Sequencing the human microbiome in health and disease

Journal:	*Human Molecular Genetics*
Manuscript ID:	HMG-2013-D-00971.R1
Manuscript Type:	4 Invited Review Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Cox, Michael; Imperial College London, National Heart and Lung Institute Cookson, William; Imperial College London, National Heart and Lung Institute Moffatt, Miriam; Imperial College, National Heart and Lung Institute
Key Words:	human microbiome, 16S rRNA, metagenomics, bacteria

Sequencing the human microbiome in health and disease

Michael J Cox*, William OCM Cookson and Miriam F Moffatt

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

*corresponding author
Tel: +44 207 351 8730
Fax: +44 207 351 8126
Email: michael.cox1@imperial.ac.uk

Abstract

Molecular techniques have revolutionised the practise 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 in 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 visualise 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-microbe 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 recognised 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 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 microflora (5, 6, 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 (Figure 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 (Figure 2A). 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 designs, which may select for or against particular groups of organisms (36, 37). An under-recognised 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, which 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 summarised 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 optimised, are crucial for optimal phylogenetic placement of test sequences. Two analysis pipelines are in common use for analysing 16S rRNA gene sequence data: QIIME (45) and Mothur (46), though there is no standardised 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 hospitalisation have no diagnosis. DNA sequencing therefore has the immediate potential to fill a major unmet clinical need.

Although identification and characterisation of disease-causing organisms is the ultimate goal (see whole genome sequencing below), 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 (Figure 2B 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 (WGS)

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 characterise 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 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 shigatoxigenic *Escherichia coli* O104:H4 outbreak occurred in Germany and isolates were sequenced, data released and the annotation 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 time-scales (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, whole genome sequencing 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 (DTC) 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.

Fluorescent *in situ* hybridisation (FISH) is useful for direct visualisation 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 localisation 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 emphasised.

The pervasive role of the human microbiota in health and disease is still largely unexplored. Three hundred years after van Leeuwenhoek first visualised 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. The authors are supported by a Wellcome Trust Senior Investigator Award to WOCMC and MFM.

References

1. Van Leuwenhoek,A. (1683) An Extract of a Letter from Mr. Anth. Van Leuwenhoek, concerning Animalcules Found on the Teeth; Of the Scaleyness of the Skin, &c. *Phil. Trans. (1683-1775),* 646–649.

2. Belshaw,R., Pereira,V., Katzourakis,A., Talbot,G., Paces,J., Burt,A. and Tristem,M. (2004) Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 4894–4899.

3. Koeth,R.A., Wang,Z., Levison,B.S., Buffa,J.A., Org,E., Sheehy,B.T., Britt,E.B., Fu,X., Wu,Y., Li,L., et al. (2013) Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.*, **19**, 576–585.

4. Claesson,M.J., Cusack,S., O'Sullivan,O., Greene-Diniz,R., de Weerd,H., Flannery,E., Marchesi,J.R., Falush,D., Dinan,T., Fitzgerald,G., et al. (2011) Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U.S.A.*, **108 Suppl. 1**, 4586–4591.

5. Duncan,S.H., Louis,P. and Flint,H.J. (2007) Cultivable bacterial diversity from the human colon. *Lett. Appl. Microbiol.*, **44**, 343–350.

6. Delwart,E. (2013) A Roadmap to the Human Virome. *PLoS Path.*, **9**, e1003146.

7. Parfrey,L.W., Walters,W.A. and Knight,R. (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front. Microbio.l*, **2**, 153, 1-6.

8. Yatsunenko,T., Rey,F.E., Manary,M.J., Trehan,I., Dominguez-Bello,M.G., Contreras,M., Magris,M., Hidalgo,G., Baldassano,R.N., Anokhin,A.P., et al. (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.

9. Wade,W.G. (2013) Pharmacological Research. *Pharmacol. Res.*, **69**, 137–143.

10. Grice,E.A., Kong,H.H., Renaud,G., Young,A.C., NISC Comparative Sequencing Program, Bouffard,G.G., Blakesley,R.W., Wolfsberg,T.G., Turner,M.L. and Segre,J.A. (2008) A diversity profile of the human skin microbiota. *Genome Res.*, **18**, 1043–1050.

11. Price,L.B., Liu,C.M., Johnson,K.E., Aziz,M., Lau,M.K., Bowers,J., Ravel,J., Keim,P.S., Serwadda,D., Wawer,M.J., et al. (2010) The Effects of Circumcision on the Penis Microbiome. *PLoS ONE*, **5**, e8422.

12. Ravel,J., Gajer,P., Abdo,Z., Schneider,G.M., Koenig,S.S.K., McCulle,S.L., Karlebach,S., Gorle,R., Russell,J., Tacket,C.O., et al. (2011) Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.*, **108 Suppl 1**, 4680–4687.

13. Bogaert,D., Keijser,B., Huse,S., Rossen,J., Veenhoven,R., van Gils,E., Bruin,J., Montijn,R., Bonten,M. and Sanders,E. (2011) Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. *PLoS ONE*, **6**,

e17035.

14. Hilty,M., Burke,C., Pedro,H., Cardenas,P., Bush,A., Bossley,C., Davies,J., Ervine,A., Poulter,L., Pachter,L., et al. (2010) Disordered microbial communities in asthmatic airways. *PLoS ONE*, **5**, e8578.

15. Charlson,E.S., Bittinger,K., Haas,A.R., Fitzgerald,A.S., Frank,I., Yadav,A., Bushman,F.D. and Collman,R.G. (2011) Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. *Am. J. of Resp. and Crit. Care Med.*, **184**, 957–963.

16. Faith,J.J., Guruge,J.L., Charbonneau,M., Subramanian,S., Seedorf,H., Goodman,A.L., Clemente,J.C., Knight,R., Heath,A.C., Leibel,R.L., et al. (2013) The Long-Term Stability of the Human Gut Microbiota. *Science*, **341**, 1237439–1237439.

17. Caporaso,J.G., Lauber,C.L., Costello,E.K., Berg-Lyons,D., Gonzalez,A., Stombaugh,J., Knights,D., Gajer,P., Ravel,J., Fierer,N., et al. (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50, 1-8.

18. Consortium,T.H.M.P. (2013) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

19. Findley,K., Oh,J., Yang,J., Conlan,S., Deming,C., Meyer,J.A., Schoenfeld,D., Nomicos,E., Park,M., NIH Intramural Sequencing Center Comparative Sequencing Program, et al. (2013) Topographic diversity of fungal and bacterial communities in human skin. *Nature*, **498**, 367–370.

20. Turnbaugh,P.J., Hamady,M., Yatsunenko,T., Cantarel,B.L., Duncan,A., Ley,R.E., Sogin,M.L., Jones,W.J., Roe,B.A., Affourtit,J.P., et al. (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

21. Fierer,N., Lauber,C.L., Zhou,N., McDonald,D., Costello,E.K. and Knight,R. (2010) From the Cover: Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6477–6481.

22. Palmer,C., Bik,E.M., DiGiulio,D.B., Relman,D.A. and Brown,P.O. (2007) Development of the human infant intestinal microbiota. *PLoS Biol.*, **5**, e177.

23. Han,Y.W., Redline,R.W., Li,M., Yin,L., Hill,G.B. and McCormick,T.S. (2004) Fusobacterium nucleatum Induces Premature and Term Stillbirths in Pregnant Mice: Implication of Oral Bacteria in Preterm Birth. *Inf. and Imm.*, **72**, 2272–2279.

24. Dominguez-Bello,M.G., Costello,E.K., Contreras,M., Magris,M., Hidalgo,G., Fierer,N. and Knight,R. (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 11971–11975.

25. Cardenas,P.A., Cooper,P.J., Cox,M.J., Chico,M., Arias,C., Moffatt,M.F. and Cookson,W.O. (2012) Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. *PLoS ONE*, **7**, e46803.

26. Tunney,M.M., Klem,E.R., Fodor,A.A., Gilpin,D.F., Moriarty,T.F., McGrath,S.J., Muhlebach,M.S., Boucher,R.C., Cardwell,C., Doering,G., et al. (2011) Use of culture and molecular analysis to determine the effect of antibiotic treatment on microbial community diversity and abundance during exacerbation in patients with cystic fibrosis. *Thorax*, **66**, 579–584.

27. Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5088–5090.

28. Lane,D.J., Pace,B., Olsen,G.J., Stahl,D.A., Sogin,M.L. and Pace,N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 6955–6959.

29. Olsen,G.J., Lane,D.J., Giovannoni,S.J., Pace,N.R. and Stahl,D.A. (1986) Microbial Ecology and Evolution: A Ribosomal RNA Approach. *Annu. Rev. Microbiol.*, **40**, 337–365.

30. Lane,D.J. (1991) 16S/23S rRNA sequencing. In Stackebrandt,E., Goodfellow,M. (eds), *Nucleic acid techniques in bacterial systematics*.

31. Muyzer,G., de Waal,E.C. and Uitterlinden,A.G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *App. and Environ. Microbiol.*, **59**, 695–700.

32. Eckburg,P.B. (2005) Diversity of the Human Intestinal Microbial Flora. *Science*, **308**, 1635–1638.

33. Desantis,T.Z., Brodie,E.L., Moberg,J.P., Zubieta,I.X., Piceno,Y.M. and Andersen,G.L. (2007) High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microb. Ecol.*, **53**, 371–383.

34. Rajilić-Stojanović,M., Heilig,H.G.H.J., Molenaar,D., Kajander,K., Surakka,A., Smidt,H. and de Vos,W.M. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.*, **11**, 1736–1751.

35. Metzker,M.L. (2009) Sequencing technologies—the next generation. *Nat. Rev. Genetics,* **11**, 31-46.

36. Sim,K., Cox,M.J., Wopereis,H., Martin,R., Knol,J., Li,M.-S., Cookson,W.O.C.M., Moffatt,M.F. and Kroll,J.S. (2012) Improved Detection of Bifidobacteria with Optimised 16S rRNA-Gene Based Pyrosequencing. *PLoS ONE*, **7**, e32543.

37. Klindworth,A., Pruesse,E., Schweer,T., Peplies,J., Quast,C., Horn,M. and Glockner,F.O. (2012) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.*, **41**, e1, 1-14.

38. Tanner, MA., Goebel, BM., Dojka, MA., and Pace, NR. (1998) Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ. Microbiol*. 1998, **64**, 3110–3113

39. Klappenbach, JA., Saxman, P.R., Cole, JR and Schmidt, TM. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res*., **29**, 181-184

40. Stackebrandt,E. and Goebel,B.M. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.*, **44**, 846–849.

41. Ashelford,K.E., Chuzhanova,N.A., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *App. and Environ. Micro.*, **71**, 7724–7736.

42. Cole, JR., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, RJ., Kulam-Syed-Mohideen, AS., McGarrell, DM., Marsh, T., Garrity, GM. *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*., **37 (suppl. 1)** D141-D145.

43. McDonald,D., Price,M.N., Goodrich,J., Nawrocki,E.P., Desantis,T.Z., Probst,A., Andersen,G.L., Knight,R. and Hugenholtz,P. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.

44. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glockner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

45. Caporaso,JG., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,FD., Costello,EK., Fierer,N., Peña,AG., Goodrich,JK., Gordon,JI., et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Meth.*, **7**, 335–336.

46. Schloss,PD., Westcott,SL., Ryabin,T., Hall,JR., Hartmann,M., Hollister,EB., Lesniewski,RA., Oakley,BB., Parks,DH., Robinson,CJ., et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environ. Microbiol.*, **75**, 7537–7541.

47. Duff,R.M., Simmonds,N.J., Davies,J.C., Wilson,R., Alton,E.W., Pantelidis,P., Cox,M.J., Cookson,W.O.C.M., Bilton,D. and Moffatt,M.F. (2013) A molecular comparison of microbial communities in bronchiectasis and cystic fibrosis. *Eur. Respir. J.*, **41**, 991–993.

48. Cox,M.J., Allgaier,M., Taylor,B., Baek,M.S., Huang,Y.J., Daly,R.A., Karaoz,U., Andersen,G.L., Brown,R., Fujimura,K.E., et al. (2010) Airway Microbiota and Pathogen Abundance in Age-Stratified Cystic Fibrosis Patients. *PLoS ONE*, **5**,

e11044.

49. Ley,R.E., Bäckhed,F., Turnbaugh,P., Lozupone,C.A., Knight,R.D. and Gordon,J.I. (2005) Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 11070–11075.

50. Legendre,P. and Legendre,L. (2012) Numerical Ecology 3rd ed. Elsevier.

51. Shade,A., Peter,H., Allison,S.D., Baho,D.L., Berga,M., Bürgmann,H., Huber,D.H., Langenheder,S., Lennon,J.T., Martiny,J.B.H., et al. (2012) Fundamentals of microbial community resistance and resilience. *Front. Microbiol.*, **3**, 417.

52. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. and Merrick,J.M. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

53. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J., et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.

54. Pallen,M.J. and Wren,B.W. (2007) Bacterial pathogenomics. *Nature*, **449**, 835–842.

55. Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A., et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 3140–3145.

56. Roetzer,A., Diel,R., Kohl,T.A., Rückert,C., Nübel,U., Blom,J., Wirth,T., Jaenicke,S., Schuback,S., Rüsch-Gerdes,S., et al. (2013) Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med.*, **10**, e1001387.

57. Rohde,H., Qin,J., Cui,Y., Li,D., Loman,N.J., Hentschke,M., Chen,W., Pu,F., Peng,Y., Li,J., et al. (2011) Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.*, **365**, 718–724.

58. Köser,C.U., Holden,M.T.G., Ellington,M.J., Cartwright,E.J.P., Brown,N.M., Ogilvy-Stuart,A.L., Hsu,L.Y., Chewapreecha,C., Croucher,N.J., Harris,S.R., et al. (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.*, **366**, 2267–2275.

59. Edwards,DJ. and Holt,KE. (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform. Exp.*, **3**, 2.

60. Loman,NJ., Constantinidou,C., Chan,JZM., Halachev,M., Sergeant,M., Penn,CW., Robinson,ER. and Pallen,MJ. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Rev. Microbiol.*, **10**, 599–606.

61. Richardson,E.J. and Watson,M. (2013) The automatic annotation of bacterial genomes. *Brief. Bioinformatics*, **14**, 1–12.

62. Willner,D., Haynes,M.R., Furlan,M., Hanson,N., Kirby,B., Lim,Y.W., Rainey,P.B., Schmieder,R., Youle,M., Conrad,D., et al. (2012) Case Studies of the Spatial Heterogeneity of DNA Viruses in the Cystic Fibrosis Lung. *Am. J. of Resp. Cell and Mol. Biol.*, **46**, 127–131.

63. Turnbaugh,P.J., Ridaura,V.K., Faith,J.J., Rey,F.E., Knight,R. and Gordon,J.I. (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science Trans. Med.*, **1**, 6ra14, 1-19.

64. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

65. Tyson,G.W., Lo,I., Baker,B.J., Allen,E.E., Hugenholtz,P. and Banfield,J.F. (2005) Genome-Directed Isolation of the Key Nitrogen Fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an Acidophilic Microbial Community. *App. and Environ. Microbiol.*, **71**, 6319–6324.

66. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.-M., et al. (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

67. Koren,O., Knights,D., Gonzalez,A., Waldron,L., Segata,N., Knight,R., Huttenhower,C. and Ley,R.E. (2013) A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Comput. Biol.*, **9**, e1002863.

68. Qin,J., Li,Y., Cai,Z., Li,S., Zhu,J., Zhang,F., Liang,S., Zhang,W., Guan,Y., Shen,D., et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

69. Duncan,S.H. (2006) Proposal of *Roseburia faecis* sp. nov., *Roseburia hominis* sp. nov. and *Roseburia inulinivorans* sp. nov., based on isolates from human faeces. *Int. J. Syst. Evol. Microbiol.*, **56**, 2437–2441.

70. Lopez-Siles, M., Khan, TM., Duncan, SH., Harmsen, HJM., Hermie, JM., Garcia-Gil, J., and Flint, HJ. (2012) Cultured Representatives of Two Major Phylogroups of Human Colonic *Faecalibacterium prausnitzii* Can Utilize Pectin, Uronic Acids, and Host-Derived Substrates for Growth. *Applied and Environ. Microbiol.*, **78**, 420-428

71. Cummings,J.H., Pomare,E.W., Branch,W.J., Naylor,C.P. and Macfarlane,G.T. (1987) Short chain fatty acids in human large intestine, portal, hepatic and venous blood. *Gut*, **28**, 1221–1227.

72. Karlsson,F.H., Fåk,F., Nookaew,I., Tremaroli,V., Fagerberg,B., Petranovic,D., Bäckhed,F. and Nielsen,J. (2012) Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Commun.*, **3**, 1245.

73. Nicholson,J.K., Holmes,E. and Wilson,I.D. (2005) Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Microbiol.*, **3**, 431–438.

74. Verberkmoes,N.C., Russell,A.L., Shah,M., Godzik,A., Rosenquist,M., Halfvarson,J., Lefsrud,M.G., Apajalahti,J., Tysk,C., Hettich,R.L., et al. (2008) Shotgun metaproteomics of the human distal gut microbiota. *ISME J.*, **3**, 179–189.

75. Poretsky,R.S., Hewson,I., Sun,S., Allen,A.E., Zehr,J.P. and Moran,M.A. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.*, **11**, 1358–1375.

76. Geißdörfer,W., Moos,V., Moter,A., Loddenkemper,C., Jansen,A., Tandler,R., Morguet,A.J., Fenoller,F., Raoult,D., Bogdan,C., and Schneider,T. (2012) High frequency of *Tropheryma whipplei* in culture negative endocarditis. *J. Clin. Micro*. **50**, 216-222.

77. Radajewski,S., Ineson,P., Parekh,N.R. and Murrell,J.C. (2000) Stable-isotope probing as a tool in microbial ecology. *Nature*, **403**, 646–649.

78. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,NN., Anderson,IJ., Cheng,J-F., Darling,A., Malfatti,S., Swan,BK., Gies,EA., *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.

79. Gareau,M.G., Sherman,P.M. and Walker,W.A. (2010) Probiotics and the gut microbiota in intestinal health and disease. *Nat. Rev. Gastroenterol. Hepatol.*, **7**, 503–514.

80. van Nood,E., Vrieze,A., Nieuwdorp,M., Fuentes,S., Zoetendal,E.G., de Vos,W.M., Visser,C.E., Kuijper,E.J., Bartelsman,J.F.W.M., Tijssen,J.G.P., et al. (2013) Duodenal Infusion of Donor Feces for Recurrent *Clostridium difficile*. *N. Engl. J. Med.*, **368**, 407–415.

81. Wilson,M.J., Weightman,A.J. and Wade,W.G. (1997) Applications of molecular ecology in the characterization of uncultured microorganisms associated with human disease. *Rev. in Med. Microbiol.*, **8**, 91.

Figure Legends

1: A schematic demonstrating the processes for 16S rRNA gene sequencing, whole genome sequencing and metagenomics. Sample collection, DNA extraction, sequencing and sequence analysis are required in all three techniques. 16S rRNA gene sequencing and whole genome sequencing involve additional steps.

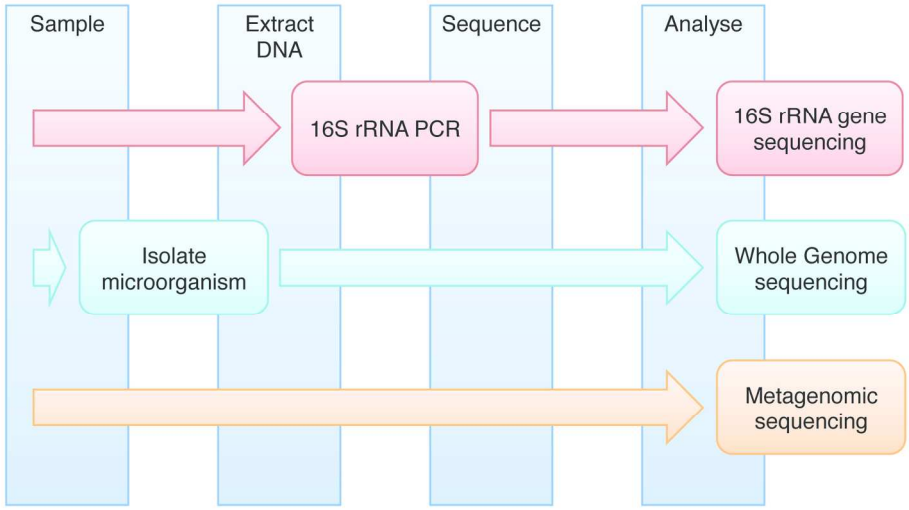
2A: The approximately 1.5 kb 16S rRNA gene of *Escherichia coli* showing the 9 variable regions that make it an ideal target as a phylogenetic marker gene.

2B: A diagram demonstrating species richness and evenness and how they describe the composition of a community. Each shape represents an individual and the colour and nature of the shape represents a different type of organism. Increased numbers of different types of organism is described as increased species richness. When no one organism is dominant, the community is described as even.

Table 1: Explanation of commonly used ecological terms in the field of microbiota research.

Term	Explanation
Evenness	A measure of the skew in abundance of community members. Is there one dominant organism or are all evenly represented?
Richness	The number of different types of organism present.
Diversity	A combination of richness and evenness; can be considered to be a summary statistic for community structure as membership, abundance and evenness are taken into account.
Simpson index	A common diversity index indicating the probability that two individuals taken at random from a population are the same. Often presented as the inverse so that increasing diversity is mirrored by an increasing index value.
Shannon index	Alternatively, Shannon entropy – another common diversity index that quantifies the uncertainty of predicting the next individual taken from a sample.
Alpha diversity	Within sample diversity.
Beta diversity	Between sample diversity.

Figure 1



A schematic demonstrating the processes for 16S rRNA gene sequencing, whole genome sequencing and metagenomics. Sample collection, DNA extraction, sequencing and sequence analysis are required in all three techniques, though 16S rRNA gene sequencing and whole genome sequencing involve additional steps.
180x127mm (300 x 300 DPI)

Figure 2A

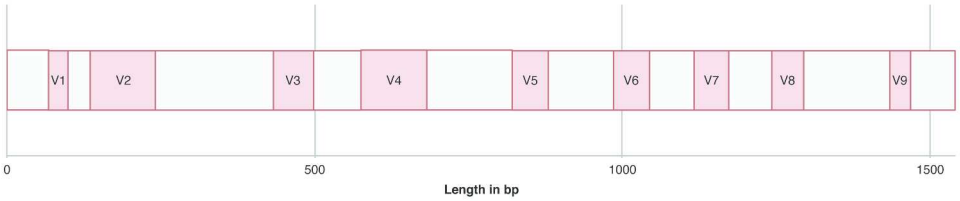
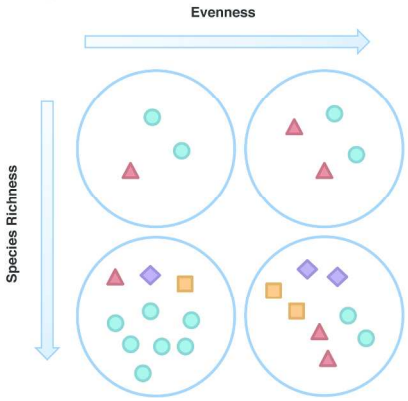


Figure 2B



2A The approximately 1.5 Kb 16S rRNA gene of Escherichia coli, showing the 9 variable regions that make it an ideal target as a phylogenetic marker gene.

2B A diagram demonstrating species richness and evenness and how they describe the composition of a community. Each shape represents an individual and the colour and nature of the shape represents a different type of organism. Increased numbers of different types of organism is described as increased species richness. When no one organism is dominant, the community is described as even.
254x177mm (300 x 300 DPI)