Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Extracting data from the muck: deriving biological insight from complex microbial communities and non-model organisms with next generation sequencing

Kevin V Solomon[1], Charles H Haitjema[1], Dawn A Thompson[2] and Michelle A O'Malley[1]

It is becoming increasingly clear that microbes within microbial communities, for which cultured isolates have not yet been obtained, have an immense, untapped reservoir of enzymes that could help address grand challenges in human health, energy, and sustainability. Despite the obstacles associated with culturing these microbes, recent advances in next-generation sequencing (NGS) have made it possible to explore complex microbial communities in their native context for the first time. Key to extracting meaning from rapidly growing NGS datasets are bioinformatics tools that assemble the sequence data, annotate homologous sequences and interrogate it to reveal regulatory patterns. Complementing this are advances in proteomics that can link NGS data to biological function. This combination of next generation sequencing, proteomics and bioinformatic analysis forms a powerful tool to study non-model microbes, which will transform what we know about these dynamic systems.

Addresses
[1] Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA
[2] Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA

Corresponding author: O'Malley, Michelle A (momalley@engineering.ucsb.edu)

Current Opinion in Biotechnology 2014, **28**:103–110

This review comes from a themed issue on **Systems biology**

Edited by **Christian M Metallo** and **Victor Sourjik**

For a complete overview see the Issue and the Editorial

Available online 4th February 2014

0958-1669/$ – see front matter,

http://dx.doi.org/10.1016/j.copbio.2014.01.007

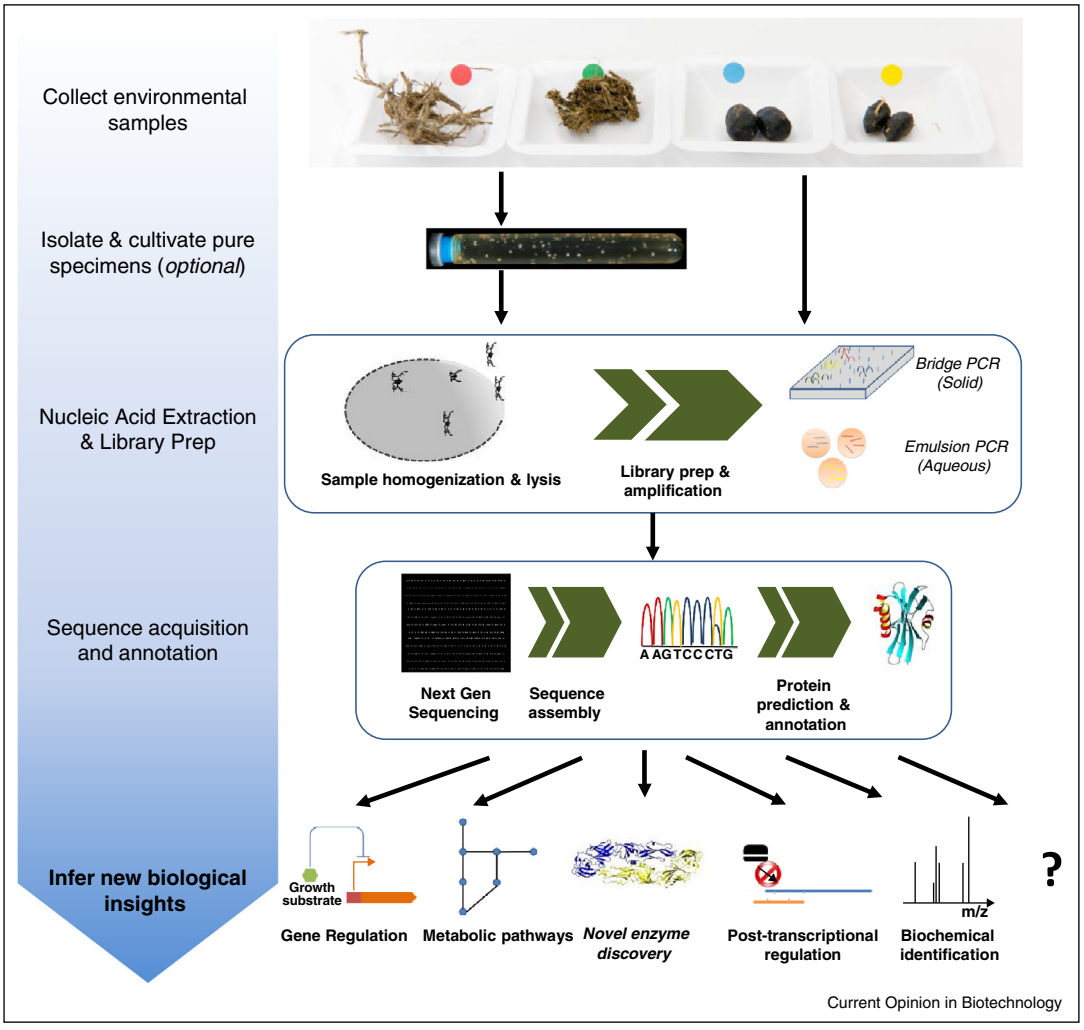
Introduction

The study of model microbes has set the foundation for what we know about molecular, cellular, and systems biology, which has revolutionized our ability to adapt and engineer microbes for bioprocessing; yet, by comparison, we know very little about non-model microbes, though they constitute 99% of the Earth's biosphere [1]. These microbes remain uncharacterized as they are fastidious, and often exist in complex communities, which have precluded their isolation and study. Microbial communities thrive in a myriad of exotic locales ranging from the extreme pressure of the ocean floor to the digestive tracts of mammals, and similarly they fulfill a number of critical, and diverse ecological roles [2•,3–5]. As such, they represent a vast untapped reservoir of enzymes for biotechnological applications such as bioenergy [2•] and drug discovery [6] and can offer new insights into the specialized function of biological systems. Next generation sequencing bypasses the need for isolation of microorganisms making possible, for the first time, the genomic and transcriptomic study of these vibrant microbial communities in their native context.

Next generation sequencing (NGS) refers to a number of technologies that generate high throughput, massively parallel sequence reads that allow whole genome assembly (WGA) or transcriptome assembly in a fraction of time compared to traditional Sanger sequencing (see [7••] for a description of common platforms that are currently used). More importantly, this is achieved at a cost that is accessible for most laboratories. All of these technologies share a common work flow, described in Figure 1. Generally, isolated DNA or cDNA libraries prepared from RNA are first fractionated into a library of small inserts. These libraries are then processed, often barcoded and amplified by either bridge PCR [8] or emulsion PCR [9], and sequenced in parallel in millions of small volume reactions. The sequence readout is typically in the form of light, which is captured and processed by high quality optics and image acquisition software. The results of these reactions are then put through a high performance computing pipeline to reassemble the individual fragments into complete gene sequences, enabling numerous downstream analyses for biological insight.

The high dynamic range, accurate base pair resolution, and exponentially plummeting costs of NGS platforms stand to empower many types of studies beyond sequence acquisition, including expression analysis by RNA sequencing (RNA-seq) and bioprospecting surveys for novel genes. As NGS only requires isolated nucleic acids, it can be extended to study the composition and function of dynamic microbial communities (see [10••] for a detailed review of applications of metagenomics/metatranscriptomics) or to analyze in detail the connection between an organism's genome/transcriptome and its apparent phenotype [2•,10••,11]. Key to this analysis has been the development of improved bioinformatics

Figure 1



NGS analytical workflow: nucleic acids are isolated either directly from crude environmental samples (in this case dung) or from isolates obtained through traditional microbial enrichment and cultivation techniques (e.g. roll tubes). Nucleic acids are typically processed to generate a library by reverse transcription (for RNA only), fragmentation of the DNA/cDNA, barcoding with adaptors for identification, amplification and sequencing, and amplification by emulsion PCR or bridge PCR. The library is then sequenced by an NGS platform and the resulting short reads assembled into a complete genome/transcriptome through assembler algorithms. The resulting sequence data are analyzed for genes and annotated. The assembled sequence can then be subjected to a number of computational and proteomic analyses to decipher key biological insights.

tools and new advances in (meta)proteomics that can help provide additional biochemical evidence to link sequence to function [11]. This review addresses the selection of appropriate NGS platforms for a given application, and discusses current tools and methods available to infer biological meaning from the expanse of NGS data. Particular emphasis will be given to examples of rumen microbial communities, which offer extraordinary promise as a source of novel lignocelluloytic enzymes.

Next generation sequencing

The last decade has given rise to a number of NGS platforms with distinct characteristics such as library insert size required, read length produced, depth (read redundancy or coverage), sequence accuracy, availability and cost [7••]. The project type and goals dictate which platforms are feasible (Table 1). For example, some platforms such as 454 Titanium are well suited for projects not requiring high sequence yield (reads per run). These include assembly of smaller genomes (< 6 Mb) and metagenomic studies such as 16S rDNA-based taxonomic profiling. In contrast, high sequence coverage is needed for *de novo* transcript assembly whereas gene expression profiling falls in the medium range. Another consideration is whether to sequence only one end of the insert (single end) versus both sides (paired end). Genome and transcriptome assembly benefits from paired end sequencing as the opposing reads are known to be linked by a sequence of defined size which aids