Running Title

High-throughput analysis of microbial communities

Title

Methods for high-throughput comparative analyses of natural microbial communities

Sarah P. Preheim[1], Allison R. Perrotta[2], Jonathan Friedman[3,4], Chris Smilie[4], Ilana Brito[1], Eric Alm[1]

[1] Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

[2] Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA

[3] Physics, Massachusetts Institute of Technology, Cambridge, MA

[4] Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA

Abstract

One of the most commonly applied metagenomics approaches is the amplification and sequencing of the highly conserved ribosomal RNA (rRNA) genes from organisms in a complex microbial community. Ribosomal RNA surveys, typically using the 16S rRNA gene for prokaryotic identification, provide information about the total diversity and taxonomic affiliation of organisms present in a sample. This

chapter covers the creation of 16S rRNA metagenomic surveys and comparative analysis of microbial communities using high-throughput techniques. It is meant to be an outline for those not familiar with 16S rRNA surveys using Illumina technology, and highlights some important considerations in study design, implementation and analysis. We begin by outlining the best practices for minimizing artifacts and errors during library construction. We also explain how to transform raw data into meaningful units for further analysis, including a novel method of distribution-based clustering. This paper will address issues specific to Illumina short read and paired-end data.

Introduction

Prokaryotic cells make up the majority of biomass on the planet (Whitman, Coleman, & Wiebe, 1998), and influence every ecosystem from the world's oceans to the human gut. Metagenomics approaches, including amplicon-based techniques targeting the conserved small subunit ribosomal RNA genes (commonly 16S rRNA in prokaryotes) facilitate research of the structure, function and stability of microbial communities (Amann, Ludwig, & Schleifer, 1995; Woese, Kandler, & Wheelis, 1990). 16S rRNA surveys of prokaryotic diversity data have greatly expanded our understanding of the microbial world, such as identifying that a majority of organisms in the environment are unculturable by standard culturing techniques (Rappe & Giovannoni, 2003) and that a large majority of the genetic diversity in a sample comes from rare organisms that are found at low abundance (Huse, Welch, Morrison, & Sogin, 2010). Additionally, 16S rRNA surveys of human-

associated microbial communities have shed light on the importance of the community in human health and diseases such as obesity (Ley et al., 2005) and malnutrition (Smith et al., 2013).

Although 16S rRNA surveys of microbial communities are widely used to characterize the composition and diversity of microorganisms present in a sample, there are many problems associated with transforming 16S rRNA sequences into proxy for species, as well as other issues that arise during the sequencing process. The preparation of samples incorporates many known biases and sources of error that may result in inaccurate and skewed compositional results. (Forney, Zhou, & Brown, 2004). DNA extraction techniques preferentially lyse certain cell types over others, resulting in an uneven representation of 16S rRNA genes in the community DNA pool (Frostegard et al., 1999). Polymerase chain reaction (PCR) amplification is known to have inherent biases regarding the percent GC of the template, amplicon length and mismatches in the primer-binding site (Polz & Cavanaugh, 1998). Additionally, single base changes and chimeras can create artificial diversity during both PCR and sequencing (Qiu et al., 2001). Steps should be taken to ameliorate these errors and biases whenever possible.

In spite of these inherent problems, 16S rRNA surveys show great promise as new techniques allow for amplicon-based studies to keep pace with DNA sequencing technology. With the ability to compare hundreds of community profiles in one sequencing run, researchers can now strive to complete comprehensive body site sampling in healthy adults with as part of the Human Microbiome project (Huttenhower et al., 2012), to sequence hundreds of thousands of samples from

across the world in the Earth Microbiome (http://www.earthmicrobiome.org/) and take on countless other projects characterizing microbial community variation in space or time.

Chapter Goals

This chapter will outline both molecular methods and bioinformatics approaches used to identify and compare microbial communities using high-throughput sequencing technologies, specifically the Illumina platforms (San Francisco, CA). There are many websites that can be used as a guide regarding the specifics of Roche-454 pyrosequencing and analysis. Illumina sequencing differs from Roche-454 pyrosequencing in the total number of reads, the total read length, error rate and paired-end sequencing approach. We will address some issues that are particularly important for Illumina sequencing, although many of the principles are similar regardless of sequencing platform. We focus on the following topics in the comparative analyses of microbial communities:

- Library Construction
- Quality control/OTU assignment/classification

Library construction

Identifying microorganisms present in a natural community using a sequencing-based 16S rRNA survey begins with the construction of a library. A library is a collection of DNA fragments that represents the sequence diversity in a sample. These fragments are enriched from the rest of the community genomic DNA

by PCR using primers which match to the microbial population or gene of interest. However, these sequences must be manipulated in a platform-specific way for sequencing. The complete molecular construct contains the genomic DNA sequences from the enrichment reaction, sequences that identify the sample it originated from (i.e. index or barcode sequences) and sequences required by the platform to adhere library fragments to the solid matrix and provide a priming site for the sequencing reaction (Fig. 1). Adding a barcode sequence to the molecular construct identifying which sample the library originated from allows for hundreds of libraries to be sequenced in the same reaction, commonly called multiplexing.

[Insert Figure 1]

There can be multiple different designs for these molecular constructs, depending on the researchers needs and priorities. One approach maximizes useable read lengths by using the same primer for both amplification and sequencing (Knight et al., 2011). We have developed an alternative approach which maximizes the flexibility of the library construct by separating the amplification of genomic DNA of interest from the addition of Illumina-specific adapters and indexing sequences (Blackburn, 2010). With this two-step PCR approach, the same indexing oligos can be used to sequence multiple different amplicons (e.g. different areas of the 16S rRNA gene or other functional genes). For example, a 96-well plate of synthesized oligos, each containing a unique index sequence can be used to sequence both bacterial 16S and eukaryotic 18S sequences. Additionally, the first few bases that are read by the sequencer can be manipulated, which provide additional indexing capacity since this region can be used as a second barcode.

Alternatively, this region may be used to improve sequencing quality for amplicon-only runs (for more information, reference the section on *Multiplexing and sequencing*).

With a two-step approach, the complete molecular construct is created in a step-wise manner (Fig. 1). Primers targeting the community genomic DNA for the enrichment step are synthesized with a binding site for the primers used during library amplification. After the first PCR reaction (Step 1), the region of interest has been enriched and the amplicon has incorporated a sequence on either end that matches part of the Illumina specific sequences. This overlap is the priming site for a second PCR reaction (Step 2) that incorporates the Illumina adapters and the indexing or barcode sequences into the molecular construct. For dual indexing capacity, an additional barcode sequence can be added between the template specific primers and the sequences overlapping the adapters.

There are steps necessary to build the molecular construct from the two-step approach described above. These are:

- Normalization
- Amplification from community genomic DNA
- Addition of Adapter and Indexing Sequences
- Purification and multiplexing

Normalizing samples before PCR cycling

Proper PCR amplification is an important part of obtaining results that minimize methodological errors. Over-cycling can promote chimeric molecules by

promoting the extension of partial amplification products when primer concentrations drop at the end of the cycle (Qiu, et al., 2001). Over cycling also tends to normalize amplicon concentrations for different templates altering their relative abundances (Polz & Cavanaugh, 1998). Therefore, we use an initial real-time PCR to visualize the reaction curves and ensure that each sample is not cycled past the exponential or log phase of the reaction. Not only can this improve data quality, but it can also be used to normalize input DNA concentrations when they might be different, ensuring that different samples are cycled similarly.

Real-time PCR is a reaction similar to subsequent PCR reactions in library construction, but visualized using an optical monitor during thermal cycling. Real-time PCR was carried out using with a CFX 96 Real-Time System (BioRad, Hercules, CA) with the following conditions: 0.5 units of Phusion (New England Biolabs, Ipswich, MA) with 1 x High Fidelity buffer, 200 μM of each dNTP, 0.3 μM of each forward and reverse primer primers and approximately 40 ng of mixed DNA template were added for each 25 μL reaction. We chose to use the V4 region of the 16S rRNA gene because these primers provide broad taxonomic coverage (Wang & Qian, 2009) and good taxonomic assignments (Soergel, Dey, Knight, & Brenner, 2012). The sequence we used for the first step amplification targets the position 515 (5' -ACACG ACGCT CTTCC GATCT YRYRG TGCCA GCMGC CGCGG TAA - 3') and 786 (5'- CGGCA TTCCT GCTGA ACCGC TCTTC CGATC TGGAC TACHV GGGTW TCTAA T-3') of the *E. coli* 16S rRNA gene, respectively (Knight, et al., 2011). Additionally, 5 X SYBR Green I nucleic acid stain (Molecular Probes, Eugene, OR) was added as a reporter for double stranded DNA abundance. Samples were cycled with the

following conditions: denaturation at 98 °C for 30 sec annealing at 52 °C for 30 sec and extension at 72 °C for 30 sec. This is carried out for 40 cycles and the threshold value was set manually to a point just above the background (Fig. 2). Using the cycle number at which the curve for each sample crossed this threshold ($C_{t,sample}$), samples are normalized to the most dilute sample (highest $C_{t,sample}$ value), with a C_t less than or equal to 20 cycles ($C_{t,lowest}$). Normalization is done by estimating the concentration of each sample relative to the most dilute sample. Templates are diluted according to the following equation:

$$1.75^{-(C_{t,sample}-C_{t,lowest})}$$

Once the samples are diluted accordingly, the first step PCR will use $C_{t,lowest}$ cycles for all diluted samples.

Amplification of the 16S rRNA gene from community DNA

There are two aspects of the first step reaction that can minimize artifacts; cycling conditions and running multiple reactions. The diluted template should be amplified under conditions similar to the real-time PCR and be limited to the cycle number identified as $C_{t,lowest}$. Although SYBR is not added, the rest of the reagents and cycling temperatures and times should be identical to the conditions in the real-time PCR reaction. Previous studies have demonstrated that the maximum amount of diversity is recovered when samples are split into multiple reactions and cycled separately (Lahr & Katz, 2009), since skewed representation of sequences can results from jackpot effects that arise in single PCR reactions. For example, cycling each library in four- 25µL replicates will improve the recovery of the total diversity

of the original sample. These separate reaction replicates (often done across different PCR plates), should be pooled before beginning purification step.

Addition of sequencer (platform-specific) specific adapters and indexing sequences

Platform specific adapters and indexing sequences are added to the amplicon through a second step PCR reaction (Fig. 1). The first step reactions are purified with Agencourt AMPure XP- PCR purification system (Beckman Coulter, Brea, CA), which can be used in small batches, or in a 96-well format, using a 96-well magnetic plate, following the manufacture's protocol. The primers in the first step reaction contain an overlapping sequence, which provides the primer-binding site for incorporating the full platform-specific sequences and a sample specific barcode sequence during the second step reaction. The conditions for the second step PCR are similar to the first step, although 4 μl of the purified first step reaction was used as a template and 0.4 μM of each PE-III-PCR-F (5'- AATGA TACGG CGACC ACCGA GATCT ACACT CTTTC CCTAC ACGAC GCTCT TCCGA TCT- 3') and the barcoded reverse primer (5' - CAAGC AGAAG ACGGC ATACG AGAT**N NNNNN NNN**CG GTCTC GGCAT TCCTG CTGAA CCGCT CTTCC GATCT -3' where N's are represent the indexing sequencing specific for each sample) was used with 9 cycles. The concentration of other reagents was the same. Samples were cycled with the following conditions: denaturation at 98 °C for 30 sec annealing at 83 °C for 30 sec and extension at 72 °C for 30 sec. Samples are again cycled in 4 x 25 μL reactions for 7-9 cycles, which is sufficient to allow the adapters and indexing sequences to

become incorporated in to the final construct. All reaction products are cleaned with the Amp-Pure magnetic beads in a manner similar to the first step reaction.

Multiplexing and sequencing

A final real-time PCR step is an accurate and high-throughput method for multiplexing samples together for sequencing. The relative concentration is determined for all samples in a set (usually 96 samples at a time) using the C_t values as described above. Samples are then diluted to the same relative concentration and pooled with equal representation. Primers used for this round of real-time PCR should be those that anneal to the Illumina adapters which mimic binding to the solid matrix. This informs of the concentration of product that will actually be sequenced in each sample and confirms that adapters have been properly added. Pooled samples are finally checked using an Agilent Bioanalyzer, which provides a quantitative histogram of the size of library fragments. Illumina sequencing requires final fragments of 200-650 bp. Often, PCR reactions are not 100% efficient, and adjustments can be made to more accurately pool samples by assuming samples increase by 1.85-fold (for example) for each cycle. If all of the samples are of a similar origin (i.e. all obtained from human stool) they can often be pooled together without a final real-time PCR, assuming uniform concentrations across all samples.

Brito 4/15/13 11:20 PM
Comment [1]: verify

Sequencing should be done with the addition of some amount of complexity in the sample when sequencing with Illumina platforms. Whether using MiSeq or HiSeq or older technology (e.g. GAII), additional diversity improves the quality of the resulting sequence. There are two options for added diversity: to spike in a small

amount of a non-amplicon sample (e.g. 20-50% phiX, standard used for QC), or generate diversity with complexity regions of different lengths. We are considering incorporating staggered primers to improve the quality from amplicon only runs by varying the length of the complexity region, although it has not been tested. Along with standard paired-end sequencing reads, a specific barcode (indexing) read can be performed for the appropriate number of barcode bases (i.e. 8 bases in our example). This is done before preparing for the reverse read. For further details please reference Illumina's Index Read protocols.

Control samples are useful for optimizing program parameters and trouble-shooting. For large projects, especially those spanning multiple sequencing lanes, control samples ensure reproducibility across lanes. A control sample may be a completely defined, mixed community (i.e. mock community) of either bacteria or DNA templates or a sample that is re-sequenced in each of the sequencing lanes (i.e. re-sequenced control), or both. The benefit of a mock community is that the diversity and sequences are known prior to sequencing, which can help with troubleshooting and optimizing program parameters during processing. The re-sequenced control can help to assess reproducibility across sequencing runs.

References

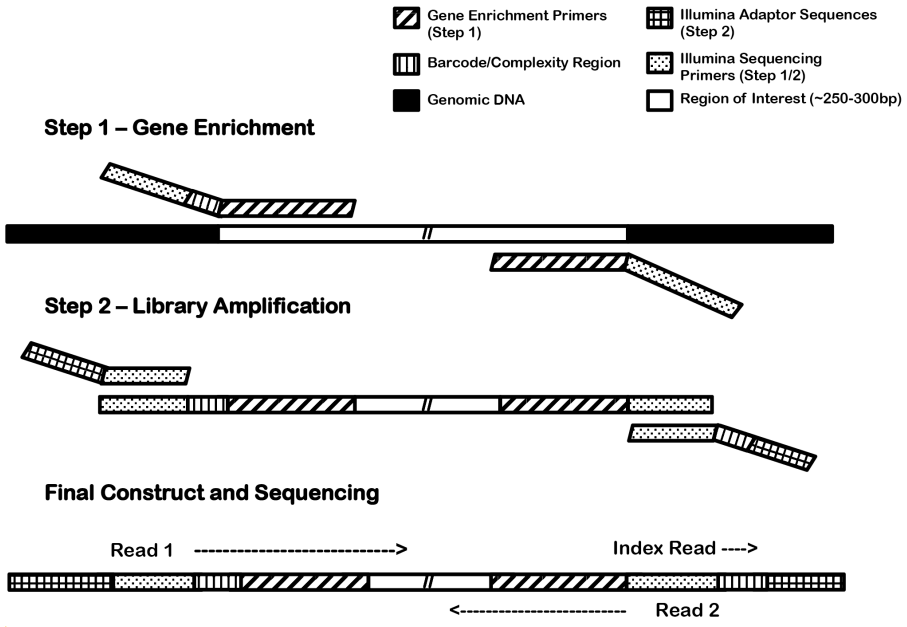
Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic Identification and in-Situ Detection of Individual Microbial-Cells without Cultivation. *Microbiological Reviews, 59*(1), 143-169.

Blackburn, M. C. (2010). *Development of New Tools and Applications for High-Throughput Sequencing of Microbiomes in Environmental or Clinical Samples.* Massachusetts Institute of Technology, Cambridge, MA.
Forney, L. J., Zhou, X., & Brown, C. J. (2004). Molecular microbial ecology: land of the one-eyed king. [Review]. *Current Opinion in Microbiology, 7*(3), 210-220.
Frostegard, A., Courtois, S., Ramisse, V., Clerc, S., Bernillon, D., Le Gall, F., et al. (1999). Quantification of bias related to the extraction of DNA directly from soils. [Article]. *Applied and Environmental Microbiology, 65*(12), 5409-5420.
Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology, 12*(7), 1889-1898.
Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207-214.
Knight, R., Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 4516-4522.
Lahr, D. J., & Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't
Research Support, U.S. Gov't, Non-P.H.S.]. *BioTechniques, 47*(4), 857-866.
Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America, 102*(31), 11070-11075.
Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. [Comparative Study
Research Support, U.S. Gov't, Non-P.H.S.]. *Applied and Environmental Microbiology, 64*(10), 3724-3730.
Qiu, X. Y., Wu, L. Y., Huang, H. S., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., et al. (2001). Evaluation of PCR-generated chimeras: Mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology, 67*(2), 880-887.
Rappe, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology, 57*, 369-394.
Smith, M. I., Yatsunenko, T., Manary, M. J., Trehan, I., Mkakosya, R., Cheng, J. Y., et al. (2013). Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor. *Science, 339*(6119), 548-554.
Soergel, D. A. W., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *Isme Journal, 6*(7), 1440-1444.
Wang, Y., & Qian, P. Y. (2009). Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. [Article]. *Plos One, 4*(10).

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America, 95*(12), 6578-6583.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a Natural System of Organisms - Proposal for the Domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America, 87*(12), 4576-4579.

Figure Legends



Unknown
Formatted: Font:14 pt, Bold

Figure 1. Overview of two-step 16S rRNA library construct. The gene enrichment step (Step 1) involves amplification of genomic DNA of interest (i.e. 16S rRNA gene) using region-specific primers. Library amplification (Step 2) involves the addition of the barcode or indexing sequence, which should be unique for each library. The final construct includes all of the adapter and cluster binding sites required by Illumina

sequencing (hatched), along with sequencing primer binding sites (dotted) and the region of genomic DNA that was enriched (white).