Project Plan

Background

Leptospirillum ferriphilum is a Gram negative, chemolithoautotrophic, and acidophilic bacteria i.e. bacteria that is found in acidic, metal-rich environments. They are obligate aerobe and gain energy only by oxidizing iron, generating ferric iron (Fe^{3+}) using ferrous iron (Fe^{2+}) as an electron donor. (Christel et al. 2018).

Due to their ability to catalyze mineral dissolution, *L. ferriphilums* are used in biomining (Christel et al. 2018). Biomining is an industrial process where living organisms extract metals from solid material. ´

L. ferriphilum is important in gold recovery and has been identified in the bio-leaching pile for the recovery of chalcopyrite coppers. In spite its importance much is unknown about this organism, thus mapping the hole genome will provide a better understanding of it physiological processes (Christel et al. 2018) and thereby improve the efficiency of biomining.

Aim

This project is about to re-produce some of the analyses in Christel et al., re-analyzing their data and re-evaluating their biological conclusions. The aim of the project is to have a deep understanding of some bioformatic methods relevant in this project and, to get familiar with bioinformatic tools and methods that are commonly used when analyzing sequencing data. One of the purposes is to become aware of the continuous development of these methods and the impact of the updates.

Methods

To different analyses will be performed. The first is to use PacBio and obtain the whole genome sequence. This will be done through *de novo* assembly of the genome from long reads. The fully assembled genome will then be annotated to investigate its synteny with a closely-related species. The second is transcriptomics and differential gene expression analysis using paired reads obtained by RNA-seq.

Table 1: Analysis 1, Genome Assembly (Data from WGS):

Analysis	Software	Running time
Genome assembly	Canu	~ 11,5 h (2 cores)
Assembly evaluation	Quast	< 15 min (1 core)
Assembly evaluation	MUMmerplot	< 5 min (1 core)
Annotation	Prokka	< 5 min (2 cores)
Annotation	eggNOGmapper	~ 1 h (HMM algorithm)
Synteny comparison	blastn	

Table 2: Analysis 2, Transcriptomics and Differential Gene Expression

(Data from RNA-Seq):

Analysis	Software	Running time
Quality control	FastQC	
Trimming	Trimmomatic	~ 15min per file, 5 files (2 Cores)
Quality control	FastQC	
Aligner	BWA	~ 5 h (2 cores)
RNA-seq reads counting	Htseq	~ 8 h
Differential Expression	Deseq2 (Rlibrary)	

Workflow

Following data analyses will be performed:

- Genome assemble of PacBio reads.
- Assembly quality assessment
- Structural and functional annotation.
- Synteny comparison with a closely related genome
- Reads preprocessing: trimming + quality check (before and after)
- Mapping and counting RNA-seq reads, and analyzing differential expression.

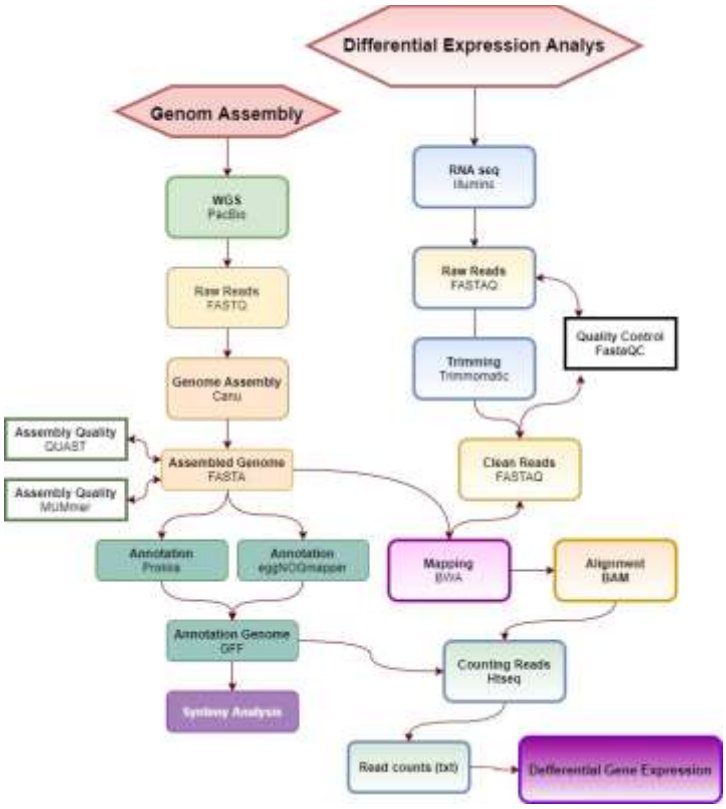

Figure 1 Workflow for analysis 1 (Genome Assembly) and analysis 2 (Differential Gene Expression)

Project organization

Data and code are separated. Folders or file names starts with a number, since by default they will be shown in alphanumerical order. It is easier to know in which order they were created when they are numerically organized. Data files, especially big data files, os compressed. I will be working both on my local computer and on UPPMAX depending on which program I will be running.

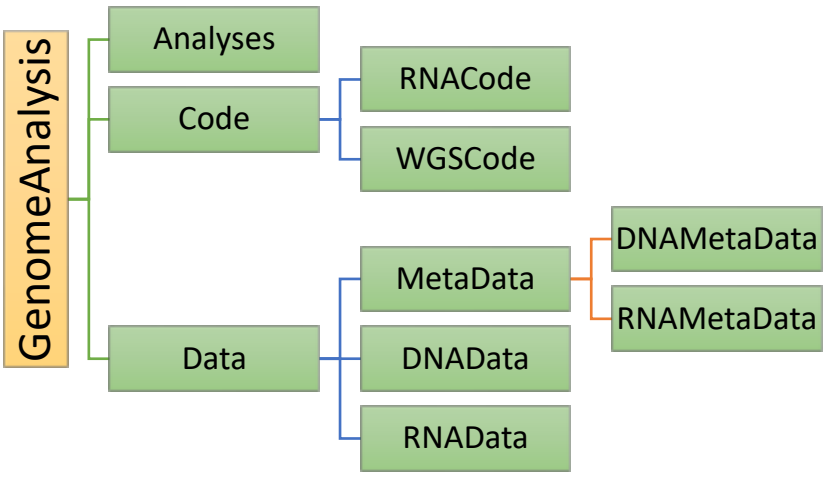

Figure 2 Data structure of the repository.

Timeplan

The timeframe of the project is 24/3-20/5 and the checkpoints of the different methods are noted in the table below.

Table 3: Analyses checkpoints

weak	**Deadline**
13	Seminar
14	Project planning
16	Genome Assembly + Genome annotation
16	Comparative genomics
17-18	RNA mapping
19	Synteny

References

Christel S, Herold M, Bellenberg S, El Hajjami M, Buetti-Dinh A, Pivkin IV, Sand W, Wilmes P, Poetsch A, Dopson M. 2018. Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species Leptospirillum ferriphilumT. Applied and Environmental Microbiology, doi 10.1128/AEM.02091-17.