OPEN

The ISME Journal (2016) 10, 1025–1028

www.nature.com/ismej

COMMENTARY

Towards biome-specific analysis of meta-omics data

Youssef Darzi[1,2,3], Gwen Falony[2,3], Sara Vieira-Silva[2,3] and Jeroen Raes[1,2,3]

The ISME Journal (2016) **10,** 1025–1028; doi:10.1038/ismej.2015.188; published online 1 December 2015

Microbial ecology has witnessed tremendous progress over the last decade empowered by meta-omics approaches and innovations in DNA/RNA sequencing as well as high-resolution mass spectrometry. In this climate, the rise of meta-omics projects (Raes, 2011) such as MetaHIT and the Human Microbiome Project, Tara Oceans, the Global Ocean Sampling Expedition and the Earth Microbiome Project aiming at unraveling the structure and function of specific microbiomes in different habitats was observed. Now that massive data generation is no longer science fiction, the bottleneck shifts to computational analysis (Falony *et al.*, 2015).

On the bioinformatics front, important efforts have already gone into the 'upstream' part of the analysis. Traditional sequence mapping, assembly, binning and clustering approaches have been scaled up for the handling of hundreds of gigabytes of sequencing data (Kim *et al.*, 2013), and the generation of biome-wide gene catalogs greatly facilitates the analysis. Also, machine learning techniques have been successfully applied on several occasions to predict disease biomarkers from meta-omics data (for example, Williams *et al.* (2009) and Zeller *et al.* (2014)). However, for metabolic pathway-based functional analysis, researchers usually still rely on 'classic' approaches developed for single genomes. To identify and quantify the biochemical functions and pathways that make up the metabolic wiring of an ecosystem and assess functional shifts upon perturbation, associations between environment, metabolism and species–function relationships, current studies usually rely on broad metabolic databases (for example, KEGG (Kanehisa *et al.*, 2014)). Despite their unquestionable merit, such resources unfortunately tend to be biased, for historical reasons, toward Eukaryotes and model organisms' metabolism. These databases thus often include pathways and pathway variants that do not exist in many ecosystems under research, or lack part of its enzymatic routes, which can be misleading when drawing conclusions and penalizing for statistical significance in large-scale studies. As a case in point, no specific pathway module for the production of butyrate can be found in the KEGG encyclopedia, despite its significant clinical importance for the gut ecosystem. Thus, using 'universal' databases often results in suboptimal functional assignment and fewer or false-positive outcomes.

Recently, novel approaches are being pursued to improve the sensitivity and specificity of functional interpretation of meta-omics data using biome-specific approaches. For instance, Le Chatelier *et al.* (2013) investigated specific metabolic shifts in the gut metagenome of obese individuals based upon inspection of 51 manually compiled gut-specific pathway modules and Sunagawa *et al.* (2015) studied ocean biochemical processes using a targeted set of markers for essential ocean biogeochemical processes. Likewise, Prestat *et al.* (2014) used manually curated HMM profiles targeted for soil biochemical pathways to improve the accuracy and the rate of functional annotation of soil metagenomic samples. However, such biome-specific approaches are more exceptions rather than the rule.

Here, we illustrate the advantages of using biome-specific approaches in an example comparative analysis of a human gut metaproteomics data set (10 samples: 4 are healthy individuals and 6 Crohn's Disease (CD) patients in remission) from Erickson *et al.* (2012) by comparing the outcome of a standard (KEGG-based) analysis versus GOmixer (Raes Lab, Ghent, Belgium), a human gut-specific metabolic pathway analysis tool that we developed for this purpose (available as an online tool and downloadable software package at: http://www.raeslab.org/gomixer/).

In short, the GOmixer workflow starts by quantifying human gut metabolic pathway modules for each sample, by mapping gene abundances on a database of predefined gut-specific modules. A module is a set of tightly related enzymatic functions that represent a cellular process with defined input and output metabolites. The modules used in GOmixer's database were manually compiled based on extensive literature searches (Le Chatelier *et al.*, (2013); Vieira-Silva *et al.*, unpublished). For a module to be considered present by GOmixer, its coverage (percentage of metabolic steps present) should be higher than a user specified threshold. Module abundance is defined as the average abundance of the metabolic steps covered for this pathway. After quantification, statistically over/under-represented metabolic modules between two groups of samples, in this case Healthy and CD patients are determined using the nonparametric Wilcoxon's rank-sum test, given that metagenomics data are generally distribution free and that the test is robust to outliers. Benjamini–Hochberg's false

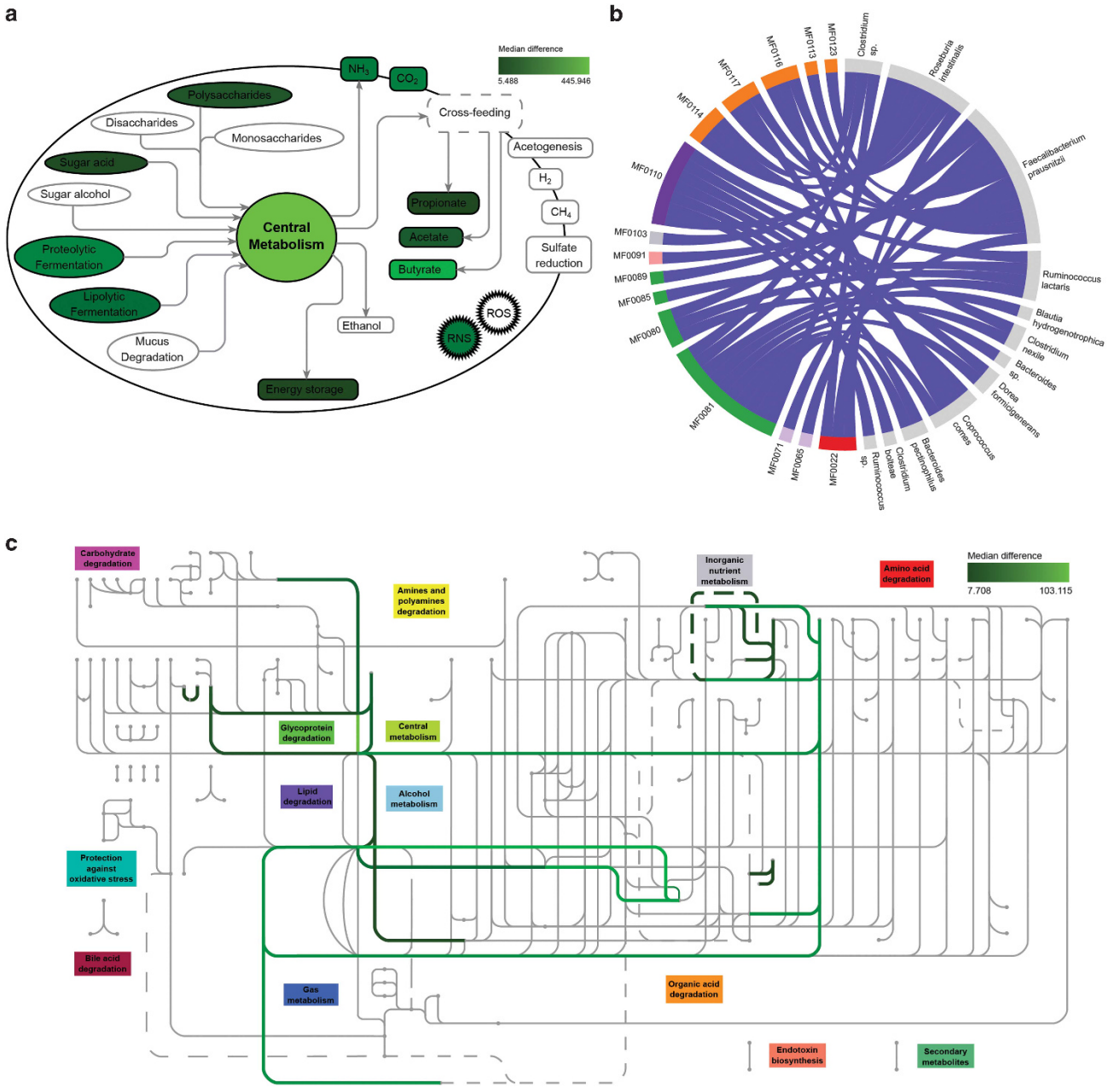

Figure 1 GOmixer analysis outcome of Erickson *et al.* (2012). (**a**) Global gut metabolic processes map that gives an overview of the major metabolic processes in the gut. The color scale reflects significantly enriched abundances in Healthy compared to Crohn's Disease (CD) subjects. (**b**) Chord plot highlighting species-function associations. Modules (MF numbers) belonging to the same global metabolic process share the same color. Association links reflect module over/under-representation in Healthy (blue) or CD (red). In this analysis all functions are over represented in Healthy. (**c**) Gut module map consisting of modules connected by their input and output compounds. They are clustered according to their hierarchical classification (for example, amino-acid degradation) and reflect the flow of compounds from top to bottom. The color scale reflects significantly enriched abundances in Healthy compared with CD subjects.

discovery rate is then used to correct for multiple testing. The results are displayed on a gut-specific global metabolic map to easily highlight trends in functionally related pathways (Figure 1).

Table 1 shows the comparison between both approaches. The results show that agreement between universal and gut-specific analyses can be found on multiple occasions. For instance, both analyses show differential expression of the Glycolysis and the Entner–Doudoroff pathways. However, several modules not relevant to the context of the human gut were also detected as significantly different using universal module-based analysis. For example, the Crassulacean acid metabolism module (M00169), which is a carbon fixation pathway in plants and M00344 (formaldehyde assimilation, xylulose monophosphate pathway), which is specific to yeast but not to prokaryotes, are both found to be down-regulated in CD patients. The reasons for these observations is the existence

Table 1 Comparison between KEGG- and GOmixer-based metaproteomics analysis of Crohn's patients vs controls

KEGG	*Gut metabolic modules*
M00001: Glycolysis (Embden–Meyerhof pathway), glucose = > pyruvate	MF0080: Glycolysis (preparatory phase)
M00002: Glycolysis, core module involving three-carbon compounds	MF0081: Glycolysis (pay-off phase)
M00008: Entner–Doudoroff pathway, glucose-6P = > glyceraldehyde-3P, pyruvate	MF0089: Entner-Doudoroff pathway I
M00003: Gluconeogenesis, oxaloacetate = > fructose-6P	MF0022: isoleucine degradation
M00061: Uronic acid metabolism	MF0065: pectin degradation-5-dehydro-4-deoxy-D-glucuronate degradation
M00166: Reductive pentose phosphate cycle, RuBP, CO2 = > glyceraldehyde-3P	MF0071: D-galacturonate degradation
M00170: C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type	MF0085: pyruvate:formate lyase
M00171: C4-dicarboxylic acid cycle, NAD, -malic enzyme type	MF0091: beta-D-glucuronide and D-glucuronate degradation
M00173: Reductive citric acid cycle (Arnon–Buchanan cycle)	MF0103: nitrate reduction (assimilatory)
M00178: Ribosome, bacteria	MF0110: glyoxylate bypass
M00183: RNA polymerase, bacteria	MF0113: acetyl-CoA to acetate
M00194: Maltose/maltodextrin transport system	MF0114: acetyl-CoA to crotonyl-CoA
M00197: Putative sugar transport system	MF0116: butyrate production via transferase
M00214: Methyl-galactoside transport system	MF0117: butyrate production via kinase
M00308: Semi-phosphorylative Entner–Doudoroff pathway, gluconate = > glyceraldehyde-3P, pyruvate	MF0123: propionate production (propanediol pathway)
M00345: Formaldehyde assimilation, ribulose monophosphate pathway	
M00357: Methanogenesis, acetate = > methane	
M00169: CAM (Crassulacean acid metabolism), light	
M00344: Formaldehyde assimilation, xylulose monophosphate pathway	

Modules highlighted in green show common modules to both analyses. Modules highlighted in red are non-gut modules that were recruited by the KEGG-based analysis only.

of enzymes found in more than one metabolic module, causing enzymes of truly present modules to sometimes yield artifact overrepresentation of other modules as well (Ye and Doak, 2009). Besides increasing the false-positive rate, these uninformative modules inflate the number of statistical comparisons to be done and thus penalize true signals when correcting for multiple testing. We explored whether approaches that aim at reducing false-positive rate by finding a minimum number of pathways that can explain all genes observed in a given metagenome can resolve these issues and reveal only gut-relevant pathways. For this reason, we reanalyzed the data using MinPath (Ye and Doak, 2009) and the KEGG database. Although this clearly improved results, non-relevant pathways (for example, the Crassulacean acid metabolism module) were still recovered (Supplementary Table S1). Moreover, when applying this approach on 1267 (human-filtered) metagenomes (healthy, diabetes and CD) used for the construction of the Integrated Gene Catalog (IGC) (meta.genomics.cn/metagene/meta/home) of the human gut microbiome, several non-relevant plant (for example, M00085: Fatty acid biosynthesis, elongation, mitochondria and M00114: Ascorbate biosynthesis, plants, glucose-6P = > ascorbate), human (for example, M00042: Catecholaminebiosynthesis and M00135: GABA biosynthesis, eukaryotes, putrescine = > GABA) and bacterial modules (for example, M00165: Reductive pentose phosphate cycle (Calvin cycle), M00376: 3-Hydroxypropionate bi-cycle) were recruited (see Supplementary Table S2 for details). However, the bigger problem lies not in overpredicting but in missing relevant modules for the collection of a complete overview of the metabolic network at hand. As an example, the GOmixer-based analysis also uncovered specific gut fermentation modules, which highlighted significant downregulation of proteolytic and lipolytic fermentation, polysaccharide degradation, and the production of short-chain fatty acids in CD patients. These metabolic processes are essential for the functioning of the human gut microbiota (Gerritsen *et al.*, 2011) and very relevant for understanding pathomechanisms; however, no detailed module level definition is available for them in the KEGG database.

Overall, this case study illustrates that using universal, generic approaches in microbiome studies are not without risk, and shows that using biome-specific databases provides substantial advantages for ecological and clinical analysis of meta-omics data sets for specialists and non-specialists alike. The reason for this is threefold: First, the specific focus of biome-specific modules allows careful hypothesis generation and tangible data analysis for non-specialists. Second, it allows moving beyond coarse-grained functional assignment to fine-grained module level assignment, which is crucial to associate bacterial species to specific metabolic roles. Third, modules with well-defined input and

output compounds are also important to integrate multiple types of omics (for example, integrating metabolomics data) and to model the ecosystem responses to perturbation or predict future behavior in longitudinal studies. This in turn will help us understand the fundamentals of the microbial ecology of a wide range of ecosystems, help us design better diagnostics (both in clinical as environmental applications), and better targeted therapies/interventions. Finally, at the time where meta-omic data generation is becoming available to all, but the analysis and hypothesis generation remains complex, biome-specific, user-friendly tools such as GOmixer will contribute to the reduction of this complexity and help drive omics-based microbial ecology forward.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank all members of the Raes lab but especially Samuel Chaffron and Gipsi Lima-Mendez for helpful discussions and constructive comments. This work is supported by the Fund for Scientific Research—Flanders, FP7 METACARDIS HEALTH-F4-2012- 305312, KU Leuven, the Rega institute and IWT. SVS is supported by Marie Curie Actions FP7 People COFUND—Proposal 267139.

Y Darzi and J Raes are at Microbiology Unit, Faculty of Sciences and Bioengineering Sciences, Vrije Universiteit Brussel, Brussels, Belgium
Y Darzi, Gwen Falony, Sara Vieira-Silva and J Raes are at VIB, Center for the Biology of Disease, Leuven, Belgium
Y Darzi, Gwen Falony, Sara Vieira-Silva and J Raes are at Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
E-mail: jeroen.raes@med.kuleuven.be

References

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G *et al.* (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**: 541–546.

Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C *et al.* (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's Disease. *PLoS One* **7**: e49138.

Falony G, Vieira-Silva S, Raes J. (2015). Microbiology Meets Big Data: the case of gut microbiota–derived trimethylamine. *Annu Rev Microbiol* **69**: 305–321.

Gerritsen J, Smidt H, Rijkers GT, de Vos WM. (2011). Intestinal microbiota in human health and disease: the impact of probiotics. *Genes Nutr* **6**: 209–240.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res* **42**: D199–D205.

Kim M, Lee K-H, Yoon S-W, Kim B-S, Chun J, Yi H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform* **11**: 102–113.

Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J *et al.* (2014). FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res* **42**: e145–e145.

Raes J. (2011). Why meta-omics should be mega-omics: on experimental design and multiple testing hell. *Environ Microbiol Rep* **3**: 19–20.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Structure and function of the global ocean microbiome. *Science* **348**: 1261359.

Williams HRT, Cox IJ, Walker DG, North B V, Patel VM, Marshall SE *et al.* (2009). Characterization of inflammatory bowel disease with urinary metabolic profiling. *Am J Gastroenterol* **104**: 1435–1444.

Ye Y, Doak TG. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* **5**: e1000465.

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI *et al.* (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**: 766.

Supplementary Information accompanies this paper on *The ISME Journal* website (http://www.nature.com/ismej)