454 LIFE SCIENCES

Roche Diagnostics

Metagenomics analysis using the Genome Sequencer™ FLX system

The Genome Sequencer™ FLX System from 454 Life Sciences™ and Roche Applied Science is a versatile sequencing platform suitable for a wide range of applications, including *de novo* sequencing and assembly of genomic DNA, transcriptome sequencing, small RNA analysis, and amplicon sequencing. The Genome Sequencer FLX is built upon 454 Sequencing™ technology that allows 200–300 base-pair read lengths and very high single-read accuracy. One application in which the technology is accelerating the field's understanding is metagenomics.

Metagenomics is the study of genomic content in a complex mixture of microorganisms (see **Table 1** for other definitions). The two primary goals of this approach are to develop a consensus of what populations of microorganisms are present (a horizontal screen) and then to identify what roles each microorganism has within a specific environment (a vertical characterization). Metagenomics samples are found nearly everywhere, including several microenvironments within the human body, soil samples, extreme environments such as deep mines and the various layers within the ocean. Therefore, the diversity of microorganisms is thought to be in the range of tens of millions to greater than hundreds of millions of species.

Recent publications based on Genome Sequencer data have shown the vast diversity of the microbial world. In one example, a survey of marine viral metagenomes from 68 sites in four regions indicated that global viral diversity is possibly a few hundred thousand viral species[1]. The vast majority of viruses identified in the survey are widespread, but the composition of the viral communities varies from region to region. In a second example, two deep mine samples were found to comprise very different communities having different metabolisms, even though the samples were collected in very close physical proximity to one another[2]. Notably, the microbes studied were completely different from other previously sequenced microbial communities. A third publication, addressing microbial diversity within the ocean, described a screening approach using sequence tags for ribosomal RNA[3]. In this study, it was estimated that the microbial diversity was up to three orders of magnitude greater than the previous estimate of 10^6 species.

Timothy Harkins[1] & Thomas Jarvie[2]

[1]Roche Diagnostics, Roche Applied Science, 9115 Hague Road, Indianapolis, Indiana 46250, USA. [2]454 Life Sciences, 20 Commercial Street, Branford, Connecticut 06405, USA. Correspondence should be addressed to T.H. (tim.harkins@roche.com).

Table 1 | Glossary of metagenomics terms

Metagenomics	The sequencing and analysis of DNA from environmental samples without the need for culturing individual, clonal organisms.
Environmental genomics	A synonym for metagenomics.
Random community genomics	Frequently used as a synonym for metagenomics. This term has been defined by some researchers as the sequencing of whole communities from the environment with or without cloning, but without screening for any functional component.
Microbial diversity	A type of metagenomics study using a sequencing approach that focuses on a specific hypervariable region of the genome (that is, the V6 region of 16S rRNA) to assess the number of different species within an environment.
Virome	The totality of viruses and their genetic molecules within an environment.
Microbiome	The total collection of microbes and their genetic material within an environment.

Although the roles of most microorganisms have yet to be discovered, several recent publications have shown that microorganisms have both competing and synergistic interactions with each other, and these interactions can change as their local environments change. For example, a study of the human gut found that there are two principal populations of bacteria, the Bacteroidetes and the Firmicutes, and their relative abundance changes as the body fat of the individual changes[4]. As body fat increases, the abundance of Firmicutes increases, and this in turn increases the capacity for energy harvest, leading to higher obesity rates.

Advantages of Genome Sequencer technology

The Genome Sequencer FLX System (**Fig. 1**) offers several advantages for metagenomics studies. The first breakthrough is that it eliminates the

APPLICATION NOTES

Figure 1 | The Genome Sequencer FLX instrument. The Genome Sequencer FLX supports a number of formats, allowing users to customize the number of samples per instrument run and the number of reads per sample. A single run can be physically divided into 2, 4 or 16 samples with 210,000 reads per sample, 70,000 reads per sample and 12,000 reads per sample.

requirement to clone DNA fragments into bacteria. Consequently, the Genome Sequencer FLX avoids the cloning bias that is introduced in sample preparation observed in Sanger sequencing–based methods.

With over 400,000 sequencing reads per instrument run, this system can facilitate extensive surveys to identify large numbers of different genes, metabolic pathways and microbial species that may be present, while providing a dramatic reduction in the cost per project. This allows researchers to approach samples and address questions that, until recently, only major genome centers could manage. As a result, the Genome Sequencer FLX System offers a powerful technology that supports research studies to answer environmental and ecological questions.

Sequencing read length is also an important factor for metagenomics studies. As most genomes in metagenomics samples are unknown, and relatively little reference sequence information is available, it is important to have read lengths that allow researchers to both assemble genomes in a *de novo* fashion and uniquely assign reads to a specific gene and/or genome. Other next-generation sequencing technologies use sequencing read lengths, termed 'microreads', that are in the range of 15–40 bp in length. Microreads are limiting in metagenomics because of the homology and the repetitive regions within one genome and across the various genomes within the sample. The ability to uniquely assign a read to one genome becomes more challenging when there are numerous genomes present within a sample. Additionally, microreads prevent the assembly of many of the reads. Previous reports have estimated that sequencing read lengths of 100 bases are near the minimum from a utility perspective[2,5]. The Genome Sequencer FLX generates sequencing read lengths of 200–300 bases, with read length being dependent upon the specific sequence characteristics.

Preparing metagenomics samples for sequencing

The following protocol for the preparation of metagenomics samples is provided courtesy of Matthew Haynes and Forest Rohwer (San Diego State University). The Genome Sequencer FLX requires approximately 5-µg samples of relatively pure DNA for metagenomics studies. These samples may be eukaryotic, prokaryotic or viral genomic DNA, plasmid DNA or cDNA. Five micrograms of double-stranded DNA with an absorbance ratio (A_{260}/A_{280}) of ~1.8, at a concentration of at least 300 ng/µl, is desirable. The ratio A_{260}/A_{230} is an indicator of nucleic acid purity and should ideally be in the range of 1.8–2.2. Values less than 1.0, often obtained from environmental samples, indicate the presence of contaminants that may interfere with enzymatic procedures. Plasmid DNA can be amplified with commercially available polymerases to produce a sufficient quantity for sequencing using the Genome Sequencer FLX.

More difficulty may be encountered in the preparation of sufficient amounts of genomic DNA from very small samples or in cases where the efficiency of DNA isolation is limited. Microbial and bacteriophage DNA have been isolated for sequencing using the Genome Sequencer FLX from a large number of environmental samples: water, sediment, soil, feces, blood, stromatolites and others[2]. Commercially available kits can be used to isolate microbial DNA from soil, stromatolites and other solid substrates; if bacteria can be pelleted, tissue DNA purification kits are effective. Isolation of bacteriophage and some eukaryotic viruses can be accomplished using cesium chloride step-gradient centrifugation, as described in a published method[6]. DNA is then extracted from viral particles with formamide and cetyltrimethylammonium bromide (CTAB) as described in the literature[7].

If 5 µg of DNA cannot be prepared directly from samples, the DNA can be amplified. For example, one commercially available method uses the ϕ29 DNA polymerase, which has a helicase activity that allows it to amplify genomic DNA using random primers and a single denaturation step. With this method, one 20-µl reaction (containing 10–100 ng of template DNA) will produce approximately 4 µg of amplified DNA. DNA produced in this process may be underrepresented in the 500–1,000 bp near termini. After amplification, the DNA can be purified using silica columns. The eluted DNA (~400 µl) is then ethanol precipitated to concentrate: add 40 µl of 3 M sodium acetate, pH 5.2, 2 µl of glycogen (10 mg/ml) and 880 µl 100% ethanol; place at –20 °C for 2–16 h, centrifuge in microcentrifuge for 10 min, remove supernatant, wash briefly with 70% ethanol and air dry before resuspending in H_2O. DNA is then ready for sequencing with the Genome Sequencer FLX.

Bioinformatics for metagenomics

The Genome Sequencer FLX sequences over 100 million bases per instrument run. For metagenomics studies this presents a very large data set with several challenges that need to be addressed. To help

http://www-ab.informatik.uni-tuebingen.de/software/megan
MEGAN, a Metagenome Analyzer, allows a single scientist to analyze large data sets and group sequencing reads into taxonomic units.

http://seed.sdsu.edu/FIG/index.cgi
The SEED database is a public resource of complete and draft genome sequences to help relate sequences to metabolic function.

http://img.jgi.doe.gov/cgi-bin/m/main.cgi
IMG/M provides tools for analyzing the functional capability of microbial communities based on their metagenome sequence, in the context of reference isolate genomes, using a variety of public functional and pathway resources.

http://camera.calit2.net/
CAMERA is a user-driven site dedicated to providing the scientific community with metagenomics data and analysis tools.

Figure 2 | Useful URLs for metagenomics analysis.

understand the available bioinformatics resources, we have listed several publicly available websites (**Fig. 2**). Typically, the first objective of the analysis is to identify which sequencing reads can be associated with sequenced genomes versus unsequenced organisms. Using an application such as MEGAN, researchers can group their data based on taxonomic level to summarize and order their results. With this approach, researchers can assess the complexity of their samples as to the diversity of microorganisms that are present. Additionally, a 16S rDNA analysis can be performed, which will identify a low number of sequencing reads that can help to determine which genera or species are present within the sample.

The next objective for many metagenomics studies is to identify the metabolic functions of a microorganism within the sample. With average Genome Sequencer FLX read lengths of 250 bases, it is possible to search against sequence databases for homologs, but the hit rate will most likely be low, in the 5–10% range (owing to the incomplete nature of the microbial sequence databases). Nevertheless, this will still provide up to 20,000–40,000 sequences to examine for known functionality. By using BLASTX, a computationally intensive application, a functional analysis can be performed to help identify the metabolic function of the organisms within the metagenomic sample.

The ultimate goal of metagenomics is to generate sequence assemblies resulting in, at a minimum, full-length genes and, preferably, complete genomes. The 250-base-pair read lengths generated by the Genome Sequencer FLX allow *de novo* assemblies of metagenomes. Generating assemblies of the sequencing reads is critical for characterizing full-length genes, discovering new genes and ultimately helping model microbial communities based upon sequence similarities.

Summary

Until recently, metagenomics analysis was limited by the cost, low throughput and inherent cloning bias of the Sanger technologies. The Genome Sequencer FLX System provides a comprehensive view of metagenomics samples with high throughput, no cloning bias, and read lengths long enough to allow diversity and functional analysis of microbial communities.

Additional information about the Genome Sequencer FLX System is available from Roche Applied Science (http://www.genome-sequencing.com). Genome Sequencer, 454, 454 Sequencing and 454 Life Sciences are trademarks of 454 Life Sciences Corporation, Branford, Connecticut, USA. License disclaimer information is available online (http://www.genome-sequencing.com).

1. Angly, F.E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
2. Edwards, R.A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7**, 57 (2006).
3. Sogin, M.L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).
4. Turnbaugh, P.J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
5. Goldberg, S.M. *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* **103**, 11240–11245 (2006).
6. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* 2.74 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989).
7. Ausubel, F.M. *et al. Short Protocols in Molecular Biology* (5th edn.) Vol. 1, 2–11 (Wiley, New York, 2002).