Bioinformatics, 32(11), 2016, 1678–1685
doi: 10.1093/bioinformatics/btv526
Advance Access Publication Date: 5 September 2015
Original Paper

OXFORD

Gene expression

metaModules identifies key functional subnetworks in microbiome-related disease

Ali May[1,2,3,*], Bernd W. Brandt[2], Mohammed El-Kebir[1,4,5], Gunnar W. Klau[1,3,5], Egija Zaura[2], Wim Crielaard[2], Jaap Heringa[1,3,†] and Sanne Abeln[1,3,†,*]

[1]Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam, Amsterdam, The Netherlands, [2]Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam, Amsterdam, The Netherlands, [3]Amsterdam Institute for Molecules Medicines and Systems (AIMMS), VU University Amsterdam, Amsterdam, The Netherlands, [4]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, USA and [5]Life Sciences, Centre for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
Associate Editor: Gunnar Ratsch

Received on April 20, 2015; revised on July 25, 2015; accepted on September 2, 2015

Abstract

Motivation: The human microbiome plays a key role in health and disease. Thanks to comparative metatranscriptomics, the cellular functions that are deregulated by the microbiome in disease can now be computationally explored. Unlike gene-centric approaches, pathway-based methods provide a systemic view of such functions; however, they typically consider each pathway in isolation and in its entirety. They can therefore overlook the key differences that (i) span multiple pathways, (ii) contain bidirectionally deregulated components, (iii) are confined to a pathway region. To capture these properties, computational methods that reach beyond the scope of predefined pathways are needed.

Results: By integrating an existing module discovery algorithm into comparative metatranscriptomic analysis, we developed metaModules, a novel computational framework for automated identification of the key functional differences between health- and disease-associated communities. Using this framework, we recovered significantly deregulated subnetworks that were indeed recognized to be involved in two well-studied, microbiome-mediated oral diseases, such as butanoate production in periodontal disease and metabolism of sugar alcohols in dental caries. More importantly, our results indicate that our method can be used for hypothesis generation based on automated discovery of novel, disease-related functional subnetworks, which would otherwise require extensive and laborious manual assessment.

Availability and implementation: metaModules is available at https://bitbucket.org/alimay/metamodules/

Contact: a.may@vu.nl or s.abeln@vu.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The human microbiome is a functional community that plays an important role in health and disease (Huttenhower *et al.*, 2012; Mitra *et al.*, 2013; Morgan *et al.*, 2012). Disorders that have been previously linked to, for instance, the gut microbiome, include cardiovascular disease (Wang *et al.*, 2011) and type 2 diabetes (Qin *et al.*, 2012). Because distinct populations of microbes from different patients can harbor remarkably similar functional capabilities (Turnbaugh *et al.*, 2009), the pathogeneses of these diseases may not necessarily be related to specific microbial population shifts from the health-associated microbiome. Phenotypic differences are therefore better explained by changes in the regulation of cellular mechanisms, for instance in the form of under- or overexpression of certain metabolic pathways or parts thereof. A prerequisite for this is the ability to identify differentially expressed functional systems using computational pipelines that can operate at the systems level. The availability of such pipelines is likely to lead to increased discovery of preventive and therapeutic targets in microbiome-related diseases.

Metagenomics and metatranscriptomics are two powerful approaches for studying, respectively, the functional capability and activity of microorganisms in environmental and clinical samples. The latter provides a more direct view into disease-associated factors since microbial activity is often clinically more relevant than microbial genetic potential. Supporting this are recent observations of many metagenomically abundant gene families being consistently downregulated at the transcriptional level, for instance, in the human gut microbiome (Franzosa *et al.*, 2014).

Comparative metatranscriptomics (meta-RNA-Seq) aims at identifying the functions that are significantly differentially expressed by microbial communities under different conditions. Because the number of active microbial genes can reach up to several millions, the functional profile of a sample is commonly constructed by mapping mRNA reads first to genes, and then to higher-level functional units such as KEGG Orthology groups (KOs, Kanehisa and Goto, 2000). KOs are organism-independent identifiers that group together proteins of similar biochemical functions. As with genes, they can be used to perform differential expression analysis using packages such as edgeR (Robinson *et al.*, 2010) and DESeq2 (Love *et al.*, 2014). Since interpreting results at KO level is not straightforward, this is typically followed by pathway or gene set enrichment analysis (Subramanian *et al.*, 2005) (as in Marchetti *et al.*, 2012). Another common approach is to use bioinformatics pipelines such as MEGAN (Huson *et al.*, 2007) or HUMAnN (Abubucker *et al.*, 2012) to map KOs to predefined functional modules and pathways, such as those in the KEGG and MetaCyc (Caspi *et al.*, 2014) databases, and then to identify the deregulated functional units (as in Franzosa *et al.*, 2014).

Pathway-based approaches provide invaluable insights into disease biology; however, they have a number of shortcomings that are critical for the identification of functional markers. First, they consider each predefined pathway individually and therefore ignore the interactions between pathways that may be crucial for discovering causal biochemical mechanisms. Second, collapsing the expression levels of KOs into pathways or modules can have two undesired outcomes: non-differentially but highly expressed functions within a pathway can obfuscate significant differences of other parts of the same pathway. Moreover, pathways that contain up- and downregulated genes may be determined as non-differentially expressed overall. Finally, significantly deregulated pathway regions that are not covered by predefined module delineations cannot be identified.

Methods that address the above limitations of pathway-based methods have been previously reported. By combining protein-protein interaction networks with gene expression data, Ideker *et al.* (2002) introduced the concept of identifying connected subsets of regulatory genes that are responsible for changes in expression under different conditions. This concept has since been extended to other settings, including networks of enzymatic (Patil and Nielsen, 2005) and regulatory elements (Oliveira *et al.*, 2008). The central idea in these methods can be summarized in three main steps: (i) constructing a mathematical network (graph) that captures the core molecular interactions in biological pathways, (ii) assigning numerical scores that reflect biological activity (or response) to the components (edges/nodes) in the network and finally (iii) performing a search on the network to identify the highly scoring subnetworks that delineate the biological hotspots from the rest of the network [for a review, see Mitra *et al.* (2013)].

The current work introduces metaModules; a novel approach that focuses on comparative metatranscriptomic datasets for identifying metabolic subnetworks of significantly deregulated orthologous gene groups. An earlier work in this direction is MetaPath (Liu and Pop, 2011), a method for finding differentially abundant enzymatic subnetworks in metagenomic datasets. The biological questions addressed by our method and MetaPath are fundamentally different: whereas MetaPath is specifically designed to find subnetworks of reactions that are either enriched or depleted as a whole in a given experimental condition, metaModules focuses on finding connected subsets of significantly deregulated ortholog gene groups. Note that the subnetworks discovered by MetaPath do not have to be composed of significantly differentially abundant or expressed elements. Instead, the statistical significance of these subnetworks is evaluated using a bootstrapping strategy. On the other hand, as it is designed with a metatranscriptomic focus, our approach uses statistics that are tailored to the properties of RNA-Seq data. Following the work by Dittrich *et al.* (2008), our method specifically searches for connected subnetworks that predominantly contain significantly differentially expressed KOs, which can be either up- or downregulated in different experimental conditions. Finally, the exact search algorithm used by metaModules, Heinz (Dittrich *et al.*, 2008), guarantees to find the optimal maximum-scoring subnetwork within the input network.

We applied metaModules to previously published metatranscriptomic datasets of two well-studied, oral microbiome-related diseases: periodontal disease and dental caries. Periodontal (gum) disease originates from an unbalanced host-microbiome interaction that leads to an enhanced host inflammatory response. If left untreated, it causes the destruction of tooth-supporting structures and may eventually result in tooth loss (Pihlstrom et al., 2005). Dental caries (tooth decay) is one of the most common chronic diseases, where the main cause is a net overproduction of acids during microbial carbohydrate fermentation in the oral cavity, leading to the demineralization of dental hard tissues (Selwitz et al., 2007). Using metaModules, we recovered significantly deregulated connected subnetworks of functions that are important in a number of mechanisms previously associated with these diseases, such as the production of metabolites that inhibit human cell growth in periodontal disease. Our results indicate that using high throughput data, our methodology can be used to generate new hypotheses automatically by uncovering novel disease-related metabolic mechanisms in microbial communities. Although we demonstrated it on microbiome-related disease datasets in this study, metaModules can be used for datasets from any type of comparative metatranscriptomic studies.

2 Methods

2.1 Metatranscriptomic data

We applied metaModules to metatranscriptomic datasets from recently published studies of periodontal disease (Jorth *et al.*, 2014) and dental caries (Peterson *et al.*, 2014). The periodontal disease (PD) dataset consisted of patient-matched healthy ($n = 3$) and diseased ($n = 3$) microbiome samples, each prepared by pooling subgingival plaque from three healthy or diseased teeth. The dental caries (DC) dataset comprised of supragingival plaque samples collected from all dental surfaces of 36 individuals who had either a caries-positive ($n = 19$) or a caries-negative ($n = 17$) oral health profile.

2.2 Preprocessing and mapping of reads

All methods within our pipeline mentioned below were used with default parameters except where noted. Sickle v1.210 (available at https://github.com/najoshi/sickle) was used to trim the 50-bp (PD) and 100-bp (DC) single-end reads using an average quality score of 15 over a sliding window of 10% of the read length. Trimmed sequences shorter than 30-bp (PD) or 60-bp (DC) and those that contained more than 10% ambiguous bases were discarded. Next, human mRNA reads were filtered by Best Match Tagger v1.1.0 (available at ftp://ftp.ncbi.nih.gov/pub/agarwala/bmtagger/) using the NCBI database of human mRNA transcripts predicted by NCBI Gnomon. Subsequently, SortMeRNA v1.99-beta (Kopylova *et al.*, 2012) was used to filter the non-coding RNAs in the remaining reads. As a second step for removing non-coding RNAs, the reads were aligned using USEARCH (v7.0.1090, -usearch_local, Edgar, 2010) to a combination of sequences from the GtRNAdb tRNA database (Chan and Lowe, 2009) and the SSURef_NR99 and LSURef sequences (release 115) from the SILVA database (Quast *et al.*, 2013). An *E*-value cutoff of 1.0 and an identity threshold of 0.5 were used for the mapping, and mapped reads with bit-scores >54 were removed as recommended by Leimena *et al.* (2013). The remaining reads were aligned to a custom KEGG database (Kanehisa and Goto, 2000) of 3 002 391 proteins from 2965 organisms. Each protein in the database was a member of at least one of the 8940 KEGG Orthology groups (KO). Each KO was further categorized under one or more of the 369 KEGG pathway maps. The nucleotide-to-protein search, where the top 10 hits were kept for the query reads, was performed using the ublast command in USEARCH v7.0.1090 with an *E*-value threshold of 1000 for the PD and of 10 for the DC dataset. For the PD dataset, a high *E*-value threshold was required for mapping a sufficient number of short reads (30- to 50-bp) in each sample; however, to minimize the ambiguity of matches in this case, a read coverage of 0.8 and an identity level of 0.7 were used as additional alignment cutoffs.

2.3 Differential expression analysis

After the reads were mapped to KEGG proteins, gene counts were estimated using a 'best-hit' approach by counting the number of times a gene was mapped to a read as the best-matching target. In the case of tied hits with the same *E*-value, the count for each top-scoring gene was incremented. Note that, given the statistical framework used for differential expression analysis is suitable, more sophisticated methods such as HUMAnN (Abubucker *et al.*, 2012), which take into account the alignment quality and gene length differences, can be used to estimate the counts.

Differential expression analysis at gene level was performed using the R package DESeq2 v1.6.1 (Love *et al.*, 2014). A multi-factor model was used for the paired samples in the PD dataset, while for the DC dataset an unpaired analysis was performed by using the default processing and analysis steps described in the DESeq2 package vignette.

KO counts were estimated by summing the counts of all genes belonging to a given KO. Differential expression analysis at KO level was performed as described for genes. The KOs with very low mean of normalized counts were discarded by the independent filtering option in DESeq2. While this filtering is optional in RNA-Seq analyses that use the raw (and not the adjusted) *P*-values in their pipeline, its use is recommended in metatranscriptomic datasets where sparse counts may well lead to inaccurate results (Luo *et al.*, 2013).

2.4 Networks

We considered two KOs functionally interacting if there was a KEGG relation defined between them in a given KEGG pathway map (e.g. acting on the same compound in two successive reactions, or sharing a direct signaling relationship). We combined these interactions in a single network, removed self-loops and duplicated edges, and obtained a global network of 6642 nodes (KOs) and 35 405 undirected edges (interactions) with a large connected component (5547 nodes and 34 348 edges). The global networks for the PD and DC datasets were obtained by taking the intersection of KOs in the global KEGG network and those retained in the two datasets, and then removing the non-interacting, isolated nodes in the resulting networks, i.e. the nodes from which there is no path to another node in the network.

2.5 Identification of maximum-scoring subnetworks

Given metatranscriptomic data from healthy and diseased individuals and an interaction network of KOs, our goal is to identify functional subnetworks of significantly deregulated KOs in microbial communities. The subnetwork identification in metaModules is based on the work by Dittrich *et al.* (2008). Here, each node in the network is assigned a negative or a positive weight reflecting the statistical significance of its differential expression (*P*-value), followed by the application of a search algorithm (Heinz) that identifies the optimal maximum-scoring connected subnetworks (MSS) in the network.

As described by Pounds and Morris (2003), a *P*-value distribution of differential expression data can be regarded as a mixture of two distributions that model a signal and a noise component: *P*-values obtained from the background noise alone will follow a uniform distribution. On the other hand, the signal component, which arises from true differences in expression, follows a Beta(a, 1) distribution. To derive node weights from *P*-values, we first fitted a beta-uniform mixture model to the distribution of *P*-values as described in Beisser *et al.* (2010). Next, the mixture (λ) and shape (a) parameters of this model, the list of *P*-values and a false-discovery rate (FDR) were given as input to the Heinz package to derive the node (KO) weights. The adjusted node score (weight) S of a *P*-value x is given by:

$$S(x) = (a-1)(\log(x) - \log(\tau(\mathrm{FDR}, \lambda, a))) \quad (1)$$

where $\tau(\mathrm{FDR}, \lambda, a)$ is the significance threshold corresponding to the specified FDR and the fitted values of λ and a, as described in Pounds and Morris (2003). The KOs with *P*-values lower than this threshold are regarded as part of the signal and are assigned positive weights, while KOs with higher *P*-values are assumed to constitute the noise and are given negative weights. Note that the directionality of deregulation does not play a role in the calculation of the weights. In other words, the resulting MSSs do not have to be dominantly up- or downregulated in a certain phenotype. However, if needed, a

one-sided test can be used in the differential expression analysis step to specifically search for subnetworks that are up- or downregulated as a whole.

After the annotation of the KO network with weights, Heinz was used to identify the MSS in each dataset. The networks were visualized using the eXamine plugin (Dinkla *et al.*, 2014) in Cytoscape (Shannon *et al.*, 2003).

2.6 Contribution of species to deregulated KOs

We determined the change in the contribution of individual organisms to the identified subnetworks in health and disease (Supplementary Fig. S1). Performing differential expression analysis at both gene and KO level allowed us to obtain the mean expression levels of genes and KOs in the healthy and diseased patient groups. Note that the genes in our approach were annotated not only with KOs but also with species using KEGG. Therefore, we were able to determine the contribution of a species in terms of its gene expression to the KOs in the MSSs in different phenotypes. In our approach, each species s consists of a set of active genes $G_s = \{g_{s,1}, g_{s,2}, \ldots\}$. Similarly, each KO i in an MSS contains one or more genes from a number of organisms $G_i = \{g_{i,1}, g_{i,2}, \ldots\}$. The log fold change in the expression of KO i for species s was then determined by:

$$\mathrm{FC}_{s,i} = \log_2\left(\frac{\sum_{g\in G_i\cap G_s} g_d}{\sum_{g\in G_i\cap G_s} g_h}\right) \quad (2)$$

where g_d and g_h are the mean expression values of gene g across diseased and healthy patient groups, respectively. Subsequently, for each dataset, we identified the 10 species that showed the largest up- and downregulation in disease across all KOs in the MSSs.

2.7 Comparison with other methods

We compared our method with MetaPath (Liu and Pop, 2011), the only method currently available for identifying differentially abundant MSSs in metagenomic datasets. We ran MetaPath with default parameters to identify MSSs in the DC dataset (the PD dataset was omitted since MetaPath is not applicable to paired samples). Among the significant MSSs reported by the algorithm for the health- and disease-associated phenotypes, we considered only the top-scoring MSS for each phenotype. These enzymatic reaction subnetworks were mapped back to a subnetwork of KOs, where the KO interactions were determined using the global reaction network and the reaction-to-KO mapping file provided with the MetaPath.

The topology of the global reaction network used by MetaPath differs from the global KO network used in metaModules. To enable a consistent comparison, we transformed MetaPath's directed network of reactions into an undirected network of KOs. We ran metaModules using an FDR = 0.15 on this network to obtain an MSS of similar size to the top-scoring MSSs obtained by MetaPath.

3 Results and discussion

3.1 Processing and mapping of reads

In total, we processed nearly two billion metatranscriptomic reads from a periodontal disease (PD, Jorth et al., 2014) and a dental caries (DC, Peterson et al., 2014) dataset. Both sets consisted of microbiome RNA samples taken from healthy and diseased conditions. The reads were processed through three filtering steps: (i) quality control, (ii) host mRNA filtering and (iii) removal of non-coding RNAs. In the PD dataset, 53% of the reads were removed after the quality control due to the large number of very short reads (15-bp) that mainly comprised two of the three healthy samples (Table 1). The percentage of non-coding RNAs in the remaining reads was remarkably low (12%). On the contrary, the vast majority of reads (93%) in the DC dataset passed the quality control; however, almost 85% of the reads were discarded after the removal of non-coding RNAs, indicating substantial non-coding RNA contamination in the samples. Both of the datasets contained similarly low levels of host mRNA contamination (<2%). Both datasets had similar levels of reads mapped to KEGG genes (<6%). The low percentage of mapped reads is possibly due to the fact that our database contained only the genes that were characterized in KEGG pathway maps. Gene and KO counts were estimated as described in Section 2.3.

3.2 From the significance of deregulation to networks

Our goal was to identify significantly differentially expressed maximum-scoring connected subnetworks (MSS) in a global network of interacting KOs. To find such subnetworks, each node (KO) in the global network was assigned a weight that reflected the statistical significance of its deregulation. These weights were calculated using the P-value distribution of KO fold changes obtained by differential expression analysis.

The subsequent filtering of KOs with low base mean expression across all samples reduced their numbers in the PD and DC datasets from 8882 and 8527 to 7105 and 5081, respectively. The numbers of KOs up- and downregulated ($P < 1$) in disease in the PD dataset were, respectively, 3870 and 3235 (Supplementary Fig. S2A). In the DC dataset, 2671 KOs were up- and 2410 were downregulated in disease (Supplementary Fig. S2B).

To transform the P-values into weights, we fitted a beta-uniform mixture (BUM) model to the distribution of P-values. The maximum-likelihood values of the mixture (λ) and shape (a) parameters of the BUM models were, respectively, 0.50 and 0.31 for the PD, and 0.48 and 0.54 for the DC dataset.

For each dataset, the parameters of the BUM model, the list of P-values and a false-discovery rate (FDR) were given as input to the Heinz package (Dittrich *et al.*, 2008), where KO weights were calculated following Equation (1). Heinz uses the FDR to calculate a significance (P-value) threshold according to which KOs with lower and higher P-values are assigned positive and negative weights, respectively (Fig. 1, secondary axis). In the PD dataset, we used a strict FDR = 0.0007 that corresponded to a very low significance threshold $P = 10^{-5}$ and constrained the number of positively scoring KOs, which in return limited the size of the resulting MSS, to an interpretable scale (approximately 40 KOs). For allowing a similar number of nodes to be positively scored in the more noisy DC dataset, we used a larger FDR = 0.11 that corresponded to a significance threshold $P = 0.003$ (Fig. 1B, red line). Both Figure 1A and B clearly show the signal (the Beta distribution at low P-values) as well as the noise component (the uniform distribution at higher P-values) in the datasets. For these FDR values, 134 nodes in the PD network (1.9%) and 152 nodes in the DC network (3%) were assigned positive weights, while the remaining nodes attained negative weights.

3.3 Significantly deregulated connected subnetworks

To obtain a global pathway map that covers the inter-pathway connections, we integrated all KEGG pathways into a global KO network of 6642 nodes (KOs) and 35 405 edges (KO interactions, Section 2.4). By taking the overlap between the global network and the KOs identified in the PD and DC datasets, we derived a global disease-specific network for each dataset, each covering more than

Table 1. Processing and mapping of reads for the two datasets

Dataset	Samples (#)	Raw reads	After QC (%)	After host mRNA removal (%)	After t/rRNA removal (%)	Mapped (%)	Genes (#)	KOs (#)
PD	6	773 306 101	46.6	44.8	33.1	4.5	2 195 939	8882
Disease	3	453 478 643	58.3	55.8	41.7	6.7	2 134 243	8872
Healthy	3	319 827 458	30.1	29.1	21.0	1.4	708 089	8290
DC	36	1 194 558 482	93.2	92.7	15.3	5.8	1 386 937	8527
Disease	19	643 025 297	93.8	93.5	14.4	5.6	1 073 778	8232
Healthy	17	551 533 185	92.4	91.8	16.3	6.1	1 079 242	8249

PD, periodontal disease; DC, dental caries.
The percentages are calculated with respect to the number of raw reads.

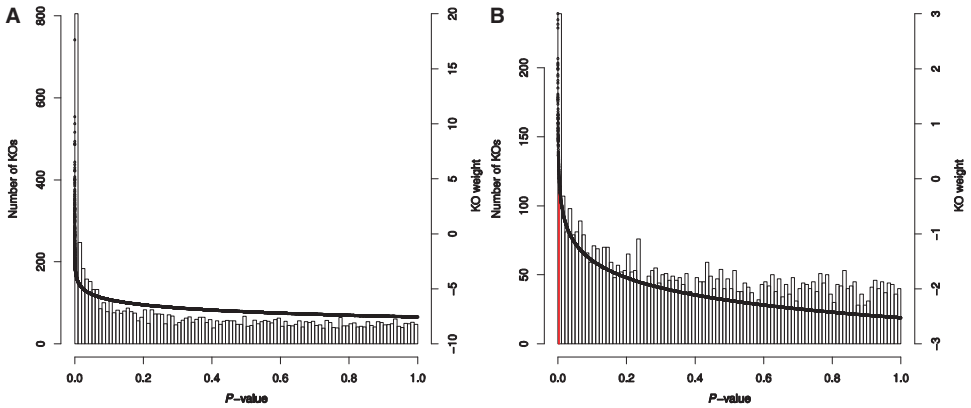

Fig. 1. The *P*-value distribution of KO fold changes (primary axis) was used to calculate the node (KO) weights (the secondary axis) in the networks. (**A**) The periodontal disease and (**B**) dental caries dataset. The red line shown for the caries dataset denotes the significance (*P*-value) threshold that separates KOs with positive and negative weights. Note that the black lines display the KO weights and not the BUM model fits to the distributions

53% of the KOs in the global network. Furthermore, each disease network included a large connected component that contained at least 77% of the nodes in the network as a whole.

Having the nodes of these global disease network annotated with weights from the previous step, we used the Heinz algorithm (Dittrich et al., 2008) and identified the connected MSS on each network.

3.3.1 Periodontal disease

The maximum-scoring subnetwork identified in the periodontal disease dataset consisted predominantly of KOs that were upregulated in disease (32 up- versus 7 downregulated in disease, Supplementary Table S1), suggesting that key mechanisms involved in disease are mostly due to overexpression of certain functions by the microbiome rather than the lessening of those found in health. The KOs in the subnetwork were members of 48 KEGG pathways in total (a single KO can be a member of multiple pathways). Nonetheless, all KOs upregulated in disease could be covered by the shown regions of six KEGG pathways (Fig. 2A). Among these, parts of the lysine degradation and butanoate, pyruvate, and sulfur metabolisms included in the subnetwork were almost entirely composed of KOs upregulated in disease.

A vast part of the periodontal disease subnetwork that was upregulated in disease covered the biochemical reactions that lead to the production of butanoate, the conjugate base of butyric acid, by two different mechanisms. The first mechanism appears to start with the 2-step breakdown of pyruvate to acetoacetyl-CoA in the pyruvate metabolism, where acetoacetyl-CoA is fed into the butanoate metabolism (Supplementary Fig. S3). In the butanoate metabolism, four KOs in the subnetwork covered the enzymatic steps necessary for the conversion of acetoacetyl-CoA to butanoate (Supplementary Fig. S4). The second mechanism was the lysine degradation pathway, where other KOs in the subnetwork (all upregulated in disease) fully covered the 6-step conversion of lysine to butanoate (Supplementary Fig. S5). The involvement of butanoate in periodontal disease has been known as an inhibiting agent for human cell growth (Chang *et al.*, 2013; Levine, 1985; Singer and Buckner, 1981). Additionally, the subnetwork KOs in the sulfur metabolism covered two of the three metabolic steps for the reduction of sulfate to sulfide. The role of sulfate-reducing bacteria in periodontitis has been previously reported as the production of cytotoxic levels of sulfide that lead to cellular destruction (Langendijk *et al.*, 1999, 2000). Other pathways that were mainly upregulated in disease included arginine, proline, glycine, serine and threonine metabolism.

Our results regarding disease-associated butanoate production largely overlap with the findings reported in the original study of the PD dataset (Jorth *et al.*, 2014), where the authors describe the involvement of lysine degradation in butanoate production. As an addition to their findings, we report the apparent interplay between the pyruvate and butanoate degradation pathways as a secondary mechanism for butanoate production, as well as the involvement of sulfide as a potential virulence factor in periodontal disease.

3.3.2 Dental caries

In contrast to periodontal disease, the maximum-scoring subnetwork identified in the dental caries dataset was mostly downregulated in disease (18 down- versus 13 upregulated KOs in disease, Supplementary Table S2). The health-associated KOs were mainly

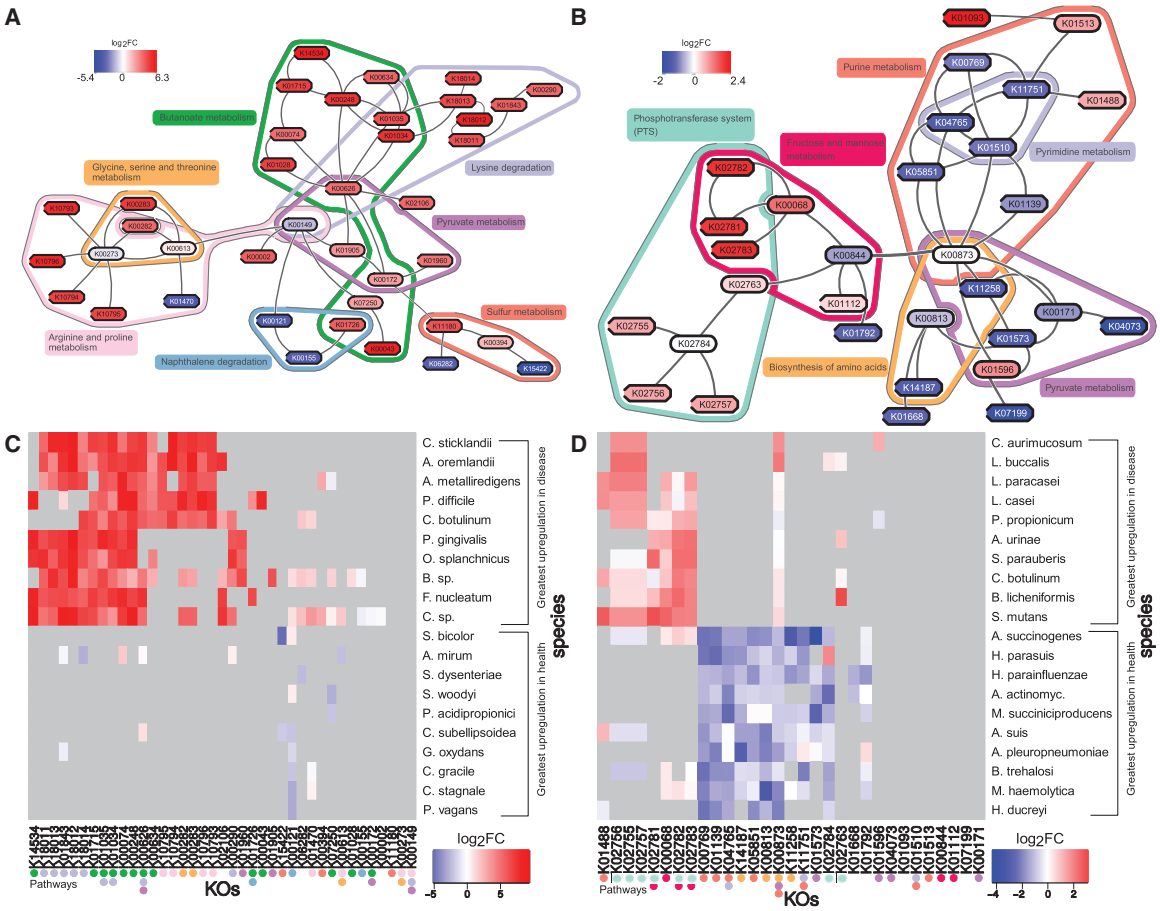

Fig. 2. The maximum-scoring subnetworks identified in (**A**) the periodontal disease and (**B**) dental caries datasets, where node colours denote the KO fold changes. For (**C**) the periodontal disease and (**D**) dental caries datasets, species that showed the greatest up- and downregulation in the expression of these functions in disease are shown. In all panels, red and blue, respectively, denote up- and downregulation in disease. Grey in (C) and (D) denotes the absence of KO expression in both health and disease. The coloured dots underneath the KOs in (C) and (D) correspond to the pathway colours as indicated in (A) and (B). The lists of KO descriptions and full species names can be found in the Supplementary Tables S1–S3

constituents of the pathways associated with amino acid biosynthesis and pyrimidine, purine and pyruvate metabolism (Fig. 2B). In the pyruvate metabolism, they covered the metabolic path for the breakdown of oxaloacetate to acetaldehyde. In contrast, oxaloacetate is converted to another substance, phosphoenolpyruvate, by the only KO in the pyruvate metabolism that was upregulated in disease (K01596), suggesting a different role for the compound in disease. The second pathway downregulated in disease was the biosynthesis of amino acids. Together with the upregulation of these pathways in health, we attribute the health-associated expression of KOs in the purine and pyrimidine metabolism to bacterial growth and proliferation. This argument might be supported by the fact that, by using a slightly higher false-discovery rate, the dental caries subnetwork could have been extended to include six other KOs (all upregulated in health) from the glycerophospholipid metabolism, the main products of which are the components of bacterial cell membranes required for growth. Earlier studies show that the metabolic activity of caries-promoting bacteria results in low pH conditions that drastically affect the survival of other oral species (Bradshaw *et al.*, 1989; Harper and Loesche, 1984). On the other hand, in healthy state, mutual interactions are necessary to obtain energy and maintain growth, for instance for the catabolism of salivary glycoproteins (Dejong and Vanderhoeven, 1987).

The disease-associated parts of the caries subnetwork included 9 KOs from the pathways of phosphotransferase system and fructose and mannose metabolism (Fig. 2B). Perhaps functionally the most interesting of these KOs are the sugar phosphotransferases involved in the beta-glucoside metabolism (K02755, K02756 and K02757). The importance of beta-glucoside metabolism for cariogenic oral bacteria has been put forward by studies showing the inhibitory effect of mutations in the beta-glucoside metabolism genes on biofilm formation, a vital process for the colonization of these bacteria on tooth surfaces (Kiliç *et al.*, 2004). A secondary set of disease-associated KOs in the caries subnetwork that might be functionally interesting are from the fructose and mannose metabolism. Four of these KOs convert sorbitol to fructose 6-phosphate. Unlike many other oral species, the well-known cariogenic bacteria *Streptococcus mutans* can metabolize sorbitol and mannitol as a carbon source (Ajdic *et al.*, 2002) In contrast, the only KO downregulated in disease (K00844) in the same pathway converts fructose to fructose 6-phosphate. Within fructose metabolism, this suggests that while in health fructose is the main carbon source, in dental caries sorbitol might be used as an additional source of carbon.

We note that the DC dataset in its original study (Peterson *et al.*, 2014) was not used to perform a differential expression analysis, but rather an overall gene expression analysis of the healthy and diseased samples as a whole. Therefore, to our knowledge, these are the first results obtained from a metatranscriptomic dataset that aim at elucidating the role of functional metabolic modules that are involved in dental caries.

3.3.3 Significant subnetworks

To assess the likelihood of obtaining the MSSs identified in the PD and DC datasets due to chance, we used a resampling procedure. We randomly shuffled the KO weights in the global disease networks, calculated the MSSs in the networks with shuffled weights, and determined the cumulative weight of the identified subnetworks ($n = 500$). In the PD dataset, none of the 500 shuffled networks resulted in an MSS that outscored the periodontal disease subnetwork. For the DC dataset, only 36 out of 500 MSSs attained cumulative scores that were higher than that of the caries subnetwork. We attribute these 36 random networks to the lower signal content and the higher FDR used in the dental caries dataset. The comparison between the cumulative weights of original subnetworks and those that resulted from resampling strongly indicates that the results reported above are not originating from random effects (Supplementary Fig. S6).

3.4 Contribution of different species to subnetworks

Previous studies of microbiome-related diseases have demonstrated that the metabolic functions associated with disease or health may be performed by particular taxonomic groups within a microbial community (Belda-Ferre *et al.*, 2012; Man *et al.*, 2011). To determine whether this applies to the functional disease subnetworks we identified in this study, we determined the lists of 10 species that showed the greatest overall increase or decline in disease in the expressions of functions that were found in the subnetworks (Section 2.6).

3.4.1 Periodontal disease

The KOs in the butanoate metabolism were upregulated by almost all of the top 10 species that showed the largest increase in the expression of periodontal disease subnetwork KOs in disease, including the well-known periodontal pathogens *Porphyromonas gingivalis* and *Fusobacterium nucleatum* (Fig. 2C). Previously, only the *F. nucleatum* species was proposed to be responsible for butanoate production (Jorth *et al.*, 2014). Our analysis suggests otherwise; a wider range of organisms seems responsible for the production of butanoate. Note that our reference metagenome database included genes from all species in KEGG, while Jorth *et al.* (2014) used genomes of 60 oral species to construct their reference database, which may explain why we were able to find a wider range of species.

Although expressing all other KOs in the butanoate metabolism, *P. gingivalis* and *Odoribacter splanchnicus* surprisingly did not express one of the key KOs in the same pathway (K00626, EC 2.3.1.9 in Supplementary Fig. S4). This suggests that in these bacteria, this function is covered by enzymes from other pathways, or that its product (acetoacetyl-CoA) is acquired from the environment.

3.4.2 Dental caries

The 10 species that showed the greatest upregulation of the KOs in the dental caries subnetwork in disease included *Streptococcus mutans* and *Lactobacillus casei*, which frequently are associated with dental caries (Fig. 2D). Interestingly, in addition to the well-known cariogenic species, we observe, as in the periodontal disease dataset, a wider range of species with a similar pathogenic expression profile in the maximum-scoring subnetwork.

A contradictory finding in this section is that *Streptococcus, Clostridium and Lactobacillus* species appear to show an increased expression of adenosine deaminase (K01488) in purine metabolism in the disease state. One of the products of this enzyme is ammonia, the production of which increases the pH in the oral cavity and therefore is generally described as an inhibitory factor on caries development (Burne *et al.*, 2012; Koopman *et al.*, 2015). This result is even more intriguing, considering that the species showing the greatest overall upregulation of the KOs in the caries module in health did not show increase in the expression of adenosine deaminase.

3.5 Comparsion with other methods

As there is currently no other method for identifying subnetworks of significantly deregulated KOs in metatranscriptomics datasets, we compared our approach with MetaPath, a method that is developed for solving a similar but fundamentally different biological problem. Unlike metaModules, MetaPath aims to find subnetworks of enzymatic reactions that are enriched or depleted as a whole in a given experimental condition.

Since the global metabolic networks natively used by the two methods are different, for consistency, we applied both methods on the MetaPath network (see Section 2.7 for details).

The comparison of their results immediately makes it clear that the two methods address distinct biological questions: whereas MetaPath finds subnetworks of fully up- or downregulated KOs, metaModules finds a subnetwork in which the vast majority of KOs is significantly deregulated (Supplementary Figs S7 and S9).

The overlap of 7 KOs between the subnetworks calculated by metaModules and MetaPath are parts of the two disease-associated pathways described in Section 3.3: fructose and mannose metabolism and phosphotransferase system (Supplementary Figs S7 and S8). This finding is in line with the above conclusions; both of these pathways were mostly significantly deregulated in a certain phenotype (caries-positive) and therefore were parts of the solution spaces of both methods.

In conclusion, these results suggest that the two methods are complementary as they address different questions.

3.6 Metabolic pathways in microbial communities

Since microbial communities can communicate and interact by exchanging metabolites (Rath and Dorrestein, 2012), metabolic pathways and in particular KOs are well-suited to study the gene expression levels in full microbial communities and changes thereof in disease-associated communities. The suitability of our approach within this framework is highlighted by the recovery of essential up- and downregulated subnetworks by the analysis of high throughput metatranscriptomic datasets. Within our datasets, we do indeed seem to find hints of species relying on other species within the community, e.g. as a source for metabolites. This means that high throughput sequencing experiments will in fact be one of the most common ways to study such species, as culturing such organisms in the laboratory will be understandably difficult (Ling *et al.*, 2015). Therefore, the computational methodology developed here to study these mechanisms in their natural communities will be important to understand the full mechanisms of microbiome-related diseases.

4 Conclusion

We developed a new methodology for comparative metatranscriptomics that facilitates the automated identification of metabolic hotspots that are significantly deregulated in disease, and recovered disease-associated functions in two oral microbiome-mediated disorders. The diseases examined in this work, periodontal disease and dental caries, have been extensively investigated using a wide range

of experimental methods in previous studies. Being able to automatically recover large parts of the known disease biology using single high throughput transcriptomic datasets from full microbial communities is an important step forward for the discovery of currently unknown factors in microbiome-mediated diseases.

Acknowledgements

The authors thank Kasper Dinkla for adapting the eXamine software.
Funding: University of Amsterdam under the research priority area 'Oral Infections and Inflammation' (to W.C.) and S.A. is supported by The Netherlands Organisation for Scientific Research (NWO).

Conflict of Interest: none declared.

References

Abubucker,S. *et al.* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, **8**, e1002358.

Ajdic,D. *et al.* (2002) Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc. Natl. Acad. Sci. USA*, **99**, 14434–14439.

Beisser,D. *et al.* (2010) BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–1130.

Belda-Ferre,P. *et al.* (2012) The oral metagenome in health and disease. *ISME J.*, **6**, 46–56.

Bradshaw,D.J. *et al.* (1989) Effects of carbohydrate pulses and pH on population shifts within oral microbial communities *in vitro*. *J. Dent. Res.*, **68**, 1298–1302.

Burne,R.A. *et al.* (2012) Progress dissecting the oral microbiome in caries and health. *Adv. Dent. Res.*, **24**, 77–80.

Caspi,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.

Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.

Chang,M.C. *et al.* (2013) Butyrate induces reactive oxygen species production and affects cell cycle progression in human gingival fibroblasts. *J. Periodont. Res.*, **48**, 66–73.

Dejong,M.H. and Vanderhoeven,J.S. (1987) The growth of oral bacteria on saliva. *J. Dent. Res.*, **66**, 498–505.

Dinkla,K. *et al.* (2014) eXamine: exploring annotated modules in networks. *BMC Bioinformatics*, **15**, 201.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, I223–I231.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Franzosa,E.A. *et al.* (2014) Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. USA*, **111**, E2329–E2338.

Harper,D.S. and Loesche,W.J. (1984) Growth and acid tolerance of human dental plaque bacteria. *Arch. Oral. Biol.*, **29**, 843–848.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res*, **17**, 377–386.

Huttenhower,C. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Jorth,P. *et al.* (2014) Metatranscriptomics of the human oral microbiome during health and disease. *mBio*, **5**, e01012–e01014.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kiliç,A.O. *et al.* (2004) Involvement of *Streptococcus gordonii* beta-glucoside metabolism systems in adhesion, biofilm formation, and in vivo gene expression. *J. Bacteriol.*, **186**, 4246–4253.

Koopman,J.E. *et al.* (2015) Stability and resilience of oral microcosms toward acidification and Candida outgrowth by arginine supplementation. *Microb. Ecol.*, **69**, 422–433.

Kopylova,E. *et al.* (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.

Langendijk,P.S. *et al.* (1999) Sulfate-reducing bacteria in periodontal pockets and in healthy oral sites. *J. Clin. Periodontol.*, **26**, 596–599.

Langendijk,P.S. *et al.* (2000) Sulfate-reducing bacteria in association with human periodontitis. *J. Clin. Periodontol.*, **27**, 943–950.

Leimena,M.M. *et al.* (2013) A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics*, **14**.

Levine,M. (1985) The role for butyrate and propionate in mediating Hela-cells growth-inhibition by human dental plaque fluid from adult periodontal-disease. *Arch. Oral. Biol.*, **30**, 155–159.

Ling,L.L. *et al.* (2015) A new antibiotic kills pathogens without detectable resistance. *Nature*, **517**, 455–459.

Liu,B. and Pop,M. (2011) MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. *BMC Proc.*, **5**, S9.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Luo,C. *et al.* (2013) A user's guide to quantitative and comparative analysis of metagenomic datasets. *Meth. Enzymol.*, **531**, 525–547.

Man,S.M. *et al.* (2011) The role of bacteria and pattern-recognition receptors in Crohn's disease. *Nat. Rev. Gastro. Hepat.*, **8**, 152–168.

Marchetti,A. *et al.* (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc. Natl. Acad. Sci. USA*, **109**, E317–E325.

Mitra,K. *et al.* (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.*, **14**, 719–732.

Morgan,X.C. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, **13**, R79.

Oliveira,A.P. *et al.* (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.*, **2**, 17.

Patil,K.R. and Nielsen,J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. USA*, **102**, 2685–2689.

Peterson,S.N. *et al.* (2014) Functional expression of dental plaque microbiota. *Front. Cell. Infect. Microbiol.*, **4**, 108.

Pihlstrom,B.L. *et al.* (2005) Periodontal diseases. *Lancet*, **366**, 1809–1820.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Qin,J. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

Rath,C.M. and Dorrestein,P.C. (2012) The bacterial chemical repertoire mediates metabolic exchange within gut microbiomes. *Curr. Opin. Microbiol.*, **15**, 147–154.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Selwitz,R.H. *et al.* (2007) Dental caries. *Lancet*, **369**, 51–59.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Singer,R.E. and Buckner,B.A. (1981) Butyrate and propionate—important components of toxic dental plaque extracts. *Infect. Immunol.*, **32**, 458–463.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. USA*, **102**, 15545–15550.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Wang,Z. *et al.* (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, **472**, 57–63.