An Efficient Pipeline for Microbiome Analysis

Daniel H. Huson*†, Benjamin Buchfink* and Hans-Joachim Ruscheweyh*‡

August 28, 2015

Microbiome analysis using metagenome shotgun sequencing is computationally challenging. A typical project may involve hundreds of samples, each represented by tens of millions of DNA sequencing reads. For functional analysis, it is necessary to align all reads against a comprehensive protein reference database, such as NCBI-NR, which currently contains over 60 million sequences.

Here we present a simple, highly efficient pipeline for analyzing metagenomic reads. Our pipeline has four parts.

1. DIAMOND [1] is used to compare all DNA sequencing reads against the NR database. Our program aligns Illumina reads against protein references at up to 20,000 times the speed of BLASTX, while achieving a similar level of sensitivity. Input is a file of reads in FASTA or FASTQ format (usually gzipped).

2. Then, for each sample, the resulting file ("DIAMOND alignment archive" file with suffix .daa) is analyzed using a new program called "daa2rma" that performs taxonomic and functional binning of all reads. As described in [2], reads are mapped to the NCBI taxonomy using the LCA algorithm and functional analysis is performed by mapping reads to SEED, COG and/or KEGG. The output is an RMA file (MEGAN "Read Match Archive" file with suffix .rma).

3. The resulting RMA files are are then made accessible via the local network or over the world wide web using a new program called MeganServer [3].

4. Project members access the RMA files remotely to interactively analyze and compare their datasets using an upcoming new version 6 of MEGAN [4] .

This pipeline minimizes computational time and disk space, and makes it easy to access results. In an ongoing project, the alignment of one billion Illumina reads against microbial

[*] Center for Bioinformatics, Universität Tübingen

[†] Life Sciences Institute, National University of Singapore

[‡] Department of Biosystems Science and Engineering, ETH Zurich (Basel), SIB Swiss Institute of Bioinformatics, Basel, and Scientific IT Services, Research Informatics, ETH Zurich (Basel)

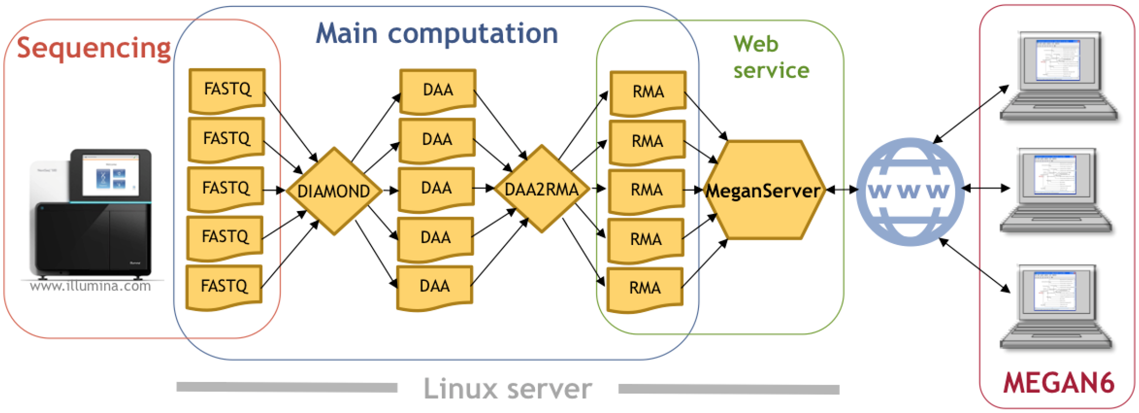

Figure 1: Microbiome analysis pipeline. FASTQ files produced by a sequencer are aligned against a protein reference database using DIAMOND. Taxonomic and functional analyses of the resulting alignments are performed using daa2rma. MeganServer publishes the resulting files to the web. Authorized project members access the files remotely using MEGAN6.

NR using DIAMOND took one day on a single server, and the subsequent taxonomic and functional analysis using daa2rma took about 4 hours. Making files accessible via MeganServer takes no additional time.

All three programs, DIAMOND, MeganServer and an alpha-test version of MEGAN6, are available from:
`http://ab.inf.uni-tuebingen.de/data/software`.

References

[1] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using DIAMOND. Nature Methods **12**, 59–60 (2015)

[2] Huson, D.H., Mitra, S., Weber, N., Ruscheweyh, H.-J., Schuster, S.C.: Integrative analysis of environmental sequences using MEGAN 4. Genome Research **21**, 1552–1560 (2011)

[3] Ruscheweyh, H.-J., Huson, D.H.: MeganServer - Easy interactive access to large-scale metagenome data. In preparation (2015)

[4] Huson, D.H., *et al.*: MEGAN6 - Microbiome analysis involving hundreds of samples and billions of reads. In preparation (2015)