MICROBIOME

WG3 Progress update

23th May 2022

WG3 Objectives and major deliverables

Objectives:

To optimise and standardize the use of state-of-the-art ML techniques, resulting in **best practice SOPs** specific to various microbiome data types, human body ecosystems and research questions. The WG3 will also investigate opportunities for **automating** the established SOPs into pipelines for translational use by clinicians and non-experts.

Major Deliverables:

D3.1: A decision tree of ML/Stats methods along with optimised parameters suitable for various data types, ecosystems and research questions (disseminated through Web-portal and GitHub).

D3.2: A publication and white-paper describing the SOPs emanating from D3.1.

D3.3: A report outlining areas suitable for automation

WG3 Objectives and major deliverables

Objectives:

To optimise and standardize the use of state-of-the-art ML techniques, resulting in **best practice SOPs** specific to various microbiome data types, human body ecosystems and research questions. The WG3 will also investigate opportunities for **automating** the established SOPs into pipelines for translational use by clinicians and non-experts.

Input:

- State-of-the-art ML/Stats methods (→ WG1)
- Benchmark data (→ WG2)

Benchmark Data: colorectal cancer (CRC) use-case

Shotgun data

- **1600 samples**, from 10 publicly available studies
- **8 countries**, over 3 continents (Europe, America, Asia)
- All sequences have been downloaded and processed the same way
  - Mapping of the reads onto the 10.4 million genes IGC2 reference catalog
  - Generation of the gene abundance profiling table (rarefaction and FPKM normalization)
  - Generation of the Metagenomic Species (MGS) abundance table from 100 marker genes
- **Metadata available**: health status and phenotype (healthy, patient, adenoma, CRC stage), country, BMI, gender, age, gene and MGS richness
- Gathered and processed by Emmanuelle Le Chatelier *et al.* (nov. 2021)

Benchmark Data: colorectal cancer (CRC) use-case

Shotgun data

BioProject	country	N all	PMID	DOI
PRJEB7774	AUT	156	25758642	DOI: 10.1038/ncomms7528
PRJEB10878	CHN	128	26408641	DOI: 10.1136/gutjnl-2015-309800
PRJEB6070	FRA	156	25432777	DOI: 10.15252/msb.20145645
PRJEB6070	GER	43		
PRJEB27928	GER	82	30936547	DOI: 10.1038/s41591-019-0406-6
PRJNA397112	IND	110	30698687	DOI: 10.1093/gigascience/giz004
PRJNA531273	IND	30	31719139	DOI: 10.1128/mSystems.00438-19
PRJNA447983	ITA	140	30936548	DOI: 10.1038/s41591-019-0405-7
PRJDB4176	JPN	645	31171880	DOI: 10.1038/s41591-019-0458-7
PRJEB12449	USA	110	27171425	DOI: 10.1371/journal.pone.0155362
	TOTAL	**1600**		

Download link: https://filesender.renater.fr/?s=download&token=521ee599-29a7-4abd-bad2-a5a4e56d4ad0

Benchmark Data: colorectal cancer (CRC) use-case

16S data

- **709 samples**, from 3 publicly available studies
- **3 countries**, over 2 continents (Europe, America)
- All sequences have been downloaded and processed the same way
  - All datasets were processed using qiime2 pipeline with DADA2 for Sequence quality control and feature table construction, and SILVA database for taxonomic assignment, then a phyloseq object was constructed.
- Info & download link: https://hackmd.io/@laurichi13/rJt3ewZut
- Gathered and processed by Laura Marcos *et al.* (dec. 2021)

WG3 Objectives and major deliverables

Objectives:

To optimise and standardize the use of state-of-the-art ML techniques, resulting in **best practice SOPs** specific to various microbiome data types, human body ecosystems and research questions. The WG3 will also investigate opportunities for **automating** the established SOPs into pipelines for translational use by clinicians and non-experts.

Addressed questions:

- Microbiome data types: 16S rRNA amplicons & shotgun metagenomics
- Ecosystems: gut
- Research question: diagnostics

WG3 progress update

- Monthly meetings from September 2021
- Exchange about the different approaches to use ML on the benchmark data
  - **Shotgun data** (Alberto Tonda *et al.*, Marta Lopes *et al.*, Julia Eckenberg, Magali Berland *et al.*, …)
  - **16S data** (Laura Marcos *et al.*, Christian Jansen *et al.*, …)
  - **Compositional data analysis** (Karel Hron & Matthias Templ)
  - **Literature review** about the ML techniques used by the community - in collaboration with WG1 (Laura Marcos, Eliana Ibrahimi & Rajesh Shigdel)
- Working on the deliverables:
  - **Decision tree from the literature**, to help to identify the most commonly used ML techniques
  - **Decision tree from optimization work** conducted by the people participating to WG3

Plan for WG3 meeting

- WG3 group members update since the last meeting
- Work on the deliverables
  - SOPs & Decision tree
- WG3 organization
  - I’ll be in maternity leave in a few weeks and WG3 will need a new leader in the meantime. Sonia Tarazona is ok to help but will have limited time on the period.
  - Anyone interested to step up?

Useful links

- **Slack channel:** https://ca18131.slack.com/
- **Shotgun dataset:** https://filesender.renater.fr/?s=download&token=521ee599-29a7-4abd-bad2-a5a4e56d4ad0
- **16S dataset:** https://hackmd.io/@laurichi13/rJt3ewZut
- **Bioinformatic processing for shotgun data:** https://ca18131.slack.com/files/UUNS11R38/F02NBMW5KSM/2021-11-17-bioinformatic-processing.pdf
- **Bioinformatic processing for 16S data:** https://ca18131.slack.com/files/U015ZFHBXEW/F02RDAMGKTJ/16sdataset_processing.pdf.pdf
- **This presentation:** https://docs.google.com/presentation/d/1a1WpW42larjBS7vNuKk7Cb5lHC56JrEI4wKycLndx3I
- **WG3 SOPs:** https://docs.google.com/document/d/1RVQY7UI3YyuAX__TNgIQ04AHXlVG5XkN5K7PvbVdm1Y