QmihR: Pipeline for Quantification of Microbiome in Human RNA-seq

Bruno Cavadas[1,2,3], Joana Ferreira[1,2], Rui Camacho[4,5], Nuno A. Fonseca[6], and Luisa Pereira[1,2,7(✉)]

[1] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal
luisap@ipatimup.pt
[2] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal
[3] Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Porto, Portugal
[4] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[5] LIAAD/INESC TEC, Porto, Portugal
[6] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK
[7] Faculdade de Medicina da Universidade do Porto, Porto, Portugal

Abstract. The huge amount of genomic and transcriptomic data obtained to characterize human diversity can also be exploited to indirectly gather information on the human microbiome. Here we present the pipeline QmihR designed to identify and quantify the abundance of known microbiome communities and to search for new/rare pathogenic species in RNA-seq datasets. We applied QmihR to 36 RNA-seq tumor tissue samples from Ukrainian gastric carcinoma patients available in TCGA, in order to characterize their microbiome and check for efficiency of the pipeline. The microbes present in the samples were in accordance to published data in other European datasets, and the independent BLAST evaluation of microbiome-aligned reads confirmed that the assigned species presented the highest BLAST match-hits. QmihR is available at GitHub (https://github.com/Pereira-lab/QmihR).

Keywords: Microbiome · RNA-seq data · Identification · Quantification

1 Introduction

A mutualist symbiotic relationship between microbes and their animal hosts has been estimated to occur for at least the last 500 million years [1]. A big impulse on our knowledge on the ‘normal’ human microbiome is being contributed by large scale studies such as the Human Microbiome Project (HMP) [2] and MetaHIT [3]. Major findings of HMP [4] indicated an overall high diversity of community members, heterogeneous in terms of within host versus between host ratio diversities, and ethnicity was amongst one of the strongest associations with microbiome. An intact microbial community is essential for a healthy development of the host [5], and several changes

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_21

to the microbiome are beginning to be described as associated with complex diseases, such as cancer [6, 7].

Initially, most studies of microbial communities depended on the sequencing of the gene coding the bacterial and archaea 16s rRNA, but the paradigm shift in sequencing technologies is also changing this analyses. Efforts have been applied to complete sequence the microbiome directly [8], and the huge amount of human-focused omics data (for e.g., international consortia such The Cancer Genome Atlas (TCGA) [9] and Genotype-Tissue Expression project (GTEx) [10]) has the potential to indirectly contribute information on the human microbiome [11]. In fact, it has been already shown [11] that human whole genome/exome (WGS/WES) and transcriptome sequences (RNA-seq) contain human-unmapped reads that match bacteria, viruses and fungi that colonize/infect the individuals. However, a technical challenge is that a large number of short reads cannot be uniquely mapped to a specific location at one genome, mapping instead to multiple locations at one or related genomes, influencing the bacterial abundance classification. This issue must be taken into account in the development of efficient pipelines, which can incorporate probabilistic methods that attribute these reads to the most abundant species already identified through unique-location mapping reads (such as RSEM [12]).

In this paper, we describe a pipeline to characterize the microbiome inferred from human-focused RNA-seq data, designed to perform a reliable classification of bacterial abundance. We assess its efficiency through a real TCGA RNA-seq dataset collected in 36 Ukrainian patients from gastric carcinoma. This dataset was selected as it can be compared with published information of the gut microbiome in European individuals, inferred from traditional techniques of 16s rRNA sequencing [13].

2 Description of the Pipeline

We designed a pipeline (Fig. 1) aiming to best characterize known microbiome communities, despite also allowing to collect reads that can be processed in BLAST for identification of new/uncommon pathogenic species. Currently, the most common microbiome species occurring in various human habitats are well characterize, rendering more efficient to design pipelines that search first for a reference panel of microbial species, and allow identification of the subset of unmapped non-human reads. HMP is a good departing database to construct these reference panels per location in the human body.

QmihR begins by trimming of reads using Trimmomatic [14]. It checks if: (1) the mean of two consecutive bases is below 20 Phred; and (2) the resulting read is smaller than 40 bases. This pre-processing step removes adapters and low quality reads, following the best practices for accurate RNA-seq expression estimates [15]. Even when using the pipeline in already indexed non-human mapped reads, we advise to perform this trimming as in our experience there are still low-quality reads classified as unmapped.

Then the global alignment of the reads against the bacterial reference database is made with Bowtie2 [16] and quantification of bacterial genera is performed through RSEM [17]. This tool takes a probabilistic approach to the quantification of reads in

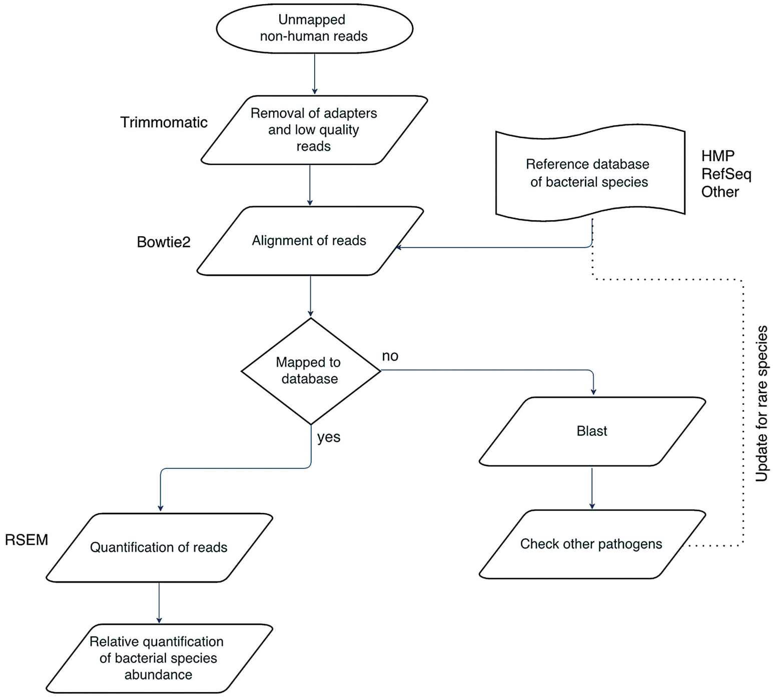

Fig. 1. Scheme of the QmihR pipeline.

cases of multi-mapping, and avoids discarding all reads that would multi-map in diverse species, conducting to a more real solution. A previous publication [18] has shown that RSEM presents the higher accuracy amongst probabilistic algorithms, guiding our choice. RSEM produces as output counts of mapped reads per gene belonging to a species (giving an indication of the most expressed genes). The pipeline takes the counts of the various genes within a species and aggregates them to produce counts of reads aligned per species, which are then normalized by the library size for the mapped reads against the bacterial reference database, as indicated in the Eq. (1).

$$normalizedcounts = \frac{counts\ gene}{\sum all\ reads\ mapped\ to\ database} \times 10^6 \tag{1}$$

3 Application to a TCGA RNA-seq Dataset

The original human-unmapped raw RNA-seq reads obtained in tumor tissue from 36 Ukrainians patients of gastric carcinomas were found in the TCGA Genomic Data Commons repository (https://gdc.cancer.gov/). The microbe reference panel used contains 194 bacterial whole genomes (one representative strain per species) collected from NCBI following the species identified by the HMP [2] in the gastrointestinal tract.

QmihR reported that the microbiome in the cohort (Fig. 2) is dominated by the genera *Bacillus* and *Pseudomonas* (around 21% and 17%, respectively), then *Escherichia* and *Enterobacter* (10–15%). The class I carcinogen *Helicobacter* reaches 3% overall frequency in Ukraine. This microbiome diversity is in accordance to published data in other European cohorts [3].

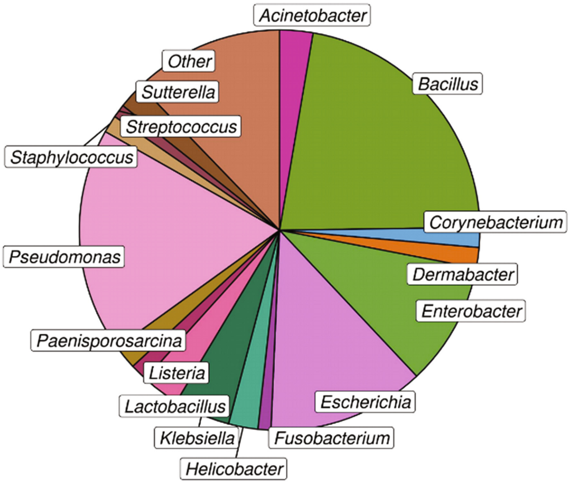

Fig. 2. Overall microbiome abundance in gastric tumor samples from Ukraine (n = 36). Only genera that passed a threshold of 1% of mean abundance are displayed in the graph, otherwise they are summed together in a class denominated as "other".

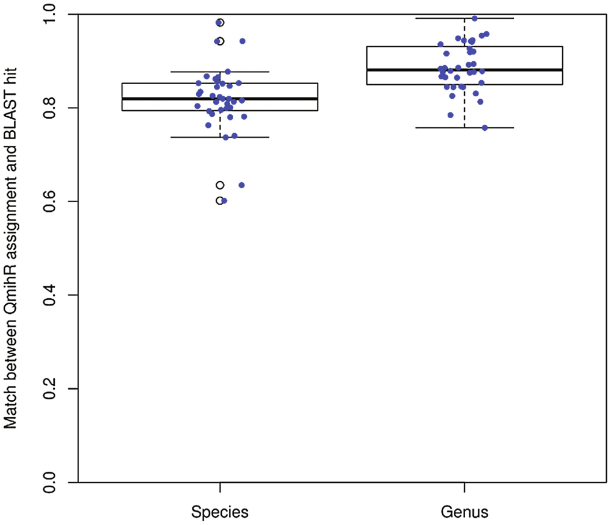

Fig. 3. Comparison of hit-species/genera matches between QmihR and BLAST for all microbe-aligned reads in the 36 gastric tumor samples from Ukraine.

In order to double-check the assignment of microbe species, we run the total amount of QmihR-assigned reads in BLAST (database downloaded on 3th February 2017, and curated for excluding sequences from uncultured species). In the Ukrainian dataset (Fig. 3), in around 82% of the reads the species identified in QmihR would also be on the list of top hits provided by BLAST, and the value raises to 88% when limiting to the genus level. We also took a closer look into the two samples with poorer results, and confirmed in BLAST that some read-pairs would align with an identity of 97–100% in the forward and 93–100% identity in the reverse in the QmihR-assigned species.

4 Benchmarking

QmihR took in average 30 min per sample to calculate the microbiome abundance, based on the reference microbe panel provided (mean 14 Gb of raw un-mapped reads in fastq format), when using an Intel Core i7-4700 2.4 GHz with 8 cores and 16 Gb of RAM. It is a fast and efficient tool that may be used in human microbiome inference from RNA-seq, in health and disease conditions.

To run the full set of unmapped reads in BLAST tool would take weeks. Even the test of running the QmihR-mapped reads in bacteria took between 2 and 8 h per sample

5 Conclusions

QmihR is a fast and efficient tool that may be used in human microbiome inference from RNA-seq, both in health and disease conditions. To our best knowledge, this is the first pipeline for quantification of the microbiome (bacterial) from RNA-seq data. A similar pipeline was developed to infer viral infection in RNA-seq TCGA samples [19], a case-study that presents, nevertheless, some differences to the situation analyzed here. Viral genomes are smaller than bacterial ones and the genes detected in the RNA-seq are the ones important for the infection and display low homology between species. In the bacteria, the reads detected in RNA-seq are mostly from rRNA genes (higher than 90%; similarly to the human genes), which display certain similarity between species, generating the multi-location read problem.

Acknowledgements. We wish to thank TCGA for the access provided to the protected data used in this work. Funds were guaranteed by the project "Advancing cancer research: from basic knowledge to application"; NORTE-01-0145-FEDER-000029; "Projetos Estruturados de I&D&I", funded by Norte 2020 – Programa Operacional Regional do Norte. I3S is financed by FEDER - Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Competitiveness and Internationalization Operational Programme (POCI), Portugal 2020, and by Portuguese funds through FCT/Ministério da Ciência, Tecnologia e Inovação in the framework of the project "Institute for Research and Innovation in Health Sciences" (POCI-01-0145-FEDER-007274).

References

1. Cho, I., Blaser, M.J.: The human microbiome: at the interface of health and disease. Nat. Rev. Genet. **13**, 260–270 (2012)
2. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., Gordon, J.I.: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature **449**, 804–810 (2007)
3. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., Wang, J.: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**, 59–65 (2010)
4. Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012)
5. Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., Gordon, J.I.: Host-bacterial mutualism in the human intestine. Science **307**, 1915–1920 (2005)
6. Thomas, R.M., Jobin, C.: The microbiome and cancer: is the ‘Oncobiome’ mirage real? Trends Cancer **1**, 24–35 (2015)
7. Brawner, K.M., Morrow, C.D., Smith, P.D.: Gastric microbiome and gastric cancer. Cancer J. **20**, 211–216 (2014). (Sudbury, Mass)
8. Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S.A., Joossens, M., Cenit, M.C., Deelen, P., Swertz, M.A., Weersma, R.K., Feskens, E.J., Netea, M.G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y.S., Huttenhower, C., Raes, J., Hofker, M.H., Xavier, R.J., Wijmenga, C., Fu, J.: Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science **352**, 565–569 (2016)
9. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Cancer Genome Atlas Research Network.: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10), 1113–1120 (2013)
10. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B.: The genotype-tissue expression (GTEx) project. Nat. Genet. **45**(6), 580–585 (2013)
11. Samuels, D.C., Han, L., Li, J., Quanghu, S., Clark, T.A., Shyr, Y., Guo, Y.: Finding the lost treasures in exome sequencing data. Trends Genet. **29**, 593–599 (2013)
12. Chandramohan, R., Wu, P.Y., Phan, J.H., Wang, M.D.: Benchmarking RNA-seq quantification tools. In: 35th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC), pp. 647–650 (2013)
13. Dicksved, J., Lindberg, M., Rosenquist, M., Enroth, H., Jansson, J.K., Engstrand, L.: Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. J. Med. Microbiol. **58**(4), 509–516 (2009)
14. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120 (2014)
15. Williams, C.R., Baccarella, A., Parrish, J.Z., Kim, C.C.: Trimming of sequence reads alters RNA-seq gene expression estimates. BMC Bioinform. **17**, 103 (2016)

16. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat Meth **9**, 357–359 (2012)
17. Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinform. **12**, 323 (2011)
18. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**(5), 525–527 (2016)
19. Tang, K.W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., Larsson, E.: The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat. Commun. **4**, 2513 (2013)