Introduction

At birth, neonatal organs are either sterile or have low numbers of bacteria, but by adulthood, there are between 10 and 100 trillion microbes residing in the human body. Microbial environments in neonates and adults of interest include the skin, nose and oropharynx, the lower respiratory tract, the gastrointestinal tract with its regional differences, and the urogenital tract with special interest in the vagina [1]. Of these habitats, microbes living in the small and large intestine of prematurely born infants have received the most attention. Fecal specimens showing reduced gut-related bacterial diversity and the presence of potential pathogens are associated with late-onset neonatal sepsis [2]. Moreover, a shift in the fecal microbiota to a predominance of the phylum Proteobacteria and a reduction in the phylum Firmicutes has been seen 72 h before the onset of necrotizing enterocolitis [3]. Understanding how microbial diversity in infants relates to neonatal infections is an important aspect of this research.

Presently, it is appreciated that the composition of intestinal microbiota is influenced by an infant’s genetic background [4, 5], vaginal or caesarean delivery [6], residence at home versus in the hospital [7–9], the adminis-
tration of antibiotics [10, 11], and a diet consisting of parenteral nutrition, breast milk, formula, fortifiers or a combination of those nutrients [12, 13]. The dynamic ecosystem developing in the gut plays major roles in nutritional, metabolic, physiologic, and immunologic processes [14–16]. Current concepts suggest that the development of intestinal microbiota have associations with the pathogenesis of obesity [17, 18], diabetes [19, 20], and allergic disease [21, 22]. Given the effects of antibiotics on the composition of microbes in the neonatal intestine and their effects on future health, proper use of these agents has vast significance.

Bacterial classification or taxonomy has been the focus in studying neonatal environments, but the field is rapidly moving towards comprehending functional relationships among microbes and between microbes and their host [23]. Future research must integrate data about the microbiota with (a) the complete set of RNA transcripts produced by environmental genomes at study sites (transcriptome), (b) the complete array of structural and functional proteins expressed by genomes in a habitat (proteome), and (c) the full complement of small molecular metabolites identified in a biological sample from a host and subject-related microbes (metabolome) [24]. Only when the classification of bacteria in habitats is integrated with their effects on cellular function will the field of metagenomics influence neonatal therapeutics.

The remainder of this article discusses metagenomics and because certain words may be unfamiliar, the Appendix has a glossary of terms that are employed in this review.

A Historical Perspective of Metagenomics

The ensuing points recount the switch from culture-based microbiology to culture-independent DNA sequencing to identify microbes in the body.

• Over 125 years ago, Koch and colleagues invented culture-based methods for bacterial isolation [25]. Laboratory medicine still uses these techniques to diagnose infection, but molecular systems are supplanting some of those procedures [26].

• Woese and Fox [27] used the small subunit of ribosomal nucleic acid (16S rRNA) of microorganisms and discovered a new domain, the Archaea. In establishing phylogenetics [28], Professor Woese gave scientists a roadmap to a new discipline.

• Pace et al. [29] proposed cloning DNA directly from environmental samples. This task was accomplished and it took microbiology further into the molecular era [30].

• While this field initially started with cloning of environmental DNA, it was followed by screening for gene expression and function in soil microbes [31]. Shotgun sequencing of environmental DNA showed the use of this approach in more robust projects [32, 33].

• Crucial to studying microbes in human habitats, Nobel Laureates Fred Sanger and Walter Gilbert had previously invented methods of DNA sequencing [34–36]. In the next decade, Nobel Laureate Kary Mullis described the polymerase chain reaction (PCR) to amplify miniscule amounts of DNA in a specimen [37]. These methods were absolutely required to advance clinical and investigative medicine, but their laborious nature initiated a search for automated DNA analyses.

• In 1996, a novel method called pyrosequencing was reported [38]. Pyrosequencing was the birth of the ‘next or 2nd generation’ of DNA analytical methods and was based on detection of pyrophosphate release when a nucleotide is incorporated into DNA [39]. Conversely, Sanger sequencing is based on chain termination with dideoxynucleotides [36]. Commercial instrumentation using pyrosequencing to analyze DNA became available in 2005, and this method is less time-consuming and more cost-effective [40].

• Two years after the pyrosequencing method was announced, Handelsman coined the term metagenome to describe cloning of the collective genomes and the biosynthetic machinery of soil microflora so their functions could be better appreciated [31]. Hence, an analytical way to study complex genomes in an environment came forth in the same timeframe as the concept of metagenomics.

The Human Microbiome Project

In 2001, Nobel Laureate Joshua Lederberg [41] devised the word ‘microbiome’ to describe the microbiota and their collective genomes as an ecologic community living in the human body. The collective bacteria and their genes in the intestine of infants is an example of a ‘neonatal microbiome’. Relman and Falkow [42] also suggested in 2001 that a ‘second human genome project’ define the interactions between microbial commensals and pathogens and their hosts. This idea gained momentum and the Human Microbiome Project (HMP) was born in 2007 [43]. In the remainder of this review, the analytical
systems commonly used to study neonatal microbiomes are explained. Thereafter, the bioinformatics used to correct errors that are inherent in the experimental methods that investigate microbiomes are summarized. Finally, how bioinformatics converts raw data to representations of an environmental microbiome in a scientific publication are depicted and described.

**Workflow Protocols Used in Studying a Neonatal Microbiome**

The term ‘workflow’ has two connotations in metagenomics. The first involves protocols used while performing genomic studies on different analytical instruments. These commercial instruments are also known as platforms. A second meaning for workflow is the application of computerized methods to obtain data from an analytical device. Workflow procedures depend on the instrument used for genomic analysis. The following automated platforms will be discussed.

**Next or 2nd-Generation Sequencing Platforms and the Year Introduced**

(a) 454 Life Sciences Analyzers: GS FLX+ Titanium & GS Junior Systems; 2005.

(b) Illumina/Solexa Analyzers: Genome IIx, HiSeq and MiSeq Systems; 2006.

(c) Life Technologies Analyzers: Ion Torrent and Ion Proton Systems; 2011.

**3rd-Generation Platforms (Single-Molecule Sequencing) and the Year Introduced**


Critically important to any protocol involving DNA analysis is the collection, the handling and the storage of samples, and quality assurance procedures must be applied [44–46]. New investigators to the field must be highly knowledgeable about the methods that yield optimal specimens for analyses. For example, the collection and storage of feces may be less detailed than that for studying intestinal tissue.

Figure 1 depicts the three major steps used during DNA analysis: preparation, immobilization, and sequencing. Immobilization is critical to ‘Next Gen’ technology because amplification of DNA is required for sufficient analytical material during sequencing. The Next Generation system used most often to study microbiota in neonatal environments is the 454 Life Sciences platform. For this reason, the methods describing the 454 system are more completely presented. The first to last step in 454 pyrosequencing are: (1) isolation of genomic DNA; (2) DNA fragments are generated [47]; (3) the ends of DNA fragments are repaired; (4) single-stranded DNA (ssDNA) fragments are used to generate an amplicon library (fig. 2a) [46, 48]; (5) the amplicon library is then pooled, cleaned, size-selected, and quantified; (6) ssDNA is then annealed to capture beads; (7) beads are emulsified in water-in-oil microreactors (one amplicon attached to
a single bead per well); (8) emulsion PCR (emPCR) creates clonal amplification of ssDNA on each bead; (9) DNA-positive beads are recovered from the microreactors and a bead is loaded along with packing beads and the reaction enzymes into a well of the fiber-optic chip (Pico Titer™ Plate); (10) the Pico Titer™ Plate is centrifuged and placed in the sequencer; (11) adenine, cytosine, guanine, or thymidine are separately added and flow over plates during sequencing, and (12) a record of nucleotide incorporation into a DNA strand is based on light emission (i.e. pyrosequencing) and a graphic image is generated (fig. 3). Quality control measures optimize pyrosequencing, while errors that occur during amplification must be rectified before data analyses [49]. In summary, 454 Life Science analyzers detect incorporation of nucleotides into rDNA using the ‘sequencing by synthesis’ approach. Since a pyrogram cannot be retained for long-term use, the record is transformed into a computerized format for storage and analysis [50, 51].

The 1,542 nucleotides that comprise 16S rRNA have conserved and variable regions (fig. 2b). This structure is exploited during pyrosequencing to define phylogenetic classifications in microbial communities [52]. Amplicons should have hypervariable regions (e.g. V1–V3) selected specifically to study a given microbial environment [46, 53–56]. The type of genome analyzer used, and whether the device has short or long read lengths, influences which region or regions of hypervariable 16S rRNA that are selected for microbial classification. If only one region (e.g. V4) is selected, then the depth of sequencing (i.e. the total number of reads) may overcome limitations associated with short reads that incompletely interrogate 16S rRNA.

**Fig. 2.** Identification of bacteria during 454 Life Sciences sequencing. **a** A fragment isolated from genomic DNA that is used in 454 pyrosequencing. The fragment has a 454 A specific adapter, a barcode (red) used to identify a particular sample (also termed the Multiplex Identifier or MID), and a forward primer complementary to a 16S rDNA region. At the opposite end of the fragment, there is a 454 B adapter, a barcode, and a reverse primer complementary to a separate region of 16S rDNA. **b** The small subunit of bacterial rRNA which has conserved (orange) and hypervariable regions (gray with numbers). The variable regions are the ‘fingerprint’ for microbial classification. Double-headed arrow lines show variable regions commonly used for microbial classification during next generation sequencing. After a 454 pyrosequencing run, a 97% or greater similarity of a sequence to a region of bacterial 16S rRNA is used customarily to assign an OTU or ‘Genus and species’.

**Next and 3rd-Generation Instrumentation for Studying a Neonatal Microbiome**

**454 Life Sciences Analyzers**

This system is described above because of its frequent use to describe neonatal microbiomes. It is probable that pyrosequencing will be used for the immediate future to study neonatal habitats.

**Illumina/Solexa Systems**

The initial steps used by the Illumina and 454 platforms are similar through the creation of a sequencing...
Thereafter, the methods are divergent. DNA fragments are attached to a flow cell rather than beads. Bridge amplification is followed by cluster generation, and finally, sequencing primers are annealed to linear DNA amplicons in the clusters (fig. 1b). Sequencing by synthesis occurs after reversible terminator bases and fluorescently labeled nucleotides are added. Non-incorporated nucleotides are washed away, a critical step before imaging because retained, unincorporated fluorescently labeled nucleotides are recorded. Following laser excitation, the emitted fluorescence from each cluster in a flow cell is captured photographically. This results in a large number of reads from the clusters. The accuracy of the Illumina MiSeq platform may decline as read length increases, and depending on the rRNA variable regions used, its shorter reads may hinder microbial classification. Nevertheless, the MiSeq platform is gaining popularity because of lower device and operating costs versus 454 pyrosequencing. A comparison of 454 Life Sciences and Illumina analytical methods to characterize a neonatal microbiome has not taken place. Defining which device performs better should be a future research goal.

**Ion Torrent Systems**

This analyzer is a 2.5-generation device because incorporation of nucleotide uses a non-optical method compared to ‘light’ detection. The platform has a semiconductor that senses release of a proton when a nucleotide is added to a complementary strand of DNA. The purchase cost is lower and the run times are shorter compared to other platforms; however, the raw error rate is higher [57]. This device has been used to study an outbreak of multidrug-resistant *Escherichia coli* in a neonatal unit [58].

The 3rd-generation platforms are called ‘a marriage of nanotechnology with molecular biology’. Two novel devices are briefly described and discussed.
PacBio RS System

The protocol creates a DNA library, but there is no amplification step. The device uses a ‘single molecule real time’ called SMRT® sequencing method. The first innovation uses four specific fluorescent labels that are attached to the terminal phosphate rather than a nucleotide base. During incorporation by a DNA polymerase, the fluorescent label on the nucleotide is cleaved leaving a native DNA strand for incorporation of the next dNTP.

The second innovation is the nanophotonic chamber called the Zero Mode Waveguide (ZMW). As a dNTP rapidly flows in and out of the ZMW (a 70-nm cylindrical metallic chamber on a glass support), light is detected as the fluorescent label is excited during nucleotide addition. The rapid incorporation of dNTPs along a single DNA strand results in long reads that are further amplified by multiple ZMV chambers. The device can potentially sequence an entire genome rapidly, but an observed raw error rate of 12.9% raises concern [57]. This platform was used to analyze the Haitian cholera epidemic [59], but it has not studied neonatal microbiomes.

Oxford Nanopore Technologies

This platform also does not use DNA amplification. The technology uses either natural, manufactured, or hybrid nanopores with a 1-nm internal diameter that are embedded in an electrically resistant membrane bilayer. There are thousands of nanopores in a membrane. An ionic current passes through the nanopore establishing a voltage across the membrane. Two types of DNA sequencing are utilized: (a) strand sequencing of single nucleotides wherein ssDNA passes through the nanopore [60], and (b) exonuclease sequencing in which single nucleotides are released enzymatically from ssDNA at the nanopore opening and trapped by a beta-dextrin adapter while flowing through the nanopore [61]. A distinctive disruption in current by the biomolecule distinguishes between adenine, cytosine, guanine, and thymidine. Oxford has two analytical systems, the GridION™ and the miniaturized MinION™. The platform has a high (4%) error rate which Oxford says will be reduced to 0.1–1.0% [62]. Oxford Nanopore devices have not studied neonatal microbiomes.

Issues associated with each genome analyzer, such as read length, error rates in reads, and cost, are becoming evident [63, 64]. Space limitations required describing analytical devices without visual aids, but we encourage readers to view videos on ‘YouTube’ which elucidate the mechanisms used by each manufacturer’s platform.

Sequencing Analysis and Data Presentation: Defining a Neonatal Microbiome

Metagenomic analyses are not straightforward. The proper computational tools, training and experience, and collaboration with biostatisticians are required to obtain quality endpoints [51]. The workflow for data analysis has three stages [23]. First, raw data in a storage format are filtered depending on the sequencing platform and the nature of the research. Programs must ascertain read quality (i.e. identify and remove substitutions, insertions, and deletions), detect and eliminate chimeric sequences, assess read length after removing low-quality bases, and remove artifacts [65–67]. Phred and other programs like DRISEE assign a quality score to each base in a sequence [68]; however, the method of acceptance or removal of bases must not be excessively conservative. Chimeras are artificial DNA sequences generated during amplification and can be falsely interpreted as novel bacteria and will inflate the apparent diversity of microbes in a sample. Thus, they must be eliminated. A major concern in 454 pyrosequencing is the correct determination of homopolymers from flow values [69]. ‘Noise flow values’, defined as light signals that are weak and unrelated to base addition, need removal (see public program: http://blog.mald.e.org/index.php/flowsim/).

The filtered reads are used in the second step that generates taxonomy and related microbial abundance information by comparison to 16S rRNA sequence databases or by using a computer program that reads operational taxonomy units (OTUs). OTUs take the place of ‘Genus and species’ in analyses of microbiomes. Named ‘Genus and species’ present in genomes may not match these exact marker sequences. Mothur [70], Greengenes [71] and other software programs facilitate in depth (i.e. Genus and species) OTU-related assignments and are being improved constantly. The assignment of sequences to OTUs is referred to as binning, and acceptance of the assignment from a 16S rRNA database often requires sequence similarity levels of 97 or 99% [72].

The third step is creation of qualitative or quantitative representations of community-related microbes. Different phylogenetic tree formats (e.g. jackknife node graphics) or principal components analytical plots show bacterial populations within or across environments [73, 74]. The abundance of OTUs is often depicted using heat maps, histograms, and pie charts (fig. 4a–c). Descriptive tools used to characterize a microbiome are often derived from the field of ecology [75–78]. These measures of bacterial populations in a habitat include rarefaction curves.
and graphic displays of alpha- and beta-diversity (fig. 4d–f). QIIME and UniFrac, respectively [79, 80], are online software programs used frequently after pyrosequencing to distinguish bacterial communities in clinical samples. QIIME generates jackknife estimators of abundance data and two- or three-dimensional principal components analytical plots (e.g. showing clusters of OTUs that are similar in an individual or a habitat). These illustrations often appear in publications that study a clinical microbiome. Examples of graphic displays produced by QIIME are present in a report of bacteria found in neonatal intensive care units [9]. UniFrac-related graphics have been used to report distortions in the gut-related microbiota rather than emergence of pathogens before the onset of late-onset sepsis in hospitalized neonates [81]. Software programs used in analyses of microbiomes can be found at the HMP website (http://hmpdacc.org/tools_protocols/tools_protocols.php).

Fig. 4. Presentation of next generation sequencing results. a A heat map that examines the proportion of bacterial taxa in individuals; a color key defines the percentage of specific OTUs to an individual’s total microbiome. b A histogram identifies different OTUs on the x-axis and abundance of a specific OTU is quantified on the y-axis. c A pie chart has a color-coded key to the bacterial taxa present in an environmental sample and the size of the slice corresponds to a percentage of taxa present in the total microbiome. The black slice represents unknown bacteria whose sequence similarity was less than 97% compared to known sequences in a 16S rRNA database. d–f Illustrations of biodiversity in environmental samples from three different subjects. The rarefaction curve (d) is an ecologic assessment tool; the technique assesses species richness in environmental specimens [78]. A steep slope without a plateau means only a fraction of species have been discovered to date, while a flattened curve indicates reasonable sampling and that additional interrogation of the habitat will yield only few species. Another ecologic tool is the Shannon Index (e). This statistical calculation measures alpha-diversity which is a reflection of the numbers of species in an environment and their relative abundances, often designated richness and evenness, respectively. Beta-diversity (f) is a comparison of the amount of species change between ecosystems or environments (e.g. oropharynx vs. intestine). Beta-diversity is influenced by the turnover of species in those habitats. Beta-diversity distance matrices can be computed and principal components analytical plots can be generated; a simple form of that illustration is shown (f). QIIME creates rarefaction curves and graphic displays of alpha- and beta-diversity (http://qiime.org/tutorials/tutorial.html). Figures modified and reproduced with author permission: Morgan XC, Huttenhower C: Chapter 12: Human microbiome analysis. PLoS Comput Biol 2012;8:e1002808.
Finally, 16S rRNA-based classification of bacteria does not indicate whether a microbe is viable and capable of either invasion or creation of metabolic effects in the host. By annotating the microbial analysis with information about the metatranscriptome, metaproteome, and metabolome in the habitat, one gains insight into activities of living microbiota. But, such analyses create significant computational challenges [50, 82]. That topic alone would require a review that surpasses the magnitude of this one.

Conclusions

The metadata created during analyses of an environmental microbiome results in visual formats that are unfamiliar to most clinicians. This review will hopefully assist neonatal caregivers to decipher current and future studies involving microbiomes. Metagenomics will continue to enhance our understanding of how late-onset neonatal sepsis and necrotizing enterocolitis occur, but large gaps in our knowledge still remain. For example, the virome is absent in the intestine at birth, but thereafter this organ acquires bacteriophages rapidly. Bacteriophages probably have a vital role in gut-related bacterial colonization, the emergence of virulence in intestinal bacteria, and even the prevention or therapy of enteric infections [83–85]. By sequentially studying the emerging microbiome in different neonatal organs and by appreciating the functional changes induced in infants by their resident microbiota [86], researchers can utilize metagenomics to advance neonatal health care.

Acknowledgments

This work was supported by Gerber Foundation grant PN002-1214-2708 and NIH grant HD057744 to M.P. Sherman and an Alfred P. Sloan Foundation award to S.T. Kelley.

Disclosure Statement

M.P. Sherman, J. Minnerly, W. Curtiss, S. Rangwala and S.T. Kelley have no financial or other conflicts of interest to disclose.

Appendix: Glossary

16S rRNA: The smaller subunit component of the prokaryotic ribosome and used as the most common taxonomic marker for microbial communities.

Adapter: A short oligonucleotide of known sequence that is typically attached to longer nucleic acids. In next generation sequencing, an adapter provides a priming site for both amplification and sequencing of the adjoining, unknown nucleic acid.

Amplicons: Segments of DNA that are products of natural or artificial amplification.

Binning: Clustering sequences based on their nucleotide composition or similarity to a reference database; used to make an assignment to a bacterial classification system.

Biodiversity: This is an ecologic measure of the complexity of a community. Alpha-diversity is the number (richness) and distribution (evenness) of taxa expected within a single population. Beta-diversity measures absolute or relative overlap of how many taxa are shared between populations. High biodiversity occurs when many taxa (high richness) are present at similar abundances (an even distribution).

Bridge Amplification and Clustering: Pre-sequencing fragments are immobilized on a flow cell via hybridization involving two adapters attached at opposing ends of the fragment and slide-anchored primers with a sequence complementary to the adapters. Using standard PCR reagents, thermal cycling in the ‘primer lawn’ produces many copies of the original fragment which localize in a tight cluster. This technology is used in Illumina genome analyzers.

Emulsion Polymerase Chain Reaction (emPCR): A method used for bead-based amplification of a library. A single adapter-bound fragment is attached to a bead and emulsified in oil containing amplification reagents; they surround a bead/fragment. Parallel amplification of single-strand fragments on millions of beads produces many amplicons for a sequencer-ready library. This method is used in 454 Life Sciences and Ion Torrent genome analyzers.

Jackknife Estimator: Jackknife node trees are depictions that estimate the uncertainty in PCoA plots and hierarchical clustering of microbial communities. The concepts are relevant to beta-diversity of habitats. Graphic displays are an output of QIIME software.

Microbiome: The total microbial community and its biomolecules within a defined environment.

Microbiota: The total collection of microbial organisms within a community, typically used in reference to an animal host.

Multiplexing: Pooling of multiple adapter-barcoded libraries into one sequencing run.

Operational Taxonomic Unit (OU): This is a terminal node designation in phylogenetic analysis of microbiota. OU is akin to ‘Genus and species’ in classic microbiology and typically uses rDNA and a percent similarity threshold for classifying microbes.

Principal Coordinates Analysis (PCoA): PCoA is a multidimensional scaling method to explore and to visualize similarities or dissimilarities of data, such as OTUs within an environment. In two- or three-dimensional plots, the matrix distances of data can be displayed. This graphic analysis is supported by QIIME software.

Pyrosequencing: A method of DNA sequencing detects the release of pyrophosphate upon nucleotide incorporation. This is the chemistry used in 454 Life Sciences genome analyzers.

QIIME (Quantitative Insights into Microbial Ecology): Public software that processes data from high-throughput 16S rRNA sequencing studies. The purpose is to provide a start-to-finish workflow that starts with multiplexed sequence reads and finishes with taxonomic and phylogenetic profiles and comparisons either in a sample or among specimens (i.e. environments).
Rarefaction Curve: A plot in which the horizontal axis represents samples (often DNA sequences) and the vertical axis represents diversity (e.g. number of distinct taxa).

Read: The primary output of DNA sequencing, consisting of a short stretch of DNA that is defined from sequencing as a region of a single DNA fragment.

Taxonomy: Hierarchical classification of life into three domains, namely eukaryotes, bacteria and Archaea. Metagenomics deals mainly with the classification of bacteria, but can include the viruses and eukaryotic microbes. The bacteria are placed in the largest to smallest groupings starting with Phylum, Class, Order, Family, Genus, and species, respectively.

UniFrac: A method developed to calculate a distance measure between microbial communities using phylogenetic information. A weighted version of the UniFrac metric assigns relative abundance of each of the taxa within the communities. Data is appropriately ‘binned’ into operational taxonomic units which can then be dealt with as taxa within the UniFrac framework.

Virome: The genomes of all the viruses that inhabit a particular organism or environment.

References


35 Sanger F, Donelson JE, Coulson AR, Kössel H, Fischer D: Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. Proc Natl Acad Sci USA 1973;70:1209–1213.


