PROJECT FINAL REPORT

Grant Agreement number: **FP7-HEALTH-2010-261376**

Project acronym: **IHMS**

Project title: **International Human Microbiome Standards**

Funding Scheme: **Coordination and Support Action**

Period covered: **from February 1 2011** **to Jan. 31 2015**

Project co-ordinator name, title and organisation:

Tel: 33 1 34 65 25 10

Fax: 33 1 34 65 25 21

E-mail: dusko.ehrlich@jouy.inra.fr

Project website address: http://www.microbiome-standards.org

PUBLISHABLE SUMMARY

1. EXECUTIVE SUMMARY

Humans live in constant association with microbes. The number of our microbial companions exceeds by at least ten-fold those of cells of our own body and the number of unique genes they encode is at least 150-fold greater than the number of genes in our own genome. This complex and dynamic microbiota has a profound influence on human health and disease. Understanding the dynamic and variable nature of human microbial communities and their impact on our bodies is a next frontier of human biology. To reach this understanding it is necessary to study the microbial communities world-wide and with appropriate approaches, which need to be standardized in order to make studies comparable and thus optimally synergistic. The overall concept of IHMS is to promote development and implementation of standard procedures and protocols across three separate but related activities:

(i) Collecting and processing of human samples.

(ii) Sequencing of the human-associated microbial genes and genomes

(iii) Organizing and analyzing the data gathered

IHMS focused on the gut microbial communities, which are most complex and numerous and are often viewed as a neglected organ, known to impact health and disease. It focused on the cutting edge approach, Quantitative Metagenomics, developed in the large FP7 project MetaHIT.

Four SOPs were developed for sample collection. They address issues of conservation of the microbial composition in function of time required to deep-freeze the samples, as such samples can be conserved indefinitely. An additional SOP has been adapted, in agreement with the recommendations of the International Human Microbiome Consortium, to identify the samples via the metadata that describe the individual from whom the sample was obtained.

Two SOPs were developed for sample processing, that is, DNA extraction, after testing the protocols used by over 20 laboratories involved in the human microbiome work world-wide. One is most useful for manual work and could be implemented by the laboratories involved in smaller scale studies. Another is apt to automation and will be useful for large scale studies in the institutions willing and able to invest to develop the process.

Three SOPs for sequencing are proposed, outlining quality control of the DNA to be sequenced, describing the procedure and the quality control of the output, the sequencing reads.

Two SOPs are recommended for the assessment of the microbial community composition, based on the sequencing data; one for taxonomic and another for functional composition.

In parallel with SOPs, a method has been developed to cluster the genes coded by the same genomes and to assemble full genomes at a high quality. This generic procedure has enabled discovery of over 500 new species and ten-fold more smaller genetic elements, such as viruses and plasmids.

All the SOPs are publicly accessible at the IHMS web site http://www.microbiome-standards.org/#SOPS.

2. A SUMMARY DESCRIPTION OF PROJECT CONTEXT AND OBJECTIVES

Context

Humans live in constant association with microbes that are present on surfaces and in cavities of the human body, and even within our cells. The number of our microbial companions exceeds by at least ten-fold those of cells of our own body and the number of unique genes they encode is at least 150-fold greater than the number of genes in our own genome. This complex and dynamic microbiota has a profound influence on human physiology, nutrition, and immunity. Understanding the dynamic and variable nature of human microbial communities should lead to deeper understanding of human biology and open avenues to modulate the microbial communities in order to improve our health and well-being.

To progress towards these ambitious goals a number of actions have already been undertaken by the international scientific community during the past few years. The most advanced of these were integrated into several programs of a large scope, the EU FP7 project MetaHIT (http://www.metahit.eu) and the NIH Human Microbiome project (http://nihroadmap.nih.gov/hmp/). Among the shared goals of these programs, the most prominent were characterization of the human-associated microbial communities and of their variations in different human pathologies. To coordinate the activities carried out within different programs, the International Human Microbiome Consortium (IHMC, http://www.human-microbiome.org/), has been constituted and formally announced in October 2008. This has facilitated identification of a clear need to standardize the procedures in the Human Microbiome research. The overall concept of IHMS is to promote development and implementation of standard procedures and protocols in three separate but related fields:

(iv) Collecting and processing of human samples.

(v) Sequencing of the human-associated microbial genes and genomes

(vi) Organizing and analyzing the data gathered

Standardization in these 3 fields is fully recognized as being of crucial importance for the development of the Human Microbiome field, as it enables comparisons of results across large studies, accelerating the progress in this arena of ever increasing importance for human health and well-being.

Objectives

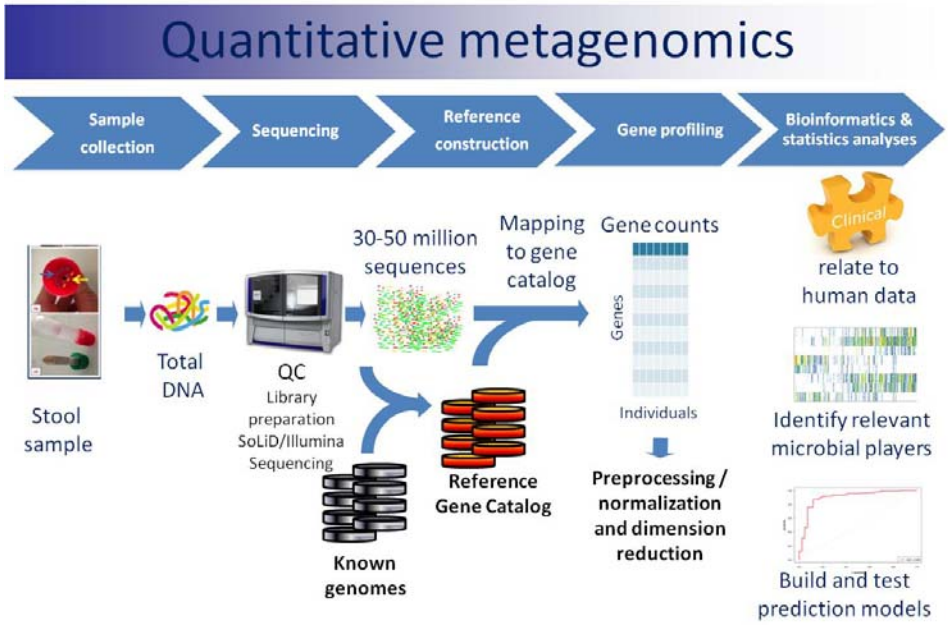
To reach this overall aim the project had the following objectives:

- Coordinate standardization of procedures and protocols within the existing Human Microbiome research programs and those yet to come.
- Gather and compare the protocols used to collect, identify and process human samples and aid to develop the standard operating procedures for sample collection, identification and processing
- Compare sequences of genes and genomes of human-associated microorganisms generated by various methodologies and approaches, aid to develop standards to define sequence quality and recommend procedures to reach the standards.
- Assess the approaches and procedures used to analyze the sequence data and the associated metadata and recommend standards for data analysis.

During the execution of the project we focused on one type of biological samples, the stools. The reasons for this choice were multiple.

1) Stool samples represent the gut microbial communities, which are most numerous and abundant in the human body. Their total mass can approach 2 kgs and they are often viewed as a neglected organ of the human body, which greatly impacts our health and disease.
2) Stool samples can be obtained in a fully non-invasive way. This greatly facilitates comparisons of large cohorts of patients and healthy individuals, which is necessary to detect alterations related to a disease. Indeed, it is unethical to submit healthy individuals to invasive interventions, when they cannot benefit from them, as is most often the case in cohort comparisons.
3) Gut microbial communities were the prime target of several large international studies, carried out by the MetaHIT consortium and HMP, but also by Chinese and European colleagues.
4) Other body site samplings were exhaustively addressed by the HMP project and the SOPs have been developed and posted. The IHMS partners deemed that re-visiting these would have represented non-justified duplication of effort and thus sub-optimal use of resources.

We also focused on the use of Quantittive Metagenomics to analyse the microbiome. The main reason was that the approach, which involves intensive sequencing of total DNA present in the stools, has the resolution well superior to the alternative, analysis of a single gene, encoding the 16 S ribosomal RNA. It informs on all the microbial genes harbored by an individual, allows to precisely determine the composition of the microbial communities at various taxonomic levels, from strains to phyla as well as the functions the community encodes.

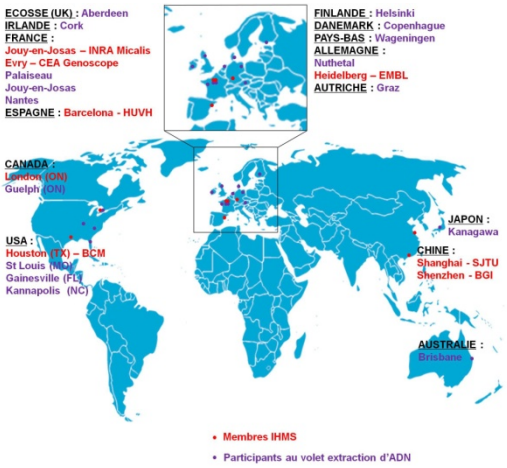


The standard operating procedures (SOPs) that are the main output of our project are obviously conditioned by these choices. As regards strictly sample collection, identification and processing, the SOPs are specific to stools. In contrast, SOPs for metagenomic DNA sequencing and data analysis, even if best adapted to the stool sample context (high microbial biomass, thus abundant DNA and a low ratio of human cells; this is different for most other human-associated microbial niches), are generic and can be useful in other contexts.

3. DESCRIPTION OF THE MAIN S & T RESULTS/FOREGROUND

Protocols to collect identify and process human samples

Ecological studies of environmental samples for the thorough characterization of their microbial communities are in essence difficult. A proper account of environmental complexity involves key steps towards the least biased representation. These involve i) sample collection procedures, ii) metadata collection and iii) nucleic acid extraction procedures. We have reviewed current practices for all three steps, based on available literature and activities of >20 research groups involved in human intestinal metagenomics, across 4 continents (Asia, Australia, Europe, North America).



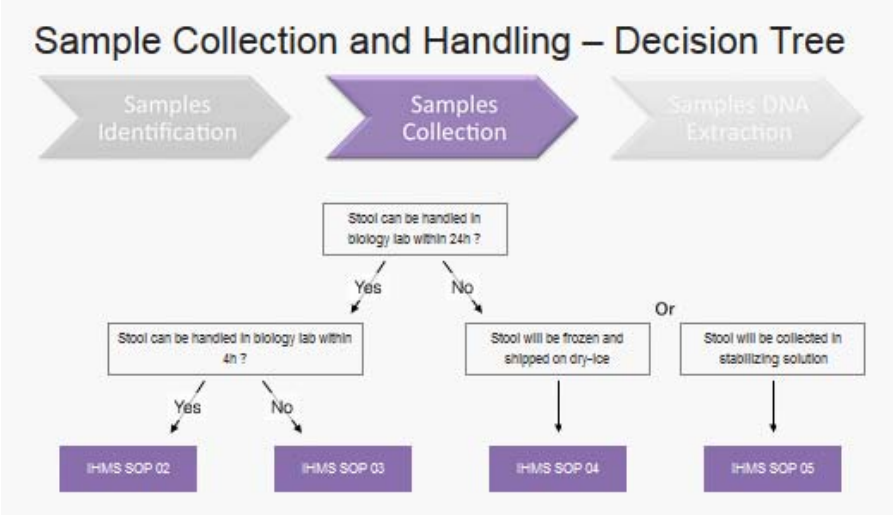
Sample collection

The procedures are most often not precisely reported and it turned out difficult to collect an exhaustive list of practices. They range from auto-collection of fecal samples at home, using a variety of collection devices and storage vessels and conditions, to collection in hospitals. Although the impact of this step in the overall process is likely to be critical, no systematic comparative assessment of practices has ever been undertaken. Yet, on the basis of what is known of the human intestinal ecosystem, it is possible to highlight key determinants of quality in sample collection procedures, shipment and storage.

Time: although quite obvious, it is essential that time be limited as much as possible between stool emission and cold storage. For samples collected at home, anticipation of delivery to a processin laboratory within less than one day is often attempted.

Temperature : Little is known for certain on the impact of temperature between the 37°C of stool in the human body and the -80°C target for storage. Precautions to limit enzymatic activity, especially in case of bacterial lysis, led to favour temporary storage and shipment at 4°C rather than ambient temperature, if freezing (at -20°) couldn't be rapidly achieved.

Anaerobiosis : It is quite evident that when storage at low temperature is bound to occur with several hours delay, ensuring anaerobic conditions will greatly limit deterioration of anaerobic bacteria and thereby preserve sample integrity.



Taking into account the review of the procedures and a long-standing field experience of the project partners, we have developed four standard operating protocols for various conditions of sample collection, accompanied by a decision tree to help select the most appropriate SOP. The main variable is time of transfer of the sample upon collection to the laboratory for processing.

- First SOP is for samples that can be transferred to the laboratory within 4 hours upon collection. The transfer can be done at room temperature, which greatly simplifies the process.
- Second SOP is for samples that can be transferred to the laboratory between 4 and 24 hours upon collection. Anaerobic conditions must be established for sample conservation. This is achieved by the addition of a commercial substance (Anaerocult) to the vessel that contains the sample. Room temperature can be used during sample transfer.
- Third SOP is for samples that can be transferred to the laboratory between 24 hours and 7 days, and be frozen immediately upon collection, at -20°. Samples must thereafter not be thawed and have to be shipped to the laboratory on dry ice.
- Fourth SOP introduces a stabilisation solution, which preserves microbial composition at room temperature. Samples can be shipped to the laboratory by currier mail.

The four SOPs conserve the composition of the stool microbial communities in a comparable manner, as assessed by quantitative metagenomics. Long term conservation (biobanking) of the samples requires in all cases storage at -80°. Storage of several separate frozen aliquots of a sample is advised, as thawing cannot be followed by re-freezing without alteration of the community composition. The SOPs and the decision tree are publicly accessible on the IHMS web site.

We have produce videos, which illustrate the procedures in a step-by-step manner, to accompany the written protocols and thus facilitate their use. The videos are also accessible on the web site..

Metadata

IHMC Common Data Fields

Field	Form	Comment
Subject identification	Code	
Sample type	stool, skin, etc.	
IHMC Member Institution		
Country		
Age	age (years)	
Gender	CV [male, female]	
Blood tests	CV [yes, no]	
Blood pressure	sys/dia mmHG	
Weight	kgs	
Height	m/cm (standing)	
Disease	CV [yes, no]	If "yes", disease specified in UMLS terms
HIV/AIDS	CV [positive, negative, no test, unknown]	
Smoking history	CV [no smoking, smoking, ex-smoker, unknown]	
Antibiotic usage	CV [yes, no]	If "yes", antibiotic specified per CV list
Notes:		
Age only, date of birth may be identifying.		
Non-US projects are using metric weight and height measurements.		
Blood tests: The standard is only yes/no if blood was taken.		
UMLS, Unified Medical Language System, http://www.nlm.nih.gov/research/umls/		
(CV=common vocabulary)		

Metadata is a general term used to encompass all descriptors that qualify the individual from which derives the stool sample. There has been to date little work on consensus definition of minimal sets of metadata, and international studies make the task complex because of the diversity of ethics requirements and protection of persons in different countries. Nevertheless, the International Human Microbiome Consortium has recommended minimal set of metadata including in essence: Age, Gender, Race, BMI, geographic localization. The recommended SOP is posted at the IHMS site.

DNA extraction procedures

Key determinants must be found in three main features of DNA extraction, bacterial lysis, contaminants removal and DNA recovery mode.

Bacterial lysis. Complex microbial communities are composed of diverse microbes that dramatically differ in their resistance to lysis, which is the primary step in environmental DNA

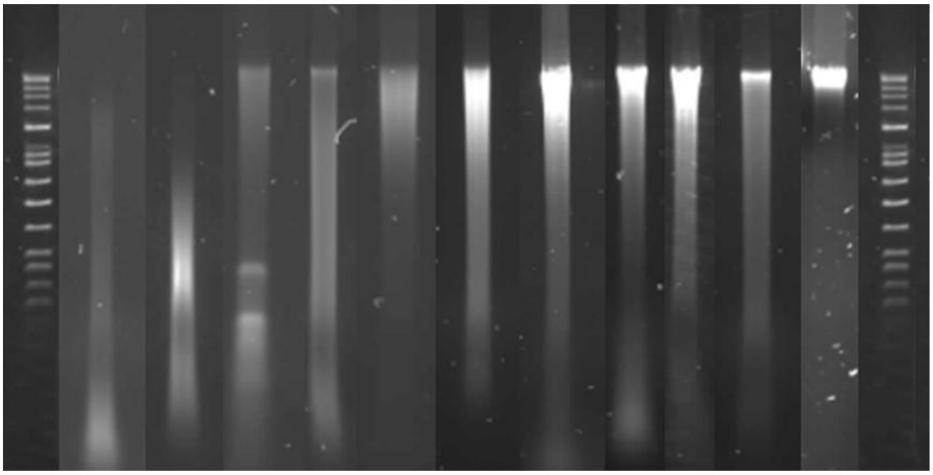
extraction. Microorganisms which are most resistant to lysis, such as gram-positive cocci and methanogenic archaea, require harsh physical and chemical treatments while those easy to disrupt, such as gram negative Bacteroidetes, may tend to yield denatured DNA under drastic treatment conditions.

Contaminants removal. The stool matrix most often contains aromatic constituents, such as scatols and fecols that may co-extract with DNA molecules. These molecules will often inhibit enzymes and require high dilutions of extracts to allow subsequent treatments (ligation or PCR amplification), required for sequencing.

DNA recovery mode. Classical mode of DNA recovery upon cell lysis involves alcohol precipitation. Yet numerous kits have been designed and are currently commercialized. They involve adsorption of nucleic acids on columns and elution-concentration.

Worlwide protocol testing

A total of 21 laboratories worldwide, including 6 IHMS partners, were participants in this study. They were provided standard samples (two stools and a bacterial mix) and returned DNA preparations obtained by their in-house protocols. Dramatic differences in the amounts (200-fold) and DNA quality (from fully degraded at worst to high molecular weight at best) was observed, as illustrated for 10 representative protocols, highlighting the importance of standardisation of this step of analysis.



82ng to 16µg DNA from 150 mg aliquots

Identification of most efficient protocols

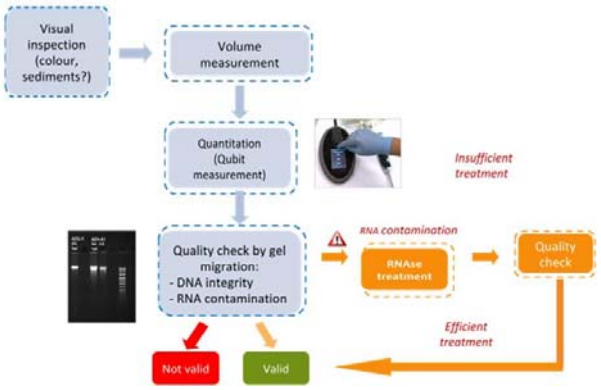
A set of 5 most efficient protocols were identified, by combining the criteria of DNA yield (>.5µg) and low degradation levels (less than 25% of DNA should be below 1.8 kb size). Of these, 3 involved use of a commercial kit (Qiagen). However, use of the kit did not guarantee high DNA quality, as 5 other protocols that also employed the kit did not perform well. Obviously, the way the kit is used impacts the quality of the resulting DNA. The 5 best protocols were re-tested in the participant laboratories, in order to assess their reproducibility and 'transferability'. Two SOPs are found to be most easy to implement and to give most reproducible results. One is adapted for manual use and can be of interest to laboratories that are involved in smaller scale microbiome studies. The other is based on a commercial DNA extraction procedure, which can be potentially fully automated and can be of interest to institutions involved in large scale studies, where the investment in automation appears justified. Both SOPs are accessible on the public IHMS web site.

DNA sequencing

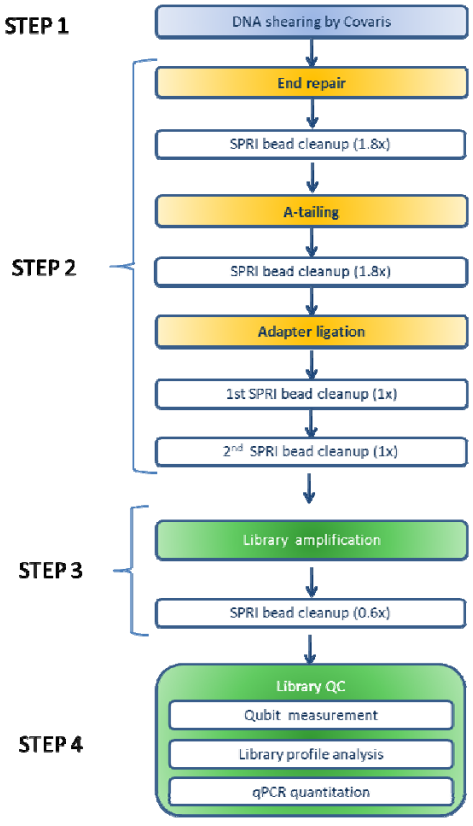
High throughput DNA sequencing is a central feature of microbiome characterization by Quantitative metagenomics. It allows, on the one hand, to construct the reference gene catalogs, by assembling the reads into longer contigs, which can be annotated, and thus to assess the genetic potential of the microbial communities from a given ecological niche. It captures, on the other hand, presence and abundance of each and every catalog gene in any individual under study and thus allows generating the gene profile of every individual. Comparison of the gene profiles of groups of individuals, in turn, allows identifying genes, species and functions associated with particular traits of interest, such as health or disease, nutrition or development, to name a few. The quality of these read-outs is critically dependent on the quality of the DNA sequencing reads. As a main output of the project we have established SOPs for DNA Sample Quality Control, for generation of Metagenomic short reads and for Quality Control of short reads.

DNA Sample Quality Control

The first SOP allows assessing the quality and the amount of the DNA to be sequenced, according to the workflow outlined in the figure. Step-by-step procedure and the decision thresholds are posted on the IHMS site.

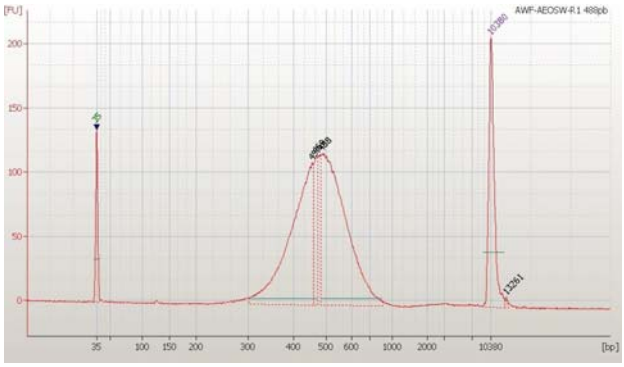


Generation of short Metagenomic sequencing reads



The second SOP describes the short read metagenomics generation in a step by step manner. The purpose of this procedure is to prepare indexed paired end libraries with 180 - 680 bp insert size DNA library that will be used for sequencing on the Illumina HiSeq2000 on 100 bp paired end lengths. Starting material is 250 ng genomic DNA extracted from fecal samples. The protocol is intended for low-throughput, manual library construction, but it is designed to be automation-friendly, easing the transition to automation if needed by throughput requirements. Genomic DNA is broken into smaller fragments via Covaris E210 instrument. Then end repair, A tailing and barcoded adaptors ligation are performed manually or by a liquid handling system. After PCR enrichment, library purification, qualitative and quantitative assessment of the library are performed. Quality control includes a Qubit measurement, library profile analysis by capillary electrophoresis and qPCR. These last steps as well as library purification are also easily amenable to automation by means of liquid handling systems.

We have initially focused on different key issues: the library preparation protocol, the extent of sequencing and the length & type (single or paired end) of sequencing, targeting the Illumina technology, used in most Quantitative Metagenomics studies. Library preparation was optimized for reproducibility and robustness, simplicity and scalability of the workflow. With these improvements, we are able to build, in parallel, 10 libraries in 5 hours on an automated system, with a reproducible high quality perfectly suitable for sequencing.

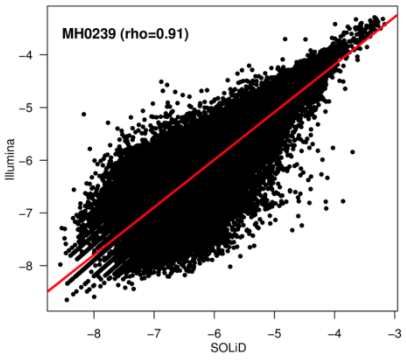


Quality Control of short reads

This SOP describes how to remove low quality reads from Illumina sequencing, how to remove unexpected fraction of reads and extra sequences due to internal calibration process. Sequencing is performed on a Illumina HiSeq2000 or HiSeq2500 sequencing. Demultiplexing of libraries is achieved using the bcl2fastq tool (version 1.8.3) from the CASAVA package (https://my.illumina.com/) , with at most 1 mistach allowed for index sequence detection. Starting from each end of the read, bases with quality < 20 are removed. 20,000 read pairs are randomly picked, and library construction adapters and sequencing primers are detected using tools from the fastx_toolkit suite (version 0.0.13,2) (http://hannonlab.cshl.edu/fastx_toolkit/). A maximum of 3 mismatches is permitted for adapter sequence recognition. A minimum of 0.5 % of the reads bearing a given adapter is mandatory to trigger adapter removal. From the whole read set, library construction adapters and sequencing primers are detected following criteria above, and for each read the sequence in 5' of the adapter is kept. The longest subsequence containing at most 1 N base is kept for subsequent steps. Reads shorter than 30 bases are discarded.

Different sequencing technologies

Two fundamentally different technologies are used for metagenomic sequencing. One, developed by Illumina, is based on DNA polymerization whereas the other, developed by Lifetech, is based on DNA ligation. We have sequenced total DNA extracted from 24 different stool samples by the two and compared the abundances of 3.3 million gut intestinal microbial genes computed from the sequences. The correlation is very good, close to 0.9, indicating that the studies using either technology can be compared.

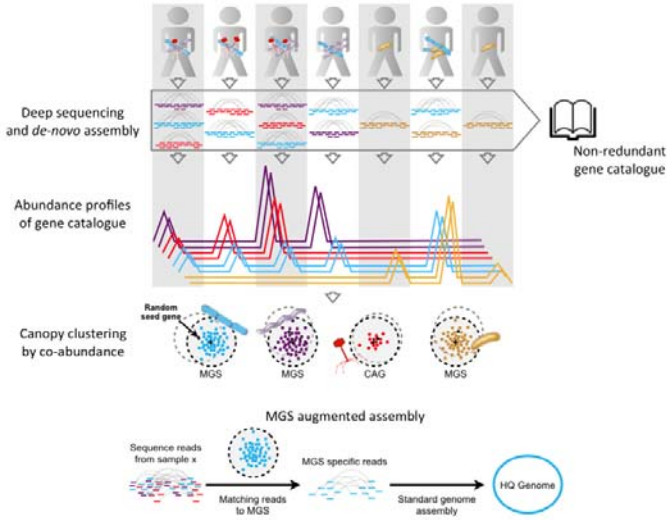


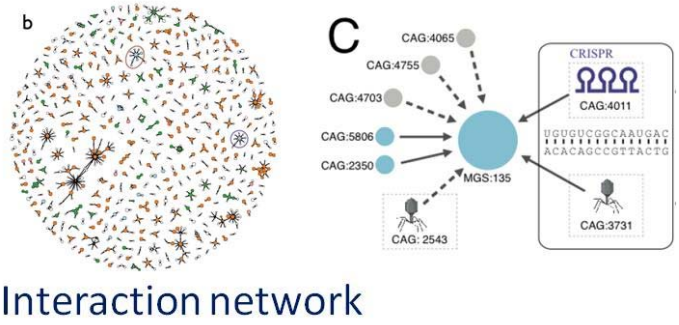
To further document this comparability, we have compared the diagnostic potential of the two methods, using the same gene-abundance based algorithm in two study populations. One was composed of 292 Danish individuals and the other of 49 French individuals; the former was analyzed by the Illumina and the latter by the Lifetech technology. A similar diagnostic accuracy, of 96% and 95% was obtained in both cases. This shows that the results of the two technologies are highly comparable.

Long contiguous reference sequence

Current high throughput sequencing methods all produce short reads, which are assembled into longer contiguous sequences. First methods to assemble metagenomics reads were established in teh MetaHIT project by the IHMS partners and were used to establish the reference gut gene catalogue. Among the approaches used to extend gene-length continuous sequences into longer contigs was deemed desirable, as the genomic neighbourhood of a gene can be very informative. The most direct approach would be to cultivate a given microbial species from the gut and sequence its genome. However, the approach is hampered by our inability to culture an overwhelming majority of teh gut microbial species. An alternative approach tested by the scientific community was isolation of single cells of a species and sequencing of their genomes. This approach turned out to be very challenging and did not yield many sequenced genomes.

To address the challenge, we developed an innovative approach, based on an exhaustive and unsupervised co-abundance gene binning across a series of complex metagenomic samples. The method is very efficient and allows clustering genes encoded by the same genomes. In such a way we have identified 741 groups of over 700 genes, that we term metagenomic species and that correspond to gut bacterial species, taht we denote metagenomics species (MGS). Furthermore, we have also discovered by the same procedure over 6000 smaller genetic elements (phages, plasmids, CRISPR…), which co-exist with the bacterial species. This is illustrated by a simple example. Obviously, a phage cannot be present in an ecosystem if its host bacterium is absent. We have organised large and small metagenomics units in a dependency network, of over 800 relations, which describes human gut microbiome at unprecedented detail.





Gene clusters corresponding to the bacterial species were used to identify the contigs and scaffolds that encode them. For a given sample the reads were aligned using Burrows-Wheeler Aligner to the MGS specific scaffolds and the mapped reads, including unmapped mates, were extracted. These reads were then corrected by Quake46 and reassembled with Velvet. As several samples were used for assembly of each MGS, the best assembly was selected based on ranking of contig N50 and the number of contigs in the assemblies. Contigs with low coverage were removed from the assemblies. The contigs and scaffolds were then filtered to 100 and 500 bp minimum lengths, respectively, and gaps in scaffolds were filled using SOAPdenovo GapCloser.

To assess the quality of the assemblies, we adopted the six criteria from the Human Microbiome Project (HMP); five address the contiguity of the assembly and one the genome completeness, by counting core genes contained in the assembly. The criteria are i) 90% of the genome assembly must be included in contigs > 500 bp, ii) 90% of the assembled bases must be at > 5 X read

coverage, iii) The contig N50 must be > 5 kb, iv) scaffold N50 must be > 20 kb, v) average contig length must be > 5 kb and vi) > 90% of the core genes must be present in the assembly. In total 360 sample-specific MGS augmented assemblies, from 247 unique MGS passed all six criteria. In addition, 139 unique assemblies passed five criteria. Over 85% of these had no closely related reference genomes and correspond to previously unknown bacterial species.
Our method, which favourably complements other approaches used to generate long contiguous metagenomics sequences, was reported in Nature Biotechnology in 2014.

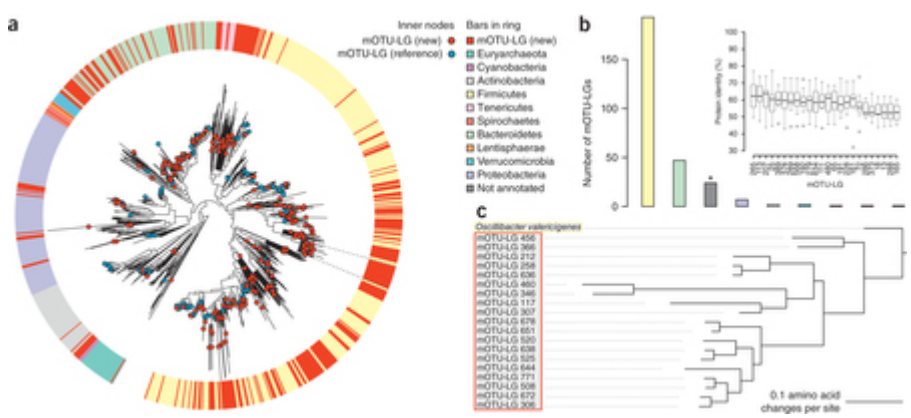
Data analysis

We have developed a quality control procedure that strikes a good balance between the amount and quality of sequence data. These standards have already been adopted within the MetaHIT project and the procedure has been used to process >1200 gut metagenome samples. Using these standards has greatly improved (i) the assembly of typically millions of Illumina reads into longer contiguous sequences, which serve as a basis for predicting microbial genes and (ii) the alignment of Illumina reads to reference sequences, which is pivotal in quantifying the taxonomic and functional composition of metagenomic samples.

Phylogenetic analysis

Phylogenetic composition is a major readout of microbiome analysis. We have developed two products to achieve assessment of the phylogenetic composition. First, we developed an algorithm that defines prokaryotic species based on suitable phylogenetic marker genes and released software that can be used to classify genomes. A web-server has been implemented to ease the use for non-experts. Second, this method was adopted to enable the phylogenetic profiling and calculation of ecological indices based on shotgun sequencing data. The main advance over previous methods is its ability to quantify more than half of the microbial species for which no sequenced reference genomes are available to date. The SOPs are posted on the IHMS web site.

Microorganisms are ubiquitous in all natural environments, including host associated habitats, where they play key roles in biogeochemical and nutrient cycling. Studying the structure and function of microbial communities is thus essential for advancing our understanding of ecosystem processes and their role in human health and disease. A common approach for profiling microbial communities involves the sequencing and classification of amplified 16S ribosomal RNA-encoding gene (16S rDNA) fragments using DNA directly isolated from environmental samples. Due to its universality for prokaryotes and the availability of large, curated reference databases[1-3], the 16S rDNA represents a powerful phylogenetic marker despite (i) biases introduced by gene copy-number variations[4-6], (ii) variability in amplification efficiency[7,8], (iii) inconsistencies when targeting different regions of this gene[9], and (iv) known issues in accurately and consistently delineating prokaryotic species[10]. In contrast to this single gene-targeted approach, shotgun sequencing of metagenomes generates millions of short reads that are randomly sampled from microbial community genomes. This approach commonly involves the alignment of reads to taxonomically annotated reference genomes[11] and results in a read-abundance distribution into taxonomic bins. However, without appropriate normalization by genome size, which has to be estimated for uncharacterized species, taxonomic abundance estimates may be highly biased. Alternatively, conserved phylogenetic marker genes, clade specific or universal, that are both present as single copies in most genomes and rarely subject to horizontal gene transfer represent ideal candidates for taxonomic profiling of environmental samples.

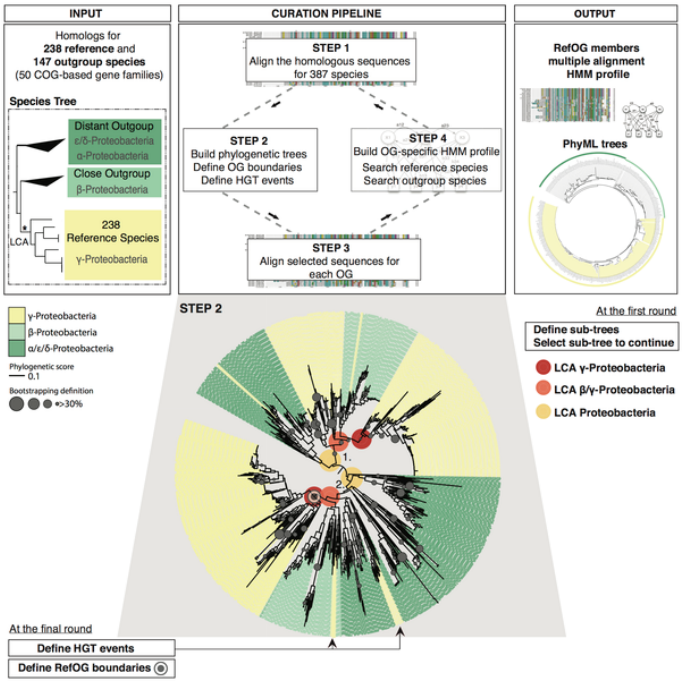


We developed a method based on single copy marker genes (MGs) that provide prokaryotic species boundaries at higher resolution compared to the 16S rDNA to estimate relative abundances of metagenomic species. It uses MG sequence from metagenomes and reference genomes that are clustered into molecular taxonomic units (mOTUs) at MG-specific species-level cutoffs. By applying this method to human faecal samples, we first determined the number and the abundance of species that were not represented by currently available genome sequences. We next exploited the availability of mOTU abundance data to link genes of common species origin into mOTU linkage groups (mOTU-LGs) based on their co-variance across multiple samples. Evaluation of the performance of our method showed that it could estimate community similarities at previously unachieved accuracy. To illustrate the utility of our method, we compared species diversities of gut microbial communities from individuals sampled in the context of the Human Microbiome Project (HMP) and the European MetaHIT initiative and also found that the majority of the differentially abundant species in samples from ulcerative colitis patients were lacking genomic resources, that is, they could not have been identified using reference-based methods.

Our method was reported in Nature Methods in 2013.

Functional analysis

Interpretation of metagenomic readouts of human microbiomes depends for applicability on the possibility of functional interpretation, which necessitates knowledge of microbial gene functions. Therefore, microbial gene functions need to be identified in any sample that is processed.
Gene functional knowledge is commonly derived from single-species investigations and therefore linked to known taxa, whereas metagenome analyses typically concern a vast multitude of poorly characterized strains. Therefore, techniques needed to transfer knowledge from laboratory systems to novel genes such as those found in metagenome reference gene catalogs, in a way which does not depend on knowledge of the exact species origin of such genes are needed. This is best accomplished by using resources encompassing orthologous groups, which are gene families radiating from single entities in appropriately-chosen ancestral microbes. Research has shown that assignment of novel genes to such groups and transferring functional annotations across genes in the same orthologous groups, provides a powerful and accurate method for annotating either assembled genomes or metagenome reference gene catalogs with functional terms, either as free descriptions or using controlled vocabularies such as the Gene Ontology or the COG functional categories.
We have assessed approaches for functional gene annotation that were used previously and outline a robust strategy for annotation of a collection of genes, either derived from a single genome or from a metagenome, with functional assignments drawing from orthologous groups.



Benchmarking orthology resources on the basis of conservation of protein function or domain architecture across orthologous versus non-orthologous proteins is inadvisable. Rather benchmarking of the reliability of an orthology resource should be done with respect to how well known evolutionary relationships in curated “gold standard” datasets are recaptured. We have therefore developed larger and also prokaryote-focused benchmark datasets which can be used to ascertain the applicability of orthologous group resource construction approaches for the purpose of functional annotation of human microbiome data. The curated gene families, called Reference Orthologous Groups, are publicly available.

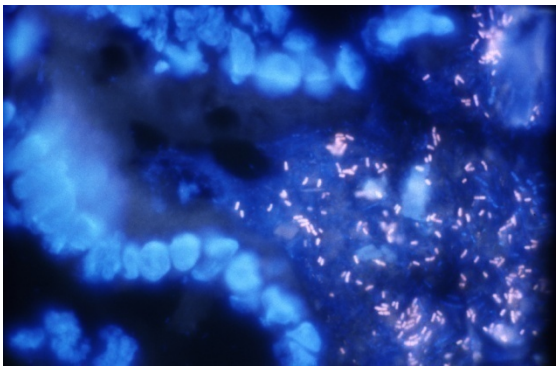
Given the choice of an orthologous group resource shown to be robust in regards to bacterial gene orthologous group assignment, there are different options for assigning novel genes, such as those called within a metagenomic reference gene catalog, to orthologous groups. Groups can be rebuilt containing novel and well-known genes. However, this is computationally intensive in many cases, and may also reduce the fidelity of the orthologous groups considerably for several reasons: introduction of low-quality, chimeric or miscalled genes and biasing the species sampling of the groups among them.

The approach we recommend as being more appropriate is to assign novel genes to the groups they show the highest similarity to, taking into account potential confounding by multi-domain protein architectures. The simplest approach involves pairwise sequence similarity searches (e.g. BLAST); these however are both less sensitive and less accurate than profile-based methods such as Hidden Markov Model (HMM) searches. In the eggNOG 4 resource, tools are available for assigning novel genes from genomes or metagenomes to this collection of orthologous groups, presently based on pairwise similarity.

Both specific studies and large-scale statistical investigation shows orthology to be generally associated with conservation of gene function, though the specific process of annotation transfer can be done in different ways. Functional annotation can in some cases be assigned to an orthologous group as a whole, recorded as such and transferred to all novel group members. This is the case with the descriptions and functional categories assigned to manually curated COG and KOG orthologous groups, as well as their automated extension in eggNOG and the descriptions and pathway membership information available in the commercially licensed KEGG database. However, not all functionally relevant information about members of a group is always captured in this summary information. Alternatives include performing enrichment analysis on functional terms assigned to annotated group members, then transfering any terms to novel members that is significantly characteristic of the group. The most complex approach lies in constructing a gene family tree for all members of an orthologous group, inspecting this tree for clades that correspond to particular specific subfunctions, and transfering annotations based on the position of a novel gene in this tree.

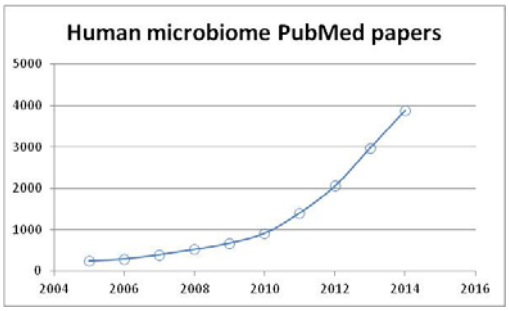
4. THE POTENTIAL IMPACT

Human gut microbiome is a neglected organ of our body



Humans live in associations with microbes, which outnumber own cells 10-fold. Most abundant and complex of these microbial communities is situated in the gut, and can represent up to 2 kg of mass. The community has a huge metabolic potential and provides to our body substances required for our health, exemplified by vitamins, amino acids or short chain fatty acids, which are important nutrients and provide regulatory control of the digestive system. It creates a barrier which protects our body from invasion by pathogens and educates our immune system thus enhancing the protection against external threats. Gut microbial community composition appears to be altered in many chronic diseases, which have been ever increasing in industrialized societies such as European Community over the past decades, suggesting the role of the communities in these diseases. A view that they represent a neglected organ of our body is thus fully justified.

It is thus not surprising that human microbiome attracts tremendous attention. That of the academic communities can perhaps best be illustrated by a spectacular exponential increase of the number of publications that target it over the past 10 years. A number of international meetings are being organized to address this theme; a prestigious International Human Microbiome Congress (the 5th was held in Luxembourg earlier this month), is among these. But industry is becoming increasingly attracted as well to the field, with a number of start-up companies (Enterome as one of the first and most visible exemples, Ceres, 2nd Genome and others). Bridging the two worlds, academic and professional, is ongoing, via meetings such as The 1st Annual Translational Microbiome Conference, organized in Boston next month or the 2nd & 3rd Microbiome Forum, which will be held in London and San Diego, in May & September, respectively.



Need for standardization of the human gut microbiome research

Assessment of the status of the gut microbiome in each and every individual underlies basically all studies of the impact of the microbiome on health and disease. The assessment of this complex organ involves an integrated succession of steps, each of which addresses different aspects of the process. First are sample collection and processing, from volunteers to the laboratory, next the generation of DNA sequences, which captures presence and abundance of microbial species and sub-species, level genomic elements and last the translation of this quantitative information into microbiome descriptors, such as gene profiles. The descriptors are used, in turn, to identify the elements of the microbiome that are associated to bioclinical parameters and are likely to impact health and disease.

The world-wide practice of microbiome assessment in different large studies did not follow common standards for any of the outlined steps. As a consequence, the results of the studies were most often different. This can be illustrated by two examples. One is the comparison of the overall microbiome richness of the individuals enrolled in the Human Microboiome Project of the NIH and the MetaHIT program of the EC. HMP participants had a much lower richness than the MetaHIT ones. Does this reflect a real biological difference between Americans and Europeans or the artifact of different methodology used? Another is the microbiome alteration reported for the type 2 diabetes patients in China and in Sweden. In the two studies patient microbiome greatly differed from that of the healthy individuals, but the differences were largely not the same. Are the alterations due to the differences in the disease on the two continents or do they reflect methodological artifacts?

The potential impact of the IHMS research

Clearly, the impressive worldwide growth of interest in the microbiome research, illustrated in the above section, cannot be sustained if the microbiome studies, which aim to address critical issues of human health and disease, do not yield reproducible answers. The main outputs of the IHMS research, Standard Operating Procedures, address this need.

Socio-economic impact

Sample collection and storage

All microbiome studies start with sample collection. Indeed, once the samples are correctly collected and stored, without altering the microbial community composition, they can either be analyzed immediately, within the context of a given study, or can be kept in biobanks, essentially indefinitely, and thus be available for further analyses in the future, in ways that could not have even been foreseen at the collection time.

The immediate impact of the sample-related SOPS generated by the IHMS consortium will be to enable studies that will be imminently comparable. On the short-to-medium term time scale, this is of utmost importance for the academia, which strives to advance knowledge and understanding of human biology. On the one hand, the results of the studies will be more easily replicated, and on the other, the possible robust differences will be identified and will guide progress toward novel biological inferences. The knowledge will therefore more efficiently advance. This will of course be one of the main drivers of the human microbiome field.

Beyond academia, study comparability is of a perhaps even greater importance for the industry, in view of the multi-billion health-related issues it contends with. Indeed, for the pharmaceutical economic area, development of microbiome bio-markers based diagnostic and prognostic approaches, which can be robustly used to assess health status of an individual and monitor response to treatments, is of a high interest.

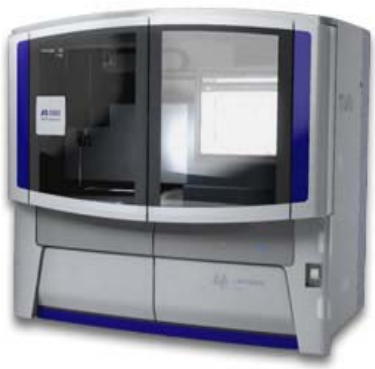
On a medium-to-longer term scale the sample-related SOPs could perhaps have even higher impact. Given that the chronic diseases progress constantly over the past decades in the industrial societies and that they can neither be prevented nor cured, identification of individuals at risk to develop them is critically important, as the risk-alleviating treatments could then be conceived and implemented. The time span to develop these diseases is nevertheless long. How could then the appropriate risk markers be identified? An answer is in the biobanking of samples from large cohorts and the analysis of samples from individuals that had adverse events over time, together with matched controls. The pre-requisite for te strategy is correct collection and storage of the samples, exactly the issue addressed by the IHMS SOPs.

Sample processing

As with SOPs for sample collection, SOPs for sample processing, that is DNA extraction, will enable studies that will be largely comparable, with an immediate short term impact. Indeed, in contrast to sample collection, which requires longer term organization congruent with the execution of clinical studies, involving possibly ethical amendments, the DNA extraction SOPs could be implemented very shortly across many labs that are already involved in the human microbiome work and the ones that are entering it. Immediate posting of the protocols on the IHMS web site, ahead of a scientific publication which is being currently prepared, is perceived as accelerator of the field. To most efficiently achieve this goal, the SOPs were presented at the recent International Human Microbiome Congress, earlier this month, very close to the end of the project.

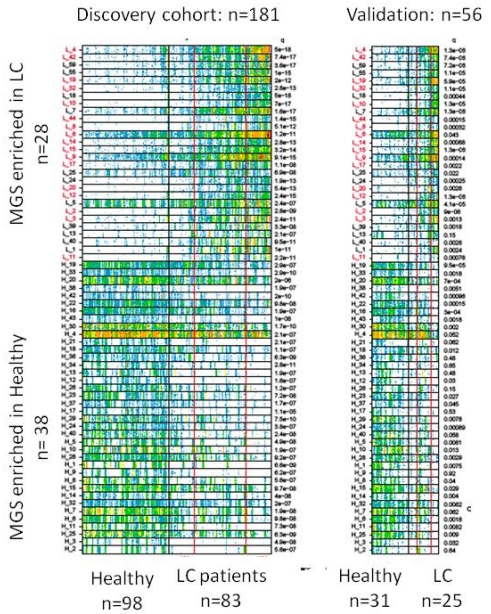
But the SOPs are not meant to freeze the field at its current state. For instance, implementation of the current protocols on automated platforms will greatly increase the throughput, decrease repetitive labour, increase the quality and reduce the cost of the process. Also, new technologies for downstream analysis, namely DNA sequencing, are anticipated and the requirements for them may be different than for the current ones, which will lead to the evolution of the SOPs. However, the impact of the IHMS will be in providing a benchmark against which further developments will be measured. In this regards, the SOPs will impact the short-to-medium term research and exploitation in the microbiome field.

DNA sequencing

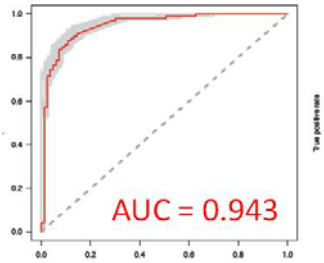


SPOs for DNA sequencing will also help to render comparability of the metagenomics studies. Indeed, the quality of the short sequencing reads impacts the downstream analysis and most notably generation of gene profiles, which is a staple of the determination of the microbiome composition. The SOPs address the quality of the DNA used for sequencing, the procedure to generate the reads and the assessment of the read quality. Of course, the impact is on a short to medium term, as it is anticipated that the sequencing

technology will continue developing further. Nevertheless, our SOPs will be useful for proper benchmarking of new technologies.



Development of the method for assembling genes from the same species into clusters by co-abundance binning and generating full genome sequence for a high fraction of the clusters will impact greatly the studies yet to come. First, for the human gut, we have already discovered numerous new species, which were previously totally unknown. For many we have shown to be associated to a disease and perhaps even more importantly to a risk of a disease. This discovery enables much deeper analysis of the alteration of the microbiome than previously possible, both on taxonomic and functional level. This alone will render microbiome field more impactful, both for knowledge generation and for practical use, such as development of robust diagnostic tests and for identifying novel targets for interventions, possibly aiming to modulate the microbiome composition and bring it closer to a healthy state. For instance, when metagenomic species were used rather than the gut microbial genes the precision of diagnostic of liver cirrhosis was increased (from 84 % to 94%) and the test was rendered simpler (15 biomarkers for the former, only 7 for the latter).



But our approach is not limited to human gut. It is totally generic and can be applied to almost any ecological system where enough individuals are examined. This was already shown for two model systems important for human health, the mouse and the pig gut microbiome (in preparation for publication, with contribution of several IHMS partners). We expect that the approach will be very broadly used in many ecosystems and impact the related fields and studies.

Data analysis

The method we have develop to assess the taxonomic composition of the microbial community will impact the microbiome research greatly, as it is more comprehensive and precise than the generally used method, which relies on the 16S gene encoding bacterial ribosomal RNA. It is also generic and can be extended to other ecosystems, as it relies on a set of taxonomically conserved genes throughout the bacterial tree of life. Similarly, the clusters of the orthologous genes and the procedures to construct them will enable functional analyses across a plethora of ecosystems.

Main dissemination activities

The main outputs of the IHMS are different SOPs. Their principal dissemination is via the IHMS web site. Since their posting at the beginning of this month they have been downloaded 184 times. Given the number of laboratories involved in this emerging human microbiome field this is a tremendous success. INRA will sustain accessibility of the SOPs through the IHMS web site which is hosted by INRA.

Beyond posting on the site, our dissemination activities were via scientific publications and presentation at the meetings. We have published IHMS results in journals of the Nature press

group, which is also a great success, given the mainly technical content of IHMS, which is protocol standardization. A paper describing the gene clustering method and genome reconstruction, published in Nature Biotechnology last year has been viewed over 10 000 times and is the 5th on the altimetrics list of all articles from that journal of comparable age. It was accompanied by an editorial, underlining the attention that the scientific community is paying to the IHMS studies. Beyond the scientific community, public media also reported on the study.

IHMS activities were also regularly reported on at international meetings. Most recently the SOPs were presented at the 5th International Human Microbiome Congress, held in Luxemburg last month.

5. THE WEB SITE

The intranet and internet website for IHMS have been developed on the domain: www.microbiome-standards.org.

The public domain (henceforth "the internet") includes the presentation of the overall project objectives and structure, descriptions of the teams involved in the project and description of the work packages. The restricted domain (henceforth "the intranet") contains project management and organization links and documents, namely Deliverables' and Milestones' descriptions as well as the documents related to the Grant Agreement and Negotiating Instructions.

Sample collection and processing SOPs, sequencing and data analysis SOPs have been established and are posted on the public web site http://www.microbiome-standards.org/#SOPS.