Use of Whole Genome Shotgun Metagenomics: A Practical Guide for the Microbiome-Minded Physician Scientist

Jun Ma, PhD[1,2] Amanda Prince, PhD[1] Kjersti M. Aagaard, MD, PhD[1,2,3,4]

[1] Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Baylor College of Medicine
[2] Department of Molecular and Human Genetics, Bioinformatics Research Lab, Baylor College of Medicine
[3] Department of Molecular and Cell Biology, Baylor College of Medicine
[4] Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, Texas

Semin Reprod Med 2014;32:5–13

Address for correspondence Kjersti M. Aagaard, MD, PhD, Department of Obstetrics & Gynecology, Division of Maternal-Fetal Medicine and Department of Molecular and Cell Biology, Department of Molecular Physiology and Biophysics, National School for Tropical Medicine, Center for Metagenomics and Microbiome Research, Center for Reproductive Medicine, John M. Eisenberg Center for Health Outcomes Research, Bionformatics Research Lab at the HGSC, Translational Biology and Molecular Medicine and Co-Director, Baylor College of Medicine MSTP program, 1 Baylor Plaza, Houston, TX 77030 (e-mail: aagaardt@bcm.edu).

Abstract

Whole genome shotgun sequencing (WGS) has been increasingly recognized as the most comprehensive and robust approach for metagenomics research. When compared with 16S-based metagenomics, it offers the advantage of identification of species level taxonomy and the estimation of metabolic pathway activities from human and environmental samples. Several large-scale metagenomic projects have been recently conducted or are currently underway utilizing WGS. With the generation of vast amounts of data, the bioinformatics and computational analysis of WGS results become vital for the success of a metagenomics study. However, each step in the WGS data analysis, including metagenome assembly, gene prediction, taxonomy identification, function annotation, and pathway analysis, is complicated by the shear amount of data. Algorithms and tools have been developed specifically to handle WGS-generated metagenomics data with the hope of reducing the requirement on computational time and storage space. Here, we present an overview of the current state of metagenomics through WGS sequencing, challenges frequently encountered, and up-to-date solutions. Several applications that are uniquely applicable to microbiome studies in reproductive and perinatal medicine are also discussed.

Keywords
- whole genome sequencing
- microbiome
- assembly
- metabolic pathways

The relationship of humans with our environmental microbes is documented throughout history. The discovery of the smallpox vaccine by Edward Jenner and the great pandemics of the Bubonic Plague and the 1918 influenza have demonstrated the volatile and parasitic side of microbes. However, we also have mutualistic and commensal relationships with the microbes in our environment. Recently, the Human Microbiome Project (HMP) consortium took on the task of documenting what constitutes a healthy microbiome.[1–4] This question has help to highlight studies demonstrating that dysbiosis of the microbiome is associated with type 2 diabetes mellitus, obesity, inflammatory bowel disease, and colorectal cancer.[5–8] Further, dysbiosis of the microbiome has been implicated as a cause for preterm birth. Gravidae that undergo preterm birth often have an intrauterine infection with increases in inflammatory cytokines,[9,10] such as IL-6 and IL-1β. Thus, these studies have begged the question of how do we establish and maintain a healthy microbiome. With the exponentially expanding interest in human microbiome research, a working knowledge of the methodology and tools used in this field is fundamental to translational research. This is notably true in reproductive and perinatal research initiatives, where there is a tremendous potential need for investigators well versed in both the technology and biology of the

Issue Theme The Microbiome and Reproduction; Guest Editors, James H. Segars, MD, and Kjersti M. Aagaard, MD, PhD

DOI http://dx.doi.org/10.1055/s-0033-1361817.
ISSN 1526-8004.

expanding field of research. While an increasing number of investigators are familiar with and employing 16S-based metagenomic approaches, there are far fewer investigators who have a working knowledge of alternative metagenomic approaches.

Before the era of massively parallel NextGen sequencing, the clone-based metagenome approach in combination with Sanger sequencing was used for early metagenomics research. First, DNA content of a genomic clone is sheared into random fragments before cloning fragments into plasmid vectors that are grown to produce monoclonal libraries containing enough genomic material for sequencing. Although Sanger sequencing produces long reads (100–2,000 bp), usually only a few selected inserts could be obtained. Thus, this process is low throughput and suffers from the limitation of assembly regions with large repeats and cloning bias. Hence, it is not surprising that NextGen sequencing (NGS) techniques have quickly replaced Sanger sequencing because of their collectively unique advantages. In addition to economical low per base cost and higher throughput, the cloning step and its inherent problems seen in Sanger sequencing methods are no longer issues for NGS techniques. Environmental samples can be sequenced directly by NGS techniques, which allows for the investigation of unculturable and low abundance species. Therefore, the comprehensive characterization of more complex and diverse microbial communities, such as microbial communities related to the human reproductive system, become feasible. NGS techniques used in metagenomics research mainly include 454 Genome Sequencer Pyrosequencing (454 Life Sciences; Roche Company, Branford, CT) for 16S rDNA sequencing, Solexa/Illumina (Illumina Inc., San Diego, CA) for whole genome shotgun sequencing (WGS) studies, and the most recent Helicos (Helicos Bio Sciences, Cambridge, MA) single-molecule sequencing technology also for WGS studies.[11]

The majority of recent studies examining the bacterial flora communities residing within humans have utilized 16S rDNA sequencing techniques. The nine variable regions of the 16s rRNA gene are flanked by conserved stretches in the majority of bacteria. This conservation can be used as targets for PCR primers with near-universal bacterial specificity. Therefore, through 454 Pyrosequencing, sequences read are obtained from one region of the 16S rRNA gene, which is then quantified and subsequently assigned a taxonomy. Thus, when compared with WGS techniques of the full length 16S rRNA gene, the coverage of each sample is much higher and many more samples are able to be run in parallel using bar-coding system. However, the downside to WGS techniques is that a small proportion of reads could be assigned to lower level taxonomy due to the shorter read length. Overall, the resolution of the community composition obtained with 16S Pyrosequencing techniques is orders of magnitude larger than Sanger sequencing with a lower per base cost.

The Illumina technology was introduced around the same time as 454 Pyrosequencing technology. The Illumina instruments produce more than 10 times the number of reads per run as the 454 GS FLX machines, albeit of much shorter lengths (less than 100 bp compared with 400–500 bp of 454 reads). The advantages of WGS sequencing on Illumina platform over 16S rDNA sequencing on 454 platform are the ability to provide information on genome assembly, species level taxonomy abundance, gene predication, and metabolic pathway reconstruction.[12] However, each stage of the analysis is complicated by incomplete coverage, the high volume of data, the short length of reads, and intrinsic errors caused by parallelism sequencing.[13,14] In this review, we will primarily focus on the bioinformatics procedure to transform Illumina-generated short reads into biologically meaningful taxonomic and functional entities (►**Fig. 1**). Recent developed tools specific for metagenomic data analysis and their application to human reproductive medicine will be discussed as well (►**Table 1**).

Genome Assembly

Genome assembly is essential for the study of gene arrangements and gene function. For assembly in a single organism, all the DNA fragments come from the same genome. However, this is not the case when it comes to metagenome assembly, and several obstacles make metagenome assembly especially challenging. For samples from an environment with low microbial abundance, the coverage on the genome is usually incomplete. Although longer gene sequences could be achieved, there is still a risk of making chimeric contigs from different operational taxonomic units (OTUs). This risk of chimeric contigs is further complicated by genomic repeats.

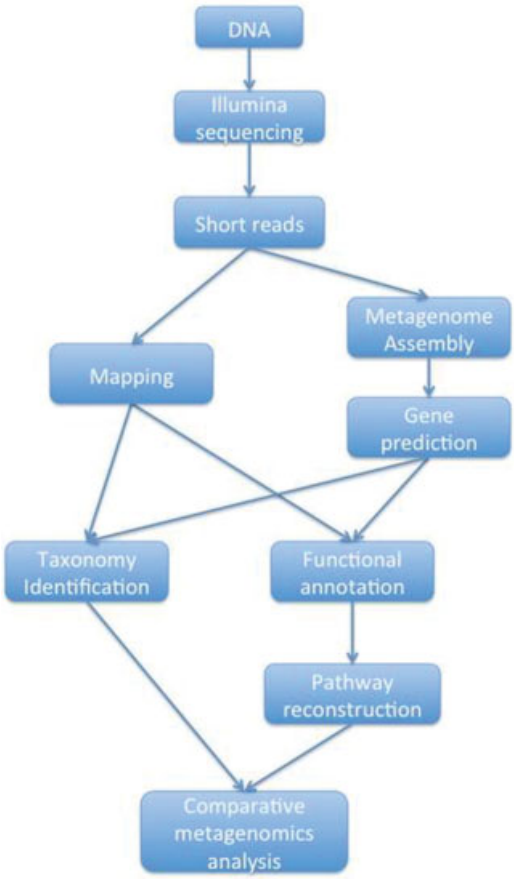

Figure 1 Flowchart of metagenomics whole genome shotgun sequencing data analysis.

Table 1 Tools for metagenomics analysis mentioned in this review

Process	Tools	Reference
Metagenome assembly	MetaVelvet Meta-IDBA GeneStitch Ray Meta Bambus 2	20 21 22 23 25
Gene prediction	MetaGene Annotator MetaGeneMark Orphelia	26 28 29
Taxonomic identification	MEGAN WebCARMA Phymm PhymmBL MetaPhyler MetaPhlAn	31 33 35 36 37
Functional and pathway annotation	BLASTX UBLAST HMMER THINK-Back HUMAnN	 40 41 42 45
Data analysis pipeline	MetAMOS MG-RAST Genboree tool kit	58 55 54

For the same reason, the assembly process could distort the species abundance as well.[15]

Reconstructing genomes without referencing a previously sequenced genome is called de novo assembly, which is proven to be hard to solve computationally (NP-hard). The traditional method used for assembly of Sanger-based sequences is Overlap Layout Consensus. An overlap graph is first constructed with each read as a node and edge representing the overlap identified between reads. The graph is thereafter analyzed to determine the paths connecting reads together to construct the genome. However, this method is not suitable to be used on the assembly of short reads generated on an NGS platform because in the worst case each read must be compared with all other reads. NGS methods usually generate an order of magnitude more reads compared with Sanger sequencing, which significantly increases the computational complexity. Most of the recently developed metagenomics assembly algorithms are based on Eulerian tour of de Bruijn graphs. In de Bruijn graphs, reads are first decomposed into fixed length k-mers. Nodes are represented by k-mers, with the reads themselves being the edges connecting the nodes (►**Fig. 2**). The overlaps are implicitly represented in the graph by paths that traverse from one read to its neighbor. The output is usually a simple path of contigs. In this way, the high number of reads does not affect the number of nodes and because repeats only appear once, the problem of high redundancy in reads is also solved. Moreover, the solution to a de Bruijn graph is an Eulerian path, and a linear-time algorithm to solve an Eulerian path does exist.

Traditional assemblers designed for the assembly of single organism genomes were initially applied to assemble metagenomes (i.e., Velvet,[16] Celera,[17] and SOAPdenovo[18]) with limited success. Recently, various Eulerian strategy-based assemblers have been developed specifically for the assembly of metagenomes.[19]

One assembler in particular is MetaVelvet[20] that was developed based on Velvet,[16] a popular assembler for single genomes. The basic idea of MetaVelvet is to take the de Bruijn graph constructed from sequences obtained from multiple species as a mixture of multiple de Bruijn subgraphs, where each subgraph represents an individual species. The mixed de Bruijn graph is then decomposed into individual subgraphs based on coverage difference and graph connectivity, and the subgraphs are subsequently used for building scaffolds.

The program Meta-IDBA[21] is used to address the issue of identifying the branches in the de Bruijn graph caused by polymorphism in similar subspecies (sp-branches) or caused by similar genomic regions shared by different species (cr-branches). Meta-IDBA first identifies and removes cr-branches in the graph, which leaves connected components corresponding to a set of subgraphs of the same species. Finally, each component is transformed into a multiple alignment of consensus sequences to represent the contigs of different subspecies.

GeneStitch[22] uses the prior knowledge of the species composition and gene contents to guide the assembly process. The idea is that the assembled contigs are similar to given reference genes. Alternatively, the contigs could be inferred from the tangled de Bruijn graph using a network matching algorithm, and if no prior sequence knowledge is available, a general dataset of genes could be used to reference the gene sets as well. With the ever increasing number of samples and reads, scalability is becoming important for metagenome assembly.

Ray Meta[23] is a method that was developed for scalable distributed de novo metagenome assembly on Ray.[24] Ray Meta does not modify the de Bruijn subgraphs as MetaVelvet and Meta-IDBA. It applies heuristics-guided graph traversals on k-mers in parallel, which is more amenable to distributed

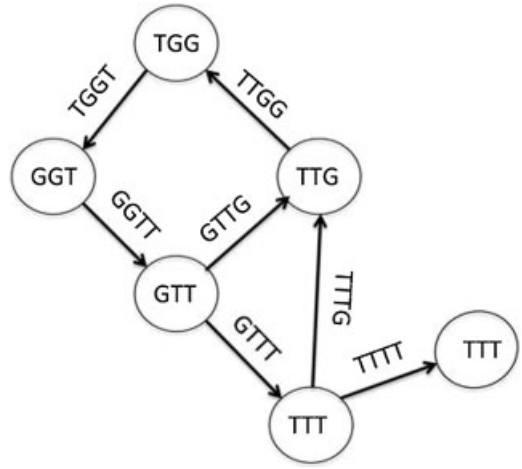

Figure 2 A sample de Bruijn graph with $k = 4$. The edges of the graph are unique subsequences of reads with length of k. The nodes of the graph represent common subsequence of length of k-1. If the suffix of one node matches with the prefix of the other node with length of k-2, the two nodes are connected. This graph consists of short reads for the consensus sequence "GTTTGGTTGT."

computing. All of these methods are claimed to yield longer contigs and more representative taxonomic representations on simulated and real data compared with assemblers designed for single genome assembly.

After the reads are assembled into contigs, the relative positions of the contigs along a genome are determined by scaffolding, a process that depends on mate-pair information. If two ends of the mate pair are in differing contigs, the two contigs are inferred to be adjacent to each other on the genome. Most of the assemblers contain module of scaffolding. This scaffolding algorithm starts with the most reliable information and gradually adds more data as long as the new information agrees with the constructed scaffold. Tools, such as Bambus 2,[25] have been developed for metagenome scaffolding. Bambus 2 can be applied to virtually all existing sequencing technologies and the output from popular assemblers.

Gene Prediction

A fundamental purpose of gene assembly is to enable gene predictions using scaffolds, such that genes can be classified into correct functions. There are two classically described methods for gene prediction: (1) train model parameters on known annotations to predict unknown annotation or (2) to train models based on homology search, which aligns sequences to gene database to find homologous sequences. However, it is not possible to apply these traditional methods directly to metagenomics data. The incomplete open reading frame (ORF) acquired from metagenome assembly often lacks start or stop codons; therefore, ab-initio programs do not work in this scenario. In addition, there is not yet a sufficient metagenomics sequence database to build a statistical model to distinguish coding from noncoding ORFs. The obvious drawback for these homology-based methods is that it only provides information for known genes.

Recent tools have been developed to address this core issue of metagenome gene prediction. MetaGene Annotator (MGA) [26] is upgraded from MetaGene.[27] First, all ORFs in MGA are extracted and scored on a model estimated from annotated genomes. Then, an optimal combination of ORFs is calculated using the scores of orientations and distances of neighboring ORFs. MGA also uses the logistic regression models of the GC content and di-codon frequencies from MetaGene. In addition, it has an adaptable ribosomal binding site model based on complementary sequences of 16S ribosomal RNA, which helps MGA to precisely predict translation start sites.

MetaGeneMark[28] uses a heuristic approach originally developed for finding genes within small fragments of anonymous prokaryotic genomes and/or highly inhomogeneous genomes. The training dataset consists of 357 bacteria and *Archaea* species. Linear regression is applied to the relevant information in the training set, such as the relationship between positional nucleotide frequencies and the global nucleotide frequencies and the relationship between the amino acid frequencies and the global GC content. The initial frequency values of the occurrence of 61 codons are calculated based on the above information and subsequently modified by the frequency of each amino acid determined by the GC content. Finally, the Markov model of a protein coding region is constructed based on the usage of all 61 codons.

Orphelia is the third recently developed tool for gene prediction,[29] and is unique in that it adopts a neural network-based method. The neural network is trained on randomly excised DNA fragments from the genomes that were used for discriminate training. The artificial neural network combines sequence features, such as monocodon usage, dicodon usage, and translational initiation sites, with ORF length and GC content to compute a posterior probability of an ORF to encode a protein.

One recent study has benchmarked these three gene prediction methods and demonstrated variable performance at different read lengths and fragment types.[26] As might be logically predicted, the authors found that longer reads result in better gene prediction. In addition, while MGA had the best sensitivity, it was the worst in specificity for most read lengths. MetaGeneMark had average sensitivity but much better specificity than MGA, and Orphelia had the lowest annotation error for longer read lengths. Therefore, the combination of several methods, screening intergenic regions for overlooked genes, and using dedicated frameshift detectors may result in better prediction accuracy.[26] Decisions as to which will perform optimally will also be dependent upon the number of reads per sample and the ratio of bacterial to human reads. This is similarly related to the human body niche of sample origin.

Taxonomy Identification

One important question to answer in metagenomics analysis is "What microbes are present?" which leads to the identification of taxonomy distribution in metagenomic samples. 16S rDNA-based surveys produce on average 10,000 sequences that range from 400 to700 bp in length per sample. Rapid taxonomic classifiers, such as the Ribosomal Database Project (RDP) classifier,[30] use these sequences to generate taxonomic distribution down to the genus level. Despite its popularity, 16S rDNA-based methods suffer from the biased estimation of microbial diversity due to the variability in copy number of the 16S gene and the PCR.

There are two key pathways enabling taxonomic identification using WGS reads. The first employs homology search against a reference gene database. For example, MEGAN[31] first performs a BLASTX search against the NCBI-NR database. Taxonomic analysis is then conducted by placing each read onto a node of the NCBI taxonomy according to the lowest common ancestor of the top hits, and the NCBI taxonomy is based on a hierarchically structured classification of all species represented in the NCBI. Instead, CARMA,[32] and the refined version WebCARMA,[33] searches all Pfam domain and protein families as phylogenetic markers to identify the source organisms of unassembled reads using hidden Markov models. Then a phylogenetic tree is reconstructed for each matching Pfam family and the corresponding query reads. Finally, the reads are classified into a higher-order taxonomy depending on their phylogenetic relationships to family members with known taxonomic affiliations. It is worth noting that only a small portion of reads have matches by BLAST against the microbial database.

An alternative strategy to homology-based approaches is to use machine learning and statistical methods to classify reads based on the composition of the DNA base signatures. The interpolated Markov models (IMMs) have been employed with success in bacterial gene classification using the GLIMMER system.[34] Compared with other methods, IMMs utilize information from sequences of different lengths and integrate the results. The program Phymm[35] demonstrates the use of IMMs in classification of metagenomic reads. In Phymm, a classifier is trained on a large amount of curated genomes. This classifier constructs probability distributions that represent the observed patterns of nucleotides characterizing each chromosome or plasmid. PhymmBL[35] demonstrates that the combination of machine learning and BLAST produces higher accuracy than either method alone.

Given the complexity of metagenomic assemblies, the taxonomic classification can also be achieved by directly using reads before assembly. Large-scale studies, for example, the HMP,[1–3] likely includes hundreds of samples. In these large-scale studies, the computational efficiency of BLAST becomes the bottleneck for the analysis process if all the reads are used for classification without assembly. Therefore, reference marker gene sets are constructed to reduce the size of the database. MetaPhyler[36] is one of these methods, which relies on 31 phylogenetic marker genes derived from existing genomes and the NCBI-NR database. Furthermore, instead of using a universal classification threshold for all genes at all taxonomic levels, MetaPhyler uses different thresholds for classifiers to the reference gene and to the taxonomic level, which results in much faster analysis. MetaPhlAn[37] first identified more than 2 million potential markers using 2,887 genomes from Integrated Microbial Genomes (IMG) system,[38] which was further refined to a catalog of 1,221 species with 231 markers per species and $>$ 115,000 markers at higher taxonomic levels. The relative abundance of each taxonomic level is made by the alignment of reads to clade-specific marker sequences in this catalog. Microbial clade abundance is then estimated by normalizing read-based counts with the average genome size of each clade. MetaPhlAn has been applied to the analysis of vaginal microbiome (posterior fornix) of asymptotic women enrolled in the HMP. As *Lactobacillus* is the dominant genus in the vaginal microbiome, it is important to further classify this genus down to the species level to reflect the detailed microbial profile. Using this strategy, all five of the signature *Lactobacillus* species could be identified by MetaPhlAn. Despite the technical difference between 16S sequencing and WGS, the estimated relative abundance is quite similar.

Functional and Pathway Annotation

After discovering the microbial consistency, the next question to be answered from WGS data are "What are these microbes able to do?". There are two issues involved in this process. The first issue is to assign functional annotation to the assembled ORF or to the reads directly. The other issue is to place the genes in the context of biological pathways, especially metabolic pathways.[39] The most straightforward way for functional prediction is by aligning query sequences to an existing reference protein database, but then one must determine which database to use. The size and contents of the databases are different, which will in turn affect the efficiency and accuracy. If one is interested in annotating as many sequences as possible, the NCBI RefSeq database would be a good choice because it has the most comprehensive collections of genomes. For this purpose, various versions of BLAST, including BLASTX and BLASTP, could be applied. However, this approach suffers from the long computation time required to search through all the homologs in reference to the database for each sequence in the dataset. To speed up the process, BLAST can be done in parallel, like the MBLASTX (MulticoreWare, St. Louis, MO) used by the HMP. Tools, such as UBLAST,[40] have been developed for high-throughput sequence classifications that are often an order of magnitude faster than BLAST in real applications, but these applications lose sensitivity. The raw quantification obtained from alignments needs to be normalized by the size of the reference coding sequences. The results from a homology search are often affected by sequence conservation due to the functional homology in different organisms. When sequences are mapped to structurally or functionally conserved region, they can easily be assigned to different species if only a similarity score is used.

A possible solution to misclassification is to adopt the more sensitive profile-based search method. This method uses databases with profiles generated from alignments of protein families that share similar functions, such as COG, Pfam, or TIGRfam. Hidden Markov–based HMMER[41] was designed to perform a fast search against profiles generated from multiple sequence alignments. Although more sensitivity is achieved this way, fewer sequences get annotated. For partial proteins generated on short contigs or unassembled sequences, a repository with patterns or motifs (i.e., PROSITE) might be used for a functionality search. If gene prediction is successful, genomic neighborhood, phylogenetic profiling, and gene coexpression analysis may provide useful information for functional prediction as well.

Pathway-based analysis has been developed to interpret the results from microarray experiments before applying the results to metagenomics data. Pathway here indicates a series connected sets of genes with nodes representing genes and lines representing their relationships (►**Fig. 3**). The significance of these pathways is decided by functional enrichment statistics (Fisher exact test) or by scoring based on the pool of genes in the sample (gene set enrichment analysis [GSEA]). One major drawback of these count-based methods is disregarding the topology of the pathways. The order of the genes in the pathway could help with the interpretations of the results. Fortunately, more complex methods have been developed to address this problem. THINK-Back,[42] stands for Knowledge-based Interpretation of High Throughput data, is a suite of tools trying to generate biologically meaningful hypothesis by using knowledge in pathway databases, such as KEGG, PANTHER, Reactome, and Biocarta. One method in THINK-Back adjusts the score generated by GSEA[43] by incorporating the appearance frequency of the genes in a KEGG database.[44] Another method takes into account the topology

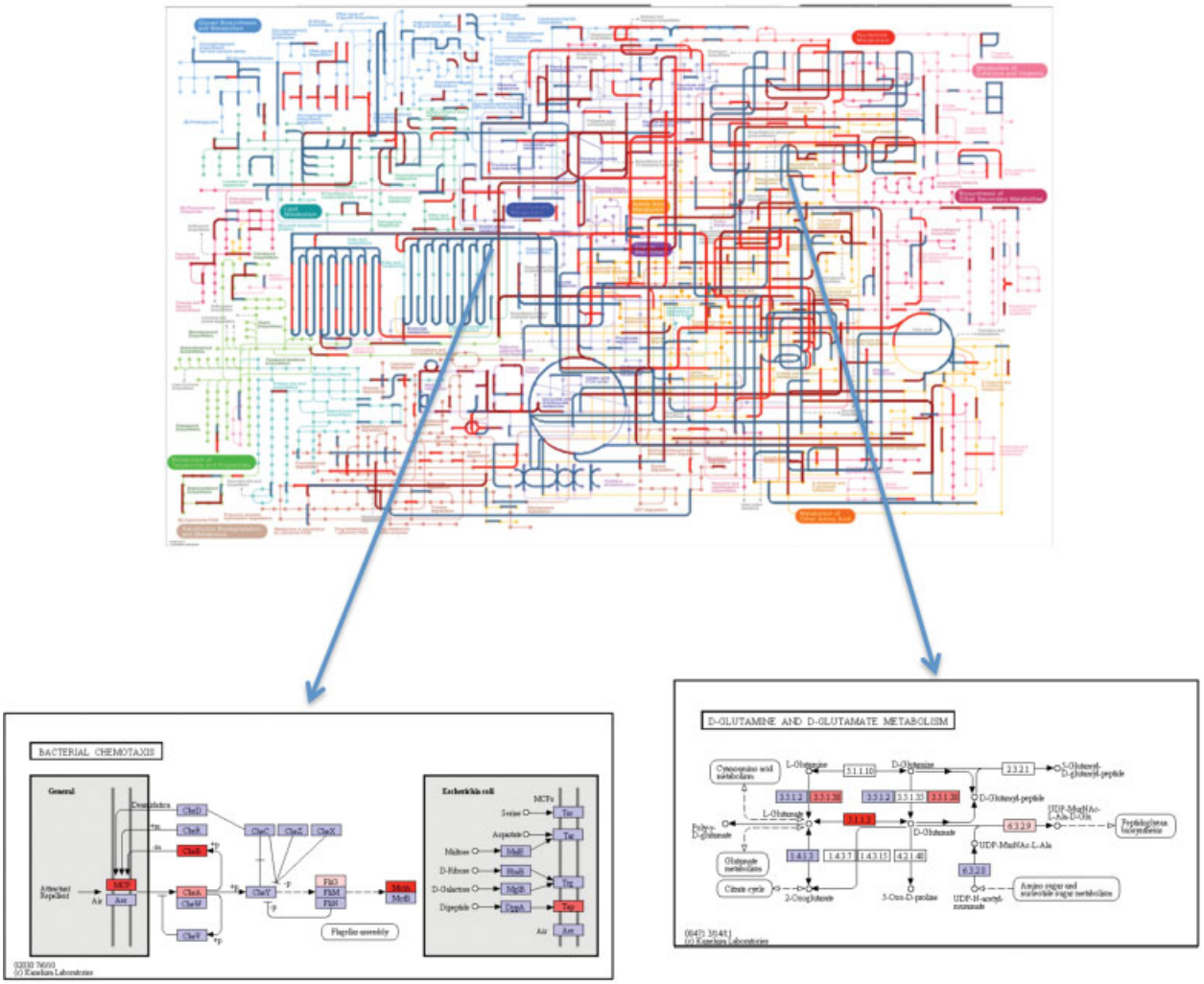

Figure 3 Mock KEGG pathway map shows the concept of pathway analysis. The figure on the top contains all the KEGG pathways involved in metabolic process. Blue and red colors indicate the enrichment of genes in either case or control group. Two examples of KEGG pathways are shown with nodes representing proteins or molecules and lines representing their biological relationship. The gradient of red indicates the average relative ratio of gene abundance between case and control samples.

of the pathways to calculate a density score, which is subsequently used for adjusting GSEA scores.[42] The pathway reconstruction using WGS data is essentially using the number of gene copies to indicate the activity of pathway, which is quite different from RNA-based microarray and RNA-seq analysis. But the idea of integrating topology pathway information into pathway analysis could still be applied to metagenomics pathway reconstruction.

As described above, most of the ORFs assembled from metagenomics reads are partial and likely contain errors caused by frame shifts; therefore, another way to perform functional annotation is to skip the gene calling altogether and use the protein coding sequences identified from the reads. In HUMAnN (the HMP Unified Metabolic Analysis Network),[45] the reconstruction of a network is accomplished by mapping the protein coding genes onto reference pathway collections, such as eggnog and KEGG orthology groups, based on their homology to the reference genes previously characterized. MinPath[46] adopts an integer programming algorithm to reconstruct "minimal pathway," which is defined as a list of functions annotated for a set of genes in a minimal pathway that includes all the gene functions. This approach avoids the problem of identification of spurious pathways and overestimation of microbial abundance. After data normalization and smoothing, pathway coverage (relative confidence of each pathway being present in the sample) and pathway abundance (relative "copy number" of each pathway in the sample) are generated and organized into a matrix-like format for postprocessing.

Comparative Metagenomics

Despite all the challenges with WGS as covered in this review, important environmental and biological questions have been investigated through comparisons of taxonomic abundance and metabolic pathway activity. Because of the dynamic nature of the microbiome, there is large variation in microbiome profiles even from samples obtained from a similar environment. Therefore, a higher number of samples need to be collected to get an accurate measure of the microbiome.

Ergo, in addition to a large amount of sequences, there are also a large number of samples with metadata. The taxonomy profile is often organized into a matrix with rows representing taxonomy (either relative abundance for each taxonomic level or OTU counts) and columns representing each sample. Depending on the complexity of the microbiome, there could be thousands of rows and columns in the matrix.

Matrix expansion yields the issue of data dimension. Dimensionality reduction becomes important to decrease the computational cost. If the taxa table is large, it will be helpful to first filter the table to remove singletons or OTUs/species only appearing in a small number of samples. As singletons or rare species may be generated by sequencing error, they are not helpful for the purpose of comparative metagenomics. Principal coordinates analysis (PCoA) is the most popular technique to perform dimensionality reduction. PCoA takes the results of β diversity comparisons that are generated using phylogenetic (UniFrac[47]) or nonphylogenetic (such as Canberra) distance on taxon table and produces a new matrix with fewer dimensions by solving eigenvalues. The direction of each axis is chosen to maximize the variation in the data. Normally, the first three coordinates are chosen to visualize the samples in three dimensions. The points that cluster together indicate that these samples have similar taxonomy profiles. An alternative, nonparametric rank-based method is to use nonmetric multidimensional scaling, which could avoid the arch effect caused by the sparsity of the matrix.[48]

If clusters are observed from a 3D PCoA plot, most biologists will be interested to know which taxa cause the differences in the microbial community depending on the metadata. Thus, some machine learning techniques could be used to answer this question. Not every statistical test should be used for every analysis, but the combination of several analyses can produce a more accurate result. The Random Forest generates a large ensemble of decision trees from a random subset of the data and a random selection of the variable. The resulting ensemble of trees is then used with a majority-voting approach to decide which sample belongs to which group. One advantage of the Random Forest is that there is no need for cross validation to get an unbiased estimate of the test set error. An out-of-bag error estimate is generated internally by a bootstrap sample from the original data. This is very useful when the sample size is small. Boruta is an all-relevant feature selection wrapper algorithm around Random Forest. It finds important features by iterative learning of the Random Forest classifier. In the end, a list of features confirmed to differentiate groups is generated.[4] LEfSe (linear discriminate analysis effect size)[49] is a recently developed tool to identify genomic features (genes, pathways, or taxa) specific to each group. LEfSe first use Kruskal–Wallis sum-rank test and Wilcoxon rank-sum test to identify the significant differential abundance with respect to the class of interest. Then linear discriminant analysis is applied to estimate the effect size of each differentially abundant feature. LEfSe also provides bar plots and cladogram plots to represent the discovered biomarkers.

Random Forest and linear discriminate analysis are both supervised machine learning methods, which means that the samples have been assigned to a group before the learning task. Unsupervised learning has been applied to metagenomics data as well to discover the hidden structures of microbiome. One effort is the introduction of the "enterotype" by Arumugam et al[50] using the human gut samples from the MetaHIT consortium.[50] Enterotypes are generated by performing clustering analysis on the gut microbial communities. The difference among three enterotypes is driven by key bacterial genera and not related with age, gender, or body weight. There is also a report about the existence of enterotype-like clusters in the vaginal microbiome community based on the abundance of bacterial species, mainly species in the *Lactobacillus* genus.[51] However, recent research, including our own, indicates that one should take precautions when performing enterotyping.[52,53] Despite various ways to generate a taxa table, clustering is a statistical approach, whose performance is affected by many factors. One recent study tried to identify the influence of various factors on enterotyping, including clustering methodology, distance metrics, OTU-picking methods, sequencing depth, and sequencing methods.[52] Using the HMP data, instead of discrete enterotypes, a smooth gradient distribution of *Bacteroides* abundances was observed in gut microbiome. For the vaginal microbiome, depending on the taxonomy level, distance metrics and scoring methods, two to five clusters are found using the HMP vaginal data. These results suggest that distance metrics and the clustering methods have the largest effect on enterotyping. At least one absolute scoring method combined with two to three distance metrics should be used to verify the existence of enterotypes.[52]

WGS Data Analysis Pipelines

As described above, the metagenomics data analysis includes assembly, gene predication and annotation, taxonomic classification, and so on, but each of these tasks is performed by specific software that requires installation, configuration, and integration. This is a daunting task even for bioinformatics experts. Most of the research groups construct analysis pipeline by picking tools for each task based on their own experience. For a laboratory without bioinformatics support, it may be difficult to perform meaningful analysis with a large amount of data. With this in mind, we have recently worked to produce single-site, publicly available tool sets.[54] Specifically designed for 16S analysis, our Genboree Microbiome tool set was deployed using the web-based Genboree workbench, which has an easy-to-use GUI interface. Users upload the sequencing file and metadata and choose the desired task analysis by clicking. Similar web tools for WGS data analysis have been developed too. MG-RAST[55] is a comprehensive web tool for both phylogenetic and functional summaries. MG-RAST is based on a modified version of the RAST (rapid annotation based on subsystem technology) server[56] upon the SEED framework, which provides automated sequence assignment by comparison with both protein and nucleotide databases. Users can upload the sequence file to the server and keep data private or public.

Like QIIME for 16S-based analysis, a similar standardized framework for WGS data analysis has been created. Smash-Community[57] is one of the early pipelines designed for 454

and Sanger data with limited capability for follow-up analysis. MetAMOS[58] is a modular and customizable framework for metagenomic assembly and analysis, which is also user friendly. The construction is built upon the AMOS open-source genome assembly framework. A collection of publicly available tools is tied together by the lightweight workflow system Ruffus, including Meta-IDBA,[21] MetaVelvet,[20] SOAP-denovo,[18] Bowtie,[59] MetaGeneMark,[28] MetaPhyler,[36] and more. The modular design enables users to check the output for each step and facilitates the integration of data generated by other tools or in different formats.

WGS Application in Human Reproductive Medicine

Despite these advances in WGS analysis, the prevailing technique used for microbiome research in the area of human reproductive medicine is still 16S rDNA sequencing. However, WGS techniques have been adopted in several recent studies in addition to the 16S sequencing. A subset of samples from HMP was subjected to Illumina sequencing, which included samples from posterior fornix.[2] The result from this study demonstrated that although each body site is characterized by signature clade, most of the metabolic pathways are evenly distributed and prevalent across both individuals and body habitats. However, this analysis revealed that the pathways related with oligosaccharide and polyol transport system are more active in posterior fornix samples. One recent study on dynamic changes of gut microbiota from first to third trimesters also used Illumina HiSEq. 2000 to examine the enrichment of specific metabolic pathways.[60] The analysis of data did not find difference in the mean relative abundance of gene categories or metabolic pathways between trimesters. Therefore, the shifts of gut microbiome during pregnancy may not be associated with metabolic changes. However, a network analysis of correlations between COG (cluster of orthologous groups) abundances across samples indicated the loss of network modularity in the third trimester, which indicates a reduction in phylogenetic diversity and a more uneven distribution of taxa.[60] The results of this study are in agreement with studies of phylogenic diversity from our laboratory.[4] As WGS techniques continue to improve and become more user friendly, this will be a powerful tool in future studies with a focus on human reproduction. For instance, in studying preterm birth, it can be challenging to detect bacteria in the amniotic fluid of patients.[61] Yet, recent studies have detected bacteria deep within human fetal membranes.[62] In addition, an independent study found that bacteria were harbored in the basal plate of the placenta.[63] Remarkably, there was not statistical significance in the presence of bacteria between preterm and term patients.[63] Thus, while previous sequencing techniques have failed to detect bacteria in the placenta, the advent on NGS techniques may help to advance our understanding of the role of the microbiome in promoting healthy, term pregnancies.

Conclusion

The rapid advancement of sequencing technology has brought both promise and challenges to the metagenomics field. We can now explore unknown environments as community genomic, ecologic niches in previously unparalleled and dynamic fashions. However, the downstream analysis currently lags behind the sequencing technology. Compared with 16S-based metagenomic sequencing, WGS generates exponentially more sequences that necessitate large storage requirements, and produce large numbers of unknown species that demand more computational resources. In this review, we have introduced several recently developed tools dedicated to metagenomics assembly, gene predication, and pathway reconstruction. There is still a high demand for more efficient and more sensitive tools to perform standardized analysis. In addition to WGS, RNA-based metatranscriptomics is also under development to provide more details on the dynamic changes in the community, which may alleviate the limitation caused by DNA-based methods. Metabolomics attempts to measure the complete set of molecules in the community, which could provide important information on the study of host–microbe interactions. In our era of "omics-based discovery science," physician scientists are increasingly called upon to work side by side with computational scientists. It is our hope that this review will provide our fellow microbiome-minded reproductive and perinatal biologists with a working knowledge of the current state of the science.

References

1 Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;486(7402):215–221
2 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012; 486(7402):207–214
3 Aagaard K, Petrosino J, Keitel W, et al. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J 2013;27(3):1012–1022
4 Aagaard K, Riehle K, Ma J, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PLoS ONE 2012;7(6):e36466
5 Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55–60
6 Joossens M, Huys G, Cnockaert M, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut 2011;60(5):631–637
7 Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe 2008;3(4):213–223
8 Sobhani I, Tap J, Roudot-Thoraval F, et al. Microbial dysbiosis in colorectal cancer (CRC) patients. PLoS ONE 2011;6(1):e16393
9 Goldenberg RL, Hauth JC, Andrews WW. Intrauterine infection and preterm delivery. N Engl J Med 2000;342(20):1500–1507
10 Nold C, Anton L, Brown A, Elovitz M. Inflammation promotes a cytokine response and disrupts the cervical epithelial barrier: a possible mechanism of premature cervical remodeling and preterm birth. Am J Obstet Gynecol 2012;208:e201–e207
11 Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLOS Comput Biol 2010;6(2):e1000667
12 Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. Brief Bioinform 2012;13(6):711–727
13 Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. Curr Opin Biotechnol 2012;23(1):64–71
14 Teeling H, Glöckner FO. Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. Brief Bioinform 2012;13(6):728–742

15 Pop M. Genome assembly reborn: recent computational challenges. Brief Bioinform 2009;10(4):354–366
16 Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18(5):821–829
17 Myers EW, Sutton GG, Delcher AL, et al. A whole-genome assembly of Drosophila. Science 2000;287(5461):2196–2204
18 Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20(2): 265–272
19 Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 2001;98(17): 9748–9753
20 Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 2012;40(20):e155
21 Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 2012;28(11):1420–1428
22 Wu YW, Rho M, Doak TG, Ye Y. Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics. Bioinformatics 2012;28(18):i363–i369
23 Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol 2012;13(12):R122
24 Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 2010;17(11):1519–1533
25 Koren S, Treangen TJ, Pop M. Bambus 2: scaffolding metagenomes. Bioinformatics 2011;27(21):2964–2971
26 Yok NG, Rosen GL. Combining gene prediction methods to improve metagenomic gene annotation. BMC Bioinformatics 2011;12:20
27 Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 2006;34(19):5623–5630
28 Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 2010;38(12):e132
29 Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 2009;37(Web Server issue):W101–105
30 Cole JR, Wang Q, Cardenas E, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 2009;37(Database issue):D141–145
31 Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17(3):377–386
32 Krause L, Diaz NN, Goesmann A, et al. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 2008; 36(7):2230–2239
33 Gerlach W, Junemann S, Tille F, Goesmann A, Stoye J. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics 2009;10:430
34 Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27(23):4636–4641
35 Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 2009;6(9):673–676
36 Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 2011;12 Suppl 2:S4
37 Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9(8):811–814
38 Markowitz VM, Chen IM, Palaniappan K, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 2012;40(Database issue):D115–122
39 De Filippo C, Ramazzotti M, Fontana P, Cavalieri D. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 2012;13(6):696–710
40 Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–2461
41 Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39(Web Server issue):W29–37
42 Farfan F, Ma J, Sartor MA, Michailidis G, Jagadish HV. THINK Back: Knowledge-based interpretation of high throughput data. BMC Bioinformatics 2012;13 Suppl 2:S4
43 Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 2005; 102(43):15545–15550
44 Ma J, Sartor MA, Jagadish HV. Appearance frequency modulated gene set enrichment testing. BMC Bioinformatics 2011;12:81
45 Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 2012;8(6):e1002358
46 Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 2009;5(8):e1000465
47 Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005; 71(12):8228–8235
48 Dinsdale EA, Edwards RA, Bailey BA, et al. Multivariate analysis of functional metagenomes. Frontiers in Genetics 2013;4:41
49 Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60
50 Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature 2011;473(7346):174–180
51 Ravel J, Gajer P, Abdo Z, Sci USA. Vaginal microbiome of reproductive-age women. Proc Natl Acad 108 Suppl 2011, 1 SRC - GoogleScholar:4680–4687
52 Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 2011;9(4):279–290
53 Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 2011; 334(6052):105–108
54 Riehle K, Coarfa C, Jackson A, et al. The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. BMC Bioinformatics 2012;13 Suppl 13:S11
55 Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010;2010 (1):pdb prot5368
56 Aziz RK, Bartels D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75
57 Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. SmashCommunity: a metagenomic annotation and analysis tool. Bioinformatics 2010;26(23):2977–2978
58 Treangen TJ, Koren S, Sommer DD, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol 2013;14(1):R2
59 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10(3):R25
60 Koren O, Goodrich JK, Cullender TC, et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 2012;150(3):470–480
61 Han YW, Shen T, Chung P, Buhimschi IA, Buhimschi CS. Uncultivated bacteria as etiologic agents of intra-amniotic inflammation leading to preterm birth. J. Clin Microbiol 2009;47(1):38–47
62 Steel JH, Malatos S, Kennea N, et al. Bacteria and inflammatory cells in fetal membranes do not always cause preterm labor. Pediatr. Res 2005;57(3):404–411
63 Stout MJ, Conlon B, Landeau M, et al. Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. Am J Obstet Gynecol 2013;208(3):226 e221–227