Project Deliverable: D-JRP-PARADISE-WP3.1

Workpackage 3

Responsible Partner: P41-SVA, P11-RKI

Contributing partners: P27-ISS, P1-ANSES

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**60 Months**

DOCUMENT MANAGEMENT

Project deliverable	D-JRP-PARADISE-WP3.1
Project Acronym	PARADISE
Author	Christian Klotz, Katja Winter, Karin Troell, Ema Östlund, Simone M. Cacciò
Other contributors	
Due month of the report	30
Actual submission month	36
Type *R: Document, report* *DEC: Websites, patent filings, videos, etc.;* *OTHER*	R **Save date**: 20-Jan-21
Dissemination level *PU: Public (default)* *CO: confidential, only for members of the consortium (including the Commission Services)*	PU **This is the default setting.** If this project deliverable should be confidential, please add justification here (may be assessed by PMT): ....
Dissemination *Author's suggestion to inform the following possible interested parties.*	OHEJP WP 1 ☒ OHEJP WP 2 ☒ OHEJP WP 3 ☒ OHEJP WP 4 ☒ OHEJP WP 5 ☐ OHEJP WP 6 ☐ OHEJP WP 7 ☐ Project Management Team ☒ Communication Team ☒ Scientific Steering Board ☒ National Stakeholders/Program Owners Committee ☐ EFSA ☒ ECDC ☒ EEA ☐ EMA ☐ FAO ☐ WHO ☐ OIE ☐ Other international stakeholder(s): ..... Social Media: ..... **Other recipient(s):** .....

Report on the *in-silico* selection of highly polymorphic sequences in *C. parvum* and *G. duodenalis* genomes

1. Background

One of the main goal of the PARADISE project is to develop new, informative typing schemes for the molecular characterization of the protozoa *Cryptosporidium parvum* and *Giardia duodenalis*. Historically, the limited knowledge about the genomes of these parasites has limited the opportunity for a rationale selection of genetic markers, and efforts have mostly focused on loci containing repetitive sequences (mini- and micro-satellites) and a few polymorphic genes. The limitations in the use of this type of markers are well recognized, and no standardized genotyping schemes are currently available.
The introduction of the Next Generation Sequencing (NGS) methodologies opened the possibility to sequence the whole genome of a given organism, and one possible use of genome data is the selection of highly polymorphic markers. It is important to recall that the analyses of the data generated by NGS experiments requires bioinformatics workflows (pipelines) that often should be tailored for the specific pathogen under analysis. For example, *Giardia duodenalis* has a tetraploid genome and isolates are known to differ in terms of allelic sequence heterogeneity, whereas *Cryprtosporidium* has a haploid genome,but the obligatory sexual phase of this parasite provides ample opportunities for recombination among isolates. As such, the specific challenges for data analysis are different for the two parasites.

The PARADISE project has been designed to address both aspects, namely 1) to provide new whole genome data for both parasites, and 2) to develop and optimize the necessary pipelines for data analysis and marker selection.This report describes the procedures that were undertaken for the first *in silico* detection of variable regions suitable for the subsequent selection of markers for both parasites.

1.1 Giardia duodenalis

An *in silico* selection pipeline has been developed to identify polymorphic sequences suitable to develop a new Multi Locus Sequence Typing (MLST) scheme for *G. duodenalis*, with a specific focus on on og te two variants associated with human infection, assemblage B. The analysis aims at selecting genomic regions with a length of 500-700 bp, that are present in all assemblage B isolates but show differences at the sequence level to allow distinguishing different genotypes. Additional selection criteria include a low level of allele sequence heterogeneity (ASH) within each sample and an equal distribution of these regions across the five chromosomes.
In total, 18 samples of *G. duodenalis* assemblage B were included in the first analysis. Genomic sequence assemblies of two isolates were publicly available on NCBI. Additional five isolates were sequenced and de novo assembled at the RKI. Genome sequences of the remaining isolates were created using reference-base methods from already available sequence data within the consortium or from sequences generated by the PARADISE project (WP2, Task 2.1).
A fully automated and comprehensive analysis workflow has been developed (Figure 1).The bioinformatic analysis is based on a pairwise whole genome alignment of the *G. duodenalis* assemblage B samples, which are then combined into a multiple sequence alignment (MSA). Based on the MSA, genomic regions with a minimum length of 1000 bp are selected based on the condition that they occur in all input genomes. Giving a window size (500-700 bp) it was evaluated whether each position, or column, of the MSA allows to differentiate the *G. duodenalis* assemblage B genomes included in the analysis. Those positions were regarded as positions of interest (POIs).
Additionally, the ASH was determined for each region in the MSA by isolate-specific mapping to the original sequencing data and by assessing the nucleotide frequencies in the respective read alignments.

Taking into account the ASH, the read coverage, the sequence length and the number of POIs, the initial hit regions were filtered and ranked. Using the data set of 18 samples, 104 putative regions were selected, and will be manually inspected and considered to conduct differentiating multiplex primer sets.

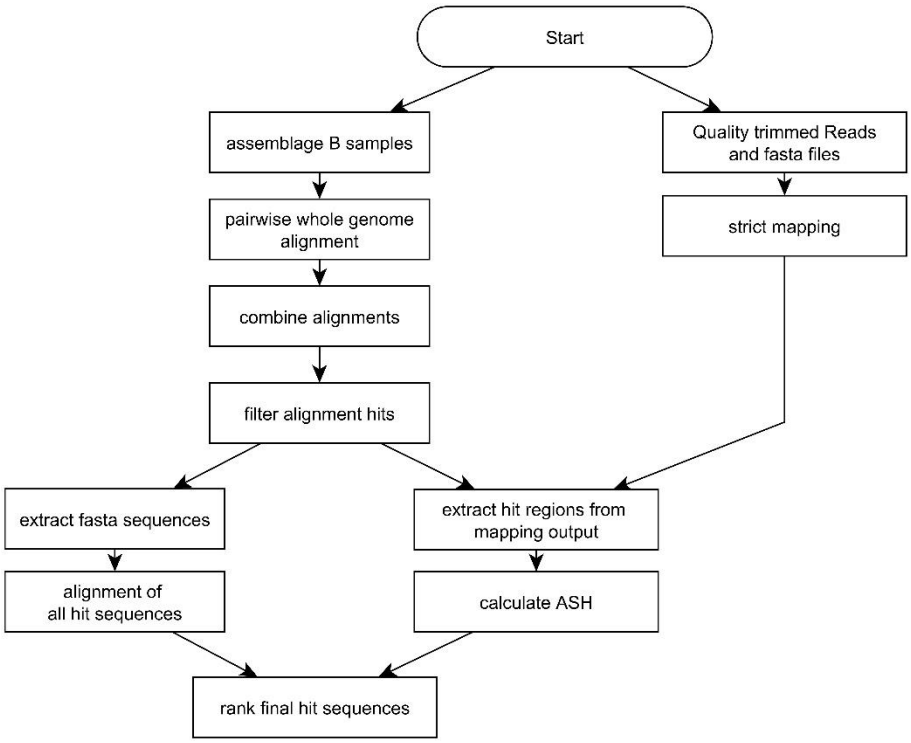

Figure 1: General workflow of the bioinformatic analysis to detect putative suitable primer regions in G. duodenalis assemblage B genomes; ASH – allele sequence heterogeneity

1.2 Cryptosporidium parvum

An *in silico* selection pipeline has been developed to identify polymorphic sequences suitable for subsequent primer selection to develop a new MLST scheme for *Cryptosporidium parvum*.
The bioinformatic analyses were conducted on 81 *C. parvum* genome sequences, which were generated by the PARADISE project (WP2, Task 2.1) or in the context of previous project invonving consortium's partners, or retrieved from public databases. Of these, 11 samples were removed due to a high level of contamination or low number of sequence reads. Therefore, 70 samples were retained for the analysis, of which 3 were treated as single end data after removing reads originating from bacterial contamination. The raw paired-end reads were trimmed of adapters and low-quality bases, and downsampled to 100x coverage based on the size of the chosen reference genome (*Cryptosporidium parvum* IOWA, release 46). The IOWA genome sequence and its annotation was used as reference for all subsequent analyses requiring a reference.
The EBI-Parasite pipeline scripts, developed in the context of a previous Horizon 2020 project (COMPARE), were used on the trimmed and downsampled reads. The pipeline steps include assembly, mapping of reads to a reference genome, variation analysis, repeats analysis, analysis of variability within coding sequences, and the generation of multiple sequence alignment of single chromosomes (Figure 2).
To select regions suitable for potential markers, the genomic regions identified by the pipeline as containing variable sites (i.e., showing Single Nucleotide Polymorphisms, SNPs) were ranked using

Simpson's diversity index, which provides a measure of how well the sequence discriminates between the samples.
The sequences were selected for manual examination based on the number of POIs within 500 bases. A multiple sequence alignment (MSA) of all 70 samples was extracted for each selected sequence. To allow primer design, additional 100 nucleotides at the 5' and 3' flanking regions were extracted and included in the MSA. A subset of 61 variable regions was retained at this stage.

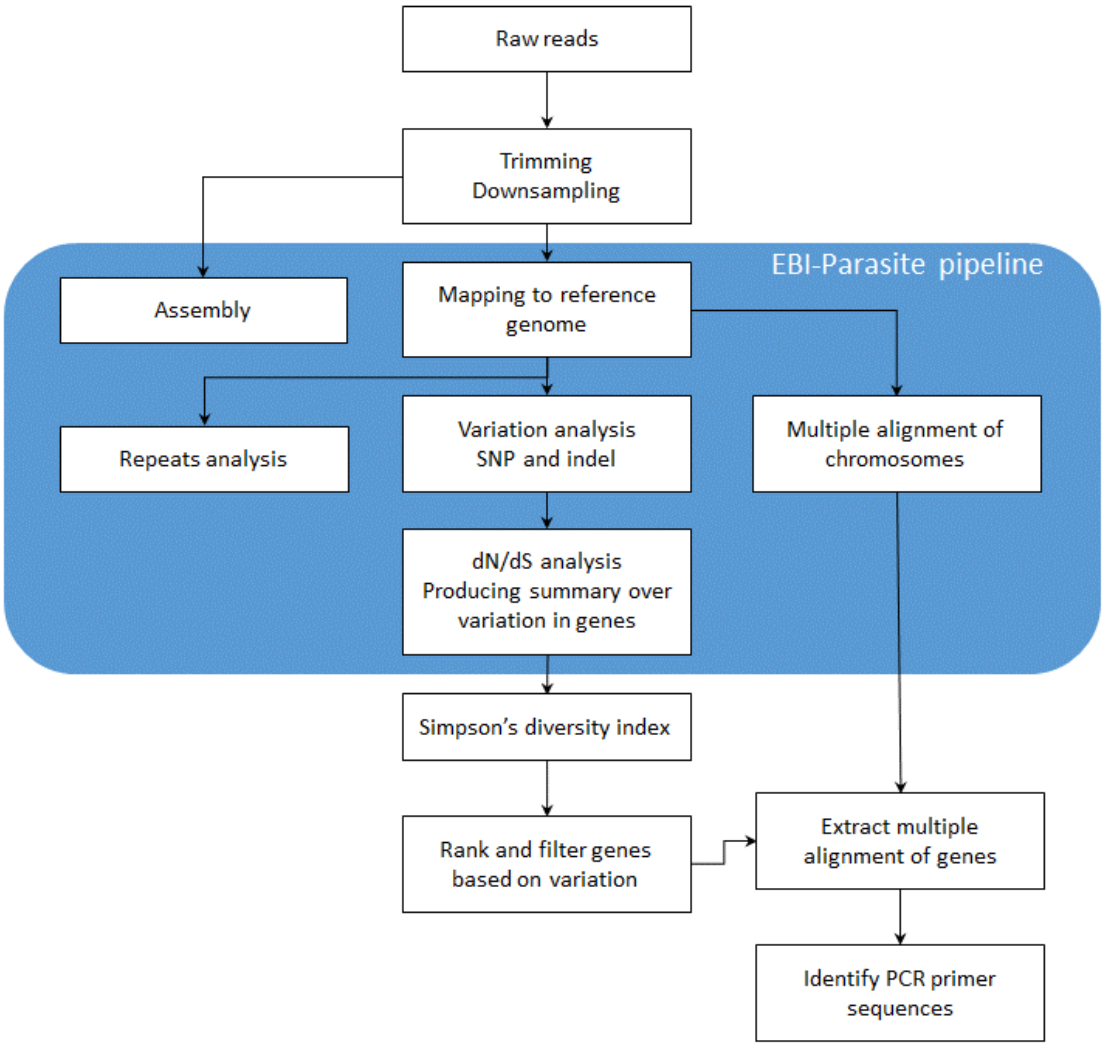

Figure 2. General workflow of the bioinformatic analysis used to setect *C. parvum* variable regions

2. Conclusions

All available genome data for *Cryptopsoridium parvum* and *Giardia duodenalis* (assemblage B) were retrieved and analyzed by two independent workflows at SVA and RKI. Selection criteria used to identify variable regions included the presence of an informative number of SNPs and of a reduced ASH in the selected sequences (for *G. duodenalis*).

At present, the in silico selection resulted in a list of 61 variable regions suitable for putative markers for *Cryptopsoridium parvum* and 104 for *Giardia duodenalis* (assemblage B). The next steps wil include the design of primer pair sequences for PCR amplification of the target regions and the experimental evaluate of the selected variable regions as putative markers in a larger set of samples (deliverable D-JRP-PARADISE-WP3.2 due M42).