

The sbv IMPROVER Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC)

Technical document

Version 2.0 Final, 9.09.2019

Table of Contents

Table of Contents

Table of Contents	1
1 Datasets description	2
1.1 Training datasets	3
1.1.1 IBDMDB/Schirmer et al. dataset	3
1.1.2 He et al. dataset	3
1.2 PMI testing dataset	3
2 QC and samples selection	4
2.1 QC pipeline	4
2.2 Sample selection	5
3 Generation of taxonomy and pathway abundances matrices	6
3.1 Taxonomy abundances matrices	6
3.2 TaxID description file	7
3.3 Pathway abundances matrices	7
3.4 PathID description file	8
4 Class labels	9
5 Data files for download	10
6 Participants' tasks and submission on the sbv IMPROVER Challenge website	16
6.1 Tasks	16
6.2 How to submit your results	17
7 Scoring	18
8 Abbreviations	19
9 Annex 1. IBDMDB data download	19
10 Annex 2. He et al. data download	19
11 References	20

1 Datasets description

For the *Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC)* (herein referred to as “Challenge”), the organizers propose the use of metagenomics data from two published human studies (1-3) as training datasets, and from one internal human study as test dataset. The datasets are provided as quality controlled raw and processed data as described below. The participants have the freedom to use additional public and private shotgun metagenomics datasets for training their models, providing that they supply a description and source of the dataset(s).

The Challenge is split into two sub-challenges:

- For the sub-challenge 1 (“MEDIC RAW”), shotgun metagenomics sequencing reads from human fecal samples are provided in the form of pair-end gzip-compressed fastq files (<SAMPLE_NAME>_<READ_NUMBER>.fastq.gz).
- For sub-challenge 2 (“MEDIC PROCESSED”), taxonomy and pathway relative abundances matrices generated by processing raw data in different analysis pipelines (section 3 for detailed description) are provided in the form of tab-delimited files.

All class labels associated with samples from both training datasets are also provided. The class labels for the testing dataset are not provided and constitute the Gold Standard of the Challenge. The training and testing datasets will be released altogether.

The summary of the datasets content can be found in Table 1.

Table 1 Datasets content summary.

Dataset	Provided for	Total samples	Non-IBD	CD	UC
IBDMDB/Schirmer et al. (2,3)	Training	54	14	23	17
He et al. (1)	Training	116	53	63	0
PMI	Testing	105	Not disclosed	Not disclosed	Not disclosed

1.1 Training datasets

The training datasets by The Integrative HMP (iHMP) Research Network Consortium (2) and He et al. (1) comprise shotgun metagenomics data from fecal samples of non-IBD, CD and UC human subjects. A detailed description on the generation of each dataset is provided in the Materials and Methods section of the respective publication.

1.1.1 IBDMDB/Schirmer et al. dataset

The first training dataset includes paired-end whole genome sequencing reads from a publicly-available longitudinal study conducted in North-America (2,3) as a part of the second edition of the human microbiome project, namely the integrative Human Microbiome Project (iHMP, <https://www.hmpdacc.org/ihmp/>). A total of 1338 paired fastq files were downloaded from <https://ibdmdb.org/tunnel/public/HMP2/WGS/1818/rawfiles> including raw data obtained from fecal samples of adults and children collected at multiple time points. Following quality checking, and after selecting one sample per adult subject, the dataset provided for training includes a total of 54 samples from 23 CD, 17 UC and 14 non-IBD subjects. Please, refer to paragraph 2.2 for details on quality check and data processing.

1.1.2 He et al. dataset

The second training dataset comprises paired-end whole genome sequencing reads from a publicly-available cross-sectional study conducted in China (1). A total of 123 paired fastq files were data downloaded from <http://gigadb.org/dataset/100317> including raw data obtained from fecal samples of adults. Following quality checking, the dataset provided for training included a total of 116 samples from 63 CD and 53 non-IBD subjects. Please, refer to paragraph 2.2 for details on the quality checking and data processing.

1.2 PMI testing dataset

The PMI testing dataset consists of 105 paired-end whole genome sequencing of fecal samples from CD, UC and non-IBD. Please, refer to paragraph 2.2 for details on quality check and data processing.

To generate the dataset, whole-genome DNA was extracted from fecal samples using the QIAamp DNA Stool Mini Kit (Qiagen) according to manufacturers' instructions. For shotgun libraries creation, DNA was fragmented using the Covaris S220 ultrasonic fragmentation system. A DNA library was created from the obtained fragments using the NEBNext Ultra II kit (NEB, USA) according to the manufacturers' instructions. This kit is designed to prepare libraries of samples containing from 500 pg

to 1 µg of DNA. The quality assessment of the obtained libraries was carried out on HiSens chips of the device 2100 Bioanalyzer (Agilent Technologies, USA). The concentration of the DNA library was determined using a Qubit 2.0 fluorometer (Invitrogen, USA). Sequencing of libraries was performed on the NextSeq 500 Illumina platform (USA).

2 QC and samples selection

2.1 QC pipeline

QC pipeline described below was used to assess the quality of samples from each dataset. The reasons behind using this pipeline were to filter out potential human and technical contaminating sequences as well as to confirm the good overall quality of remaining sequencing reads.

The schematic representation of the QC pipeline is presented on the Figure 1.

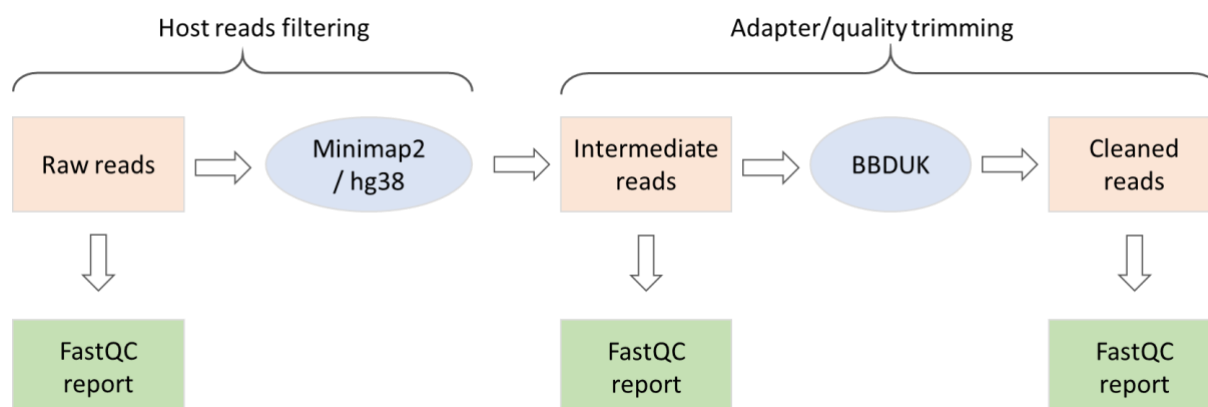


Figure 1: Schematic representation of the QC pipeline.

Raw reads were mapped on the human genome (hg38) using Minimap2 (version 2.8) aligner (PMID: 29750242) with the option `-a` (CIGAR and output alignment in SAM format) and `-x sr` (short single-end reads without splicing). Unmapped reads were collected using SAMtools (version 1.7) `view` (4) command with the `-f 12` SAM flag (read unmapped, mate unmapped). Obtained reads were subjected to contaminants and adapter trimming with the BBDuk program of the BBTools toolkit (version 37.99) (Bushnell B. BBDuk short read aligner, and other bioinformatics tools. 2014 <http://sourceforge.net/projects/bbmap>.) with the *k* size set to 23. QC reports for the raw, pre- and post-trimming reads were generated using FastQC (version 0.11.6) software (Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and gathered together using MultiQC (version 1.7) module for Python (5).

2.2 Sample selection

Sample selection was based on the sample metadata and the results of the aforementioned QC pipeline. One common criterion for all samples selected for this project was the age of the subject to whom the sample belong (older than 18 years) as adult subjects. There were additional criteria specific to each dataset (Table 2).

Table 2 Criteria used for the samples selection

Dataset	Provided for	Reads pairs in the post-QC data	Subjects' age	Additional criteria
IBDMDB/Schirmer et al.	Training	$>10 \times 10^6$	≥ 18 years old	One time-point sample (the earliest one) per subject
He et al.	Training	$>20 \times 10^6$	≥ 18 years old	Included samples that are not marked as "host contaminated" in the original research metadata
PMI	Testing	$>20 \times 10^6$	≥ 18 years old	Confirmed IBD diagnosis as CD or UC

3 Generation of taxonomy and pathway abundances matrices

3.1 Taxonomy abundances matrices

Taxonomy abundances matrices for all three datasets (training datasets 1 and 2 and testing dataset) were generated using the final output of the aforementioned QC pipeline (see paragraph 2.1).

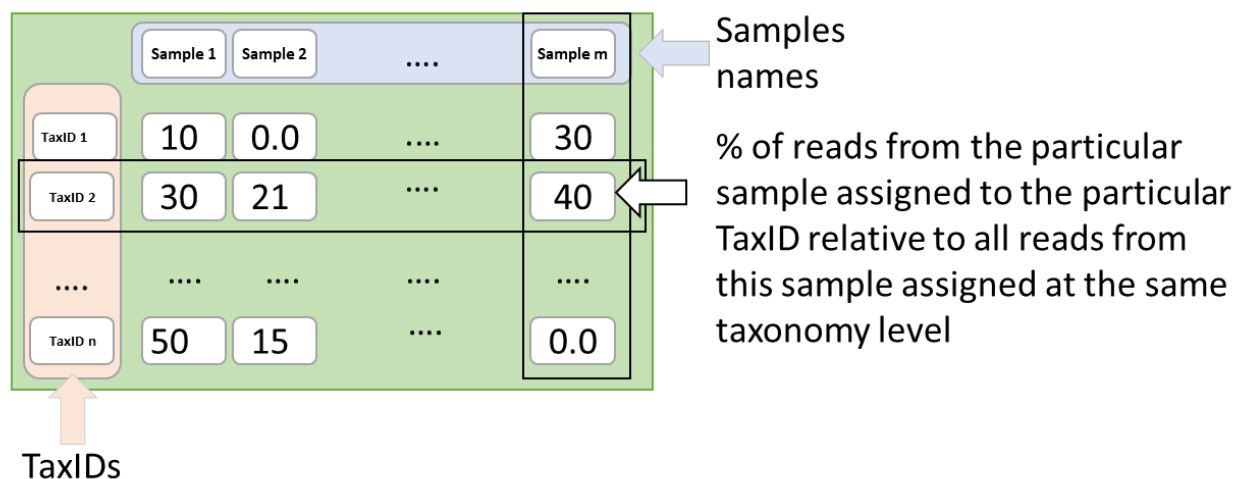


Figure 2: Schematic representation of the taxonomy matrix structure.

Each sample was subjected to taxonomic profiling using Kraken2 (6) for the initiate reads classification and Bracken (version 2.0) (<https://doi.org/10.7717/peerj-cs.104>) for the classification correction at the Species, Genus, Family, Order, Class, Phylum and Superkingdom levels. Kraken2 reference database was build using the *--standard* option that enforces the download of the RefSeq (7) bacteria/archaeal genomes, RefSeq plasmid sequences, RefSeq complete viral genomes and GRCh38 human genome (database built in February 2019). Bracken database was built using the read length equals to 100 bp for training datasets and 75 bp for testing dataset and *k*-mer length equals to 35 bp. Bracken correction was performed with the minimum of 10 reads required for a classification at the specified rank.

For each dataset, sample-associated relative abundance profiles have been organized as a taxonomy matrix. A schematic representation of such matrix is shown on the figure 2. Each taxonomy matrix is a tab-separated file. The column names represent the sample identification number. The first column contains the TaxID associated with relative abundances reported for each samples at Species, Genus, Family, Order, Class, Phylum and Superkingdom levels. The relative abundances (ranging from 0 to 100%) calculated for a sample corresponds to the percentage of reads assigned to a specific taxon relative to the total number of reads classified for all taxons at a specific level of taxonomy.

3.2 TaxID description file

In addition to the taxonomy abundances matrices, Challenge participants are provided with a “TaxID description” file that contains the taxonomy rank and full name associated with each TaxID as shown in the figure 3.

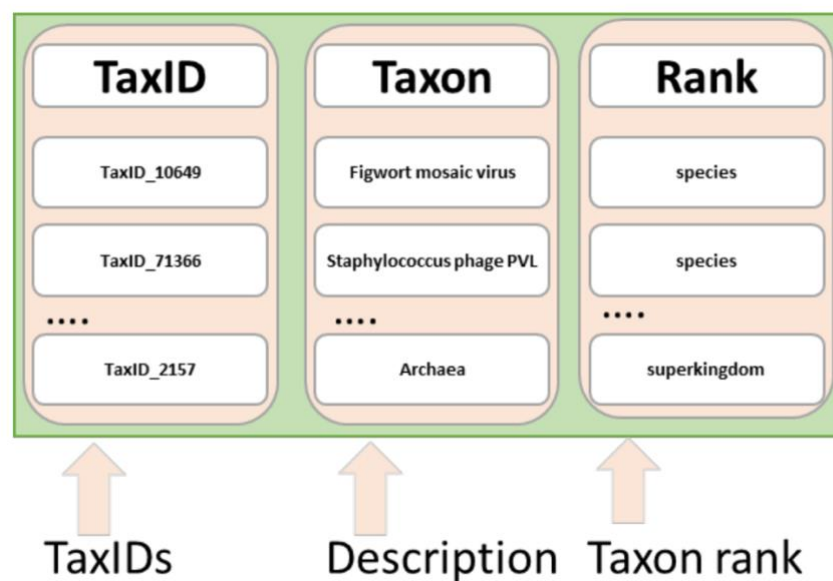


Figure 3: Schematic representation of the TaxID description file.

3.3 Pathway abundances matrices

Pathway abundances matrices for all three datasets (training datasets IBDMDDB/Schirmer et al. and He et al. and PMI dataset) were generated using the Biobakery’s “wmgx” pipeline (8) starting from the raw reads, using default settings and reference databases, except the 16S database that was generated using a text search for “16S” in the NCBI nucleotide database, selecting all sequences which belong to “Fungi”, “Protists”, “Bacteria”, “Archaea” and “Viruses”, with a range of length between 700 and 2000 base pairs and storing those sequences into a fasta file. More specifically, the HUMAnN2 component of the Biobakery (9) pipeline computed pathway abundances for each sample by associating reads with MetaCyc reaction pathways, stratified where possible by species. Pathway abundance files generated for each sample using the Biobakery pipeline were joined into a single matrix with the sample identification numbers as column names and the unique pathways identification number as row names

as illustrated in the figure 4. When pathway abundance was missing for a sample, the pathway abundance value was set to 0.

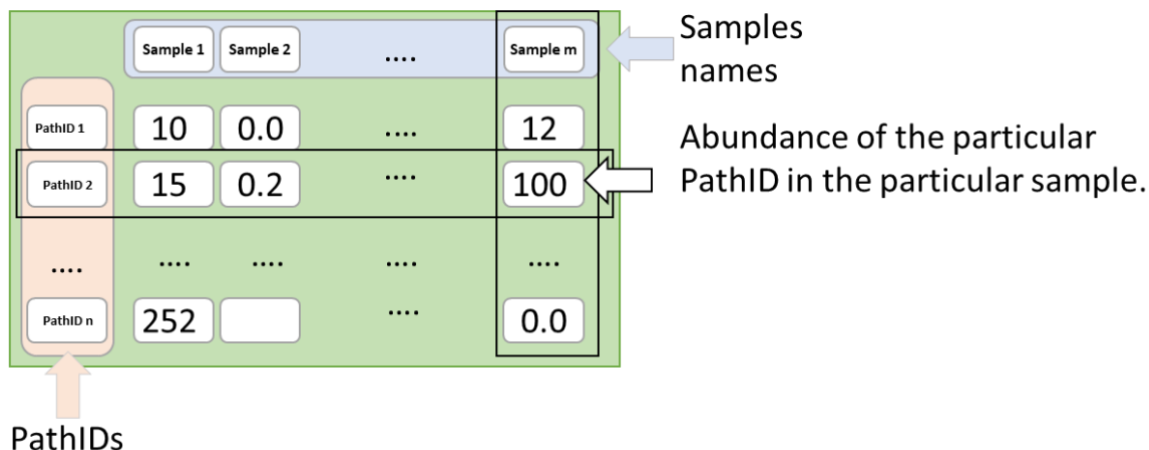


Figure 4: Schematic representation of the pathway abundances matrix structure.

3.4 PathID description file

In addition to the pathway abundances matrices, Challenge participants are provided with a “PathID description” file that contains the full pathway information associated with each PathID as shown in the figure 5.

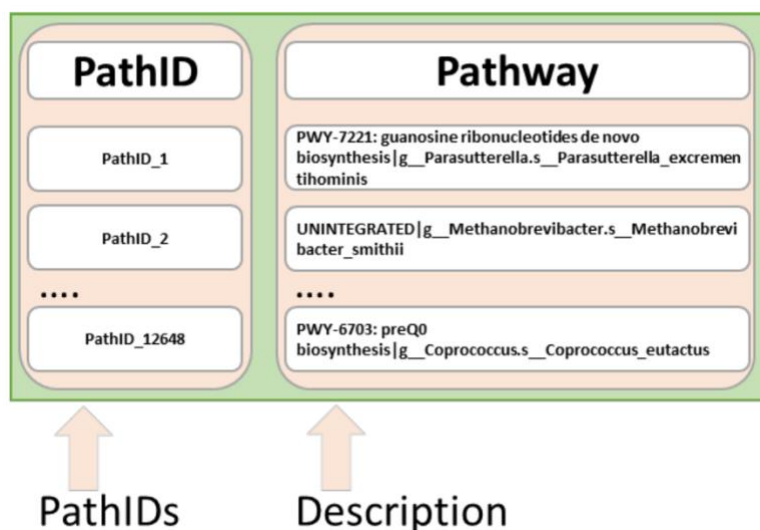


Figure 5: Schematic representation of the PathID description file.

4 Class labels

Class labels associated with each QC selected sample from subjects diagnosed with UC or CD and without IBD (non-IBD) are provided for both training datasets by the Challenge organizers (Table 3 below). Participants can additionally use the publicly available extended metadata accessible at https://ibdmdb.org/tunnel/products/HMP2/Metadata/hmp2_metadata.csv (IBDMDB dataset) and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624284/bin/gix050_Table_and_Supplementary_Table_CD.xlsx (He et al. dataset) for both training datasets. Depending on the training dataset, the extended metadata include information such as subject age, body mass index and gender.

5 Data files for download

Data proposed for training predictive classification models and provided for testing to predict class labels are summarized in the figure 6 below.

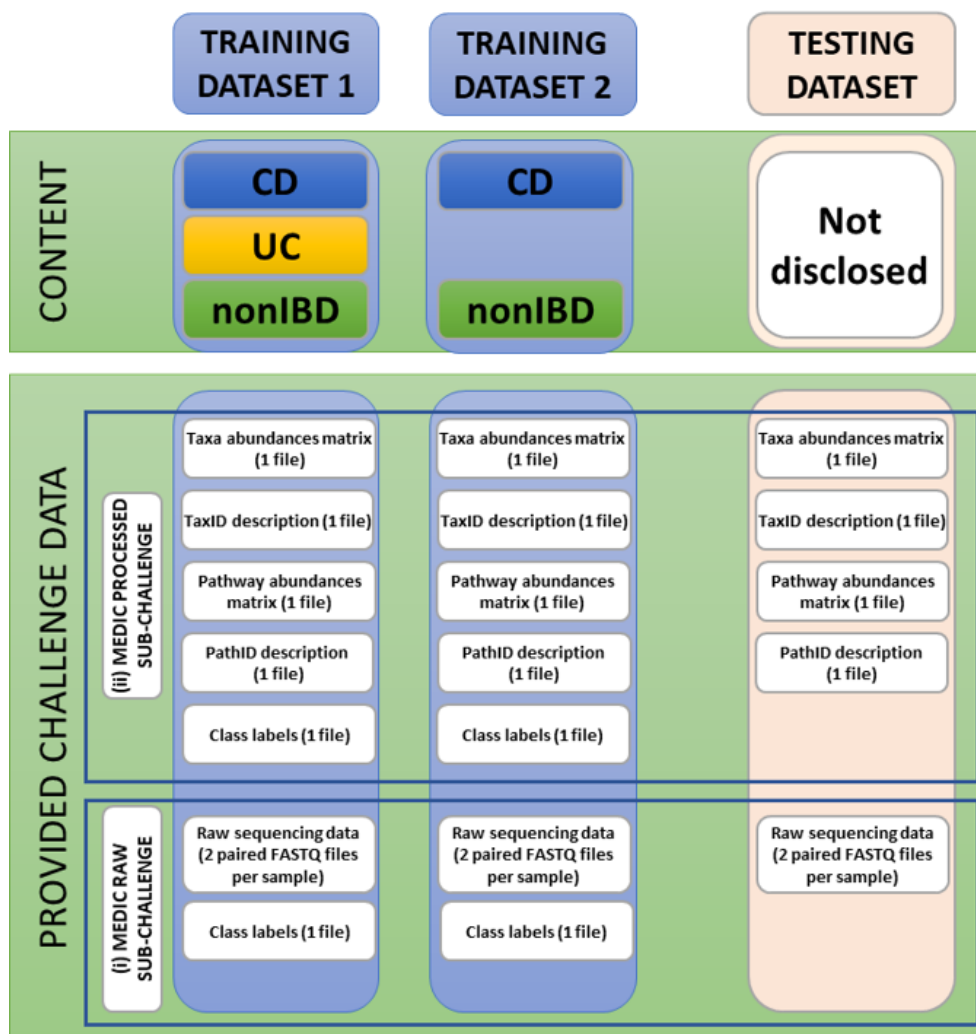
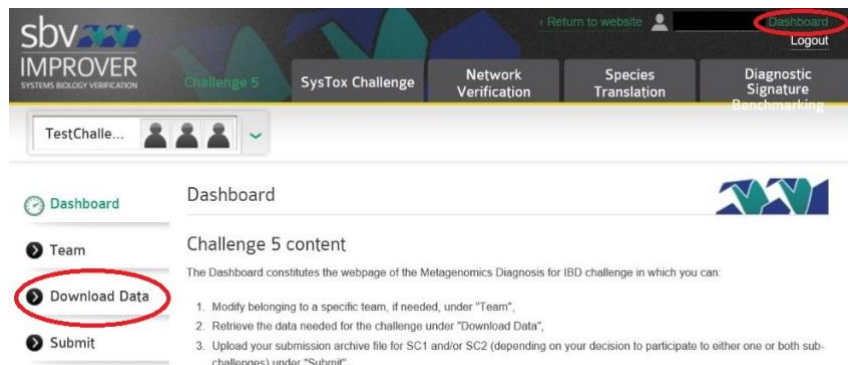


Figure 6. Schematic view of the shotgun metagenomic sequencing datasets provided for training and test including (i) raw (FASTQ files) for the sub-challenge 1 and (ii) processed data for the sub-challenge 2 in the form of taxonomy and pathway abundances matrices.

To download the challenge data, you will need to register to access the Dashboard located at the top of the sbv IMPROVER website. You will first need to create or join an existing Team. Then, you will have the access to download the Challenge data, and later submit your predictions as shown in the snapshot below.



Instructions for data download is provided for Sub-challenge 1 and Sub-challenge 2 separately.

Dashboard
Team
Download Data
Submit
FAQs / Help

Download Data

Sub-challenge 1: MEDIC - RAW

Please download the raw sequencing data for the training set 1, directly from the public repository, as described in the [Technical Document](#) (Annex 1). Link to data: <https://ibdmdb.org/tunnel/public-HMP2/WGS/1819/rawfiles>

Fastq files for each sample can be downloading from The Inflammatory Bowel Disease Multomics Database (<https://ibdmdb.org/>) using the following set of commands (for Linux OS):

```
wget https://ibdmdb.org/tunnel/static/HMP2/WGS/1819/<SAMPLE_NAME>.tar tar -xzf <SAMPLE_NAME>.tar
```

Where the <SAMPLE_NAME> is one of the samples names provided in the Class labels file available for download below.

Training set 1	One text file containing the class labels.	Download
Schirmer et al.	Class_labels_Schirmer.txt	

Please download the raw sequencing data for the training set 2, directly from the public repository, as described in the [Technical Document](#) (Annex 2). Link to data: <http://gigadb.org/dataset/100317>

Fastq files for each sample can be downloaded from the Sequencing Read Archive (SRA, <https://trace.ncbi.nlm.nih.gov/Traces/sra/>). Prior to the data downloading the SRA Toolkit utility should be installed. See <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> for the installation instructions. After the SRA Toolkit installation each dataset can be downloaded using the fastq-dump tool of SRA Tools utility:

```
fastq-dump --split-files --gzip <SAMPLE_NAME>
```

Where the <SAMPLE_NAME> is one of the samples names provided in the Class labels file available for download below.

Training set 2	One text file containing the class labels.	Download
He et al.	Class_labels_He.txt	

Test set

PMI Data	Raw sequencing data: two-paired FASTQ files per sample in a tar file. The tar archive (264 Gb) contains 210 fastq.gz files - 2 for each of the test samples.	Download
	test_dataset.tar	

Sub-challenge 2: MEDIC - PROCESSED

Test set	One zip file (testset_subchallenge2_files.zip) of 9 tab-delimited txt files as follows:	Download
PMI Data	TrainingSchirmer_TaxonomyAbundance_matrix.txt TrainingSchirmer_PathwayAbundance_matrix.txt Class_labels_Schirmer.txt TrainingHe_TaxonomyAbundance_matrix.txt TrainingHe_PathwayAbundance_matrix.txt Class_labels_He.txt TestingDataset_TaxonomyAbundance_matrix.txt TestingDataset_PathwayAbundance_matrix.txt TaxID_Description.txt PathID_Description.txt	

The sbv IMPROVER Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC)

Technical document

Version 2.0 Final, 9.09.2019

Table 3 Summary of training and test data files provided for the Challenge.

Sub-Challenge (SC)	Data type	File description	IBDMDB/Schirmer et al.	He et al.	PMI
	Provided for	NA	Training	Training	Testing
	Groups	NA	23 CD 17 UC 14 non-IBD	63 CD 0 UC 53 non-IBD	CD UC non-IBD
Sub-challenge 1	All data files described below (except <u>*Raw Data</u> proposed to be used for training) are available for download in the Dashboard (3 buttons)				

SC1	*Raw data	Paired end data for each sample, fastq format (gzip-compressed)	Download instructions in the ANNEX 1 below Accessible from their original location: https://ibdmdb.org/tunnel/public/HMP2/WGS/1818/rawfiles	Download instructions in the ANNEX 2 below Accessible from their original location: http://gigadb.org/dataset/100317	Archive (tar)
SC1	Class labels	1 tab-delimited txt file that includes the list of QC-selected samples associated with their respective class labels	For training	For training	Gold standard (not provided)
Sub-challenge 2	All data files described below are included into a zip archive file available for download in the Dashboard (1 button)				
SC2	Processed data	1 Taxonomy abundances matrix tab-delimited txt file 1 Pathway abundances matrix tab-delimited txt file	yes	yes	yes
SC2	TaxID and PathID description	1 TaxID description tab-delimited txt file 1 PathID description tab-delimited	yes	yes	yes

		txt file			
SC2	Class labels	1 tab-delimited txt file that includes the list of QC-selected samples associated with their respective class labels	Same file as the one provided for SC1	Same file as the one provided for SC1	Gold standard (not provided)
	Reference		(2,3)	(1)	No publication

The sbv IMPROVER Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC)

Technical document

Version 2.0 Final, 9.09.2019

6 Participants' tasks and submission on the sbv IMPROVER Challenge website

The participants have the **freedom to participate in either one or both sub-challenges**.

To be eligible for scoring, participants must comply with the submission requirements described in this section 6 - *Participants' tasks and submission on the sbv IMPROVER challenge website* below

- ✓ Submission completeness
- ✓ Compliance with requested data formats
- ✓ Compliance with the related sections of the Challenge Rules (Link below)

6.1 Tasks

- 1) **CLASS PREDICTIONS** - You will provide a confidence value in the interval [0,1] and with 1 being the highest confidence, that the sample belongs to "class 1":
 - IBD (class 1) vs non-IBD (class 2)
 - UC (class 1) vs non-IBD (class 2)
 - CD (class 1) vs non-IBD (class 2)
 - UC (class 1) vs CD (class 2)

To train your predictor models, you can use raw (sub-challenge 1) or processed (sub-challenge 2) shotgun metagenomics data from two independent published studies (Table 1). You can also use any additional private/public datasets you find suitable.

- 2) **FEATURES/SIGNATURE** - You will provide the list of selected features (a subset of TaxIDs or PathIDs) used in their classification prediction model(s) applied on the test dataset, and their associated value of importance (optional), if available (e.g. variable of importance for features obtained using random forest based approaches).
- 3) **METHOD DESCRIPTION** - You will describe your approaches by providing sufficient information to allow reproducibility. Your write-up should include:
 - A description of the feature extraction method(s)
 - A description of the classification/machine learning approach(es) used.

The description should include publication references if the method is published, software and versions, software parameters used, parameters/coefficients of the model applied for prediction, description of metagenomics pipeline for processing shotgun sequencing raw data.

A **write-up [template](#)** provides guidance regarding minimal information needed.

6.2 How to submit your results

1. Submission completeness including all prediction files, list of selected features files and description of the computational approach in a write-up contained in a zip file for the submission on the sbv IMPROVER website as described in Table 4 below. The zip file name will include the prefix “SC1” or “SC2” to allow submission discrimination to either one or the other sub-challenge such as, for example, SC1_xxxx.zip or SC2_xxxx.zip (xxxx can be any type of text without special character). Names and content of any file part of the submission **must NOT contain any contact detail and participant names** to ensure that the submission remains anonymous for the scoring.

Table 4. Description of the total of files contained in one zip file per submission for the respective sub-challenge

A zip archive file including **all submission template files (with expected file formats and names)** for sub-challenge 1 and 2 can be downloaded from [here](#)





Sub-challenge	1 - “MEDIC RAW”	2 - “MEDIC PROCESSED”	
A complete Submission – zip	1 zip file that includes 9 files in total (described below in rows 1,2 and 3)	1 zip file that includes 17 files in total (described below in rows 1,2 and 3)	
1) Classification prediction files - Tab-delimited	4 files	- 4 files for the four 2-class comparisons for the use of the taxonomy matrix	- 4 files for the four 2-class comparisons for the use of the pathway matrix
2) Selected features (=signature) files - Tab-delimited	4 files	- 4 files for the four 2-class comparisons for the use of the taxonomy matrix	- 4 files for the four 2-class comparisons for the use of the pathway matrix
3) Write-up – PDF (template provided as .docx)	1 file	1 file	

Important note: in case you wish to propose different solutions to the challenge (e.g. use different methods, parameters) you are allowed to provide several submissions per team, however:


- Each complete submission corresponds to one zip file containing 9 and 17 files for sub-challenge 1 and 2, respectively.
- Although all your submissions will be scored and ranked, only your submission showing best performance will be retained for eligibility as best performer
- Submissions to the “MEDIC PROCESSED” sub-challenge 2 for which raw metagenomics data has been used will be reassigned to the “MEDIC RAW” sub-challenge 1.

2. Compliance with requested data formats (Table 4 above).
3. Compliance with the related sections of the [Challenge Rules](#)

When ready, you will provide your submission on the Dashboard of the sbv IMPROVER website after logged in and having *created* or *joined* a Team.

 Dashboard
  Team
  Download Data
  Submit

Submit your prediction (See section 6.2 of the Technical Document)



Submit your answers to either sub-challenge 1 **and/or** sub-challenge 2 as a zip file (with the filename SC1_XXXX.zip or SC2_XXXX.zip) using the same upload button below. For details on the data preparation see the [Technical document](#), and for the proper file-format please download the [Template files](#) for participants.

Although all your submissions will be scored and ranked, only your submission showing best performance will be retained for eligibility as best performer.

Each complete submission corresponds to one zip file containing 9 and 17 files for sub-challenge 1 and 2, respectively.

Once your data is uploaded, it will be automatically validated. If you receive an error message, please make the required changes and then upload the zip archive again.

Sub-Challenge 1

Submission SC1	One file (zip archive), with the proper content, corresponding to sub-challenge 1 (8 files + 1 write-up) is required.	No file uploaded	Upload file
-----------------------	---	------------------	-----------------------------

Sub-Challenge 2

Submission SC2	One file (zip archive), with the proper content, corresponding to sub-challenge 2 (16 files + one write-up), is required.	No file uploaded	Upload file
-----------------------	---	------------------	-----------------------------

7 Scoring

Participants' eligibility for scoring of their predictions is conditional to their compliance with the submission requirements and Challenge rules detailed above in the section 6 - *Participants' tasks and submission on the sbv IMPROVER Challenge website* of this document, as well as the related sections on the [Challenge Rules](#).

The scoring procedure will be conducted as described in the [Scoring](#) section of the website.

8 Abbreviations

IBD – Inflammatory bowel disease

CD – Crohn's disease

UC - ulcerative colitis

non-IBD – samples from subjects without CD or UC diagnosed

WGS – whole genome sequencing

DNA - Deoxyribonucleic acid

9 Annex 1. IBDMDB data download

Fastq files for each sample can be downloading from The Inflammatory Bowel Disease Multi'omics Database (<https://ibdmdb.org/>) using the following set of commands (for Linux OS):

```
wget https://ibdmdb.org/tunnel/static/HMP2/WGS/1818/<SAMPLE\_NAME>.tar  
tar -xf <SAMPLE_NAME>.tar
```

Where the <SAMPLE_NAME> is one of the samples names provided in the Class labels file

10 Annex 2. He et al. data download

Fastq files for each sample can be downloaded from the Sequencing Read Archive (SRA, <https://trace.ncbi.nlm.nih.gov/Traces/sra/>). Prior to the data downloading the SRA Toolkit utility should be installed. See <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> for the installation instructions.

After the SRA Toolkit installation each dataset can be downloaded using the fastq-dump tool of SRA Tools utility:

```
fastq-dump --split-files --gzip <SAMPLE_NAME>
```

Where the <SAMPLE_NAME> is one of the samples names provided in the Class labels file

11 References

1. He, Q., Gao, Y., Jie, Z., Yu, X., Laursen, J. M., Xiao, L., Li, Y., Li, L., Zhang, F., Feng, Q., Li, X., Yu, J., Liu, C., Lan, P., Yan, T., Liu, X., Xu, X., Yang, H., Wang, J., Madsen, L., Brix, S., Wang, J., Kristiansen, K., and Jia, H. (2017) Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1-11
2. Integrative, H. M. P. R. N. C. (2014) The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276-289
3. Schirmer, M., Franzosa, E. A., Lloyd-Price, J., McIver, L. J., Schwager, R., Poon, T. W., Ananthakrishnan, A. N., Andrews, E., Barron, G., Lake, K., Prasad, M., Sauk, J., Stevens, B., Wilson, R. G., Braun, J., Denson, L. A., Kugathasan, S., McGovern, D. P. B., Vlamakis, H., Xavier, R. J., and Huttenhower, C. (2018) Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* **3**, 337-346
4. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079
5. Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048
6. Wood, D. E., and Salzberg, S. L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46
7. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745
8. McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., Segata, N., and Huttenhower, C. (2018) bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235-1237
9. Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., and Huttenhower, C. (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**, 962-968