Data mining the human gut microbiota for therapeutic targets
Matthew Collison, Robert P. Hirt, Anil Wipat, Sirintra Nakjang, Philippe Sanseau and James R. Brown

Abstract
It is well known that microbes have an intricate role in human health and disease. However, targeted strategies for modulating human health through the modification of either human-associated microbial communities or associated human-host targets have yet to be realized. New knowledge about the role of microbial communities in the microbiota of the gastrointestinal tract (GIT) and their collective genomes, the GIT microbiome, in chronic diseases opens new opportunities for therapeutic interventions. GIT microbiota participation in drug metabolism is a further pharmaceutical consideration. In this review, we discuss how computational methods could lead to a systems-level understanding of the global physiology of the host–microbiota superorganism in health and disease. Such knowledge will provide a platform for the identification and development of new therapeutic strategies for chronic diseases possibly involving microbial as well as human-host targets that improve upon existing probiotics, prebiotics or antibiotics. In addition, integrative bioinformatics analysis will further our understanding of the microbial biotransformation of exogenous compounds or xenobiotics, which could lead to safer and more efficacious drugs.

Keywords: microbiome; computational biology; bioinformatics; drug discovery

INTRODUCTION
Studies of the human genome have led to key advancements in biomedical science and drug discovery. However, it is increasingly apparent that determinants of our health are not solely controlled by our own genomes. Rather, many disease pathologies involve the interplay between the human body, the external environment and the complex communities of microorganisms residing on the mucosal surfaces of our respiratory tract [1], urogenital tract [2], gastrointestinal tract (GIT) [3], and skin [4]. The complement of microbial cells coinhabiting any individual human being, the microbiota, exceeds at least 10-fold the number of human origin cells [5,6]. Furthermore, the gene collective of this residing microbial community, the microbiome, exceeds by at least 100 times the complement of genes present in the human nuclear genome [7]. While our knowledge of infectious diseases and the causative bacterial, viral and eukaryotic pathogens is well established, the roles of complex non-pathogenic microbiota communities in sustaining health or promoting disease are only recently studied.

The advent of sensitive, high-volume DNA sequencing and metabolomics technologies has lead to a rapid expansion of data sets from microorganism populations associated with various chronic disease phenotypes. Major public funding initiatives, such as the US National Institutes of Health (NIH) Human Microbiome Project, initiated in 2007 [8],
and the EU MetaHIT Consortium, begun in 2008 [9], are driving the characterization of microbiomes from hundreds, soon thousands, of individuals of different ages, geographical, dietary and disease backgrounds. New methodologies are also being developed for modulating the microbiota in model organisms to better understand their associations with disease phenotypes. Together these rich data sources are driving further innovation in computational biology and bioinformatics databases related to the human microbiota (Figure 1).

Historically, chronic and infectious diseases have been distinct and non-overlapping areas of drug discovery. However, rapidly emerging science in human–microbe interactions is opening new therapeutic paradigms where viral, bacterial and parasitic infections could be treated via human-host gene targets. Conversely, certain pathogens or members of the microbiota contribute to the progression and pathology of chronic diseases; thus, opening treatment options that might include anti-bacterial or anti-viral therapeutics (Figure 2). Recently, we presented a computational biology strategy for identifying potential human–host targets for the treatment of infectious diseases [10]. Here, we discuss computational approaches for the identification and prioritization of targets linked to the human microbiota for treatment of chronic diseases. In this review, we emphasise GIT microbiome computational analyses because of the pharmacological significance of this microbial community in both chronic diseases and drug metabolism. However, the various computational methodologies and databases discussed here in the context of GIT microbiome studies can be more widely applied to the analysis of other microbial communities and their associated pathologies.

**THE GUT MICROBIOTA IN HUMAN HEALTH AND DRUG METABOLISM**

With >200 m² of mucosal surface area and a nutrient-rich environment, the GIT hosts the majority of the human microbiota [5,11] (Figure 3). Absorption
in the distal gut results in \(~10\%\) of the metabolites in the host systemic blood flow being of bacterial origin [12]. Changes in gut microbiota have been linked with numerous GIT and other systemic diseases [13]. The broad implications of GIT microbiota on human physiology and ready access to this community via faecal sampling (about 60% of faecal material is microbial biomass [13,14]) have driven studies probing the status of the GIT microbial ecosystem in healthy and diseased populations.

In healthy individuals, changes in the GIT microbiota have been associated with host genetics [15], aging [16] and dietary patterns [17,18]. A recent population study involving 39 individuals from different cultures, geographical locations, races, as well as gut-associated and dietary disease patients found that subjects’ GIT microbial communities could be segregated into three statistically robust clusters [19]. These clusters, known as enterotypes, were also found to be consistent across major populations, including 85 European [7] and 154 American [18] individuals. Enterotypes were differentiated by the relative abundances of three bacterial genera, \textit{Bacteroides}, \textit{Prevotella} (both of the phylum Bacteroidetes) and \textit{Ruminococcus} (of the phylum Firmicute) [19]. The stratified nature of the enterotype data indicate that the microbial ecology in each individual’s gut establishes a stable and structured biodiversity that is independent of nationality, continent, sex, age or other physical phenotypic factors such as body mass index.

Other studies suggest that long-term rather than short-term diet are important in shaping GIT microbial communities. Comparisons between children from Europe on a typical Western diet—high in animal protein and fat—and children from Burkina Faso in Africa on a low animal protein and high carbohydrate diet found the \textit{Bacteroides} enterotype was higher in Europeans while the \textit{Prevotella} enterotype predominated in African children [20]. Another recent study also found that animal fat and high protein versus carbohydrate-rich diets correlated with the \textit{Bacteroides} and \textit{Prevotella} enterotypes, respectively [17]. In the same study, controlled feeding of 10 subjects with high-fat, low-fibre versus low-fat, high-fibre diets produced detectable changes in

\textbf{Figure 3:} The pervasive presence of the microbiota across human surfaces and its integration with the host biology. The left-hand side lists key attributes and influences underlying the human host genome and the right-hand side lists key attributes and influences underlying the microbiota. The figure is split into three horizontal subsections illustrating environmental factors (boxes distinguishing the biotic and the abiotic elements), genetic and cellular feature (including the source of variability) and disorders.
their microbiome within 24 h. However, overall individuals’ enterotypes remained stable for the duration of the 10-day study, suggesting that long term rather than transient dietary trends determine the ecological structure of the gut microbial communities.

Other factors such as environmental exposure also influence GIT microbiota biodiversity and microbiome composition. Babies delivered naturally have microbial communities most similar to that of the mother’s vagina, while the microbiota from neonates born by caesarean section closely resembles the mother’s skin bacteria [21]. A fascinating case of lateral gene transfer (LGT) is the occurrence of marine bacterial genes encoding the carbohydrate-active enzymes that metabolize algal polysaccharides in the microbiomes of Japanese individuals whose diets include a high content of seaweed [22]. Further understanding of the microbiota variation across human populations, as well as the environmental and genetic factors shaping it, will be important in designing future microbiome-targeted therapeutics.

One central aspect of the human microbiota symbiosis is the dialogue between the microbiota and the immune system [23,24]. The microbiota contributes to the development of both the mucosal and systemic immune systems. It is now appreciated that the loss of homoeostasis in the GIT immune system plays a central role in numerous disease conditions, including inflammatory bowel disease (IBD) [13,25,26]. Homoeostasis of the mucosal immune system requires tolerance to the residential microbiota and regulation to avoid overgrowth and invasion of internal tissues. Microbiota interactions with the innate and adaptive immune systems contribute to the establishment of an optimal equilibrium, whereas disturbances can lead to dysbiosis [27] and disease states through the development of intestinal inflammation [28,26]. This is reflected on the host side by genetic predispositions to IBD, which point to the importance of the immune system and microbial sensing. For example, mutations in human genes NOD2, ATG16L1 and those encoding defensins are known predispositions to IBD [28,26].

The homoeostasis of the GIT is also affected by the presence of eukaryotic parasites. Many parasitic protists, for example, Entamoeba, Giardia and Blastocystis, and parasitic animals, such as helminths worms, have specialized in colonizing the human gut, thereby affecting the GIT immune status. Parasitic protist infestations may be contributing factors to increased incidents of IBD and other chronic diseases [29]. In the case of helminths, human GIT interactions can positively influence immune homoeostasis (reflecting host–parasite adaptations) by modulating the immune system towards an optimal anti-inflammatory status [30]. Noticeable changes over the past 50 years in antibiotic use, hygiene (especially the eradication of helminths in most of the Western world) and diet (increase in proportions of meat and animal fat) are all potential contributing factors to disturb this balance established over millennia. The hygiene hypothesis suggests that the rapidly changing human lifestyle throughout the Western world has resulted in predisposition to multiple prevalent diseases due to imbalances in immune-microbiota coupling [31]. A better understanding of the contribution of the different partners in dysbiosis, and the cross-talk taking place between them, represents important opportunities to develop new strategies to treat chronic diseases, such as IBD and asthma [30,31].

While the natural variation of the human microbiota has yet to be fully determined, significant changes in gut microbial communities have been associated with several diseases, including type II diabetes [32], obesity [33], fatty liver disease [34], irritable bowel syndrome (IBS) [35] and IBD [36]. Obesity studies in animal models involving reciprocal transplants of faecal contents from obese to germ-free mice confirm the role of gut microbiota in controlling body weight and energy homoeostasis [37]. Furthermore, animal obesity models support linkages between GIT microbiota dysbiosis, low-grade inflammation and diet-induced type II diabetes [38]. In human patients with infective endocarditis, treatment with vancomycin has shown an increased weight gain, which is thought to be caused by dysbiosis due to colonization of Lactobacillus sp., a bacterium intrinsically resistant to this antibiotic [39]. Clinically, antibiotic usage has also been associated with increased incidence of Crohn’s disease and ulcerative colitis in both adults [40] and children [41]. The status of gut microbiome is increasingly being studied in a wide variety of further clinical conditions such as post small bowel transplantation [42], colorectal cancers [43] and malnutrition–immunity imbalances [44]. Other areas of interest are the potential roles of the GIT microbiota in diseases of distal organs, including asthma [45] and behavioural disorders [46]. For these diseases and others, the GIT microbiome is an intriguing therapeutic target.
because of its potential to dramatically expand the ‘human’ druggable genome, presently estimated to be around 20% of the human proteome [47,48]. Here, a key contribution can be made through computational approaches to link specific changes in microbial communities with physiological changes in the host that reflect disease phenotypes.

 Besides being a potential therapeutic target for chronic disease, the vast metabolic potential of the gut microbiota also plays a key role in the metabolism or biotransformation of xenobiotics, including many marketed drugs. More than 30 drugs are known substrates of bacterial enzymes in the GIT [49], which can have considerable impact on drug development. A tragic example is the reported deaths of several patients co-prescribed a new antiviral drug, sorivudine, along with an oral 5-fluorouracil, which were attributed to secondary drug metabolites generated by gut microflora [50]. Selective inhibition of gut microbiota enzymes could be a potential strategy to improve drug efficacy and safety. As an example, the colon cancer chemotherapeutic CPT-11 has a dose-limiting side effect of severe diarrhoea caused by reactivation of the prodrug by gut bacteria producing the enzyme β-glucuronidase. In rodent models, Wallace et al. [51] introduced an inhibitor of bacterial β-glucuronidase, which allowed for higher CPT-11 dosing [51]. Interestingly, β-glucuronidase is not essential for bacterial viability, so inhibition of this enzyme blocked the drug metabolism function while minimizing disturbance of the gut microbial community. The development of selective modulators of bacterial enzymes or species responsible for drug biotransformation could be an intriguing strategy for improving the clinical efficacy and safety profiles of particular drugs. The efficacies of many widely used drugs, including statins, are likely determined by both microbiota and host genetic factors [52], which could prompt integration of microbiome, human genetics and metabolomics into future personalized medicine initiatives [53,54].

**MODULATING THE MICROBIOTA**

Targeted therapeutics for manipulating the microbiota are still nascent and rudimentary. Prebiotics and probiotics are the most commonly marketed generic supplements for GIT disorders [55]. However, our understanding of their mechanisms of symptom relief is limited. Probiotics are usually supplements of single bacterial strains that integrate into the broader microbiota with limited global GIT effect [56,57], unless the microbiota is temporarily significantly depleted such as after antibiotic treatments [58]. Prebiotics, nutrients aimed at stimulating the growth of specific microbial species, have shown greater potential for manipulating the environmental pressures that shape the microbial ecology, especially in the developmental stages of life [59]. However, the effects of these supplements are short term and can be overshadowed by the overall diet [55]. It is widely accepted that a full complement of biodiversity is important for a healthy microbiota and single species supplements or substitutions have little effect on the long-term host microbiota phenotype [3,19]. Responsive probiotics are, as far as we are aware, the closest the field has come to targeted microbial therapies [60]. By introducing iron-responsive probiotics, the microbiota can be prepared for trauma and internal bleeding, such as prior to surgery.

Other more extreme therapies are intentional infections by parasitic helminths worn [61] and faecal transplants [62], which have been tested as alternatives to invasive surgery in IBD patients. In preliminary clinical trials, these approaches seem to somewhat regulate the host gut inflammatory response [62], although more thorough controls are desirable and the duration of relief is unclear.

Traditionally, antibiotics have been considered the most common treatment against infectious disease, microbial disorders and inflammation. However, growing evidence for microbial contribution to health and advanced understanding of complex microbial diseases have resulted in a re-evaluation of some antibiotic [63,64] and immunosuppressant treatments [65]. Antibiotics have been considered as a poor choice for GIT microbiota modification because of tolerance issues associated with long-term dosing and the lack of bacterial species specificity. However, antibiotics can positively modulate chronic disease conditions, such as diabetes and obesity, at least in rodent models [66]. Desirable pharmacological properties for a potential GIT microbiota modulator would be selective bacterial species activity and high bioavailability in the gut. Intriguingly, these are precisely the type of molecules that are considered to be failed candidates in antibiotic drug development [67]. The fact that >80% of the gut microbial species cannot be cultured using conventional laboratory methods restricts the use of high-throughput compound screening campaigns to discover anti-microbiota compounds [68]. However,
the development of in vitro human gut models [69] as well as using bacteria-phyllum-specific antibiotics, e.g. compounds effective against Gram-positive Firmicutes but not Gram-negative Proteobacteria, might accelerate the development of narrow spectrum drugs for microbiota modulation.

Another potential avenue for therapies is targeting host genes involved with microbiota cross talk. Our knowledge of human receptors engaged in the maintenance of the GIT microbial community balance is still evolving [71]. Toll-like receptors (TLRs) are responsible for cellular responses against bacterial infections, initiation of inflammation, production of antimicrobial peptides, maturation of antigen-presenting cells and activation of cellular repair and survival pathways [72]. TLR2 and TLR4 are primary sensors of pathogenic bacteria but are also important in maintaining microbiota homeostasis in the gut. Disruptions of TLR and nucleotide-binding oligomerization domain (NOD) pathways have been associated with colorectal cancer [72], IBD [73] and other intestinal diseases [27]. Microorganisms synthesize a wide range of low-molecular weight signalling molecules, many of which are similar to human- or eukaryotic-produced metabolites [74]. Using computational meta-analysis of data sources, such as human gene expression, bacterial metagenomics and metabolomics, molecules responsible for human-host bacterial crosstalk could be discovered and form the basis for future drug design.

COMPUTATIONAL APPROACHES TO TARGET MINING THE MICROBIOME

The ‘target’ for therapeutic intervention in the microbiome is a broad concept ranging from molecular entities, such as specific genes or proteins, to pathogenic species or dysbiosis [27]. Furthermore, the complex pathology often associated with microbiota-related diseases means multiple mechanisms can be linked and several different ‘targets’ can be isolated through studies of the same condition.

The international community is tackling human microbiome research through two complementary sequencing programmes. First, a large number of reference genomes from selected microbial taxa found in different human body sites are being cultured (where possible) and sequenced in their entirety [8,75]. Second, bacterial culture-independent metagenomic sequencing is being used to investigate natural microbial ecological adaptations and enterotypes that characterize a human anatomical site or condition. Having both a rich collection of reference genomes as well as broad metagenomic survey data from controlled population studies are crucial for determining microbiota and disease associations. Historically, clinical microbiology has relied upon culture-dependent methods to seek evidence in support of one of Koch’s key postulates [76] that one pathogen or pathogenic genetic feature results in a single disease disorder. However, a chronic disease might be the result of interactions involving subtle changes across many different microbiota species and communities. Thus, combined reference genome and culture-independent metagenomic analyses are necessary to study complex microbial ecologies at both the molecular and community levels in order to illuminate linkages between microbiome and specific disease phenotypes.

Various commercial next-generation sequencing (NGS) platforms have been used for metagenomics surveys. Several well-developed, publicly available software tools, such as QIIME [77] and MGrast [78], provide the means for identifying species and genes from metagenomic data as well as quantifying biodiversity both within and between samples. A thorough discussion of the usage and parameterization of primary analysis software tools (for assembly and initial annotation) is provided elsewhere [79]. Here we will concentrate on secondary computational analysis geared towards the characterization of microbiome phenotypes.

Reference genomes and comparative genomics

Determining the taxonomic content and gene function of a metagenomic data set primarily depends on the availability and annotation of reference genome sequences. Ideally, the reference genome catalogue for the human microbiota would provide an exact functional homologue of every metagenomic read in a data set. Therefore, given the associated ecological conditions, human anatomical regions, hygiene status and medical history, each mapped metagenomic sequence read could be assigned some physiological context and molecular function based on the functional annotation of the reference gene [70,80,81]. However, not all genes have complete functional annotations, and inferring function from sequence similarity is one of the major challenges in this field. Moreover, many genes lack functional
Metagenomics

The emergence of metagenomics facilitates the exploitation of gene set and taxonomic relationships by providing sequence information of the entire microbiome. Metagenomics is to identify features of the high-dimensional data set that characterize the microbiome and microbe-host phenotype. In metagenomic analyses, functional annotations allow us to study the taxonomic diversity and the functional profile between complex microbial communities. The main aim of comparative metagenomics is to identify features of the high-dimensional data set that characterize the microbiome and microbe-host phenotype. In metagenomic analyses, functional annotations allow us to study the taxonomic diversity and the functional profile between complex microbial communities.

Comparative analysis of metagenomes can be applied to identify similarities and differences in the taxonomic and functional profiles of different microbial communities. This can help to identify potential functional modules that are enriched in specific habitats or hosts, and to understand the role of these modules in the ecology and function of the microbiome.

Table 1: A survey of software for secondary and comparative analysis of metagenomics data sets

<table>
<thead>
<tr>
<th>Software feature</th>
<th>iPath2.0</th>
<th>Pathway projector</th>
<th>KEGG atlas (KAS)</th>
<th>metaSHARK</th>
<th>MEGAN v3</th>
<th>HUMAnN</th>
<th>CoMet</th>
<th>MetaStats</th>
<th>MetaPath</th>
<th>QIIME</th>
<th>LEfSe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Annotation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Taxonomic assignment</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Functional assignment</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Comparative analysis</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Visualization</td>
<td>KEGG atlas</td>
<td>KEGG atlas</td>
<td>KEGG atlas</td>
<td>ViewSHARK</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td></td>
</tr>
<tr>
<td>Computational analysis</td>
<td>Lightweight (cloud)</td>
<td>Lightweight (cloud)</td>
<td>Lightweight (cloud)</td>
<td>Middle weight (cloud)</td>
<td>Middle weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>Highly scalable KEGG custom</td>
<td>Highly scalable KEGG custom</td>
<td>Highly scalable KEGG custom</td>
<td>Scalable MSNR SEED KEGG eggNOG (M5NR)</td>
<td>Scalable MSNR SEED KEGG eggNOG (M5NR)</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td></td>
</tr>
<tr>
<td>Database reference</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: A survey of software for secondary and comparative analysis of metagenomics data sets

<table>
<thead>
<tr>
<th>Software feature</th>
<th>iPath2.0</th>
<th>Pathway projector</th>
<th>KEGG atlas (KAS)</th>
<th>metaSHARK</th>
<th>MEGAN v3</th>
<th>HUMAnN</th>
<th>CoMet</th>
<th>MetaStats</th>
<th>MetaPath</th>
<th>QIIME</th>
<th>LEfSe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Annotation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Taxonomic assignment</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Functional assignment</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Comparative analysis</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes'</td>
<td>Yes'</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Visualization</td>
<td>KEGG atlas</td>
<td>KEGG atlas</td>
<td>KEGG atlas</td>
<td>ViewSHARK</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td>Multiple custom</td>
<td></td>
</tr>
<tr>
<td>Computational analysis</td>
<td>Lightweight (cloud)</td>
<td>Lightweight (cloud)</td>
<td>Lightweight (cloud)</td>
<td>Middle weight (cloud)</td>
<td>Middle weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td>Heavy weight (cloud)</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>Highly scalable KEGG custom</td>
<td>Highly scalable KEGG custom</td>
<td>Highly scalable KEGG custom</td>
<td>Scalable MSNR SEED KEGG eggNOG (M5NR)</td>
<td>Scalable MSNR SEED KEGG eggNOG (M5NR)</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td>Scalable KEGG MetaCyc CAZy TCDB VFDB MEROPS</td>
<td></td>
</tr>
<tr>
<td>Database reference</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td>KEGG custom</td>
<td></td>
</tr>
</tbody>
</table>

*Given a BLAST output file to confirm assignments. *Not on the default, however, can be done with customized data sets. *Main databases used, however, not an exhaustive list.
bacterial and archaeal microbes as well as quantification of sequence abundances [85]. Such culture-independent methods of bacterial species analysis are especially important for environmental surveys where >99% of microbial species can be unculturable [83]. Taxonomic diversity can also be determined from shotgun sequenced metagenomics data sets using analysis pipelines [86] which perform SSU-rRNA sequence identification and phylogenetic tree reconstruction to infer taxonomy [87].

Initial studies of microbial communities in natural environments concentrated on the taxonomic diversity of species present and simple trends in their abundances. Using early versions of software packages, such as MGrast [78] and MEGAN [88], these studies demonstrated fluctuations in marker microbes over time, such as during fasting and feeding cycles [89,90]. However, due to the vast biodiversity of the gut, which contains an estimated 5000 microbial species, and high inter-individual variability it became clear that more refined analyses and advanced visualization were required [7].

Now community biodiversity is measured by using multiple phylogenetic metrics such as UniFrac distances, which is the relative phylogenetic branch distances of unshared species against both shared and unshared branches in the compared microbial data sets [91]. Microbial biodiversity is usually evaluated using a variety of metrics such as absolute, normalized and/or percentage abundances. Visualization of these biodiversity metrics can involve the use of heat maps, histograms and phylogenetic trees (Figure 4), depending upon the data curation and analysis goals. Recent developments include three-dimensional (3D) heat maps [92] and interactive phylogenetic representations of the biodiversity [93].

The problem however remains; high-dimensional data sets show numerous significant shifts in
biodiversity and multi-species trends, which are almost impossible to distinguish across large population studies by manual analysis. To solve this problem, statistical extraction is crucial and multivariate analysis is required to model biodiversity. Principal Coordinate Analysis (PCoA) is similar to principal component analysis (PCA) but performs better on community data that has patchy distribution of species, such as microbiomes. PCoA as implemented in MGraST [78] and QIIME [77] has successfully been used as an extension of biodiversity metrics to model comparative ecologies. This statistical method has enough power to help identify both an individual with forensic reliability [94] and recurrent trends from multiple population studies [7,19]. Figure 5 shows a PCoA comparison of microbiota biodiversity from a small population of obese and lean mice. A major limitation using microbiome sampling is that the single-sample approach does not reflect the dynamic nature of the microbiome over time and body site, and therefore models incorporate a large amount of noise. A recent population study addressed this issue by daily sampling of two individuals at four body sites over several months [90].

Figure 5: PCoA using unweighted UniFrac scores of nine mice gut microbiomes (five lean, red squares, and four obese, blue circles), a subset of data from Crawford et al. [125]. (A) Shows PCoA of the first and second principal components. (B) Shows the PCoA of the first- and third-principal components. (C) Shows PCoA of the second and third principal components. Figures generated in QIIME [77].
PCoA to cluster sample sets, they found pronounced variation in microbiota species composition over time, suggesting that the core microbiome is highly dynamic [90]. This type of analysis may help identify microbe–host interactions characterizing a phenotypic trait that could be important for disease.

Given the capacity to investigate the microbiota ecology and its variations across time, these analyses could be extended to identify specific biomarkers of disease. Metastats [95] is a package for the identification of differential representations of genes or taxa, based on two sample t-tests and false discovery rate, which produces P-values for differential abundance. LEfSe [96] is a further development of this approach to account for biological consistency in subcategorized data sets using linear discriminate analysis. LEfSe is capable of identifying biomarkers of specified conditions at multiple levels in data sets, such as taxa or functional data sets, and grade the biomarker by statistical significance. Figure 6 illustrates the identification and prioritization of biomarkers characteristic of the ulcerative colitis microbiota in mice, identifying variation in taxonomic composition that characterize both the healthy microbiota and the ulcerative colitis microbiota.

Functional diversity—What are they doing?
The dynamic nature of the GIT environment and microbial genomes means that the adaptation of the gut microbiota is rapid. Individual bacterium can acquire new genetic features through LGT that might complement, shift or replace existing metabolic functions. This has lead to the definition of the pan-genome for microbes [97], which is especially relevant in the highly complex environment of the gut [98,99]. However, metagenomic data with its typically low sequence coverage leading to short contigs and numerous singletons are not ideal to investigate LGT as the full-genomic context of a gene is required to investigate its origin in detail. As demonstrated in Figure 7, the relative abundance of phyla varies widely across individual samples. However, the relative abundance of major functional categories remains almost equal, which implies that gene functions are collectively redistributed across the microbiota in each individual. This adaptation allows the microbiota to fully encompass the environmental niche and ensure the fitness of the entire microbial population. These functional adaptations to environmental stimuli at a finer level are thought to represent the defining characteristics of the microbiota determining pathophysiology of disease. Hence, targeting functional variation at a finer granulation could identify important features corresponding to valuable drug targets, such as proteases and other enzymes degrading xenobiotics, biochemical stressors and toxins [100].
Functional diversity has been visualized at two levels of abstraction: metabolic pathway reconstruction and functional categorization (Figure 8). In metabolic pathway reconstruction, metagenomic reads are mapped onto known pathways, which can then be compared across classification data sets to give a microbial community, rather than species, level overview of metabolic capacity. Functional categorization uses layers of subcategories to view functional trends at different levels, which again can be compared across classification data sets. By comparing metagenomic sequences from a microbial community to reference genomes of a known isolation source, sets of genes and gene functions important to a particular habitat can be revealed [70]. Analysing the distribution of genes and gene functions across different habitats helps generate hypotheses for the level importance of a gene set in a particular environment.

Similar to taxonomic data sets, the differential representation of genes and gene functions can also be established using the software Metastats [95] and LEfSe [96]. The scale of gene representation is, however, approximately 3000 times greater than that of taxa; therefore, clustering of genes by functional pathways or categorical function can allow for more powerful analyses. MetaPath is customized for statistical identification of over-represented metabolic pathways by using a greedy search algorithm to iterate over Metastats’s differential abundance results [101]. Predictive relative metabolic turnover (PRMT) is a recent extension of this approach that takes into account the metabolic transformations of the pathways as well as sequence abundance. This can be used for approximate system modelling, thereby predicting metabolite usage based on metagenomics data sets [102]. Equally, this approach could be reversed to predict microbial function based on the metabolites present. This level of metagenomics–metabolomics coupling has the potential to be a powerful tool for modelling

Figure 7: Comparison of taxonomic variability against functional variability in six human microbiota samples, a subset of data Turnbaugh et al. [18]. (A) Shows phylum-level bacterial variability by percentage abundance. (B) Shows gene function variability across the major functional categories by percentage abundance.
the host–microbiota metabolite systems in disease data sets.

However, system modelling is highly dependent on robust assignments of gene function across any particular microbiome, which is a serious limitation at present. In one study, an average of 25% of all predicted open reading frames (ORFs) from human gut metagenomic data were shown to have no significant similarity to any protein-coding sequences known from existing complete genomes [103]. Similarly, it was shown that >50% of ORFs identified in human gut metagenomic samples could not

Figure 8: Visualization of functional metagenomics across six mice microbiota (three lean, one overweight and two obese), a subset of data from Turnbaugh et al. [18] (A) Shows comparative abundance of the major functional categories (abundance has been normalized to compensate for data set sizes). (B) Shows comparative metabolic pathways mapped onto the KEGG atlas in the background; purple lines indicate metabolic pathways present in both populations, blue lines indicate metabolic pathways only present in the lean population and red lines indicate metabolic pathways present in only the overweight and obese population. Figure generated in MGrast [78].

be assigned to any known conserved functional region in the Pfam database [104]. In a study of perhaps one of the largest human population to date, it was shown that around a third of microbiome sequence reads were unmapped to any known genes and another third mapped to hypothetical genes [7]. Furthermore, the functional annotations of existing proteins databases can be inaccurate and inconsistent [105]. This highlights the limitations of functional inference when only a section of the data sets have some known molecular assignment; hence, boosting the taxonomic coverage of reference genomes and enriching their functional annotations are primary and fundamental tasks underpinning fruitful metagenomic analysis [8].

In addition to relying on pair-wise sequence similarity to reference genome sequences, profile-based approaches to identify known conserved functional regions in various protein databases, such as Pfam, Interpro and PRIAM, can also be used to assign potential functions to metagenomic reads. Such analyses include highly sensitive profile–profile-based searches [106,107] used to investigate divergent protein domains or families identified from metagenomic data [70] that otherwise reveal no similarity in pair-wise comparison or simple profile analyses. Notably, protein domain databases are originally derived from the analyses of reference genomes and are continually updated as new sequences from both complete genome and metagenomic sequences become available [70,108,109]. Thus, periodic re-analysis of published metagenomic data sets could illuminate new bacterial species or gene functions in those study conditions as reference genome and protein databases mature over time.

Most metagenomics analysis software packages provide references to multiple databases for annotation. For example, the CoMet web server [110] is an easy-to-use comparative metagenomic package for analysing such data sets. The tool provides a complete pipeline for ORF finding, functional annotation (Pfam, FIGfam and Gene Ontology) and comparative statistical analysis of multiple metagenomic data. However, the majority of these databases are only sparsely populated with environment-specific genes, and the divergent nature of gut microbes provides significant challenges in mapping reads to known-function genes.

Comparative studies of the microbiome typically involve a large number of protein sequences and thus will require a significant amount of computing power. The tremendous amount of sequence data from reference genomes (across all habitats) and metagenomic surveys of the human microbiome necessitates the development of efficient tools for comparative sequence analysis. Several high-throughput computational frameworks employing Grid and Cloud computing have been developed, e.g., QIIME [77] and Microbase [111], with an aim to reduce the amount of time taken to complete compute intensive analysis tasks on several million protein sequences. The bioinformatics workflows developed using high-throughput computing frameworks have already been shown to be satisfactory for facilitating post-genomics studies that require actions to be performed in a systematic and automated fashion [112]. Further development of high-throughput computational bioinformatics workflows will continue to be important given the increasing complexity of metagenomic human microbiome data. An alternative and complementary approach is to implement speed-optimized sequence analyses tools, which will also enable more efficient metagenome analyses [110].

MOVING BEYOND SEQUENCE DATA
Beyond microbial genomics and metagenomics, other studies have investigated human–host genomics, epigenetics, proteomics, and metabolomics as well as microbiota metabolomics [113,114]. Strategies to optimize the integration of the microbiome, host-derived data sets and metabolomics to investigate associations of host–microbe genetic traits with disease conditions will allow a more global understanding of host–microbiota interactions in human health and disease [114]. These multi-dimensional human–host–microbial data sets will allow us to investigate the microbiome-encoded metabolic pathways and their relationships with human physiological and environmental variables [114]. This implies a predictive nature between the parameters and could allow, in the future, a wider inference from simple sample analysis to extrapolate much broader fields. For example, measurements from multiple areas of the intestine and host tissues coupled with metagenomics sampling and predictive models could be used to develop detailed behavioural analysis of microbial trends and phenotypic behaviour in response to stimuli such as drugs. Ultimately, we would expect this type of analysis
to highlight multiple targets for intervention which can be matched up with the compound repositories in silico to search for drug repurposing opportunities [115].

DATA INTEGRATION APPROACHES

For certain diseases, there has been a great deal of activity in the development of tools for the integrative analysis of therapeutics and their targets [115,116]. Integration across the previously described data sets related to microbiome effects on human health will be especially challenging yet critical for the advancement of microbiome targeted therapeutics. Effective strategies for controlling and manipulating our gut microbiota could benefit enormously by taking into account temporal changes in human and microbiome gene expression that could be correlated with other factors such as the host’s immune status and genotype as well as environmental conditions. Thus, we can expect to see integrative microbiome models that include information derived from global strategies for measuring gene expression such as RNAseq. These models may, in turn, lead to the identification of biomarkers and expression profiles that are indicative of the microbiota status in terms of disease involvement and responsiveness to therapeutic agents.

Data integration, not a new problem in bioinformatics, is already an active research area in the field of molecular and medical informatics. Software specific for the analytic integration of various data relating to the microbiome are now starting to appear. Some tools such as HUMAnN [117] already partially address this challenge by integrating functional and taxonomic analysis of microbiota. The Integrated Microbial Genomes (IMG) system also provides a Web portal that facilitates the analysis of genomic sequence in an integrative environment [118].

Conventional approaches such as data warehouses coupled with tools for multi-dimensional statistical analyses still have the potential to provide rich integrative environments that will allow biologists to explore the interdependencies between many factors simultaneously. More recently, graph-based approaches to representing biological data have started to appear and are valuable for the systems-level analysis [119,120]. These graphs can be of different levels of abstraction from functional networks [121] to highly annotated and computationally amenable semantic graphs (Ondex) [120]. In the future, dynamic systems modelling may play a large part in helping us to understand the complexity of the microbiome.

Efforts to integrate data necessary for microbiome analysis may start to reveal deficiencies in upstream processes such as DNA sequence annotation. For example, the clinical status of a patient from where a particular microbiome was sampled might be used as an annotation. This would require not only new sequence annotation formats but also careful blinding of patient data in order to maintain confidentiality. As such, we may find that new standards are required to capture the metadata required for microbiome analysis. Moreover, these types of data may need to address other components of a microbiome sample. For example, in addition to the microorganisms, the metabolite profile linked with a microbiome sample will also need to be measured, described and stored in a database in a form amenable to integration [122].

CONCLUDING REMARKS

The human microbiota is a major consideration in sustaining health and fighting disease. The potential therapeutic implications of modulating the microbiota are enormous; however, this remains a young field of science and drug discovery. As research in this field expands, data sources will become more mature and the complex nature of the human microbiota interaction will become better defined. Our understanding of the host–microbiome super-organism may develop to the point where microbial-targeted therapies are a major consideration in pharmaceuticals and personalized medicine. The computational infrastructure required to support this growing field will involve solutions based on a combination of massively parallel data handling, refined data-mining strategies and customized platforms for data integration.

Future advancements in microbiota-directed therapeutics will depend upon partnerships between academia and industry. There is a growing realization in the pharmaceutical industry that leveraging new knowledge from ever more complex biomedical data sets will require creative pre-competitive initiatives to enhance public data sources [123]. Certainly, there will be significant opportunities to develop such public–private partnerships around the microbiome and human
health. Deeper understanding of the variation in microbiota in healthy populations as well as developing more solid evidence linking microbiota status with disease phenotypes will mainly come from public-funded research programmes. The discovery and development of targeted therapeutics, especially small-molecule-based drugs, and the sponsoring of clinical trials will require industrial-level resources. Thus, computational biology and bioinformatics will also serve as an important bridge between these public and private efforts to exploit the human microbiome for the benefit of human health.

**FUNDING**

EPSRC grant (EP/G037620/1) for the Engineering Doctorate in Biopharmaceutical Process Development (to M.C.).

**Acknowledgements**

The authors wish to thank GlaxoSmithKline and Newcastle University for providing technical expertise and funding which enabled this research. The authors would also like to thank the four anonymous reviewers for their valuable and detailed feedback on this manuscript.

**References**


