Journal: Advances in Wound Care

Online Publication Date: 2014

Article Type: Critical Review

Title: Next-generation sequencing: a review of technologies and tools for wound microbiome research

Authors: Brendan P Hodkinson, PhD[1], Elizabeth A Grice, PhD[1]

Author Affiliations: [1]Department of Dermatology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania

Corresponding Author Information: Elizabeth A Grice (T: +215-898-3179; F: +215-573-2033; E-mail: egrice@upenn.edu)

Abstract

Significance: The colonization of wounds by specific microbes or communities of microbes may delay healing and/or lead to infection-related complication. Studies of wound-associated microbial communities (microbiomes) to date have primarily relied upon culture-based methods, which are known to have extreme biases and are not reliable for the characterization of microbiomes. Biofilms are very resistant to culture and are therefore especially difficult to study with techniques that remain standard in clinical settings.

Recent Advances: Culture-independent approaches employing next-generation DNA sequencing have provided researchers and clinicians a window into wound-associated microbiomes that could not be achieved before and has begun to transform our view of wound-associated biodiversity. Within the past decade, many platforms have arisen for performing this type of sequencing, with various types of applications for microbiome research being possible on each.

Critical Issues: Wound care incorporating knowledge of microbiomes gained from next-generation sequencing could guide clinical management and treatments. The purpose of this review is to outline the current platforms, their applications, and the steps necessary to undertake microbiome studies using next-generation sequencing.

Future Directions: As DNA sequencing technology progresses, platforms will continue to produce longer reads and more reads per run at lower costs. A major future challenge is to implement these technologies in clinical settings for more precise and rapid identification of wound bioburden.

Table of Contents

1.0 Scope and Significance

Humans are known to host diverse, complex communities of micro-organisms that include bacteria, archaea, micro-eukaryotes and viruses. A breach in the epithelial barrier is a port of entry for microorganisms, and all wounds are contaminated to some degree by these typically commensal microbes along with others from the environment. Contamination can lead to colonization, infection (which can be recurrent), delayed healing, and potentially amputation. Next-generation sequencing provides a window into wound-associated microbial communities (microbiomes) with a reasonable cost and timeframe. In this review, we outline the current technologies and highlight some of their applications with regard to wound microbiome research.

2.0 Translational Relevance

Research into wound microbiomes to date has relied heavily on culture-based methods, which have dominated the field for decades, even though these methods are known to introduce major biases[1]. Until very recently, culture-free methods for studying microbial communities relied on imprecise fingerprinting techniques or molecular cloning followed by Sanger sequencing. While Sanger sequencing can provide an accurate picture of community composition, generating data sets large enough to allow community-wide comparisons (e.g., those designed to discern microbiome-based biomarkers) has often been time- and cost-prohibitive. With the advent of high-throughput next-generation sequencing, characterizing numerous microbial communities has become feasible and cost-effective.

3.0 Clinical Relevance

The communities of microbes associated with wounds can potentially cause recurrent infection and/or delayed healing, and may profoundly affect the local and systemic immune

response in patients[2, 3]. Biofilms, which commonly form on orthopedic hardware and may form on chronic wounds, are very resistant to culture and are therefore especially difficult to study with the culture-based techniques that remain standard in clinical settings. The future of wound care may incorporate knowledge of microbiomes gained from next-generation sequencing, to more precisely identify colonizing/infecting microbiota, and to guide management and treatment.

4.0 Discussion

4.1 What are the different next-generation sequencing platforms?

Below we introduce the five major platform types that have been used for microbiome studies (see also Table 1 and Fig. 1). This should provide a comprehensive overview of the technologies to orient those attempting to navigate the literature or design new studies. Although there are additional next-generation sequencing platforms, these are not covered in detail here because they are not currently known to be in use for microbiome research.

4.1.1 454 (Roche) GS FLX(+). 454 Life Sciences (a Roche company) brought the first next-generation sequencing technologies to market, with the overall approach being introduced in 2005[4]. The 454 family of platforms has been used ever since for a great variety of applications, and its long reads have made it especially appealing for studies of microbiomes, since longer reads can generally be identified with greater accuracy and precision.

The overall approach for 454 is pyrosequencing-based. The sequencing preparation begins with lengths of DNA (e.g., amplicons or nebulized genomic/metagenomic DNA) that have specific adapters on either end, created by using PCR primers with adapter sequences or by ligation; these are fixed to tiny beads (ideally, one bead will have one DNA fragment) that are suspended in a water-in-oil emulsion. An emulsion PCR step is then performed to make multiple

copies of each DNA fragment, resulting in a set of beads in which each one contains many cloned copies of the same DNA fragment. A fiber-optic chip filled with a field of micro-wells, known as a PicoTiterPlate, is then washed with the emulsion, allowing a single bead to drop into each well. The wells are also filled with a set of enzymes for the sequencing process. At this point, sequencing-by-synthesis can begin, with the addition of bases triggering pyrophosphate release, which produces flashes of light that are recorded to infer the sequence of the DNA fragments in each well as each base type is added.

Currently, the most advanced chemistry/platform combination in this family (GS FLX+ System with the GS FLX Titanium Sequencing Kit XL+) can produce approximately one million reads per run with reads up to 1,000 bases in length (mode read length: 700 bases). Paired-end sequencing is available, which produces pairs of reads, each of which begins at one end of a given DNA fragment. Samples can be multiplexed as long as the library is prepared in such a way that different molecular barcodes are found between the adapter sequences and the sequences of interest derived from the samples. With this approach, the indexing barcodes will appear at the beginning of each sequence (‘in-line’), allowing each sequence to be assigned to a sample bioinformatically.

One shortcoming of the 454 approach is that it frequently misidentifies the length of homopolymers (stretches of nucleotides in which all bases are identical). Additionally, this technology is often considered to be cost-ineffective when compared to other next-generation sequencing technologies because, given a limited budget, one can produce many more sequences with Illumina, SOLiD, or Ion Torrent. However, for some applications that require longer read lengths, it remains the most cost-effective platform. Although the technology broke new ground when it was introduced, 454 Life Sciences will no longer support the platform after 2016.

4.1.2 Illumina (Solexa) GA/HiSeq/MiSeq/NextSeq. Illumina produces the most widely-used family of platforms. The technology was introduced in 2006 (http://www.illumina.com/technology/solexa_technology.ilmn) and was quickly embraced by many researchers because a larger amount of data could be generated in a more cost-effective manner. Over the years, read lengths have increased so that many of those who initially would have only used 454 have switched over to Illumina platforms due to the cost-effectiveness of the technology[5-7].

Although it is a sequencing-by-synthesis method whose release followed quickly on the heels of 454, the Illumina approach differs notably from 454 in two major ways: (a) it uses a flow cell with a field of oligonucleotides attached, instead of a chip containing individual micro-wells with beads, and (b) it does not involve pyrosequencing, but rather reversible dye-terminators. The dye termination approach resembles the 'traditional' Sanger sequencing [8]. It is different from Sanger, however, in that the dye terminators are reversible, so they are removed after each imaging cycle to make way for the next reversible dye-terminated nucleotide [9].

Sequencing preparation begins with lengths of DNA that have specific adapters on either end being washed over a flow cell filled with specific oligonucleotides that hybridize to the ends of the fragments. Each fragment is then replicated to make a cluster of identical fragments. Reversible dye-terminator nucleotides are then washed over the flow cell and given time to attach; the excess nucleotides are washed away, the flow cell is imaged, and the terminators are reversed so that the process can repeat and nucleotides can continue to be added in subsequent cycles.

Currently, the longest reads produced on an Illumina platform can be found on the MiSeq, which can produce paired-end reads that are 300 bases in length each. The platform with the greatest output overall is the HiSeq 2500, producing 4 billion fragments in a paired-end

fashion with 125 bases for each read in a single run. Illumina has recently released the HiSeq X Ten, which is an array of ten HiSeq machines sold as a unit, for higher throughput than ever before. Another recent release is the NextSeq 500, which is being marketed as the first high-throughput desktop sequencer. Multiplexing for Illumina sequencing is typically handled differently from the 'in-line' barcoding approach pioneered by 454, although this option is available. Illumina sequencing often involves a separate indexing read, which requires a separate indexing primer. An additional indexing read can be run using the adapters found on the lawn of the flow cell, making it possible to employ 'dual indexing' for a greater degree of sample multiplexing.

4.1.3 Applied Biosystems SOLiD. This type of sequencing was introduced in 2007 [10] and has not reached the same level of popularity as the 454 and Illumina platforms for microbiome research. Although it does not provide the read lengths achievable through either of the previous platforms, and is not as high throughput as the Illumina HiSeq, its utility has been demonstrated for microbiome applications [11].

The SOLiD process begins with an emulsion PCR step akin to the one used by 454, but the sequencing itself is entirely different from the previously described systems. Sequencing involves a multi-round, staggered, di-base incorporation system. DNA ligase is used for incorporation, making it a 'sequencing-by-ligation' approach, as opposed to the 'sequencing-by-synthesis' approaches mentioned previously. Mardis [12] provides a thorough overview of the complex sequencing and decoding processes involved with using this system.

The SOLiD 5500xl W Genetic Analyzer produces up to 3 billion reads per run with reads that are 75 bases long. Paired-end sequencing is available, but with the second read in the pair being only 35 bases long. Multiplexing of samples is possible through a system akin to the one used by Illumina, with a separate indexing run; while standard in-line molecular barcode

sequencing would be possible, the short reads make this inadvisable. Although it can generate large numbers of sequences in a run, the persistent short read length has greatly limited its utility.

4.1.4 Ion Torrent PGM/Proton. By 2010, 454 had carved out a niche of providing longer reads, while Illumina and SOLiD had demonstrated the ability to provide massive numbers of sequences all in one shot. At this point, each company began to produce platforms that would cater toward a new type of customer: the researcher who could benefit from next-generation technologies but does not require data sets of the magnitude possible through the standard platforms. Within a short period of time, the 454 GS Junior, the Illumina MiSeq, and the SOLiD FlowChip were all released and geared toward those wanting something more scaled down. Ion Torrent entered the market in 2010 with the Personal Genome Machine (PGM), claiming to be the first company to truly bring next-generation sequencing to the masses by making it feasible and affordable for smaller laboratories.

The Ion Torrent system begins in a manner similar to 454, with a plate of micro-wells containing beads to which DNA fragments are attached. It differs from all of the other systems, however, in the manner in which base incorporation is detected. When a base is added to a growing DNA strand, a proton is released, which slightly alters the surrounding pH. Micro-detectors sensitive to pH are associated with the wells on the plate, which is itself a semi-conductor chip, and they record when these changes occur. As the different bases are washed sequentially through, additions are recorded, allowing the sequence from each well to be inferred.

The Ion Proton platform currently produces the highest output, with up to 50 million reads per run that have read lengths of ~200 bases, while the PGM (which has an output that is about an order of magnitude lower as far as read count) has the longest reads at ~400 bases.

One interesting feature, however, is that fragments longer than those that can be fully sequenced through this system are currently removed through a size-selection step, making it impossible to sequence the ends of longer fragments. Bi-directional sequencing is available, but 'pairing' the reads themselves does not seem to be reliable with this technology in its current state [13]. Multiplexing is possible through the standard in-line molecular barcode sequencing. Like 454, Ion Torrent is also susceptible to homopolymer-related errors. The Ion Torrent approach can be quite effective for generating microbiome data [14, 15], although the strict size selection imposed and the lack of reliable mate-pairing for bidirectional reads hinder this technology from being more widely adopted by microbiome researchers.

4.1.5 PacBio RS. Pacific Biosciences (PacBio) uses a Single-Molecule Real-Time (SMRT) sequencing approach. Although Helicos BioSciences produced the first single-molecule sequencing platform, PacBio has had much greater commercial success and currently leads the way for single-molecule sequencing. When the PacBio technology was first released, there was a great deal of concern regarding the high error-rates in base calls. However, the company has since incorporated Circular Consensus Sequencing (CCS) into their system, which has greatly reduced error rates by allowing fragments to be sequenced repeatedly and thereby checked for errors.

The PacBio sequencing system involves no amplification step, setting it apart from the other major next-generation sequencing systems. The sequencing is performed on a chip containing many zero-mode waveguide (ZMW) detectors. DNA polymerases are attached to the ZMW detectors and phospholinked dye-labeled nucleotide incorporation is imaged in real time as DNA strands are synthesized. PacBio's RS II C2 XL currently offers both the greatest read lengths (averaging around 4,600 bases) and the highest number of reads per run (about 47,000). The typical 'paired-end' approach is not used with PacBio, since reads are typically long enough that fragments, through CCS, can be covered multiple times without having to

sequence from each end independently. Multiplexing with PacBio does not involve an independent read, but rather follows the standard 'in-line' barcoding model.

4.2 What are the major microbiome applications of next-generation sequencing?

The various sequence-based, culture-independent microbiome studies typically have many elements in common, and a similar workflow is necessary for each (Fig. 2). However, the particular questions being addressed will guide the experimental design and the methodology for generating, processing and interpreting data. The main approaches used for examining and characterizing microbiomes are outlined below.

4.2.1 Amplicon-based profiling. Methods that employ the sequencing of amplicon populations allow one to construct detailed community profiles of microbiota samples based on the relative abundances of the taxa that they contain. The diverse sequences from a single gene found in each of the organisms can serve as proxies for the taxa that they represent. Downstream analyses of sequence libraries can be performed to discern whether there are correlations between certain factors and (a) particular taxa or (b) shifts in overall community structure.

4.2.1.1 Bacterial community profiling (16S amplicon sequencing). The best-studied part of the human microbiome is the bacterial portion. Bacteria make up the majority of the organisms on and in the human body and well-established procedures and workflows are in place for their study. By far the most popular genomic region for studying bacterial diversity is the gene encoding the RNA for the ribosomal small subunit, typically known as '16S.' This gene is ideal for a number of reasons: (a) it is present in all bacteria; (b) it contains stretches within it that are nearly universal in sequence throughout all bacteria, and (c) it contains hyper-variable regions that are widely divergent between different taxa. The pattern of extremely conserved regions interspersed with hyper-variable regions makes it possible to target the gene with primers and

also use it to identify taxa with some level of precision [16]. Primers that will universally anneal to the bacterial 16S region are used to PCR amplify the diverse fragments of the gene found in the different organisms of a given DNA sample. In this way, a population of 16S amplicons is produced that reflects the composition of the community of organisms in a given sample.

4.2.1.2 Fungal community profiling (ITS, LSU & SSU amplicon sequencing). For surveys of fungal diversity, there is less of a universal consensus on the gene of interest. The three most commonly used loci are all ribosomal, and are known as the internal transcribed spacer (ITS), large subunit (LSU) and small subunit (SSU) regions. It is noteworthy that 'SSU' is the more general term for the gene that is called '16S' in bacteria, although it is typically called '18S' in Eukaryotes since it has a larger molecular weight. Of the three most commonly-used ribosomal amplicons, ITS is the most effective locus for providing species-level identifications. In fact, this locus is now commonly used as the fungal 'species barcode' region because it nearly always contains a sufficient level of variation for species differentiation [17]. The LSU and SSU loci are more conserved, and are therefore quite effective for phylogenetically-based microbiome analyses [18, 19].

4.2.2 Shotgun sequencing methods. While the amplicon sequencing methods described above work well for broad characterization and comparison of communities, they contain inherent biases that come from the use of specific primers and multiple cycles of amplification. Shotgun methods allow profiling of the whole community (including viruses, archaea, and micro-eukaryotes) based on fragments from throughout the genomes/transcriptomes of the diverse organisms contained therein. Perhaps most importantly, this type of approach can give direct information regarding function. In certain sample types, obtaining a large enough quantity of DNA can be difficult, and should be taken into consideration when deciding which sample preparation and sequencing protocols are to be used.

4.2.2.1 Metagenomics (shotgun DNA sequencing). Shotgun metagenomic sequencing makes it possible to examine both the taxonomic composition and the functional genetic potential of a community. Since there are no markers that work across all of life (including viruses), this approach is currently the only way to profile whole microbial communities. Metagenomics generally allows for more accurate determination of the relative abundances of different organisms, since it typically involves little or no DNA amplification, which can introduce biases. Most often metagenomic shotgun sequencing is used to understand the functional potential of communities, which is typically inferred by querying sequence reads against databases, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways database [20] and/or COG (Clusters of Orthologous Groups of proteins) database functional categories [21]. An example of the information gleaned from metagenomic shotgun sequencing lies in a study of the gut metagenome and its association with obesity. Obese mice were observed to have an increased ratio of Firmicutes to Bacteroidetes, and this compositional shift translated into changes in the metabolic potential of the gut microbiota, where the obese mouse gut metagenome was enriched with genes for energy harvest [22]. One weakness of metagenomic shotgun sequencing is that analysis often involves comparison between different parts of different genomes, making accurate classification (functional and/or taxonomic) a crucial step. Classification can be unreliable, though, since the number of whole genome reference sequences currently available is limited. Traditional microbial whole genome sequencing relies on the ability to culture the micro-organism, which can be difficult as most micro-organisms do not thrive under standard culture conditions or in isolation. However, developing technologies, such as co-culturing single cells in gel microdroplets [23], are promising solutions to this obstacle. Advances in single-cell genomics and culturing technologies in addition to high-throughput sequencing advances should help grow reference genome databases. Finally, for samples where metagenomic shotgun sequencing may not be feasible, prediction of metagenome

functional content is possible with tools such as PICRUSt, which relies on marker gene data and reference genomes to infer composite metagenomes [24].

4.2.2.2 Metatranscriptomics (RNA transcript sequencing). With metatranscriptomic sequencing, the full range of actively transcribed genes can be examined in any context. This approach makes it possible to take a snapshot of the activity happening at the molecular level in the organisms found in a sample. The transcripts can be genes from both the host and the members of the microbiome, so one can begin to examine host-microbe interactions and determine not only how the microbiome itself behaves, but how the host may react to the members of the microbiome. Researchers have even been able to show collaboration between the host and associated microbes for performing specific functions that are critical for host survival [25]. The problems of classification mentioned above regarding metagenomics hold here as well when examining transcription in the diverse assemblage of organisms in the microbiome.

4.3 What bioinformatics tools and skills are needed for analysis?

A number of open source software packages integrate the analysis steps for next-generation microbiome sequence data. The two main programs used are QIIME [26] and mothur [27], both of which provide automated scripts/commands for performing complex steps while remaining customizable to many different types of data sets and experimental designs. The CloVR [28] and MG-RAST [29] programs provide an even greater deal of automation (the former actually including QIIME and mothur commands), but offer less customizability. Many of these packages provide tutorials and documentation on their websites that are useful in orienting the user to the different workflows and processing and analysis steps that are available.

4.3.1 Sequence Pre-Processing. Typically, the initial output of next-generation sequencing is formatted in a way that is specific to the platform. Mothur and QIIME take most file types produced by sequencing platforms (e.g., .sff files produced by the 454 platform) and can perform the majority of the pre-processing steps described below from there. If necessary, files produced by the sequencing platform can be converted to FASTQ format using either software produced by the platform's manufacturer or custom scripts. Then one must consider not only basic file format (e.g., FASTQ, FASTA, etc.), but also the arrangement of the files (e.g., whether barcodes are contained in a separate file or are in-line at the beginning of each sequence) and the way in which data are encoded in the definition lines. Some of the programs that are useful for this, in addition to mothur and QIIME, are: fastx-toolkit [30] and TagCleaner [31]. However, these tools may fall short of providing complete solutions for sequence processing, and custom processing scripts may need to be written (using, e.g., BASH, Python, and/or Perl).

Removal of low quality sequences is imperative in microbiome studies, as variation introduced by error will inflate diversity estimations and suggest the presence of novel organisms. Sequences suspected to contain raw sequencing errors should be discarded, and different parameters have been described to aid in detection of these sequences [32]. Sequencing of a mock community, made up of known micro-organisms in known quantities, in parallel with experimental samples can provide an estimate of error rate. Additionally, chimeras that are produced during PCR amplification steps should be identified and removed from the dataset using tools designed for this purpose, such as ChimerSlayer or UCHIME [33, 34].

4.3.2 Assembly. For shotgun-style methods, one important step in the preparation of the data set is assembling the reads into longer stretches of DNA (i.e., contigs and/or scaffolds) based on the consensus of overlapping sequence reads. When assembling multiple genomes from samples with many different organisms (as is typically the case for microbiome studies that employ shotgun DNA sequencing), specialized assembly algorithms are required so that

false/chimeric assemblies are minimized. Some programs that are geared toward assembly from metagenomic data are MetaVelvet [35], IDBA-UD[36], MetaPar [37], and MetAMOS [38]. Assembly is a challenge for heterogeneous genomes when micro-organisms are present in low abundance and thus only incomplete coverage can be achieved. These challenges are compounded by the fact that reference genomes are not available for most micro-organisms. For amplicon-based methods, assembly is often necessary when a paired-end approach has been used. To join overlapping pairs of sequences, specialized programs such as PANDAseq [39] and PEAR [40] have been written, but this task can also be performed within QIIME and mothur.

4.3.3 Characterization. To make biological sense out of the sequence data generated through next-generation technologies, one can begin by determining the within-sample (alpha-) diversity, the between-sample (beta-) diversity, the taxonomic composition, and the functional composition of the communities being studied. QIIME and mothur are ideal for determining diversity metrics and assigning taxonomy to amplicon sequences. MG-RAST allows a big-picture look at both the taxonomic and functional composition of a data set, but with limited customizability. Two programs that allow more detailed and customizable functional assessments of shotgun data are MEGAN [41] and BLAST2GO [42]. Pros and cons of different approaches to calculating operational taxonomic units, assigning taxonomy, inferring phylogeny, and calculating diversity metrics are extensively described in other reviews[43, 44].

4.3.4 Statistics & Visualization. After the broad characterization of microbial communities, the next steps are (a) to test for correlations/associations between aspects of the microbiome and various factors and (b) to visualize the results. The programs QIIME, mothur, and MG-RAST provide some tools for statistical analysis and visualization. More advanced analyses and visualization can be performed in R[45]; other software packages that provide the ability to perform basic statistical analysis and data visualization are Matlab, SAS, SPSS, and Stata. While R is typically considered to have a steep learning curve, a strong background in

programming is not necessarily required. For those with programming experience, Python can prove quite useful (especially in conjunction with SciPy and NumPy) for this purpose.

5.0 Summary

In the past decade, next-generation sequencing has enabled researchers to answer questions that were previously intractable. The market potential of this technology has spawned numerous platforms in a relatively short period of time, and new platforms are constantly being developed. As technology progresses, a major goal will be to fill in the sequencing space with platforms that can produce longer reads and more reads per run (i.e., add to the upper right portion of Fig. 1). The area of microbiome research has benefitted greatly from the advent of next-generation sequencing, and is one discipline that has grown by leaps and bounds in recent years as a result. A variety of computational tools and software packages have been developed to deal with data from next-generation sequencing platforms. Studies that utilize culture-independent next generation sequencing approaches are beginning to provide valuable insight into the composition, diversity, and dynamics of wound bioburden, and its relationship to impaired healing and development of infection-related complication. A major challenge in the future will be bringing this technology to the clinic as a potential diagnostic and/or prognostic tool.

6.0 Take-Home Messages

1. Next-generation sequencing has made it cost- and time-effective to fully characterize wound-associated microbial communities.

2. There are currently five major next-generation sequencing platform families used in microbiome studies, all of which have strengths and weaknesses that must be weighed when designing an experiment.

3. Amplicon-based methods are effective for characterizing and comparing the overall taxonomic/phylogenetic composition of bacterial and fungal communities.

4. Shotgun sequencing methods are effective for characterizing communities of micro-organisms including viruses, archaea, and non-fungal eukaryotes, and allow one to investigate (a) functional potential of the organisms, by examining genomic DNA (metagenomics), or (b) functions being performed in the cells, by looking at RNA transcripts (metatranscriptomics).

5. Bioinformatics associated with next-generation sequencing can typically be divided into the following major categories (listed in order): sequence pre-processing, sequence assembly, community characterization, hypothesis testing (within a statistical framework), and data visualization.

7.0 Acknowledgements and Funding Sources

The authors thank the members EAG's laboratory for discussions of next-generation sequencing platforms and the associated bioinformatics. BPH was funded by an NIH/NIAMS T32 Grant Award to the University of Pennsylvania Department of Dermatology (AR007465).

8.0 Author Disclosure and Ghostwriting

The authors do not declare any conflicts of interest. The authors do not declare any ghostwriter contributions.

9.0 About the Authors

BPH is a Post-Doctoral Research Fellow in the Department of Dermatology at the University of Pennsylvania Perelman School of Medicine and is a member of EAG's laboratory. He received his PhD in 2011 from Duke University where his studies focused on the use of next-generation

sequencing for microbiome research. **EAG** is an Assistant Professor of Dermatology at the University of Pennsylvania Perelman School of Medicine. She received her doctorate in Human Genetics in 2006 from Johns Hopkins University, and performed post-doctoral work at the National Institutes of Health. Her research interests focus on genomic and metagenomic analyses of cutaneous host-microbe interactions.

10.0 References

1. Scales BS and Huffnagle GB. The microbiome in wound repair and tissue fibrosis. *The Journal of pathology* 229: 323-331, 2013.
2. Grice EA and Segre JA. Interaction of the microbiome with the innate immune response in chronic wounds. *Adv Exp Med Biol* 946: 55-68, 2012.
3. Grice EA, Snitkin ES, Yockey LJ, Bermudez DM, Liechty KW, and Segre JA. Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proc Natl Acad Sci U S A* 107: 14799-14804, 2010.
4. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, and Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380, 2005.
5. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G, and Schmidt AH. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics* 15: 63, 2014.
6. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, and Ravel J. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* 2: 6, 2014.
7. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, and Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*, 2012.
8. Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463-5467, 1977.
9. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara

ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53-59, 2008.
10. Peckham HE, McLaughlin SF, Ni JN, Rhodes MD, Malek JA, McKernan KJ, and Blanchard AP. SOLiDTM Sequencing and 2-Base Encoding. In: *American Society of Human Genetics*, 2007, p. 2624.
11. Mitra S, Forster-Fromme K, Damms-Machado A, Scheurenbrand T, Biskup S, Huson DH, and Bischoff SC. Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. *BMC Genomics* 14 Suppl 5: S16, 2013.
12. Mardis ER. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* 9: 387-402, 2008.
13. Ho A. *Next-generation sequencing: Acquisition, analysis, and assembly*: University of New Mexico, 2013.
14. Milani C, Hevia A, Foroni E, Duranti S, Turroni F, Lugli GA, Sanchez B, Martin R, Gueimonde M, van Sinderen D, Margolles A, and Ventura M. Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. *PLoS One* 8: e68739, 2013.
15. Yergeau E, Lawrence JR, Sanschagrin S, Waiser MJ, Korber DR, and Greer CW. Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Appl Environ Microbiol* 78: 7626-7637, 2012.
16. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, and Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* 82: 6955-6959, 1985.
17. Seifert KA. Progress towards DNA barcoding of fungi. *Molecular Ecology Resources* 9 Suppl s1: 83-89, 2009.
18. Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, Nabel CS, Hill DA, Artis D, Bachman MA, Custers-Allen R, Grunberg S, Wu GD, Lewis JD, and Bushman FD. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol* 13: R60, 2012.
19. Hodkinson BP and Lendemer JC. Next-generation sequencing reveals sterile crustose lichen phylogeny. *Mycosphere* 4: 1028-1039, 2013.
20. Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32: D277-280, 2004.
21. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, and Natale DA. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41, 2003.
22. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, and Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027-1031, 2006.
23. Fitzsimons MS, Novotny M, Lo CC, Dichosa AE, Yee-Greenbaum JL, Snook JP, Gu W, Chertkov O, Davenport KW, McMurry K, Reitenga KG, Daughton AR, He J, Johnson SL, Gleasner CD, Wills PL, Parson-Quintana B, Chain PS, Detter JC, Lasken RS, and Han CS. Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Res* 23: 878-888, 2013.
24. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, and Huttenhower C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 31: 814-821, 2013.

25. Tartar A, Wheeler MM, Zhou X, Coy MR, Boucias DG, and Scharf ME. Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes. *Biotechnology for Biofuels* 2: 25, 2009.
26. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, and Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7: 335-336, 2010.
27. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, and Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537-7541, 2009.
28. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, and Fricke WF. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12: 356, 2011.
29. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, and Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386, 2008.
30. Pearson WR, Wood T, Zhang Z, and Miller W. Comparison of DNA sequences with protein sequences. *Genomics* 46: 24-36, 1997.
31. Schmieder R, Lim YW, Rohwer F, and Edwards R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11: 341, 2010.
32. Schloss PD, Gevers D, and Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6: e27310, 2011.
33. Edgar RC, Haas BJ, Clemente JC, Quince C, and Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194-2200, 2011.
34. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methe B, DeSantis TZ, Petrosino JF, Knight R, and Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 21: 494-504, 2011.
35. Namiki T, Hachiya T, Tanaka H, and Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 40: e155, 2012.
36. Peng Y, Leung HC, Yiu SM, and Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420-1428, 2012.
37. Kim M, Ligo JG, Emad A, Farnoud F, Milenkovic O, and Veeravalli VV. MetaPar: Metagenomic Sequence Assembly via Iterative Reclassification. *arXiv* 1311.3932, 2013.
38. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, and Pop M. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol* 14: R2, 2013.
39. Masella AP, Bartram AK, Truszkowski JM, Brown DG, and Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13: 31, 2012.
40. Zhang J, Kobert K, Flouri TX, and Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30: 614-620, 2014.
41. Huson DH, Auch AF, Qi J, and Schuster SC. MEGAN analysis of metagenomic data. *Genome Res* 17: 377-386, 2007.

42. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, and Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676, 2005.
43. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, and Knight R. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet* 13: 47-58, 2012.
44. Hamady M and Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* 19: 1141-1152, 2009.
45. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.

11.0 Tables

Table 1. Summary of the five major next-generation sequencing platform families. The average read length is given for the platform/chemistry combination in each family that provides the longest reads.

Platform Family	Clonal Amplification	Chemistry	Highest Average Read Length
454	Emulsion PCR	Pyrosequencing (seq-by-synthesis)	700 bp (paired-end sequencing available)
Illumina	Bridge Amplification	Reversible dye-terminator (seq-by-synthesis)	300 bp (overlapping paired-end sequencing available)
SOLiD	Emulsion PCR	Oligonucleotide 8-mer chained ligation (seq-by-ligation)	75 bp (paired-end sequencing available)
Ion Torrent	Emulsion PCR	Proton detection (seq-by-synthesis)	400 bp (bidirectional sequencing available)
PacBio	N/A (single molecule)	Phospholinked fluorescent nucleotides (seq-by-synthesis)	8,500 bp

12.0 Figures

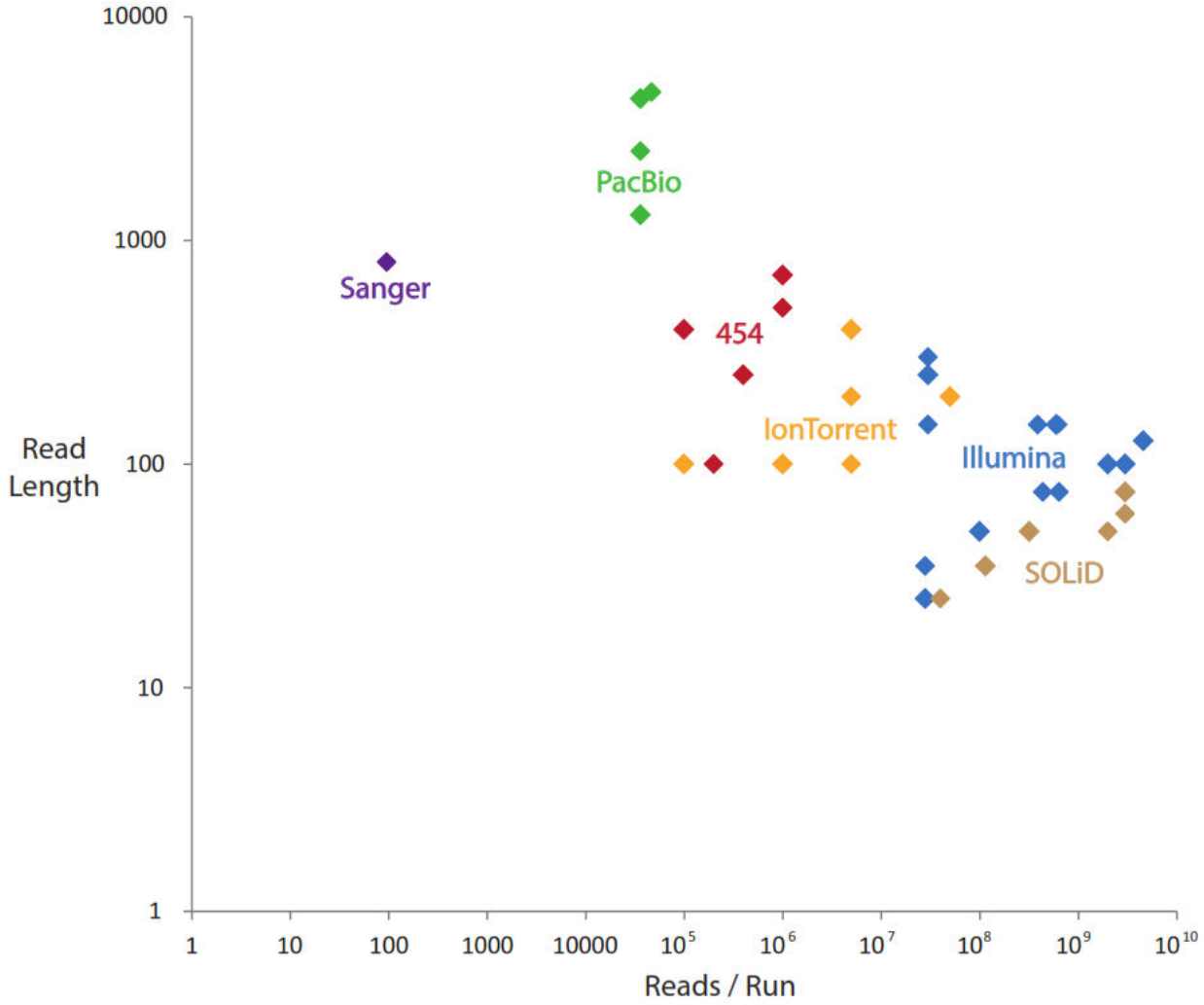

Fig. 1. Sequencing space based on read length (in bases) and number of reads per run. Points represent official platform/chemistry combination releases and are color-coded based on the platform family.

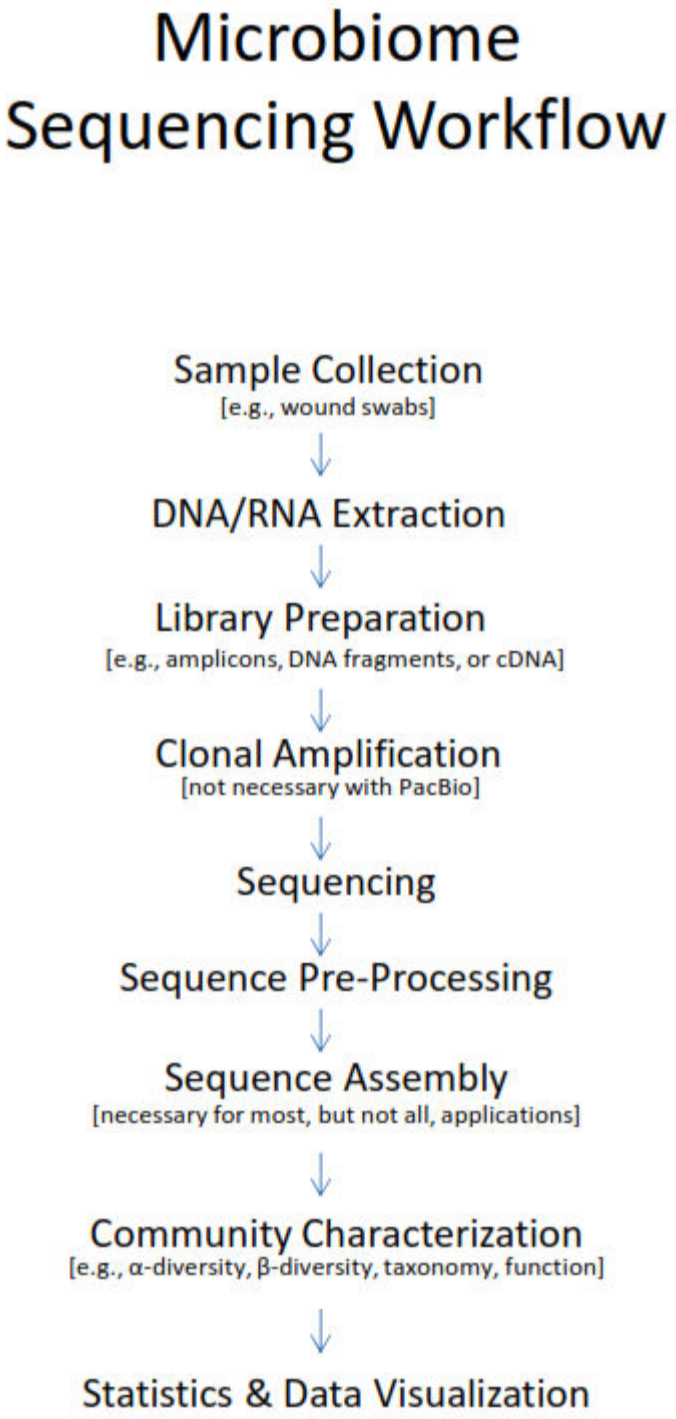

Fig. 2. Standard sequencing workflow for microbiome research.