Quantitative metagenomics: form reads to biomarkers

Edi Prifti*[1], **Emmanuelle Le Chatelier***[1], **Nicolas Pons***[1], Mathieu Almeida[2], Amine Ghozlane[1], Florian Plaza[1], Ndeye Aram Gaye[1], Pierre Leonard[1], Jean-Michel Batto[1] and Dusko Erlich[1]

[1] INRA, Institut National de la Recherche Agronomique, US1367 Metagenopolis, 78350 Jouy-en-Josas, France

[2] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA.

*** These authors contributed equally to the work**

Introduction. The study of complex microbial ecosystems has been increasingly improved with the advent of metagenomics[1,2]. Human microbiome and in particular human gut microbiota has been increasingly investigated over the last few years. This highly diverse ecosystem whose collective genome exceeds 100-fold the size of the human genome[3], provides the host with vital functions[4] resulting from a long co-evolution of prokaryotes and eukaryotes. The role of gut microbiota in human health and disease has received unprecedented attention over the past few years[5] and several complex chronic diseases, such as obesity[6], inflammatory bowel disease[7], liver cirrhosis[8], type-I[9], and type-2 diabetes[10,11] have been associated with gut microbiota.

Quantitative metagenomics (QM), where whole DNA from a given ecosystem is extracted and sequenced, as opposed to targeted metagenomics (16S), where only targeted sequences are amplified and sequenced, allows measuring accurately the presence and abundance of all DNA sequences and generates large amount of data. Sequenced reads are mapped and counted onto ecosystem-representative reference gene sets or individual genomes and result in very big sparse matrices with millions of variables. Even though the number of metagenomics data-mining tools is growing[12,13] many issues concerning data processing and statistical analyses are still to be tackled. Comparative and validation studies are also needed to remove general confusion on which are the right tools to use. For instance the usefulness of many tools developed for the analyses of 16S rRNA is extrapolated without such formal proof for QM data, which have other properties.

Methods. Here we discuss our experience with different data processing techniques and analytical approaches that have been proposed or adapted to explore QM data in identifying gut microbial biomarkers associated with complex diseases. Pre-processing of count matrices is a crucial step that alters irreversibly the data for downstream analyses. The main purpose being to remove technical variability and noise such as that due to variation in sequencing depth from one sample to another and make gene and sample profiles comparable among each other. Normalization, rarefaction and filtering are some of the techniques that we have been using and which yield satisfying results. Unfortunately well-formalized normalisation techniques that work well in transcriptomics are not adapted for QM data and other specific

methods are still to be formalized and tested. Another important topic in QM is the dimension reduction as for instance the concept of metagenomic species (MGS) which was developed in the lab based on the co-abundance clustering of the gene profiles (Nielsen et al, Nature Botech, in press). This technique allows reducing more than 1000-fold the complexity of the dataset and applying more powerful statistics.

Applications. We applied these different processing and analytical methods, implemented in a suite of tools, Meteor Studio, MetaOMineR and Metaprof, to real gut microbiome data in a number of different studies[14,15]. We compared for instance the gut microbiome of the liver cirrhosis patients with that of healthy controls and the associated microbiome signal was very strong. Will only a small number of identified species we were able to discriminate very accurately patients from controls (Qin et al, Nature, in press).

Keywords. quantitative metagenomics, pre-processing, biomarkers, gut microbiome

References

1 Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annual review of genetics* **38**, 525-552, doi:10.1146/annurev.genet.38.072902.091216 (2004).

2 Sleator, R. D., Shortall, C. & Hill, C. Metagenomics. *Letters in applied microbiology* **47**, 361-366, doi:10.1111/j.1472-765X.2008.02444.x (2008).

3 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

4 Nicholson, J. K. *et al.* Host-gut microbiota metabolic interactions. *Science* **336**, 1262-1267, doi:10.1126/science.1223813 (2012).

5 Walsh, C. J., Guinane, C. M., O'Toole, P. W. & Cotter, P. D. Beneficial modulation of the gut microbiota. *FEBS letters*, doi:10.1016/j.febslet.2014.03.035 (2014).

6 Clarke, S. F. *et al.* The gut microbiota and its relationship to diet and obesity: new insights. *Gut microbes* **3**, 186-202, doi:10.4161/gmic.20168 (2012).

7 Elson, C. O. & Cong, Y. Host-microbiota interactions in inflammatory bowel disease. *Gut microbes* **3**, 332-344, doi:10.4161/gmic.20228 (2012).

8 Bajaj, J. S. *et al.* Altered profile of human gut microbiome is associated with cirrhosis and its complications. *Journal of hepatology*, doi:10.1016/j.jhep.2013.12.019 (2013).

9 Wen, L. *et al.* Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **455**, 1109-1113, doi:10.1038/nature07336 (2008).

10 Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99-103, doi:10.1038/nature12198 (2013).

11 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).

12 Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PloS one* **7**, e47656, doi:10.1371/journal.pone.0047656 (2012).

13 Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology* **14**, R2, doi:10.1186/gb-2013-14-1-r2 (2013).

14 Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585-588, doi:10.1038/nature12480 (2013).

15 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).