Accepted Manuscript

Title: Bioinformatics for NGS-based Metagenomics and the Application to Biogas Research

Authors: Sebastian Jünemann, Nils Kleinbölting, Sebastian Jaenicke, Christian Henke, Julia Hassa, Johanna Nelkner, Yvonne Stolze, Stefan P. Albaum, Andreas Schlüter, Alexander Goesmann, Alexander Sczyrba, Jens Stoye

PII: S0168-1656(17)31598-5
DOI: http://dx.doi.org/10.1016/j.jbiotec.2017.08.012
Reference: BIOTEC 7992

To appear in: *Journal of Biotechnology*

Received date: 20-3-2017
Revised date: 8-8-2017
Accepted date: 9-8-2017

Please cite this article as: Jünemann, Sebastian, Kleinbölting, Nils, Jaenicke, Sebastian, Henke, Christian, Hassa, Julia, Nelkner, Johanna, Stolze, Yvonne, Albaum, Stefan P., Schlüter, Andreas, Goesmann, Alexander, Sczyrba, Alexander, Stoye, Jens, Bioinformatics for NGS-based Metagenomics and the Application to Biogas Research.Journal of Biotechnology http://dx.doi.org/10.1016/j.jbiotec.2017.08.012

Bioinformatics for NGS-based Metagenomics and the Application to Biogas Research

Sebastian Jünemann[1,3†*], Nils Kleinbölting[1†], Sebastian Jaenicke[1,2], Christian Henke[1], Julia Hassa[1], Johanna Nelkner[1], Yvonne Stolze[1], Stefan P. Albaum[1], Andreas Schlüter[1], Alexander Goesmann[2], Alexander Sczyrba[1,3] and Jens Stoye[1,3]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

[2]Bioinformatics and Systems Biology, Justus-Liebig-Universität, Gießen, Germany

[3]Faculty of Technology, Bielefeld University, Bielefeld, Germany

[†]Contributed equally

[*]Corresponding author, jueneman@cebitec.uni-bielefeld.de

Highlights

- Introduction into 16S rRNA gene amplicon analysis, read- and assembly-based whole genome shotgun metagenomics, suitable bioinformatics approaches and tools.
- Presentation of bioinformatic workflows for metagenomics exemplarily illustrated by applications on microbial community analysis of biogas fermenters.
- EMGB and MGX, two comprehensive bioinformatic platforms for metagenomics provided by de.NBI (BiGi).

Abstract

Metagenomics has proven to be one of the most important research fields for microbial ecology during the last decade. Starting from 16S rRNA marker gene analysis for the characterization of community compositions to whole metagenome shotgun sequencing which additionally allows for functional analysis, metagenomics has been applied in a wide spectrum of research areas. The cost reduction paired with the increase in the amount of data due to the advent of next-generation sequencing led to a rapidly growing demand for bioinformatic software in metagenomics. By now, a large number of tools that can be used to analyze metagenomic datasets has been developed. The Bielefeld-Gießen center for microbial bioinformatics as part of the German Network for Bioinformatics Infrastructure bundles and imparts expert knowledge in the analysis of metagenomic datasets, especially in research on microbial communities involved in anaerobic digestion residing in biogas reactors. In this review, we give an overview of the field of metagenomics, introduce into important bioinformatic tools and possible workflows, accompanied by application examples of biogas surveys successfully conducted at the Center for Biotechnology of Bielefeld University.

Keywords: metagenomics; high-throughput 16S rRNA gene amplicon sequencing; next-generation sequencing; biogas; anaerobic digestion; de.NBI

1. Introduction

Metagenomics addresses the analysis of the genomic content of complete microbial communities and provides insights into their structure and function, thereby yielding information on organisms that cannot easily be cultured (Handelsman et al., 1998). In the past decade, metagenomics based on next-generation sequencing (NGS) data became a rapidly growing research field in microbiomics portrayed most impressively by ambitious consortia projects investigating habitats as complex and diverse as the human body (Turnbaugh et al., 2007) or even the earth (Gilbert et al., 2014). To keep pace with this rapid progress, bioinformatics grew similarly offering by now a myriad of algorithms, tools, and software suites to address all kinds of tasks related to NGS based metagenomics ranging from pre-processing, analysis, evaluation, and graphical representation. Within the German Network for Bioinformatics Infrastructure (de.NBI), an effort to centralize, harmonize, and ease the use of bioinformatic services for the research community, the Bielefeld-Gießen center for microbial bioinformatics (BiGi) can look back on a decade of applied metagenomics and the development of bioinformatics solutions related to this research field.

The purpose of this review is to give an overview of metagenomics in general, its evolvement along the past decade with respect to different sequencing technologies, and the bioinformatics solutions and workflows for different analysis approaches. Appropriate examples on the microbial analysis of biogas fermenters illustrate the applicability of presented technologies and workflows. Therefore, this review is structured as follows: After an introduction to metagenomics, insights into anaerobic digestion and biogas plants will be given, covering important aspects such as the operating principle, the involved metabolic processes, the microbial composition, and the main research goals. Building on data driven application scenarios, the bioinformatic reviewing section is divided into two subsections starting with 16S rRNA gene based community analysis (also often designated as targeted metagenomics), followed by whole genome shotgun metagenomics, including read-based approaches and more recent metagenome assembly and binning strategies. In the following section, both approaches are exemplified by applications from Bielefeld biogas research. New solutions offered to the community by BiGi for analyzing metagenomic samples following both read and assembly based approaches will be presented thereafter. Finally, a short outlook to future developments is given in the concluding remarks.

Even though this review is build upon the example of biogas research, it is to be noted that the herein presented technologies as well as the bioinformatic solutions and workflows are, of course, applicable to other habitats and ecosystems.

1.1 Sequencing and Metagenomics: a historical outline

It is widely accepted that the microbial diversity is tremendous and that the vast majority of these organisms, approximately 99%, can not easily be cultured and therefore not studied using traditional approaches (Streit and Schmitz, 2004). Cultivation-independent approaches, such as denaturing gradient gel electrophoresis (Ferris et al., 1996; Muyzer et al., 1993,), terminal restriction fragment length polymorphisms (Liu et al., 1997; Osborn et al., 2000), fluorescence in situ hybridization (Antón et al., 1999; Llobet-Brossa et al., 1998) and various types of PCR like quantitative real-time PCR have been used for decades to analyze single organisms or whole environmental communities culture-free (Fierer et al., 2005; Kolb et al., 2003). But these techniques are limited with respect to the resolution they can be applied to, i.e. the information detail on the whole microbial consortia and their individual members achieved with those techniques are limited, thus rendering a deep description of complex or diverse communities challenging. With the beginning of the new millennium, pioneering culture-free whole-genome shotgun (WGS) sequencing efforts have been made to investigate habitats previously thought to be almost inaccessible due to their extreme conditions or rich diversity, as for example the acid mine drainage (Tyson et al., 2004) and the Sargasso Sea (Venter et al., 2004). Albeit being path-breaking projects, they rested on direct cloning of environmental DNA (Pace et al., 1985) in conjunction with capillary sequencing which still was expensive and time-consuming to conduct at high depths at this time. However, the advent of NGS technologies shortly thereafter revolutionized the sequencing market and sequence based research, thus leading to a dramatic drop in sequencing costs while simultaneously simplifying the sequence library preparation by omitting the cloning step altogether. This paved the way for sequence-based analysis of all kinds of environments and microorganisms, most notably those that are impossible to grow in pure culture, as for instance symbionts or fossilized remains of ancient organisms. In NGS, DNA material extracted from an organism or an environmental sample is fragmented by WGS fragmentation and then each fragment is sequenced on a micrometer scale in parallel where each individual sequencing reaction is recorded separately. The actual advantage over traditional Sanger capillary electrophoresis sequencing is that NGS instruments can carry out millions of sequencing reactions in a high-throughput massively parallel manner and are capable to decipher all DNA contained in a mixture in one sequencing effort. The first NGS instrument broadly available was the GS20 sequencing-by-synthesis (SBS) pyrosequencing device by 454 Life Science in 2004 (a Roche company since 2007) and its successor the 454 GS FLX in 2005, offering a read length of 100 to 150 bp and a throughput of 20 Mb per run. The 454 platform underwent a continuous improvement with the 454 GS FLX Titanium upgrade in 2008 and the 454 GS FLX+ in 2009, capable of producing over 600 Mb in a single run with a read length of up to 1 kbp. Another push to the sequencing market was achieved by Solexa releasing the Genome Analyzer in 2006 as a direct competitor to the GS FLX. This SBS instrument, resembling in essence the Sanger procedure by utilizing fluorescent reversible terminator nucleotides, offered 1 gigabase (Gb) of data in one run, yet with a very short read length of 35 bp but with the possibility to sequence paired-end (PE) data, effectively doubling the read length. Illumina, which purchased Solexa in 2007, developed a wide range of improvements and instruments from mid to large scale output in the following years (e.g., the GAIIx, the HiSeq family, and the popular benchtop variant MiSeq). To date, Illumina's sequencing instruments are without doubt the most widely used on the market due to their superior per-base cost efficiency and high sequencing accuracy. Among their platforms they achieve read lengths between 2x125 and 2x300 bp and a throughput between multiple Gb (in a few hours) and several terabases (in almost a week).

Of course, several other companies released sequencing instruments of their own, for example Pacific Bioscience imaging fluorescent nucleotides during the single molecule real-time SBS process with their RS and RS II instruments, the sequencing by ligation of dinucleotides in the colour-space operating SOLiD sequencers marketed by Applied Biosystems (meanwhile acquired by Thermo Fisher Scientific), the ion semiconductor chip

sequencing method measuring pH changes while performing SBS as part of Thermo Fisher's, formerly Life Technologies', Ion Torrent sequencers, the so-called third-generation sequencing platforms as for example Oxford Nanopore's MinION performing 'strand-sequencing' by utilizing ionic current sensing of single unlabeled ssDNA molecules passing through a protein nanopore, and many more (Goodwin et al., 2016; Hodkinson and Grice, 2015; Kulski, 2016).

This overall development contributed to the leverage of sequencing based research in all areas of life science. However, as the technology grew faster than their computational counterpart, in fact outperforming Moore's Law many times over, unforeseen amounts of sequencing data were generated that posed a real challenge for data analysis as well as for bioinformatics in general. Sophisticated bioinformatic tools able to process millions of reads in reasonable time were not required before and therefore simply not available. Thus, the main focus of early NGS based metagenomic studies with respect to bioinformatic processing was in essence handling these massive amounts of data. Thereby, processing pipelines were kept simple, i.e. getting from raw sequencing reads to results directly, by conducting extensive BLAST (Camacho et al., 2009) searches on potent compute clusters to compare either 16S rRNA gene sequence tags to ribosomal databases or to detect environmental gene tags (EGTs) by searching WGS reads against various nucleotide or protein databases (Angly et al., 2006; Edwards et al. 2006; Gill et al., 2006; Turnbaugh et al., 2006). Another difficulty with the emerged NGS instruments were their inherent platform specific error profiles. Possible biases and artifacts introduced during the sequencing process were not sufficiently understood. However, it could be shown that they can have, if left untreated, a falsifying impact on the study results (Berry et al., 2011; Dohm et al., 2008; Gomez-Alvarez et al., 2009; Huse et al., 2007; Kunin et al., 2010; Nakamura et al., 2010; Taub et al., 2010). Over time and with increasing sequencing depth, adequate quality treatment as error filtering or error correction became a fundamental step in metagenomics and was incorporated in standard processing pipelines by means of specialized bioinformatics tools (see (Yang et al., 2013; Zhou et al., 2014) for example). In addition, new tools have been developed to process and analyze specifically metagenomic data sets, ranging from pre-processing, different kinds of taxonomic classification, the reconstruction of genomes by metagenome assembly, the clustering of assembled genomic fragments, the clustering of single reads into taxonomic units or to functional groups, and many more (Teeling and Glöckner, 2012; Kim et al., 2013). Even though this resulted in more complex and tool-rich pipelines - further demonstrating the bioinformatic bottleneck in metagenomics - this allowed for more elaborated studies. It became possible not only to taxonomically describe a microbial community or to link metagenomic reads to genes, orthologous groups, or metabolic pathways but also to investigate extreme environments and complex community hosting habitats (Xie et al., 2011) and, most interestingly, to perform comparative metagenomics that made it possible to e.g., find conformities with respect to ecological roles of the distinct deep-sea sediment and soil microbial communities (Quaiser et al., 2011), identify specific shifts in the bacterial community structure of the saliva microbiome that constitutes a cariogenic factor (Yang et al., 2012), and that led to a more complete understanding of the prevalent and rare organisms involved in anaerobic biomass degradation and biomethanation in biogas fermenters thus paving the way for future process optimization (Jaenicke et al., 2011).

1.2 Introduction to research on biogas producing microbial communities

The use of alternative and renewable energy sources like wind, water and sun is important regarding global warming, increasing energy demand and limited fossil fuels. Another important and promising renewable energy source is biomass (Weiland, 2010). In Germany, the share of consumed electricity from renewable resources in 2015 was 31.5%, of which biogas had a percentage of 16.8% (BMWi, 2016). With 8,856 biogas plants (BGPs) with an installed electric output of 4,018 MW, able to supply 8.4 million households with electricity (German Biogas Association, 2015), Germany is one of the major biogas producing countries in the European Union (EurObserv'ER, 2015).

Biogas is the product of the naturally occurring anaerobic digestion of biomass in many habitats, like soils and sediments. Biogas consists mostly of methane and carbon dioxide as well as small amounts of nitrogen, hydrogen sulfide, ammonia and carbon monoxide (Weiland, 2003). In production-scale BGPs, biogas is produced by a complex anaerobic microbial community from different substrates like organic waste, silage and manure. To generate electricity and heat, the biogas is captured and combusted in a combined heat and power unit.

BGPs can be distinguished by their process technology and process parameters. These include the fed substrates, the process temperature and fermenter type. The fermentation process itself can be distinguished by the organic dry matter content (oDM) of the fed substrates into wet and dry fermentation. Dry fermentation is characterized by an oDM content of 15 to 35%; often mono-fermentation of renewable biomass e. g. maize silage is performed. In wet fermentation with an oDM content below 15%, the fed substrates are often mixtures from different sources. Due to stirring and pumping of the fermenter content, possible process inhibitors are diluted and the digestion is evenly distributed throughout the fermenter (FNR, 2013; Stolze et al., 2015; Weiland, 2003).

The production of biogas can take place under different process temperatures, which influences the biogas production process, like the microbial community composition and the process stability (Stolze et al., 2016). Thermophilic BGPs are operated at 45 to 60°C and are characterized by a lower community diversity, but feature a faster and more efficient substrate turnover rate and a higher biogas and methane output (Levén et al., 2007; Maus at al., 2016a; Maus et al., 2016b; Xia et al., 2014; Zhang et al., 2015). On the other hand, they have a higher energy demand, a less stable process performance and the microbial community is more prone to fluctuations. Mesophilic BGPs are operated at 35 to 42°C; they have a lower risk of ammonia inhibition and the process as well as the microbial community are more stable against, for example, temperature fluctuations (Levén et al., 2007; Weiland, 2010). Mesophilic BGPs are often fed with renewable energy crops in contrast to thermophilic reactors which are recommended for industrial food residues and organic household waste degradation, due to the sanitizing effect of higher temperatures (Weiss et al., 2008). In Germany, the majority of BGPs is operated under wet and mesophilic fermentation conditions (Stolze et al., 2015; Weiland, 2010).

The process of biogas formation is divided into four continuously and simultaneously performed phases: hydrolysis, acidogenesis, acetogenesis and methanogenesis. These steps are performed by a complex anaerobic microbial community. The first three phases are performed by bacteria and the last step is performed by archaea. During the hydrolysis, complex biopolymers like cellulose, starch, proteins and lipids are degraded to oligo- and monomers by hydrolytic bacteria of the phyla Bacteroidetes and Firmicutes. Notably, many species of the Firmicutes belong to the genus *Clostridium*. During acidogenesis, oligo- and monomers are further converted to volatile fatty acids (VFAs) like acetate, propionate, and butyrate by acidogenic bacteria of the phyla Firmicutes, Chloroflexi, Bacteroidetes and Proteobacteria. During the next step, the acetogenesis, the VFAs are converted into acetate, CO_2 and H_2 by acetogenic Bacteria - species of the phyla Firmicutes and Proteobacteria, and the genera *Syntrophomonas*, *Syntrophobacter*, *Pelotomaculum* and *Thermoanaerobacter*. The last step of anaerobic digestion is the methanogenesis, which is solely performed by Archaea that usually belong to the orders

Methanomicrobiales, Methanosarcinales and Methanobacteriales of the phylum Euryarchaeota. In this stage, the products of the previous stages are used for methane formation. Methanogenesis can be separated in different pathways using different substrates for methane production: the hydrogenotrophic methanogenesis (CO_2 and H_2), the acetoclastic pathway (acetate) and the methylotrophic path (methylated compounds) (Christy et al., 2014; Venkiteshwaran et al., 2015; Weiland, 2010).

In the past, BGPs and the involved microbial communities were studied with the aim to identify key organisms. To characterize the structure and the composition of the microbial communities, biological approaches like cultivation based methods, 16S rRNA gene amplicon sequencing, denaturing gradient gel electrophoresis (DGGE), and metagenome sequencing and analysis have been used (Kröber et al., 2009; Ortseifen et al. 2016; Sundberg et al. 2013; Weiss et al., 2008;).

For the bacterial part of the community, species belonging to the class Clostridia and Bacteroidetes were identified to be most dominant in mesophilic BGPs, followed by the class Bacilli, members of the phylum Proteobacteria, and the class Spirochaetes. In addition, the classes Thermotogae and Synergistetes were highly present in thermophilic BGPs (Lebuhn et al., 2014; Maus et al., 2016b; Stolze et al., 2015; Weiss et al., 2008; Zakrzewski et al., 2013). For the archaeal part of the community, members of the orders Methanomicrobiales, Methanosarcinales and Methanobacteriales were found to be most dominant in mesophilic BGPs and Methanothermobacter and Methanosarcina were shown to be most prevalent in thermophilic BGPs. All methanogenic archaeal species identified in BGPs so far belong to the phylum Euryarchaeota (Bremges et al., 2015; Kröber et al., 2009; Maus et al., 2016a; Weiss et al., 2008).

However, the majority of members participating in the biogas production process is still unknown (Campanaro et al., 2016; Maus et al., 2016b; Stolze et al., 2016; Treu et al., 2016) and a large fraction of microorganisms from environmental habitats is not easily cultivable (van der Lelie et al., 2012). To get deeper insights into complex microbial communities in BGPs, WGS metagenomics using assembly and binning approaches to reconstruct complete genomes of the community, as well as single cell genomics are promising developments.

2. Community analysis based on 16S rRNA gene amplicons

When studying an environmental community, one of the first and fundamental question concerns the taxonomic composition of the community and the phylogenetic relationship of their members. One approach for taxonomic classification is to study specific marker genes which are present in a wide range of organisms. Most commonly, this approach involves the characterization of the 16S small subunit ribosomal RNA (rRNA) gene, which enabled researchers for decades to analyze different and complex habitats following the traditional approach of cloning the 16S rRNA gene or fragments thereof into a vector and subsequent Sanger sequencing of the insert DNA (Giovannoni et al., 1990; Pace et al., 1997; Tringe and Rubin, 2005). The 16S rRNA gene is particularly useful as it is ubiquitous in prokaryotes since the 16S rRNA is an integral part of the ribosome, evolves at a constant rate, thereby also serving as a phylogenetic clock, and consists of highly conserved as well as hypervariable regions (HVRs) V1 to V9 facilitating both the amplification by universal primers and a high specificity to distinguish between organisms (Janda and Abbott, 2007). However, as this was a laboratory intense and time consuming procedure, early surveys usually analyzed only 10 to 100 clones per sample which restrained the achievable resolution of the full community spectrum. Enabling direct sequencing of DNA libraries, NGS instruments led to a renaissance of amplicon studies (Tringe and Hugenholtz, 2008) constituting at present a cost and effort efficient routine for microbiological research (Huse et al., 2008; Sogin et al., 2006) to investigate bacterial communities, even up to several hundreds of samples in parallel in one sequencing run. It has since been successfully applied to all different types of habitats (Eckburg et al., 2005; Gill et al., 2006; Quaiser et al., 2011; Weiss et al., 2008) and revealed relationships of microbiomes to different human diseases like periodontitis (Jünemann et al., 2012), Crohn's disease (Erickson et al., 2012), and obesity and inflammatory bowel disease (Greenblum et al., 2012). Due to the shorter read lengths of NGS, only a single HVR or a combination of neighbouring HVRs of the rRNA gene is amplified by PCR and consecutively sequenced (Liu et al., 2008). Albeit it could be shown that short sequencing reads and sub-stretches of the 16S rRNA gene allow for an accurate taxonomic classification ((Liu et al. 2007), it is obvious that the restriction on HVRs naturally limits the taxonomic resolution compared to full length genes (Schloss 2010). Researchers must also be aware of the fact that different targeted HVRs of the same 16S rRNA gene eventually result in different classification results (Youssef et al. 2009; Mizrahi-Man et al. 2013).

The pre-processing of raw sequencing reads prior to the subsequent analysis, mostly referred to as quality checking (QC), is common for sequence based surveys but especially essential for amplicon datasets. In this process, low quality and erroneous reads as well as amplification artifacts are removed which can significantly improve the analysis accuracy (Bakker et al., 2012; D'Amore et al., 2016; Huse et al., 2007; Lee et al., 2012; Quince et al., 2011; Schirmer et al., 2015). In addition, QC is essential to prevent an overestimation of the community species diversity which, if omitted, is likely to be introduced during sequence clustering by forming artificial taxonomic units based on erroneous or artificial reads (Huse et al., 2010; Kunin et al., 2010; Schloss et al., 2011). For each step in QC, a wide range of tools are available for which some examples will be given in the following. Briefly, QC usually comprises (not necessarily in this order): [i] the removal of any sequencing adapters, amplification primers, and multiplex identifiers (Bolger et al., 2014; Martin, 2011; Sturm et al., 2016), [ii] filtering or trimming of low quality reads (Joshi and Fass, 2011; Li, 2011; Modolo and Lerat, 2015; Schmieder and Edwards, 2011), [iii] screening for wrong amplification targets by seed alignments (Caporaso et al., 2010a, 2010b; Schloss et al., 2009), [iv] de-noising or error correction of reads (Edgar and Flyvbjerg, 2015; Huse et al., 2010; Quince et al., 2011), [v] trimming to equal length as a preparation for sequence clustering (see step ii), [vi] read de-replication (the process of identifying identical reads and compressing them into one representative in conjunction with an abundance count) in order to accelerate downstream processing (Edgar, 2010; Fu et al., 2012; Ghodsi et al., 2011), and [vii] the detection and removal of artificial chimeric sequence as a result of cross hybridization of different DNA fragments during PCR (Edgar, 2016; Haas et al., 2011; Wright et al., 2012). For PE

data, an assembly step is integrated merging the forward and reverse read at their overlapping region (given that the overlap is long and the quality high enough). This can be done depending on the data quality either right at the beginning of the QC or at subsequent steps (Magoč and Salzberg, 2011; Zhang et al., 2014).

After the generation of error-free (and de-replicated) reads, the actual analysis is initiated encompassing in essence three major blocks: clustering into operational taxonomic units (OTUs), taxonomic classification, and statistical evaluation. OTU clustering has the advantage to group sequences without prior knowledge of their reference taxonomy which is important when analyzing yet unknown communities. This can be done on any convenient sequence similarity threshold but most commonly a 97% cutoff is chosen with respect to the 16S rRNA gene sequence similarity between two species (Stackebrandt and Goebel, 1994). Clustering of large-scale sequencing data from a distance matrix containing all pairwise differences, i.e. hierarchical clustering, requires high computational costs. Other techniques, as for instance heuristic greedy online clustering, where sequences either form new clusters or join existing ones while being processed top-down based on abundance profiles (Edgar, 2010; Fu et al., 2012; Ghodsi et al., 2011,) are to be preferred and represent the quasi standard in current amplicon pipelines. However, care must be taken with regard to the data quality, the actual clustering method, and the clustering threshold (Chen et al., 2013). Taxonomic classification, on the one hand, can be conducted (on single read or OTU level) reference-based by similarity searches against 16S rRNA gene reference databases (e.g., RDP (Cole et al., 2014), SILVA (Pruesse et al., 2007), or Greengenes (DeSantis et al., 2006)) or they can be placed in a phylogenetic tree thereby inferring the phylogenetic relationship to known members (Matsen et al., 2010; Stamatakis, 2014). Of course, when relying on direct reference assignments the classification effort depends on the completeness and correctness of the database, which cannot always be guaranteed while at the same time too many hits of ambiguous queries can pose an issue. On the other hand, reference-free approaches build upon a reference data set of which specific features are extracted (mostly k-mer composition and frequencies) and used either in a machine learning process or in probability models (such as naïve Bayesian) to infer the taxonomy of a query sequence (DeSantis et al., 2011; Wang et al., 2007). They stand out when it comes to processing speed and the classification of sequences not present in reference databases. On the downside, however, their accuracy is determined by an optimal training set and the choice of an appropriate confidence threshold (Mizrahi-Man et al., 2013). Finally, a series of ecological and statistical analyses is usually carried out to assess e.g., the sampling effort by rarefaction analysis, to estimate species richness and species diversity on different ecological scales as defined by alpha, beta, and gamma diversity, to compare sample conditions by means of principal component analysis or multidimensional scaling, or to compare two communities phylogenetically using, for example, the UniFrac distance (Lozupone and Knight, 2005). For conducting statistical evaluations in particular, but also for all steps enumerated so far, several software suites are available that bundle these algorithms and tools into one platform. Of these, it has been shown that the two popular software suites mothur (Schloss et al., 2009) and QIIME (Caporaso et al., 2010a) are superior to their competitors by means of, i.a., their flexibility, offered processing and analyzing methods, supported input and output formats, and supported sequencing technologies (D'Argenio et al., 2014; Nilakanta et al., 2014; Plummer et al., 2015).

Despite the proven efficiency of the 16S rRNA gene in amplicon surveys, several disadvantages remain. First of all, this approach can only target the bacterial and archaeal domain and - as this approach is based on PCR - is prone to biases introduced by the amplification primers (Sipos et al., 2007; Suzuki and Giovannoni, 1996), and the choice of the targeted HVR can restrict the specificity as well as the comparability between different studies (Mizrahi-Man et al., 2013; Soergel et al., 2012). Next, the resolution of the 16S rRNA gene is limited, especially for incomplete genes, thereby rendering distinction between closely related species difficult, and also with respect to extremely low abundant organisms (Fox et al., 1992). Horizontal gene transfer of 16S rRNA genes can also lead to wrong phylogenetic or taxonomic conclusions (Schouls et al. 2003). Last but not least, varying ranges

of gene copy numbers (GCN) in different species may introduce a bias on taxon abundances (Acinas et al., 2004). Even though this can be bioinformatically addressed by correcting the inferred abundances with the aid of reference GCNs (Angly et al., 2014), in the presence of amplification biases the premise, i.e. amplification rates correspond to GCNs, cannot be guaranteed.

3. Whole genome shotgun metagenomics

In contrast to amplicon based studies, WGS metagenomics allows amplification free insights into the composition as well as the potential function of the underlying community. With continuously decreasing sequencing costs paired with greatly improved throughput WGS metagenomics became increasingly reasonable and applicable. Of course, QC is also important for WGS metagenomics and in principle carried out analogous to the steps outlined above in the 16S rRNA gene amplicon workflow description, just omitting the steps dealing in particular with amplification related errors and artifacts.

In general, there are two approaches for WGS metagenomics: read-based and assembly- based metagenomics. The former aims to classify single reads with regard to taxonomy and function. It is well suited to answer questions related to the taxonomical composition of a sample or related to the presence or absence of organisms, genes or metabolic pathways. A fragment recruitment of metagenome sequences to known reference genomes or genes can be done to examine coverage and variation of those. In assembly-based metagenomics, reads are first de novo assembled to contigs and hereafter clustered into so-called genome bins during a binning process. Thereby, it is possible to reconstruct genomes of (even yet unknown) highly abundant taxa from a metagenomic sample. For this purpose, the corresponding workflow includes an assembler that is well suited for the reconstruction of long contigs and a genome binner to cluster such sequences from the same organism. For taxonomic profiling, the assembly is followed by a taxonomic classification, also referred to as taxonomic binning. If the metabolic potential is of special interest, a gene prediction, functional annotation, and metabolic reconstruction is done on assembled contigs. It is to be noted that binning, in terms of generally separating sequences into groups, is, of course, not limited to assembled contigs per se and can also be performed on unassembled reads. An overview of possible workflows and their individual steps as described in the following is given in Figure 1.

3.1 Fragment recruitment

Basically, fragment recruitment is the mapping of all metagenomic reads to one or more selected references, which can be complete genomes, genomic bins, or sequences of specific genes of interest. This method is a simple approach to determine the presence or absence of specific organisms and useful to identify highly conserved regions and to gain insights into genomic variations like SNPs, insertions, and deletions or to examine the diversity of genes and gene families. It has been used for the first time in the global ocean survey (Rusch et al., 2009) relying on the alignment tool BLAST. An example for a dedicated fragment recruitment tool is FR-HIT (Niu et al., 2011), which outperforms BLAST in terms of runtime and recruits significantly more reads than classic read mapping applications.

3.2 Taxonomic classification

Taxonomic classification of metagenomic reads is related to 16S rRNA marker gene analyses in the way that it assigns sequences to taxonomic groups, thereby deducing a community profile. Accordingly, such classification can either be done reference-based or reference-free.

There exist many and well established tools for reference-based classification, most of them relying on local alignments and of which some will be highlighted in the following. MG-RAST (Meyer et al., 2008) first predicts genes in the dataset using FragGeneScan (Rho et al., 2010), clusters translated amino acids with UCLUST (Edgar, 2010) and maps representative sequences against a custom M5nr database (Wilke et al., 2012) using a variant of BLAT (Kent et al., 2002), the BLAST-like alignment tool, for taxonomic (and functional) annotation. MEGAN (Huson et al., 2016) uses a lowest common ancestor (LCA) approach and assigns the LCA taxonomy of all BLAST hits having a score similar to the best hit to each sequence. CARMA3 (Gerlach and Stoye, 2011), an extension of

CARMA (Krause et al., 2008a), implements a reciprocal BLAST search to improve classification accuracy, but also provides a HMMER3-based variant to search against the Pfam database (Finn et al., 2014). Taxator-tk (Dröge et al., 2015) uses a combined approach of sequence segmented similarities to a reference set and a hereupon approximated phylogenetic tree to taxonomically classify sequences with high precision. All these similarity based methods usually offer a high classification resolution and accuracy but also lack processing speed.

A possibility to greatly speed up the classification is to replace the direct alignment of a query against a reference database with a fast lookup method of fixed-length k-mers extracted from the query in a hash-based index structure built from the references. Classification is then done by inferring a combined taxonomy based on the individual k-mer matches of a query to the prebuilt index. Examples implementing this approach are Kraken (Wood and Salzberg, 2014) or Clark (Ounit et al., 2015). Kraken assigns the LCA to ambiguous k-mers already in the index database and classifies reads based on the highest-weighted root-to-leaf path of a tree containing all matched k-mer taxa, thereby achieving fast processing time. Clark only uses target-specific k-mers of a reference set that are uniquely characterizing each target on a predefined taxonomic level. Reads are then assigned to a target, i.e. classified, for which it shares the highest number of k-mers in a taxonomic tree at this given level. This technique increases the sensitivity but, on the other hand, the classification is bound to a specific taxonomic level. One approach to reduce the high memory requirements of k-mer indexing structures was implemented in the classifier Centrifuge, exploiting a highly compressed Burrows-Wheeler transformed Ferragina-Manzini index (Kim et al., 2016). For the classification, short exact matches between a read and the index are identified and maximally extended. Then, scores for each species hit are assigned while giving greater weight to longer segments and reporting the highest scoring taxas. Additionally to the classification of individual reads, Centrifuge is able to estimate the abundance of each detected taxa at any taxonomic rank through an Expectation-Maximization algorithm.

Reference-free classifiers also rely upon sequence composition, where composition features, mostly k-mer frequencies, are used, for instance, in supervised machine learning approaches for taxonomic classification. In PhylopythiaS+ (Gregor et al., 2016), a Support Vector Machine is trained on a set of reference sequences, manually or automatically incorporated by marker gene detection from the sample. Then, k-mer frequencies of lengths between 4 to 6 within a read are used to predict the taxon on this model. PhyloPythiaS+ achieves a high precision and recall in datasets with deep branches at larger taxonomic distances (Sczyrba et al., 2017). Other techniques for machine learning approaches are, for example, a naive Bayes classifier as implemented in NBC (Rosen et al., 2011) or Interpolated Markov Models used by Phymm and PhymmBL (Brady and Salzberg, 2009). While these classifiers do not require near full length or complete reference sequences they often require relatively long query sequences in order to achieve sufficient entropy on the composition feature (Dröge and McHardy, 2012). For a more complete listing of taxonomic classifiers and an evaluation of their performance please see Peabody et al. (2015).

Accuracy and sensitivity of the classification, especially for the aforementioned methods, can be improved with longer reads or contigs. Thus, an assembly of the metagenome prior to the classification might be a good strategy, although it increases overall computation time and especially memory requirements.

3.3 Metagenomic Assembly

Downstream bioinformatics analysis like gene prediction can be eased, if short reads are assembled into longer, contiguous sequences (contigs). There are many good performing assemblers for reads originating from a single prokaryotic genome available (Earl et al., 2011; Jünemann et al., 2014; Magoc et al., 2013). The difficulty of metagenome assemblies lies in biological complexity, i.e. a mixture of ambiguous and unambiguous genomic elements from different genomes at varying coverages. This makes it particularly hard to differentiate

homologous regions of different strains or to resolve paralogous and repetitive regions, increasing the possibility of intragenomic mis-assemblies or even intergenomic chimeric assemblies (Luo et al., 2012; Mende et al., 2012). Most current assembly algorithms for single genomes and metagenomes can be classified either as Overlap Layout Consensus (OLC) or de Bruijn Graph (DBG) assemblers (for methodological details please see e.g., Nagarajan and Pop (2013) or Li et al. (2012)).

By now, a range of specialized short read metagenome assemblers exist, all utilizing different techniques to deal with the data complexity. MetaVelvet (Namiki et al., 2012), Meta-IDBA (Peng et al., 2011) and its enhanced version IDBA-UD (Peng et al., 2012) work by partitioning the DBG based on k-mer coverage and separately assembling the subgraphs to reduce the complexity. IDBA-UD additionally iterates over several values of k such that missing k-mers in low-depth regions can be supplemented using contigs from previous iterations and aligns PE reads to contigs followed by local assembly to improve the assembly. Yet, they are very memory expensive. Ray Meta (Boisvert et al., 2012) makes extensive use of distributed resources and works by heuristically traversing the DBG and requires less amounts of RAM compared to MetaVelvet and Meta-IDBA. In addition, Ray Meta also provides the possibility to perform a taxonomic profiling (Ray Communities) based on k-mer co-occurrences in the DBG and in a reference set, like the GreenGenes taxonomy (McDonald et al., 2012). One of the more recent metagenome assemblers is MEGAHIT (Li et al., 2016). It uses special succinct DBGs and exhibits a memory footprint that does not considerably exceed the size of the input data. As of major version 1.0, newly introduced CPU based algorithms eliminated GPU dependencies, making it accessible to a wider range of users. A final example is metaSPAdes (Nurk et al., 2016), which is built upon the commonly used SPAdes genome assembler (Bankevich et al., 2012) utilizing paired and multisized DBGs combining graphs of different values of *k*. MetaSPAdes focuses on constructing longer consensus sequences, thereby giving less weight to specific features of rare strains. Of course, the choice of the k-mer size is critical for the result of the DBG metagenome assembler, as could be shown in a recent metagenome assembler evaluation (Sczyrba et al., 2017). Also, assemblers using a range of k-mers featured an overall better performance compared to single k-mer assemblers. This is because larger k-mers tend to facilitate the reconstruction of highly abundant genomes, whereas smaller k-mers are better suited for low abundant genomes.

The reconstruction of complete or draft genomes by metagenome assembly, important to fully characterize yet unknown species or to decipher taxa specific characteristics, is even more challenging, although it has shown to be feasible (Hess et al., 2011; Iverson et al., 2012; Luo et al., 2012; Nielsen et al., 2014). Obviously, only the most abundant taxa can be recovered to a large extent, further facilitated when no closely related strains are present in the dataset. For this purpose, PRICE (Ruby et al., 2013) has been developed, which uses an initial set of seed contigs to focus downstream local assemblies of related reads that are consecutively elongated and merged in several steps and iterations.

After assembly, the assembly quality, in particular the number of assembly errors and mis-assemblies, should be assessed, as an erroneous assembly can negatively influence downstream processing, such as a genome binning (Nielsen et al., 2014). A first assessment can be achieved by mapping of all reads back to the assembly. Here, the mapping fraction and the identification of mislocated PE read mates are used to evaluate the assembly effort, which can be done basically with any convenient mapping software. A dedicated software that calculates various metrics to evaluate metagenome assemblies is MetaQUAST (Mikheenko et al., 2016), a modification of the genome assembly evaluation tool QUAST (Gurevich et al., 2013). Besides common metrics (e.g. N50), it provides a reference based evaluation to identify mis-assemblies by mapping contigs to references and reporting statistics for each reference separately. If, as usual, the references are not known a priori, a de novo evaluation can be performed, where appropriate references are identified on the basis of BLASTn hits of contigs against 16S rRNA gene references of the SILVA database and thereafter downloaded from NCBI.

Finally, it is feasible to perform a binning prior to an assembly in order to reduce both complexity as well as compute and memory usage by assembling the bins independently, thus making exceptional large metagenomic datasets computationally manageable in the first place. This might also help to reduce the formation of chimeric assemblies. Another possible approach is the pre-assembly binning that aims to facilitate the reconstruction of individual genomes (Cleary et al., 2015).

3.4 Genome binning

Metagenome assemblies result in sequence fragments of varying length with their origin in distinct genomes. Genome binning aims to group these fragments by utilizing different techniques to cluster related fragments into so-called bins. This reference free genome binning uses inherent characteristics of the sequence like GC content or k-mer frequencies, which vary across organisms, and often makes use of machine learning methods. MaxBin (Wu et al., 2016) uses tetranucleotide frequencies and contig abundances to genomically bin sequences, achieving high precision and recall (Sczyrba et al., 2017). MetaBAT (Kang et al., 2015), MetaWatt (Strous et al., 2012), and CONCOCT (Alneberg et al., 2014) also use tetranucleotide frequencies, MetaBAT and CONCOCT coverage and PE linkage in addition. All three are a good choice, if a larger fraction of the dataset should be included in the binning (Sczyrba et al., 2017). Although no taxonomy assignment is done initially, bins identified in this step can be analyzed further, for example in a taxonomic classification or functional characterization.

3.5 Gene prediction and functional annotation

To gain insights beyond taxonomic composition, the metagenomic data needs to be functionally annotated. Fragment recruitment, as outlined above, utilizing a database of functionally annotated genes or proteins constitutes a straightforward approach for doing so, after which annotations exhibiting a certain coverage could be mapped, for instance, to metabolic pathways, for instance using KEGG (Kanehisa et al., 2017).

Another approach is the de novo annotation of metagenomic reads or assembled metagenomic contigs, initiated by a gene prediction step. Albeit this is a common task in genomics and genome annotation pipelines, single genome gene prediction tools are not well suited for metagenomic datasets, because of the diversity of sequence composition, sequence errors, and sequence length (Hoff, 2009a). Several de novo coding sequence (CDS) prediction programs are available that have been developed to cope with these problems, as e.g., Prodigal in metagenomic mode (Hyatt et al., 2012), MetaGene (Noguchi et al., 2006), MetaGeneMark (Zhu et al., 2010), Orphelia (Hoff et al., 2009b), FragGeneScan-Plus (Kim et al., 2015) or MetaGun (Liu et al., 2013). Downstream functional annotation of predicted CDSs is, in essence, following the standard procedures in classic genomics. This can be done, on the one hand, by searching for homology (using BLAST for example) against specific databases like COG (Tatusov et al., 2000), EggNOG (Huerta-Cepas et al., 2016) or KEGG orthology groups. On the other hand, Hidden Markov Models can be used to assign protein sequences to protein families using e.g., HMMer (Eddy et al., 2011) and Pfam. Moreover, metagenomic analysis platforms like MG-RAST and IMG/M (Markowitz et al., 2014) provide dedicated annotation components whereas established prokaryotic annotation systems, such as GenDB (Meyer et al., 2003), can and have been used to annotate metagenomic contigs.

4. Review of exemplary studies on biogas microbiomes

In the following, several selected application examples are briefly summarized focusing on the bioinformatic analysis subdivided in 16S rRNA gene or WGS based surveys, respectively. Thereby it is shown how the chosen sequencing platform in conjunction with the applied analysis methods could give profound insights into questions regarding the composition and function of biogas residing microbial communities.

4.1 Application examples on 16S rRNA gene based community analysis

Next to surveys solely relying on 16S rRNA gene amplicons, a still widely used application scenario for them is to serve as a complement to verify and refine results obtained by other approaches such as WGS metagenomics or metatranscriptomics. Such an example is the study by Zakrzewski et al. (2012), which made use of 454-pyrosequencing to profile the transcriptionally active members and their most actively transcribed functions from a production-scale agricultural BGP operating at mesophilic temperatures (Jaenicke et al., 2011; Krause et al., 2008a; Schlüter et al., 2008). 16S rRNA gene amplicons were used to evaluate the taxonomy profiles deduced from the metatranscriptome 16S rRNA sequences. For this, the HVRs V3 and V4 were amplified and sequenced on the GS FLX platform using the FLX Titanium chemistry. Downstream processing was a simplified version of the pipeline outlined in chapter [2]. Reads were first length filtered and screened for primers and multiplex identifiers using the RDP Amplicon Sequence Pipeline (Wang et al., 2007). Consecutively, QIIME was used to verify reads for correctly spanning the V3-V4 region, as well as for OTU clustering and taxonomic profiling. These profiles then revealed a discrepancy of Archaea between the 16S rRNA gene amplicon and the 16S rRNA metatranscriptome reads, where Archaea were overrepresented in the latter one, indicating either an increased transcriptional activity of Archaea or a bias of the amplification primer (Zakrzewski et al., 2012). On the other side, high accordance for dominant taxa between both approaches could be identified. Not only is this a strong mutual confirmation but also indicates that, in this case, the prevalent taxa are also the transcriptionally active ones.

Likewise, the study by Stolze et al. (2015) made also use of 16S rRNA gene amplicons to confirm differences between two BGPs, one operating at wet and one at dry fermentation condition, as revealed by comparative metagenomics. Obtained metagenomic profiles showed a high accordance on higher taxonomic ranks. However, identified differences in archaeal sub-communities were subsequently analyzed in detail at a higher taxonomic resolution following the MiSeq 16S rRNA gene amplicon approach based on the HVR region V4 and 2x300 bp PE reads. The analysis was mainly done with QIIME and USEARCH (Edgar, 2010) and the quality control was much more sophisticated in comparison to the previous study, reflecting the acquired insights on this topic (essentially covering step [i], [ii], [iv], [v], and [vi] as described in chapter [2]). After OTU clustering representative sequences were taxonomically described using the RDP Classifier. Further phylogenetic analysis on selected Archaea was done by aligning the representative sequences using Infernal (Nawrocki and Eddy, 2013) with the small subunit rRNA model from Rfam (Nawrocki et al., 2015). This formed the basis for a phylogenetic tree reconstruction with FastTree (Price et al., 2009). The classification profiles and the archaeal phylogenetic trees of the 16S rRNA gene amplicons could further elucidate the composition of archaeal sub-communities, which are specific to either dry or wet fermentation conditions in both samples, i.e. different prevalence of members of the genus *Methanoculleus*.

To further study different operating conditions of BGPs, one thermophilic BGP operating at higher (54°C) and three mesophilic BGPs operating on lower (40°C) temperature, their microbial communities were characterized by Stolze et al. (2016) using 16S rRNA gene amplicons again in conjunction with WGS metagenomics. After data

were generated on the MiSeq system and extensively pre-processed, QIIME and USEARCH were used for OTU clustering and classification, here reference based using the Greengenes database. OTU representatives were phylogenetically described using ARB and the SINA aligner (Quast et al., 2013; Yilmaz et al., 2013). Further, community profiles were corrected for 16S gene copy numbers using Copyrighter (Angly et al., 2014). Obtained results demonstrated a significant difference between the four BGPs' taxonomic profiles on the phylum level, especially the phylum Thermotogae being present only in the thermophilic BGP. In addition, differences between the mesophilic BGPs were found as well for the phyla Fusobacteria, Spirochaetes, and Cloacimonetes. However, distinctive OTUs for these phyla present only in one or two of the four BGPs could not unambiguously be described taxonomically due to the lack of suitable reference genomes, therefore missing insights into their functional role. This, ultimately, demonstrates the limitations of a 16S rRNA based analysis and highlights the need to analyze the complete metagenome in order to deduce the full functional potential of a microbial community.

4.2 Application examples on WGS metagenomics

One of the earliest insights into the metagenome of a biogas producing community was achieved by Schlüter et al. (2008), when metagenomic assemblers were not publicly available yet. Here, a fermentation sample from the first fermenter of a BGP operating at mesophilic temperatures was taken to analyze the structure, gene content, and metabolic capabilities of the residing microbial community. Sequences were generated on the GS FLX System and assembled using the GS De Novo Assembler (Margulies et al., 2005). Reads were assigned to COG categories using BLASTx. Assembled short contigs were taxonomically and functionally annotated by means of gene names, gene products, EC numbers, and COG categories using the SAMS platform (Bekel et al., 2009). Longer contigs were functionally annotated with the GenDB pipeline. To identify contigs encoding cellulosome proteins, tBLASTn against a custom database was performed. Finally, metagenomic reads were mapped to the *Methanoculleus marisnigri* JR1 reference genome sequence using BLASTn. The results of this study revealed that within the COG domain related to metabolisms, the most prevalent categories were energy production and conversion, and carbohydrate as well as amino acid transport and metabolism, respectively, corresponding to a genetic profile characteristic for an anaerobic microbial community. Members of the class Clostridia, Bacteriodetes, Bacilli, and Methanomicrobia dominated the community and the genus *Methanoculleus* was found to be the most abundant and best covered methanogenic Archaea. The different detected clostridial species were shown to be important for hydrolysis of cellulose. This study also demonstrated that 142 Mbp of sequencing data were not sufficient to cover the whole complexity of such a rich community and that metagenome assembly was challenging (only 16.04% of total bases could be assembled).

In a more recent study, Bremges et al. (2015) addressed the limitations of such early studies to decipher also the rare microbial spectra of BGPs by increasing the sequencing depth by more than ten times. The metagenome and metatranscriptome of an agricultural production-scale BGP, operated under mesophilic wet fermentation conditions, were sequenced on Illumina's GAIIx and MiSeq platforms, resulting in more than 23 Gbp and 12 Gbp, respectively. After read trimming with Trimmomatic (Bolger et al., 2014), a metagenome assembly was done using Ray Meta with k-mer sizes from 21 to 61, of which the assembly with the best size, contiguity and percentage of reads mapped back to the assembly was chosen. Genes were predicted on the contigs using Prodigal in metagenomic mode and functionally annotated with the KEGG Orthology using BLAST. To associate the metatranscriptome with the metagenome, the number of reads originating from both forming a subset of the predicted genes were counted using BEDtools (Quinlan and Hall, 2010) and the genes associated to methanogenesis pathways were identified. The majority of genes involved in these pathways could be assembled and for many of those an active gene expression was indicated, especially for those related to hydrogenotrophic

methanogenesis. This high coverage dataset has also been used in a metaproteome study highlighting the usefulness of metagenomics to related research areas (Kohrs et al., 2015).

While these two studies focused on BGPs under mesophilic conditions, a study by Maus et. al. (2016b) aimed to characterize the microbial community and ecology of a thermophilic BGP in a polyphasic survey, i.e. the combination of traditional cultivation and molecular characterization techniques, complemented by metagenome and metatranscriptome analyses. In contrast to the assembly based approach in the study by Bremges et al. (2015), a read-based approach has been applied here. Total microbial DNA and RNA were extracted from samples of two digesters and sequenced on the Illumina MiSeq instrument . Reads obtained were PE merged using Flash and consecutively analyzed using the MGX platform. Taxonomic classification was done by searching for 16S rRNA gene sequences in the metagenome or transcripts thereof in the metatranscriptome and compared to the RDP database following MGX's '16S pipeline'. For analyzing the gene content of the community, metagenomic reads were functionally annotated using a BLASTx search against the COG database, whereas carbohydrate-active enzymes were identified in particular using MGX's dbCAN pipeline (Yin et al., 2012). Additionally, fragment recruitment using FR-HIT was performed on the complete metagenomic dataset in order to determine species affiliated to the strains *Clostridium cellulosi* str. DG5, *Herbinix hemicellulosilytica* str. T3/55^{T}, and *Defluviitoga tunisiensis* str. L3. Altogether, the thermophilic biogas producing microbial community was found to be different from mesophilic ones with the most metabolically active fermentative Bacteria identified as members of the genera *Defluviitoga*, *Clostridium* cluster III, and *Tepidanaerobacter*, and for the less diverse methanogenic Archaea as members of the genus *Methanoculleus*. The results also confirm that strains highly related to the reference strain *D. tunisiensis* L3 (Maus et al., 2016c) play a key role within the community of the thermophilic biogas-production plant. Finally, profile differences of the metagenome and metatranscriptome, most prominently in the abundant and metabolically active genera *Defluviitoga*, *Clostridium* and *Methanoculleus*, also showed that a high metabolic or transcriptional activity does not necessarily determine an equal high metagenomic abundance, or vice versa.

To further elucidate the observed differences between the three mesophilic and the one thermophilic BGPs, as outlined in the previous sub-chapter, a metagenomic assembly was pursued by Stolze et al. (2016). For this, PE data from Illumina's HiSeq were first quality checked with the JGI's QC pipeline, yielding 328 Gbp of data. After assembly with Ray Meta, coding sequences on contigs larger than 1 kb were predicted using Prodigal which were then compared to the NCBI database with DIAMOND's BLASTp mode (Buchfink et al., 2015) and loaded into MEGAN5 for taxonomic classification. Binning was done with MetaBAT and the resulting bins were checked with CheckM (Parks et al., 2015) for completeness, strain heterogeneity and contaminations. Genome bins were also taxonomically classified with Taxator-tk. Five genome bins belonging to the four phyla Thermotogae, Fusobacteria, Spirochaetes, and Cloacimonetes (also showed distinct differences in the 16S rRNA amplicon data) were further analyzed in GenDB 2.0 for KEGG pathway mapping and in EDGAR 2.0 (Blom et al., 2016) for comparative gene analyses. Insights into the genetic potential and putative roles of yet uncharacterized biogas residing organisms could be obtained for these selected genome bins. The genetic comparison of the putative Fusobacteria and Cloacimonetes species (represented by their corresponding bins) to their closest references revealed that they are putatively involved in amino acid fermentation, whereas the one represented by the Spirochaetes bin most likely ferments sugars.

5. Bioinformatic solutions offered by BiGi

The de.NBI Bielefeld-Gießen Resource Center for Microbial Bioinformatics has a long-time experience in 16S rRNA gene amplicon and metagenomic analysis and several dedicated tools have been developed and successfully applied to research projects in this course. Many of those tools were initiated as a direct consequence of the lack of appropriate bioinformatics solutions able to answer the questions raised by the researchers. Over time and with a rapidly growing bioinformatic fundament, the request on computational metagenomics shifted from single task problem solvers to integrated and graphic rich platform solutions. In this light and besides general consulting and workflows on demand, the de.NBI center BiGi offers two platforms for the analysis of metagenomic datasets, MGX and EMGB.

MGX is an advanced open-source application for the analysis of unassembled shotgun metagenome data and provides a comprehensive set of high-throughput analysis pipelines aimed at the taxonomic classification of metagenome datasets including recent tools like e.g., Kraken (Wood et al., 2014) or Centrifuge (Kim et al., 2016). Functional analysis capabilities of MGX include, among others, the characterization of environmental communities by assignment of EC numbers and the identification of COGs or Pfam domains in the unassembled metagenome sequences. Based on these results, users are able to perform a variety of statistical analysis tasks, like e.g., rarefaction analysis, PCA/PCoA, clustering, or the computation of biodiversity indices. In addition, MGX supports the creation of fragment recruitments. The intuitive and easy-to-use graphical user interface was implemented in Java (Figure 2), is available for all major operating systems (Windows, Linux, Mac OS X), and allows to create interactive as well as high-quality charts based on taxonomic and functional profiling results. Currently, two MGX servers are hosted and maintained by BiGi at the JLU Gießen and at Bielefeld University, where sizeable high-performance compute clusters are available for the analysis of large metagenome datasets. MGX is accessible at https://mgx-metagenomics.github.io and projects can be applied for by contacting mgx@computational.bio.uni-giessen.de.

The Elastic MetaGenome Browser (EMGB), on the other hand, provides pipelines to analyze metagenomic datasets and metagenome assemblies and offers a responsive web interface to visually inspect the data (Figure 3). EMGB is a web application for browsing especially large annotated metagenomic datasets in real time and features an interactive phylogenetic tree, live GO term statistics as well as dynamically highlighted KEGG pathways for each dataset. In addition, a dedicated comparative viewer displays aggregated versions of these visualizations to draw comparisons across many different datasets at once. Versatile filtering options assist in finding the genes of interest. EMGB is implemented as a responsive single-page application using HTML5 standards as well as the AngularJS framework and supports desktop browsers as well as mobile devices. Data is fetched via the ElasticSearch REST interface. The ElasticSearch server backend is a real-time search engine that enables fast access to large datasets. Therefore, data is automatically pre-processed after it is imported into EMGB as a single JSON file. Projects and access for EMGB can be applied for by contacting emgb@cebitec.uni-bielefeld.de.

6. Concluding remarks

To conclude this review, we hope that the workflows and application examples outlined herein will serve researchers firstly approaching metagenomics or the research on energy production by means of biomass degradation in biogas plants as a solid baseline to design and execute their projects, especially for the metagenomic and anaerobic digestion related community in Germany. We also hope, that this review, the presented bioinformatics expertise, and the offered solutions by BiGi will contribute to the overall de.NBI effort to promote bioinformatics in life sciences. Of course, any review is a snapshot in time and we anticipate that scientific questions as well as their methodological approaches will change in the near future. Promising new third-generation sequencing technologies, as for instance the Oxford Nanopore's single-molecule sequencing platform MinION, are already approaching the research community and have the potential to revolutionize the life sciences yet again (Laver et al., 2015). In particular with respect to the assemblage of genomes and metagenomes, the theoretically very high achievable read length has the potential to render short read assemblers obsolete one day (Goodwin et al., 2015). However, metagenomics in general and metagenome assemblies in particular remain computationally challenging and demonstrate the need for a bioinformatic infrastructure that provides appropriate compute resources. This is even more relevant for researchers without access to dedicated compute hardware adequate for such analyses at their home institution. A possible solution to centralize such resources while democratizing the utilization of them are cloud infrastructures, for which prototypes have already been established successfully for the area of bioinformatics, as for instance the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) in the United Kongdom (Connor et al., 2016). In a similar effort, the de.NBI cloud, which is currently being established in Germany and of which BiGi is one of the host providers, pursues a similar purpose aiming in particular to provide infrastructure and services for bioinformatics in life sciences. In this manner, a central upcoming objective with respect to the tools and pipelines presented here, as well as platform based applications and their access to high performance compute resources, will be their integration into the de.NBI cloud, thus persisting the methodological approaches presented in this review.

Funding

Funding: This work was supported in parts by grants of the German Federal Ministry of Education and Research (BMBF) of the project ‘Bielefeld-Gießen Center for Microbial Bioinformatics - BiGi’ (Grant number 031A533) within the German Network for Bioinformatics Infrastructure (de.NBI). ASCH, JH and JN acknowledge the German Federal Ministery of Food and Agriculture (BMEL) for financial support via the Fachagentur für Nachwachsende Rohstoffe e.V. (FNR) of the joint research projects Biogas-Core (FKZ 22017111) and BMP-III (FKZ 22404015).

Contribution

Conceived and designed the review: SPA ASCZ AG JS. Drafted the manuscript: SJÜ NK. Provided biological background: ASCH JH JN YS. Contributed bioinformatic platforms: SJA CH. Wrote the paper: SJÜ NK SJA CH JH JN YS. All authors read and approved the final manuscript.

References

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. and Polz, M. F., 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J. Bacteriol. 186, 2629-2635.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F. and Quince, C., 2014. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144-1146.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A. and Rohwer, F., 2006. The marine viromes of four oceanic regions. PLoS Biol. 4, e368.

Angly, F. E., Dennis, P. G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P. and Tyson, G. W., 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2, 11.

Antón, J., Llobet-Brossa, E., Rodríguez-Valera, F. and Amann, R. 1999. Fluorescence *in situ* hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. Environ. Microbiol. 1, 517-523.

German Biogas Association, 2015. Biogas statistics.

Bakker, M. G., Tu, Z. J., Bradeen, J. M. and Kinkel, L. L., 2012. Implications of pyrosequencing error correction for biological data interpretation. PLoS One 7, e44357.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455-477.

Bekel, T., Henckel, K., Küster, H., Meyer, F., Mittard Runte, V., Neuweger, H., Paarmann, D., Rupp, O., Zakrzewski, M., Pühler, A., Stoye, J. and Goesmann, A., 2009. The Sequence Analysis and Management System - SAMS-2.0: data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies. J. Biotechnol. 140, 3-12.

Berry, D., Ben Mahfoudh, K., Wagner, M. and Loy, A., 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol. 77, 7846-7849.

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C. and Goesmann, A., 2016. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res. 44, W22-W28.

BMWi, 2016. Development of renewable energy sources in germany 2015.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. and Corbeil, J., 2012. Ray Meta: scalable *de novo* metagenome assembly and profiling. Genome Biol. 13, R122.

Bolger, A. M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Brady, A. and Salzberg, S. L., 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods 6, 673-676.

Bremges, A., Maus, I., Belmann, P., Eikmeyer, F., Winkler, A., Albersmeier, A., Pühler, A., Schlüter, A. and Sczyrba, A., 2015. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. GigaScience 4, 33.

Buchfink, B., Xie, C. and Huson, D. H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59-60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Campanaro, S., Treu, L, Kougias, P. G., De Francisci, D., Valle, G. and Angelidaki, I. 2016. Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. Biotechnol. Biofuels, 9:26.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R., 2010a. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335-336.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L. and Knight, R., 2010b. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26, 266-267.

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. and Zhao, H., 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS One 8, e70837.

Christy, P. M., Gopinath, L. and Divya, D., 2014. A review on anaerobic decomposition and enhancement of biogas production through enzymes and microorganisms. Renew. and Sustain. Energy Rev. 34, 167-173.

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S. and Alm, E. J., 2015. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat. Biotechnol 33, 1053-1060.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. and Tiedje, J. M., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, D633-D642.

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M. J., Richardson, E., Ismail, M., Elwood-Thompson, S. and others, 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. Microbial Genomics 2.

D'Argenio, V., Casaburi, G., Precone, V. and Salvatore, F., 2014. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. BioMed Res. Internatl. 2014, 325340.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069-5072.

DeSantis, T. Z., Keller, K., Karaoz, U., Alekseyenko, A. V., Singh, N. N., Brodie, E. L., Pei, Z., Andersen, G. L. and Larsen, N., 2011. Simrank: Rapid and sensitive general-purpose k-mer search tool. BMC Ecology 11, 11.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36, e105.

Dröge, J. and McHardy, A. C., 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief. Bioinform. 13, 646-655.

Dröge, J., Gregor, I. and McHardy, A. C., 2015. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics 31, 817-824.

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., Shakya, M., Podar, M., Quince, C. and Hall, N., 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17, 55.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, İ., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S.-P., Wu, W., Chou, W.-C., Srivastava, A., Shaw, T. I., Ruby, J. G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Yin, S., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I. and Paten, B., 2011. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Res. 21, 2224-2241.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. and Relman, D. A., 2005. Diversity of the human intestinal microbial flora. Science 308, 1635-1638.

Eddy, S. R., 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7, e1002195.

Edgar, R. C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

Edgar, R. C., and Flyvbjerg, H., 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics, btv401.

Edgar, R., 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv , 074252.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C. and Rohwer, F., 2006. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7, 57.

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N. C., Fraser, C. M., Hettich, R. L. and Jansson, J. K., 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 7, e49138.

EurObserv'ER, 2015. The state of renewable energies in europe. EurObserv'ER Report 15, 1-88.

Ferris, M. J., Muyzer, G. and Ward, D. M., 1996. Denaturing gradient gel electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat community. Appl. Environ. Microbiol. 62, 340-346.

Fierer, N., Jackson, J. A., Vilgalys, R. and Jackson, R. B., 2005. Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. Appl. Environ. Microbiol. 71, 4117-4120.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. and Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222-D230.

FNR, 2013. Biogas - an introduction.

Fox, G. E., Wisotzkey, J. D. and Jurtshuk, P., 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. 42, 166-170.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152.

Gerlach, W. and Stoye, J., 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. 39, e91.

Ghodsi, M., Liu, B. and Pop, M., 2011. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics 12, 271.

Gilbert, J. A., Jansson, J. K. and Knight, R., 2014. The Earth Microbiome project: successes and aspirations. BMC Biol. 12, 69.

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.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. and Field, K. G., 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature 345, 60.

Gomez-Alvarez, V., Teal, T. K. and Schmidt, T. M., 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME J. 3, 1314-1317.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C. and McCombie, W. R., 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 25, 1750-1756.

Goodwin, S., McPherson, J. D. and McCombie, W. R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333-351.

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. Proc. Natl. Acad. Sci. U.S.A. 109, 594-599.

Gregor, I., Dröge, J., Schirmer, M., Quince, C. and McHardy, A. C., 2016. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ 4, e1603.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072-1075.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Consortium, H. M., Petrosino, J. F., Knight, R. and Birren, B. W., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21, 494-504.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5, R245-R249.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z. and Rubin, E.M. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331, 463-67.

Hodkinson, B. P. and Grice, E. A., 2015. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv. in Wound Care 4, 50-58.

Hoff, K. J. 2009a. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics 10, 520.

Hoff, K. J., Lingner, T., Meinicke, P. and Tech, M., 2009b. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res. 37, W101-W105.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. and Bork, P., 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, D286-D293.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. and Welch, D. M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8, R143.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A. and Sogin, M. L., 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 4, e1000255.

Huse, S. M., Welch, D. M., Morrison, H. G. and Sogin, M. L., 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering.. Environ. Microbiol. 12, 1889-1898.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J. and Tappu, R., 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. 12, e1004957.

Hyatt, D., LoCascio, P. F., Hauser, L. J. and Uberbacher, E. C., 2012. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28, 2223-2230.

iTagger (2017). Available: https://bitbucket.org/berkeleylab/jgi_itagger/. Accessed 2017 March 19.

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L. and Armbrust, E. V., 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine *Euryarchaeota*. Science 335, 587-590.

Jaenicke, S., Ander, C., Bekel, T., Bisdorf, R., Dröge, M., Gartemann, K.-H., Jünemann, S., Kaiser, O., Krause, L., Tille, F., Zakrzewski, M., Pühler, A., Schlüter, A. and Goesmann, A., 2011. Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. PLoS One 6, e14519.

Janda, J. M. and Abbott, S. L., 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J. Clin. Microbiol. 45, 2761-2764.

Joshi, N. and Fass, J., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]. Available at https://github.com/najoshi/sickle. Accessed 2017 March 19.

Jünemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J. and Harmsen, D., 2012. Bacterial Community Shift in Treated Periodontitis Patients Revealed by Ion Torrent 16S rRNA Gene Amplicon Sequencing. PLoS One 7(8), e41606.

Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J. and Harmsen, D., 2014. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. PLoS One 9, e107014.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361.

Kang, D. D., Froula, J., Egan, R. and Wang, Z., 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165.

Kent, W. J., 2002. BLAT - the BLAST-like alignment tool. Genome Res. 12, 656-664.

Kim, D.; Hahn, A. S.; Wu, S.-J.; Hanson, N. W.; Konwar, K. M. and Hallam, S. J., 2015. FragGeneScan-plus for scalable high-throughput short-read open reading frame prediction: Computl. Intell. in Bioinform. and Computl. Biol. (CIBCB) 2015 IEEE Conference on, 1-8.

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

Kim, D., Song, L., Breitwieser, F. P. and Salzberg, S. L., 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721-1729.

Kohrs, F., Wolter, S., Benndorf, D., Heyer, R., Hoffmann, M., Rapp, E., Bremges, A., Sczyrba, A., Schlüter, A. and Reichl, U., 2015. Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities. Proteomics 15, 3585-3589.

Kolb, S., Knief, C., Stubner, S. and Conrad, R., 2003. Quantitative detection of methanotrophs in soil by novel pmoA-targeted real-time PCR assays. Appl. Environ. Microbiol. 69, 2423-2429.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A. and Stoye, J., 2008. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 36, 2230-2239.

Kröber, M., Bekel, T., Diaz, N. N., Goesmann, A., Jaenicke, S., Krause, L., Miller, D., Runte, K. J., Viehöver, P., Pühler, A. and Schlüter, A., 2009. Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. J. Biotechnol. 142, 38-49.

Kulski, J. K., 2016. Next-generation sequencing - an overview of the history, tools, and 'omic'applications, in: Kulski, J. K. (Ed.), Next Generation Sequencing - Advances, Applications and Challenges. Intech.

Kunin, V., Engelbrektson, A., Ochman, H. and Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. 12, 118-123.

Laver, T., Harrison, J., O'neill, P., Moore, K., Farbos, A., Paszkiewicz, K. and Studholme, D. J., 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. 3, 1-8.

Lebuhn, M., Munk, B. and Effenberger, M., 2014. Agricultural biogas production in Germany-from practice to microbiology basics. Energy, Sustain. and Soc. 4, 10.

van der Lelie, D., Taghavi, S., McCorkle, S. M., Li, L.-L., Malfatti, S. A., Monteleone, D., Donohoe, B. S., Ding, S.-Y., Adney, W. S., Himmel, M. E. and Tringe, S. G., 2012. The metagenome of an anaerobic microbial community decomposing poplar wood chips. PLoS One 7, e36740.

Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R. and Cary, S. C., 2012. Groundtruthing next-gen sequencing for microbial ecology--biases and errors in community structure estimates from PCR amplicon pyrosequencing. PLoS One 7, e44224.

Levén, L., Eriksson, A. R. B. and Schnürer, A., 2007. Effect of process temperature on bacterial and archaeal communities in two methanogenic bioreactors treating organic household waste. FEMS Microbiol. Ecol. 59, 683-693.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H. and Lam, T.-W., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102, 3-11.

Li, M., 2011. DUK-A Fast and efficient Kmer based sequence matching tool. Lawrence Berkeley Natl. Lab.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. and Fan, W., 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief. Funct. Genomics 11, 25-37.

Liu, W. T., Marsh, T. L., Cheng, H. and Forney, L. J., 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl. Environ. Microbiol. 63, 4516-4522.

Liu, Y., Guo, J., Hu, G. and Zhu, H., 2013. Gene prediction in metagenomic fragments based on the SVM algorithm. BMC Bioinformatics 14 Suppl 5, S12.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D. and Knight, R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis.. Nucleic Acids Res. 35, e120.

Liu, Z., DeSantis, T. Z., Andersen, G. L. and Knight, R., 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36, e120-e120.

Llobet-Brossa, Rosselló-Mora and Amann, 1998. Microbial Community Composition of Wadden Sea Sediments as Revealed by Fluorescence *In Situ* Hybridization. Appl. Environ. Microbiol. 64, 2691-2696.

Lozupone, C. and Knight, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ Microbiol. 71, 8228-8235.

Luo, C., Tsementzi, D., Kyrpides, N. C. and Konstantinidis, K. T., 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. 6, 898-901.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J. and Salzberg, S. L., 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29, 1718-1725.

Magoč, T. and Salzberg, S. L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963.

Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., Huntemann, M., Billis, K., Varghese, N., Tennessen, K., Mavromatis, K., Pati, A., Ivanova, N. N. and Kyrpides, N. C., 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res. 42, D568-D573.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, pp-10.

Matsen, F. A., Kodner, R. B. and Armbrust, E. V., 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11, 538.

Maus, I., Kim, Y. S., Wibberg, D., Stolze, Y., Off, S., Antonczyk, S., Pühler, A., Scherer, P. and Schlüter, A., 2016a. Biphasic study to characterize agricultural biogas plants by high-throughput 16S rRNA gene amplicon sequencing and microscopic analysis. J. Microbiol. Biotechnol. 27(2), 321-334.

Maus, I., Koeck, D. E., Cibis, K. G., Hahnke, S., Kim, Y. S., Langer, T., Kreubel, J., Erhard, M., Bremges, A., Off, S., Stolze, Y., Jaenicke, S., Goesmann, A., Sczyrba, A., Scherer, P., König, H., Schwarz, W. H., Zverlov, V. V., Liebl, W., Pühler, A., Schlüter, A. and Klocke, M., 2016b. Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. Biotechnol. Biofuels 9, 171.

Maus, I., Cibis, K. G., Bremges, A., Stolze, Y., Wibberg, D., Tomazetto, G., Blom, J., Sczyrba, A., König, H., Pühler, A. and others 2016c. Genomic characterization of Defluviitoga tunisiensis L3, a key hydrolytic bacterium in a thermophilic biogas plant and its abundance as determined by metagenome fragment recruitment. J. Biotechnol. 232, 50-60.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610-618.

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., Raes, J. and Bork, P., 2012. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One 7, e31386.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. and Pühler, A., 2003. GenDB - an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31, 2187-2195.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A. and others, 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386.

Mikheenko, A., Saveliev, V. and Gurevich, A., 2016. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088-1090.

Mizrahi-Man, O., Davenport, E. R. and Gilad, Y., 2013. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. PLoS One 8, e53608.

Modolo, L. and Lerat, E., 2015. UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. BMC Bioinformatics 16, 137.

Muyzer, G., de Waal, E. C. and Uitterlinden, A. G., 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. 59, 695-700.

Nagarajan, N. and Pop, M., 2013. Sequence assembly demystified. Nat. Rev. Genetics 14, 157-167.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N. and Kanaya, S., 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 39, e90.

Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y., 2012. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res. 40, e155.

Nawrocki, E. P. and Eddy, S. R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933-2935.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J. and Finn, R. D., 2015. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 43, D130-D137.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E. and others, 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822-828.

Nilakanta, H., Drews, K. L., Firrell, S., Foulkes, M. A. and Jablonski, K. A., 2014. A review of software for analyzing molecular sequences. BMC Res. Notes 7, 830.

Niu, B., Zhu, Z., Fu, L., Wu, S. and Li, W., 2011. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes.. Bioinformatics 27, 1704-1705.

Noguchi, H., Park, J. and Takagi, T., 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 34, 5623-5630.

Nurk, S., Meleshko, D., Korobeynikov, A. and Pevzner, P., 2016. metaSPAdes: a new versatile *de novo* metagenomics assembler. arXiv preprint arXiv:1604.03071.

Ortseifen, V., Stolze, Y., Maus, I., Sczyrba, A., Bremges, A., Albaum, S. P., Jaenicke, S., Fracowiak, J., Pühler, A. and Schlüter, A. 2016. An integrated metagenome and-proteome analysis of the microbial community residing in a biogas production plant. J. Biotechnol. 231, 268-279.

Osborn, A. M., Moore, E. R. and Timmis, K. N., 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ. Microbiol. 2, 39-50.

Ounit, R., Wanamaker, S., Close, T. J. and Lonardi, S., 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.. BMC Genomics 16, 236.

Pace, N. R., Stahl, D. A., Lane, D. J. and Olsen, G. J., 1985. Analyzing natural microbial populations by rRNA sequences. ASM News 51, 4-12.

Pace, N. R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734-740.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. and Tyson, G. W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043-1055.

Peabody, M. A., Van Rossum, T., Lo, R. and Brinkman, F. S. L., 2015. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities.. BMC Bioinformatics 16, 363.

Peng, Y., Leung, H. C. M., Yiu, S. M. and Chin, F. Y. L., 2011. Meta-IDBA: a *de Novo* assembler for metagenomic data. Bioinformatics 27, i94-101.

Peng, Y., Leung, H. C. M., Yiu, S. M. and Chin, F. Y. L., 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420-1428.

Plummer, E., Twin, J., Bulach, D. M., Garland, S. M. and Tabrizi, S. N., 2015. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. J. of Proteomics & Bioinformatics 8, 283.

Price, M. N., Dehal, P. S. and Arkin, A. P., 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26, 1641-1650.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. and Glöckner, F. O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188-7196.

Quaiser, A., Zivanovic, Y., Moreira, D. and López-García, P., 2011. Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. ISME J. 5, 285-304.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-D596.

Quince, C., Lanzen, A., Davenport, R. J. and Turnbaugh, P. J., 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38.

Quinlan, A. R. and Hall, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Rho, M., Tang, H. and Ye, Y., 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38, e191.

Rosen, G. L., Reichenberger, E. R. and Rosenfeld, A. M., 2011. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics 27, 127-129.

Ruby, J. G., Bellare, P. and Derisi, J. L., 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 3, 865-880.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M. and Venter, J. C., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5, e77.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43, e37.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541.

Schloss, P. D. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. 6, e1000844.
Schloss, P. D., Gevers, D. and Westcott, S. L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS one 6, e27310.

Schlüter, A., Bekel, T., Diaz, N. N., Dondrup, M., Eichenlaub, R., Gartemann, K.-H., Krahn, I., Krause, L., Krömeke, H., Kruse, O., Mussgnug, J. H., Neuweger, H., Niehaus, K., Pühler, A., Runte, K. J., Szczepanowski, R., Tauch, A., Tilker, A., Viehöver, P.

and Goesmann, A., 2008. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J. Biotechnol. 136, 77-90.

Schmieder, R. and Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-864.

Schouls, L. M., Schot, C. S. and Jacobs, J. A., 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. J. Bacteriol. 185, 7241-7246.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and others, 2017. Critical Assessment of Metagenome Interpretation- a benchmark of computational metagenomics software. bioRxiv , 099127.

Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K. and Nikolausz, M., 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol. Ecology 60, 341-350.

Soergel, D. A. W., Dey, N., Knight, R. and Brenner, S. E., 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 6, 1440-1444.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U.S.A. 103, 12115-12120.

Stackebrandt, E. and Goebel, B., 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 44, 846-849.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., Siebner, C., Pühler, A. and Schlüter, A., 2015. Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. Biotechnol. Biofuels 8, 14.

Stolze, Y., Bremges, A., Rumming, M., Henke, C., Maus, I., Pühler, A., Sczyrba, A. and Schlüter, A., 2016. Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. Biotechnol. Biofuels 9, 156.

Streit, W. R. and Schmitz, R. A., 2004. Metagenomics - the key to the uncultured microbes. Curr. Opin. Microbiol. 7, 492-498.

Strous, M., Kraft, B., Bisdorf, R. and Tegetmeyer, H. E., 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers Microbiol. 3, 410.

Sturm, M., Schroeder, C. and Bauer, P., 2016. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinformatics 17, 208.

Sundberg, C., Al-Soud, W. A., Larsson, M., Alm, E., Yekta, S. S., Svensson, B. H., Sørensen, S. J. and Karlsson, A. 2013. 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. Microbiol. Ecol. 85, 612-626.

Suzuki, M. T. and Giovannoni, S. J., 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. 62, 625-630.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33-36.

Taub, M. A., Corrada Bravo, H. and Irizarry, R. A., 2010. Overcoming bias and systematic errors in next generation sequencing data. Genome Med. 2, 87.

Teeling, H. and Glöckner, F. O., 2012. Current opportunities and challenges in microbial metagenome analysis - a bioinformatic perspective. Brief. Bioinform. 13, 728-742.

Treu, L., Kougias, P. G., Campanaro, S., Bassani, I. And Angelidaki, I.2016. Deeper insight into the structure of the anaerobic digestion microbial community; the biogas microbiome database is expanded with 157 new genomes. Bioresour. Technol. 260, 6.

Tringe, S. G. and Rubin, E. M., 2005. Metagenomics: DNA sequencing of environmental samples. Nat. Rev. Genet. 6, 805-814.

Tringe, S. G. and Hugenholtz, P., 2008. A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. 11, 442-446.

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.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I., 2007. The human microbiome project. Nature 449, 804-810.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37-43.

Venkiteshwaran, K., Bocher, B., Maki, J. and Zitomer, D., 2015. Relating Anaerobic Digestion Microbial Community and Process Function. Microbiol. Insights 8, 37-44.

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. and others, 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261-5267.

Weiland, P., 2003. Production and energetic use of biogas from energy crops and wastes in Germany. Appl. Biochem. Biotechnol. 109, 263-274.

Weiland, P., 2010. Biogas production: current state and perspectives. Appl. Microbiol. Biotechnol. 85, 849-860.

Weiss, A., Jérôme, V., Freitag, R. and Mayer, H. K., 2008. Diversity of the resident microbiota in a thermophilic municipal biogas plant. Appl. Microbiol. Biotechnol. 81, 163-173.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E. M., Kyrpides, N., Mavrommatis, K. and Meyer, F., 2012. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics 13, 141.

Wood, D. E. and Salzberg, S. L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.

Wright, E. S., Yilmaz, L. S. and Noguera, D. R., 2012. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Appl. Environ. Microbiol. 78, 717-725.

Wu, Y.-W., Simmons, B. A. and Singer, S. W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605-607.

Xia, Y., Wang, Y., Fang, H. H., Jin, T., Zhong, H. and Zhang, T. 2014. Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. Sci. Rep. 4, 6708.

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., Wang, X., Chen, S., He, G., Xiao, X. and Xu, A., 2011. Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME J. 5, 414-426.

Yang, F., Zeng, X., Ning, K., Liu, K.-L., Lo, C.-C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., Chain, P. S., Xie, G., Ling, J. and Xu, J., 2012. Saliva microbiomes distinguish caries-active from healthy human populations. ISME J. 6, 1-10.

Yang, X., Chockalingam, S. P. and Aluru, S., 2013. A survey of error-correction methods for next-generation sequencing. Brief. Bioinform. 14, 56-66.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. and Glöckner, F. O., 2013. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Res. D643-D648.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F. and Xu, Y., 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445-W451.

Youssef, N., Sheik, C. S., Krumholz, L. R., Najar, F. Z., Roe, B. A. and Elshahed, M. S. 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. Appl. Environ. Microbiol. 75, 5227-5236.

Zakrzewski, M., Goesmann, A., Jaenicke, S., Jünemann, S., Eikmeyer, F., Szczepanowski, R., Al-Soud, W. A., Sørensen, S., Pühler, A. and Schlüter, A., 2012. Profiling of the metabolically active community from a production-scale biogas plant by means of high-throughput metatranscriptome sequencing. J. Biotechnol. 158, 248-258.

Zakrzewski, M., Bekel, T., Ander, C., Pühler, A., Rupp, O., Stoye, J., Schlüter, A. and Goesmann, A., 2013. MetaSAMS - a novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets. J. Biotechnol. 167, 156-165.

Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614-620.

Zhang, Q., Shuwen, G., Zhang, J., Fane, A. G., Kjelleberg, S., Rice, S. A. and McDougald, D. 2015. Analysis of microbial community composition in a lab-scale membrane distillation bioreactor. J. Appl. Microbiol. 118, 940-953.

Zhou, Q., Su, X. and Ning, K., 2014. Assessment of quality control approaches for metagenomic data analysis. Sci. Rep. 4, 6957.

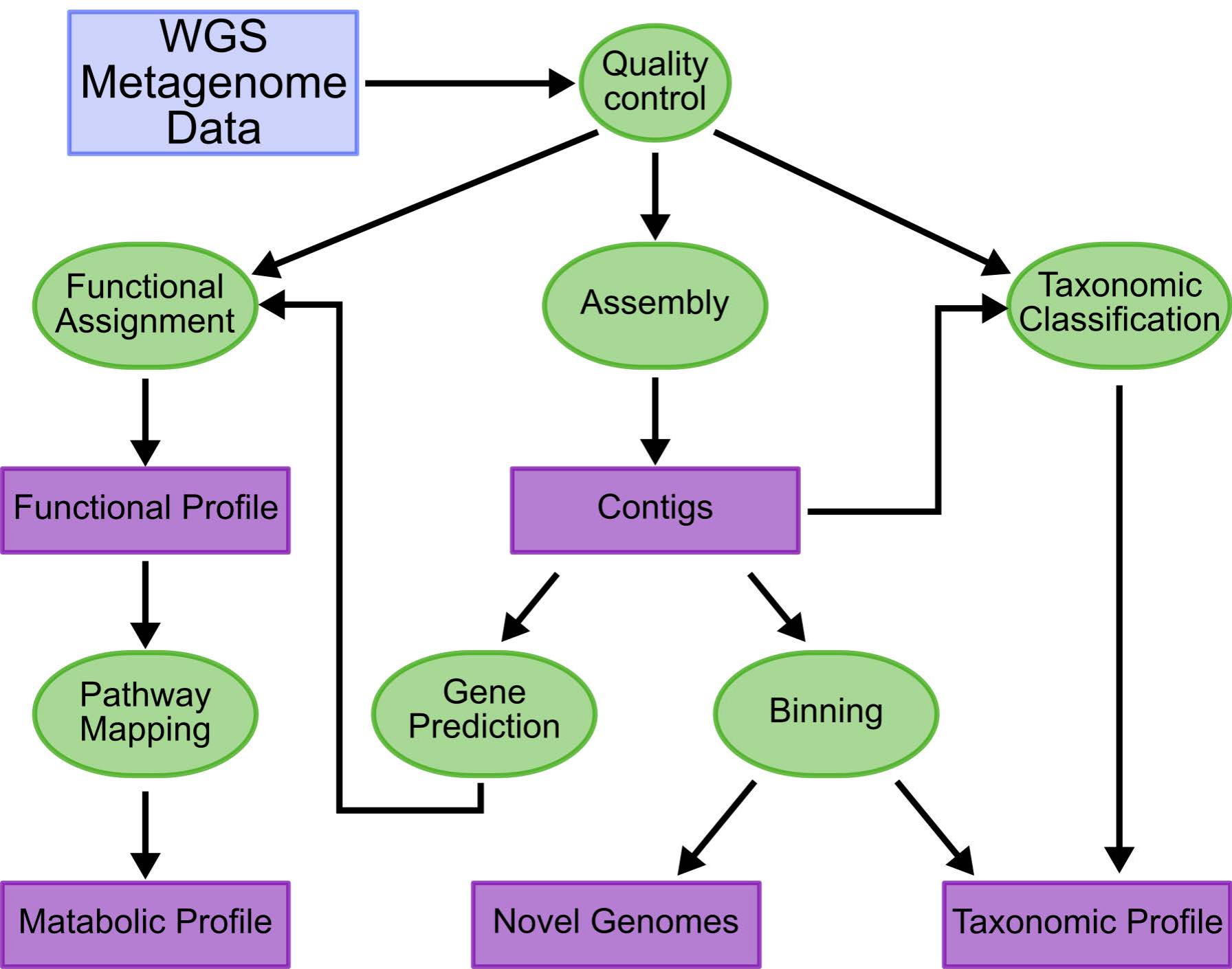
Zhu, W., Lomsadze, A. and Borodovsky, M., 2010. *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res. 38, e132.

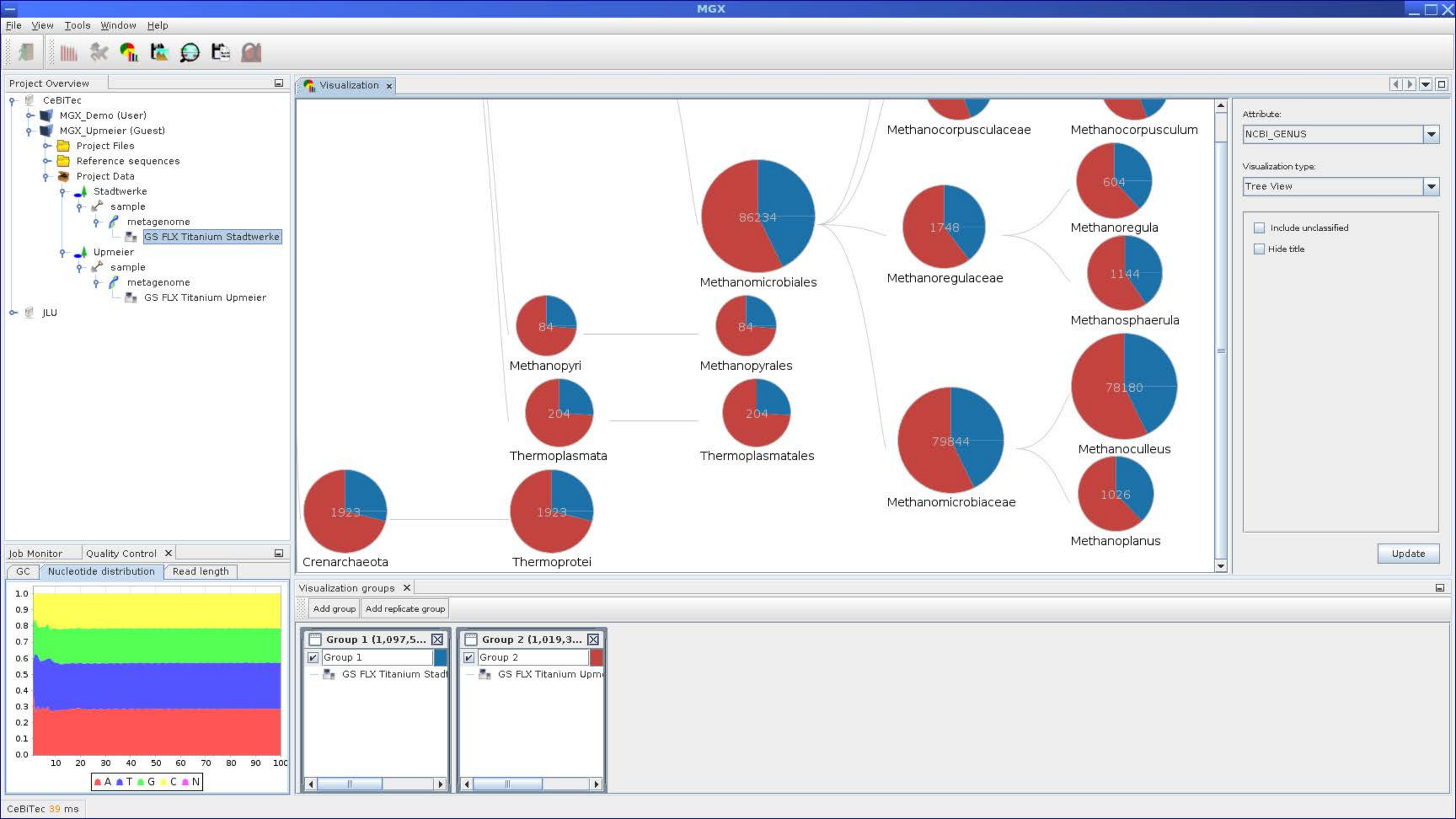
Figures

Figure 1: Schematic overview of the major steps in metagenomic workflows. Square boxes represent data and results, and oval boxes represent processing steps. For more details please see the manuscript sections 3.1 to 3.5.

Figure 2: MGX interactive data visualization. The MGX application allows to create highly interactive charts based on the analysis of unassembled metagenome datasets. The main component of the figure displays a comparative taxonomic plot for two different biogas plants.

Figure 3: The EMGB website is divided into several tabs. A dataset tab comprises of the filter boxes functionality at the top (**a**), the phylogenetic tree view on the left (**b**) and the tables and visualizations at the center of the page (**c**). Additional dataset tabs can be added in the "+" tab (**d**) and all loaded datasets can be compared in the rightmost tab "Comparison" (**e**).


WGS Metagenome Data
Quality control
Functional Assignment
Assembly
Taxonomic Classification
Functional Profile
Contigs
Pathway Mapping
Gene Prediction
Binning
Matabolic Profile
Novel Genomes
Taxonomic Profile


MGX
File View Tools Window Help
Project Overview
CeBiTec
MGX_Demo (User)
MGX_Upmeier (Guest)
Project Files
Reference sequences
Project Data
Stadtwerke
sample
metagenome
GS FLX Titanium Stadtwerke
Upmeier
sample
metagenome
GS FLX Titanium Upmeier
JLU
Visualization
Methanocorpusculaceae
Methanocorpusculum
86234
Methanomicrobiales
1748
Methanoregulaceae
604
Methanoregula
1144
Methanosphaerula
84
Methanopyri
84
Methanopyrales
204
Thermoplasmata
204
Thermoplasmatales
79844
Methanomicrobiaceae
78180
Methanoculleus
1026
Methanoplanus
1923
Crenarchaeota
1923
Thermoprotei
Attribute:
NCBI_GENUS
Visualization type:
Tree View
Include unclassified
Hide title
Update
Job Monitor
Quality Control
GC
Nucleotide distribution
Read length
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
10 20 30 40 50 60 70 80 90 100
A T G C N
Visualization groups
Add group
Add replicate group
Group 1 (1,097,5...
Group 1
GS FLX Titanium Stadt
Group 2 (1,019,3...
Group 2
GS FLX Titanium Upme
CeBiTec 39 ms

EMGB 2

plant1234

Comparison

a

d

e

GO: nitrogen compound metabolic process

GO ID is exactly GO:0006807

OR GO Lineage is exactly GO:0006807

Count is in range 3 to 1000

% Identity is greater than 70

b

- Bacteria (47,396)
  - Firmicutes (28,613)
  - uc_Bacteria (6,826)
  - Bacteroidetes (4,384)
    - Bacteroidia (3,242)
    - uc_Bacteroidetes (976)
    - Flavobacteriia (122)
      - Flavobacteriales (114)
        - Flavobacteriaceae (112)
        - uc_Flavobacteriales (2)
        - Cryomorphaceae (0)
          - Fluviicola (0)
          - Owenweeksia (0)
          - uc_Cryomorphaceae (0)
      - uc_Flavobacteriia (8)
    - Cytophagia (32)
    - Sphingobacteriia (12)
  - Proteobacteria (3,126)

c

Table | GO Stats | Pathways | Gene Count Ratio

Genes 11 to 20 (10,275 total)
Page 2 of 1,028

	All Subject Titles	E-value	% Identity	Length
contig-139628000014_12	DNA methyltransferase [Cecembia lonarensis]	1.00e-42	81.7	109
contig-139628000014_13	Type I restriction-modification system, DNA-methyltransferase subunit M [Mariniradius saccharolyticus]	5.00e-40	70.1	117
contig-871000055_3	MULTISPECIES: DNA-directed RNA polymerase subunit alpha [Alcaligenes]	0	88.1	327