NFDI4Microbiome: extended abstract for the NFDI conference

2019-03-28

https://doi.org/10.5281/zenodo.2614002

1. Formal details

Planned title of the consortium: National Research Data Infrastructure for Microbiome Research

Acronym of the planned consortium: NFDI4Microbiome

Lead institution or facility: ZB MED - Information Centre for Life Sciences

Name and work address of a contact person (including email address and institutional affiliation):

Prof. Dr. Konrad U. Förstner
foerstner@zbmed.de
ZB MED - Information Centre for Life Sciences
Gleueler Straße 60
50931 Cologne
Germany

Members of the planned consortium (including institutional affiliation, without address)

- Prof. Dr. Konrad U. Förstner, ZB MED - Information Centre for Life Sciences
- Prof. Dr. Jörg Overmann, German Collection of Microorganisms and Cell Cultures (DSMZ)
- Prof. Dr. Peer Bork, European Molecular Biology Laboratory (EMBL)
- Prof. Dr. Alfred Pühler, German Network for Bioinformatics Infrastructure (de.NBI)
- Prof. Dr. Jens Stoye, Genome Informatics, Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University
- Prof. Dr. Alice McHardy, Helmholtz Centre for Infection Research (HZI)

Participants in the NFDI conference (names, institutional affiliation and email address; max. 3 persons)

- Prof. Dr. Konrad U. Förstner, ZB MED - Information Centre for Life Sciences
- Prof. Dr. Alfred Pühler - German Network for Bioinformatics Infrastructure (de.NBI)
- Dr. Alexander Sczyrba, Computational Metagenomics, Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University

Research area of the planned consortium (research area according to the DFG classification system (not subject areas):

21 (Biology) and 22 (Medicine)

Participating research institutions (without address)

- ZB MED - Information Centre for Life Sciences
- DSMZ - German Collection of Microorganisms and Cell Cultures
- EMBL - European Molecular Biology Laboratory
- Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University

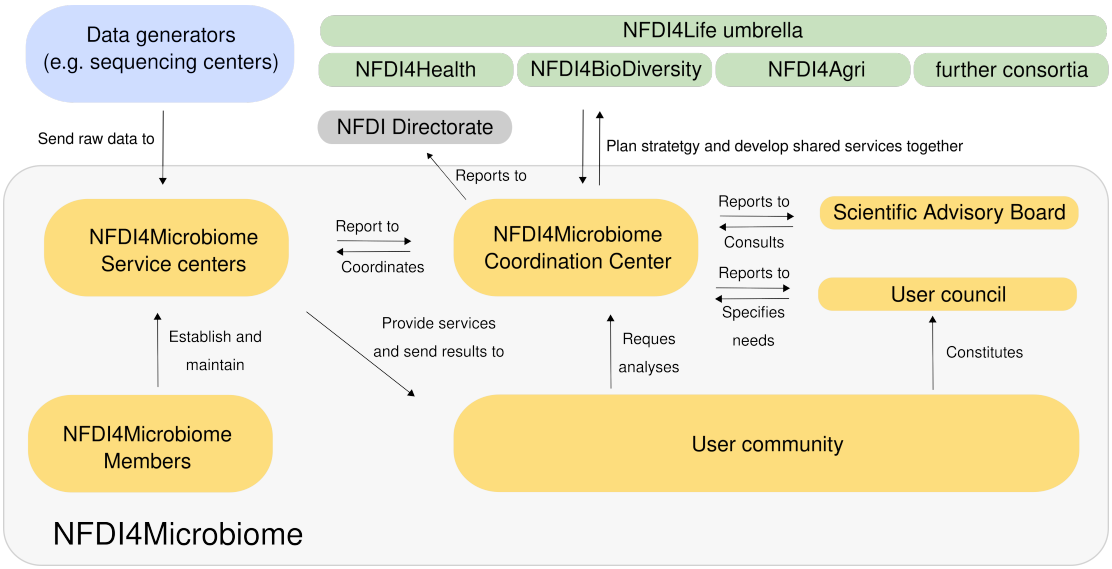
Participating infrastructure facilities and/or potential information service providers (without address)

- ZB MED - Information Centre for Life Sciences
- de.NBI - German Network for Bioinformatics Infrastructure

Planned proposal submission date (2019, 2020, 2021):

2019

Overview diagram or organisational chart for the planned consortium



2. Subject-specific and infrastructural focus of the planned consortium

2.1 Key questions/objectives of the consortium

Scientific background: A large fraction of existing microorganisms is associated with plants, animals, and human beings, where they typically exert essential functions. However, the complexity of these microorganisms and their symbiotic interactions is yet poorly understood. Microbiota (including viruses) have a strong impact on many aspects of human life, starting from health to ecologically relevant processes. Given climate changes our understanding of ecosystems needs to improve dramatically in order to be able to act against men-made issues and microbiota have been largely ignored. One of the major medical threads according UNO is antibiotic resistance. Both seemingly uncoupled important issues require a better understanding

of the microbial world. Furthermore, several uncultured species are very precious study subjects for the identification of compounds with relevance for biotechnology and medicine, and still need to be explored. There one of the biggest challenge in the understanding of microbiota lies in the complexity of numerous biotic interactions between the specific strains of a microbiota and their abiotic environmental factors. Mapping and deciphering those molecular interactions and the underlying regulatory mechanisms is a crucial step towards an understanding and usage of microbiota. However, gathering sufficient data for the analysis of these microbiomes and their interactomes is a challenging and rewarding task. While genomics and other omics aproaches deliver a pleathora of data, to understand them they need to be organised and research needs to be supported. Data gathered at separate molecular levels (DNA, RNA, protein, metabolite) from different and maybe all available bacterial strains has to be processed and integrated to decipher the functioning of the individual species in the microbiome, the complexity of the microbiome overall and its contributions to human health and environmental benefits.

Objectives: The vision of the NFDI4Microbiome consortium is to make the analysis of diverse microbiome related data consistent, reproducible and accessible to all fields of the life science community. It will assist researchers with different scientific challenges to understand microbial communities and the interaction between the species in them. For this purpose the consortium will provide the computational infrastructure as well as analytical tools for the community to compile, analyze and store various types of data with the aim to decipher the interspecies interactions on a molecular level. NFDI4Microbiome will enable efficient and reproducible processing of metagenomes, meta-transcriptomes, meta-proteomes and meta-metabolomic information as well as of data from single cell sequencing. It will enrich this data by metadata from databases and by knowledge automatically extracted from literature and make the data interoperable. The increased understanding of microbiota interactomes and bacterial interspecies interactions facilitated by this infrastructure will be beneficial for biotechnology, agriculture, ecology, and medicine.

2.2 Known needs/current status of research data management in the relevant discipline/subject-specific relevance of the planned consortium

2.2.1 From a research perspective

The need for efficient analysis of microbiome related data exists in different distinct research fields. For instance, the impact of the microbiomes on the human body - most prominently in the gut - on human health is demonstrated to be significant, but the interaction mechanisms are currently far from being entirely understood. Similarly, microbiomes have a tremendous effect on the physiology of agriculturally relevant plants and animals. The importance of microbial species on the ecosystem of the oceans is another showcase of how an understanding of interaction of microbial species with biotic and abiotic entities is crucial for our survival on this planet. These exemplary research fields have very different scientific questions and origins but share the need for efficient analysis of microbiomes and the molecular interaction of their members. Often researcher do not have the capabilities to deal with high-throughput technologies and are overwhelmed by the size as well as the complexity of the multi-Omics data and required processing steps. To address these shared needs by a community, that is fragmented over numerous fields, common standards and a consistent research infrastructure to store and translate the data into insights is required. Such a solution would also make comparative analysis as well as search and reuse of data sets across research fields possible.

2.2.2 In terms of available information providers and services

There are currently internationally some information service providers that address parts of the required analysis workflow. For example the Joint Genome Institute (JGI) offers the platform "Integrated Microbial Genomes & Microbiomes (IMG/M)" that assists in the annotation, analysis and distribution of microbiome datasets sequenced at JGI and offers the service to scientists world wide if they agree with the IMG/M data release policy. Several web-services like MEGAN provide functions like taxonomic or functional analysis for metagenomics data sets. Until now, there is no unified service starting from the data generation to the integration of the diverse data sets nor for the exploration of the functional interaction between microbiota, which is essential for the understanding of their underexploited benefits.

2.3 Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes

The consortium will provide infrastructure for the consistent and complete analysis of data that is required for a holistic understanding of microbiomes and the interactions of their species. Service centers of the consortium will directly receive raw data from collaborating data generating facilities (sequencing and mass spectrometer centers) for analyses ordered by users. The data will be analyzed, combined, enriched and finally submitted to repositories (see 2.4). The computational infrastructure will be provided by the consortium member de.NBI (German Network for Bioinformatics Infrastructure).

2.4 Description of data types and of underlying data processing / data analysis methodologies

The intended infrastructure will cover the whole analysis workflow from raw data to the generation of biological insights. Data from high-throughput technologies (second and third generation sequencing devices for DNA and RNA and mass spectrometry for proteins and metabolites) - ideally directly received from the four recently established sequencing centers and other institutions - will be stored and associated with a rich set of metadata that describes the sampling conditions. The data will undergo quality control and primary processing (e.g. quality trimming) and then will be channeled into user selected analysis workflows for example with read mapping, read assembly, sequence binning, species detection, gene annotation and enrichment with data from databases and linking to relevant literature. Integration and correlation of the different data types will be performed to generate models of interactions and to give a systems perspective on species and their interactions. The final results will be semantically enriched by metadata to have machine readable data publication. The workflow will also include the easy submission of the raw as well as processed data and their associated metadata to community accepted repositories like NCBI Short Read Archive (SRA) or EBI European Nucleotide Archive (ENA) for long term preservation.

2.5 Planned implementation of the FAIR principles and information about any existing policies or guidelines in the relevant discipline

NFDI4Microbiome will fully comply to the FAIR (Findable, Accessible, Interoperable, Re-usable)[1] principles and promote Open Science with all its facets. As part of this, the consortium will define

[1] *https://www.force11.org/group/fairgroup/fairprinciples*

a required, rich set of metadata that describes the sampling conditions and will allow only the submission of data after this metadata was provided and quality controlled. Data stewards will assist in the process of compiling the metadata. Members of our consortium have been leading the development of the International Human Microbiome Standards (IHMS)[2] and NFDI4Microbiome will promote similar standards for other microbiome sources together with the community. Furthermore, the consortium will encourage contributors to choose rather permissive licenses (ideally CC0)[3] for the submitted data sets in order to avoid legal barriers for data sharing. The consistent management of the submitted data as well as the contributed rich annotation with metadata form the core to the powerful search of the original data (and given results) and to the efficient reuse and comparison by the research community. NFDI4Microbiome aims to follow good software engineering practice to generate a software stack based on FLOSS (Free/Libre/Open Source Software) and all developed software will be made public under OSI (Open Source Initiative) compliant licenses. For this purpose NFDI4Microbiome will cooperate with the RSE4NFDI.

2.6 Planned measures for user participation and involvement

NFDI4Microbiome has strong foundations in the microbiome community and is consulted by representative organisations from the community including the German Association for General and Applied Microbiology (VAAM). Currently the process to reach out to further members of the community and the creation of a questionnaire to get a structure overview of the needs are ongoing. Once established, NFDI4Microbiome will install a user council and a scientific advisory board to have dedicated feedback channels in order to adapt the operations to the requirements and needs of the community.

2.7 Existing and intended degree of networking of the planned consortium

The NFDI4Microbiome consortium is tightly connected with the umbrella consortium NFDI4Life and several sister consortia including NFDI4BioDiversity, NFDI4Health and NFDI4Agri and aims to collaborate closely with further future NFDI4Life related consortia. Together these consortia will work on solutions in shared areas of interest for example in the definition of new standards and common technical infrastructure in case of same data types.

2.8 Additional information

Show cases/experiences: Members of the consortium have provided computational infrastructure, analytical tools and workflows for the project consortia SIMBA (Sustainable Innovation of Microbiome Applications in Food System)[4] and Virus-X[5].

[2] *http://www.microbiome-standards.org*
[3] *https://creativecommons.org/publicdomain/zero/1.0/*
[4] *https://www.luke.fi/en/projects/simba/*
[5] *http://virus-x.eu/*