DOMINO

Funded by the European Union

DELIVERABLE

D3.2: Design and Final version of the FMD

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060218

Deliverable 3.2 - Design and Final Version of the FMD

Deliverable Nr.	D3.2
Due date	August 2024
Submission date	August 2024
Deliverable type	R
Dissemination level	Public
Work package	WP 3
Author(s)	Nicola Segata, Paul Cotter, Hrituraj Dey, Federica Pinto, Simone Anzà, Vitor Heidrich, Liam Walsh, Katie Falà, Viyal Soni, Joseph Ancla, Orla O'Sullivan

Document version	1.0		
Grant agreement	101060218	**Duration**	60 months
Start Date	March 2023	**End date**	February 2028

DOMINO

D3.2

Contributors

NAME	ORGANISATION
Nicola Segata	UNITN
Paul Cotter	TEAGASC
Hrituraj Dey	UNITN
Vitor Heidrich	UNITN
Liam Walsh	TEAGASC
Katie Falà	TEAGASC
Federica Pinto	UNITN
Orla O'Sullivan	TEAGASC
Simone Anzà	UNITN
Claudia Mengoni	UNITN

Revision history

VERSION	DATE	REVIEWER	MODIFICATIONS
X	DD/MM/YY	Name	Modifications
Y	DD/MM/YY	Name	Modifications

D3.2

Table of contents

1. Executive summary

For the work related with this deliverable, we performed a major development step for the curatedFoodMetagenomicDatabase (cFMD), comprising a total of 2,533 metagenomes spanning 59 food metagenomic datasets (Carlino et al. 2024). Specifically, cFMD has been improved with the inclusion of functional profiles of the metagenomes and eukaryotic metagenome-assembled genomes. As per this milestone, this improved version of the cFMD (v1.2.0), also referred to as Food Microbiome Database (FMD), is now available to DOMINO Partners (https://github.com/SegataLab/cFMD). In addition, during the course of the DOMINO project, 22 additional new datasets publicly available through NCBI have been identified in the literature. In combination, the datasets have 404 samples across a range of foods. Other 6 food metagenomic datasets including 188 samples have been generated within DOMINO. All these datasets are being processed and curated and will expand FMD by almost 20% by the end of 2024.

2. Background

Fermented products, including items like kefir, yogurt, raw cheese, kombucha, ginger beer, and wine, represent culturally relevant foods and can hold a significant place in the average person's diet. This

widespread consumption has spurred both industry and academia to take a keen interest in comprehending the underlying mechanisms at play. Understanding the relationships within these microbial communities allow for the development of safer products while minimizing setbacks such as fermentation failure. In this regard, members of DOMINO previously generated the curatedFoodMetagenomicData (cFMD) database, consisting of both publicly available and recently generated food (shotgun) metagenomes that covers curated metadata, taxonomic profiles, as well as reconstructed genomes. The initial cFMD iteration comprises a total of 2,533 metagenomes linked with 59 datasets with wide-ranging food types (cheese, meat, fermented vegetables, wine, kefir, etc.). Maintenance, dissemination and data expansion of the cFMD is being carried out over the course of DOMINO, generating a more comprehensive database called **Food Microbiome Database (FMD)**. Current efforts include the identification, and generation of newly published sequencing datasets which include representative metagenomes of multiple fermented food substrates (dairy, cereals, meats and vegetables) to ensure the continued expansion of the FMD. Deliverable D3.2 includes the final version of the FMD, which comprises the implementation of an improved ontology, the inclusion of functional profiles and the addition of eukaryotic metagenome-assembled genomes that were overlooked in the previous version of the database. Despite D3.2 being defined as the final version, FMD will be constantly updated and implemented along with DOMINO, with new datasets and data newly generated within the project.

3. Methods

3.1. Retrieval of eukaryotic metagenome-assembled genomes from food

Eukaryotic metagenome-assembled genomes (MAGs) were retrieved by assessing the 17,009 MAGs that did not pass the minimum threshold to be considered at least a medium-quality prokaryotic MAG according to checkM v1.1.3 (Parks et al. 2015). Taxonomy for the assembled MAGs were assigned for the genomes that had an average nucleotide identity of > 95% based on the 17,438 publicly available eukaryotic genomes present in NCBI.

The quality control for eukaryotic MAGs fraction assessed using BUSCO v.5.6.1 (Manni et al. 2021) resulted in 787 MAGs (392 high-quality (HQ) and 395 medium-quality (MQ) MAGs, keeping a threshold of completeness > 90% and contamination < 5% for HQ and completeness ≥ 50% and contamination < 5% for MQ). The 787 MQ/HQ eukaryotic MAGs were clustered into species-level genome bins (SGBs) at 95% average nucleotide identity and assigned a taxonomy based on 17,438 publicly available fungal genomes present in NCBI.

3.2. Functional profiling of food metagenomes

Functional profiles were retrieved using HUMAnN v3 10 (Beghini et al. 2021), using pangenomes annotated with UniRef90 on all species detected with MetaPhlAn v3 10 (—metaphlan-options "-t rel_ab —index v30_CHOCOPhlAn_201901").

3.3. Metagenomic assembly of DOMINO datasets

MAGs were generated by performing de novo metagenomic assembly on each metagenome independently using a validated pipeline for prokaryotes (Pasolli et al. 2019). Initially, metagenomic assembly was carried out using MEGAHIT v1.1.1 (Li et al. 2015). Contigs shorter than 1000 bp were removed. The remaining contigs were then aligned against the original raw data using Bowtie2 v2.2.9 to determine coverage information (Langmead and Salzberg 2012). Subsequently, contigs were binned using MetaBAT v2.12.1 (Kang et al. 2019). Quality control of the resulting putative genomes was performed with CheckM v1.1.3, keeping only HQ (completeness >90% and contamination <5%) and MQ (completeness ≥50% and contamination <5%) MAGs, according to established standards (Bowers et al. 2017). LQ MAGs from this step underwent eukaryotic quality control as previously described (see section 3.1).

3.4. Strategy for the integration of new food metagenomic datasets in FMD

The expansion of FMD with new food metagenomic datasets identified from the literature will follow the same strategy followed in its previous iteration (**Figure 1**). In detail, following download of metagenomic data and curation of food-associated metadata, datasets will be uniformly processed through a dedicated pipeline exploiting a set of validated and state-of-the-art tools (mainly developed and maintained by Segata Lab). Following data preprocessing to remove low-quality and contaminating reads (https://github.com/SegataLab/preprocessing), MAGs will be generated through a validated metagenomic assembly pipeline applied to each metagenome separately (Pasolli et al. 2019), which will be followed by MAG quality-checking to discard low-quality genomes. These quality-checked genomes will be exploited to expand the microbial genomes database MetaRefSGB, which will be followed by unsupervised clustering of genomes into species-level genome bins (SGBs) according to their relative genomic distances (those are further split into known and unknown SGBs (kSGBs and uSGBs, respectively), depending on whether the SGB contains also isolate representatives). Next, this updated version of MetaRefSGB will be used as input to the ChocoPhlAn pipeline in order to recognize SGB-specific marker genes to be used for metagenomic taxonomic profiling with MetaPhlAn 4 (Blanco-Míguez et al. 2023). Functional profiling with to-be-released

HUMAnN 4 will also be performed. Finally, functional and taxonomic profiles will be made publicly available along with sample-specific metadata in the next expansion round of FMD.

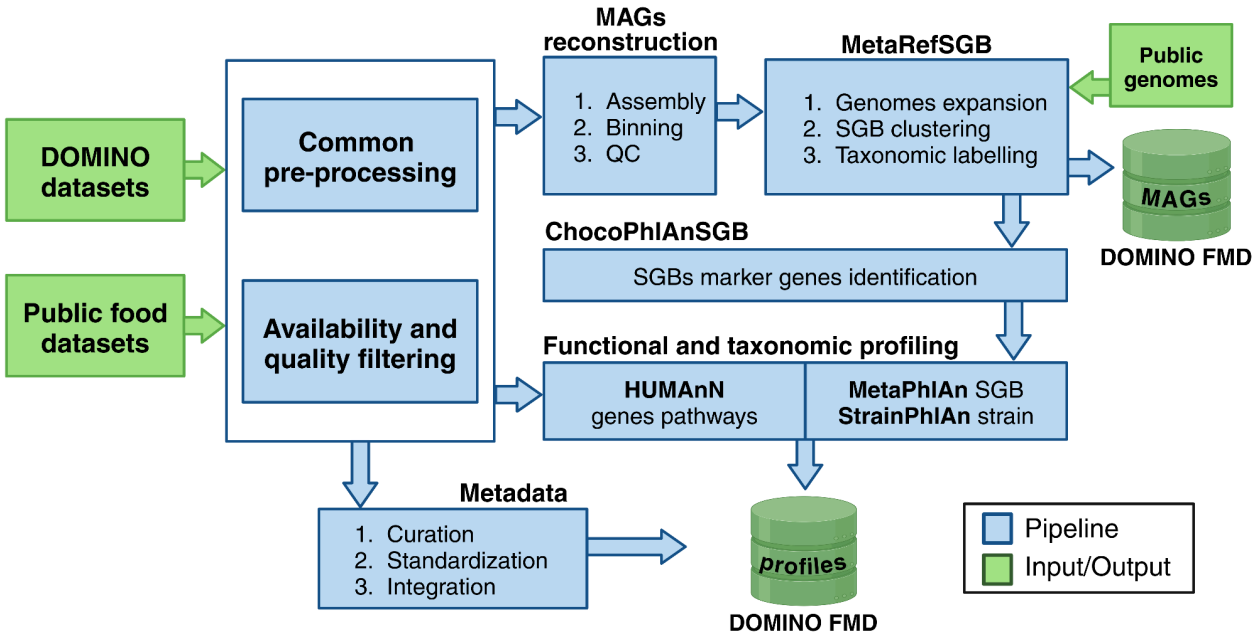

Figure 1. Food Microbiome Database (FMD) workflow.

4. Results & Discussion

4.1. Improvement of metadata ontology

The existing ontology framework, which integrates elements from both the MASTER and cFMD ontologies, will undergo further refinement to address gaps identified in previous iterations. This refinement will be particularly tailored to plant-based foods and their associated processes. The enhancement process will include the standardization of metadata categories across various types of fermented foods, ensuring alignment with widely recognized ontologies such as FoodON (https://foodon.org), OntoBiotope (http://bibliome.jouy.inra.fr/demo/alvisdb/obt/browse2), and FOBI (Castellano-Escuder et al. 2020) Feedback from Case Study Leaders in WP4 will also be incorporated to ensure relevance and applicability. Furthermore, the ontology will evolve throughout WP4 and will be progressively implemented during the project's course. The final version of this ontology will be included in the FMD for DOMINO, scheduled for release in Deliverable 3.8 in month 60. This iterative approach will ensure the ontology remains dynamic and responsive to the project's needs, thereby facilitating more accurate and comprehensive data integration, analysis, and interpretation.

4.2. Inclusion of eukaryotic metagenome-assembled genomes

During DOMINO, we included 787 eukaryotic MAGs recovered from food metagenomes in FMD (**Figure 2**). After clustering at SGB-level, these MAGs spanned 108 eukaryotic SGBs, 78 of which are known and 30 unknown SGBs (kSGBs and uSGBs, respectively). Among the widespread fungal taxa the most frequently reconstructed SGB was *S. cerevisiae* (n = 191 MAGs), mostly recovered from fermented beverages (n = 149) and alcoholic beverages (n = 29). Among the uSGBs, the most represented were assigned to the families *Aspergillaceae* (EUK10000018), *Mucoraceae* (EUK10000045), and *Aspergillaceae* (EUK10000063), all encompassing 4 MAGs each.

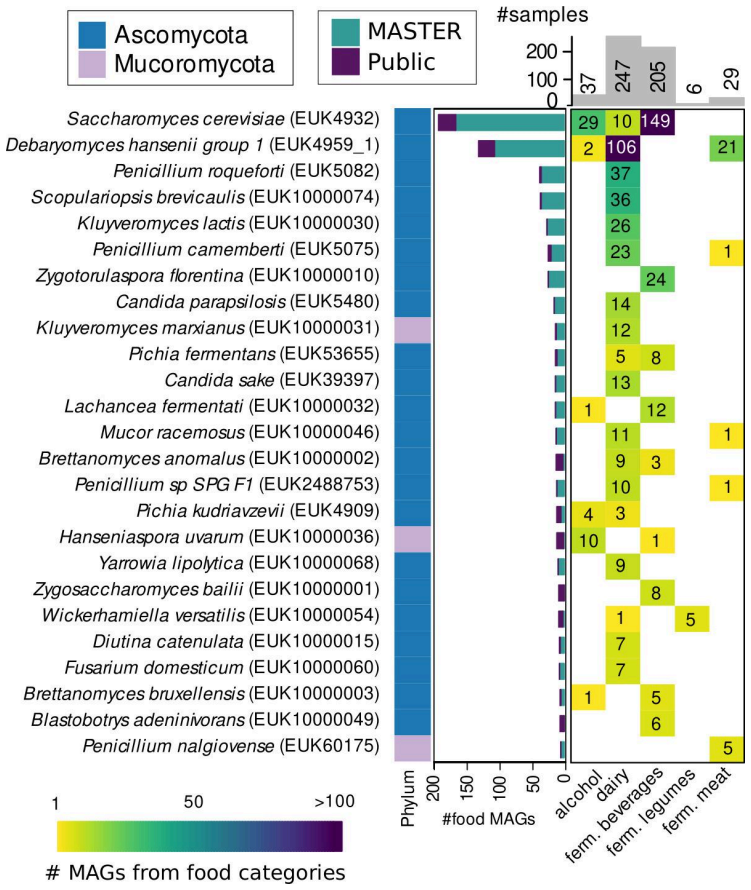

Figure 2: Top 25 eukaryotic SGBs with the highest number of MAGs recovered from food metagenomes. The number of food eukaryotic MAGs is reported in total and for each food category with at least five MAGs, along with the number of samples from which these MAGs were recovered. Highest reported MAGs were found in alcohol and dairy followed by other food categories such as fermented beverages, fermented legumes and fermented meat.

4.3. Inclusion of functional profiles

During DOMINO, we generated UniRef90 gene family abundances as well as metabolic pathway abundances and coverages of the 2,512 quality-controlled food metagenomes currently present in the database. These functional profiles are now publicly available for all food samples within FMD.

4.4. Metagenomic assembly of DOMINO datasets to be integrated

The metagenomic assembly of new DOMINO datasets involves a total of 188 samples from 6 different partners (INRAE n = 24; UNITO n = 24; TFTAK n = 37; IRD n = 40; CSIC n = 37; TEAGASC n = 26) and belonging to different food types: fermented legumes (INRAE), fermented table olives (UNITO), fermented veggies (TFTAK), fermented cereals (IRD), apple pomace (CSIC), and water kefir (TEAGASC). After quality-checking, we have generated 325 HQ and 557 MQ prokaryotic MAGs from these newly sequenced DOMINO food metagenomes. We also retrieved 117 HQ and 134 MQ eukaryotic MAGs from those datasets, totalling 1,133 quality checked reconstructed genomes to be integrated into FMD.

Furthermore, UniTrento & TEAGASC are coordinating for hybrid (short+long read sequencing technology) assembly of new isolates from foods. In total, 233 isolates from 6 partners (INRAE n = 20, UNITO n = 53, TFTAK n = 41, IRD n = 41, CSIC n = 42) have been processed and sequenced with short-read sequencing technology by UniTrento. This data is being combined with long-read data from the same isolates produced by TEAGASC to produce high-quality, circularized, microbial genomes. These assemblies will be incorporated to the genomic database of FMD and leveraged during the next round of its expansion.

4.5. Curation of new food metagenomic datasets to be integrated

In addition to the metagenomic datasets generated during DOMINO, during the course of the DOMINO project, 22 additional new public datasets to be added to FMD have been identified (**Table 1**). In combination, the datasets have 404 samples across a range of foods. Samples include cocoa beans, cheese, kefir, kombucha, beer, sausages, as well as more niche, region-specific fermented foods such

as pozol (Mexican beverage), peron namsing, keem and jalebi (a fermented soybean product, a fermented beverage, and a wheat-based confectionery from India, respectively). To meticulously assess each public study, we documented the methodology of every publication, specifically focusing on the employed sequencing platform and diverse DNA extraction methods, encompassing commercial kits. Moreover, each dataset holds a study accession and NCBI bioproject number with the data publicly available through the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home). These additional metagenomic datasets will undergo metagenomic assembly to generate MAGs, which will be used to improve genomic databases that are the basis for taxonomic and functional profiling tools. Using these enriched databases, food metagenomes will undergo taxonomic and functional profiling. Finally, MAGs and microbiome profiles will be made publicly available through FMD.

STUDY_ID	FOOD_CATEGORY	FOOD_TYPE	COUNTRY	NCBI_BIOPROJECT	YEAR	SEQUENCING_METHOD	DNA_EXTRACTION_KIT	#_OF_SAMPLES
FontanaF_2023	dairy	raw_milk_cheese	Italy/ITA	PRJNA865096	2023	Illumina NextSeq 500	DNeasy PowerFood Microbial Kit (Qiagen)	128
Franciosal_2021	fermented_meat	fermented_sausages	Italy/ITA	PRJNA636619	2021	Illumina NextSeq 500	RNeasy Power Microbiome kit (QIAGEN Group)	15
GonzalezOrozcoB_2023	dairy	milk_kefir	USA	PRJNA843464	2023	Illumina (unspecific)	Wizard Genomic DNA Purification Kit (Promega)	1
	fermented_beverages	kefir_grains	USA	PRJNA843464	2023	Illumina (unspecific)	DNeasy PowerSoil Kil (Qiagen)	1
MotaGutierrezJ_2021	fermented_seeds	cocoa_beans	Mexico/MEX	PRJNA627182	2021	Illumina NextSeq 500	MasterPure Complete DNA & RNA purification kit(Illumina, San Diego, CA)	20
SaakC_2023	dairy	cheese_rind	USA	PRJNA778418	2023	Illumina NovaSeq 6000	No specific kit used - phenol-chloroform method used	18
YangC_2021	dairy	cheese	China/CHN; Italy/ITA; Greece/GRC	PRJNA683931	2021	Illumina HiSeq 2000	DNeasy PowerFood Kit (Omega Bio-tek Inc., Norcross, GA)	8
SalgadoTS_2021	dairy	milk_kefir	Mexico/MEX	PRJNA704713	2021	Illumina NextSeq 500	PowerSoil® DNA Isolation Kit (Mobio-Qiagen, Netherlands)	2
LopezSanchezR_2023	fermented_grain	pozol	Mexico/MEX	PRJNA648868	2023	Illumina NextSeq 500	PowerSoil DNA Isolation kit (catalogue no. 12888–50), PowerMax Soil DNA Isolation kit (catalogue no. 12988–10 y) and UltraClean Microbial DNA Isolation kit (catalogue no. 12224–250)	4
YapM_2020	dairy	bovine_raw_milk/ human_milk	Ireland/IRE	PRJEB38099	2020	Illumina NextSeq 500	DNeasy PowerSoil Pro kit (Qiagen, West Sussex, United Kingdom)	39
OlgaP_2019	fermented_beverages	kombucha	Brazil/BRA	PRJNA63680; PRJNA636891;PRJNA	2019	Illumina HiSeq 2500 (For all the projects)	No specific kit used	9

				6370; PRJNA63701; PRJNA636837				
QuijadaN_2022	dairy	hard_cheese	Austria/AUT	PRJEB48589	2022	Illumina HiSeq 2500	innuSPEED Bacteria/Fungi RNA kit	2
ShangpliangH_2023_a	fermented_milk	laal_dahi	India/IND	PRJNA914731	2023	Illumina Novaseq 6000	No specific kit used	2
ShangpliangH_2023_b	fermented_flour	jalebi_batter	India/IND	PRJNA914730	2023	Illumina Novaseq 6000	No specific kit used	2
KharnaiorP_2023	fermented_soybean	peron_namsing	India/IND	PRJNA914873	2023	Illumina Novaseq 6000	No specific kit used	4
YouL_2022	fermented_milk	koumiss	China/CHN	PRJNA687995	2022	Illumina HiSeq Xten	DNeasy® PowerFood® Microbial Kit (QIAGEN, Germany)	23
LimaC_2020	fermented_seeds	cocoa_beans	Brazil/BRA	PRJNA552479	2020	Illumina HiSeq Xten	No specific kit used = Extraction from lyophilized cacao pulp	14
DecadtH_2024	dairy	gouda_cheese	Belgium/BEL	PRJEB64331	2024	Illumina NextSeq 500	No specific kit used	19
YasirM_2022	non-dairy fermented	pickles	Saudi Arabia/SAU	PRJNA816670	2022	Illumina HiSeq 2500	DNeasy PowerSoil Pro Kit (Qiagen, Germany)	18
FalardeauJ_2023	dairy	surface_ripened_soft_cheeses	Canada/CAN	PRJNA863305	2023	Illumina (unspecified)	Neasy PowerFood microbial kit (Qiagen, Inc.,Toronto, ON, Canada	56
AlmeidaO_2020	fermented_seeds	cocoa_beans	Brazil/BRA	PRJNA527768	2022	Illumina NextSeq 500	PowerSoil® DNA Isolation Kit (Mobio-Qiagen, Netherlands)	9
YuY_2022	fermented_seeds	mustard	China/CHN	PRJNA780248	2022	Illumina Novaseq 6000	E.Z.N.A.® Stool DNA Kit (D4015-02, Omega, Inc., United States)	9
TomarS_2023	alcohol	keem	India/IND	PRJNA763767	2023	Illumina NovaSeq 6000	QIAamp® PowerFecal® Pro DNA Kit (Qiagen, Germany)	1

Table 1: Overview of new shotgun metagenomics datasets selected for inclusion in the FMD, including pertinent details such as Food substrate, NCBI Bioproject number and publication year. Publicly available datasets were obtained from a literature search, followed by quality inspection based on full text screening.

5. Conclusion

For this milestone, an improved first version of the FMD (comprising a total of 2,533 metagenomes spanning 59 datasets) is now available to DOMINO Partners. On top of the prokaryotic metagenome-assembled genomes and metagenomic taxonomic profiles, the current version of the database includes newly added eukaryotic metagenome-assembled genomes, functional profiles of the metagenomes, and improved metadata ontology. The FMD will be continuously updated with more food metagenomes (as well as metagenome-assembled genomes and metagenomic profiles) throughout the course of the DOMINO project, both with datasets from the literature and datasets generated within DOMINO.

References

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. "Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4." *Nature Biotechnology*, February. https://doi.org/10.1038/s41587-023-01688-w.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

Carlino, Niccolò, Aitor Blanco-Míguez, Michal Punčochář, Claudia Mengoni, Federica Pinto, Alessia Tatti, Paolo Manghi, et al. 2024. "Unexplored Microbial Diversity from 2,500 Food Metagenomes and Links with the Human Microbiome." *Cell* 0 (0). https://doi.org/10.1016/j.cell.2024.07.039.

Castellano-Escuder, Pol, Raúl González-Domínguez, David S. Wishart, Cristina Andrés-Lacueva, and Alex Sánchez-Pla. 2020. "FOBI: An Ontology to Represent Food Intake Data and Associate It with Metabolomic Data." *Database: The Journal of Biological Databases and Curation* 2020 (January). https://doi.org/10.1093/databa/baaa033.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies." *PeerJ* 7 (July):e7359.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." *Molecular Biology and Evolution* 38 (10): 4647–54.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W.

Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes." *Genome Research* 25 (7): 1043–55.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." *Cell* 176 (3): 649–62.e20.

Coordinator, Project manager and Work Package leader

Prof. Nicola Segata University of Trento (+39) 0461 285218 nicola.segata@unitn.it

Prof. Paul Cotter Teagasc Food Research Centre Tel +353 (0) 2542694 paul.cotter@teagasc.ie