MoU JRC – FVG

AREA 1 – Mobility Scheme

JRC proposed Activity

Reference no.:	**26**
JRC Directorate	F – Health, Consumers and Reference Materials
Unit	F.7 – Knowledge for health and consumer safety
Location	JRC, Ispra (Italy)
Short description of the activities of the Unit	The mission of F.7 is to support EU policies on consumers, food safety and health by mapping, collating, analysing, quality checking and communicating in a systematic and digestible way all the relevant scientific data, methods, tools and knowledge available worldwide respect to their impact on policy.
Title of the JRC proposed Activity:	**Development of a sequence-signatures-fishing bioinformatics pipeline.**
Short description of the proposed activity:	Public genomic, metagenomic, metatranscriptomics and sequencing data in general are becoming an invaluable resource for meta-analysis, allowing to quickly increasing our knowledge for answering to the most diverse questions while reducing the need to generate new data. Indeed, large scale data produced within a specific study are being often used to answer questions which go beyond the original scope for which the data were initially produced. This is an inherent quality of large scale genomic data. Starting from these considerations, it is here proposed to develop a bioinformatics pipeline to identify specific user-selected sequence markers and signatures from large and public collections of sequencing data. Specifically, the main scientific interest is the capability of recognize sequence fingerprints as molecular markers in sequence data deriving from the most diverse environments and samples to answer to questions such as: is there any synthetic sequence in (meta-)genomics data? Is there any association between mobile elements activity and specific diseases or phenotypic traits? Has a specific virus infected a given cohort of individuals? To answer these and many other similar questions, once identified the specific sequence signatures, it is fundamental to fish them in a big collection of sequencing data. This activity hopefully will culminate in the development of a computational tool that will assist in doing that, implemented as a modular bioinformatics pipeline. Specific care needs to be given in the choice of the algorithm to use for the search to optimize the sequence search in a huge database. For instance it will explore the possibility to use algorithms such as Sequence Bloom Trees that have been proven to allow such searches over a reasonable amount of time (https://www.nature.com/articles/nbt.3442). However, specific

	study will be performed at the beginning of the Collaboration on the most updated literature to choice the most updated and optimized strategy. The pipeline will be modular allowing for great flexibility and user manoeuvring for infinite expandability over the time based on community needs and requirements. Two test case searches will be used as pilot and proof of concept into the initial development: 1) search for identification of fingerprints of artificial sequences in large whole metagenomics sequencing datasets from environmental samples; 2) search for identification of target site duplication (TSD) as marker of retro-virus/transposon activity in large metatranscriptomics sequence datasets from neurodegenerative disorder disease samples.
Required profile of the Partner Institution:	University or Research Institution recognised as scientific centre of excellence within the national and international academic scene, with relevant research activities in Mathematics, Computational Biology, Bioinformatics applied to Omics and Neuroscience. High quality scientific works carried out by its researchers are expected to be published regularly in leading international journals with a high impact factor, including the most prestigious scientific journals. Occurrence of collaboration agreements with other world's leading research institutes and universities is also highly desired.
Indicative required profile of the researcher/expert (that will implement the activity)	Expert in bioinformatics/computational biology, with special skills on data integration and harmonization, development of tools, methods and databases for large-scale functional genomics data analysis.