Universiteit
Leiden
The Netherlands

Impact of nitrogen fertilization on the soil microbiome and nitrous oxide emissions

Cassman, N.A.

Citation
Cassman, N. A. (2019, April 17). *Impact of nitrogen fertilization on the soil microbiome and nitrous oxide emissions*. Retrieved from https://hdl.handle.net/1887/71732

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/71732

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/71732 holds various files of this Leiden University dissertation.

Author: Cassman, N.A.
Title: Impact of nitrogen fertilization on the soil microbiome and nitrous oxide emissions
Issue Date: 2019-04-17

Chapter 6

Genome-resolved metagenomics of sugarcane vinasse bacteria

Noriko A. Cassman, Kesia S. Lourenco, Janaina B. do Carmo, Heitor Cantarella and Eiko E. Kuramae

Published as:

Cassman NA, Lourenco KS, do Carmo JB, Cantarella H & Kuramae EE, 2018. "Genome-resolved metagenomics of sugarcane vinasse bacteria." *Biotech for Biofuels*. 11(1): 48.

Abstract

Background: The production of 1L of ethanol from sugarcane generates up to 12 L of vinasse, which is a liquid waste containing an as-yet uncharacterized microbial assemblage. Most vinasse is destined for use as a fertilizer on the sugarcane fields because of the high organic and K content; however, increased N_2O emissions have been observed when vinasse is co-applied with inorganic N fertilizers. Here we aimed to characterize the microbial assemblage of vinasse to determine the gene potential of vinasse microbes for contributing to negative environmental effects during fertirrigation and/or to the obstruction of bioethanol fermentation.

Results: We measured chemical characteristics and extracted total DNA from six vinasse batches taken over 1.5 years from a bioethanol and sugar mill in Sao Paulo State. The vinasse microbial assemblage was characterized by low alpha diversity with 5 to 15 species across the six vinasses. The core genus was *Lactobacillus*. The top six represented bacterial genera across the samples were *Lactobacillus*, *Megasphaera* and *Mitsuokella* (Phylum Firmicutes, 35 – 97% of sample reads); *Arcobacter* and *Alcaligenes* (Phylum Proteobacteria, 0 – 40%); *Dysgonomonas* (Phylum Bacteroidetes, 0 – 53%); and *Bifidobacterium* (Phylum Actinobacteria, 0 – 18%). Potential genes for denitrification but not nitrification were identified in the vinasse metagenomes, with putative *nir*K and *nos*Z genes the most represented. Binning resulted in 38 large bins with between 36.0 and 99.3% completeness, and five small mobile element bins. Of the large bins, 53% could be classified at the phylum level as Firmicutes, 15% as Proteobacteria, 13% as unknown phyla, 13% as Bacteroidetes and 6% as Actinobacteria. The large bins spanned a range of potential denitrifiers; moreover, the genetic repertoires of all the large bins included the presence of genes involved in acetate, CO_2, ethanol, H_2O_2, and lactose metabolism; for many of the large bins, genes related to the metabolism of mannitol, xylose, butyric acid, cellulose, sucrose, “3-hydroxy fatty acids and antibiotic resistance genes were present. In total, 21 vinasse bacterial draft genomes were submitted to the genome repository.

Conclusions: Identification of the gene repertoires of vinasse bacteria and assemblages supported the idea that microbiological variation of vinasse might lead to varying patterns of N_2O emissions during fertirrigation. Furthermore, we uncovered draft genomes of novel strains of known bioethanol contaminants, as well as draft genomes unknown at the phylum level. This study will aid efforts to improve bioethanol production efficiency and sugarcane agriculture sustainability.

6.1 Introduction

Sao Paulo State contains a total of 5.7 million hectares of land planted with sugarcane. These fields supply the input for Brazil's large bioethanol industry, which is the second largest producer of bioethanol worldwide (UNICA). Brazil has more than 300 sugarcane processing plants, including sugar mills (producing only sugar), mills with distillery plants (sugar and ethanol production), and independent distilleries (only ethanol production). In the 2013/2014 season, the total ethanol production was 13.9 thousand m^3 (UNICA, 2013/2014 harvest). The major by-product of sugarcane ethanol production is vinasse; up to 12 L of vinasse is generated per liter of ethanol [1]. Sugarcane vinasse consists of water (about 93%) and organic compounds, and contains K, Ca and Mg, though the amount of these components depends on the characteristics of the input sugarcane and subsequent processing steps [2]. The major organic components of sugarcane vinasse are low molecular-weight organic compounds, mainly glycerol, lactic acid, ethanol, and acetic acid [3]. In general, vinasse has a low pH of around 4 and high chemical oxygen demand of 100 - 500 g per L.

The large volumes of vinasse and its chemical properties of high organic and K content have led to its widespread reuse as a fertilizer supplement for sugarcane crops. Most often the vinasse is sprayed onto the fields, which is a process termed fertirrigation. This method is low-cost and contributes to net energy savings in sugarcane bioethanol production cycles because the vinasse is locally transported and applied [4]. Benefits of using vinasse as fertilizer include improved short-term soil quality, crop production and crop quality [5-8]. However, negative effects include decreasing long-term soil fertility (lead leaching, N immobilization) and increasing greenhouse gas emissions, especially the emission of N_2O when used in conjunction with an N fertilizer [2, 9-12]. These effects depend on the soil and environmental characteristics as well as vinasse-specific nutrient contents (reviewed in [12]). The increased N_2O emissions from vinasse fertirrigation may be due to the stimulation of soil microbes by vinasse-derived organic material (i.e. a form of priming) or the activity of vinasse-derived cells containing genes in N_2O-producing pathways[8].

Nitrous oxide emissions are produced through two main microbially-mediated processes in soil: nitrification (NH_4^+ to NH_2OH to NO_3^-) and denitrification (NO_3^- to NO_2^- to NO to N_2O to N_2). Nitrification is carried out by microbes containing the ammonia monooxygenase enzyme, which is encoded by the gene *amo*A, and generally used as a functional marker of nitrifiers. Denitrifier bacteria may contain the nitrite reductase genes *nir*S and *nir*K, the nitric oxide reductase gene *nor*B and/or the nitrous oxide reductase gene *nor*B, which each encode for

the enzymes involved in the respiration of nitrite to nitric oxide to nitrous oxide to dinitrogen gas, respectively. The abundance of the different microbes containing denitrification genes, and the abundance of these genes when measured as functional markers, is known to correlate with the actual N_2O emission rates from soils [8]. While much is known regarding the chemical characteristics of vinasse, there are only a few indirect studies of its biotic components despite recent attention to the environmental effects of its use in fertirrigation.

The microbiota present in vinasse likely encompasses the microorganisms present in the bioethanol production process. The most common raw material for ethanol production in Brazil is the mixture of diluted molasses and cane juice, used in the distilleries annexed to sugar producing mills. The ethanol pipeline starts with crushing the unwashed sugarcane stalk to separate the sugarcane juice from the pulpy stalk residue known as bagasse. The sugarcane juice is heated and clarified with lime; the clarified juice is concentrated in an evaporator at 115 degrees C followed by vacuum boiling pan, at which point sugar and molasses crystallize. By centrifugation, the sugar crystals are separated from the mother liquor. This liquor is again crystallized in vacuum pans and then passed through continuous sugar centrifuges. The last residual solution is called molasses, which has high sucrose content suitable for ethanol production. The raw material for ethanol production is a mixture of unsterilized sugarcane juice, molasses and water [13]. The fermented material is then distilled at temperatures of at least 78 ^{0}C to separate the ethanol from the remaining waste vinasse. This vinasse is then transported via open channels or trucks to the sugarcane site for fertirrigation. The mixed sugarcane juice is fermented using proprietary *Saccharomyces cerevisiae* strains through two methods: batch (85% of distilleries as of 2011) or continuous fermentation (15%). In batch processing, the fermentation occurs in parallel, while in continuous fermentation the process occurs in series (reviewed in [14]). In either method, the yeast cells are treated with sulfuric acid, antibiotics, hop products and/or chemical biocides to reduce bacterial contamination, recovered by centrifugation, and reapplied to the fermentation tanks. This recycling step occurs between 400 and 600 times in a harvest season and despite the antibacterial treatment, bacteria remain the major contaminants.

The main bacterial contaminants of the bioethanol pipeline are lactic acid bacteria, which tend to dominate the samples taken from the ethanol pipeline in the steps prior to vinasse [15, 16]. These bacteria, in particular Lactobacillus species, compete with the commercial yeast strains for sugar or form exopolyssacharides that floccule yeast cells [17-19]. Contamination by bacteria – through sucrose competition, flocculation of the commercial yeast strain or fer-

mentation inhibition – can lower the efficiency of the bioethanol process by up to 30% [16, 20]. Furthermore, because of the antibiotic treatment of the yeast cells during the recycling step, contaminant bacteria may be a source of antibiotic resistance genes, as has been recently reported in a field study [21]. Other sources of contamination are wild yeast strains from the input sugarcane stalks, which are not sterilized prior to the production pipeline [22]. To date, no studies have investigated the presence of bioethanol pipeline contaminants in vinasse.

Here we investigated concurrently the chemical and microbial properties of vinasse in order to characterize the vinasse assemblage. We explored metagenomic data taken from vinasse samples over 1.5 years of production from a bioethanol mill in Piracicaba, SP, Brazil. The mill processes sugarcane produced in the region within a rough 40 km radius. Vinasse is distributed by trucks for fertirrigation during the harvest season (April to November). To characterize the microbial assemblage of this vinasse, we sequenced total DNA from six vinasse samples. We analyzed the resulting 18 shotgun metagenomes through metagenomics and differential abundance binning. To investigate the potential for N_2O emissions from fertirrigation with vinasse, special attention was given to sequences and reconstructed genomes annotated as genes involved in N_2O-related metabolism. Furthermore, we also identified genes relating to bioethanol production concerns to identify future research directions. To date this is the first culture-independent study of the vinasse microbial assemblage. Our main questions were (1) what are the overall and sample-wise taxonomic and functional characteristics of the vinasse microbial assemblages? and (2) what is the potential of the vinasse microbes for N_2O emissions, obstruction of fermentation and/or antibiotic resistance?

6.2 Materials and Methods

6.2.1 Sampling description

The bioethanol mill from which we sampled is in the region of Piracicaba in SP, Brazil. The mill takes in sugarcane from the region and produces sugar and ethanol. We obtained six time points of vinasse taken from transport trucks prior to their departure to the fields for chemical and molecular analyses. The trucks hold about 10,000 L of vinasse. Prior to sampling, the vinasse was held in the trucks for 24 hours. Of the vinasse, 0.5 L sampled from the truck was immediately kept at 4 degrees C. The six sampling dates were 13/11/2013 (A, Nov. 2013), 13/12/2013 (B, Dec. 2013), 15/07/2014 (C, July 2014), 15/08/2014 (D, Aug. 2014), 14/10/2014 (E, Oct. 2014) and 10/11/2014 (F, Nov. 2014). The dates of the vinasse sampling corresponded to summer (October, November and December) or

winter (July and August) sugarcane harvests. Because each vinasse was a random assemblage of contaminants from the bioethanol process, we considered each time point an independent measure for statistical analysis.

6.2.2 Chemical analyses, DNA extraction, and qPCR quantification and sequencing

For each vinasse sample, 500 ml was used for chemical analyses. The remaining three subsamples of 100 ml per time point were used for DNA extraction. First, the samples were centrifuged at 10,621 x g (Sigma 2-16P) for 10 min to separate cells from the liquid. Total DNA was extracted from the pellets with the MoBio PowerSoil kit according to the manufacturer's instructions. Between 553 and 5310 ng was sent for sequencing (**Additional file 1**). The DNA was prepared as a MiSeq Illumina paired-end library and sequenced (3 replicates x 6 samples = 18 metagenomes) or used for quantitative PCR of genes that encode for the enzymes involved in the sequential biochemical steps leading to N_2O production (*amo*A, *nir*K, *nir*S, *nor*B) or removal (*nos*Z). The qPCR reactions were performed in a 96-well plate (Bio-Rad) using CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad). The qPCR reaction, primers combinations and thermal cycler conditions of each gene amplification are listed in **Additional file 2.** The qPCR data was acquired at 72 °C and melting curve analysis was performed to confirm specificity. Amplicon sizes were checked by agarose gel electrophoresis. Samples were analyzed with two technical replicates. Reaction efficiency varied from 80 to105% and R^2 values ranged from 0.94 to 0.99.

6.2.3 Metagenome processing and read-based sample comparisons

Bioinformatics processing was performed on a Linux server (Linux-3.13.0-76-generic-x86_64-with-Ubuntu-14.04-trusty) with 64 nodes and 250 GB RAM. Processing was performed in a Snakemake v3.7.1 workflow or with bash or perl scripts (available upon request). The 18 shotgun metagenomes were checked for tag sequences and evaluated for statistics using FastQC v0.10.1 (Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and PRINSEQ-lite version 0.20.4 [23]. Raw reads were filtered out using PRINSEQ if they had more than 1% of ambiguous (N) characters, had a mean quality score of less than 25 or were exact duplicates. Reads were trimmed at the 3'end if the mean quality score was less than 20 within a sliding window size of 10 (clean reads). The clean paired-end reads were used in further analyses unless otherwise noted. The raw paired-end reads were merged using PEAR v0.9.5; these merged

read ends were trimmed by quality and filtered out if the merged read had more than 1% ambiguous characters (parameters: -q 20 –n 0.01) with PEAR (merged reads) [24]. For downstream normalization of annotation counts, calculations of average genome size per sample were carried out using MicrobeCensus [25]. To compare the metagenomes directly, sample distances were determined from the partial de Bruijn assembly of the clean forward reads using MetaFAST 0.1.0 (revision 57253a1) [26].

6.2.4 Taxonomy, phylogeny and alpha diversity

To characterize the taxonomic composition, functional potential and diversity of the microbial assemblages in the vinasse samples, we profiled the metagenomes using different databases. First, the merged reads were uploaded to the metagenome analysis platform MG-RAST version 3.6 [27]. The metagenomes were compared using the default presets to the RefSeq or Subsystems databases to obtain taxonomic or functional profiles, respectively. Refseq annotations, including eukaryota, bacteria, archaea and viruses, were determined using the last common ancestor approach. The MG-RAST taxonomic (phylum-level) and functional (Subsystems Level 1) profiles were analyzed with the Statistical Analysis of Metagenome Profiles (STAMP) software [28]. Taxonomic or functional level abundances significantly different among vinasse samples were evaluated using ANOVA. The Tukey-Kramer post-hoc test with a 95% confidence interval and the Benjamini-Hochberg correction was used to identify differing phyla or Subsystems Level 1 category abundances between the vinasse metagenomes with significance determined at corrected $p<0.001$ or 0.05, respectively. The taxonomic profiles at genus level were kept to visualize the relative abundance of genera across samples.

Because the metagenomes were well-represented in the MG-RAST databases, we further characterized the taxonomy and functional potential of the metagenomes using metaphlan2 version 2.6.0 and humann2 version 0.9.9 pipelines [29,30]. For metaphlan2 analysis, we used the “relab” analysis with the “--ignore_eukaryotes” flags to obtain taxonomic profiles. To gain an overall view of the taxonomy present in the vinasse samples and the phylogenetic relationships between the species in the samples, the average taxonomic distributions of the vinasse samples from metaphlan2 were visualized as a cladogram using Graphlan [31]. To examine the taxonomic profiles of vinasse across samples, these were visualized through heatmaps with average linkage clustering on Euclidean distances using hclust2. For the humann2 analysis, we annotated the forward clean reads against the UniRef90 database [32]. Pathway abundances were visualized exclud-

ing the “UNMAPPED” and “UNKNOWN” categories using hclust2 heat maps with average linkage clustering on Euclidean distances. To obtain a measure of alpha diversity, we ran metaphlan2 with previously mentioned flags on samples rarified to the smallest library size (280,161 reads).

To infer the phylogenetic relationships between the organisms present in the vinasse samples, full-length 16S rRNA genes were recruited from the vinasse metagenome reads using REAGO version 1.1 on forward clean reads truncated to 201 bp [33]. The resulting full-length 16S rRNA vinasse sequences were aligned and taxonomically classified against the SSU 128 SILVA reference database using SINA [34,35]. The five nearest neighbors for each full-length 16S rRNA sequence were downloaded in addition to two *Verrucomicrobia* outgroup sequences. The 16S rRNA sequences were aligned without gaps using ClustalW in MEGA7 (121 sequences in total)[36]. A neighbor-joining tree was created with evolutionary distances computed using the Maximum Composite Likelihood method [37,38]. Phylogenetic distances were evaluated with bootstrap tests (1000 replicates) [39]. To obtain a measure of alpha diversity we recruited full-length 16S rRNA genes using REAGO as above on the rarified metagenomes. Further, we evaluated a measure of genus-level relative abundance across samples by mapping the metagenome reads to the extracted 16S sequences grouped by taxonomic affiliation using bowtie2. These abundances were calculated as percentages of the number of aligned pairs from the total number of metagenome reads per sample.

6.2.5 Putative denitrification and nitrification gene abundances

To investigate the potential for N_2O emissions from the vinasse samples, we used two approaches: 1) metagenome read matching to profile HMMs of denitrification and nitrification genes and 2) recruitment of denitrifying and nitrifying genes from the reads. Profile HMMs for the *amo*A_*AOA, amo*A_*AOB, nir*K, *nir*S, *nor*B, *nosZ, nosZ_atypical_1* and *nosZ_atypical_2* genes were downloaded from the Functional Gene Repository (FUNgene version 8.3; available at http://fungene.cme.msu.edu/). Reads were translated to protein sequences with the “meta” setting using Prodigal version 2.6.2. The ORFs were queried for HMM matches using HMMsearch (command: hmmsearch --noali -o <filename.fasta> <gene.hmm> <filename.fasta>; available at https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch). The HMM matches were normalized to reads per kilobase per genome equivalent (RPKG = (# mapped reads / HMM gene length (KB)) / genome equivalents). The RPKG normalization accounts for genome size, library size and gene length biases, allowing for gene and sample comparisons.

In parallel, the gene-targeted assembler pipeline megagta version 0.1_alpha was used to recruit full-length genes from the metagenomes [33,40]. Gene-targeted assemblies (i.e. recruitments) were carried out on *amo*A_AOA, *amo*A_AOB, *nir*S, *nir*K, *nor*B_*cNor*, *nor*B_*qNor*, *nos*Z and *nos*Z_a2 genes using megagta. Further, to infer alpha diversity, the ribosomal *rpl*B gene was recruited from the rarified metagenomes.

6.2.6 Cross-assembly and binning

We evaluated the performance of three assemblers (Ray-meta [41], Megahit [42] and metaSpades[43]) in cross-assembling the 18 vinasse metagenomes; the best cross-assembly was that from the metaSPADES assembler version 3.8.2 based on assembly characteristics evaluated using MetaQUAST (QUAST Version 3.0, build 07.07.2015 05:57 [44]). The 18 metagenomes were cross-assembled with metaSpades using kmer sizes 77, 99 and 127. The sample reads were mapped to the cross-contigs using bowtie2 to obtain cross contig abundances [45] . The final metaSPADES cross-assembly was binned using three tools for comparison: CONCOCT (with anvio version 2.3.2), Metabat [46] and MaxBin2 version 2.1.1 [47]. The contig annotation tool (CAT version 2) was used to determine the taxonomic affiliation of all ORFs identified in each bin using prodigal to find ORFs and diamond blastp against the NCBI-nr database [48]. CAT taxonomy results were formatted using custom Perl scripts and visualized with TreeMap to aid with the taxonomic characterization of the bins. Because more genomes with >90% completeness and coherent taxonomies were found from the MaxBin2 binning, these were selected for downstream analysis. CheckM was used to check the original MaxBin2 bins [49]. These bins were manually refined using anvi'o version 2.3.2 based on cross-contig taxonomy (from CAT), hierarchical clustering of the cross-contigs and sample coverage information [50]. The relative sample abundances of the bins were noted as the percent of sample reads recruited to the bin out of the total sample reads recruited to all the bins (i.e. percent recruitment anvi'o results).

The "good bins" were identified as having >90% completeness and <10% redundancy. Further "interesting bins" were further identified as those with >20% completeness and <10% redundancy and/or coherent contig taxonomies. Functional annotation of the "good and interesting bins" were carried out using prokka with the "kingdom" flag set to bacteria or viruses depending on the taxonomic classification [51]. To characterize the bins by their potential functional type, prokka annotation results were mined for lines matching EC numbers of KEGG enzymes of compounds related to bioethanol production interests and N_2O emis-

sions. These KEGG compounds were acetate (C00033), cellulose (C00760), xylose (C00181), lactose (C00242), caprylic acid (C01585, carbon dioxide (C00011), diacetyl (C00741), hydrogen peroxide (CC00027), lactaldehyde (C05999) and phenyllactate (C05607). The lists of EC numbers were obtained by querying the KEGG REST API on each compound ID. Keyword searches of "3-hydroxy" fatty acids, "cyclic dipeptide," antibiotic "resistance," and nitrification and denitrification genes were additionally used to identify the potential presence of these functions in the bins.

In parallel, to confirm potential denitrification and nitrification gene presence, bin sequences were compared to HMMs of nitrification and denitrification genes from FunGene as described previously but with the prodigal setting "single." The HMM matches were normalized by bin size (number of ORFs and total number of bp in ORFs) and HMM length in bp.

6.3 Results

6.3.1 Vinasse chemical characteristics and metagenome overview

The chemical characteristics of the vinasse samples are listed in **Table 1**. Average pH was low at 4.4±0.4, ranging between 3.9 (D) and 4.8 (C). Total organic carbon averaged 29±1.8 g L^{-1} and ranged between 25.7 (B) and 31.4 g L^{-1} (D). Total N averaged 0.64±0.15 g L^{-1}, while that of P and K was 0.16±0.07 and 3.43±1.02, respectively. The C/N ratio averaged 42±13 and ranged between 19 (F) and 57 (C). After processing, the 18 vinasse metagenomes contained a total of 2,126 Mbp distributed into 7.82 million reads. The number of reads ranged between 280,161 and 542,208 sequences per sample with between 77 and 150 Mbp (**Additional file 1**). When the metagenome distances were compared using partial de Bruijn assembly, A and C were most similar, followed by F, followed by E; least similar were B and last D (**Additional file 3**).

Table 1. Chemical characteristics of the six vinasse samples.

Group Name	Sampling Date	pH	C org	N tot	$N\text{-}NH_4^+$	$N\text{-}NO_3^-$	P	K	C:N
			g L^{-1}	g L^{-1}	mg L^{-1}	mg L^{-1}	g kg^{-1}	g kg^{-1}	
A	Nov. 2013	4.7	28.2	0.53	65.8	17.6	0.08	2.9	53
B	Dec. 2013	4.1	25.7	0.53	63.4	10.8	0.17	2.6	49
C	July 2014	4.8	28.8	0.51	45.7	8.8	0.11	3.5	57
D	Aug. 2014	3.9	31.4	0.89	41.6	4.1	0.23	4.7	35
E	Oct. 2014	4.2	29.6	0.74	37.7	6.8	0.10	2.1	40
F	Nov. 2014	4.7	30.3	1.57	75.9	6.6	0.25	4.8	19

6.3.2 Taxonomic characterization

When compared to the M5NR database containing eukaryota, bacteria, archaea and viruses on MG-RAST (**Additional file 4**), 21 to 55% of the merged reads could be classified. Of the classified reads, 96 to 100% were annotated as bacteria. The top phyla present in the vinasse samples with relative abundances greater than 1% and/or that significantly co-varied among the samples (ANOVA at $p<0.001$ and Kruskal-Wallis post-hoc test) were Firmicutes (35 to 97% of merged reads), Bacteroidetes (0.8 to 53%), Actinobacteria (0.4 to 17.5%) and Proteobacteria (0.3 to 39.4%; **Additional file 5**). The "core" phylum observed in all vinasse samples was Firmicutes. Similarly, when compared to the metaphlan2 marker gene database containing bacteria, archaea and viruses (excluding eukaryotes), between 68 and 100% of classified reads were identified as bacteria and 0 to 32% as viruses (**Figure 1**). The previous four main bacterial phyla again dominated the vinasse samples: Firmicutes (48 to 100% of classified reads), Actinobacteria (0 to 19%) and Proteobacteria (0 to 18%), as well as viruses (0 to 32%; **Figure 2**). The most abundant bacterial genera were Lactobacillus (Phylum Firmicutes), Megasphaera (Firmicutes), Mitsuokella (Firmicutes) and Bifidobacterium (Actinobacteria). Further supporting these taxonomic results, the full-length 16S rRNA genes recruited from the vinasse metagenomes were classified as Bifidobacterium (Phylum Actinobacteria), Olsenella (Phylum Actinobacteria), Prevotella (Phylum Bacteroidetes), Lactobacillus (Phylum Firmicutes), Megasphaera (Phylum Firmicutes), Mitsuokella (Phylum Firmicutes) and *Comamonas* (Phylum Proteobacteria) genera (**Additional file 6 and 13**).

When the samples were clustered based on the MG-RAST taxonomic profiles at phylum level, E and C formed a cluster while A, F, and D were separated based on the first principal component and B was separated based on the second (**Additional file 7**). When the metaphlan2 profiles were clustered at the level of class, order, family and genus, samples A, C and F formed a cluster while B, D and E formed a separate cluster (**Figure 2**).

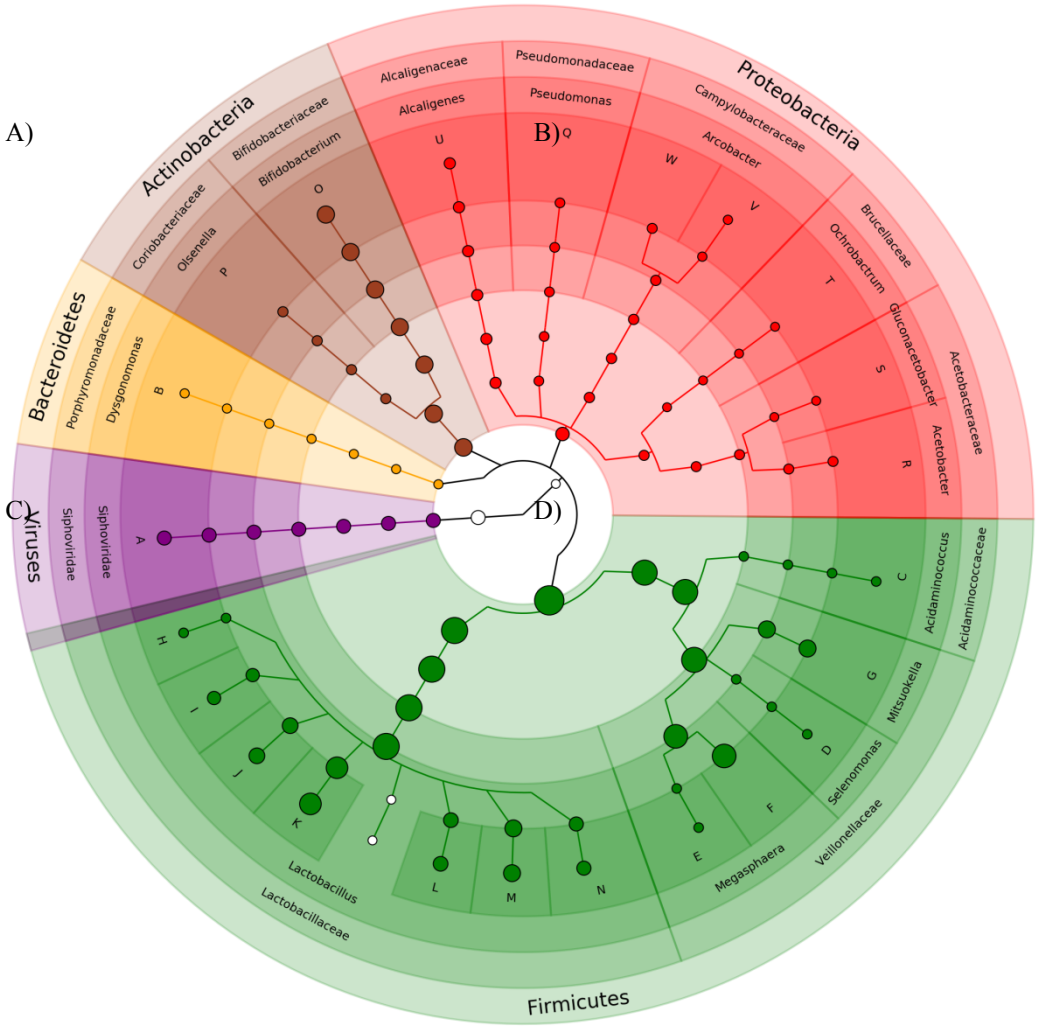


Figure 1. Average abundance of taxa in the vinasse samples. The metagenomes were analyzed using metaphlan2 and visualized with GraPhlan. Node sizes correspond to average relative abundance across the vinasse metagenomes while colors correspond to phylum. Species are noted with letters: A=*Lactobacillus* phage Lc Nu, B=*D. mossii*, C=*A. intestini*, D=*S. bovis*, E=*M. elsdenii*, F=*Megasphera* unclassified, G=*Mitsuokella* unclassified, H=*L. salivarius*, I=*L. equicursoris*, J=*L. delbrueckii*, K=*L. amylovorus*, L=*L. mucosae*, M=*L. fermentum*, N=*L. vini*, O=*B. thermophilum*, P=*Olsenella* unclassified, Q=*Pseudomonas* unclassified, R=*Acetobacter* unclassified, S=*Gluconacetobacter* unclassified, T=*Ochrobactrum* unclassified, U=*A. faecalis*, V=*A. butzleri* and W=*Arcobacter* unclassified.

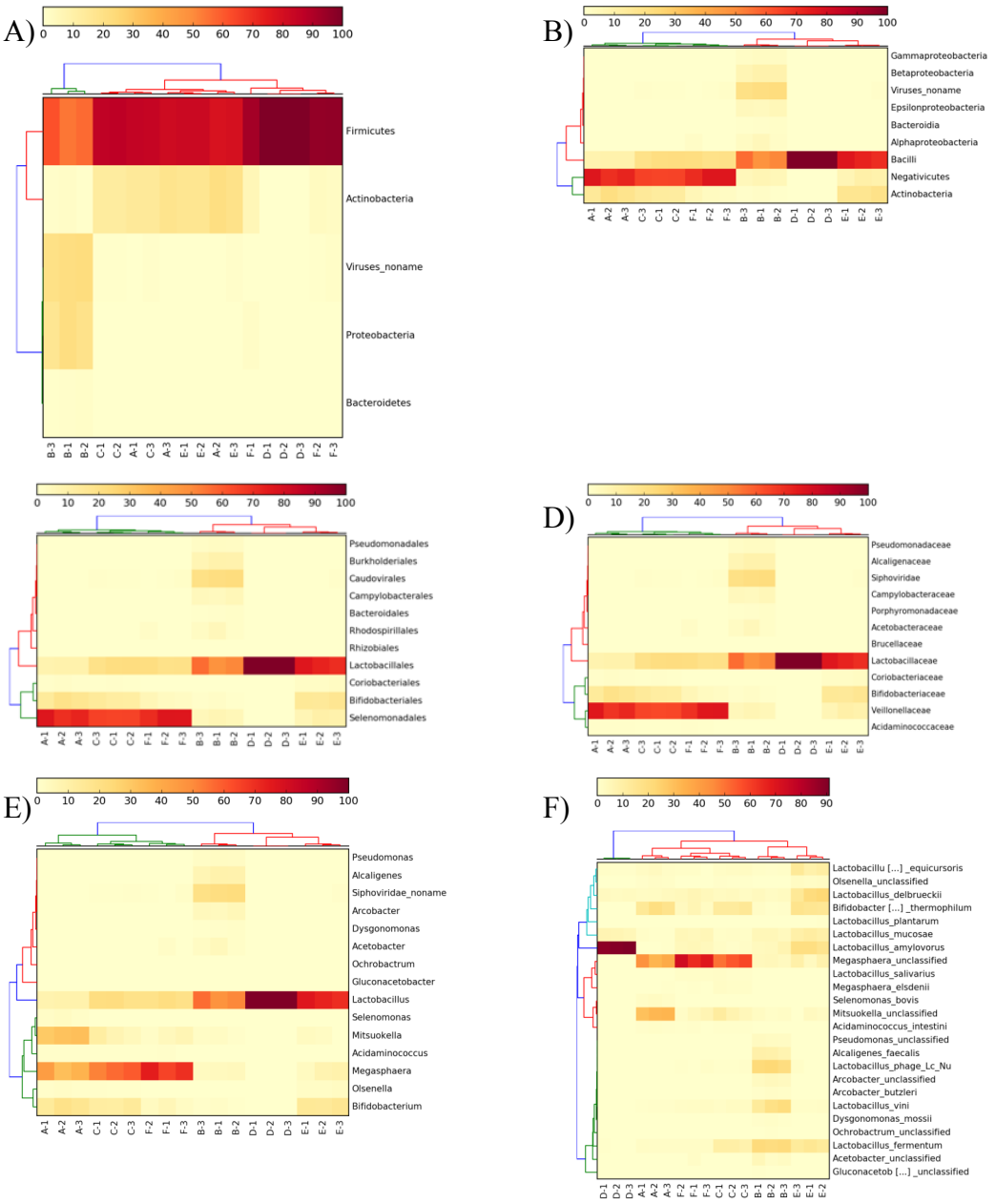


Figure 2. Taxonomic distributions across the vinasse samples at the level of A) Phylum, B) Class, C) Order, D) Family, E) Genus and F) Species. The taxonomic group and sample profiles were clustered using hclust2 from metaphlan2 results.

6.3.3 Functional potential characterization

When compared to the M5NR databases through MG-RAST, the percentage of sequences with ORFs that could be classified into functional categories ranged between 16 and 42% (**Additional file 8**). At Subsystems Level 1, the top significantly different categories were Carbohydrate metabolism, Clustering-based subsystems, Amino Acids and Derivatives, Miscellaneous, Protein metabolism, DNA metabolism, RNA metabolism, Cofactors/vitamins, Cell wall/capsule, Phages/ prophages and Nucleosides and Nucleotides. When sample distances were determined using the functional profiles at Subsystems Level 1, C, A, B and F formed a cluster while E was separated based on the first principal component and D was separated based on the second (**Additional file 9**). When the vinasse metagenomes were analyzed using the humann2 framework, abundant pathways were found in sample D, which was dominated by one *Lactobacillus* – the top abundant pathways included PWY-5100: pyruvate fermentation to acetate and lactate II and PWY-7219: adenosine ribonucleotides de novo biosynthesis (**Additional file 10**). Combining the real-time PCR, gene recruitment and gene mapping results, the vinasse metagenomes had few to no genes matching nitrification genes; in contrast, a range of denitrification genes was found (**Figure 3**). Sample B presented the most diversity of denitrification genes, with *nir*K, *nir*S, *nor*B and *nos*Z present based on the recruitment and mapped results. The presence of putative *nos*Z was supported in all samples except D. In addition, putative *nir*K was found in all samples except F.

6.3.4 Alpha diversity of the vinasse samples

Several methods were employed to obtain estimates of the alpha diversity of the vinasse samples (**Table 2**). The normalized effective number of species from MG-RAST averaged 29±14 and ranged between 3 (D) and 53 (B) species. When metaphlan2 was applied to the rarified vinasse samples, the number of classified species averaged 11±3 and ranged between 5 (D) and 14 (B) species. Partial 16S rRNA fragments recruited from the rarified samples using REAGO averaged 10±4 and ranged between 4 (D) and 17 (E). Further, when the *rpl*B gene was recruited from the rarified samples with megaGTA analysis, and average of 17±3 fragments could be found across the vinasse samples with between 13 (E) and 22 (C) *rpl*B fragments identified. When the 16S rRNA gene was amplified using real time PCR of the vinasse samples, the number of genes per kg of dry matter averaged 12e12±9e12 and ranged between 0.8 (E) and 25.7 (A); the gene abundance of 18S rRNA gene averaged 100e3±71e3 and ranged between 17e3 (D) and 208e3 (B).

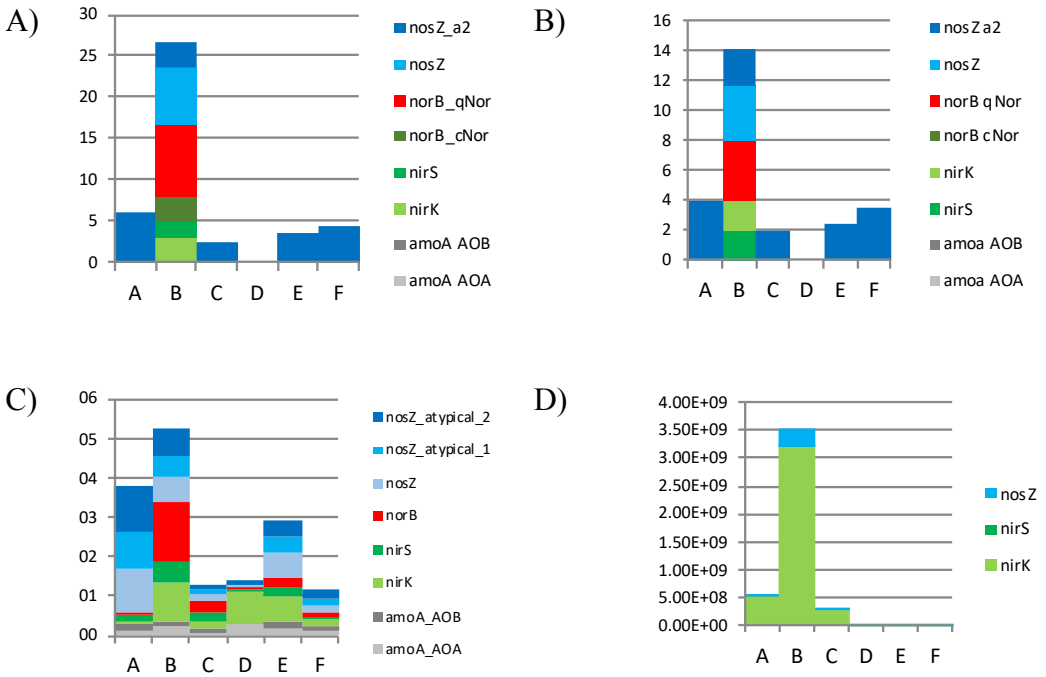


Figure 3. Putative gene abundances in the vinasse metagenomes**.** Partial gene fragments were recruited from the vinasse metagenomes using megagta on A) all reads and B) rarified reads. In parallel, vinasse metagenomes were compared to profile HMMs and the number of matches was normalized to C) reads per kilobase per genome equivalent (RPKG). In D) the gene copy numbers from real time PCR of the *nos*Z, *nir*S and *nir*K genes are depicted. Note that no qPCR of the *norB* gene was made.

Table 2. Alpha diversity estimates of the vinasse samples. Diversity was quantified by the number of partial genes recruited (REAGO and megaGTA), or the estimated number of species (metaphlan2 and MG-RAST) from the vinasse metagenomes; results from real-time PCR of the 16S gene was also included. Rarified forward reads were used as input for metaphlan2, reago and megagta analysis; merged reads were used in the MGRAST analysis and these results were normalized by library size.

	REAGO	megaGTA	metaphlan2	MGRAST	qPCR
Sample Name	# recruited 16S rRNA genes	# recruited rplB genes	# Species	Effective # species	# 16S rRNA copies (/ 1000000) kg dry matter^{-1}
A	13 ± 2	21 ± 2	10 ± 1	37 ± 1	25750 ± 13900
B	10 ± 3	16 ± 3	14 ± 1	47 ± 4	16839 ± 11664
C	12 ± 1	22 ± 2	13 ± 0	38 ± 1	16281 ± 1104
D	4 ± 0	15 ± 2	5 ± 0	3 ± 0	10749 ± 3336
E	17 ± 2	13 ± 1	10 ± 0	20 ± 1	839 ± 840
F	6 ± 2	17 ± 1	12 ± 1	29 ± 3	1135 ± 1142

6.3.5 Bin characteristics, taxonomy and functional types

The cross-assembly resulted in 221,975 cross-contigs totaling 216 Mbp. Of the cross-contigs, 40,815 were longer than 1Kbp, and 40,186 of these could be binned. After refining the bins, 20,825 cross-contigs remained distributed within the 36 good or interesting large bins (0.6 to 3.9 Mbp; hereafter referred to as the large bins). The large bins represented 39 to 68% of the sample reads. Fifty-eight percent of the large bins were classified at the phylum level as Firmicutes, 8% as Bacteroidetes, 17% as Proteobacteria, 11% as Unknown and 6% as Actinobacteria (**Table 4**). Overall, the GC percent of these bins ranged between 28 and 66%. Of the large bins, 24 were potential denitrifiers and three potential nitrifiers. The presence of genes related to acetate, CO_2, ethanol, H_2O_2 and lactose metabolism were found in all large bins while the potential presence of genes related to Lactaldehyde, mannitol, xylose, butyric acid, cellulose, diacetyl, phenyllactate, sucrose and "3-hydroxy" was variable across the large bins (**Table 5**). Last, when multidrug resistance was identified in the bin annotations, all large bins but Unknown-19 and Lactobacillus-30 contained these genes. In addition to the large bins, eight small bins (0.03 to 0.20 Mbp) lacking bacterial marker gene presence were found (**Table 3** and **Additional files 11 and 12**). The largest of the small bins, 4.2 and 8.1 were most abundant in samples E and D, respectively.

Table 3. The "good and interesting" vinasse bin characteristics and relative sample abundances (indicated by heatmap per sample).

Bin Ids	A	B	C	D	E	F	Length (Mbp)	# Contigs	N50	GC (%)	Completeness	Redundancy
1	6	1	2	0	0	1	2.37	209	23171	53	92	2
2	14	2	1	0	4	3	3.45	547	10519	49	94	4
3	8	1	18	0	2	13	2.19	352	10534	53	97	2
4.1	0	0	0	1	8	0	0.05	19	3352	37	0	0
4.2	0	0	0	0	5	0	0.14	14	32202	29	0	0
5	3	0	4	0	2	1	1.91	298	11078	60	94	3
6	4	1	2	0	1	12	2.42	539	6347	44	91	2
8.1	0	0	0	1	5	0	0.20	39	52952	36	0	0
8.2	0	1	0	9	1	0	0.03	17	1553	40	0	0
8.3	0	0	1	0	3	0	0.07	27	2920	47	0	0
8.4	0	1	0	0	0	0	0.02	10	2442	41	0	0
8.5	0	0	0	0	1	0	0.03	13	3198	39	0	0
8.6	0	0	0	1	1	0	0.03	16	3273	45	1	0
9	2	0	2	0	1	0	1.96	407	6384	60	90	6
10	1	1	1	1	3	0	2.07	183	27583	47	96	1
12	3	0	2	0	5	1	2.34	485	7489	66	91	5
13	2	0	2	0	0	0	1.48	605	2710	63	71	12
14	2	0	1	0	0	1	1.88	947	2190	52	74	15
15	2	0	2	0	1	1	2.16	1017	2297	53	76	9
16	1	1	1	0	0	13	3.02	307	22370	42	99	1
18	2	0	1	0	2	0	1.48	893	1665	63	69	22
19	1	0	1	0	3	0	2.01	917	2351	62	60	12
20	0	0	1	0	1	1	1.16	387	3766	54	64	2
21	0	0	1	1	2	1	1.78	298	9992	50	88	1
23	0	1	1	3	2	1	1.90	373	7041	47	95	4
24	0	2	1	0	1	0	1.78	220	13204	53	98	4
25	1	1	1	1	0	2	1.75	729	2822	40	91	9
26	1	1	2	0	1	1	2.12	1236	1733	41	66	15
27	0	1	1	45	3	1	1.94	262	11858	38	99	1
28	0	3	0	0	0	0	2.11	340	8670	38	96	5
29	0	1	0	2	0	0	1.02	439	2658	50	88	9
30	0	3	1	3	2	3	3.94	2021	1897	31	47	16
31	0	0	8	0	0	0	3.29	1595	2241	54	79	37
32	0	0	0	6	0	0	2.00	104	208993	36	99	1
33	0	0	0	1	3	0	1.71	447	4850	48	96	9
34	0	3	0	0	0	0	2.68	343	12289	60	92	1
35	0	4	0	0	0	0	2.72	259	13750	43	96	6
36	0	5	0	0	0	0	1.70	647	2989	49	67	5
37.1	0	10	0	0	0	0	3.08	1326	2603	57	77	24
37.2	0	4	0	0	0	0	1.22	784	1533	58	34	6
38	0	9	0	0	0	0	2.99	488	9382	60	89	3
39	0	0	0	4	0	0	1.90	205	15510	47	99	4
40.1	0	4	0	0	0	0	1.61	190	11919	28	97	2
40.2	0	2	0	0	0	0	0.57	271	2044	27	15	1

Table 4. Taxonomy of the "good and interesting" vinasse bins based on CAT classification.

Bin Id	K	Phylum	Class	Order	Family	Genus	Species
1	B	F	Nega	Selenomonadales	Veillonellaceae	Mitsuokella	U
2	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	P. mutisaccharivorax/Unclassified
3	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera	U/M. elsdenii
4.1	V				Caudovirales	Siphovirales	Lactobacillus phage Ldl1
4.2	B	U/B	U/B	U/B	U	U	U
5	B	A	Acti	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	U
6	B	B	Bact	Bacteroidales	Prevotellaceae/U	Prevotella/U	U
8.1	B/V	U	U	U	U/Caudovirales	U/Siphoviridae	U/Lactobacillus phage LdL1
8.2	B	F	B	Lactobacillales	Lactobacillaceae	Lactobacillus	U
8.3	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	P. histicola/U
8.4	B	F	Bac	Lactobacillales	Lactobacillaceae	Lactobacillus	U
8.5	B	B	U/Bact				
8.6	B	F	Bac	Lactobacillales	Lactobacillaceae	Lactobacillus	U
9	B	U/A					
10	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. equicursoris/Unclassified
12	B	A/U					
13	B	U/A					
14	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera	U/sp. DJF_B143
15	B	F	Nega	Selenomonadales	Veillonellaceae	Dialister	U/D. succinatiphilus
16	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	U
18	U/B						
19	B	U/A					
20	B	F	Clos	Clostridiales	Eubacteriaceae/U	Pseudoramibacter/U	U/P. alactolyticus
21	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. delbrueckii/U
23	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. mucosae/U
24	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. fermentum/U
25	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U/L. vini
26	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U/L. agilis
27	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U
28	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. vini
29	B	F	Clos	Clostridiales	Clostridiaceae	Clostridium	sp. CAG:568/U
30	B/A	F/E	Bacilli/Metha	Bacteroidales/Methanobacteriaceae	Lactobacillaceae/Methanobactericeae	Lactobactillus/Methanobrevibacter	U
31	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera/U	U
32	B	F	Nega	Lactobacillales	L	L	U
33	B	F	B	Lactobacillales	L	L	L. secaliphilus/U
34	B	F	B	Lactobacillales	L	L	U/L. manihotivorans
35	B	F	B	Lactobacillales	L	L	L. bifermentans/U
36	B	P	Beta	Burkholderiales	Alcaligenaceae	U	U
37.1	B	P	B	Burkholderiales	Alcaligenaceae	Alcaligenes	A. faecalis/U
37.2	B	P	B	B	A	Alcaligenes	U/A. faecalis
38	B	P	B	B	Comamonadaceae	U/Comamonas	U/C. kerstersii

K=Kingdom, F=*Firmicutes*, B=*Bacteroidetes*, A=*Actinobacteria*, P=*Proteobacteria*, E=*Euryarchaeota*, U=Unknown, Bact=*Bacteroidia*, Nega=*Negativicutes*, Clos=*Clostridia*, Mega=*Megasphaera*, Lact=*Lactobacillales*, Metha=*Methanobacteria*

Table 5. Putative gene repertoires of the large vinasse bins. Keyword searches of prokka annotation results (grey, "Y") were supplemented in the case of the N_2O metabolism-related genes with hmm profile search results (colors). Substrates for the genes related to N_2O metabolism are included.

																NO_2					
Bins	Acetate	CO_2	Ethanol	Lactaldehyde	Mannitol	Xylose	Butyric acid	Cellulose	Diacetyl	H_2O_2	Lactose	Phenyllactate	Sucrose	3-hydroxy	ABR	*nirK*	*nirS*	*norB*	*nosZ*	*nosZ_1*	*nosZ_2*
1	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
2	Y	Y	Y		Y	Y	Y	Y		Y	Y		Y	Y Y						Y	Y
3	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y Y							
5	Y	Y	Y					Y		Y	Y		Y	Y							
6	Y	Y	Y					Y		Y	Y		Y	Y Y							
9	Y	Y	Y		Y	Y		Y		Y	Y		Y	Y Y							
10	Y	Y	Y	Y	Y	Y		Y		Y	Y		Y	Y Y							
12	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
13	Y	Y	Y		Y			Y		Y	Y		Y	Y							
14	Y	Y	Y		Y			Y		Y	Y			Y Y							
15	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
16	Y	Y	Y		Y	Y	Y	Y		Y	Y		Y	Y Y							
18	Y	Y	Y		Y					Y	Y			Y							
19	Y	Y	Y					Y		Y	Y	Y	Y								
20	Y	Y	Y	Y						Y	Y			Y Y							
21	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
23	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	Y Y							
24	Y	Y	Y			Y	Y	Y		Y	Y		Y	Y Y							
25	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y		Y	Y Y							
26	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y Y							
27	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
28	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y Y							
29	Y	Y	Y					Y		Y	Y			Y							
30	Y	Y	Y					Y		Y	Y	Y	Y								
31	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y							
32	Y	Y	Y		Y			Y		Y	Y	Y	Y	Y Y							
33	Y	Y	Y		Y	Y		Y	Y	Y	Y		Y	Y Y							
34	Y	Y	Y		Y		Y	Y		Y	Y	Y	Y	Y Y							
35	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y Y							
36	Y	Y	Y		Y			Y		Y	Y		Y	Y Y							
37.1	Y	Y	Y		Y		Y	Y		Y	Y		Y	Y Y							
37.2	Y	Y	Y					Y		Y	Y		Y	Y Y		Y				Y	
38	Y	Y	Y		Y			Y		Y	Y	Y	Y	Y Y							
39	Y	Y	Y		Y	Y	Y	Y		Y	Y		Y	Y Y							
40.1	Y	Y	Y		Y			Y		Y	Y			Y Y							
40.2	Y	Y	Y					Y		Y	Y			Y							

ABR = antibiotic resistance; Notes: no genes related to the metabolism of caproic acid were found in the bin annotations. No *amo*ABC, *hao*, *nxr nor*, *nir*S genes were found in the bin annotations, but the *amo*A AOA gene was identified in Bin 23 and 40.1 and the *amo*A AOB gene was identified in Bin 33 by HMM matches.

6.4 Discussion

Here, we explored concurrently the chemical and microbiological characteristics of vinasse produced over 1.5 years from one bioethanol mill in Sao Paulo State. The aims were to characterize, for the first time, the taxonomy and potential functions of the microbial assemblage in vinasse; we further recovered draft genomes from vinasse bacteria. We combined metagenomic analyses with binning techniques to characterize the vinasse assemblages and bacteria, respectively. We discuss below both potential ethanol pipeline contamination traits of vinasse bacteria and the potential ecology of vinasse fertirrigation. The vinasse chemical characteristics fell within the range of other sugarcane vinasses [52, 53]. Different vinasse inputs are known to contribute different nutrition; this is taken into account in that vinasse fertirrigation is applied depending on the amount of K present in the input vinasse [54]; however, that different vinasse inputs contribute different bacteria was not known until now. The different nutrient contents of vinasse originate from the differences in the input of sugarcane stalks to the bioethanol production process; this might also be the source of the vinasse bacteria.

The vinasse draft genomes most likely represented the bacteria that survived the selective bottleneck of the bioethanol production pipeline. The potential for bacteria found in vinasse originating from later steps in the bioethanol pipeline, such as the truck from which we sampled the vinasse, was considered a minor source of bacteria due to the large capacity (10,000 L), making this a negligible source of bacteria. The core genus found in the vinasse samples was Lactobacillus (Phylum Firmicutes), which is a previously known ubiquitous ethanol pipeline contaminant due to its tolerance of low pH [15]. Other known contaminants found prior to the distillation stage that we observed in our vinasse samples included representatives of the Acetobacter, Bacillus, Bifidobacterium, Clostridium, Gluconacetobacter, Lactobacillus and Pseudomonas genera [16, 55, 56]. Strikingly, we identified members of the genera Megasphaera and Mitsuokella that have not previously been reported as bioethanol pipeline contaminants. Members of the genus Megasphaera and Mitsuokella are Gram negative ruminant fermenters that have been found in pig hindguts, cow rumen and human dental plaque and feces; gram-positive Bifidobacterium have also been used as probiotics in humans and are found in the gut, vagina and mouth of mammals and bovine rumens. Whether these bacteria interact with each other within each vinasse sample – e.g. Megasphaera and Mitsuokella utilizing lactose provided by Lactobacillus – is unknown, as is the direction of the interactions.

Uncovering the physiological mechanisms by which these particular bacteria survive the selection bottlenecks of the bioethanol process was outside the scope of the current research since our goals were to characterize fully the metagenomic data. However, we speculated that plausible protective mechanisms are biofilm formation [16, 57], strain-dependent temperature tolerance, and unknown pipeline management considerations. For the latter, the distillation material might not homogenized, thus creating pockets of lower temperatures where the bacteria can remain. Other management considerations that might affect the viability of bacterial cells are length of time exposed to the distillation temperature and the highest temperature reached. Evaluating the physiology of cultured isolates from vinasse, which can be done building upon the work described here, is an interesting topic for further research.

Here, using differential abundance binning, we successfully obtained 21 draft genomes from vinasse bacteria likely representing bioethanol contaminants. We confirmed that roughly half of the vinasse bins were of the genus Lactobacillus (Phylum Firmicutes), which is the most ubiquitous bacterial bioethanol pipeline contaminant [16]. We also uncovered contaminants with up to 70% of sample coverage from the Prevotella (Phylum Bacteroidetes), Megasphaera (Phylum Firmicutes), and Mitsuokella (Phylum Firmicutes) genera, which have not been as well-studied. Five of the draft genomes were from bacteria unknown at the phylum level. Furthermore, most of the bins recovered here were partly uncharacterized at the species level, supporting the idea that we obtained genomes from novel strains of bioethanol contaminants. Studies of bioethanol contaminants to date have used culture-based methods, which do not capture the entire microbial diversity; or profiling of 16S rRNA genes, which does not capture the functional potential of the contaminants [13, 14]. Bacterial contaminants in general are known to compete with commercial yeast strains, lowering ethanol yield; contaminants may also flocculate with the yeast or produce compounds such as acetate, butyric acid or lactose which might inhibit yeast fermentation [16]. Many bins contained sucrose metabolism-related genes, suggesting that these might compete with the commercial yeast strain for sugarcane sucrose. Annotation of the bins revealed the potential presence of bioethanol contaminant genes related to the metabolism of acetate, ethanol, mannitol, cellulose, hydrogen peroxide, lactose, sucrose and 3-hydroxy fatty acids. These results support the idea that vinasse bacteria are an additional source in identifying likely bioethanol process contaminants.

Interesting bins included Lactobacillus/Methanobrevibacter-bin30 and Arcobacter/Methanobrevibacter-bin40.2, which contained cross-contigs annotated as both bacterial and archaeal. Methanobrevibacter is an archaeal genus whose

methanogenic members are often found in vertebrate guts consuming the end products of bacterial fermentation. Finding them here suggests that this interaction might also be present in vinasse. In addition, we binned potential phage genomes, which suggest that phages are present in the fermentation tanks along with the host contaminants. The large phage genome bin 8.2 was most abundant in vinasse sample D, corresponding to a low diversity assemblage with a dominant bin, suggesting that the host of this phage was L. amylovorous-bin27. The phage bins 4.1, 4.2 and 8.1 were all most abundant in vinasse sample D, corresponding to a more diverse assemblage of bacterial hosts across the phyla Firmicutes and Bacteroidetes. These associations suggest that phage lysis may be a factor controlling bacterial population sizes in the fermentation tanks. Attention has recently been paid to using phage therapy to control bacterial contamination in bioethanol pipelines [58-59].

In addition to investigating the potential for vinasse bacteria to be contaminants in the production of bioethanol, we evaluated the potential for vinasse bacteria to contribute to N_2O emissions during fertirrigation. Vinasse fertirrigation can be considered a disturbance on the soil microbial community; the success of the vinasse assemblage in the soil likely depends on the connectivity (e.g. strength and direction of the vinasse species interactions). Pitombo et al.[11] identified significantly abundant bacterial genera under treatments of vinasse compared to unfertilized control plots using 16S rRNA marker abundance, and the significantly differentially abundant genera in the plots amended with vinasse included the vinasse bacteria (as identified here) Lactobacillus, Bacillus, Prevotella, Gluconacetobacter, Megasphaera, Mitsuokella and Acetobacter [11]. Further, unpublished research suggests that vinasse bacteria on a field experiment may persist at low abundances. These results together suggest that vinasse bacteria may successfully invade the soil microbial community. Furthermore, the vinasse bacteria may transfer to the sugarcane stalks during plant growth and at harvest time become the contaminants that are inputted with the sugarcane to the ethanol processing pipeline. In support, a survey of the bacteria associated with the sugarcane plant found the vinasse taxa Bacillus, Acetobacter and Gluconacetobacter as part of the “core” sugarcane microbiome [59]. While this is interesting speculation, we note that caution should be taken because the referenced studies were few and based on gene marker surveys at higher taxonomic levels, which hinders robust and precise interpretation. We recommend further research into the ecological interactions of vinasse bacteria with the soil bacterial community at the species or strain level during fertirrigation with vinasse.

Actual N_2O emissions from a soil are the result of the sequential biochemical processes nitrification and denitrification carried out collectively by the microbial communities in a soil. The total rate of N_2O emissions through nitrification or denitrification is controlled by carbon availability, moisture, oxygen availability, pH, temperature, and nitrate concentrations. These factors limit enzyme activity, gene transcription levels and microbial cell growth [61]. Furthermore, the abundance of the genes involved in the production (amoA, nirK, nirS, norB) or removal (nosZ) of N_2O is correlated with the actual N_2O emissions [62]. In the case of vinasse fertirrigation, if many denitrifiers invade a soil conducive to denitrification, we would expect more denitrification to occur. Whether net N_2O or N2 (the end product of denitrification) emissions would occur would depend on the number and genetic repertoires of the invading bacteria. Therefore, vinasse containing an assemblage with a higher partial (containing nirK, nirS and/or norB) to full (containing at least nosZ) denitrifier ratio may lead to higher N_2O emissions during fertirrigation.

Four phyla (Firmicutes, Actinobacteria, Proteobacteria and Bacteroidetes) were represented across the vinasse samples, although at the genus level the diversity of each assemblage fluctuated. The samples could generally be classified as dominated by Megasphaera (A, C, F) or Lactobacillus (B, D, E) at the genus level. The second assemblage (B) was the most diverse; it was dominated by Lactobacillus and containing, uniquely compared to the other time points, Proteobacteria such as Alcaligenes, as well as phage (Lactobacillus phage Lc Nu). The least diverse assemblage was D, containing mostly Lactobacillus and phage. Of the 22 potential vinasse denitrifiers, two were potential complete denitrifiers (containing nirK or nirS, norB and nosZ) and eight were potential incomplete denitrifiers (containing nirK or nirS and norB). The abundances of these potential denitrifiers varied across timepoint, suggesting varied effects on N_2O during vinasse fertirrigation with different vinasses. For example, the Lactobacillus-bin 27 dominated to 97% of the sample D abundance, and this contained a putative nirK gene; when this vinasse is sprayed onto the fields, one would expect nitrate degradation and an increase in N_2O or N2 depending on the gene content of the endogenous microbial community. Another abundant potential denitrifier present in sample A (Prevotella-Bin 2) contained only potential nosZ, suggesting that if the vinasse A were to be used in fertirrigation, the actual emission of N_2O may be reduced due to the further reduction of N_2O into N2. Furthermore, vinasse denitrifiers might directly contribute to the N_2O emissions observed when vinasse is added in conjunction with a nitrate fertilizer [63]. This suggests that vinasse application in conjunction with a reduced nitrogen source such as ammonium sulfate may be a feasible management practice to reduce N_2O production. Further research investigat-

ing the microbes involved in N_2O emissions during fertirrigation with vinasse would greatly aid in steering future vinasse management strategies.

Vinasse fertirrigation has raised human health concerns that vinasse bacteria may carry antibiotic resistance genes (ARGs) [21]. These genes can enter the soil resistome and can be transferred using horizontal gene transfer to other soil bacteria, with potential spreading of antibiotic resistance genes to soil-derived human pathogens. Here a search of the annotation results of the recovered vinasse bins found multidrug resistance genes in 34 of the 36 large bins. Surprisingly, no drug resistance genes were found in the phage bins; this may indicate, that the phages from which these genomes were not prophage that confer auxiliary metabolic genes in the form of antibiotic resistance to the vinasse bacteria. These results warrant further study of the fate of ARG's from vinasse during fertirrigation.

While significant progress has been made in metagenome assembly and binning, some caveats should be noted to the bins we recovered here. Misassembly and misbinning can occur and bias the final results, in our case identifying relevant genes present in the bins. We addressed these issues by comparing three assemblers and three binning tools and choosing the best of each. Further, we used large kmer sizes for the final cross-assembly, and this successfully allowed MaxBin2 to bin to the level of species. We additionally used the manual refinement feature of anvi'o to improve the bins. Because bins with low completeness as determined by the presence of universal marker genes can still contain useful information regarding potential gene content, we used all useful bins to characterize the vinasse assemblage. Eight bins could not be refined, and these represent unbinned vinasse bacterial genome content; however, the information from this genomic material was characterized in the metagenomic analyses. We included several different methods for each analysis to supplement each other as database coverage and read length can bias results based on sequence alignment. Moreover, we used the metagenomic analysis to complement the bin results. Interestingly, comparing the qPCR, putative gene abundances and gene recruitment results suggested that the qPCR primers we used do not cover the entire diversity of vinasse bacteria or alternately that the HMM results may be biased toward false positives.

Here we used metagenomic analysis and genome binning to characterize in depth the assemblage of six vinasse samples from one bioethanol mill. We identified previously unknown vinasse taxa compared to taxa identified through culture- or 16S rRNA survey-based studies of the ethanol processing pipeline steps prior to vinasse. Furthermore, we obtained 21 draft genomes and 8 phage or mobile element genomes from vinasse, which to our knowledge is the first study to do so. Vinasse bacteria included mainly putative denitrifiers, which may directly affect

soil N_2O or N_2 emissions when applied during fertirrigation, although more research is needed into the ecological interactions during this event. In the vinasse bins we found the putative presence of antibiotic resistance genes and genes affecting yeast fermentation, which potentially implicate vinasse bacteria in negative impacts on human health and bioethanol production, respectively. We suggest that monitoring the vinasse assemblage is a promising option to screen both for bioethanol production contaminants and to identify vinasse batches which, when added to the fields during fertirrigation, may lead to higher N_2O emissions. Because of the decreasing costs of high-throughput sequencing, we suggest that monitoring of vinasse assemblages can be widely implemented to improve sugarcane bioethanol production sustainability.

6.5 Declarations

Availability of data and material

The datasets generated and analyzed during the current study are available in the MG-RAST repository with identification numbers described in **Additional file 8**. The metagenomes are additionally found under NCBI BioProject id: PRJNA435511. Draft genomes are available on Zenodo (DOI: 10.5281/zenodo.1194340).

Competing interests

The authors declare that they have no competing interests.

Abbreviations

MGRAST metagenomics analysis server
***nir*K, *nir*S** nitrite reductase genes
***nor*B** nitric oxide reductase gene
***nos*Z** nitrous oxide reductase gene
***amo*A** ammonium monooxygenase gene
ABR antibiotic resistance gene

Authors' contributions

NAC, EEK, HVV designed the study. KSL, JDC, HC collected the data. NAC processed and analyzed the data, and wrote the paper. All the authors contributed with the discussion, read and approved the final manuscript.

Acknowledgements

The authors thank André C. Vitti for his assistance in collecting the vinasse samples and for collecting information on the sugar and ethanol mill pipeline. This work was supported by The Netherlands Organization for Scientific Research (NWO-729.004.003), Sao Paulo State foundation (FAPESP-2013/50365-5, FAPESP BEPE-2014/24141-5, FAPESP-2013/12716-0). Publication number 6473 of the NIOO-KNAW, Netherlands Institute of Ecology.

6.6 Additional files

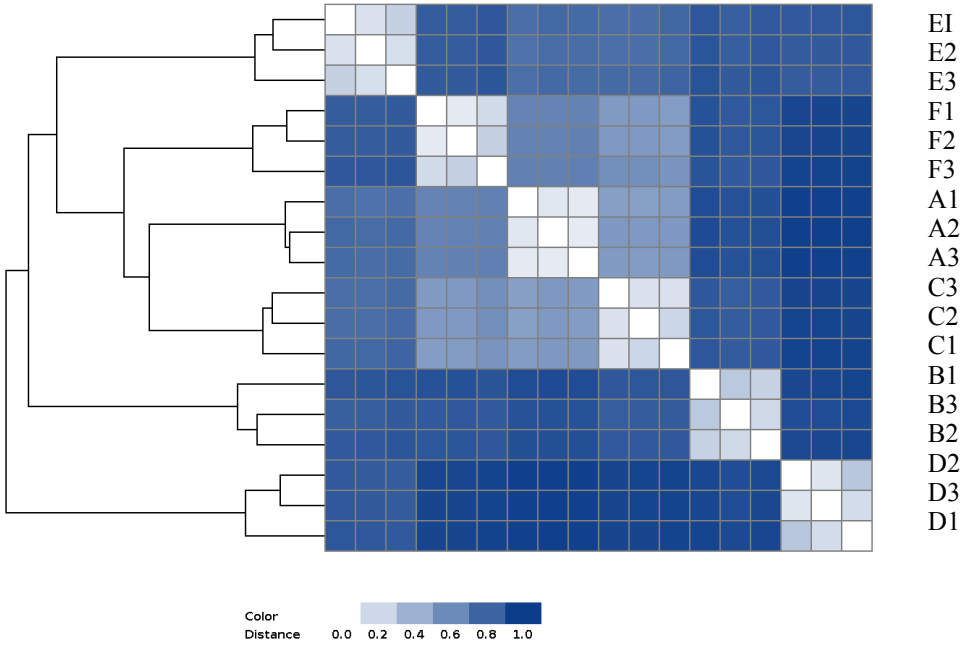
Additional file 1. Data description of the 18 vinasse metagenomes.

Date Sampled	Sample Name	Sample Id	DNA Weight (g)	DNA Conc. (ng/µl)	# reads	Forward # bases (Mbp)	Reverse # bases (Mbp)	Reads mapped to cross-contigs (%)
Nov. 2013	A-1	1V1-1	0.297	32.5	461,784	131	117	92.77
	A-2	1V1-2	0.250	39.6	454,721	130	115	92.69
	A-3	1V1-3	0.295	20.4	469,527	130	115	92.19
Dec. 2013	B-1	1V2-2	0.300	39.2	417,625	110	94	73.20
	B-2	1V2-3	0.295	53.1	517,039	142	131	73.67
	B-3	1V2-4	0.292	38.5	542,208	150	139	74.67
July 2014	C-1	2V1-1	0.301	13.9	362,499	100	89	92.38
	C-2	2V1-2	0.267	14.2	501,511	138	123	92.53
	C-3	2V1-3	0.291	14.4	432,207	119	107	92.12
Aug. 2014	D-1	2V2-1	0.295	17.2	489,336	138	132	95.12
	D-2	2V2-2	0.291	18.6	280,161	77	74	95.33
	D-3	2V2-3	0.294	21.0	351,407	971	935	95.3
Oct. 2014	E-1	3V1-1	0.294	6.19	363,382	981	914	91.30
	E-2	3V1-2	0.280	5.57	434,111	117	108	91.18
	E-3	3V1-3	0.290	7.22	472,732	135	124	91.47
Nov. 2014	F-1	3V2-1	0.294	7.29	444,056	122	114	91.14
	F-2	3V2-2	0.285	6.64	500,636	136	128	91.29
	F-3	3V2-3	0.292	5.53	323,376	883	794	91.66

Additional file 2. Primers and thermocycler conditions used in gene abundance analysis by real time qPCR of the vinasse samples.

Target gene	Primer	Primer Sequence	Size (bp)	12 µL of reaction	Thermal profile
AOA *amoA*	Arch-amoAF	5'-STAATGGTCTGGCTTAGACG-3'	635 491	6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.125 µL of each primer (10 pmol), 1.75 µL of BSA and 4 µL of DNA (3 ng).	95°C-5 min.; 40x 95°C-10s, 64°C-10s, 72°C-20s
	Arch-amoAR	5'-GCGGCCATCCATCTGTATGT-3'		6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.125 µL of each primer (10 pmol) and 4 µL of DNA (3 ng).	95°C-10min.; 40x 95°C-10s, 65°C-25s,
AOB *amoA*	amoA1F	5'-GGGGTTTCTACTGGTGGT-3'			
	amoA2R	5'-CCCCTCKGSAAAGCCTTCTTC-3'			
AOA *amoA*	Arch-amoAF	5'-STAATGGTCTGGCTTAGACG-3'	635	6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.125 µL of each primer (10 pmol), 1.75 µL of BSA and 4 µL of DNA (3 ng).	95°C-5 min.; 40x 95°C-10s, 64°C-10s, 72°C-20s
	Arch-amoAR	5'-GCGGCCATCCATCTGTATGT-3'			
	amoA2R	5'-CCCCTCKGSAAAGCCTTCTTC-3'			
*Nos*Z [3]	nosZ2F	5'-CGCRACGGCAASAAGGTSMSSGT-3'	267	6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.250 µL of each primer (10 pmol), 1.20 µL of BSA and 4 µL of DNA (1.25 ng).	95°C-5 min.; 40x 95°C-10s, 64°C-10s, 72°C-20s
	nosZ2R	5'-CAKRTGCAKSGCRTGGCAGAA-3'			
*nir*K [4]	NirK876	5'-ATYGGCGGVAYGGCGA-3'	165	6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.250 µL of each primer (10 pmol), 1.50 µL of BSA and 4 µL of DNA (1.25 ng).	95°C-5 min.; 40x 95°C-15s, 62°C-15s, 72°C-20s
	NirK1040	5'-GCCTCGATCAGRTTRTGGTT-3'			
*nir*S [5]	nirScd3aF	5'-GTSAACGTSAAGGARACSGG-3'	425	6 µL of Sybrgreen Bioline SensiFAST SYBR non-rox mix, 0.250 µL of each primer (10 pmol), 1.20 µL of BSA and 4 µL of DNA (1.25 ng).	95°C-5 min.; 40x 95°C-10s, 63°C-10s, 72°C-20s
	nirSR3cd	5'-GASTTCGGRTGSGTCTTGA-3'			

1. Francis CA, Roberts KJ, Beman JM, Santoro AE & Oakley BB. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Natl Acad Sci USA. 2005;102:14683–88.
2. Rotthauwe JH, Witzel KP & Liesack W. The Ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-Scale analysis of natural ammonia-oxidizing populations. Appl Environ Microbiol. 1997;63: 4704 12.
3. Henry S, Bru D, Stres B, Hallet S & Philippot L. Quantitative detection of the nosZ gene, encoding nitrous oxide reductase, and comparison of the abundances of 16S rRNA, *nar*G, *nir*K, and *nos*Z genes in soils. App. Environ Microbiol. 2006;72: 5181–89.
4. Henry S, Baudoin E, López-Gutiérrez JC, Martin-Laurent F, Brauman A & Philippot L. Quantification of denitrifying bacteria in soils by *nir*K gene targeted real-time PCR. J Microbiol Methods. 2004;59:327–35.
5. Throbäck IN, Enwall K, Jarvis A & Hallin S. Reassessing PCR primers targeting *nir*S, *nir*K and *nos*Z genes for community surveys of denitrifying bacteria with DGGE. FEMS Microbiol Ecol. 2004;49:401–17.



Additional file 3. Hierarchical clustering of the vinasse metagenomes based on partial de Bruijn graph overlap from Metafast analysis. Replicates were most similar to each other.

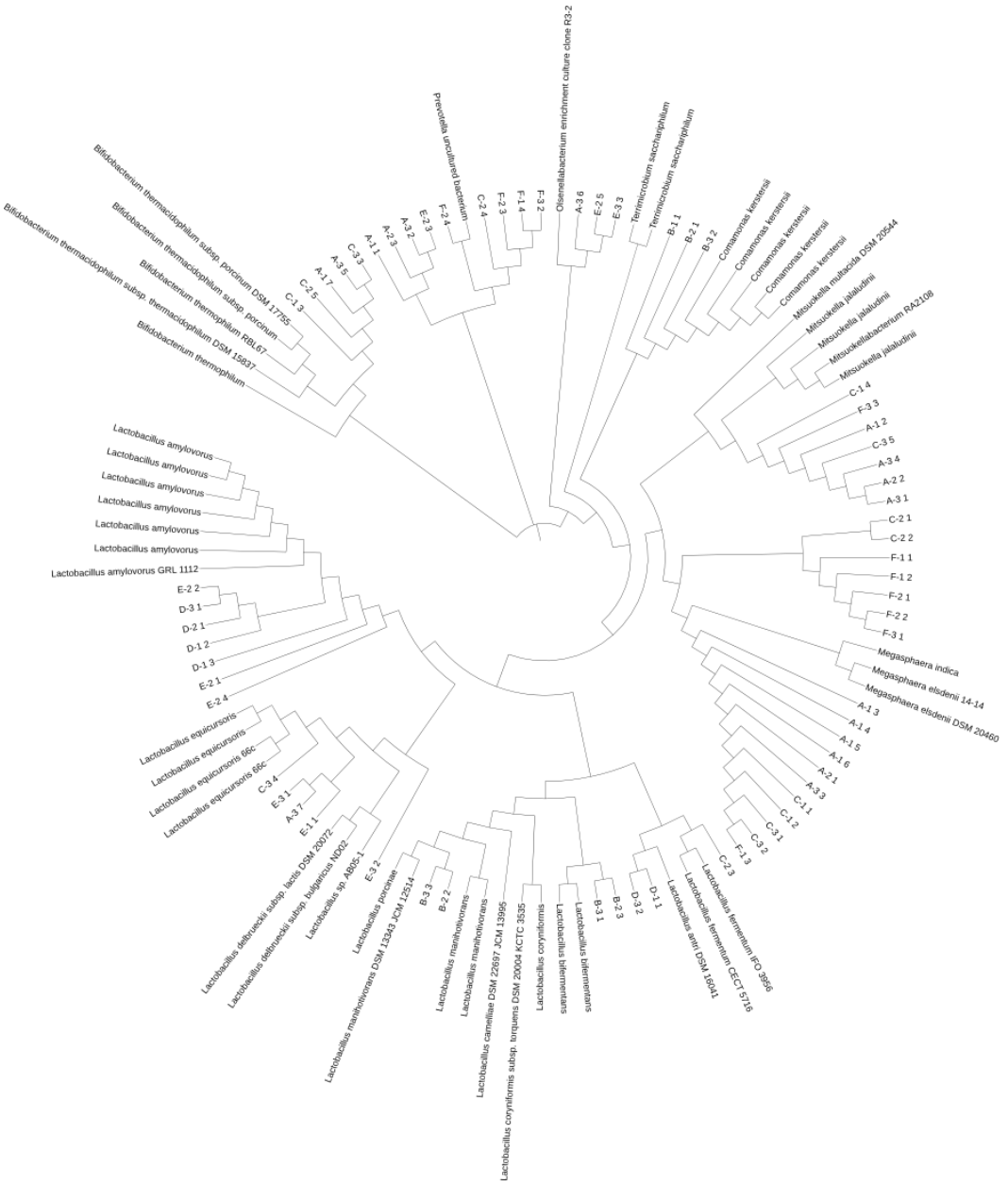
Additional file 4. Data description of the merged vinasse metagenomes uploaded to MG-RAST.

MG-RAST ID	Sample Date	Sample Name	Sample ID	Percent merged	# merged reads	Avg. merged read length	# merged bases (Mbp)
4678764.3	Nov. 2013	A-1	1V1-1	87.98%	236896	592,17	85
4678762.3		A-2	1V1-2	88.72%	230245	592,38	83
4678758.3		A-3	1V1-3	87.40%	245213	590,59	83
4678765.3	Dec. 2013	B-1	1V2-2	93.07%	263198	582,17	86
4678752.3		B-2	1V2-3	95.62%	359330	580,67	116
4678749.3		B-3	1V2-4	95.15%	376765	581,71	124
4678755.3	July 2014	C-1	2V1-1	88.74%	193318	590,16	64
4678760.3		C-2	2V1-2	91.90%	275242	589,16	91
4678754.3		C-3	2V1-3	91.96%	252621	588,15	83
4678766.3	Aug. 2014	D-1	2V2-1	95.97%	402629	581,89	139
4678761.3		D-2	2V2-2	95.88%	235413	577,50	79
4678753.3		D-3	2V2-3	96.49%	300148	576,57	100
4678756.3	Oct. 2014	E-1	3V1-1	94.29%	258649	584,54	84
4678751.3		E-2	3V1-2	92.39%	293786	585,17	96
4678759.3		E-3	3V1-3	91.44%	305881	589,63	111
4678757.3	Nov. 2014	F-1	3V2-1	93.26%	320201	586,67	109
4678763.3		F-2	3V2-2	95.37%	383266	583,43	126
4678750.3		F-3	3V2-3	84.59%	269124	575,92	94

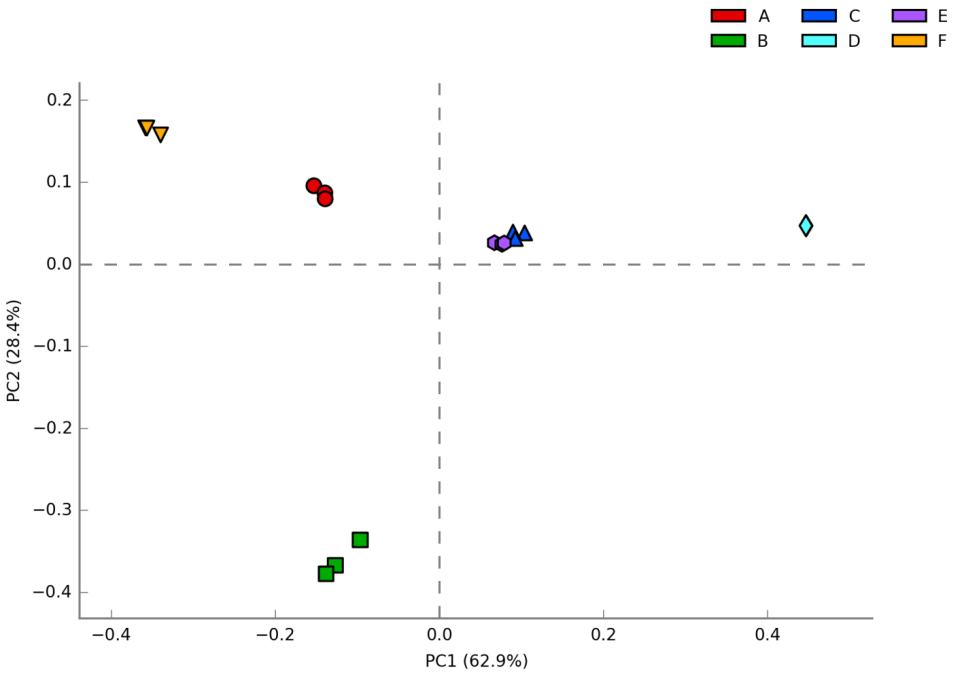
Additional file 5. Taxonomic distribution of the merged vinasse metagenomes from MG-RAST annotation against RefSeq database. Phyla with average relative abundance greater than 1% across all samples were included. Samples with significantly different phyla between groups (Tukey-Kramer post-hoc test, 95% confidence interval, $p < 0.001$) are indicated by different letters.

			Average proportion of sample					
Phylum	p-values (corrected)	Effect size	A	B	C	D	E	F
Firm.	1.94E-14	0.998	44.5±0.4 a	39.2±1.7 b	61.2±0.9 c	97.0±0.0 d	58.8±0.5 e	35.4±0.3 f
Bacter.	4.09E-14	0.998	29.4±1.5 a	9.5±0.2 b	11.3±0.7 c	0.8±0.0 d	11.5±0.3 e	52.8±1.1 f
Actino.	1.88E-09	0.985	15.9±1.5 a	2.1±0.1 b	17.5±1.4 a	0.4±0.0 b	17.8±0.6 a	3.3±0.8 b
Proteo.	7.19E-14	0.997	0.8±0.0 a	39.4±1.8 b	0.8±0.0 a	0.3±0.0 a	1.7±0.0 a	1.4±0.0 a

Effect sizes and corrected p-values were calculated using ANOVA on mean relative abundance of phyla in sample groups using the Benjamini-Hochberg multiple test correction in STAMP.



Additional file 6. Phylogenetic relationships between full-length 16S rRNA genes reconstructed from the vinasse metagenomes using REAGO. The 16S rRNA sequences from vinasse and reference sequences were aligned using ClustalW. The neighbor-joining tree was created with MEGA and visualized using iTol ignoring branch lengths.

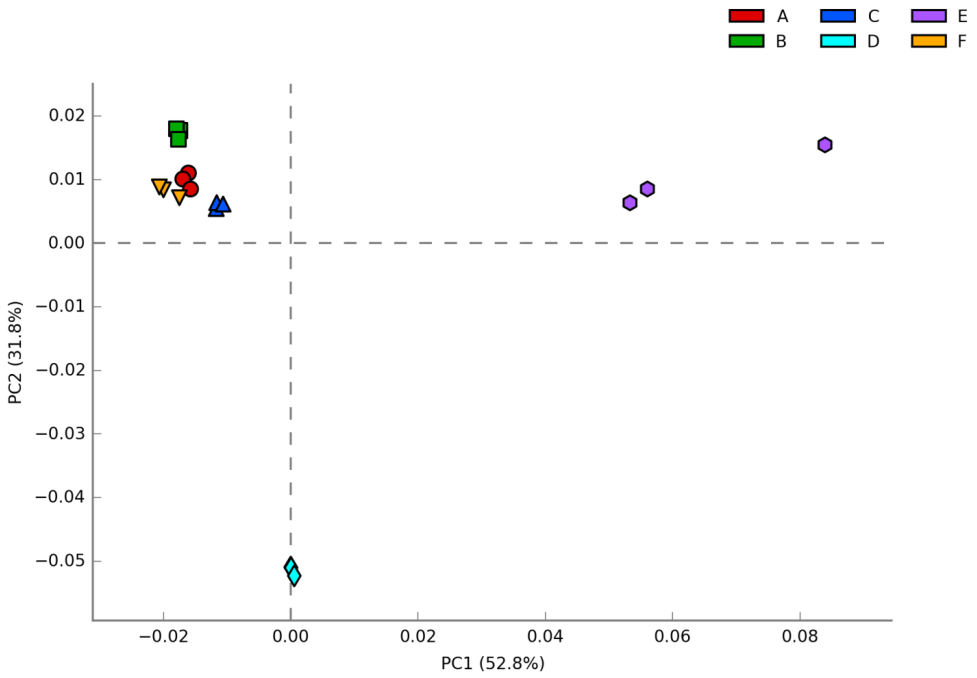


Additional file 7. Principal component analysis of the phylum-level abundance distributions of the vinasse metagenomes. Relative abundance profiles were determined using MG-RAST against the RefSeq database and phyla membership was determined using the Last Common Ancestor algorithm.

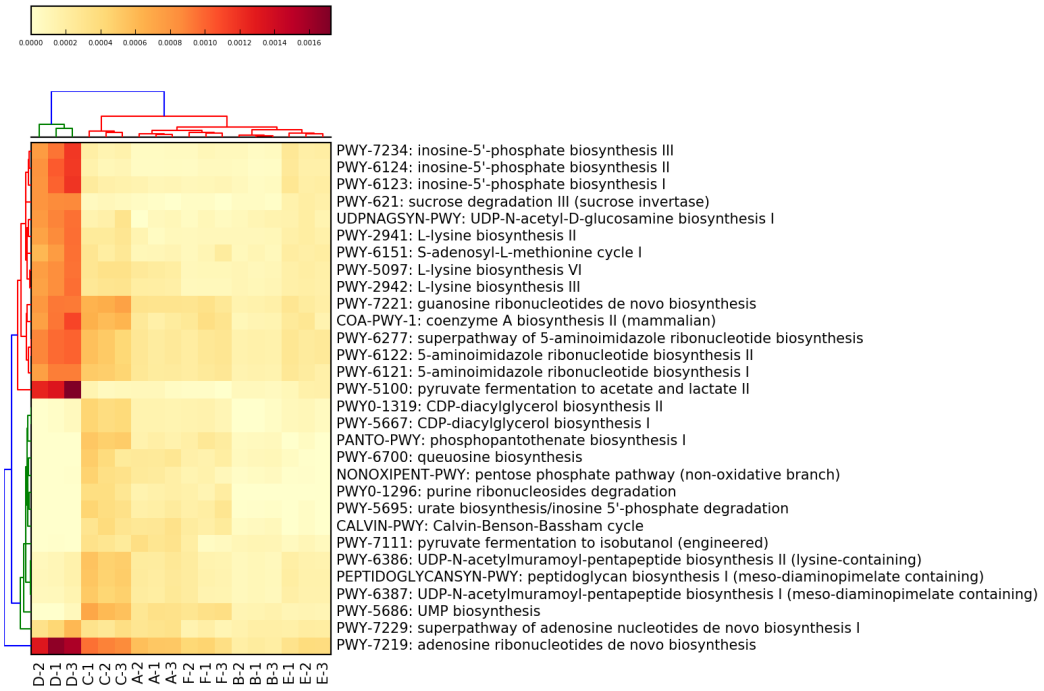
Additional file 8. Functional potential characterization of the vinasse metagenomes from MG-RAST annotation against the Subsystems database**.** Only the subsystems at level 1 with average relative abundance greater than 2% across all samples were included. Significantly different subsystems at level 1 between sample groups (Tukey-Kramer post-hoc test, 95% confidence interval, $p < 0.05$) are indicated by different letters.

			Relative sample abundance					
Subsystems Level 1 Category	p-values (corrected)	Effect size	A	B	C	D	E	F
Carbohydrates	2.84E-08	0.967	16.3±0.3 ac	12.9±0.2 b	15.5±0.1 a	15.5±0.1 a	13.7±0.3 b	17.0±0.5 c
Clustering-based subsystems	4.37E-09	0.977	14.5±0.2 a	14.6±0.0 a	15.2±0.1 b	17.2±0.0 c	14.8±0.2 ab	14.2±0.2 a
Amino Acids and Derivatives	6.35E-13	0.997	9.2±0.1 ac	9.4±0.0 a	9.3±0.1 a	4.7±0.1 b	7.0±0.2 d	8.9±0.0 c
Miscellaneous	8.50E-05	0.863	7.6±0.1 a	7.5±0.0 a	7.6±0.1 a	7.3±0.0 b	7.0±0.2 c	7.3±0.1 bc
Protein Metabolism	2.25E-08	0.969	7.1±0.1 a	6.9±0.1 a	7.8±0.1 c	8.3±0.1 d	7.0±0.1 ab	6.7±0.2 b
DNA Metabolism	3.86E-11	0.991	5.5±0.1 a	5.4±0.0 a	5.4±0.1 a	7.3±0.1 c	5.8±0.0 b	5.7±0.0 b
RNA Metabolism	3.43E-07	0.949	5.7±0.1 a	5.7±0.1 a	5.9±0.1 a	7.0±0.1 c	5.7±0.2 a	5.1±0.1 b
Cofactors, Vitamins	3.71E-11	0.991	5.6±0.0 a	5.8±0.1 a	5.1±0.1 b	3.5±0.0 c	4.0±0.2 d	5.2±0.0 b
Cell Wall and Capsule	3.04E-06	0.923	4.7±0.1 ac	4.4±0.1 bc	4.6±0.1 c	4.9±0.1 a	4.2±0.1 b	4.9±0.0 a
Phages and Prophages	8.82E-08	0.960	2.9±0.1 a	2.3±0.0 a	3.1±0.0 a	2.4±0.0 a	10.0±1.4 b	2.6±0.1 a
Nucleosides and Nucleotides	4.42E-10	0.984	3.2±0.0 a	2.7±0.0 b	3.3±0.0 a	3.7±0.1 c	4.1±0.1 d	2.9±0.1 b

Effect sizes and corrected p-values were calculated using ANOVA on mean relative abundance of Subsystems level 1 in sample groups using the Benjamini-Hochberg multiple test correction in STAMP.



Additional file 9. Principal component analysis of the Subsystems Level 1 category abundances of the vinasse metagenomes. The colors correspond to time point. Profiles were determined against the Subsystems database using MG-RAST and relative abundances of phyla were calculated out of the total number of sample reads.



Additional file 10. Functional potential profiles of the top 30 pathways across the vinasse samples, excluding “unmapped” and “uncategorized” results. The functional group and sample profiles were clustered using hclust2 from humann2 analysis against the UniRef90 database.

Additional file 11. All vinasse bin characteristics and relative sample abundances (indicated by heatmap per sample). Bin id's highlighted in green indicate "good" bins; yellow id's indicate "interesting" bins, and red id's indicate "bad" bins. Continued on next page.

Bin Id	A	B	C	D	E	F	Length (Mbp)	# Contigs	N50	GC (%)	Completeness (%)	Redundancy (%)
1	6	1	2	0	0	1	2.4	209	23171	53	92	2
2	14	2	1	0	4	3	3.5	547	10519	49	94	4
3	8	1	18	0	2	13	2.2	352	10534	53	97	2
5	3	0	4	0	2	1	1.9	298	11078	60	94	3
6	4	1	2	0	1	12	2.4	539	6347	44	91	2
7	9	2	3	0	4	2	4.9	2521	1944	51	63	40
9	2	0	2	0	1	0	2.0	407	6384	60	90	6
10	1	1	1	1	3	0	2.1	183	27583	47	96	1
11	7	2	4	0	2	10	5.8	2606	2396	44	36	9
12	3	0	2	0	5	1	2.3	485	7489	66	91	5
13	2	0	2	0	0	0	1.5	605	2710	63	71	12
14	2	0	1	0	0	1	1.9	947	2190	52	74	15
15	2	0	2	0	1	1	2.2	1017	2297	53	76	9
16	1	1	1	0	0	13	3.0	307	22370	42	99	1
17	2	0	1	0	1	0	1.4	928	1434	60	43	14
18	2	0	1	0	2	0	1.5	893	1665	63	69	22
19	1	0	1	0	3	0	2.0	917	2351	62	60	12
20	0	0	1	0	1	1	1.2	387	3766	54	64	2
21	0	0	1	1	2	1	1.8	298	9992	50	88	1
22	2	1	3	0	1	12	5.3	2375	2491	36	39	7
23	0	1	1	3	2	1	1.9	373	7041	47	95	4
24	0	2	1	0	1	0	1.8	220	13204	53	98	4
25	1	1	1	1	0	2	1.8	729	2822	40	91	9
26	1	1	2	0	1	1	2.1	1236	1733	41	66	15
27	0	1	1	45	3	1	1.9	262	11858	38	99	1
28	0	3	0	0	0	0	2.1	340	8670	38	96	5
29	0	1	0	2	0	0	1.0	439	2658	50	88	9
30	0	3	1	3	2	3	3.9	2021	1897	31	47	16
31	0	0	8	0	0	0	3.3	1595	2241	54	79	37
32	0	0	0	6	0	0	2.0	104	208993	36	99	1
33	0	0	0	1	3	0	1.7	447	4850	48	96	9
34	0	3	0	0	0	0	2.7	343	12289	60	92	1
35	0	4	0	0	0	0	2.7	259	13750	43	96	6
36	0	5	0	0	0	0	1.7	647	2989	49	67	5
36.2	0	4	0	0	0	0	1.4	868	1558	48	38	5
36.3	0	1	0	0	0	0	0.5	245	2503	48	10	0
37.1	0	10	0	0	0	0	3.1	1326	2603	57	77	24
37.2	0	4	0	0	0	0	1.2	784	1533	58	34	6
37.3	0	2	0	0	0	0	0.7	506	1236	49	16	1

Bin Id	A	B	C	D	E	F	Length (Mbp)	# Contigs	N50	GC (%)	Completeness (%)	Redundancy (%)
38	0	9	0	0	0	0	3.0	488	9382	60	89	3
39	0	0	0	4	0	0	1.9	205	15510	47	99	4
39.2	0	0	0	0	0	0	0.1	76	1387	47	5	1
40.1	0	4	0	0	0	0	1.6	190	11919	28	97	2
40.2	0	2	0	0	0	0	0.6	271	2044	27	15	1

Additional file 12. All vinasse bin taxonomic affiliations based on CAT. Bin id's highlighted in green indicate "good" bins; yellow id's indicate "interesting" bins, and red indicate "bad" bins. Continued on next page.

Bin Id	Kingdom	Phylum	Class	Order	Family	Genus	Species
1	B	F	Nega	Selenomonadales	Veillonellaceae	Mitsuokella	Unclassified
2	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	P. mutisaccharivorax/Unclassified
3	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera	Unclassified/M. elsdenii
4.1	V				Caudovirales	Siphovirales	Lactobacillus phage Ldl1
4.2	B	U/B	U/B	U/B	U	U	U
5	B	A	Acti	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	Unclassified
6	B	B	Bact	Bacteroidales	Prevotellaceae/Unclassified	Prevotella/Unclassified	Unclassified
7	B	A/B/F					
8.1	B/V	U	U	U	U/Caudovirales	U/Siphoviridae	U/Lactobacillus phage LdL1
8.2	B	F	B	Lactobacillales	Lactobacillaceae	Lactobacillus	U
8.3	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	P. histicola/U
8.4	B	F	Bac	Lactobacillales	Lactobacillaceae	Lactobacillus	U
8.5	B	B	U/Bact				
8.6	B	F	Bac	Lactobacillales	Lactobacillaceae	Lactobacillus	U
9	B	U/A					
10	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. equicursoris/Unclassified
11	B	B/F/U					
12	B	A/U					
13	B	U/A					
14	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera	U/sp. DJF_B143
15	B	F	Nega	Selenomonadales	Veillonellaceae	Dialister	U/D. succinatiphilus
16	B	B	Bact	Bacteroidales	Prevotellaceae	Prevotella	U
17	B	A/F					
18	U/B						
19	B	U/A					
20	B	F	Clos	Clostridiales	Eubacteriaceae/U	Pseudoramibacter/U	U/P. alactolyticus
21	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. delbrueckii/U
22	B	B/F					

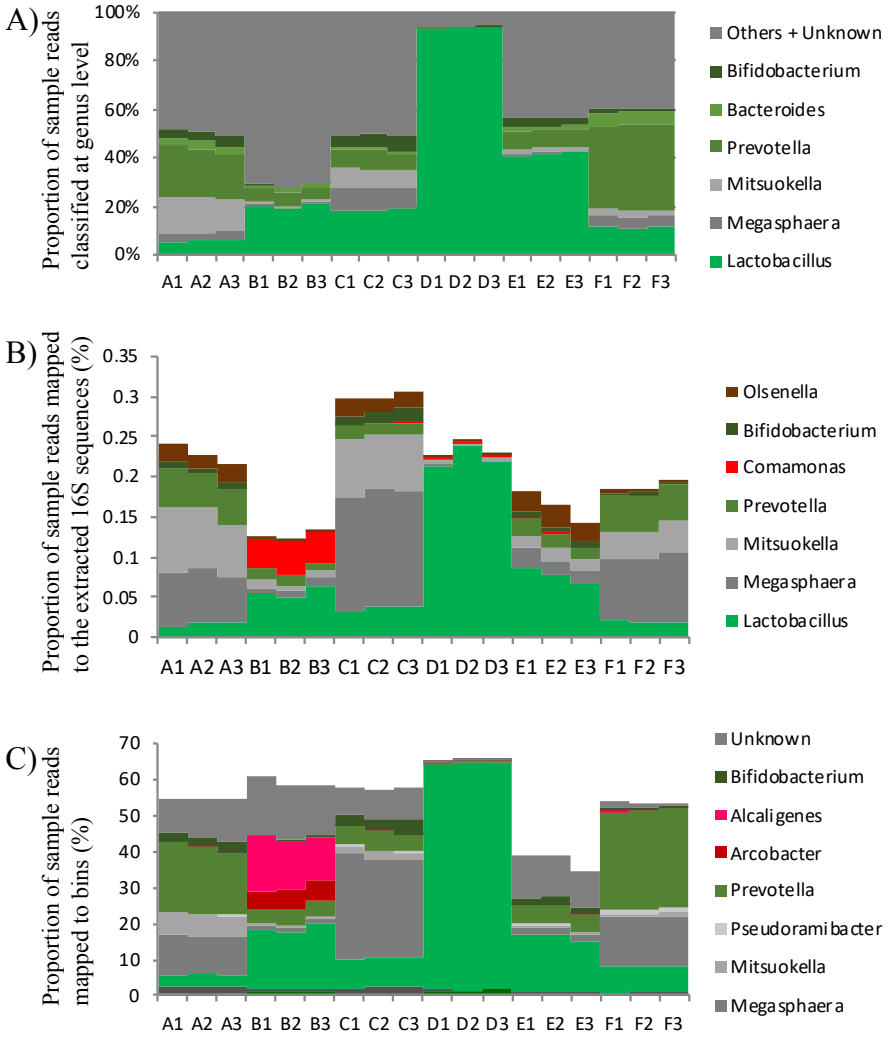
F = *Firmicutes*, B = *Bacteroidetes*, A = *Actinobacteria*, P = *Proteobacteria*, E = Euryarchaeota, U = Unknown, Bact = *Bacteroidia*, Nega = *Negativicutes*, Clos = *Clostridia*, Mega = *Megasphaera*, Lact = *Lactobacillales*

Additional file 12 cont'd.

Bin Id	Kingdom	Phylum	Class	Order	Family	Genus	Species
23	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. mucosae/U
24	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. fermentum/U
25	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U/L. vini
26	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U/L. agilis
27	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	U
28	B	F	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	L. vini
29	B	F	Clos	Clostridiales	Clostridiaceae	Clostridium	sp. CAG:568/U
30	B/A	F/E	Bacilli/Metha	Bacteroidales/Methanobacteriaceae	Lactobacillaceae/Methanobactericeae	Lactobactillus/Methanobrevibacter	U
31	B	F	Nega	Selenomonadales	Veillonellaceae	Megasphaera/U	U
32	B	F	Nega	Lactobacillales	L	L	U
33	B	F	B	Lactobacillales	L	L	L. secaliphilus/U
34	B	F	B	Lactobacillales	L	L	U/L. manihotivorans
35	B	F	B	Lactobacillales	L	L	L. bifermentans/U
36	B	P	Beta	Burkholderiales	Alcaligenaceae	U	U
36.2	B	P/F					
36.3	B	F	B	Lactobacillales	L	L	U/L. manihotivorans
37.1	B	P	B	Burkholderiales	Alcaligenaceae	Alcaligenes	A. faecalis/U
37.2	B	P	B	B	A	Alcaligenes	U/A. faecalis
37.3	B/A	P/E					
38	B	P	B	B	Comamonadaceae	U/Comamonas	U/C. kerstersii
39	B	F	B	Lactobacillales	L	L	U/L. panis
39.2	B	F	B	Lactobacillales	L	L	L. panis/L. ponis/U
40.1	B	P	E	Campylobacterales	Campylobacteraceae	Arcobacter	A. skirrowii/U
40.2	B/A	P/E	E/Metha	Campylobacterales/Methanobacteriales	Campylobacteraceae/Methanobacteriaceae	Arcobacter/Methanobrevibacter	U

F = *Firmicutes*, B = *Bacteroidetes*, A = *Actinobacteria*, P = *Proteobacteria*, E = Euryarchaeota, U = Unknown, Bact = *Bacteroidia*, Nega = *Negativicutes*, Clos = *Clostridia*, Mega = *Megasphaera*, Lact = *Lactobacillales*

Additional file 13. Comparison of genus abundances across samples from the A) MGRAST, B) extracted 16S and the C) bin taxonomy results. The colors correspond to genus and phyla as Phylum Firmicutes (green), Proteobacteria (red), Actinobacteria (brown) and Bacteroidetes (orange).



6.6 References

1. Amorim,HV, Lopes ML, de Castro Oliveira JV, Buckeridge MS, Goldman GH. Scientific challenges of bioethanol production in Brazil. App Microbiol Biotech. 2011;91:1267-1275.
2. Christofoletti CA, Escher JP, Correia JE, Marinho JFU, Fontanetti CS. Sugarcane vinasse: environmental implications of its use. Waste Manage. 2013;33:2752-2761.
3. Parnaudeau V, Condom N, Oliver R, Cazevieille P, Recous S. Vinasse organic matter quality and mineralization potential, as influenced by raw material, fermentation and concentration processes. Biores Tech. 2008;99:1553-1562.
4. Moore CCS, Nogueira AR, Kulay L. Environmental and energy assessment of the substitution of chemical fertilizers for industrial wastes of ethanol production in sugarcane cultivation in Brazil. Intl J Life Cycle Ass. 2017;22:628-643.
5. Jiang ZP, et al. Effect of long-term vinasse application on physico-chemical properties of sugarcane field soils. Sugar Tech. 2012;14:412-417.
6. Zhou M, Luo Y, Zhou Z, Gong D, Zhou X. The effect of alcohol waste liquid (as a top dressing) on growth and yield of sugarcane. Guizhou Ag Sci. 2008;36:102-103.
7. YunChuan M, YanPing Y, Qiang L, YangRui L. Effects of vinasse on the quality of sugarcane and key enzymes in sucrose synthesis. SW China J Agric Sci. 2009;22:55-59.
8. Yang SD, Liu JX, Wu J, Tan HW, Li YR. Effects of vinasse and press mud application on the biological properties of soils and productivity of sugarcane. Sugar Tech. 2013;15:152-158.
9. Navarrete AA, et al. Multi-analytical approach reveals potential microbial indicators in soil for sugarcane model systems. PloS One. 2015;10:e0129765.
10. do Carmo JB et al. Infield greenhouse gas emissions from sugarcane soils in Brazil: effects from synthetic and organic fertilizer application and crop trash accumulation. GCB Bioenergy. 2013;5:267-280.
11. Pitombo LM et al. Exploring soil microbial 16S rRNA sequence data to increase carbon yield and nitrogen efficiency of a bioenergy crop. GCB Bioenergy. 2015.
12. Moran-Salazar R, et al. Utilization of vinasses as soil amendment: consequences and perspectives. SpringerPlus. 2016;5:1-11.
13. Rein P. Proc S African Sugar Tech Ass. 196-200.
14. Lopes ML, et al. Ethanol production in Brazil: a bridge between science and industry. Brazilian J Microbiol. 2016;47:64-76.
15. Costa OY, et al. Microbial diversity in sugarcane ethanol production in a Brazilian distillery using a culture-independent method. J Ind Microbiol Biotechnol. 2015;42:73-84.
16. Brexó RP, Santana AS. Impact and significance of microbial contamination during fermentation for bioethanol production. Ren Sust Energy Rev. 2017a;73:423-434.
17. Cabrini KT, Gallo CR. Yeast identification in alcoholic fermentation process in a sugar cane industry unit of São Paulo state, Brazil. Scientia Agricola. 1999;56:207-216.
18. Lucena BT, et al. Diversity of lactic acid bacteria of the bioethanol process. BMC Microbiol. 2010;10:298.
19. de Souza RB, et al. The consequences of *Lactobacillus* vini and *Dekkera bruxellensis* as contaminants of the sugarcane-based ethanol fermentation. J Ind Microbiol & Biotech. 2012;39:1645-1650.
20. Alcarde V, Yokoya F. Effect of the bacterial population on flocculation of yeasts isolated from industrial processes of alcoholic fermentation. STAB-Açúcar, Álcool e Subprodutos. 2003;21:40-42.
21. Braga LP, et al. Vinasse fertirrigation alters soil resistome dynamics: an analysis based on metagenomic profiles. BioData Mining. 2017;10;17.
22. Antonangelo ATB, Alonso DP, Ribolla PE, Colombi, D. Microsatellite marker−based assessment of the biodiversity of native bioethanol yeast strains. Yeast. 2013;30:307-317.
23. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27:863-864.

24. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30:614-620.
25. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biology. 2015;16:51.
26. Ulyantsev VI, Kazakov SV, Dubinkina VB, Tyakht AV, Alexeev DG. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. Bioinformatics. 2016;32:2760-2770..
27. Meyer F, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386.
28. Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. Bioinformatics. 2010;26:715-721.
29. Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811-814.
30. Abubucker S, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comp Biol. 2012;8:e1002358.
31. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015;3:e1029.
32. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23:1282-1288.
33. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics. 2015;31:35-43.
34. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:590-596.
35. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823-1829.
36. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol & Evol. 2016;33:1870-1874.
37. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molec Biol & Evol. 1987;4:406-425.
38. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Nat Aca Sci USA. 2004;101:11030-11035.
39. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evol. 1985:783-791.
40. Li D, et al. ISBRA 2016, Minsk, Belarus. Proceedings June 5-8, 2016. 309 (Springer).
41. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012;13:R122.
42. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674-1676.
43. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comp Biol. 2012;19:455-477.
44. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072-1075.
45. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-359.
46. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
47. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605-607.
48. Cambuy DD, Coutinho FH, Dutilh BE. Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. bioRxiv. 2016:072868.

49. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043-1055.
50. Eren AM, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
51. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068-2069.
52. Moraes BS, Zaiat M, Bonomi A. Anaerobic digestion of vinasse from sugarcane ethanol production in Brazil: Challenges and perspectives. Ren & Sust Energy Rev. 2015;44:888-903.
53. Reis RCE, Hu B. Vinasse from sugarcane ethanol production: better treatment or better utilization? Front Energy Res. 2017;5:7.
54. Botellho RG, Christofoletti CA, Correira JE, Tornisielo VL. Environmental implications of using waste from sugarcane industry in agriculture. Prod Cons & Agr Manage. 2014;91.
55. Beckner, M., Ivey, M.L., Phister, T.G. (2011) Microbial contamination of fuel ethanol fermentations. Letters in applied microbiology 53, 387-394.
56. Bonatelli, M.L., Quecine, M.C., Silva, M.S., Labate, C.A. (2017) Characterization of the contaminant bacterial communities in sugarcane first-generation industrial ethanol production. FEMS microbiology letters 364.
57. Brexó, R.P., Sant'Ana, A.d.S. (2017b) Microbial interactions during sugar cane must fermentation for bioethanol production: does quorum sensing play a role? Critical Reviews in Biotechnology, 1-14.
58. Solomon E B, Okull D. (Google Patents, 2008).
59. Roach DR, Khatibi PA, Bischoff KM, Hughes SR, Donovan DM. Bacteriophage-encoded lytic enzymes control growth of contaminating Lactobacillus found in fuel ethanol fermentations. *Biotechnology for biofuels* **6**, 20 (2013).
60. De Souza RSC, et al. Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. Sci Rep. 2016;6:28774.
61. Wallenstein MD, Myrold DD, Firestone M, Voytek M. Environmental controls on denitrifying communities and denitrification rates: insights from molecular methods. Ecol Appl. 2006;16:2143-2152.
62. Philippot L, Andert J, Jones CM, Bru D, Hallin S. Importance of denitrifiers lacking the genes encoding the nitrous oxide reductase for N_2O emissions from soil. Glob Chang Biol. 2011;17:1497-1504.
63. Shade A, et al. Fundamentals of microbial community resistance and resilience. Front Microbiol. 2012;3:417.