



An integrated solution for handling of rRNA sequence metadata by combining academic and commercial software

R. Kottmann¹, G. Moraru², C. Moraru², J. Peplies³, F. O. Glöckner¹

¹ Max Planck Institute for Marine Microbiology, Germany - ² Heracle Software, Germany - ³ Ribocon GmbH, Germany

Software
Heracle



Ribocon

Motivation

Facts:

- 215,540,553,360 nucleotides in 114,475,051 entries for the current EMBL release 94 (March 2008)
- more than **1,200,000 entries** are represented by the small-subunit **ribosomal RNA** (see statistics at www.arb-silva.de)
- only a very limited amount of secondary data, so-called **metadata**, is yet available in the public databases
- particularly for environmental microbiologists, retrieval of metadata attached to sequence information is **extremely valuable** for subsequent analysis of selected sequences from the public databases and data interpretation.

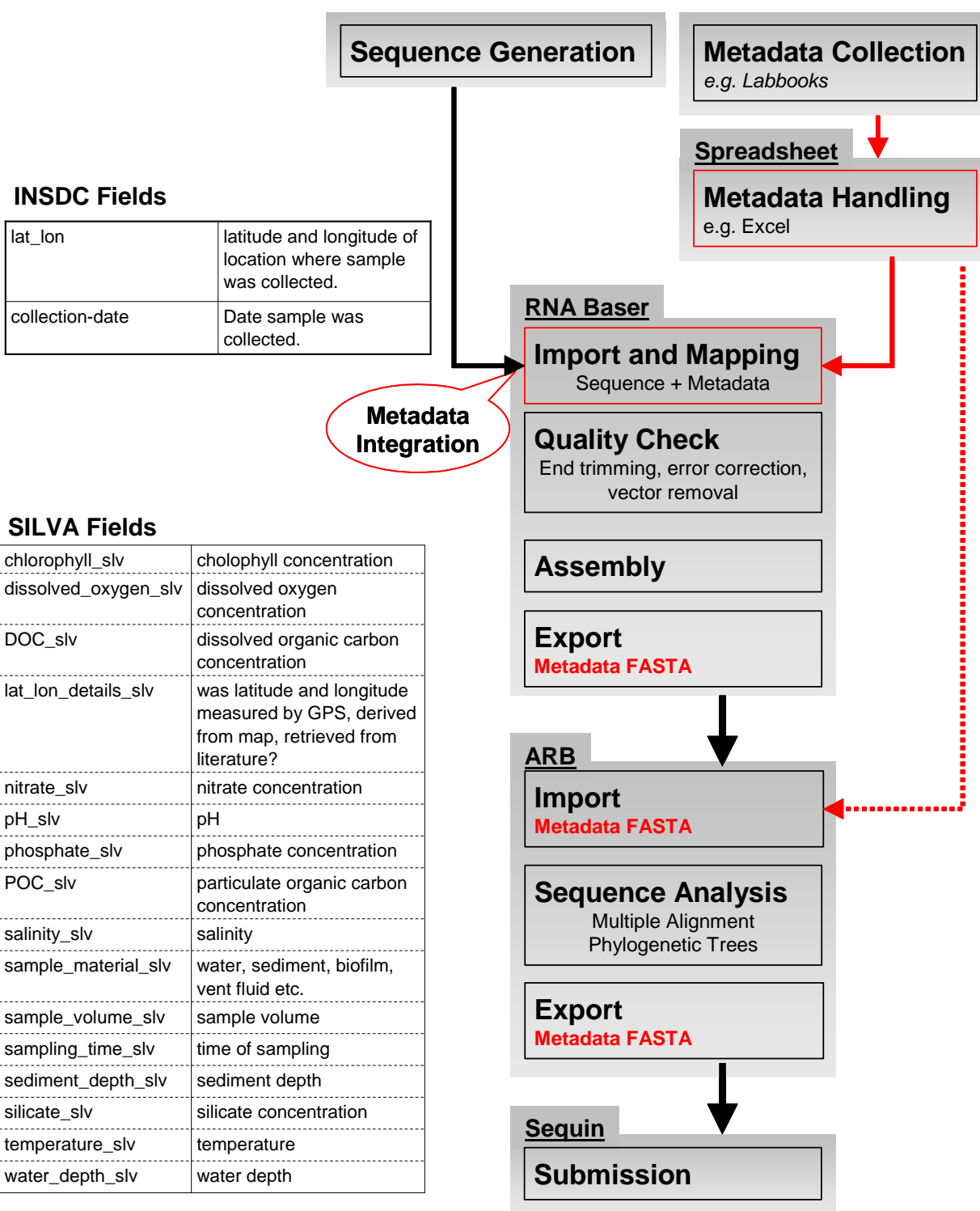
What are metadata?

- additional data related to the primary sequence information
- example: environmental data such as temperature, pH, and salinity or exact geographic position of a sampling site

Problem:

This information is often recorded by the investigators during sampling. However, due to missing standards and missing software solutions which allow for the fast and efficient merging of sequence and metadata for later easy submission, the metadata normally do not find their way into the public databases.

Our intention is to raise awareness for this limitation and to facilitate the process by providing an integrated solution combining standard academic and commercial software tools.



Standard Workflow

- the generation of raw sequence information in the laboratory is followed by a **multistep process** finalized by sequence submission to the public databases
- for initial processing of the 16S rRNA fragments, routinely powerful **commercial software** tools are used (trimming, quality check, editing, and assembly)
- a common tool for classification and phylogenetic analysis of ribosomal RNA sequences is represented by the **free** software package **ARB** and the corresponding **SILVA** databases (www.arb-home.de & www.arb-silva.de)
- for batch submission **free** software tools such as **Sequin** of NCBI (www.ncbi.nlm.nih.gov/Sequin) are available
- for data exchange during the whole process, the **standard FASTA** format is typically used, not including any kind of metadata!

New: Metadata Integration

- **early integration** of the metadata into the workflow!
- metadata stay **attached** to the sequence information in a Sequin compatible FASTA until submission!
- joined initiative including Heracle Software which currently offers the commercial software tool **DNA Baser** (www.dnabaser.com)
- currently, a modified version for rRNA metadata integration (mapping) is in preparation, called **RNA Baser** (www.rnabaser.com)
- freely-available solutions for initial metadata handling based on spreadsheet software such as Microsoft Excel will be offered

Left: Example for a collection of metadata fields (environmental parameters) as introduced by the SILVA (www.arb-silva.de) rRNA databases

Right: Typical workflow of rRNA sequence data from generation to submission (**black**: current standard; **red**: metadata integration).

RNA Baser and spreadsheet tools will become available in early summer 2008!

Check www.rnabaser.com and www.arb-silva.de