DIET SHIFTS PROVOKE COMPLEX AND VARIABLE CHANGES IN THE METABOLIC NETWORKS OF THE RUMINAL MICROBIOME

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

SARA WOLFF

Dr. Gavin Conant, Thesis Supervisor

May 2016

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

DIET SHIFTS PROVOKE COMPLEX AND VARIABLE CHANGES IN THE METABOLIC NETWORKS OF THE RUMINAL MICROBIOME

presented by Sara Wolff, a candidate for the degree of Master of Science, and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Gavin Conant

Professor William Lamberson

Professor Chris Pires

ACKNOWLEDGEMENTS

I cannot express enough thanks to my committee for the support and encouragement they have provided me along the way: Dr. Gavin Conant[1,2], my advisor and committee chair, who guided the project and whose office was always open whenever I required assistance, Dr. William Lamberson[1], who assisted with the experimental design for this project and Dr. Chris Pires[3], for helpful advice and conversations along the way.

In addition, I would like to acknowledge and thank Tasia Taxis[9], who worked with me to develop the initial data handling methods for the Illumina data as well as Melinda J. Ellison[4] and Kristi M. Cammack[4], who performed the initial taxonomic study on this group of animals. Our orthology pipeline was developed by Michael Baraboo[6] and Hyuk Jin Lee[3], and debugged and processed by Yue Hao[2], Katherine Burch[7] and Taylor Mauer[8]. Data samples were collected by Rebecca R. Cockrum[5] and Kathy J. Austin[5]. And a special thanks to Huan Truong[2], who provided assistance with the use of R and principal component analysis. I would also like to thank M. Becchi[2], J. Decker[1], R. Schnabel[1], J. Taylor[1], R. Patil[1] and A. Ravelo[3] for helpful discussions.

This project was supported USDA National Research Initiative (NRI) grant # 2011-68006-30185 (MJE, RRC, KJA, WRL, KMC, GCC), National Science Foundation grants NSF-CCF-1421765 (HT, GCC) and NSF-DBI-1358997 (MB, HJL, KB, TM, GCC) and European Union Framework 7 project GplusE (GCC).

[1]Division of Animal Sciences, [2]Informatics Institute, and [3]Division of Biological Sciences, University of Missouri-Columbia, U.S.A., [4]Department of Animal Science,

University of Wyoming, Laramie, WY, U.S.A., [5]Department of Animal Science, Colorado State University, Fort Collins , CO, U.S.A., [6]Department of Computer Science, Truman State University, Kirksville, MO, U.S.A., [7]Department of Psychology, Truman State University, Kirksville, MO, U.S.A., [8]Department of Biology, Kenyon College, Gambier, Ohio, U.S.A. and [9]National Animal Disease Center, ARS, USDA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

RFI: residual feed intake

FORG: Forage-based diet

CONC: Concentrate-based diet

OTU: Operational taxonomic unit

PCA: Principal Component Analysis

DIET SHIFTS PROVOKE COMPLEX AND VARIABLE CHANGES IN THE METABOLIC NETWORKS OF THE RUMINAL MICROBIOME

Sara Wolff

Dr. Gavin Conant, Thesis Supervisor

Abstract

Grazing mammals rely on their ruminal microbial symbionts to convert plant structural biomass into metabolites they can assimilate. To explore how this complex metabolic system adapts to the host animal's diet, we inferred a microbiome-level metabolic network from shotgun metagenomic data. Using comparative genomics, we then linked this microbial network to that of the host animal using a set of interface metabolites likely to be transferred to the host. When the host sheep were fed a grain-based diet, the induced microbial metabolic network showed several critical differences from those seen on the evolved forage-based diet. Grain-based (e.g., concentrate) diets tend to be dominated by a smaller set of reactions that employ metabolites that are nearer in network space to the host's metabolism. In addition, these reactions are more central in the network and employ substrates with shorter carbon backbones. Despite this apparent lower complexity, the concentrate-associated metabolic networks are actually more dissimilar from each other than are those of forage-fed animals. Because the both groups of animals were initially fed on a forage diet, we propose that the diet switch drove the appearance of number of different microbial networks, including a degenerate network characterized by an inefficient use of dietary nutrients.

Introduction

Ruminant animals, including sheep, are able to subsist on diets that cannot sustain most mammals because microbial symbionts, living in their rumen, degrade cellulose and other plant structural compounds into volatile fatty acids (VFAs), amino acids, and other nutrients that the host animal can use (1-4). However, many of the details of this association are poorly understand: many of the microbes involved have not or cannot be grown in culture, and even those that have been cultured have been comparatively little studied (5-8).

Previous studies characterizing the microorganisms in the rumen and their metabolism have used both culturing approaches as well as PCR amplification of small-subunit ribosomal RNA genes (rrs) (7,9). Though culturing allows us to characterize individual taxa, it limits the search to organisms that grow well as monospecies cultures. Some cultures can also have undergone mutation and adaptation to the cultured environment (10,11). RNA fingerprinting techniques such as Western blots performed on various metagenomic (i.e. environmental) samples can be used to visually differentiate microbial environments on a gel, however this information comes without the benefit of precise sequence identification (12). The amplified sequencing of 16S rDNA genes from microbial samples to identify taxa present can only identify organisms for which a primer is available or in pre-defined databases of taxa (13-16). The more recent technique of shotgun sequencing all DNA within a metagenomic sample offers an opportunity to obtain a more diverse genetic sampling of the microbes within the gut. If one could assemble the DNA from such a sample, it would conceivably be possible to create a taxonomic profile of organisms from that sample. However, there are significant hurdles

to genome assembly, including the lack of a complete catalog of gut microbes as reference genomes, the propensity of microbes to mutate and undergo lateral gene transfer, and the high degree of similarity between related microbes (13,17). In addition, several studies have provided evidence that taxonomic differences may mask similar functionality of the rumen metabolism as a whole (18-21). This observation implies that while the organisms in the rumen may be different between individual animals, they are providing similar metabolic benefits in regards to the host-microbe interface. In other words, the microbial genes present in the rumen may be similar among different animals, despite the differences in microbe taxa present (22,23).

While many metagenomic studies rely on taxonomy alone, others emphasize the importance of comparing taxa to metabolic or genetic differences in groups of individuals (24,25), re-enforcing the idea that microbial ecosystems are too complex to be defined by one phenotype or gene (8,16,23,26,27). Because of the difficulties in genome assembly based on metagenomic data, many studies instead rely on mapping metagenomic reads to databases of genes such as that of the Kyoto Encyclopedia of Genes and Genome (KEGG) to identify, for instance, enzyme-coding genes (16,23,24,26) Of course, this approach is constrained by the number and nature of the genes in the database.

Another challenge to rumen metabolic analysis is the possibility that rare, uncharacterized members of the rumen environment provide flexibility in adapting to allow the digestion of new food: the effect of these rare organisms in diet shifts and their effect on the host-rumen metabolic equation deserve consideration (28),(29). An additional hurdle to analysis of rumen metabolism has been the limitations in the databases of microbial enzymes and in their ability to link microbial and host metabolism

(12,30). However, projects such as MetaCyc have begun to remedy this deficiency (31,32).

While 16S rDNA studies have given us good surveys of the taxonomic makeup of the rumen, our knowledge of the metabolism of these organisms, either individually or collectively, is considerable less advanced, as is our understanding of how that metabolism links to that of the host.

Gene-centric methods, in which sequencing data is matched to a database of microbial genes, thus can provide complement to taxonomic approaches. The benefit in this approach is that comparisons can be made between identified taxa and identified genes, and also that further systems biology approaches can identify relationships between those genes. This may lead to the association of certain microbial genes with diseases (18,20,25,33). A recent study of the human gut microbial ecosystem in inflammatory bowel disease (IBD) and obesity by Greenblum *et al*. (26) employed this approach in studying microbial metabolism by mapping reads to an enzyme database.

Our end goal was to enhance our understanding of rumen metabolism and the relation of that metabolism to the host, and relying on any one of these methods leaves out key components of the metabolic puzzle. In our previous work (21), we performed a metagenomic analysis of the rumens of two steers. Our intent was to study the metabolic network of the rumen, or rather the set of metabolic reactions by which the microbes break down food substances into smaller compounds that can be absorbed and used by the host animal for its own metabolic processes. We found that while these two individuals showed a surprising degree of taxonomic distinctiveness, their metabolic

networks, particularly parts that interface with the host, were rather similar. Here, we employ the same sequencing strategy and computational pipeline to explore how a diet change alters the structure of the microbial metabolic network. We sampled the rumen fluid of 16 sheep, 8 of which were fed a grain-based concentrate diet and 8 of which were on an alfalfa-based forage diet. By mapping metagenomic reads to an enzyme database, we inferred a metabolic network for the rumen microbes. We also inferred a similar network for sheep using comparative genomics. We found that the two diets differed mainly in abundance of metabolic reactions, rather than the presence or absence of specific reactions. In addition, the concentrate diet reactions were closer to the host's metabolism, were more centrally located in the network, had substrates with shorter carbon backbones, and had animals that were more dissimilar from each other than were the forage diet animals.

Methods

Animal Use and Rumen Sample Collection

As previously described (34), rumen fluid was sampled from 16 growing wethers of Rambouillet, Hampshire, and Suffolk breeds. After 24 days on a shared diet of primarily hay supplemented by corn, eight animals were randomly assigned to a pelleted, concentrate-based diet (CONC: main component corn), while the remaining animals were fed a pelleted, forage-based diet (FORG: main component alfalfa). Animals were acclimated to their respective diets over a 25-day period. Rumen fluid samples were collected at slaughter immediately following a 49-day feed trial, and frozen at -80° C.

DNA Extraction & Library Preparation

Thawed 1ml rumen samples plus sterilized zirconia (0.3 g of 0.1 mm) and silicon (0.1 g of 0.5 mm) beads and 1 ml of lysis buffer were homogenized with a Mini-Beadbeater-8 (three minutes), incubated at 70°C (15 minutes + mixing every 5 minutes), and centrifuged at 4°C (16,000g for 5 minutes). The resulting supernatant was transferred to new tubes and fresh lysis buffer was added. This process was repeated a second time, and supernatants pooled. The QIAamp DNA Stool Mini Kit (Qiagen, Santa Clarita, CA, USA) was then used to precipitate nucleic acids and to remove RNA and proteins.

Genomic libraries were constructed these 16 DNA samples using Illumina's DNA sample kit according to the manufacturer's recommended protocol (34). Genomic DNA was sheared to generate fragments of 300bp using Diagenode BioRuptor and standard methods. The resulting 3' and 5' overhangs were removed, adenosine nucleotide added to

3’ ends, and Illumina adapters ligated. Fragments of 420 base pair (bp) were then size selected on agarose gel and recovered as described in the Illumina protocol. Qubit assays were used to quantify each purified library, and fragment size was confirmed by the Agilent BioAnalyzer High Sensitivity DNA assay.

Illumina Sequencing

Each sample was sequenced on an Illumina HiSeq 2000 with four samples pooled per flowcell lane, resulting in 100 base-pair, paired end sequences, with mean insert size of 309 bp. All raw sequence reads are available from NCBI’s short read archive (Project SRP028527). We quality filtered the paired-end reads by truncating each read after the first run of three bases where the phred quality score was less than 15 (35). If one or both reads in a pair had an average quality score less than 25 or was shorter than 85 bases, the read pair was omitted. After this filtering, 96 gigabases of sequence remained. Samples from the CONC animals ranged from 7.8 – 54.9 million paired reads per animal, with 0.02% - 0.63% reads matching to the sheep genome. Samples from FORG animals ranged from 16.8 – 49.2 million paired reads per animal with 0.01% - 5.17% reads matching to the sheep genome (Table 1).

OTU Analysis

As previously described (34), the filtered reads were compared to the sequences of the Ribosomal Database Project (36) using Bowtie (37) to identify fragments of 16S rDNA genes. We retained hits that matched sequences in the database at 97% or greater identity. OTUs were defined by single-linkage clustering (38) in the original database:

read pairs that mapped to exactly one such OTU were considered instances of that OTU in their respective animal. The result was a total of 349 OTUs, a complete analysis of which is presented in Ellison et al. (34).

Metabolic network inference from MetaCyc

MetaCyc is a small-molecule metabolism database that includes more than 2000 microbial metabolic networks (31). Enzyme sequences from these networks are annotated to reactions, including their respective substrates and products. From these data, we created an enzyme-centered database in which each enzyme was linked to every metabolic reaction that it might catalyze.

Using these reactions and their respective metabolites and catalyzing protein sequences, we inferred a global metabolic network for all bacteria and archaea in MetaCyc (21). We then linked that network to that of the host animal by using comparative genomics to infer a sheep metabolic network. For both networks, nodes were defined as metabolic reactions (catalyzed by enzymes): edges connect nodes that have a metabolite in common (circles and lines, respectively, in Figure 1). Reactions that did not share a metabolite with any other reaction were not included in our network. For the microbial network, any two reactions with identical metabolites in MetaCyc were merged, resulting in a total of 6140 reactions/nodes.

There is no ovine network in MetaCyc, and the bovine metabolic network only contains 1404 annotated reactions. Thus, to create an ovine metabolic network, we began with the human metabolic network, which has 2863 annotated reactions. To this network, we added any bovine reactions not already present in the human reconstruction. We then

used enzyme orthology to infer whether each enzyme in this merged network could be identified in the ovine genome. We used our previously described synteny-based orthology inference package (39,40) to infer in human-ovine and ovine-bovine orthology pairs based on Ensembl release 75 (41). We identified 15562 human-ovine and 15479 bovine-ovine ortholog pairs. Any human or bovine enzyme with an ortholog in the ovine genome was assumed to be present in sheep network. A required VFA reaction involving butyrate was missing from MetaCyc. Thus, to these ovine orthologs of 1834 human reactions and 154 bovine reactions, we added a pseudo-reaction converting butyrate to butyryl-CoA, giving a total of 1989 nodes in the ovine host metabolic network.

For both the microbial and host networks, metabolites that occurred in a large number of reactions, or "currency metabolites" were removed. Because there is no universal definition of "currency metabolite" (42), we ran all analyses three times using currency cutoffs of 25, 50 and 100 (in other words, not creating edges for metabolites occurring in 25 or more, 50 or more of 100 or more reactions; 43). Henceforth, the corresponding networks will be referred to as N_{25}, N_{50} and N_{100}. There were 261, 206 and 174 metabolites defined as currency metabolites in networks N_{25}, N_{50} and for N_{100}, respectively.

Mapping translated reads to the microbial metabolic network

To infer which of the MetaCyc reactions could be detected among the microbes from our shotgun sequences, we first translated our paired reads in all 6 possible open reading frames (ORFs). The resulting paired amino acid sequences were discarded unless both had a translated ORF longer than 29 residues. These translated reads can be more

efficiently searched than DNA sequences (21): using our custom search tool based on the SeqAn library, we searched the translated reads for 7-residue identical matches to a database sequence (44). Whenever two such matches were found in the same read/database pair, we locally aligned the pair using the Smith-Waterman algorithm (45). Alignments having >80% amino acid identity over 80% of the metagenomic ORF for both members of a read pair were retained. For efficiency, we used a compressed version of the MetaCyc database where sequences more than 97% identical to other sequences were omitted (21), yielding a reference database of 760,000 enzymes.

Using these database hits, we then assigned reads to metabolic network nodes. Enzymes can catalyze more than one reaction. We handled network node assignments in these cases in one of three ways. First, in cases where reads mapped to sequences catalyzing the same reactions, or when they also mapped to sequences catalyzing a subset of these reactions, the union of those reactions was recorded as a single node. Second, in cases where read pairs mapped to multiple non-related reactions, the read pair was discarded, as the reaction identity was impossible to ascertain. Third, if the forward and reverse reads resulted in ORFs that did not map to the same enzyme, that pair was also discarded.

VFA Analysis

We sought to assess if the read counts mapped to enzymes were correlated with the observed metabolite levels for three volatile fatty acids (VFAs). To obtain the VFA concentrations, 5ml rumen fluid samples from each animal were centrifuged for 10 minutes at 3000g. The resulting supernatant was added to a 25% metaphosphoric acid

solution containing 2-ethyl butyric acid (2.0mg/mL). The final ratio of rumen fluid to metaphosphoric acid solution was 5:1. We then incubated the samples for 30 minutes on ice and centrifuged once more, this time for 30 minutes. The resulting supernatant samples were then added to 1mL vials for analysis by gas liquid chromotography. VFA concentrations were then determined with an Agilent 6890 gas chromatograph following standard procedures as required for this machinery. Note that data for animal 1127 was absent for technical reasons. The Spearman's correlation of normalized read counts and VFA concentrations were computed in R (46). For reference, we note that butyrate appeared in the reactions of 7 nodes, propionate in 12, and acetate in 79.

Residual feed intake measures

The 16 animals used for metagenomic sequencing were selected based being in the extremes of their efficiency with which they converted nutrients into body mass, e.g., their residual feed intake (34). Thus, RFI was calculated as the deviation of true feed intake from expected feed intake. Expected feed intake was determined by regressing actual feed intake on daily gain and metabolic midweight (47).

Network Interface

In order to define an interface between sheep and microbial metabolic networks, we required a list of compounds that are potentially transferred from the rumen microbes to the host, which we henceforth refer to as "interface metabolites." We used three different sets of metabolites to link the microbial and ovine networks. The first set (VFA) consisted of the three most abundant VFAs produced in the rumen, which generally

account for 95% or more of total VFAs in that environment: acetate, propionate, and butyrate (48). Recall that it was necessary to add a reaction converting butyrate to butyryl-CoA to the host network (green node in Figure 1). To construct the second interface set we added the 20 universal amino acids to the VFAs (VFA + AA). Finally, the third set consisted of VFA + AA plus a set of metabolites known to be absorbed from the gut into cells in the digestive tract of humans (ALL). This list started with a list of 404 exchangeable human metabolites from the human metabolic network of Duarte et al., (49). Of these compounds, 234 matched the MetaCyc database. The majority of the remaining unmatched compounds are complex eukaryotic polysaccharides and lipopolysaccharides that are not synthesized by microbes (data not shown). An additional 30 compounds were excluded for being inorganic, DNA/RNA nucleotides, or CO_2, NAD or NADP. The final ALL interface set thus consists of 204 metabolites.

Edges between the host metabolic network and microbial metabolic network were defined where a node in one network shared an interface metabolite with a node in the other network, regardless of whether that interface metabolite was also defined as a currency metabolite. The result of this procedure is a single metabolic network, as illustrated in Figure 1.

Layering of Metabolic Networks

The interface metabolites implicitly define a concept of distance between the host and microbial networks. To compute this distance, we used an $O(n^2)$ version of Dijkstra's algorithm (50), where n is the number of nodes in the graph, to find the shortest path between all pairs of nodes in the network. For a given microbial (or host) node, the

distance of that node to the host (or microbial) network is simply the length of the shortest path between it and whichever node or nodes in the other network is found by the algorithm to be nearest to it. These distances partition the network into layers, where the first layer on the host and microbe sides includes reactions using an interface metabolite (which are at distance 0 from the other network; Figure 1). Each layer further from the center is another reaction removed from sharing a metabolite with the other network.

Node read density analysis

Each reaction (node) with at least one mapped read pair was analyzed by using the two-sample Wilcoxon test to test null hypothesis of no difference in normalized read count for the two diets (FORG or CONC). We then applied a 5% false-discovery rate correction (FDR) to the set of *P*-values produced by this test (51). We used the non-parametric Wilcoxon test because it is unclear what distribution the mapped read counts follow. Nodes with a significant difference between FORG (green) or CONC (red) were assigned to a color gradient based on the log of the ratio of normalized read counts (Figure 1). Nodes without a significant difference in normalized read count are shown in black.

To assess if the two diets differed in how reads were distributed among nodes, we fit a three state distribution to the normalized number of reads mapped to each node for each diet. The distribution had one proportion of nodes with no reads mapped, a second with 1 read mapped and a third proportion where the number of reads mapped to a node was assumed to follow a log-normal distribution. We used this somewhat complex distribution because visual inspection of our data suggested that nodes with 0 or 1 reads

mapped induced a clearly tri-model form to the reads count distribution. We first individually fit this distribution to the read counts from each diet separately. We then compared the sum of the log-likelihoods from this approach to the log-likelihood of fitting the combined read count data to a single distribution.

Network structure analyses

We calculated the Pearson correlation between the mapped read counts and network layer for all combinations of networks and interface metabolite sets. We also determined the correlation for the proportion of differentially abundant enzyme genes for each layer (Table 2).

We next calculated the mean network layer for the mapped reads across for the two diets across the different interface metabolite sets and networks (e.g., N_{25}, N_{50}, and N_{100}; Table 3). To assess whether this mean was significantly different for the FORG and CONC diets, we adopted a randomization approach. First, we pooled all of the mapped reads across the two diets. Then, for each of 1000 randomizations, we selected at random from these pooled reads the same number of reads as had originally been mapped to the eight FORG animals. (The remainder of the reads were assigned to CONC). We then calculated the difference in mean layer number for the pseudo-concentrate and pseudo-forage reads for each randomization. If no more than 5% of the randomized differences in mean layer are greater than the observed real difference, that is evidence that the two diets differ in mean layer.

We also analyzed the effect of diet on the metabolic network structure using four statistics:

- "Carbon sum," the total number of carbon atoms appearing in the reaction associated with that node,
- Betweenness-centrality, namely the total number of shortest paths between any pair of network nodes that pass through the selected node (52),
- Node degree, or the node's total edge count, and
- Clustering coefficient (53), which indicates the degree to which nodes clustered

For each node, we weighted these values by the normalized number of reads mapped to that node and then computed the difference in the mean of these four statistics across all nodes between the two diets. To assess whether the diets differed in the statistics, we again used a randomization approach. As before, we randomly reassigned a number of reads to each diet equivalent to the number observed in the real data. We then computed the difference in mean statistic for all four statistics. Any statistic where the observed difference for the real data was greater than that in 95% (950 randomizations) of the randomized datasets was considered significant (Table 3).

Principal Components Analysis

Principal Component Analysis (PCA) was performed with the R statistical package (46), using the normalized read counts as measurements and sampled animals as experiments. In our previous work, we identified normalized counts of 349 OTUs in these sixteen animals (34). We thus performed a separate PCA on them in the same format.

Pairwise distances between samples in OTU and node distribution

For each sampled animal, we defined an OTU distribution vector and a node distribution vector *v*, the elements of which are defined as:

$$v_i = \frac{r_i}{\sqrt{\sum_{j=0}^{n} (r_j)^2}} \qquad (1)$$

where r_i is the number of reads mapped to node/OTU *i* and the denominator scales the resulting vector to unit length. We then computed standard Euclidian distances between all pairs of vectors v_i and v_j from the animals *i*,*j* for both nodes and for OTUs.

To assess if the differences between pairs of animals given the same diet were larger than could be explained by sampling variance, we pooled all reads from each diet, randomly reassigned them proportionally to the animals, and recomputed the pairwise distances 1000 times (to obtain a p-value of .001). From these resamplings, we computed distributions of minimum, maximum and mean pairwise differences within each diet and between the two diets, as well as the correlation between node and OTU distances. These randomized correlation values are occasionally high, mostly likely because differences in the number of mapped reads between animals can generate outliers and hence spurious associations. Because we compared the Pearson correlation observed from the real data to those seen in the randomized data, our approach should not suffer from the violations of the normality assumptions seen with using Pearson correlation statistics with non-normal data.

Results

Metagenomic sequencing of rumen fluid from 16 sheep

We extracted microbial DNA from the rumen fluid from sixteen sheep and then shotgun sequenced those DNA on an Illumina GenomeAnalyzer II, yielding 100 base pair, paired end reads. These reads were not significantly contaminated with host DNA (Table 1). In our previous work, we identified 349 operational taxonomic units (OTUs) in the rumens of these animals (34): here we explore the metabolic reactions encoded by these metagenomic samples, using the same Illumina files as were used for the taxonomic data, and the MetaCyc database (31) as a reference.

Metabolic network inference

Using the enzyme and reaction data in MetaCyc, we defined a reaction-centric metabolic network where enzyme-catalyzed reactions are nodes, and any two nodes that share a metabolite are connected by an edge (left side of Figure 1). Translated metagenomic sequences were mapped to merged metabolic network for the microbes in the rumen (Figure 1). Approximately 8.9 million read pairs across the 16 animals mapped to one or more MetaCyc reactions (Table 1).

Animal VFA concentration and metabolic network structure have some association

VFAs make up approximately 70% of the dietary energy requirements for sheep (48) as biosynthetic precursors to compounds absorbed in the lower digestive tract (4). We were curious if the patterns of reads mapped to metabolic nodes/reactions would recapitulate any of the measured metabolite levels. Encouragingly, if unsurprisingly,

acetate, the most common VFA in the rumen (3), was also the VFA mostly commonly seen as a metabolite in the network. To make a more general analysis, we computed the Spearman's rank correlation between the measured concentration of each VFA in each animal and the total number of reads mapped to reactions involving that VFA in that same animal.

A significant correlation between read count and concentration was seen for propionate (Spearman's $\rho = 0.518$; $P=0.025$; Figure 2B), but not for acetate or butyrate (Spearman's $\rho = 0.332$, $P=0.11$, Figure 2A; and Spearman's $\rho = 0.175$, $P = 0.27$; Figure 2C; respectively). The lack of association between read count and acetate may be because acetate is also key metabolic intermediate and hence many reactions internal to the microbial cells involve it. Butyrate, on the other hand, is the rarest of the three metabolites both in concentration and in mapped reads, making detecting associations difficult.

Metabolic network structure varies by diet

The two diets differ in the relative density of reads mapped to each node. We fit log-normal distributions to the number of reads mapped to each node (Materials and Methods). There is a significant difference in the distribution mapped to the FORG and CONC animals ($P<10^{-10}$, likelihood ratio test with 5 degrees of freedom): more reads are mapped to each node in the CONC network than the FORG one (53.8 reads per node per 10^6 reads mapped and 41.4 reads per node for 10^6 reads mapped for CONC and FORG, respectively), with less overall spread in the number of reads mapped per node in the CONC animals (log-variance 2.7 and 2.9, respectively). Similarly, the per-node variation

in reads mapped within a diet group is less than that between diets for almost all nodes (Figure 3). Note that we are comparing these distributions for the two diets after read count normalization, meaning that a lower mean number of reads per node in one diet necessarily implies a higher variance in reads per node for that diet.

Figures 1B and 1C illustrate that the FORG animals show less animal-to-animal variation in the (normalized) number of reads mapped to a given node than do the CONC animals (Mann-Whitney-Wilcoxon test, $P<10^{-10}$). However, fewer total reads were mapped to nodes for the CONC animals, so it is possible that this difference in variation is due to smaller sample sizes. To test this possibility, for each animal, we randomly drew 100,000 of its mapped reads and then recomputed the Mann-Whitney-Wilcoxon test on 1000 of these resampled datasets. In all cases, the concentrate-fed animals still showed significantly greater variance ($P<10^{-9}$).

To link the host and microbial metabolic network, we defined three sets of *interface* metabolites that might be exchanged between the two networks (VFA, VFA+AA, and ALL; *Materials and Methods*). The VFA set contains the three volatile fatty acids used by ruminants as their primary energy source, VFA+AA adds to these the twenty amino acids; while ALL is a large set of metabolites defined based on human cellular metabolism (*Materials and Methods*). Host and microbial network nodes were connected if they both employed an interface metabolite (Figure 1). We then sorted the merged metabolic network based on the distance between each node in one subnetwork (host or microbe) and the nearest node in the other, resulting in the layered structure of Figure 1. Intuitively, reactions that involve interface metabolites are found in the innermost layer of each subnetwork (distance 0 from the other subnetwork): they

exchange compounds with the other network. The node color in the microbial network (right) indicates differential enzyme abundance between the two diets.

For the two larger sets of interface metabolites (VFA+AA and ALL) there is a general (though not invariable) negative correlation between the number of reads mapped to a reaction and that reaction's distance from the host subnetwork (Table 2), similar to the pattern we previously observed in a bovine system (21). There is no significant evidence that the proportion of nodes that differ between the diets differs by layer (Table 2). However, the layer number that an average read falls into is significantly higher for the FORG samples than for the CONC ones ($P<0.001$; Table 3), meaning that the FORG reads map more often to reactions distant from the host network than do reads from the CONC samples.

The network structure induced by the two diets also differs (Table 4). Reads from the CONC diet map more often to nodes with higher degree (e.g., larger number of edges) and higher betweenness-centrality (meaning that these nodes lie on "key paths" through the network; *Materials and Methods*). CONC reads also map to nodes with higher clustering coefficients. On the other hand, the mean number of carbon atoms involved in a node's reaction is significantly larger for the FORG reads ($P<0.001$ by network randomization, Table 4). These results are in accord with the observation that CONC reads generally map to layers closer to the host network than do FORG reads, since these nearer reactions are also more central in the network and hence involve smaller metabolites.

Diet differences are not driven by reaction presence/absence.

While the two groups of animals show many differences in the compositions of the microbial metabolic networks, these differences are not primarily driven by the absence of reactions in one group that are present in the other. Of the 2767 nodes (e.g., reactions) observed under either diet, 2254 were observed in both, while only 292 were specific to the forage-fed animals and 221 to the concentrate-fed ones. Moreover, most of these apparent differences are due to sampling effects: only 30/292 apparent differences for forage-fed animals and 47/221 of those in concentrate were significantly different and $P<0.01$ according to our previous test. Instead, most of these differentially present nodes simply represent very rare reactions: the mean number of reads mapped to nodes exclusive to FORG and CONC animals was 3.6 and 9.2, respectively, while the maximum number of reads mapped to a differentially present node was 318 (Figure 4).

Principal components analysis identifies metabolic differences in animals fed a concentrate diet

We used principal component analysis (PCA) to explore how diet interacts with the measured animal-to-animal variation in microbial taxa (OTUs) and reactions (nodes). For taxonomy, diet is the predominant driver of animal to animal differences (PC1 and PC2; Figure 5A), with these two components accounting for 92% of the total variance. However, in the case of the metabolic network nodes, most of the variation is accounted for by a PC that does not differentiate based on diet but rather that separates three animals fed a concentrate diet from the remaining 13 animals (Figure 5B, PC1). These three animals (identifiers 1220, 1239 and 1348) were three of the four CONC animals that

were inefficient in their growth relative to the mass of food consumed: in other words their residual feed intake (RFI) was high (Figure 5B) (34), and all showed very similar node profiles to each other with high variation in the number of reads mapped per node (e.g., had some nodes with very large numbers of mapped reads and others with many fewer reads mapped).

Concentrate-fed animals show large pairwise distances between each other in both taxonomy and metabolic network

We sought to further explore this apparent difference between the two diets in the inter-diet variability with a pairwise distance analysis. For both the number of reads mapped to nodes and to OTUs, the FORG animals show small and relatively uniform pairwise differences, although these differences are still larger than can be explained by sampling differences ($P<0.001$). For the CONC animals there was a visually much wider spread of distances both for the OTUs and for the nodes ($P<0.001$; Figure 5D). Interestingly, the high RFI animals showed relatively lower pairwise differences than did other comparisons. On the other hand, some of the pairwise differences between low RFI (high efficiency) CONC animals were as large as for pairs of animals fed differing diets (Figure 5D).

Discussion

Our goal in this project was to study how differing diets altered rumen microbe metabolism by comparing the metabolic networks of ruminal microbes from animals fed a concentrate diet to that of animals fed a forage diet. We initially expected to find that, despite the wide variation in microbial taxonomy seen with differing diets (21,34), the metabolic properties of these communities would remain roughly similar. However, we found numerous differences between nodes of the FORG and CONC diets.

There were several major differences in the metabolic networks of the FORG and CONC animals. These differences manifested themselves, however, not as presence or absence of particular reactions but rather as large differences in abundance for particular reactions between diets. We found that grain-based diets were dominated by a smaller set of reactions that were closer to the host's metabolic network, were more centrally located in the network, and that used substrates with shorter carbon backbones (Tables 3 and 4).

Distance from host network:

The animals on the CONC diet were employing reactions, and hence metabolites, that were closer to those that the host could use directly than were the animals on the FORG diet. This difference makes sense, as the forage diet was high in plant cellulose, which requires more reactions to break down than does the more accessible, concentrate feed (Table 3). It also demonstrates that our metabolic samples display at least some predictable patterns in their relationship to the host network, and are not merely random samples of the database.

Network centrality:

On average, the forage-fed animals had microbial reactions that used more carbon atoms than did the concentrate-fed animals. Forage-fed animals also had reactions that were, on average, of lower degree in the network and with lower betweeness-centrality. This also makes sense, as the concentrate diet is expected to employ compounds that are closer to the host's metabolism (Table 4).

An alternative possibility for the higher centrality of the concentrate network is that the reactions for the concentrate diet are better represented in our database. Metacyc itself is a sampling of the well-studied reactions, and those reactions necessarily tend to be metabolically central. In addition, there may be a bias towards the reactions of central-metabolism in the literature because the metabolites involved are well-defined. In contrast, even the well-studied reaction of the decomposition of cellulose is difficult to represent in a computer because it involves an indeterminate number of glucose molecules linked together. In fact, we believe that both this bias in the database and true signals of diet differences contribute to the observed differences of Figure 1D. Two facts speak against the idea that bias alone explains our results: first, the database used is the same for both diets, and second more reads were actually mapped to reactions for the FORG diet (Table 1).

Variability among animals within each diet

Our CONC animals were less similar to each other than were the FORG animals (Figure 5). It is tempting to ascribe this difference to the fact that the FORG diet is more similar to the evolved diet of these animals. And indeed, while foregut fermentation has

convergently evolved in herbivores multiple times (54,55), ruminants represent the most complete and dramatic adaptation to allow efficient digestion of fibrous plant diets (56). Nonetheless, we believe that an equally likely explanation lies in the structure of our feeding experiment. The animals were all originally on a forage-like diet, so it is possible that the switch to the CONC diet allowed more animal-to-animal variation to arise. Effectively the microbes in different host animals "discovered" different solutions to the same metabolic problem. In addition, some of the CONC animals reached a relatively efficient metabolic state after the diet change, while others switched to a more degenerate state where they were unable to use the nutrients in the diet as efficiently. This difference is apparent in our principle component analysis (PCA; Figure 5B), where three high residual feed intake (RFI) animals (e.g., the inefficient animals) on a concentrate diet grouped outside the other 13 animals studied. The larger variability in the concentrate diet can also be seen in the distribution of pairwise distances (Figure 5D), which also illustrates that this pattern is in fact not confined to the three high RFI animals above but is a general pattern. We hesitate to over-interpret these efficiency data because the sample size is quite small. It is nonetheless interesting that the low RFI (more feed efficient) animals appear to also have a more efficient rumen-host interface (Figure 5D).

Implications

A better understanding of the rumen microbiome has a number of interesting implications, both for ecology and organismal biology in general and for a number of practical problems. On the practical side, better matching of feed and microbial communities has the potential for creating improved products such as milk and meat with

higher levels of conjugated linoleic acid (57,58), improved feed efficiency (6,15,59), improvements in animal health and welfare (60), and in benefits for the environment, such as by reducing the amount of methane produced by the beef and dairy industry (61). Beyond agricultural concerns, an enhanced understanding of the metabolic properties of microbial environments may assist with alternative fuel source creation, improvements in soil microbiology and ecology, and environmentally friendly quick composting techniques for landfill reductions (8,32,62,63). There are also numerous implications for human health in relation to the microbiome, such as the study of obesity and inflammatory bowel disease (IBD) as related to the human gut microbiome (18,24,26,27,64). These ideas are only a small sample of positive outcomes that may result from having a more complete understanding of rumen microbial-host metabolic relationships.

Pitfalls and potential improvements

It is important to note that because we used a shotgun sequencing approach, our methods assemble a profile of microbial metabolic potential, but not necessarily one that measures the relative intracellular levels of the various enzymes. It would be interesting to attempt a similar analysis using RNA sequencing data, where we could more accurately assess which enzymes in the network are actually highly expressed. In addition, as is expected for any study employing a database, there is a bias in our results towards the information contained within the database: in our case toward the consideration of known metabolic enzymes (8). It is also possible that the variation observed between samples might be due to differing locations of rumen sampling.

However, previous work (65) suggests that the effect of location is much smaller than the effect of a different host animal. Since the major open question from this study is whether all of the CONC effects observed are due to the fact of the diet shift, rather than its nature, further analyses of both the time-course of adaptation to a new diet and of how the diet selected alters that time course, would be desirable. The correlations between metabolite levels and microbial read counts in the VFA analysis were relatively weak, mostly likely because of the small number of host animals surveyed. Larger animal samples, as well as measurements of other metabolites, such as methane, should provide even greater insight into the metabolic structure of this ecosystem.

Conclusion

In this study we illustrate how a metabolic network perspective offers new insights into how the microbiome of ruminant animals responds to changes in diet. Network perspectives expand our understanding of such microbial ecosystems in several ways. The analysis of metabolic data adds a new axis to comparisons between diets, hosts or other experimental or natural factors. More importantly, the network approach allows for hypothesis development and testing, moving the field beyond the descriptive and correlative.

Table 1: Read Statistics by Animal

Animal ID	Diet	Total reads[a]	# (%) reads that hit to nodes[b]	# (%) reads hit to sheep genome[c]	#OTU found[d]
1003	FORG	16,779,099	271,571 (1.62%)	58,830 (0.35%)	109
1009	FORG	35,930,923	761,728 (2.12%)	1,855,867 (5.17%)	161
1127	FORG	41,120,479	570,177 (1.39%)	63,121 (0.15%)	137
1208	FORG	44,823,544	1,012,146 (2.26%)	18,155 (0.04%)	140
1248	FORG	22,698,997	673,066 (2.97%)	2,804 (0.01%)	127
1366	FORG	18,124,115	396,240 (2.19%)	42,582 (0.23%)	119
1397	FORG	32,221,706	657,645 (2.04%)	62,527 (0.19%)	137
7505	FORG	47,234,706	651,061 (1.38%)	1,175,563 (2.49%)	177
1026	CONC	29,835,213	642,601 (2.15%)	7,564 (0.03%)	108
1101	CONC	54,927,600	1,008,933 (1.84%)	44,188 (0.08%)	142
1111	CONC	26,710,771	362,506 (1.36%)	167,325 (0.63%)	137
1220	CONC	7,800,938	109,479 (1.40%)	1,226 (0.02%)	75
1239	CONC	42,216,924	751,236 (1.78%)	24,083 (0.06%)	138
1348	CONC	13,577,697	245,424 (1.81%)	12,233 (0.09%)	102
1396	CONC	18,274,753	278,963 (1.53%)	7,330 (0.04%)	124
7429	CONC	30,236,139	553,079 (1.83%)	65,498 (0.22%)	135

a: Total paired reads sequenced prior to quality filtering
b: Number and percent of total reads that were mapped to nodes according to our criteria (*Materials and Methods*).
c: Number and percent of total reads that mapped to the sheep genome at 80% percent identity.
d:Number of distinct OTUs identified previously in these sequences (34).

Table 2: Read count/Network Level Correlation

		Read count vs. level (FORG)		Read count vs level (CONC)		Proportion of differential reads and level	
Group[a]	**Currency Cutoff[b]**	**Pearson's *r*[c]**	**P-value[d]**	**Pearson's *r*[c]**	**P-value[d]**	**Pearson's *r*[e]**	**P-value[d]**
VFA	N_{25}	0.230	0.576	0.219	0.549	**0.635**	**0.033**
	N_{50}	-0.166	0.105	-0.200	0.085	-0.197	0.700
	N_{100}	-0.199	0.068	**-0.276**	**0.035**	-0.221	0.324
VFA_AA	N_{25}	**-0.461**	**0.012**	**-0.619**	**0.003**	0.060	0.413
	N_{50}	**-0.255**	**0.006**	**-0.446**	**0.000**	0.287	0.325
	N_{100}	**-0.281**	**0.006**	**-0.461**	**0.000**	-0.039	0.322
ALL	N_{25}	0.168	0.364	-0.037	0.162	0.465	0.222
	N_{50}	**-0.506**	**0.002**	**-0.595**	**0.000**	-0.476	0.604
	N_{100}	**-0.425**	**0.017**	**-0.580**	**0.001**	-0.620	0.817

a: Interface metabolite set (*Materials and Methods*)
b: Network (e.g., currency cutoff; *Materials and Methods*)
c: Correlation of the number of reads mapped to a layer and that layer's distance to the host network (i.e., negative correlations imply fewer reads mapped to layers more distant from the host). Calculated for FORG and CONC, respectively.
d: *P*-value for the test of the hypothesis that the correlation in read count or in differential reads is larger than can be explained by chance; assessed by randomization of layer numbers with respect to read counts. Bold values are significant at $P=0.05$.
e: Correlation of the proportion of nodes with differential read abundance (*Materials and Methods*) between FORG and CONC animals verses the nodes' layer number.

Table 3: Diet and network position

Group[a]	Currency Cutoff[b]	Mean layer: FORG[c]	Mean layer: CONC[c]	Real Difference: mean layers[d]	*P*[e]
VFA	N_{25}	1.99	1.82	0.17	< **0.001**
	N_{50}	1.87	1.74	0.13	< **0.001**
	N_{100}	1.78	1.64	0.14	< **0.001**
VFA_AA	N_{25}	0.828	0.66	0.16	< **0.001**
	N_{50}	0.87	0.75	0.12	< **0.001**
	N_{100}	0.91	0.80	0.11	< **0.001**
ALL	N_{25}	0.51	0.45	0.06	< **0.001**
	N_{50}	0.46	0.41	0.04	< **0.001**
	N_{100}	0.50	0.47	0.04	< **0.001**

a: Interface metabolite set (*Materials and Methods*)
b: Network (e.g., currency cutoff; *Materials and Methods*)
c: Mean layer number for the reads mapped from FORG or CONC animals, respectively.
d: Difference between the mean layer for FORG and CONC
e: *P*-value for the test of the hypothesis that the two diets do not differ in mean layer. For this test, reads were randomly reassigned to diets and the mean layers recomputed 1000 times (*Materials and Methods).* Values significant at *P*=0.05 shown in bold.

Table 4: Network structure and diet

Statistic[a]	Network[b]	FORG[c]	CONC[c]	Real Difference (FORG-CONC) [d]	Mean Random Difference[e]	*P*[f]
Carbon Sum	N_{25}	33.225	32.438	0.787	0.002	**< 0.001**
	N_{50}	33.225	32.438	0.787	0.001	**< 0.001**
	N_{100}	33.225	32.438	0.787	0.001	**< 0.001**
Betweenness Centrality	N_{25}	9118.7	9423.4	304.7	1.069	**< 0.001**
	N_{50}	18318.2	20949.1	2630.9	2.148	**< 0.001**
	N_{100}	18542.9	22280.6	3737.6	1.100	**< 0.001**
Degree	N_{25}	3.544	3.622	0.078	<0.001	**< 0.001**
	N_{50}	9.172	9.316	0.143	<0.001	**< 0.001**
	N_{100}	12.154	12.984	0.830	0.001	**< 0.001**
Clustering Coefficient	N_{25}	0.882	0.886	0.004	<0.001	**< 0.001**
	N_{50}	0.859	0.862	0.003	<0.001	**< 0.001**
	N_{100}	0.846	0.847	0.001	<0.001	**< 0.001**

a: Network statistic compared between diet (*Materials and Methods*). Carbon sum: Total number of carbon atoms involved in a reaction. Betweenness-centrality: Number of shortest paths crossing through a node. Degree: Number of edges of a node. Clustering coefficient: Number of fully connected triangles a node participates in over the total number of edges.
b: Network (e.g., currency cutoff; *Materials and Methods*)
c: Mean value per read for the selected network statistic for FORG or CONC animals, respectively.
d: Difference between the mean statistic value for FORG and CONC
e: Mean difference between the two diets when reads are randomly assigned to the two diets.
f: *P*-value for the test of the hypothesis that the two diets do not differ in mean statistic. For this test, reads were randomly reassigned to diets and the network statistics recomputed 1000 times (*Materials and Methods).* Values significant at *P*=0.05 shown in bold.

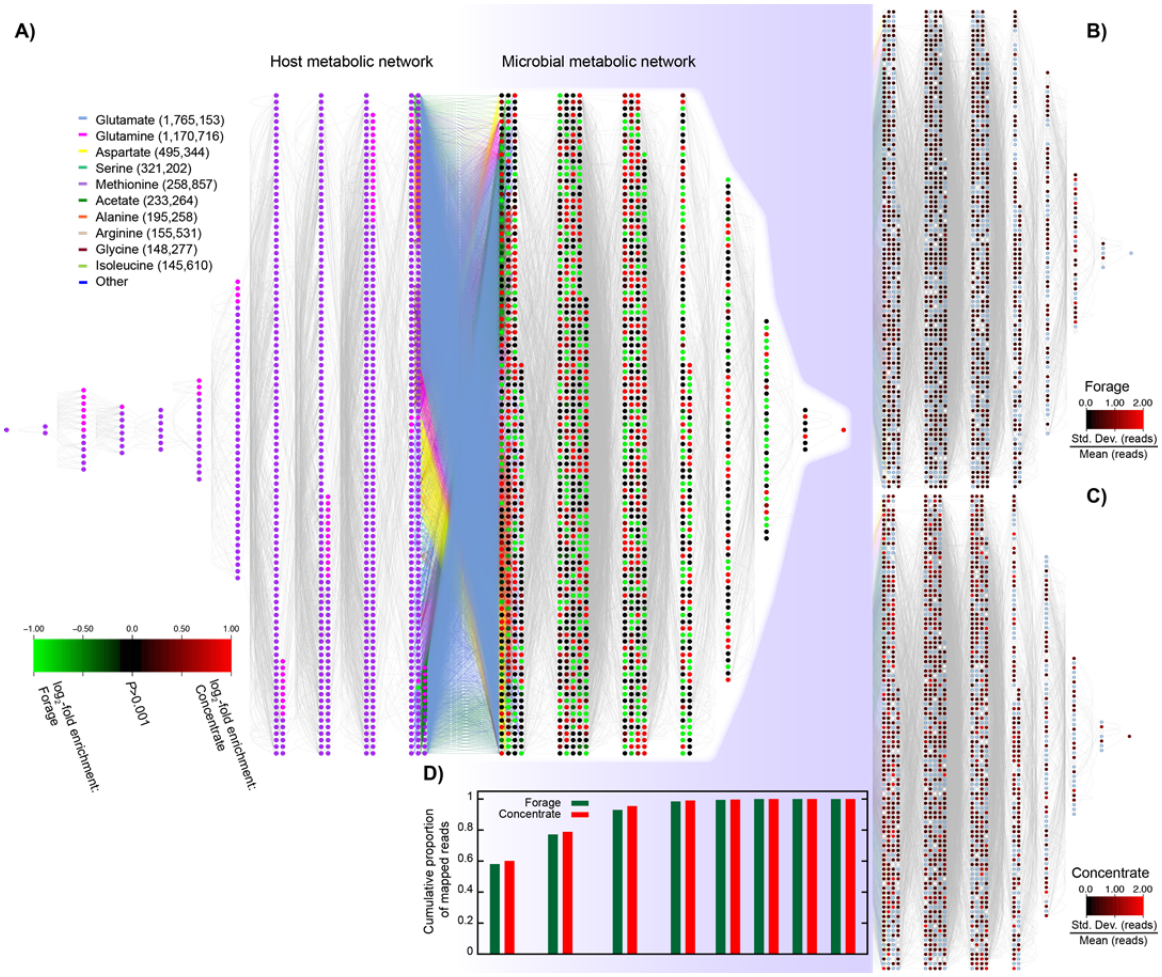

Figure 1: Merged host/microbe metabolic network. **A)** Each node (circle) is a reaction in the host genome (left) or microbial metagenomes (right). Host nodes are colored purple if derived from an orthology association between sheep and humans, magenta if from an ovine/bovine relationship or green for the case of the added buyrate-employing pseudo-reaction (Materials and Methods). Edges are shared metabolites (network N_{50}). In the center are 23 potentially shared compounds between the host and microbes (set VFA+AA; Materials and Methods): the 10 most frequent metabolites (by microbe read count to their respective reactions) are individually colored. Nodes are organized by their distance from the other subnetwork: hence nodes employing an interface metabolite are at the center with a distance of 0. Microbial nodes are colored based on the normalized $\log_2$-fold difference in read count between the two diets (green: overabundant in the FORG diet; red: overabundant in CONC). Nodes whose normalized read counts did not differ significantly between the diets are shown in black (Materials and Methods). **B)** The right half of part **A**, recolored based on the normalized animal-to-animal variance in read count for the FORG animals (see scale bar). **C)** Same as for **B** but for the CONC animals. **D)** Histogram giving the cumulative proportion of the total mapped reads for the two diets (green: FORG; red: CONC) at each level of the network. Note that the FORG animals have proportionally more reads mapped to more distant layers of the network.

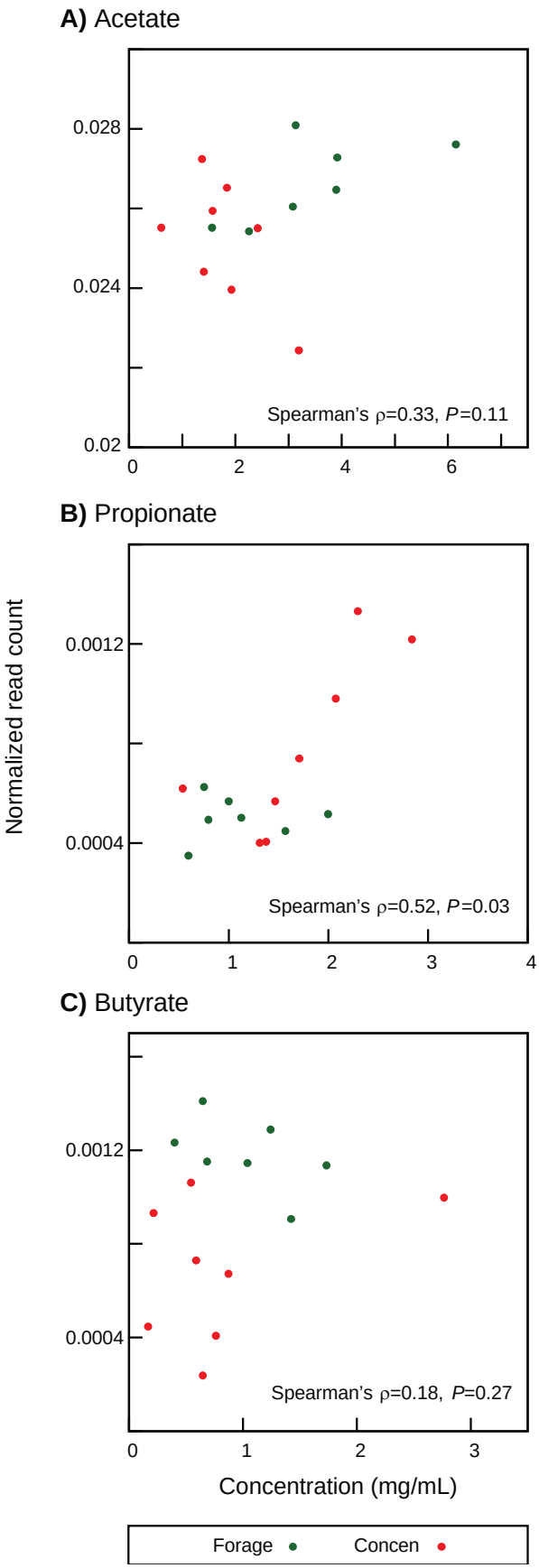

Figure 2: Association of the concentrations of three volatile fatty acids (VFAs) and of reads mapping to reactions involving them. On *x* is the concentration (mg/ml) of the VFA, on *y* is the normalized read count from reactions employing that metabolite (e.g., proportion of total mapped reads involving that metabolite). **A)** Acetate, **B)** Propionate, **C)** Butyrate.

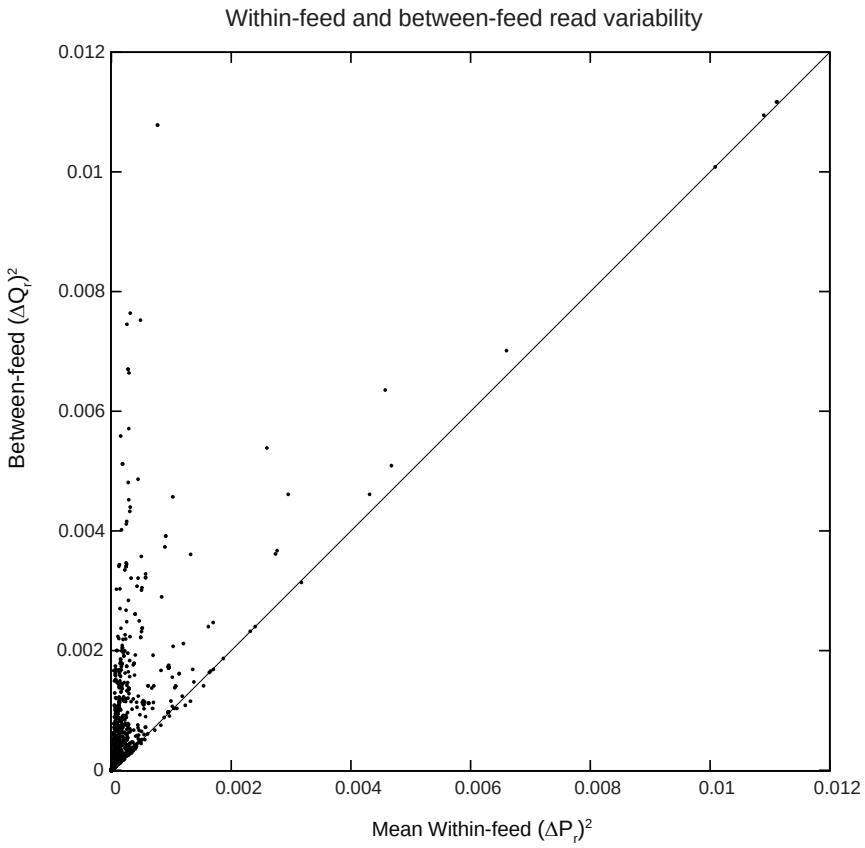

Figure 3: Within-feed animal to animal variability is less than the variability between the two feeds. For each metabolic network node, we calculated P_r for each animal: the proportion of the total mapped reads assigned to that node. We next calculated the variability between animals fed the same diet (*x*-axis) as:

$$(\Delta P_r)^2 = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(P_i - P_j\right)^2 \Big/ \left(\frac{\sum_{i=0}^{m} P_i}{m} \cdot \frac{n(n-1)}{2}\right)$$

Where *n* is the number of animals in a particular feed-group and *m* is the total number of animals. The quantity plotted is thus the mean squared difference between animals in the proportion of reads mapped to that node for a particular feed, divided by the mean proportion of reads mapped to that node across both feeds. We averaged this value for the two feeds to obtain the values on the *x*-axis above. Similarly, we calculated the mean variability between animals fed different diets for the *y*-axis:

$$(\Delta Q_r)^2 = \sum_{i=0}^{n} \sum_{j=0}^{n} \left(P_i - Q_j\right)^2 \Big/ \left(\frac{\sum_{i=0}^{m} P_i}{m} \cdot n^2\right)$$

Where P_i is the proportion of reads mapped in animal *i* from the first feed and Q_j is the proportion of animals mapped in animal *j* of the second group. Thus, (ΔQ_r)^2 is the variability in proportion of mapped reads between the two feeds, normalized by the average proportion of reads mapped for that node. A line of $y=x$ is shown for reference: as is clear, it is almost invariably the case that there is more variability between the two diets than within a diet, as would be expected if the two diets differ in the types of enzymes needed to digest them.

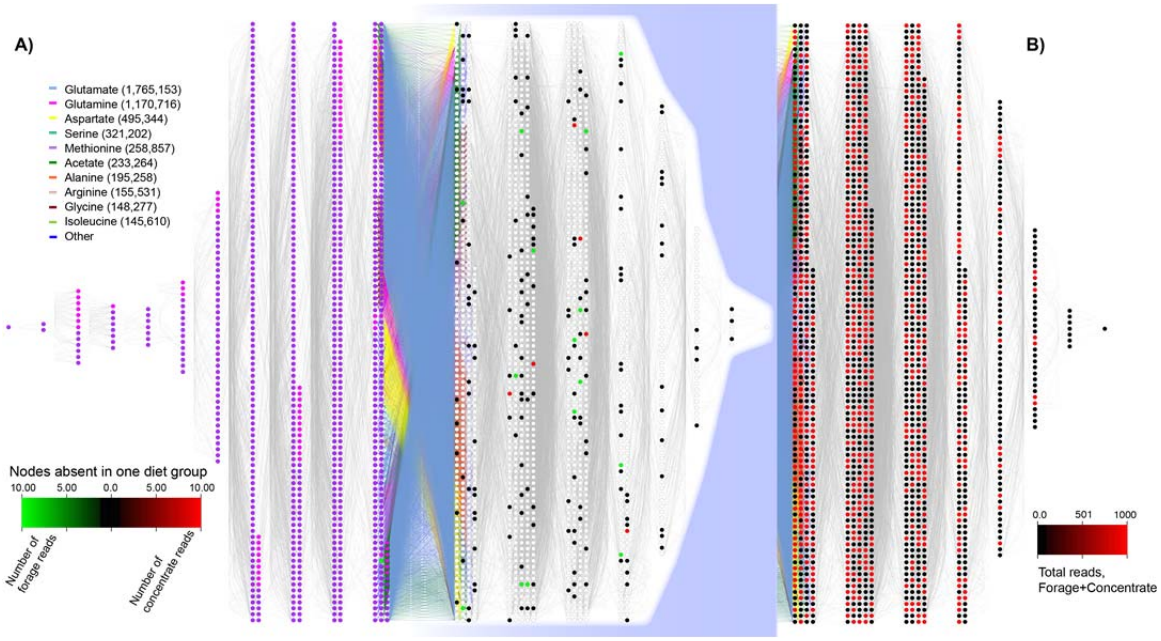

Figure 4: Node presence/absence does not explain the differences between the FORG and CONC diets. **A)** Shown are all of the nodes present in FORG and absence in CONC (green) or vice-versa (red). The color intensity is proportion to the total number of reads seen in the diet where the node is present. Note relatively low number of reads that result in a fully colored node (lower left). The maximum of reads mapped to any node in this panel is 318. **B)** For comparison, the same network topology with the total number of reads mapped across both diets for all nodes. Note the much greater range of read counts in this version of the network (lower right).

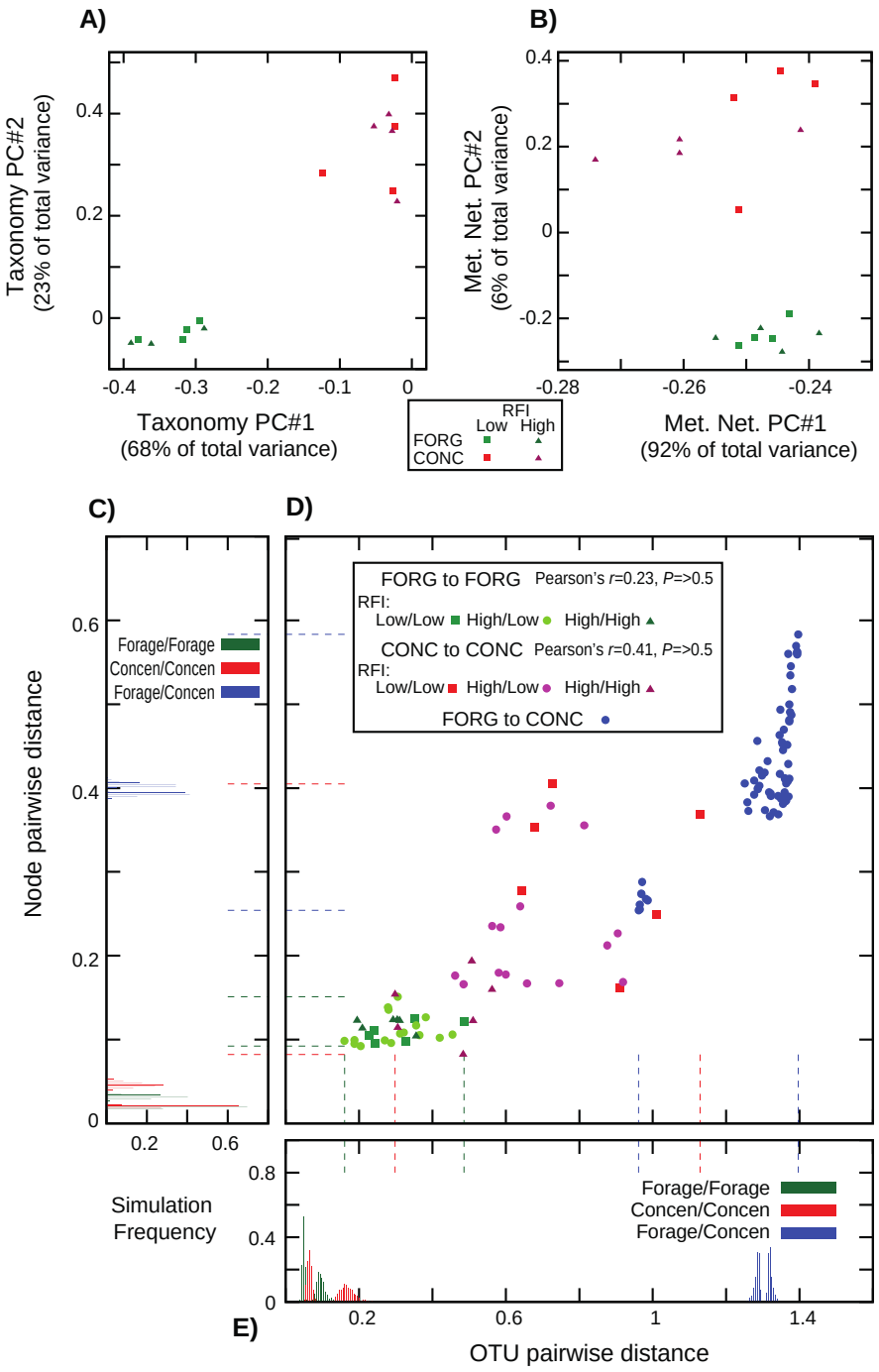
A)
B)
Taxonomy PC#2
(23% of total variance)
Taxonomy PC#1
(68% of total variance)
Met. Net. PC#2
(6% of total variance)
Met. Net. PC#1
(92% of total variance)
RFI
Low High
FORG
CONC
C)
D)
Node pairwise distance
Forage/Forage
Concen/Concen
Forage/Concen
FORG to FORG Pearson's r=0.23, P=>0.5
RFI:
Low/Low High/Low High/High
CONC to CONC Pearson's r=0.41, P=>0.5
RFI:
Low/Low High/Low High/High
FORG to CONC
Simulation
Frequency
E)
OTU pairwise distance

Figure 5: Animal-to-animal taxonomic and network differences. **A)** Principle component analysis of the OTU distributions across the 16 animals. The first two principle components (PCs) are shown, comprising 92% of the total variance. FORG animals are shown in green and CONC in red. Visually, it seems clear that the diet difference explains most of the variation in OTU distributions. **B)** Principle component analysis of the distribution of reads mapped to metabolic network nodes. The first two principle components (PCs) are shown, comprising 98% of the total variance. However, diet is no longer the main source of variation. Instead, principle component 1 separates three CONC animals (numbers 1220, 1239 and 1348; high RFI) from the rest of the dataset. Inspection of the node-level data suggests that these three animals are unusual in that they have higher than usual node-to-node variation in the number of mapped reads (namely a few nodes with a large number of mapped reads) and are also highly correlated with each other, unlike some of the other CONC animals with rather different profiles. **C)** Pairwise differences in distribution of reads mapped to OTUs (*x*-axis) and nodes (*y*-axis). FORG to FORG comparisons are shown in green, CONC to CONC in red, and FORG to CONC in blue. For each animal, a vector representing all mapped reads was normalized to unit length and then standard Euclidian distances computed for it and all other animals (Materials and Methods). For each of the three groups we computed the Pearson's correlation of OTU and node distance and compared that value to that seen from randomized datasets (Materials and Methods). **D)** Minimum and maximum pairwise OTU distances seen when reads were randomly and proportionally reassigned to each animal. On the *x*-axis is the same distance scale as *x* in **C**, on the *y*-axis is the proportion of simulations with a given minimum/maximum (Materials and Methods). The color scheme is as for **C.** Dashed lines give the minimums and maximums seen in the real data of **C**. **E)** As for **D**, except with the node distances.

References

1. Stevens, C.E. and Hume, I.D. (1998) Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. *Physiological Reviews*, **78**, 393-427.
2. Mackie, R.I. (2002) Mutualistic fermentative digestion in the gastrointestinal tract: diversity and evolution. *Integr Comp Biol*, **42**, 319-326.
3. Russell, J.B. and Rychlik, J.L. (2001) Factors that alter rumen microbial ecology. *Science*, **292**, 1119-1122.
4. Hooper, L.V., Midtvedt, T. and Gordon, J.I. (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual review of nutrition*, **22**, 283-307.
5. Kim, M., Morrison, M. and Yu, Z. (2011) Status of the phylogenetic diversity census of ruminal microbiomes. *FEMS Microbiol Ecol*, **76**, 49-63.
6. Guan, L.L., Nkrumah, J.D., Basarab, J.A. and Moore, S.S. (2008) Linkage of microbial ecology to phenotype: correlation of rumen microbial ecology to cattle's feed efficiency. *FEMS Microbiol Lett*, **288**, 85-91.
7. Kobayashi, Y. (2006) Inclusion of novel bacteria in rumen microbiology: need for basic and applied science. *Animal science journal*, **77**, 375-385.
8. Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T. *et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463-467.
9. Morgavi, D.P., Kelly, W.J., Janssen, P.H. and Attwood, G.T. (2013) Rumen microbial (meta)genomics and its application to ruminant production. *animal*, **7**, 184-201.
10. Papadopoulos, D., Schneider, D., Meier-Eiss, J., Arber, W., Lenski, R.E. and Blot, M. (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences*, **96**, 3807-3812.
11. Deng, Y. and Fong, S.S. (2011) Laboratory evolution and multi-platform genome re-sequencing of the cellulolytic actinobacterium Thermobifida fusca. *Journal of Biological Chemistry*, **286**, 39958-39966.
12. Raes, J. and Bork, P. (2008) Molecular eco-systems biology: towards an understanding of community function. *Nature reviews. Microbiology*, **6**, 693-699.
13. Ley, R.E., Peterson, D.A. and Gordon, J.I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, **124**, 837-848.
14. Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R. *et al.* (2008) Evolution of Mammals and Their Gut Microbes. *Science*, **320**, 1647-1651.
15. Myer, P.R., Smith, T.P.L., Wells, J.E., Kuehn, L.A. and Freetly, H.C. (2015) Rumen Microbiome from Steers Differing in Feed Efficiency. *PLoS ONE*, **10**, e0129174.
16. Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M. and Nelson, K.E. (2006)

Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355-1359.
17. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.
18. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A. and Affourtit, J.P. (2009) A core gut microbiome in obese and lean twins. *nature*, **457**, 480-484.
19. Human Microbiome Project Consortium, T. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207-214.
20. Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., Giannoukos, G., Boylan, M.R., Ciulla, D. and Gevers, D. (2014) Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences*, **111**, E2329-E2338.
21. Taxis, T.M., Wolff, S., Gregg, S.J., Minton, N.O., Zhang, C., Dai, J., Schnabel, R.D., Taylor, J.F., Kerley, M.S., Pires, J.C. *et al.* (2015) The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Research*.
22. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F. and Yamada, T. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59-65.
23. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554-557.
24. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027-1031.
25. Greenblum, S., Chiu, H.-C., Levy, R., Carr, R. and Borenstein, E. (2013) Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Current opinion in biotechnology*, **24**, 810-820.
26. Greenblum, S., Turnbaugh, P.J. and Borenstein, E. (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, **109**, 594-599.
27. Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M. and Henrissat, B. (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology*, **8**, e1002358.
28. Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends in microbiology*, **14**, 257-263.
29. Weimer, P.J. (1998) Manipulating ruminal fermentation: a microbial ecological perspective. *Journal of Animal Science*, **76**, 3114-3122.
30. Karlsson, F.H., Nookaew, I., Petranovic, D. and Nielsen, J. (2011) Prospects for systems biology and modeling of the gut microbiome. *Trends in biotechnology*, **29**, 251-258.

31. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, **42**, D459-471.
32. Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P. *et al.* (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A*, **106**, 1948-1953.
33. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y. and Shen, D. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55-60.
34. Ellison, M.J., Conant, G.C., Cockrum, R.R., Austin, K.J., Truong, H., Becchi, M., Lamberson, W.R. and Cammack, K.M. (2014) Diet alters both the structure and taxonomy of the ovine gut microbial ecosystem. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **21**, 115-125.
35. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186-194.
36. Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, **37**, D141-145.
37. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
38. Powell, A.J., Conant, G.C., Brown, D.E., Carbone, I. and Dean, R.A. (2008) Altered patterns of gene duplication and differential gene gain and loss in fungal pathogens. *BMC Genomics*, **9**, 147.
39. Bekaert, M. and Conant, G.C. (2011) Copy number alterations among mammalian enyzmes cluster in the metabolic network. *Molecular Biology and Evolution*, **28**, 1111-1121.
40. Conant, G.C. (2009) Neutral evolution on mammalian protein surfaces. *Trends in Genetics*, **25**, 377-381.
41. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G. and Fitzgerald, S. (2014) Ensembl 2014. *Nucleic acids research*, **42**, D749-D755.
42. Hudson, C.M. and Conant, G.C. (2011) Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. *BMC Evolutionary Biology*, **11**, 89.
43. Pérez-Bercoff, Å., McLysaght, A. and Conant, G.C. (2011) Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Molecular BioSystems*, **7**, 3056-3064.
44. Doring, A., Weese, D., Rausch, T. and Reinert, K. (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

45. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195-197.
46. R Development Core Team. (2008) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
47. Cammack, K.M., Leymaster, K.A., Jenkins, T.G. and Nielsen, M.K. (2005) Estimates of genetic parameters for feed intake, feeding behavior, and daily gain in composite ram lambs. *J Anim Sci*, **83**, 777-785.
48. Bergman, E. (1990) Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiological reviews*, **70**, 567-590.
49. Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R. and Palsson, B.Ø. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences, U.S.A.*, **104**, 1777-1782.
50. Dijkstra, E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269-271.
51. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.
52. Hahn, M.W. and Kern, A.D. (2005) Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Molecular Biology and Evolution*, **22**, 803-806.
53. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.
54. Kornegay, J.R., Schilling, J.W. and Wilson, A.C. (1994) Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. *Molecular Biology and Evolution*, **11**, 921-928.
55. Stewart, C.-B., Schilling, J.W. and Wilson, A.C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, **330**, 401-404.
56. Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R. and Gordon, J.I. (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, **6**, 776-788.
57. Or-Rashid, M.M., Odongo, N.E. and McBride, B.W. (2007) Fatty acid composition of ruminal bacteria and protozoa, with emphasis on conjugated linoleic acid, vaccenic acid, and odd-chain and branched-chain fatty acids1. *Journal of Animal Science*, **85**.
58. Huws, S.A., Kim, E.J., Cameron, S.J.S., Girdwood, S.E., Davies, L., Tweed, J., Vallin, H. and Scollan, N.D. (2015) Characterization of the rumen lipidome and microbiome of steers fed a diet supplemented with flax and echium oil. *Microbial Biotechnology*, **8**, 331-341.
59. Pitta, D.W., Indugu, N., Kumar, S., Vecchiarelli, B., Sinha, R., Baker, L.D., Bhukya, B. and Ferguson, J.D. (2016) Metagenomic assessment of the functional potential of the rumen microbiome in Holstein dairy cows. *Anaerobe*, **38**, 50-60.
60. Lee, W.-J. and Hase, K. (2014) Gut microbiota-generated metabolites in animal health and disease. *Nat Chem Biol*, **10**, 416-424.

61. Denman, S.E. and McSweeney, C.S. (2015) The Early Impact of Genomics and Metagenomics on Ruminal Microbiology. *Annual Review of Animal Biosciences*, **3**, 447-465.
62. Weimer, P.J. and Kohn, R.A. (2016) Impacts of ruminal microorganisms on the production of fuels: how can we intercede from the outside? *Applied Microbiology and Biotechnology*, **100**, 3389-3398.
63. Cheng, F., Sheng, J., Dong, R., Men, Y., Gan, L. and Shen, L. (2012) Novel Xylanase from a Holstein Cattle Rumen Metagenomic Library and Its Application in Xylooligosaccharide and Ferulic Acid Production from Wheat Straw. *Journal of Agricultural and Food Chemistry*, **60**, 12516-12524.
64. Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. and Gordon, J.I. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A*, **102**, 11070-11075.
65. Ross, E.M., Moate, P.J., Bath, C.R., Davidson, S.E., Sawbridge, T.I., Guthridge, K.M., Cocks, B.G. and Hayes, B.J. (2012) High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. *BMC genetics*, **13**, 53.