Host_microbe_mapper allows to analyse and interpret the expression of dual RNA-seq measurements and reveals potential microbial contaminations in the data

Thomas Nussbaumer[1,2,3]

[1] CUBE – Division of Computational Systems Biology Dep. of Microbiology and Ecosystem Science, University of Vienna, Austria.
[2] Chair and Institute of Environmental Medicine, UNIKA-T, Technical University of Munich and Helmholtz Center Munich, Neusäßer Str. 47, 86156 Augsburg.
[3] Institute of Network Biology (INET), Helmholtz Zentrum München (HMGU), German Research Center for Environmental Health, 85764 Neuherberg, Germany.

Abstract

Next-generation sequencing technologies provide a wealth of sequencing data. To handle these data amounts, various tools were developed over the last years for mapping, normalisation and functional analyses. To support researchers in the interpretation of expression measurements originating from dual RNA-seq studies and from host-microbe systems in particular, the computational pipeline *host_microbe_mapper* can be applied to quantify and interpret dual RNA-seq datasets in host-microbe experiments. The pipeline with all the required scripts is stored at the Github repository (https://github.com/nthomasCUBE/host_microbe_mapper).

Background

Different RNA-seq technologies generate huge amounts of data in a cost-efficient manner. To support researchers during this analysis, various computational methods can process the data such as methods for quantification and functional interpretation. For the mapping of single or paired-end reads, the tools STAR [1], Tophat [2] and *ballgawn* [3] are commonly applied in eukaryotic genomes whereas *bowtie* [4] and *bwa* [5] are commonly used for the mapping of microbial reads.

After reads have been assigned to the genome, various ways to interpret the data exist, but it needs a normalisation step before. Whereas in the *cufflinks'* [6] package, there are options to directly obtain FPKM (Fragments per Kilobase Million) counts that consider the gene length and the paired-end information, for other tools it needs to include additional programs to obtain raw read counts which can be extracted with tools such as *featureCounts* [7] or HT-seq [8]. This allows to normalize the data into TMM (Transcripts Per Million Mapped) or RPKM (Reads Per Kilobase Million) in order to perform *e.g.* the detection of differentially expressed genes or to obtain clusters of co-expressed genes with *WGCNA* [9]. Differentially expressed genes can be computed with tools such as *EdgeR* [10], *Noiseq* [11], and *DESeq2* [12] whereas *cufflinks* and *ballgawn* allow to detect differentially expressed genes as part of their pipelines. Most of the mentioned tools have been included into the current pipeline.

An additional aspect of the mapping of expression can be the removal of ribosomal RNA, which can contribute a major proportion of reads in a sequencing run. Reads of ribosomal origin can be discarded from most experiments with available kits (*e.g.* RiboZero Gold kit [13]) or *in silico* by tools such as *SortMeRNA* [14]. Otherwise, it might be of major interest to detect putatively unexpected contaminations in the data. To analyse contaminations based on 16S rRNA genes, we have also included *SortMeRNA* [14] to classify reads, that belong to eukaryotic or prokaryotic ribosomal genes. To reveal or confirm the availability of certain microbes it might be of interest to reconstruct the entire 16S or 18S ribosomal gene from the transcriptomic data. For this aim, there are tools such as *REAGO* [15] that allows to assemble these reads. Otherwise, for dual RNA-seq experiments, it can be crucial to verify, that the pathogen is the main microbial source in the dataset and host response might not active because of other microbes.

Before the entire read set is used to obtain an insight into the data, it needs preliminary access to the data to grasp a first insight into the dataset and to define most appropriate parameters that can be applied on the entire dataset. This requires much more time and computational resources in the following even so a test set might already give a good overview of the data already. This can be ideally done by randomly selecting reads (*e.g.* 100,000) from the dataset. To ease the usage of mapping dual RNA-seq data, we provide a

simple graphical user interface that allows to define the reference genomes and to define whether the full or a subset should be used. Despite a bunch of tools that provide a guideline for mapping dual RNA-seq experiments, the pipeline as described in this study has the advantage that is can be run directly on the data without need for major changes as we demonstrate in this study. Calculations can be also run in parallel by using the SLURM batch system. We demonstrate the pipeline by selecting running the experiments on various published different datasets.

Material and Methods

Data access

RNA-seq datasets and respective genomic sequences from human, mouse and the pathogen *Neisseria meningitidis* were obtained from NCBI server and Gene Bank identifiers are summarized in **Supplemental Table 1** and were used to obtain artificial read datasets with the ArtificialFastqGenerator.jar by using the coding sequences of the respective genomes [16]. All datasets represent paired-end sequences with 300 nt inserts. In total, 20 million reads were generated for each of the hosts as well as two million reads from the microbe *Neisseria meningitidis*. The respective demo files are included on the Github page.

Pipeline construction

The whole pipeline is summarized in a single batch script, labelled as *host_microbe_mapper*, that contains all the relevant commands starting from the pre-processing of the data, the mapping of host and microbial reads and finally the data normalisation. All scripts are stored in the Github repository (https://github.com/nthomasCUBE/host_pathogen_mapping). Additional prepared scripts can be used to obtain the differentially expressed genes. Therefore, we integrated scripts to perform differential expression analyses by comprising methods such as *EdgeR*, *NOISeq* and *DESeq2*, however these scripts might need additional information and mapping files between samples and conditions. The whole functionality and mapping was performed on the Leibnitz Rechenzentrum Cluster.

Results and Discussion

General description of the pipeline

The entire workflow of the program is given in **Figure 1**. A user can provide reads either in the raw FASTQ format or in the compressed format (in the GZ format) by defining the path to the directory that contains them by running the graphical user interface. Each sample within

that directory is serially processed. Before reads are mapped to the respective genomes, we perform a quality assessment by checking the FASTQ sequences with *FastQC* [17], followed by adapter removal with *cutadapt* [18]. Next, reads are cleaned from ribosomal reads by considering the *SortMeRNA* pipeline [14] that reports reads matching to ribosomal genes from both, eukaryotes and prokaryotes and are assembled with REAGO [15]. Next, reads are mapped first to the microbe and remaining reads are then mapped to one or two hosts. This is case, when *e.g.* human DNA material is grafted into a mouse model, thereby three different organisms can be simultaneously mapped. **Supplemental Table 2** lists the amount of reads in total that were mapped for the individual datasets and the reads that remain after adapter removal, pre-processing and mapping. These information are collected inside a project-specific Excel file to provide an intuitive overview of the mapping results and furthermore visualisations are made to depict how many of the reads are uniquely mapped, multiple mapped or remain unmapped (**Figure 2A**). For the mapping of eukaryotic reads, STAR [1] or *Tophat* [2] are used and can be selected by the user. Before, reads are mapped to the microbe by using either *BWA* or *bowtie*. Finally, remaining reads are used and are normalised and then differentially expressed gene are applied (*EdgeR* and/or *DESeq2)*. For each individual step in listed in separate script files.

Analyses of the dual RNA-seq datasets from published studies

Artificial datasets were generated (see Material and Methods) including dual RNA-seq datasets, that contain a low, median and a high number of microbial reads being present at 1%, 5% or 20%. We used human and mouse expression transcriptomic datasets to simulate these reads from the respective human, mouse genes and pathogen. Then, we applied the *host_microbe_mapper* pipeline and compared known expression counts to the expression, that was from measured from the various mapping tools.

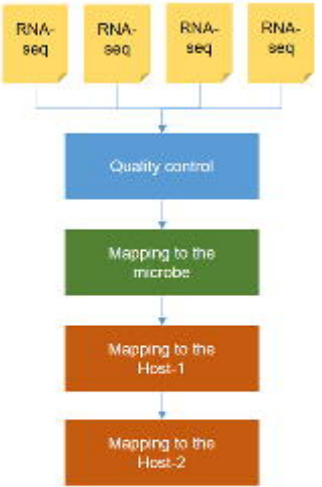
Possibility to assign the 16S genes from the transcriptome dataset directly

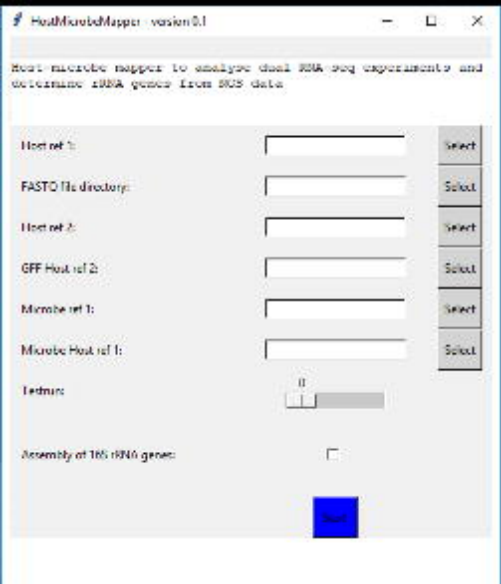
As an additional feature, *host_microbe_mapper* can extract reads from the transcriptome dataset itself to assess which microbes appear in the original raw data, *e.g.* to reveal putative contamination or to detect *e.g.* endophytes, that co-exist with the host. We extracted the classified reads from *SortMeRNA* to extract and then to assemble the entire 16S rRNA genes which can be then compared to reveal existing microbial taxa from the data. To test the pipeline, we have used again reads from the host but now added and simulated the reads from the different microbes. To simulate this, we included microbial datasets from other studies and checked then whether we were able to find them again. Typically use cases for this method would be experiments, that were run in the same batch and thereby might, given misleading barcoding, contain a mixture of reads from different experiments.

Figure 1. Illustration of the next-generation sequencing (NGS) pipeline covering the steps from the integration of the data to the mapping to the microbe and furthermore the mapping to the reference genomes from both hosts.

Figure 2. Graphical User Interface for mapping RNA-seq data for the eukaryotic and prokaryotic genomes preparing the mapping files that can be run then on the server.

1. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
2. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
3. Frazee, A.C., et al., *Ballgown bridges the gap between transcriptome assembly and expression analysis.* Nat Biotechnol, 2015. **33**(3): p. 243-6.
4. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
5. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
6. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
7. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
8. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
9. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
10. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
11. Tarazona, S., et al., *Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package.* Nucleic Acids Res, 2015. **43**(21): p. e140.
12. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
13. Benes, V., J. Blake, and K. Doyle, *Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA.* Nat Meth, 2011. **8**(11).
14. Kopylova, E., L. Noe, and H. Touzet, *SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.* Bioinformatics, 2012. **28**(24): p. 3211-7.
15. Yuan, C., et al., *Reconstructing 16S rRNA genes in metagenomic data.* Bioinformatics, 2015. **31**(12): p. i35-43.
16. Frampton, M. and R. Houlston, *Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines.* PLoS One, 2012. **7**(11): p. e49110.
17. Andrews, S., *FastQC - A quality control tool for high throughput sequence data.* 2017.
18. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* Bioinformatics in Action, 2012. **17**(1): p. 10-12.


RNA-seq
RNA-seq
RNA-seq
RNA-seq
Quality control
Mapping to the microbe
Mapping to the Host-1
Mapping to the Host-2
Gene – Functional categories
e.g. effectors
e.g. immunity genes
e.g. candidate from literature


HostMicrobeMapper - version 0.1
Host microbe mapper to analyse dual RNA Seq experiments and determine sRNA genes from NGS data
Host ref 1:
Select
FASTQ file directory:
Select
Host ref 2:
Select
GFF Host ref 2:
Select
Microbe ref 1:
Select
Microbe Host ref 1:
Select
Threads:
Assembly of 16S rRNA genes:
Start