clades (103).

CONCLUSION

Integrating multiple molecular approaches, including high-throughput community profiling, targeted profiling, and in situ visualization, may provide deeper insight into food fermentations at a systems level. Statistically supported trending of microbial communities and expression profiles with fermentation outcomes (e.g., evolution of desirable sensory characteristics or of spoilage) may help identify marker genes or microbial taxa associated with specific activities. This may lead to improved strain selection for microbial inoculation and enhanced management strategies for fermentation control. Microbial ecology of these foods may no longer consider the fermentation itself, but also how the surrounding environment impacts this process. This may be viewed at both the local level—the microbial ecology of surfaces and processing equipment within the complete production facility and their relationship to the product—as well as the macro-scale—how region and climatic conditions impact the microbial succession of fermentations or raw ingredients between geographically disparate areas. Microbial community profiles generated by NGS and other emerging technologies may even be used to determine the source of foods or raw materials, including validating food products with protected-origin status (e.g., by European law). Incorporating geospatial community profiling into the systems-level framework facilitates the process of strain selection by increasing the rate of bioprospecting and marker identification, and improves opportunities for process improvement through better understanding of the relationships between substrate conditions, process parameters, microbial activity, and fermentation outcomes.

Modern molecular tools provide the opportunity for high-resolution analysis of fermentation systems, supported by the growing quality and availability of public sequence data and bioinformatic tools. NGS, in particular, is revolutionizing the way we approach microbial ecology, promising groundbreaking discoveries in the field of food microbiology. Many challenges remain, however, and researchers must carefully select the techniques, molecular targets, and analysis tools most appropriate to their experimental purpose.

REFERENCES

Next-generation approaches to the microbial ecology of food fermentations
Nicholas A. Bokulich and David A. Mills

Chem. 59, 117-121.


http://bmbreports.org
Next-generation approaches to the microbial ecology of food fermentations
Nicholas A. Bokulich and David A. Mills


