MICROBIOME

Metagenome-wide association studies: fine-mining the microbiome

Jun Wang[1,2] and Huijue Jia[2]

Abstract | Metagenome-wide association studies (MWAS) have enabled the high-resolution investigation of associations between the human microbiome and several complex diseases, including type 2 diabetes, obesity, liver cirrhosis, colorectal cancer and rheumatoid arthritis. The associations that can be identified by MWAS are not limited to the identification of taxa that are more or less abundant, as is the case with taxonomic approaches, but additionally include the identification of microbial functions that are enriched or depleted. In this Review, we summarize recent findings from MWAS and discuss how these findings might inform the prevention, diagnosis and treatment of human disease in the future. Furthermore, we highlight the need to better characterize the biology of many of the bacteria that are found in the human microbiota as an essential step in understanding how bacterial strains that have been identified by MWAS are associated with disease.

Microbiome
The ensemble of microbial genomes and products at a given site.

Microbiota
The ecological community of microorganisms at a given site.

16S rRNA gene amplicon sequencing
Amplification and sequencing of the variable regions in 16S ribosomal RNA genes for the taxonomic profiling of bacteria and archaea in a sample.

[1]*iCarbonX, Shahe Industrial Zone, No.4018 Qiaoxiang Road, Nanshan District, Shenzhen 518083, China.*
[2]*Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Shenzhen 518083, China.*

wangjun@icarbonx.com; jiahuijue@genomics.cn

doi:10.1038/nrmicro.2016.83
Published online 11 Jul 2016

Metagenomic shotgun sequencing enables the genetic information of the entire community of microorganisms at a given site to be surveyed. The human microbiome, in particular the human gut microbiome, far exceeds the host in number of genes[1–3] and is under intense investigation because of its central role in metabolism and immune modulation[4–6]. For the most part, associations between the microbiota and human disease have been studied using 16S rRNA gene amplicon sequencing[4]. These studies have suggested that dysbiosis of the gut microbiota may be a key environmental risk factor for many human diseases, although the severity of the dysbiosis varies according to the disease. However, the data produced by 16S rRNA gene amplicon sequencing are of limited use in several respects, owing to a poor taxonomic resolution and an absence of information about the function of the microbiome. Metagenomic shotgun sequencing, in which the full complement of genes that are present in the microbiome are sequenced rather than just a single taxonomic marker gene, is able to overcome these limitations by providing information about the abundances of genes in functional pathways and at all taxonomical levels. However, until recently, metagenomic shotgun sequencing had been prohibitively expensive and the data generated had been challenging to analyse, owing to the substantial complexity compared to data generated by 16S rRNA gene amplicon sequencing.

Continued reduction in sequencing costs, developments in bioinformatic tools and the accumulation of functional knowledge has now made metagenomic shotgun sequencing more accessible as a tool to study the human microbiome. Accordingly, several recent studies have used metagenomic shotgun sequencing to survey the microbiome, as represented by metagenomic data, for associations with disease. The study designs used for these metagenome-wide association studies (MWAS) have largely been modelled on genome-wide association studies (GWAS) that identify genetic variants in the human population that are associated with a phenotype, often a disease. In the current forms of MWAS, the relative abundance of a gene in a metagenome is used to establish an association with a disease of interest (BOX 1; FIG. 1), typically after the genes are first grouped into strain-level clusters known as metagenomic linkage groups (MLGs), metagenomic clusters (MGCs) or metagenomic species (MGS), which reduces the dimensionality of the data (BOX 1; FIG. 1e; TABLE 1).

In this Review article, we summarize recent findings from MWAS that have shown associations between the microbiome and diseases such as type 2 diabetes, obesity, colorectal cancer and rheumatoid arthritis. Furthermore, we discuss strategies for verifying the associations that have been identified by MWAS and establishing whether these associations represent causal relationships in which components of the microbiome contribute to disease. Finally, we consider how MWAS might be used in the future to shed further light on the aetiology of disease and to inform the development of preventive and therapeutic interventions.

Microbiome–disease associations

Type 2 diabetes. The incidence of type 2 diabetes, a metabolic disorder that is characterized by hyperglycaemia and insulin resistance, has been increasing rapidly over the past few decades — for example, in China, more than 10% of adults are estimated to be affected by the disease[7]. Although some genetic risk factors for type 2 diabetes have been identified, they account for only a small portion of the disease risk and, as factors that do not vary significantly with time, cannot explain the rapid rise in the incidence of the disease. As the gut microbiome has recently been shown to affect host physiology[4,5], it has been proposed to be an environmental factor that contributes to the risk of developing type 2 diabetes. Indeed, before MWAS, PCR analysis and 16S rRNA gene amplicon sequencing of the gut microbiota of a small cohort of individuals with type 2 diabetes identified decreased levels of taxa from the phylum Firmicutes and the class Clostridia compared with controls[8].

Type 2 diabetes was the first disease for which a suspected association with the microbiota was studied by MWAS[9]. In this first MWAS, metagenomic sequencing of hundreds of stool samples enabled the identification of genes from the gut microbiota that were differentially

Box 1 | Designing and carrying out an MWAS

Metagenomic sequencing and assembly

An important requirement in the design of metagenome-wide association studies (MWAS) is that a sufficient volume of sequencing data is obtained to enable reliable quantitative comparisons between samples, as the number of genes that are detected in any given sample increases with the volume of sequencing data until saturation. Obtaining sufficient data is especially challenging for samples from the gut mucosa, mouth, skin, vagina and placenta, which can be dominated by sequencing reads from the host, unlike faecal samples, in which sequencing reads from the host account for no more than 1% of the total[3,105,106]. Depending on the degree of microbial diversity in a sample, even the several Gb of raw sequence data commonly used for faecal samples may be insufficient for MWAS that use non-faecal samples. However, the experimental removal of host DNA without affecting the microbial content of the samples remains a challenging task, although bioinformatics tools can be used to remove host sequences after sequencing. Following quality control, the sequencing reads are *de novo* assembled into a set of contigs that together comprise the metagenome. The development of high-throughput sequencing methods that produce longer reads is expected to improve the sequencing, assembly and analyses of metagenomic samples.

Microbial reference gene catalogues

A non-redundant gene catalogue can be constructed from a metagenome assembly by predicting genes from the assembled contigs and removing highly similar genes across samples, which are considered to be redundant[1–3,9,107]. The abundances of genes, taxa and functions in metagenomic datasets can then be quantified by alignment to such a gene catalogue (or an existing microbial reference gene catalogue) to identify associations with a disease of interest. We currently recommend that a high-quality reference gene catalogue, such as that produced for the gut microbiome[3], is used as a starting point for MWAS and as a basis for comparisons between diseases, populations or individual animals[1,3,9,13,18] (FIG. 1d). It should be noted that constructing a gene catalogue from a poor-quality metagenome assembly can artificially increase the number of genes[3]. For human faecal samples, the quality of existing reference gene catalogues is high, owing to the volume of metagenomic data that has been generated from this body site and continued efforts to improve assembly[3,45] (TABLE 1). However, for body sites other than the gut, metagenomic information remains very limited[2,45,106,108] (FIG. 1b).

Taxonomy from metagenomic data

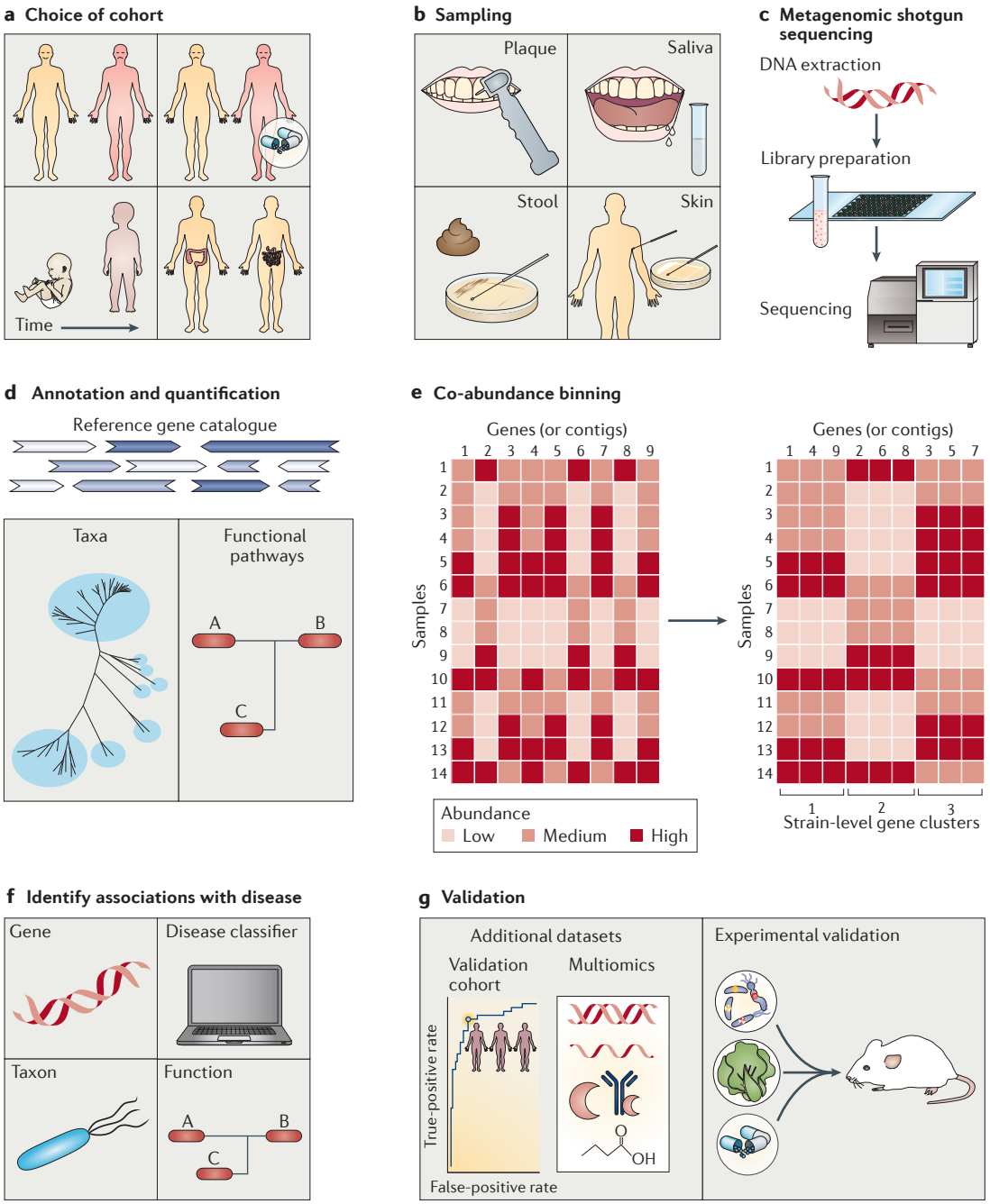
One goal of an MWAS may be to identify associations between the disease that is being investigated and specific taxa. To increase taxonomic resolution to the level of strains, genes in a metagenomic dataset can be clustered according to the genome of origin. This approach is based on the idea that genes from the same microbial genome are physically linked and, as such, should have the same pattern of abundance variation to one another across many samples (FIG. 1c). Using different correlation coefficients and computational algorithms, methods have been developed to organize microbial genes in metagenomic data into strain-level clusters known as, depending on the method used, metagenomic linkage groups (MLG)[9], metagenomic clusters (MGC)[13] or metagenomic species (MGS)[18,93]. Sequence alignment to existing microbial genome sequences are typically used to assign genes or strain-level clusters to taxa[3,109], although it should be noted that the annotation criteria have not been unified for MLG, MGC and MGS[9,13,18,19,51,93,106]. The alignment of conserved single-copy genes and strain-specific regions of the genome may be more useful than other loci for taxonomic annotation[96,110–113]. For example, the metagenomic operational taxonomic unit (mOTU) method, which assigns taxonomic annotations based on 10 conserved single-copy genes, was shown to be more accurate than the use of the 16S rRNA gene for species assignment[110]. Incorporation of additional information, such as GC content and tetra-nucleotide frequency, might help to separate clusters that could not be resolved using abundance variation, which has been a particular problem in datasets with a small number of samples[53,92].

Controlling for phenotypes in MWAS

MWAS should ideally include extensive metadata that enable factors that influence the microbiota to be controlled for. Cohorts should then be matched according to these metadata rather than using statistical regression methods that would undermine the power of MWAS to control for confounding factors. For colorectal cancer and type 2 diabetes, the disease signal (that is, the effect size) in MWAS of faecal samples is sufficiently strong to be detectable over the background variation that could be caused by other factors. However, for associations that are identified by MWAS that have a small effect size, a larger sample size might be required to distinguish the association from background variation.

Dysbiosis
An imbalance of the microbiota at a body site that is caused by an overgrowth of pathogenic microorganisms or a lack of commensal micoorganisms.

Contigs
Contiguous DNA sequences that are assembled from shorter, overlapping sequencing reads.



abundant between individuals with type 2 diabetes and controls. The study followed a two-stage procedure similar to that routinely used in GWAS, in which an initial set of genes was identified in a discovery cohort and then filtered by validation in a verification cohort (TABLE 1). The validated genes were then clustered into MLGs (each with at least 100 genes) according to co-variations in gene abundance across the samples (BOX 1; FIG. 1e). Specifically, MLGs that were annotated as butyrate-producing bacteria, including Clostridiales sp. SS3/4, *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and *Roseburia inulinivorans*, were depleted in samples from patients with type 2 diabetes, whereas MLGs that corresponded to *Bacteroides* sp. 20_3, *Clostridium hathewayi*, *Clostridium ramosum*, *Clostridium symbiosum* and *Eggerthella lenta* were enriched in these samples. Functional gene analysis of the metagenomic data using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database suggested that the gut microbiome in the cohort with type 2 diabetes was enriched for genes that function in the membrane transport of sugars, the transport of branched-chain amino acids (BCAAs), methane metabolism, degradation and metabolism of xenobiotics, sulfate reduction (hydrogen sulfide biosynthesis) and resistance to oxidative stress. By contrast, there was a relative depletion in genes that are associated with functions in bacterial chemotaxis, flagellar assembly, butyrate biosynthesis and the metabolism of cofactors and vitamins (FIG. 2).

◀ Figure 1 | **Identifying associations using MWAS.** Although metagenome-wide association studies (MWAS) can, in principle, be used to study associations between the microbiome and any trait, studies to date have focused on identifying associations between the microbiome and disease. **a** | A typical cohort to be studied by MWAS would include a group of healthy individuals (top left panel, yellow) and a group of individuals with a disease (top left panel, red). However, MWAS can also be used to compare the microbiomes of individuals in a longitudinal study: before and after a certain intervention, such as a drug treatment (top right panel) or dietary intervention (not shown); or in a natural process, such as the development of an infant (bottom left panel) or the progression of a disease (not shown). Finally, an MWAS may be designed to compare the microbiomes at different body sites for a cohort of individuals with a disease (bottom right panel). **b** | The microbiomes of samples that are taken from different body sites, such as oral plaque, saliva, stool (representing the gut microbiome) or skin, can be studied by MWAS. **c** | DNA extraction, library preparation and metagenomic shotgun sequencing of the samples generates a dataset of sequencing reads. Bioinformatics tools (not shown) are used to assemble the metagenomic reads into contigs. **d** | Genes that are predicted from contigs are compiled into a gene catalogue, or an existing reference gene catalogue that is representative of the data could be readily used. The relative abundance of a gene can be quantified by determining the number of sequencing reads that align to that gene in the reference catalogue. Furthermore, phylogenetic or functional annotation and grouping of the predicted genes allows the quantification of microbial taxa or functional pathways in the samples and comparisons between samples. **e** | Genes (or contigs, which can contain several genes and intergenic regions) that have abundances that co-vary in samples can be clustered into strain-level taxonomic units (known, according to the clustering algorithm used, as metagenomic linkage groups (MLGs), metagenomic clusters (MGCs) or metagenomic species (MGSs). **f** | Associations with a disease can be identified for individual microbial genes, taxa or functions. In addition, classifiers can be constructed using supervised machine learning to assign each sample to a certain category, such as healthy or diseased. **g** | Associations that are identified by MWAS can be validated using additional metagenomic datasets, such as samples from additional cohorts or timepoints, or using other forms of omics data. For studies that seek to identify causal relationships between a disease and the microbiome, associations that are identified by MWAS can be used to suggest hypotheses for further investigation by animal models. These experiments may involve the microbial transplant of specific species or sets of species, and/or the study of the response of the microbiome to dietary changes or drug treatment.

The depletion of butyrate-producing bacterial strains and butyrate biosynthesis genes in the gut microbiomes of individuals with type 2 diabetes may relate to the ability of butyrate to increase the secretion of glucagon-like peptide 1 (GLP1) and peptide YY, as the functions of these peptides include the promotion of intestinal gluconeogenesis, which leads to better control of glucose and energy homeostasis[10,11]. Furthermore, short-chain fatty acids (SCFAs), including butyrate, have recently been shown to limit autoimmune diabetes by controlling the expression of an immunomodulatory antimicrobial peptide by pancreatic β-cells, which demonstrates a direct influence of the gut microbiota on the pancreas[12].

Corroborating the MWAS findings from this first study[9], which analysed Chinese men and women, the second MWAS of type 2 diabetes, which analysed European women (mostly from Sweden), identified genes for membrane transporters and oxidative stress resistance as over-represented in the gut microbiomes of individuals with type 2 diabetes, whereas genes for flagellar assembly and riboflavin (vitamin B_2) metabolism were depleted in the gut microbiomes of individuals with type 2 diabetes[13] (TABLE 1). Several clostridial species were enriched in individuals with type 2 diabetes and *Roseburia*_272 was depleted in individuals with type 2 diabetes in both the Chinese and the European cohorts[9,13], which suggests common features of the gut microbiome in individuals with type 2 diabetes from the two cohorts.

In the European cohort, correlations were reported between the abundances of specific members of the gut microbiome and clinical indices that are related to type 2 diabetes. Notably, the abundance of *C. clostridioforme* positively correlated with the levels of triglycerides and C-peptide, whereas the abundance of *Lactobacillus gasseri* positively correlated with fasting blood glucose and glycated haemoglobin (HbA1c)[13]. By contrast, clostridial MGCs that were depleted in individuals with type 2 diabetes negatively correlated with the levels of C-peptide, insulin and triglycerides[13].

Interestingly, individuals with type 2 diabetes that were treated with the drug metformin had decreased levels of *Clostridium* spp. and *Eubacterium* spp., and increased levels of Enterobacteriaceae, compared with individuals with type 2 diabetes that were not treated with metformin[13]. Similar observations were made in rats with insulin resistance and obesity that was induced by a diet high in fat and sucrose[14], and in a study that combined a Danish cohort with the two previous cohorts, which also found that individuals who were treated with metformin had decreased levels of *Intestinibacter*, which is a newly defined genus that used to be included in the *Clostridium* genus[15] (FIG. 2; TABLE 1). Among members of the Enterobacteriaceae, the abundance of *Escherichia coli* was shown to be correlated with the levels of GLP1 in individuals who were treated with metformin[13]. Studies using mouse models have also linked the therapeutic effects of metformin to changes in the microbiota[16,17], and future MWAS of type 2 diabetes may help to identify other changes in the microbiota that are associated with metformin treatment in humans.

To establish whether the gut microbiome provided sufficient information to distinguish between healthy individuals and individuals with type 2 diabetes, a disease classifier was constructed through supervised machine learning of the data that were obtained from the Chinese cohort. The classifier selected 50 gut microbial genes (TABLE 1), which enabled 345 samples from the type 2 diabetes and control cohorts to be classified with an area under the receiver operating characteristic curve (AUC) of 0.81 (REF. 9). When applied to an additional set of samples, an index calculated from the abundances of the 50 genes showed a significant difference between 11 individuals with type 2 diabetes and 12 healthy individuals. These results suggested for the first time that the gut microbiome could be used to distinguish between samples from healthy individuals and individuals with a disease, in this case type 2 diabetes.

Obesity. Another metabolic disease to be studied by MWAS is obesity, which has been studied in cohorts from Denmark and France[18,19] (TABLE 1). Previously, 16S rRNA gene amplicon sequencing had identified a lower gut microbial diversity in obese individuals, in addition to a higher Firmicutes-to-Bacteroidetes ratio[20,21]. In the cohort from Denmark, which was stratified into

Short-chain fatty acids
(SCFAs). Fatty acids that have fewer than six carbon atoms. In the context of the microbiome, SCFAs usually refer to acetate, propionate and butyrate, which are produced by various species of bacteria.

Metformin
A biguanide drug that is commonly prescribed as a treatment for type 2 diabetes.

Supervised machine learning
Machine learning in which the training data are labelled (for example, as cases or controls). Using the training data, the algorithm learns to classify new data according to these labels.

Table 1 | **Diseases studied by MWAS**

Cohort size (body site; country of residence)	Gene catalogue size	Sequencing reads mapped	Clustering method	Examples of key taxa	Examples of key functions	Disease classifier features	Feature selection method	Refs
Type 2 diabetes (T2D)								
71 individuals with T2D and 74 controls for stage I; 100 individuals with T2D and 100 controls for stage II; 11 individuals with T2D and 12 controls for validation (Gut; China)	4.3 million genes	77% for stage I, 72% for stage II	MLG	*Bacteroides* sp. 20_3 and *Clostridium hathewayi* in individuals with T2D; *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and Clostridiales sp. SS3/4 in controls	Membrane transport of sugars and BCAA transport in individuals with T2D; bacterial chemotaxis, butyrate biosynthesis and metabolism of cofactors and vitamins in controls	50 genes	mRMR	9
53 individuals with T2D, 49 individuals with IGT and 43 controls (Gut; Sweden)	2,382 reference genomes	Not mentioned	MGC	*Clostridium clostridioforme* in individuals with T2D; *Roseburia*_272 in controls	Membrane transporters and oxidative stress resistance in individuals with T2D; flagellar assembly and riboflavin metabolism in controls	50 MGCs	Random forest	13
75 individuals with T2D, 31 individuals with T1D, 277 controls published by the MetaHIT consortium and 461 published samples from the Chinese and Swedish studies; 30 individuals with T2D for validation (Gut; Denmark)	Not mentioned	Not mentioned	MGS, mOTU	*Clostridium bolteae* and *Parabacteroides distasonis* in untreated individuals with T2D; *Escherichia coli* in individuals with T2D who were treated with metformin; *Roseburia* spp., *Subdoligranulum* spp. and Clostridiales spp. in controls	Production of butyrate, propionate, LPS and H_2 in individuals with T2D who were treated with metformin	1 genus for metformin-treated individuals with T2D versus controls; 63 genera for untreated individuals with T2D versus controls	SVM	15
Atherosclerosis								
12 individuals with atherosclerosis and 13 controls (Gut; Sweden)	2,382 reference genomes	28%	None	*Collinsella* spp. in individuals with atherosclerosis; *Roseburia* spp. and *Eubacterium* spp. in controls	Peptidoglycan synthesis in individuals with atherosclerosis; phytoene dehydrogenase in controls	None	None	49
Obesity								
169 individuals who are classified as obese and 123 controls (Gut; Denmark)	3.3 million genes	58%	MGS-like	*Bacteroides* spp. and *Ruminococcus gnavus* in samples with low gene counts; *F. prausnitzii*, *Butyrivibrio* spp. and *R. inulinivorans* in samples with high gene counts	Degradation of β-glucuronide and aromatic amino acids in samples with low gene counts; production of organic acids and H_2 in samples with high gene counts	4 species for samples with low versus high gene counts; 9 species for obese versus non-obese individuals	Enumeration	18
38 individuals who are classified as obese and 11 overweight individuals (Gut; France)	3.3 million genes	57%	MGS-like	Decrease in *Eubacterium rectale* and *Bifidobacterium* spp. during the calorie-restriction phase	Not mentioned	6 species	Enumeration	19

Table 1 (cont.) | **Diseases studied by MWAS**

Cohort size (body site; country of residence)	Gene catalogue size	Sequencing reads mapped	Clustering method	Examples of key taxa	Examples of key functions	Disease classifier features	Feature selection method	Refs
Liver cirrhosis								
98 individuals with liver cirrhosis and 83 controls (Gut; China)	2.7 million genes	42%	MGS	*Streptococcus anginosus* and *Veillonella atypica* in individuals with liver cirrhosis; *Faecalibacterium prausnitzii* and *Coprococcus comes* in controls	Assimilation or dissimilation of nitrate to or from ammonia, and denitrification in individuals with liver cirrhosis; histidine metabolism, ornithine biosynthesis and carbohydrate metabolism in controls	15 genes	mRMR	51
Colorectal cancer (CRC)								
53 individuals with CRC, 42 individuals with adenoma and 61 controls; 38 individuals with CRC, 5 new and 292 published controls for validation (Gut; France)	Not mentioned	Not mentioned	mOTU (single-copy genes)	*Fusobacterium* spp., *Porphyromonas asaccharolytica* and *Peptostreptococcus stomatis* in individuals with CRC; *Eubacterium eligens* in controls	Host cell wall carbohydrates in individuals with CRC; fibre-degrading enzymes and fibre-binding domains in healthy controls	22 species	LASSO	31
41 individuals with CRC, 42 individuals with adenoma and 55 controls; 5 individuals with CRC, 5 individuals with adenoma and 8 controls for validation (Gut; Austria)	3.5 million genes	76%	MLG	*Bacteroides* spp., *Fusobacterium* spp. and *Peptostreptococcus stomatis* in individuals with CRC; *Bifidobacterium animalis* in controls	Utilization of amino acids and host glycans in individuals with CRC; metabolism of sugars and dietary fibre in controls	15 MLGs for controls versus individuals with carcinoma; 10 MLGs for controls versus individuals with advanced adenoma	Random forest	32
74 individuals with CRC and 54 controls; 16 individuals with CRC and 24 controls for validation; 47 individuals with CRC and 109 controls for qPCR (Gut; Hong Kong and Denmark)	4.3 million genes	67%	MLG, mOTU	*Fusobacterium nucleatum*, *Peptostreptococcus stomatis*, *Parvimonas micra*, *Solobacterium moorei* in individuals with CRC; *Eubacterium ventriosum* in controls	Leucine degradation and guanine nucleotide biosynthesis in individuals with CRC	2 genes	None	33
Rheumatoid arthritis								
77 individuals with rheumatoid arthritis, 80 controls and 40 after- treatment samples; 17 individuals with rheumatoid arthritis and 17 controls for classifier verification (Gut; China)	5.9 million genes	80%	MLG	*Lactobacillus* spp. and *Bacteroides* spp. in individuals with rheumatoid arthritis; *Haemophilus* spp. and *Klebsiella pneumoniae* in controls	Reductive acetyl-CoA pathway in individuals with rheumatoid arthritis; arginine transport systems in controls	8 MLGs	Random forest	45

Table 1 (cont.) | **Diseases studied by MWAS**

Cohort size (body site; country of residence)	Gene catalogue size	Sequencing reads mapped	Clustering method	Examples of key taxa	Examples of key functions	Disease classifier features	Feature selection method	Refs
Rheumatoid arthritis (cont.)								
54 individuals with rheumatoid arthritis, 51 controls and 37 after-treatment samples (Dental plaque; China)	3.2 million genes	77%	MLG	*Lactobacillus salivarius* and *Atopobium* spp. in individuals with rheumatoid arthritis; *Haemophilus* spp. and *Aggregatibacter* spp. in controls	Methionine salvage pathway in individuals with rheumatoid arthritis; arginine transport systems in controls	6 MLGs	Random forest	45
51 individuals with rheumatoid arthritis, 47 controls and 24 after-treatment samples (Saliva; China)	3.2 million genes	71%	MLG	*Lactobacillus salivarius* and *Veillonella* spp. in individuals with rheumatoid arthritis; *Haemophilus* spp. and *Prevotella intermedia* in controls	Menaquinone biosynthesis in individuals with rheumatoid arthritis; arginine transport systems in controls	2 MLGs	Random forest	45

BCAA, branched-chain amino acid; IGT, impaired glucose tolerance; LASSO, least absolute shrinkage and selection operator; LPS, lipopolysaccharide; mOTU, metagenomic operational taxonomic unit; MetaHIT, Metagenomics of the Human Intestinal Tract; MGC, metagenomic cluster; MGS, metagenomic species; MLG, metagenomic linkage group; mRMR, minimum redundancy maximum relevance; SVM, support vector machine.

individuals who were obese (body mass index (BMI) >30), overweight (BMI 25–30) or lean (BMI <25), metagenomic sequencing of faecal samples showed that the gut microbiomes of obese individuals were more likely to have low gene counts than high gene counts[18]. Notably, low gene counts, which can also be viewed as low gene richness, were correlated with physiological indicators such as higher levels of body fat, insulin resistance, dyslipidaemia and inflammation[18,19]. Furthermore, obese individuals who had a gut microbiome with a low gene richness had put on more weight during the past 9 years than obese individuals who had a gut microbiome with high gene richness[18].

The differences in gene richness also seemed to correspond to differences in bacterial species richness, as measured using taxonomical marker genes[18]. Butyrate-producing bacteria, such as *F. prausnitzii*, *Butyrivibrio* spp. and *R. inulinivorans*, as well as *Akkermansia* spp. and the methanogenic archaeon *Methanobrevibacter smithii*, were depleted in the gut microbiomes that had low gene richness, whereas *Bacteroides* spp. and *Ruminococcus gnavus* were more abundant in these gut microbiomes. Furthermore, the study showed that the species composition of the gut microbiome may be a better predictor of obesity than human genetic factors, as a disease classifier that was based on nine bacterial strains (MGS-like; TABLE 1) was able to distinguish between lean and obese individuals with an AUC of 0.78, whereas a disease classifier that was based on 32 human genomic loci could only distinguish between the two cohorts with an AUC of 0.58.

In the cohort from France, a 6-week energy-restricted high-protein diet that was followed by a 6-week weight-maintenance diet led to an increase in the gene richness of the gut microbiome and also reduced body fat, cholesterol and inflammation. However, even by the end of the intervention, the individuals who had gut microbiomes with low gene richness still had higher levels of these physiological indicators than the individuals who had gut microbiomes with high gene richness[19]. Thus, the gut microbiome may not only be able to stratify individuals according to obesity, but may also predict the response of individuals to dietary intervention.

Colorectal cancer. Colorectal cancer is among the three most frequently diagnosed forms of cancer worldwide and is a leading cause of cancer mortality[22,23]. Most cases of colorectal cancer are sporadic but are often preceded by the development of dysplastic adenomas that then progress into malignant forms; this progression is referred to as the adenoma–carcinoma sequence[24]. Prior to the study of colorectal cancer by MWAS, the association of the disease with *Fusobacterium* spp.[25–28] and SCFAs[29,30] in the gut microbiota had already been described. MWAS using faecal samples have enabled the identification of other microbial markers that may facilitate the early diagnosis of colorectal carcinomas or benign adenomas that may progress to colorectal carcinomas[31–33]. Three MWAS identified strains of *Bacteroides* spp., *Fusobacterium* spp., *Parvimonas micra*, *Peptostreptococcus stomatis* and *C. symbiosum* as overrepresented in faecal samples from patients with colorectal cancer, whereas strains of *Bifidobacterium* spp. and *Streptococcus* spp. were depleted in these samples[31–33] (TABLE 1). The enrichment of these species was common to the composition of the gut microbiome in individuals with colorectal cancer in all three cohorts, despite several technical differences between each study, such as different enrolment criteria and different methods of species annotation. Interestingly, *Fusobacterium* spp., *P. micra* and *P. stomatis* are known to be anaerobes that reside in the oral cavity and are usually rare in the gut[31–33]. The abundances of some of these strains had already changed significantly in patients with advanced

Area under the receiver operating characteristic curve
(AUC). The area under a receiver operating characteristic (ROC) curve of true-positive rates versus false-positive rates, which depicts the performance of a binary classifier. AUCs typically range between 0.5 and 1, corresponding to a random and a perfect classification, respectively.

Dyslipidaemia
An abnormal amount of lipids in the blood.

Adenomas
Benign tumours that are formed from glands or that have characteristics of glands.

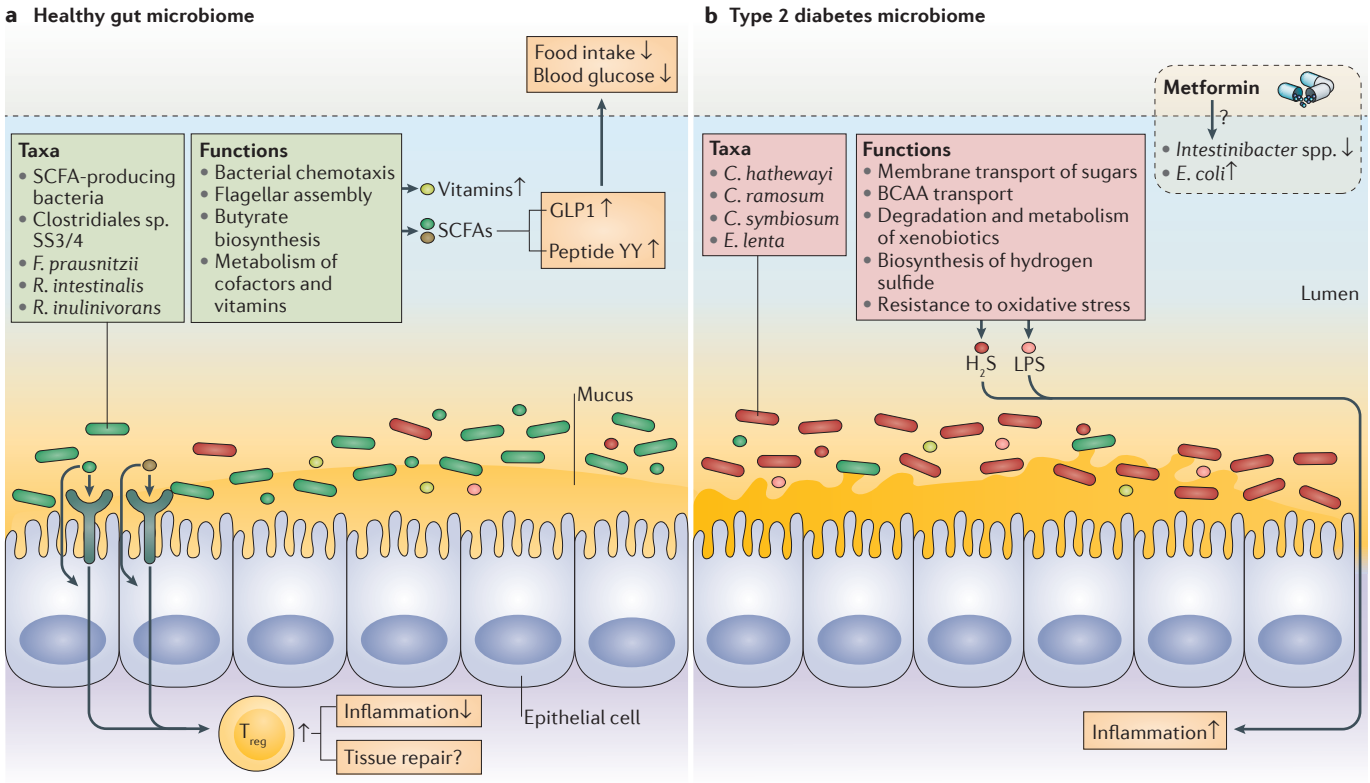


Figure 2 | **Changes to the gut microbiome that are associated with type 2 diabetes. a** | In healthy individuals, the gut microbiome is enriched for taxa that are associated with an increased capacity for the production of metabolites, such as short-chain fatty acids (SCFAs)[13,14], that promote intestinal integrity and energy homeostasis through absorption by the gut epithelium and signalling through host receptors to induce regulatory T cells (T_{reg}), which restricts inflammation and may even promote tissue repair[37]. SCFAs also stimulate the secretion of glucagon-like peptide 1 (GLP1) and peptide YY by intestinal L cells (not shown) to control glucose homeostasis and regulate food intake[10,11]. These taxa and functions tend to be depleted in the gut microbiomes of individuals with type 2 diabetes or obesity[13,14,18]. **b** | In individuals with type 2 diabetes, metagenome-wide association studies (MWAS) suggest that changes to the gut microbiome are associated with metabolic dysfunction and inflammation. For example, an increased potential for the production of hydrogen sulfide and lipopolysaccharide (LPS) could stimulate inflammation. However, the gut microbiomes of individuals with type 2 diabetes who were treated with the anti-diabetic drug metformin showed a decrease in the abundance of *Intestinibacter* spp. and an increase in the abundances of species in the Enterobacteriaceae family, such as *Escherichia coli*, compared with individuals with type 2 diabetes who did not receive metformin treatment. The increase in the abundance of *E. coli* seemed to correlate with an increase in the secretion of GLP1. BCAA, branched-chain amino acid; *C. hathewayi*, *Clostridium hathewayi*; *C. ramosum*, *Clostridium ramosum*; *C. symbiosum*, *Clostridium symbiosum*; *E. lenta*, *Eggerthella lenta*; *F. prausnitzii*, *Faecalibacterium prausnitzii*; *R. intestinalis*, *Roseburia intestinalis*; *R. inulinivorans*, *Roseburia inulinivorans*.

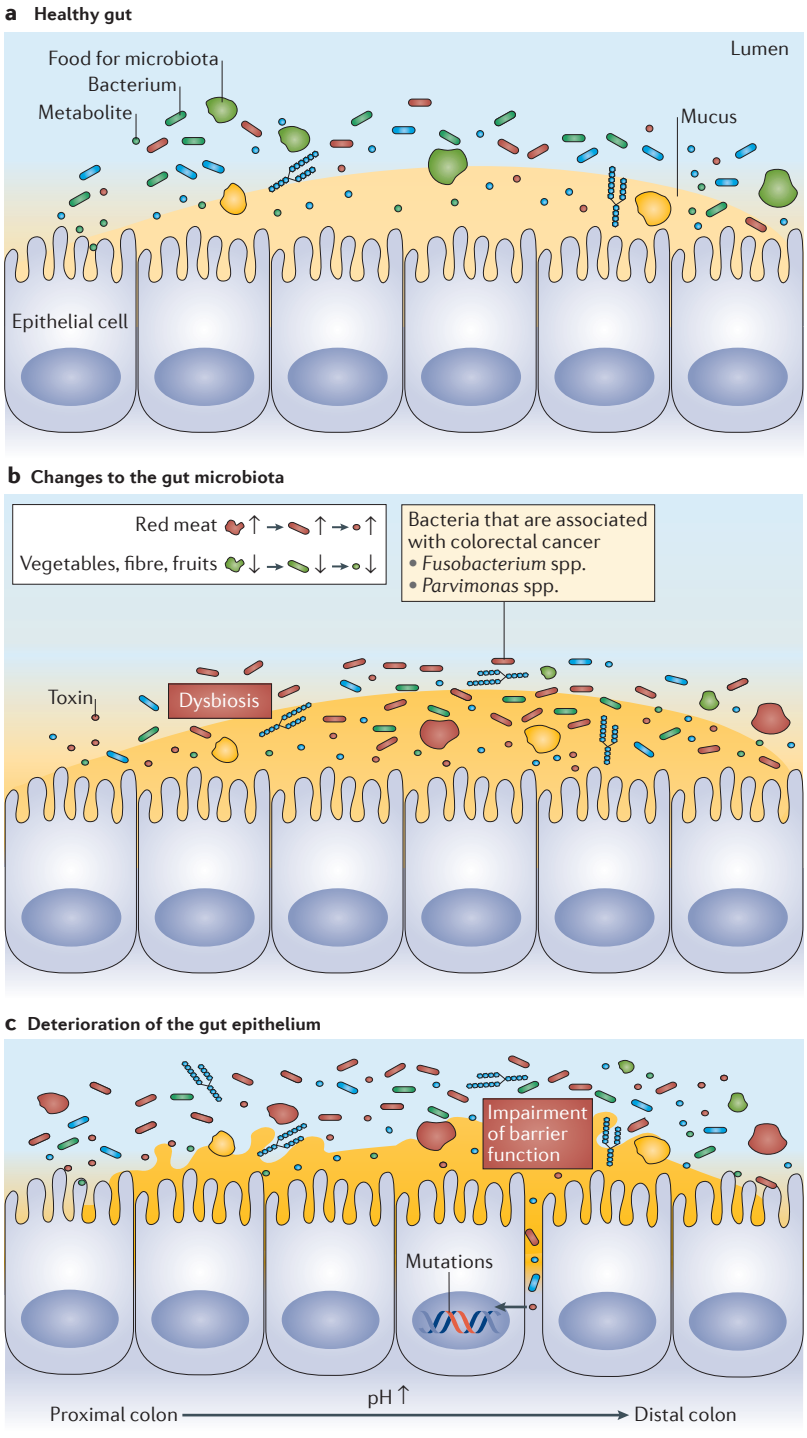
adenoma (≥1cm), and a classifier could be constructed that was able to distinguish between individuals with advanced adenomas and controls based on the composition of the gut microbiome[32]. Not only were similarities observed between the changes in the composition of the gut microbiota in individuals with colorectal cancer and individuals with adenomas, but the changes in the composition of the gut microbiota in individuals with colorectal cancer were also similar to the changes in the composition of the gut microbiota that have been reported for individuals with inflammatory bowel diseases (IBD) and mouse models of colitis-associated colorectal cancer[31,32]. Functional analysis of the metagenomic data suggested that the gut microbiome in individuals with colorectal cancer had a decreased capacity for the metabolism of sugars and dietary fibre compared with healthy individuals, but was better able to metabolize amino acids and host glycans and might also have a greater potential to produce toxins and carcinogenic metabolites[31,32] (FIG. 3).

One interesting direction for MWAS would be to compare the local gut microbiome in the tumour with the local gut microbiome in the adjacent mucosa, which have to date only been quantitatively compared using 16S rRNA gene amplicon sequencing or PCR[25–27,34,35]. Notably, *Fusobacterium* spp. were only found in tumour samples, whereas *Bacteroides* spp., *Parvimonas* spp. and *Peptostreptococcus* spp. were found both in tumour samples and in samples from the adjacent mucosa, although with a higher relative abundance in tumour samples[35,36]. It should be noted that differences in community composition may not be indicative of bacterial species that drive carcinogenesis, as such 'driver' bacterial species might have been outcompeted by 'passenger' bacteria

that gain a growth advantage as the tumour progresses through different stages[36]. Future work to better define the species in the gut microbiota that are associated with colorectal cancer would benefit from larger studies that examine different stages of colorectal cancer, differences in the location of tumours and/or differences in the age and gender groups of individuals with the disease[31–33,35].

The evidence that is emerging from MWAS in humans and from experiments in animal models suggests that the gut microbiome is a hub that integrates known risk factors that are associated with colorectal cancer, such as the consumption of red meat and smoking[24,32,37], and that influences the efficacy of therapies for the treatment of the disease[38,39]. How might these factors interact with the gut microbiota? One possibility is that the interactions relate to gut microbial metabolites that are key to a healthy gut epithelium, such as butyrate, or that might be genotoxic, such as secondary bile acids (reviewed in REFS 37,40). The consumption of fruits and vegetables possibly helps to favour the fermentation of dietary fibre, rather than host mucin, by the microbiota in the distal colon. The SCFAs that are produced through the fermentation of dietary fibre, together with lactic acid that is produced by lactic acid-producing bacteria, could also help to maintain a relatively low pH that might limit amino acid fermentation and the growth of pathogens[31,32] (FIG. 3). By contrast, a diet that is rich in red meat probably nurtures a gut microbiota that increases the likelihood of developing colorectal cancer, with increased bile secretion by host cells and an excess of iron available for pathogenic bacteria. The gut microbiota may also hold the key to the development of a personalized treatment for colorectal cancer, as several bacterial species that are found in the gut have been shown to positively or negatively affect therapies that are used to treat colorectal and other cancers in mouse models[32,38,39]. For example, administering a specific bacterium or its antigen into a mouse model of melanoma increased the success of immunotherapy[41,42].

Rheumatoid arthritis. Rheumatoid arthritis is a common autoimmune disorder that causes progressive disability and systemic complications. The concordance rates of rheumatoid arthritis are 15–30% in monozygotic twins and 5% in dizygotic twins, which suggests that genetic factors alone cannot account for the risk of developing



◀ Figure 3 | **Model for a gut microbial basis for the development of colorectal cancer. a**,**b** | Associations that were identified by metagenome-wide association studies (MWAS)[31,32] suggest that bacterial species that are usually of low abundance in the gut, and the toxins that they produce, could become more abundant in response to lifestyle or dietary changes, such as an increase in the consumption of red meat and a decrease in the consumption of fruits, vegetables and fibre. Some bacterial species that are most commonly described as anaerobic oral bacteria, such as *Fusobacterium* spp. and *Parvimonas micra*, have been identified by MWAS as being associated with colorectal cancer[31–33,35]. Functional changes in the gut microbiome might involve an increase in the production of carcinogens through processes such as amino acid fermentation and the metabolism of bile acids[37]. By contrast, bacterial species that produce the metabolites butyrate and lactate, which facilitate the maintenance of the colonic epithelium, can be depleted in the gut microbiomes of individuals with colorectal cancer. **c** | Dysbiosis of the gut microbiota can result in an impairment of gut barrier function, which increases the exposure of the gut epithelium to microorganisms and their metabolites[37,40]; some of these metabolites are mutagens that might promote carcinogenesis.

the disease[43]. Autoantibodies that are associated with rheumatoid arthritis can be detected in individuals years before the onset of the symptoms of the disease (that is, joint pains), and mucosal environments, such as the gut, mouth and lungs, have been suspected to be the initial site of inflammation before the onset of disease symptoms[44]. This suggests that early changes in the microbiota might be detectable before the development of symptoms, which provides a diagnostic use for marker genes that have been identified by MWAS.

Whereas other diseases have been studied using MWAS of faecal samples, which represent the gut microbiome, rheumatoid arthritis is the first disease to be studied using MWAS of the oral microbiome, although the gut microbiome has also been examined by MWAS for this disease[45]. The analysis of dental and salivary microbiomes alongside gut microbiomes showed that the difference between individuals with rheumatoid arthritis and healthy controls was greater in the oral microbiome than in the gut microbiome[45,46], which is consistent with the epidemiological link that has been established between rheumatoid arthritis and periodontitis[43,47]. This analysis also showed that changes in the taxonomical and functional composition of the microbiome that are associated with rheumatoid arthritis partially overlapped between the dental, salivary and gut microbiomes[45] (FIG. 4; TABLE 1) and that the abundances of specific bacterial taxa were correlated with the levels of serum markers of rheumatoid arthritis, including rheumatoid factor and anti-cyclic citrullinated peptide autoantibodies. Thus, both the gut and the oral microbiomes reflect, if not contribute to, the pathophysiology of rheumatoid arthritis.

As with type 2 diabetes, obesity and colorectal cancer, the gut microbiomes of individuals with rheumatoid arthritis could be distinguished from those of healthy individuals based on a disease classifier (TABLE 1). Furthermore, based on MLGs, a disease classifier could also be constructed to distinguish between the dental or salivary microbiomes of individuals with rheumatoid arthritis and those of healthy individuals, despite taxonomic differences between the gut and oral microbiomes. If classification based on the microbiomes of two of the three body sites was used to overrule the small number of misclassifications based on the microbiome of the third body site, almost all of the fully sampled individuals (68 out of 69) could be correctly classified as individuals with rheumatoid arthritis or as controls[45] (FIG. 4).

The disease classifier for the dental microbiome was particularly useful for evaluating the treatment effect of disease-modifying antirheumatic drugs (DMARDs). Dental samples from patients with low disease activity following treatment were often classified as healthy, and the abundances of several MLGs that are usually depleted in individuals with rheumatoid arthritis were higher in the dental microbiomes of patients who showed good or moderate improvement than in the dental microbiomes of patients who did not improve (as measured by the European League Against Rheumatism (EULAR) response criteria), which suggests that the recovery of a healthy dental microbiome might be an additional measure for evaluating DMARD treatments[45]. Preliminary results on the effects of different DMARDs on the oral and gut microbiomes were also obtained, and the outcome of treatment with DMARDs was predicted based on the microbiome of samples that were taken before treatment[45]. Given the long preclinical phase of rheumatoid arthritis, and the challenges in successfully tailoring drug treatment for each individual, microbiome-based diagnosis and prognosis may enable the development of exciting new possibilities for the management of the disease[43,45,48] (BOX 2).

Other diseases. A small number of MWAS have been reported for diseases other than type 2 diabetes, obesity, colorectal cancer and rheumatoid arthritis (TABLE 1). In one study, individuals with symptomatic atherosclerosis were found to have a lower abundance of genes that encode phytoene dehydrogenase in the gut microbiome compared with healthy individuals, and a concomitant decrease in serum β-carotene[49]. Associations between serum triglycerides, low-density and high-density lipoproteins (LDLs and HDLs) and specific members of the gut microbiota have also been identified in MWAS of rheumatoid arthritis or colorectal cancer, as well as in a study of a general population cohort that was based on 16S rRNA gene amplicon sequencing, although no MWAS has yet directly examined cardiovascular diseases[32,45,50]. Another MWAS identified bacterial taxa that were enriched in individuals with liver cirrhosis and found that many of these taxa, including *Veillonella* spp., *Streptococcus* spp. and *Haemophilus parainfluenzae*, were of suspected oral origin; a disease classifier for liver cirrhosis was also constructed, based on 15 gut microbial genes[51] (TABLE 1).

What is a healthy microbiome? Some bacterial species and bacterial functions have been associated with healthy controls in MWAS for more than one disease, which suggests that these taxa and functions might be features of a 'healthy' microbiome (TABLE 1). Notably, SCFA-producing bacteria, such as *R. inulinivorans* and *F. prausnitzii*, have been associated with a healthy gut microbiome in MWAS of several diseases, including type 2 diabetes and obesity (FIG. 2). Bacteria in the genus *Bacteroides*, which is the most abundant genus of bacteria in the gut and is associated with natural delivery and breast-feeding as infants[3,52,53], include both taxa that are associated with a healthy microbiome and taxa that are associated with some diseases (TABLE 1). On the one hand, bacteria from this genus help to metabolize various plant-derived sugars, which provides a benefit to the host; conversely, they can also forage for host glycans[54,55] and produce toxins, which are functions that might contribute to the development of colorectal cancer[31–33] (FIG. 3). Similarly, *Lactobacillus salivarius* and *Bifidobacterium dentium* are enriched in the gut microbiomes of individuals with rheumatoid arthritis[45], even though the genera *Lactobacillus* and *Bifidobacterium* are generally regarded as comprising beneficial species that train the immune system and restrict the growth of other bacterial species.

Periodontitis
Inflammation of the tissue that surrounds the teeth, which leads to the progressive loss of the alveolar bone and the loosening or loss of teeth.

Rheumatoid factor
An autoantibody against the constant region (known as the fragment crystallisable (Fc) region) of immunoglobulin G.

Anti-cyclic citrullinated peptide autoantibodies
Autoantibodies against proteins that contain the modified amino acid citrulline. Cyclic citrullinated peptides are used to clinically detect these antibodies.

A strain-level characterization of the metabolisms and host interactions of members of genera such as *Bacteroides*, *Lactobacillus* and *Bifidobacterium* may help to elucidate the full complexity of their seemingly contradictory roles in human health and disease[56,57].

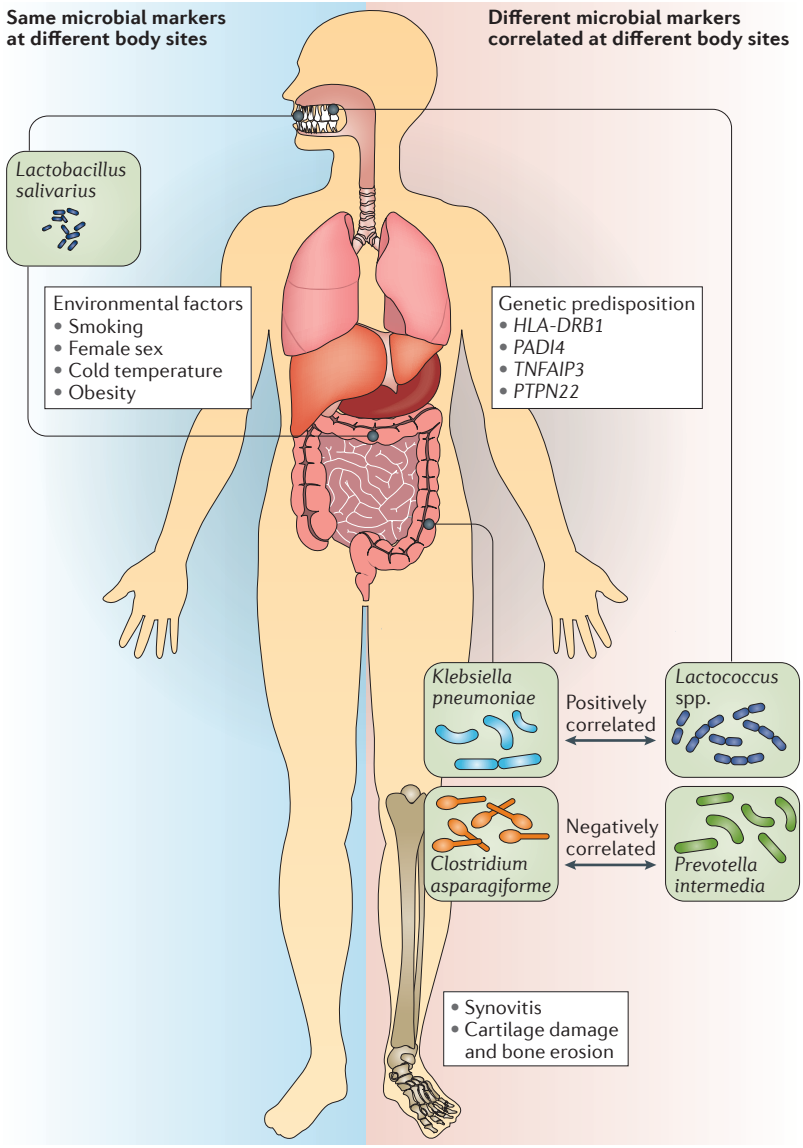


Figure 4 | **The oral and gut microbiomes of individuals with rheumatoid arthritis.** The microbiome might interact with both genetic and environmental factors that influence the risk of developing rheumatoid arthritis[45,46]. Using metagenome-wide association studies (MWAS) to examine both the oral and gut microbiomes of individuals with rheumatoid arthritis has shown an overlap between the microbiomes from the two body sites, with an enrichment of several bacterial species, including *Lactobacillus salivarius*, at both sites. Sets of bacterial species were also shown to have correlated changes in abundance between the oral and gut microbiomes of individuals with rheumatoid arthritis: for example, the abundance of *Klebsiella pneumoniae* in the gut microbiome was positively correlated with the abundance of *Lactococcus* spp. in the oral microbiome, whereas the abundance of *Clostridium asparagiforme* in the gut microbiome was negatively correlated with the abundance of *Prevotella intermedia* in the oral microbiome[45]. As such, sampling at one body site may reveal information about the microbiome at another site[45]. *HLA-DRB1*, major histocompatibility complex, class II, DRβ1; *PADI4*, peptidyl arginine deiminase 4; *PTPN22*, protein tyrosine phosphatase non-receptor type 22; *TNFAIP3*, tumour necrosis factor-α-induced protein 3.

The signatures of health and disease in the microbiome may be clearer at the functional level than at the taxonomic level. In the anaerobic environment of the gut, an increased capacity to tolerate oxidative stress is a sign of dysbiosis of the microbiome, as it is indicative of the presence of aerobic bacteria and/or activation of the host immune system (FIG. 2). Another functional indicator of dysbiosis of the gut microbiome is the reduction of sulfate or sulfite into hydrogen sulfide, which may also ameliorate dysbiosis[9,18,45,58]. For populations with a diet that is rich in resistant starch, which is fermented by the microbiota, the production of hydrogen gas as an end product of fermentation, and the subsequent use of hydrogen to produce methane, are major functions of the gut microbiome; however, these functions were identified by MWAS as perturbed in individuals with low gene counts in the gut microbiome[3,18]. Finally, the SCFAs butyrate and, to some extent, propionate are also associated with a healthy gut microbiome, as shown in MWAS of diseases such as type 2 diabetes, obesity and colorectal cancer. These metabolites are a major source of energy for gut epithelial cells and, as such, help to maintain a healthy gut environment. Furthermore, SCFAs confer additional benefits to the health of the host by inhibiting histone deacetylation, signalling through host receptors and inducing the differentiation of regulatory T cells[37] (FIG. 2). These associations between a healthy gut and the functions, rather than the taxa, of the gut microbiome demonstrate the effectiveness of functional information provided by MWAS for studying disease mechanisms, as opposed to single-gene taxonomical marker surveys, from which function can only be inferred.

From association to causality?

The success of MWAS does not only depend on the findings that are obtained using the method itself, but also on follow-up studies that investigate whether microbiome–disease associations can be validated in other cohorts, what the mechanistic bases are for these associations and whether the associations are causal (although the identification of non-causal associations can also be useful in developing diagnostic markers; BOX 2). Thus, an MWAS may show that a gene has 'guilt by association' with a disease, but its 'conviction' requires further evidence from all possible sources (FIG. 1g). It should be noted that prior knowledge of the identified markers is usually very limited, as many genes in the microbiome have not yet been functionally characterized.

Additional evidence to support disease associations. Analogous to GWAS, additional cohorts can be used to validate associations between microbial genes and a disease, which increases the sample size and reduces the problem of technical issues, and such validation cohorts have been used in MWAS of type 2 diabetes, colorectal cancer and rheumatoid arthritis[9,15,31–33,45] (TABLE 1). However, unlike the human genome, the microbiome of any given individual is subject to substantial variation, which adds to the difficulties of validation — but also raises hopes for the potential of intervention to ameliorate diseases through the modulation of the microbiota.

Box 2 | **MWAS: from bench to bedside**

Diagnostic markers

Whether causal or not, associations that are identified by metagenome-wide association studies (MWAS) could be a starting point for the development of non-invasive tests that use microbiome-based marker genes for the diagnosis of diseases before the development of symptoms (FIGS 1,4; TABLE 1). For colorectal cancer and rheumatoid arthritis, faecal or oral microbiome-based tests may be more powerful than conventional screening techniques, and quantitative PCR (qPCR) or other relatively rapid tests on only a few microbial genes would suffice[31–33,45].

Patient stratification and precision medicine

Markers that are identified by MWAS may be characteristic of only a subgroup of individuals with a disease[3,18,32,45], which might inform the tailoring of treatment. For example, gene markers in the microbiome might be predictive of the probable outcome of a specific therapy[19,32,45,91,114], as has been shown for disease-modifying antirheumatic drugs (DMARDs) in individuals with rheumatoid arthritis. Continuing to monitor these markers during the course of treatment could help to further optimize the treatment plan. For the effective screening of large groups of individuals, or even the whole population, mobile applications could be used to collect dietary, lifestyle and phenotypic information at regular time intervals from the same individual[115]. These data could then be used to generate recommendations for intervention strategies that might help to prevent or delay the development of disease. An example of a metric that might be better managed using data from the microbiome is blood glucose level, which can vary markedly between healthy individuals in response to the same meal[116,117]. Personalized dietary intervention that takes the gut microbiota, diet, physical activity, blood parameters and other metrics into consideration could then be used to help prevent the onset of metabolic disease[117].

Treatment of diseases

Dietary intervention or microbial transplant (whether of faeces, defined mixtures of bacterial strains or single strains) could be designed to target markers of disease or treatment success that have been identified by MWAS (reviewed in REF. 118). Alternatively, new drugs could be developed to target the products of these marker genes[71,114,119,120].

An additional technical challenge when designing a validation study with an independent cohort is that human cohorts are commonly heterogeneous with respect to factors such as geographical origin, BMI and age (BOX 1). Validation analyses using independent cohorts or using meta-analyses across studies are expected to become routine as more metagenomic data become available for more diseases, but these will require the sharing of metadata and data standardization.

Other forms of omics data, such as metatranscriptomic, metaproteomic and metabolomic data (together with host genetic data, which can also shape the composition of the microbiome[56,59–65]; FIG. 1g), can provide further support to candidate microbiome–disease associations that have been identified by MWAS (reviewed in REF. 66). Future studies may use metatranscriptomic and metaproteomic data to identify RNAs and proteins that are differentially enriched in the microbiomes of disease cohorts. These data may need to be analysed in combination with metagenomic data to control for the total number of bacterial cells, if the studies are to distinguish between genes that have upregulated expression and genes for which more abundant transcripts simply reflects a larger number of bacterial cells. Another promising form of omics data for the study of the microbiome is metabolomics, although a database of microbiota-contributed metabolites is currently not available[67]. In an analysis of data from an MWAS of obesity[19], metabolomic modelling of five representative bacterial species before and after a weight-reducing dietary intervention enabled the prediction of changes in the levels of SCFAs and amino acids in faeces and serum[68]. An important benefit of multiomic data is that these data provide evidence that microbial markers that have been identified by MWAS are active in the body.

Finding evidence for causality in disease associations. To progress from identifying microbiome–disease associations to identifying the functions of the microbiome in disease, experiments are required that are able to examine causality between the microbial markers that are identified by MWAS and the disease of interest. These experiments can include the study of longitudinal cohorts, experiments that use animal models and *in vitro* functional studies. Longitudinal cohorts enable the study of samples that are taken from the same individuals before the onset of disease and during disease development (FIG. 1a), especially from preclinical or high-risk groups. Such an approach would be very useful for establishing whether the enrichment of microbial genes occurs before the development of disease phenotypes at a relevant body site, which would be expected for genes that have a causal role. Longitudinal studies for testing microbiome–disease associations could be even more powerful when comparing samples before and after interventions, such as drug treatment, dietary change or microbial transplant[19,45,69–72] (FIG. 1). One example of a longitudinal study that used a drug treatment to modulate the gut microbiota examined the ability of a berberine-containing herbal mixture to alleviate type 2 diabetes[69]. Changes to the gut microbiota, such as a substantial enrichment of *F. prausnitzii*, could be observed after 4 weeks of treatment; importantly, these changes preceded the detection of the alleviation of the symptoms associated with type 2 diabetes, which were not detected until 8 weeks of treatment, which suggests that the therapeutic effects of the treatment might have been mediated by the gut microbiota. These results also extend the MWAS finding that the butyrate-producing bacterium *F. prausnitzii* was depleted in the gut microbiomes of individuals with type 2 diabetes[9,15] (FIG. 2).

The use of animal models (FIG. 1g) enables well-controlled study designs that minimize the influence of environmental and host factors that can confound the analysis of data from human cohorts, and also enables numerous physiological parameters, body sites and cell types to be measured that may not yet be accessible in humans. Together, these benefits mean that the effects of genetics, diet and microbial community function can be better separated in animal models than in human cohorts[56,73]. However, it should be noted that the microbiomes of mice from different sources can differ substantially in composition[74], which should be taken into account when designing these studies. Furthermore, mouse models should be used to investigate the exact strain that has been identified by MWAS, even if related strains are more readily available (FIG. 1g).

Germ-free mice are commonly used as animal models for studies of the microbiome, as they provide

Guilt by association
A concept from genome-wide association studies (GWAS) that describes associations in such studies as 'guilt' of a gene for a trait of interest, which means that the gene is of interest for further investigation.

a clean microbial background that enables the specific study of microorganisms of interest. However, the results that are obtained using germ-free mice should be interpreted with caution, as these mice are metabolically, immunologically and neurologically abnormal, owing to physiological changes such as an overproduction of corticosterone, an increase in the level of triglycerides and an impairment of the blood–brain barrier[75–79]. As an alternative to germ-free mice, specific pathogen-free mice may be used as animal models. Results that are obtained using these mice may be of more translational value than those that are obtained using germ-free mice; although, colonization of specific pathogen-free mice can be more challenging than germ-free mice, this can be addressed by the use of mutant mice that are deficient for specific genes[57,72]. Studies using animal models that more closely resemble humans, such as pigs or minipigs, or simpler — and therefore more experimentally tractable — model organisms, such as nematodes, are also important for functional investigations of the microbiome[74,80–82].

An example of pioneering work that used an animal model of the gut microbiota is the demonstration of a causal role for the gut microbiota in obesity. Germ-free mice that had received a faecal microbial transplant from obese human donors had a greater increase in adiposity than germ-free mice that had received a faecal microbial transplant from lean human donors[73,83]. Furthermore, in agreement with associations that were identified by MWAS[18,19] (TABLE 1), several species of bacteria, including *Akkermansia muciniphila* and *F. prausnitzii*, have been shown to have functions that prevent obesity in mouse models or in mouse gut organoids[84].

Although longitudinal studies of human cohorts and studies that use animal models can help to establish a causal link for associations that are identified by MWAS, elucidating the mechanism that underlies the association may rely on functional studies that are aimed at understanding the microbiology of individual bacterial strains, as well as their interactions with each other and with host molecules[85–87], cells and tissues[84,88]. Such studies would also help to characterize the microbial genes and products that drive the development or progression of disease, as these genes and products may not yet have known functions in existing databases[89]. For example, *in vitro* studies of the Gram-positive bacillus *E. lenta*, which has been identified in MWAS that investigated type 2 diabetes and rheumatoid arthritis, found that an operon that encodes two genes that are predicted to be bacterial cytochromes inactivates the cardiac drug digoxin and that this inactivation is inhibited by arginine[90,91]. Differences in digoxin metabolism between strains of *E. lenta* or between different individuals could then be explained by the presence or absence of this operon, although the relevance, if any, to type 2 diabetes or rheumatoid arthritis remains unknown[9,45].

Together, longitudinal studies of human cohorts, studies that use animal models and *in vitro* functional studies of microorganisms and microbial products may help to establish how associations that are identified by MWAS relate to the aetiology of a disease. As the field of MWAS is still in its infancy, with only a small number of studies that have been reported to date and with relatively small samples sizes, follow-up studies that have investigated causality of MWAS findings have not yet been reported, although, as mentioned above, animal models that use faecal transplants or antibiotic treatment have shown the involvement of the gut microbiota in the development of some diseases. Therefore, elucidating causal associations between the microbiome and disease remains an important challenge.

Conclusions and future perspectives

Despite sample sizes that are orders of magnitude smaller than GWAS, MWAS have been able to identify microbial markers that are associated with complex diseases, including type 2 diabetes, obesity, liver cirrhosis, colorectal cancer and rheumatoid arthritis (BOX 2; TABLE 1). Notably, the production of butyrate has been repeatedly identified as a feature of a healthy gut microbiome; preliminary associations have also been made for numerous other functions and genes but require further investigation.

Improving the power of MWAS to identify associations between the microbiome and disease will depend on resolving challenges in sampling, sequencing, bioinformatics analyses and the functional characterization of the microbiome. For example, the taxonomic annotation of metagenomic datasets (BOX 1) will improve as more microbial genome sequences become available from ongoing sequencing efforts, such as those using genome-resolved metagenomic assemblies[53,92–94] and single-cell sequencing methods[95,96], which promise to substantially increase the availability of genomes for species that cannot be cultivated in the laboratory. These advances may enable the construction of a comprehensive reference genome catalogue that could be used for high-resolution analyses as an alternative to the clustering of reference genes. Genome-based analyses will have to take into account the effects of horizontal gene transfer by mobile genetic elements, such as plasmids, transposons and bacteriophages[3,93,97–101], which may be relatively common in crowded environments such as the human gut[98]. Such elements are among the genetic variants in the microbiome, which also include SNPs, indels and copy number variations (CNVs)[102–104], that high-resolution studies could investigate for associations with a host phenotype. The availability of a reference genome catalogue would also help to resolve the gap between taxon-level and function-level analyses, which are largely separated in the gene catalogue approach used by current MWAS (BOX 1). The gap may be further narrowed through the incorporation of multiomic data, which provide more informative measurements of community function. Finally, future MWAS may also look for associations between diseases and eukaryotic viruses or eukaryotic microorganisms.

The ultimate goal of MWAS is to inform improved diagnostics or therapeutic (or preventive) interventions, which may include changes to the diet or lifestyle of an individual, or clinical interventions, such as microbial

Specific pathogen-free mice
Laboratory mice that are free of particular pathogens that could interfere with experiments. The excluded pathogens include both viral and bacterial pathogens.

Organoids
Organ-like structures that are grown in the laboratory.

Mobile genetic elements
DNA sequences that can be transferred between genomes or between loci of the same genome.

transplant or drug treatment. However, such clinical benefits will depend on a more detailed definition of the healthy microbiome, which can vary substantially between individuals. Currently, information on the healthy microbiome is limited, owing to relatively small sample sizes and metadata that are incomplete or that have not been shared with the public. One important use of an improved definition of the healthy microbiome may be the design of microbial transplant treatments. For those transplants that use faeces from a donor, a well-defined set of criteria that denote a healthy microbiome may be useful in donor selection and could be used to maintain a bank of quality-controlled donor samples. The criteria may also be useful for determining which strains to include in laboratory cultivated microbial cocktails for use in microbial transplant treatments. When combined with a more comprehensive understanding of the 'healthy' microbiome, we expect that improved sampling, sequencing, bioinformatics analyses and functional characterization will empower future MWAS to have many applications in human health and disease.

1. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
This study details the first gene catalogue of the human gut microbiome that is assembled from next-generation sequencing data.
2. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
3. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
This study details a high-quality reference gene catalogue that is compiled from 1,267 samples across three continents, and identifies many differences in the gut microbiome between healthy Chinese and Danish individuals.
4. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258–1270 (2012).
5. Sommer, F. & Bäckhed, F. The gut microbiota — masters of host development and physiology. *Nat. Rev. Microbiol.* **11**, 227–238 (2013).
6. Marchesi, J. R. *et al.* The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330–339 (2015).
7. Xu, Y. *et al.* Prevalence and control of diabetes in Chinese adults. *JAMA* **310**, 948–959 (2013).
8. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5**, e9085 (2010).
9. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
The first MWAS, which establishes the MLG method and identifies associations between the gut microbiome and type 2 diabetes.
10. Brubaker, P. L. & Anini, Y. Direct and indirect mechanisms regulating secretion of glucagon-like peptide-1 and glucagon-like peptide-2. *Can. J. Physiol. Pharmacol.* **81**, 1005–1012 (2003).
11. Zhou, J. *et al.* Peptide YY and proglucagon mRNA expression patterns and regulation in the gut. *Obesity (Silver Spring)* **14**, 683–689 (2006).
12. Sun, J. *et al.* Pancreatic β-cells limit autoimmune diabetes via an immunoregulatory antimicrobial peptide expressed under the influence of the gut microbiota. *Immunity* **43**, 304–317 (2015).
13. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
14. Pyra, K. A., Saha, D. C. & Reimer, R. A. Prebiotic fiber increases hepatic acetyl CoA carboxylase phosphorylation and suppresses glucose-dependent insulinotropic polypeptide secretion more effectively when used with metformin in obese rats. *J. Nutr.* **142**, 213–220 (2012).
15. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
16. Shin, N.-R. *et al.* An increase in the *Akkermansia* spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* **63**, 727–735 (2014).
17. Lee, H. & Ko, G. Effect of metformin on metabolic improvement and gut microbiota. *Appl. Environ. Microbiol.* **80**, 5935–5943 (2014).
18. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
19. Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585–588 (2013).
20. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
21. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
22. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).
23. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
24. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet* **383**, 1490–1502 (2014).
25. Kostic, A. D. *et al.* Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
26. Castellarin, M. *et al. Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
27. Kostic, A. D. *et al. Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
28. Gur, C. *et al.* Binding of the Fap2 protein of *Fusobacterium nucleatum* to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity* **42**, 344–355 (2015).
29. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).
30. Weir, T. L. *et al.* Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS ONE* **8**, e70803 (2013).
31. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
32. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
33. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* http://dx.doi.org/10.1136/gutjnl-2015-309800 (2015).
34. Marchesi, J. R. *et al.* Towards the human colorectal cancer microbiome. *PLoS ONE* **6**, e20447 (2011).
35. Nakatsu, G. *et al.* Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* **6**, 8727 (2015).
36. Tjalsma, H., Boleij, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nat. Rev. Microbiol.* **10**, 575–582 (2012).
37. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661–672 (2014).
38. Iida, N. *et al.* Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* **342**, 967–970 (2013).
39. Viaud, S. *et al.* The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science* **342**, 971–976 (2013).
40. Schwabe, R. F. & Jobin, C. The microbiome and cancer. *Nat. Rev. Cancer* **13**, 800–812 (2013).
41. Vetizou, M. *et al.* Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
42. Sivan, A. *et al.* Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).
43. McInnes, I. B. & Schett, G. The pathogenesis of rheumatoid arthritis. *N. Engl. J. Med.* **365**, 2205–2219 (2011).
44. Demoruelle, M. K., Deane, K. D. & Holers, V. M. When and where does inflammation begin in rheumatoid arthritis? *Curr. Opin. Rheumatol.* **26**, 2264–2271 (2014).
45. Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905 (2015).
This study extends MWAS to the oral microbiome, and identifies potential markers in the oral and gut microbiomes for rheumatoid arthritis and its treatment by drugs.
46. Scher, J. U. *et al.* Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2**, e01202 (2013).
47. Scher, J. U. *et al.* Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum.* **64**, 3083–3094 (2012).
48. Deane, K. D. & El-Gabalawy, H. Pathogenesis and prevention of rheumatic disease: focus on preclinical RA and SLE. *Nat. Rev. Rheumatol.* **10**, 212–228 (2014).
49. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* **3**, 1245 (2012).
50. Fu, J. *et al.* The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* **117**, 817–824 (2015).
51. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 859–864 (2014).
52. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509**, 357–360 (2014).
53. Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
This study details the largest cohort of infants for which gut microbiomes have been longitudinally profiled for one year from birth; draft genomes were assembled from each sample through the binning of contigs instead of genes.
54. Sonnenburg, J. L. *et al.* Glycan foraging *in vivo* by an intestine-adapted bacterial symbiont. *Science* **307**, 1955–1959 (2005).
55. Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* **10**, 323–335 (2012).
56. Kashnap, P. *et al.* Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on gut microbiota. *Proc. Natl Acad. Sci. USA* **110**, 17059–17064 (2013).
57. Lee, S. M. *et al.* Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426–429 (2013).
This study identifies glycan-utilizing colonization factors in *Bacteroides* spp. that are responsible for saturable colonization of individual *Bacteroides* species in mice.
58. Motta, J.-P. *et al.* Hydrogen sulfide protects from colitis and restores intestinal microbiota biofilm and mucus production. *Inflamm. Bowel Dis.* **21**, 1006–1017 (2015).
59. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl Acad. Sci. USA* **107**, 18933–18938 (2010).
60. Benson, A. K. Host genetic architecture and the landscape of microbiome composition: humans weigh in. *Genome Biol.* **16**, 203 (2015).
61. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).

62. van Opstal, E. J. & Bordenstein, S. R. Rethinking heritability of the microbiome. *Science* **349**, 1172–1173 (2015).
63. Rausch, P. *et al.* Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and *FUT2* (*Secretor*) genotype. *Proc. Natl Acad. Sci. USA* **108**, 19030–19035 (2011).
64. Pickard, J. M. *et al.* Rapid fucosylation of intestinal epithelium sustains host–commensal symbiosis in sickness. *Nature* **514**, 638–641 (2014).
65. Goto, Y. *et al.* Innate lymphoid cells regulate intestinal epithelial cell glycosylation. *Science* **345**, 1254009 (2014).
66. Franzosa, E. A. *et al.* Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13**, 360–372 (2015).
67. Sridharan, G. V. *et al.* Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat. Commun.* **5**, 5492 (2014).
68. Shoaie, S. *et al.* Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell. Metab.* **22**, 320–331 (2015).
69. Xu, J. *et al.* Structural modulation of gut microbiota during alleviation of type 2 diabetes with a Chinese herbal formula. *ISME J.* **9**, 552–562 (2014).
70. Lukens, J. R. *et al.* Dietary modulation of the microbiome affects autoinflammatory disease. *Nature* **516**, 246–249 (2014).
71. Hsiao, E. Y. *et al.* Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451–1463 (2013).
72. Atarashi, K. *et al.* T_{reg} induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* **500**, 232–236 (2013).
73. Ridaura, V. K. *et al.* Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
This study shows that co-housing mice that have received microbial transplants from an obese twin with mice that have received microbial transplants from a lean twin prevents the development of obesity-associated phenotypes.
74. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
This paper details the first gene catalogue for the gut microbiome of laboratory mice, which reports differences from the human gut microbiome as well as between mice.
75. Bäckhed, F., Manchester, J. K., Semenkovich, C. F. & Gordon, J. I. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc. Natl Acad. Sci. USA* **104**, 979–984 (2007).
76. Mukherji, A., Kobiita, A., Ye, T. & Chambon, P. Homeostasis in intestinal epithelium is orchestrated by the circadian clock and microbiota cues transduced by TLRs. *Cell* **153**, 812–827 (2013).
77. Olszak, T. *et al.* Microbial exposure during early life has persistent effects on natural killer T cell function. *Science* **336**, 489–493 (2012).
78. Braniste, V. *et al.* The gut microbiota influences blood–brain barrier permeability in mice. *Sci. Transl Med.* **6**, 263ra158 (2014).
79. Erny, D. *et al.* Host microbiota constantly control maturation and function of microglia in the CNS. *Nat. Neurosci.* **18**, 965–977 (2015).
80. Lacombe, A. *et al.* Lowbush wild blueberries have the potential to modify gut microbiota and xenobiotic metabolism in the rat colon. *PLoS ONE* **8**, e67497 (2013).
81. Hildebrand, F. *et al.* A comparative analysis of the intestinal metagenomes present in guinea pigs (*Cavia porcellus*) and humans (*Homo sapiens*). *BMC Genomics* **13**, 514 (2012).
82. Cabreiro, F. & Gems, D. Worms need microbes too: microbiota, health and aging in *Caenorhabditis elegans*. *EMBO Mol. Med.* **5**, 1300–1310 (2013).
83. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
84. Lukovac, S. *et al.* Differential modulation by *Akkermansia muciniphila* and *Faecalibacterium prausnitzii* of host peripheral lipid metabolism and histone acetylation in mouse gut organoids. *mBio* **5**, e01438-14 (2014).
85. Kim, K., Lee, S. & Ryu, C.-M. Interspecific bacterial sensing through airborne signals modulates locomotion and drug resistance. *Nat. Commun.* **4**, 1809 (2013).
86. Cuskin, F. *et al.* Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* **517**, 165–169 (2015).
87. Stowell, S. R. *et al.* Microbial glycan microarrays define key features of host-microbial interactions. *Nat. Chem. Biol.* **10**, 470–476 (2014).
88. Wang, X. *et al.* Cloning and variation of ground state intestinal stem cells. *Nature* **522**, 173–178 (2015).
89. Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C. & Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell. Metab.* **20**, 731–741 (2014).
90. Dobkin, J. F., Saha, J. R., Butler, V. P., Neu, H. C. & Lindenbaum, J. Inactivation of digoxin by *Eubacterium lentum*, an anaerobe of the human gut flora. *Trans. Assoc. Am. Physicians* **95**, 22–29 (1982).
91. Haiser, H. J. *et al.* Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
92. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
93. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
This study reports the use of MGSs to assemble genomes, 238 of which met the Human Microbiome Project (HMP) high-quality draft genome standard.
94. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* **34**, 64–69 (2015).
95. McLean, J. S. *et al.* Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl Acad. Sci. USA* **110**, E2390–E2399 (2013).
96. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
97. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
98. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
99. Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
100. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222 (2013).
101. Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).
102. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2012).
This study represents the first analysis of genomic variations, such as SNPs, in the gut microbiome.
103. Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat. Commun.* **4**, 2151 (2013).
104. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–594 (2015).
105. Aagaard, K. *et al.* The placenta harbors a unique microbiome. *Sci. Transl Med.* **6**, 237ra65 (2014).
106. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).
107. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7**, e47656 (2012).
108. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
109. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
110. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
111. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
112. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012).
113. Freitas, T. A., Li, P.-E., Scholz, M. B. & Chain, P. S. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* **43**, e69 (2015).
114. Spanogiannopoulos, P., Bess, E. N., Carmody, R. N. & Turnbaugh, P. J. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nat. Rev. Microbiol.* **14**, 273–287 (2016).
115. Gill, S. & Panda, S. A. Smartphone app reveals erratic diurnal eating patterns in humans that can be modulated for health benefits. *Cell. Metab.* **22**, 789–798 (2015).
116. Kovatcheva-Datchary, P. *et al.* Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell. Metab.* **22**, 971–982 (2015).
117. Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
This study shows that the composition of the gut microbiota, integrated with other parameters, can be used to predict the blood glucose level of an individual after a certain meal and facilitate dietary interventions.
118. Olle, B. Medicines from microbiota. *Nat. Biotechnol.* **31**, 309–315 (2013).
119. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
120. Wang, Z. *et al.* Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* **163**, 1585–1595 (2015).
This study identifies a chemical analogue of choline that shows success in inhibiting the production of trimethylamine by the gut microbiota.

Acknowledgements
This study was supported by the Natural Science Foundation of China (grants 30890032, 30725008 and 30811130531), the Shenzhen Municipal Government of China (grants JSGG20140702161403250, DRC-SZ[2015]162 and CXB201108250098A), the Danish Strategic Research Council (grant 2106-07-0021) and the Ole RØmer grant from the Danish Natural Science Research Council and Solexa project (272-07-0196). The authors thank their colleagues at BGI, Shenzhen, China, especially J. Li, Z. Lan, S. Liang, H. Xie, D. Zhang, X. Luo, M. Arumugam and K. Kristiansen, for their help in the preparation of this Review. The authors also thank Y. Xie at Michigan State University, East Lansing, USA, for helpful discussions regarding this manuscript.

Competing interests statement
The authors declare competing interests: see Web version for details.