Sequencing the human microbiome in health and disease

Michael J. Cox*, William O.C.M. Cookson and Miriam F. Moffatt

Molecular Genetics and Genomics Section, National Heart and Lung Institute, Imperial College London, Dovehouse Street, London SW3 6LY, UK

Received July 31, 2013; Revised and Accepted August 8, 2013

Molecular techniques have revolutionized the practice of standard microbiology. In particular, 16S rRNA sequencing, whole microbial genome sequencing and metagenomics are revealing the extraordinary diversity of microorganisms on Earth and their vast genetic and metabolic repertoire. The increase in length, accuracy and number of reads generated by high-throughput sequencing has coincided with a surge of interest in the human microbiota, the totality of bacteria associated with the human body, in both health and disease. Traditional views of host/pathogen interactions are being challenged as the human microbiota are being revealed to be important in normal immune system function, to diseases not previously thought to have a microbial component and to infectious diseases with unknown aetiology. In this review, we introduce the nature of the human microbiota and application of these three key sequencing techniques for its study, highlighting both advances and challenges in the field. We go on to discuss how further adoption of additional techniques, also originally developed in environmental microbiology, will allow the establishment of disease causality against a background of numerous, complex and interacting microorganisms within the human host.

INTRODUCTION

Microorganisms associated with the human body have been studied for many years in both health and disease. The first Human Microbiome Project perhaps began when Antonie van Leeuwenhoek scraped ‘gritty matter’ from between his teeth and became the first to visualize bacteria, or ‘animalcules’ in dental plaque in 1683 (1).

Since then, research on human-associated microorganisms has, for the pragmatic reason of combating infectious disease in human, veterinary and agricultural settings, tended to focus on pathogens. In addition, human-associated miocrobiome research has been limited because many bacteria are difficult to grow in the laboratory. Now techniques pioneered in environmental microbiology are being applied to human diseases and are revealing complex interactions between microorganisms themselves and with their human hosts. This holds promise for a new understanding of infectious disease and for diseases not previously recognized to have a microbial component.

With the advent of high-throughput sequencing substantial numbers of samples can be processed rapidly and cost effectively. These technological advances have led to an interest in the human as a super-organism made up of interacting human and microbial components. Such interactions may be complex and occur at many levels that extend well beyond the traditional models of host/pathogen and immune-virulence. Five to 8% of the human genome, for example, consists of endogenous retroviruses (2); gut bacteria may increase the risk of cardiovascular disease by the metabolic degradation of L-carnitine (3) and the gut microbiota may confer good health in the elderly by as yet unknown mechanisms (4). Thus, many aspects of human well-being may be influenced by our associated, integrated and ubiquitous microbiota.

In this review, we will introduce three key techniques in the field and speculate on implications for future research in human disease.

THE HUMAN MICROBIOTA

Microbiome literally means small biome, the ecosystem comprising all microorganisms in a particular environment together with their genes and environmental interactions. The assemblage of microorganisms themselves is referred to as the microbiota or microbial community and can include bacteria, archaea, viruses, phage, fungi and other microbial eukarya. Microflora

*To whom correspondence should be addressed. Tel: +44 2073518730; Fax: +44 2073518126; Email: michael.cox1@imperial.ac.uk

© The Author 2013. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
can also be found as a general term in the literature, but *flora* refers specifically to plants, rather than microbes and so the alternative terms are preferable. The human microbiota consists of microorganisms that exist upon, within or in close proximity to the human body. Bacteria are the most well-studied group of microorganisms in this context but archaea, viruses and eukarya such as fungi also account for a high proportion of the human microbiota (5–7). Parts of the body with significant bacterial populations include the gut and the oral cavity (8,9), the skin (10), urogenital tracts (11,12), the nasopharynx (13) and body sites canonically considered sterile such as the lower respiratory tract (14,15).

Organisms in these locations may be present stably over long periods (16) and may be considered endemic or may be transient (17). Composition of the microbiota tends to be defined by the body site (18) indicating the presence of different selection pressures. For example, moisture is an important driver of the community structure found on skin (19). In addition, the human microbiota has been shown to be individual, so twins share a similar but non-identical microbiota (20) and the use of microbiota analyses in forensic analysis has been suggested (21).

In designing studies of the human microbiota, variability over time and the individuality of composition necessitates longitudinal sampling, as an individual in their healthy state is their own best disease control. Large, cross-sectional study cohorts are nonetheless important when longitudinal sampling is not feasible.

Acquisition of the human microbiota is believed to start at birth (22). Bacteria may be present in amniotic fluid (23), whilst the bacteria in the newborn gut have been shown to be influenced by delivery mode with the microbiota sourced from the mother’s vagina during delivery or from skin with caesarean section (24). Accumulation of further organisms continues during infancy and childhood as demonstrated by longitudinal studies of the gut (8) and cross-sectional studies of the respiratory tract (25).

It is difficult to use culture methods to isolate more than a small percentage of the microorganisms known to be present in any environment. Isolating a wider range of organisms to be able to study them in detail takes considerable effort (26). The molecular techniques described below offer an alternative means to gather substantial detail about individual organisms and entire communities, whilst circumventing the selection biases inherent in isolation by culture (Fig. 1).

**16S rRNA GENE SEQUENCING**

16S rRNA gene sequencing has been the first molecular tool to be generally applied to the human microbiota. It gives a quantitative description of the bacteria present in a complex biological mixture, allowing investigation of whole communities and the identities of their constituent members.

The small subunit ribosomal or 16S rRNA gene (*rrnA*) is a highly conserved component of the transcriptional machinery of all DNA-based life forms. Phylogenetic mapping of *rrnA* variation was first used to establish the three domain structure of life (27). The conserved nature of the gene was subsequently exploited to develop more rapid methods for determining relationships between organisms directly from environmental DNA and RNA extracts (28,29).

The 16S rRNA gene consists of conserved and variable regions (Fig. 2). The variable regions allow discrimination between different microorganisms. 16S rRNA gene methods rely on the PCR (polymerase chain reaction) using ‘universal’ primers targeted at the conserved regions and designed to amplify as wide a range of different microorganisms as possible (30). This is followed by assaying the amplified fragment of the gene, originally with molecular fingerprinting approaches such as denaturing gradient gel electrophoresis (31) or by cloning and sequencing of the PCR products (32).

A number of different phylogenetic microarrays, consisting of multiple probes designed to discriminate sub-groups of organisms have also been used (33,34). The probes and the organisms targeted must be pre-selected, limiting the utility of arrays to detect novel organisms.

The coupling of 16S rRNA PCR with next-generation sequencing makes possible the study of many samples at low cost (35). One of the most significant limitations of 16S rRNA sequencing is the introduction of biases by PCR primer design, which may select for or against particular groups of organisms (36,37). An under-recognized problem with the use of the PCR is that bacterial contamination of reagents is common and requires extensive controls at many levels of the process (38). The 16S rRNA operon is also present in between one and fifteen copies in bacterial genomes. This may additionally influence the apparent relative abundance of an organism (39).

Analysis of 16S rRNA data relies on the clustering of related sequences at a particular level of identity and counting the number of representatives of each cluster. Since molecular methods for determining the identity of bacteria and archaea do not map directly to the classic biochemically determined taxonomies, clusters of similar sequences are referred to as operational taxonomic units (OTUs). OTU counts are summarized in a table of relative abundances for each organism in each sample and these tables are used for downstream analyses. A level of 97% sequence identity is frequently chosen as being representative of a species and 95% for a genus when using partial 16S rRNA gene sequences. Some bacteria and archaea will only be identified at the genus or family, rather than species level (40).

Accuracy of identification is dependent on the reference database chosen. Five percent of the 16S rRNA sequences in GenBank (the NIH sequence database) may be erroneous (41).
Curated databases such as The Ribosomal Database Project (42), GreenGenes (43) and SILVA (44), where sequences undergo quality assessment and alignments are manually optimized, are crucial for optimal phylogenetic placement of test sequences.

Two analysis pipelines are in common use for analyzing 16S rRNA gene sequence data: QIIME (45) and Mothur (46), though there is no standardized way of applying these pipelines to datasets.

An important deliverable of 16S rRNA gene sequencing is the identification of microorganisms that cause disease. Current microbial diagnostics provide information about the presence or absence of known pathogens in patient samples, but the culture-based techniques are much more targeted and selective when compared to 16S rRNA gene sequencing (47). More than 50% of cases of pneumonia in children and adults requiring hospitalization have no diagnosis. DNA sequencing therefore has the immediate potential to fill a major unmet clinical need.

Although identification and characterization of disease-causing organisms is the ultimate goal (see section Whole-Genome Sequencing), measures of the microbial community structure, such as species richness, community evenness and diversity, can reveal a great deal about dynamics and selection pressures experienced by the system (Fig. 3 and Table 1). Association of these parameters with relevant environmental and clinical measurements can give important insight into states of health and disease (48,49). Increased richness, evenness and diversity can be associated with stable, longer established or less active ecosystems (50). Microbial community stability, resistance to environmental pressures such as diet and antibiotic use, and resistance to invasion with pathogens are also likely to be important in human disease states affecting the bowel, mouth, lungs, skin and vagina (51).

### WHOLE-GENOME SEQUENCING

Complete genome sequencing is the foundation for the comprehensive understanding of an organism’s function. Bacteria were the first free-living organisms to undergo complete genome sequencing, with *Haemophilus influenzae* being completed in 1995 (52). As of July 2013, the National Center for Biotechnology Information’s microbial genome site listed 2552 complete genomes, although the bacteria sequenced in their entirety have been highly selected, with multiple genomes of commonly cultured clinical strains and an absence of some entire phyla (53).

It is now feasible to map all the genes that characterize a particular group of organisms, their ‘pangenomes,’ by sequencing a broad range of isolates from different sources (54). This reveals the genes that are core and that define the genomes of a particular group as well as those that are accessory, perhaps possessed by a single isolate or a subset with a particular lifestyle or pathology. It also allows inference of how pathogenicity evolves within
lineages of organisms that consist of both pathogenic and non-pathogenic organisms.

Molecular epidemiology where isolates associated with a particular disease outbreak are typed and tracked to pinpoint their source also benefits from whole-genome sequencing (WGS). Multi-locus sequence typing (MLST) relies on the sequencing of multiple house-keeping genes for a particular species (55). This can reveal important details about transmission and sources of an outbreak that can inform future responses, although it may suffer from a lack of resolution (56). Epidemiological studies have almost exclusively been conducted on stored isolates after an outbreak, limiting their usefulness in altering the course of an outbreak as it occurs.

High-throughput sequencing may bridge these gaps as it becomes feasible to generate whole-genome sequences faster than the outbreak source can be tracked. Sequencing may deliver detailed information about virulence factors, antibiotic resistance and strain source quickly enough to influence outbreak responses. The additional genomic information allows much higher resolution of strains than MLST. In 2011, a shiga-toxigenic Escherichia coli O104:H4 outbreak occurred in Germany and isolates were sequenced, data released and the assembly crowd-sourced within a week (57). Application of WGS to a neonatal methicillin resistant Staphylococcus aureus outbreak revealed a transmission event missed by other techniques and did so in clinically relevant timescales (58).

Annotation of sequence data and comparative genomics is onerous and not currently within the realm of clinical diagnostic or epidemiological laboratories. Raw sequence data must undergo quality control prior to assembly of the curated reads as a representative genome. Coding regions and their putative functions are identified during annotation and then the assembled sequence deposited in public databases. Finally, the new genome may be considered in the context of previously sequenced microorganisms (59,60). Tools for facile annotation and analysis are in development, and may allow translation of the methodology into routine clinical practice (61). Importantly, WGS still requires the isolation of the organisms concerned which is not always feasible.

METAGENOMICS
Metagenomics describes the direct sequencing of the total DNA extracted from a microbial community and combines elements of the above two approaches.

Identification of organisms present is improved relative to 16S rRNA gene sequencing, and organisms such as phage and viruses that do not have a phylogenetic marker gene can be assessed (62). This is tempered by an increased sequencing effort to detect less common organisms, relative to 16S rRNA gene sequencing. Functional annotation allows the comparison of the physiological capabilities between communities and environmental conditions (63). With sufficient representation of the organisms present it may be possible to reconstruct the metabolic and biogeochemical pathways between microorganisms and to use this to direct isolation attempts (64,65), gaining insight into the fundamental functioning of the ecosystem.

Metagenomics is beginning to be applied to the human microbiota. The healthy human gut has been postulated to consist of three metagenomically defined enterotypes, typified by relative dominance of particular groups of organisms: Prevotella, Ruminococcus and Bacteroides spp. (66). Start-up companies and research projects are now offering to do for your gut microbiota what direct to consumer genetics companies do for the human genome, classifying your faecal sample into an enterotype. It is unclear, however, what the gut enterotype might mean to the consumer, as other studies have suggested continuous measures of community structure are more representative of gut microbial diversity and that enterotypes may be artefacts of the analytic methods employed (67).

More recently, metagenome wide association studies (MWAS) have begun to emerge. In a study of Type 2 diabetes (T2D), although microbial composition of the gut varied between individuals, it did not vary greatly between cases and controls. Some functional differences were observed in the gut microbiota, with a relative decline in butyrate-producing bacteria found in T2D (68). Butyrate is a key compound in host microbiota metabolism in the gut, being a bacterial metabolite produced by key strains such as Roseburia spp. and Faecalibacterium prausnitzii (69,70) as well as a preferred carbon source for gut epithelium cells (71). For atherosclerosis, MWAS has shown depletion of the same organisms in cases of disease, accompanied by genes involved in peptidoglycan synthesis (72).

FUTURE DIRECTIONS
Many of the existing molecular studies of the human microbiota are hypothesis generating and associate particular OTUs or organisms with clinical measurements. These may yield useful biomarkers of disease but fall short generally of establishing true causality. Causality may be established by borrowing further techniques from environmental microbial ecology. For example, 16S rRNA sequencing, WGS and metagenomic studies may be combined with approaches with that allow activity of the community to be measured, including metabolomics (73), metaproteomics (74) and metatranscriptomics (75).

Little is known about the consistency of the microbiota at particular sites in different populations and environments. It is likely that genetic and epidemiological approaches that include systematic quantification of microbial communities will be able to identify the host factors that support healthy microbial communities as well as those that predispose to disease.

Fluorescence in situ hybridization is useful for direct visualization and identification of microorganisms in human tissue (76). This is important since DNA and RNA extracts often use samples such as sputum and stool, which do not give good information on the localization and spatial heterogeneity of the organisms. Techniques such as stable isotope probing (77) are also available that simultaneously reveal function, identity and activity of microorganisms in close to in situ conditions. Finally, single cell genomics shows promise as a technique that effectively allows isolation of an organism without culture (78).

Treatments targeted at the microbiota for the moment tend to focus on probiotics, introduced bacteria such as Lactobacillus spp., or prebiotics, compounds that enrich the growth of particular organisms deemed to be of benefit (79). Faecal transplants are perhaps amongst the more surprising methods for amending microbial communities. In the case of chronic Clostridium difficile
infection, transplanting a faecal slurry from a healthy donor to the patient’s bowel has demonstrated good efficacy (80), although the importance of pre-screening the transplant community for potential pathogens should be emphasized.

The pervasive role of the human microbiota in health and disease is still largely unexplored. Three hundred years after van Leeuwenhoek first visualized his dental calculus, only 50% of the microorganisms present in the oral microbiota can be cultivated. This is the highest proportion of any of the human host niches (81). The challenge remains for human microbial ecologists to establish causal relationships between the microbiota and the human host in the varied microbial niches of the body and varying disease states, with the ultimate goal of translation into real clinical benefit.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their constructive comments. W.O.C.M.C and M.F.M. are recipients of a Wellcome Trust Joint Senior Investigator Award, which also supports M.J.C.

Conflict of Interest statement. None declared.

REFERENCES


