Working together for
a world without waste

Feasibility Report

A rapid and low-cost method for assessing AD plant health through identification of functional microbial communities

A feasibility report from the ‘Driving Innovation in AD’ programme which looks at a rapid and low-cost method for assessing AD plant health through identification of functional microbial communities.

Project code: OIN001-426
Research date: March – June 2012 **Date:** October 2012

WRAP's vision is a world without waste, where resources are used sustainably.

We work with businesses, individuals and communities to help them reap the benefits of reducing waste, developing sustainable products and using resources in an efficient way.

Find out more at www.wrap.org.uk

This report was commissioned and financed as part of WRAP's 'Driving Innovation in AD' programme. The report remains entirely the responsibility of the author and WRAP accepts no liability for the contents of the report howsoever used. Publication of the report does not imply that WRAP endorses the views, data, opinions or other content contained herein and parties should not seek to rely on it without satisfying themselves of its accuracy.

Written by: Matthew Wade, Newcastle University, Civil Engineering & Geosciences
Project Partners: Agrivert, GWE Biogas, Lower Reule Bioenergy, BiogenGreenfinch, CPI

Front cover photography: Laboratory experiment

While we have tried to make sure this report is accurate, we cannot accept responsibility or be held legally responsible for any loss or damage arising out of or in connection with this information being inaccurate, incomplete or misleading. This material is copyrighted. You can copy it free of charge as long as the material is accurate and not used in a misleading context. You must identify the source of the material and acknowledge our copyright. You must not use material to endorse or suggest we have endorsed a commercial product or service. For more details please see our terms and conditions on our website at www.wrap.org.uk

Abstract

DNA sequencing and the ability to characterise the order of nucleotide bases along a DNA fragment has been possible for nearly 40 years. Next Generation Sequencing (NGS) technology has recently emerged as a defining tool that offers rapid and cost-effective characterisation of DNA. It has pushed the boundaries of what may be achieved in microbial ecology, and offers the potential for gaining a broader and deeper understanding of the anaerobic digestion process through effective identification of the principle micro-organisms involved at each metabolism step of the anaerobic digestion pathway. Observing how the microbial populations react to process perturbations and understanding the overall microbial diversity and abundance of the AD process over time is key to improving system performance from a biological perspective.

The central aim of this feasibility study was to assess the viability of Next Generation Sequencing, specifically the Ion Torrent platform, for providing a deeper insight into the microbial ecology of full-scale anaerobic digester systems. The study represents one of the first instances that NGS has been used for characterising microbial community structure of these systems. A total of three samples from a full-scale farm-based anaerobic digester were collected over different days and were sequenced using the Ion Torrent Personal Genome Machine. The time required for sequencing was less than two hours at a cost of around £1400. The results indicated that the microbial community of the anaerobic digester was stable over the period of sampling and that NGS was able to identify key groups of organisms relating to anaerobic digestion performance, such as methanogenic archeae. Phase 2 of the project intendeds to use the technology and analysis methods across a range of AD systems to demonstrate the potential of NGS for varying plant designs/configurations, operating conditions, feedstocks and digester functionality. As such, the expected demonstrations will vary in their objectives, tailored to suit the individual requirements of the AD plants.

It is expected that the outcomes from Phase 2 will help develop a platform for a commercial service that offers sequencing and bioinformatics to AD operators, subsequently leading to better AD design and process optimisation.

Executive summary

DNA sequencing and the ability to characterise the order of nucleotide bases along a DNA fragment has been possible for nearly 40 years. In 2005, a transformation in sequencing technology occurred with the development of the sequencing-by-synthesis method by 454 Life Sciences, known as pyrosequencing. 454 and other Next Generation Sequencing (NGS) technologies have since emerged as the defining tool in DNA characterisation at a greater speed and accuracy and at a lower cost.

The aim of NGS is to make DNA sequencing simpler, faster and cheaper. This is resonant to the needs of the anaerobic digestion community who have not had the exposure to sequencing technology as with other sectors due to the complex nature and specialist needs of DNA sequencing, the relative duration of sample processing and high sequencing costs.

To process and analyse the data generated by NGS, informatics and mathematical techniques (termed bioinformatics) are required to handle large amounts of data and to visualise and report upon the significant information contained in the raw sequence files. To ensure that the bioinformatics processing does not become a bottleneck in the overall workflow and to maintain the rapid turnaround between sample and delivery of results to the plant operator, high-performance/high-throughput computing is typically employed to maximise the computer processing capacity whilst minimising run time and cost.

NGS has pushed the boundaries of what may be achieved in microbial ecology. It offers the potential to gain a broader and deeper understanding of the anaerobic digestion process through effective identification of the principle micro-organisms involved at each metabolism step. By observing how the microbial populations react to process perturbations an understanding of the overall microbial diversity and abundance within the AD process can be achieved.

This feasibility study represents one of the first instances that NGS has been used for characterising microbial community structure of an anaerobic digestion system. The sequencing data can be used to correlate specific microbial activity with both environmental and performance related conditions to develop a knowledge-base for aiding decision-making and, ultimately, facilitate optimisation of digester operation.

The central aim of the feasibility study was to assess the viability of Next Generation Sequencing, specifically the Ion Torrent platform, for providing a deeper insight into the microbial ecology of full-scale anaerobic digester systems. The study represents one of the first times that NGS has been used for characterising microbial community structure of these systems. A total of three samples from a full-scale farm-based anaerobic digester were collected over different days and sequenced using the Ion Torrent Personal Genome Machine. The time required for sequencing was less than two hours at a cost of around £1400.

The results indicated that the microbial community of the anaerobic digester was stable over the period of sampling and that NGS was able to identify key groups of organisms relating to anaerobic digestion performance, such as methanogenic archeae.

The demonstration phase (Phase 2) of the project aims to develop a more rigorous approach for AD plants to exploit and benefit from NGS technology. Samples will be collected from a range of AD facilities within the waste sector, covering different operating configurations (e.g. one or two phase), feedstocks, scales etc. Process data will be collated from available on-site measurements and supplemented with laboratory chemical analysis where necessary in order to correlate process information with biological data generated by NGS. Bioinformatics and statistical data analysis tools will be used to process this information, with the aim to generate useable knowledge for addressing AD plant performance issues.

It is expected that the outcomes from Phase 2 will help develop a platform for a commercial service that offers sequencing and bioinformatics to AD operators, subsequently leading to better AD design and process optimisation.

Contents

Glossary

AD	Anaerobic Digestion
bp	Base pair
DGGE	Denaturing Gradient Gel Electrophorisis
FISH	Fluorescence In-Situ Hybridization
Mb	Megabases
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
PGM	Personal Genome Machine
RDP	Ribosomal Database Project
SFF	Standard Flowgram File

1.0 Introduction and background

1.1 Company/consortium

Environmental Engineering Group, Civil Engineering & Geosciences, Newcastle University

The School of Civil Engineering and Geosciences (CEG) at Newcastle University sees the current challenges of science and engineering at the interface between technological, human and natural systems. This has resulted in a research strategy that integrates engineering with biogeochemical and social sciences, placing us at the forefront of interdisciplinary research under the umbrella of Earth Systems Science Engineering and Management (ESSEM). The proposed research will benefit greatly from the complementary internationally recognised expertise and facilities currently available in CEG, and aims to exploit these synergies fully. This includes world-leading research into theoretical microbial ecology, carbon-efficient wastewater treatment, geochemistry of contaminated soils and sediments, petroleum biogeochemistry, climate modelling, and engineering tools for adaptation to climate change.

Civil Engineering was the highest achieving unit of assessment in NU in the last Research Assessment Exercise (RAE 2008), and the second best in the UK in terms of Research Power. Water and environmental engineering research are particularly strong, resulting in prestigious*: funding awards* – e.g. Marie-Curie Excellence Grant (ECOSERV; €1.8M); EPSRC Platform Grant, "Generic Unifying Concepts in Wastewater Treatment Design" and its renewal (GR/S59543/01; £412K, and EP/F008473/1; £775K) on which I am Co-I; EPSRC Platform Grant, "Earth Systems Engineering: Sustainable systems engineering for adapting to global change (EP/G013403/1;£1.3M); and a partner in the EPSRC STREAM, Industrial Doctorate Centre for the Water Sector; *esteem*; e.g. Queen's Anniversary Prize in Higher Education 2005 awarded to the HERO group of Prof Younger/Dr Jarvis; Fellow of the Royal Academy of Engineering; Profs. O'Connell, Younger, and Hall International Society for Microbial Ecology, Young Investigator Award to Prof Head; and Fellowship awards to; Dr. Davenport (RCUK), Prof. Curtis (BBSRC), Dr. Dawson (EPSRC), Dr. Fowler (NERC), and Dr. Jarvis (Environment Agency); and *high impact publications* in *Nature* and *PNAS*.

CEG has outstanding laboratory facilities central to the success of this project. Those for microbiology include laboratories equipped with leading-edge technologies for cultivation and cultivation-independent methods such as; an Ion Torrent Personal Genome Machine (Next Generation Sequencer), a confocal laser scanning microscope (CLSM; EPSRC, £200K), epifluorescence microscopes, a flow cytometer, and three quantitative real-time PCR thermocyclers for quantification of specific microbial populations, and specialised software for databasing and statistical analyses of microbiological data. There are also excellent facilities for both inorganic and organic chemistry, including a suite of advanced gas- and liquid-chromatography mass spectrometers and pyrolysis mass spectrometers, which will be bolstered by a state-of-the-art triple quadrupole liquid chromatography mass spectrometer (LC-MS^2) for the quantitative analysis of pollutants at very low concentrations if the proposal is successful. There are excellent computing facilities for climate, geomatic, hydrological modelling; and experimental facilities in the River Eden, a Defra Demonstration Catchment for the Water Framework Directive, of relevance to this research.

Environmental Engineering at Newcastle University has a long, proud and successful history, since it was established 60 years ago making it one of the oldest such groups in the country. The importance of biology was recognised from its inception, being one of the first to include microbiological laboratories. It has a long and proud tradition of research and practice in anaerobic digestion ever since its scientists published the first gas-chromatography method for the quantification of methane in Nature in 1956.

1.2 Introduction to your technology

Background

Traditionally, knowledge of anaerobic digestion has arisen from an empirical understanding of the process increasingly supplemented by on-line measurements that can provide a close to real-time description of the performance of the system. However, these process measurements are only indicative of the true state of the plant and operational decisions are often based on a combination of operator experience and a periodic review of the process data. Absence of critical process data or reception of erroneous measurements can result in sub-optimal plant performance or process failure. Furthermore, modern engineered systems tend to rely on process measurements for modelling, control and optimisation of critical and sensitive components of the process, and this ability is currently absent from most anaerobic digestion systems.

Some effort has been made in recent years to improve performance of bioprocesses such as activated sludge plants and, more recently, anaerobic digesters through a combination of empirical and mechanistic means. Empirical or data-driven methods are founded on the knowledge that the underlying process phenomena are of infinite complexity and measurement data can best represent this. Mechanistic methods aim to develop simplified representations of the process mechanisms that can be used to generate new knowledge or predict behaviour in a system. The two methods are not distinct, mechanistic models rely on observations in order to apply assumptions, rules and determine parameters, whereas empirical models must contain some relational element describing the expected influence of inputs on outputs.

The monitoring and diagnosis of anaerobic digestion performance is complex due to the large number of integrated processes and operating parameters involved, as well as the various chemical and compound transformations that occur at different time scales. In order to optimise process performance it is necessary to identify the microbiological organisms present at each process step and to characterise the reactor hydrodynamics and microbial population kinetics, accordingly.

In many AD reactors, microbial identification is not undertaken due to several limiting factors including cost, lack of technical know-how/availability and over-reliance on traditional operational practices. However, as the performance of bioreactors is ultimately dependent on the characterisation of the active microbial populations, it is becoming evident that a greater understanding of the composition and health of these communities during operation can lead to major performance improvements in the system.

Techniques such as Fluorescence In-Situ Hybridization (FISH) have been employed to detect the presence of broad groups of performance related microorganisms (e.g. methanogens). However, the technique is time consuming and requires design of probes that target specific organisms based on assumptions of their presence in any given sample. Hence, process diagnosis may be achieved, but there is a corresponding level of uncertainty and a lag-time of days between sample and result.

Next Generation Sequencing

DNA sequencing technology has been used in the identification of microbial populations for a number of years. However, studies have been limited to laboratory scale identification of, generally, the dominant species under specific process conditions (e.g. process start-up). With Next Generation Sequencing (NGS) technology pushing the boundaries of what may be achieved in microbial ecology, the potential for gaining a broader and deeper understanding of the anaerobic digestion process is becoming a reality.

The aim of NGS is to make DNA sequencing simpler, faster and cheaper. This is resonant to the needs of the anaerobic digestion community who have not had the exposure to sequencing technology as with other sectors due to the complex nature and specialist needs of DNA sequencing, the relative duration of sample processing and high sequencing costs.

To process and analyse the data generated by NGS, informatics and mathematical techniques (termed bioinformatics) are required to handle large amounts of data and to visualise and report upon the significant information contained in the raw sequence files. To ensure that the bioinformatics processing does not become a bottleneck in the overall workflow and to maintain the rapid turnaround between sample and delivery of results to the plant operator, high-performance/high-throughput computing is typically employed to maximise the computer processing capacity whilst minimising run time and cost.

1.3 Proposal (technology/concept) background

1.3.1 Where – origins of technology?

DNA sequencing and the ability to characterise the order of nucleotide bases along a DNA fragment has been possible for nearly 40 years. Chain termination or Sanger sequencing was developed in 1977 and has been the principle method for determination of DNA sequences for the majority of the intervening years. The increasing necessity for sequencing within the medical, public health and biological fields resulted in a growth in laboratories installed with automated sequencers and the number of trained staff. In 2005, a transformation in sequencing technology occurred with the development of the sequencing-by-synthesis method by 454 Life Sciences. The first pyrosequencing platform was 300 times cheaper than Sanger technology and, as importantly, was able to produce data equivalent to several hundred Sanger sequencers with only a single operator.

The technological progress in next-generation sequencing has been characterised by rapid development, increasing performance and dramatic reduction in costs. This development has led to a number of so-called 2nd-generation sequencing platforms reaching the market (i.e. 454, Illumina, SOLiD, Ion Torrent), each providing greater depth or breadth of sequencing, and at ever decreasing cost to the customer (Table 1). Table 2 gives a more detailed comparison of the four 2nd-generation sequencing platforms, indicating that Ion Torrent is relatively inexpensive. However, no sooner has 2nd-generation sequencing become established within the scientific community than the promotion of a 3rd-generation of sequencers has begun. These systems (GridION, PacBio RS, Qdot) offer single molecule DNA sequencing in real-time, with the potential to fundamentally change how scientists and engineers work on biological problems.

Table 1 Comparison of generational sequencing technologies

	Clone-Seq.	**2nd Gen Seq.**	**3rd Gen Seq. (Potential)**
No. reads	<100	1,000,000 +	120-1000 bases/min > 6 hours 10 billion bases per day
Generation time	~2 days	6 hours – 2 days	
Cost per GBase	1000$	<200$	40$

Table 2 Performance and cost comparison of 2nd generation sequencing technologies

	Ion Torrent	454 Sequencing	Illumina	SOLiD
Sequencing Chemistry	Ion semiconductor sequencing	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Emulsion PCR	Bridge amplification	Emulsion PCR
Mb per run	100	100	600,000	170,000
Time per run	1.5 hours	7 hours	9 days	9 days
Read length	200 bp	400 bp	2x150 bp	35x75 bp
Cost per run	$ 350 USD	$ 8,438 USD	$ 20,000 USD	$ 4,000 USD
Cost per Mb	$ 5.00 USD	$ 84.39 USD	$ 0.03 USD	$ 0.04 USD
Cost per instrument	$ 50,000 USD	$ 500,000 USD	$ 600,000 USD	$ 595,000 USD

1.3.2 What – what has been achieved to date?

This project represents one of the first instances that NGS has been used for characterising microbial community structure of an anaerobic digestion system. However, the Environmental Engineering group have successfully analysed samples from other engineered biological systems, such as wastewater treatment plants and microbial fuel cells using 454 pyrosequencing (GS-FLX and Titanium) and more conventional methods such as qPCR, FISH and DGGE.

A handful of publications have appeared in recent years, applying 454 technology to the characterisation of microbial community metagenomes in biogas plants (Schlüter *et al.*, 2008, Kröber *et al.*, 2009 and Sundberg *et al.*, 2011). Most recently, Wirth *et al.* (2012) have characterised AD systems using a short-read NGS platform (SOLiD), comparing the results to the longer read 454 technology. These existing studies can be used by subsequent work for comparative analysis.

1.3.3 Why – Why this technology?

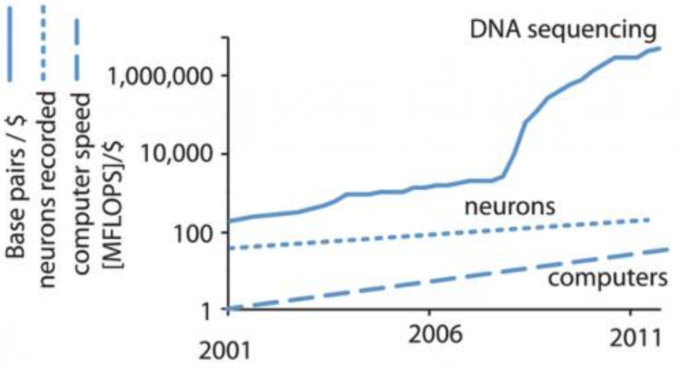
In terms of microbiological characterisation, existing scientific studies have merely scratched the surface in identifying the microbial communities present in anaerobic digestion systems. Previous studies have been restricted to focusing on a single process or sample and were carried out at the lab-scale level with limited temporal monitoring. This is generally due to the cumbersome, time consuming and highly expensive technology available for DNA extraction and sequencing. The value of Next Generation Sequencing over comparable techniques can be viewed in terms of performance and cost as follows:

- ***Better*** – The technology will allow for additional and previously undetermined information to be gathered by extraction and sequencing of the microbial DNA from samples collected in the AD process. This will allow for a more thorough analysis of the microbial communities present in the system, including their abundance, diversity and influence on performance. AD process optimisation can be achieved by having a more reliable insight into the behaviour and structure of microbial populations that directly relates to process performance and, subsequently, to engineering principles. It is the *knowledge* that is generated by analysing and interpreting the data that will add real and novel value to the way that industries such as AD address process design and operational challenges in the future.
- ***Quicker*** – The continuing development of NGS technology means that, currently, a sequencing workflow - from raw sample to basic analysis – takes under eight hours with

minimal manual laboratory time and a relatively simple protocol. The speed and accuracy of the process means that more samples can be analysed and the results obtained quickly to provide a better understanding of changes in the microbiology of the plant over time. Together with process monitoring measurements or periodic spot sampling, the DNA sequencing analysis can be used as a process diagnostic tool, an indicator of plant health or to provide a deeper understanding of the microbial behaviour of the system, i.e. application of *knowledge* to engineering.

- ***Cheaper*** – Cost is often the critical factor for implementing novel technologies into existing processes. This has been the case with DNA sequencing whereby analysis costs were prohibitively high prior to the advent of second-generation sequencing technology in early 2008. Since then, per Megabase (1Mb = 1000 DNA nucleotides), cost of sequencing has dropped dramatically from around 1000$ to under 1$ (Figure 1), resulting in the potential for much greater sequencing effort to be undertaken across many applications.

Figure 1 The changing economics of high-throughput sequencing in comparison to computing (Moore's Law) and neuroscience (Stevenson & Kording's Law)



1.4 Application of your technology into the UK AD industry now and into the future

Sequencing technology aims to implement a rapid and cost-effective method for determining the principle micro-organisms involved at each metabolism step of the anaerobic digestion pathway. By observing how the microbial populations react to process perturbations an understanding of the overall microbial diversity and abundance within the AD process can be achieved. The sequencing focuses on the 16S rRNA gene as it is highly conserved between different species of bacteria and archeae, with hyper-variable regions making it suitable for species identification. However, as the technology develops, targeting of specific genes or whole genomes may become of greater interest.

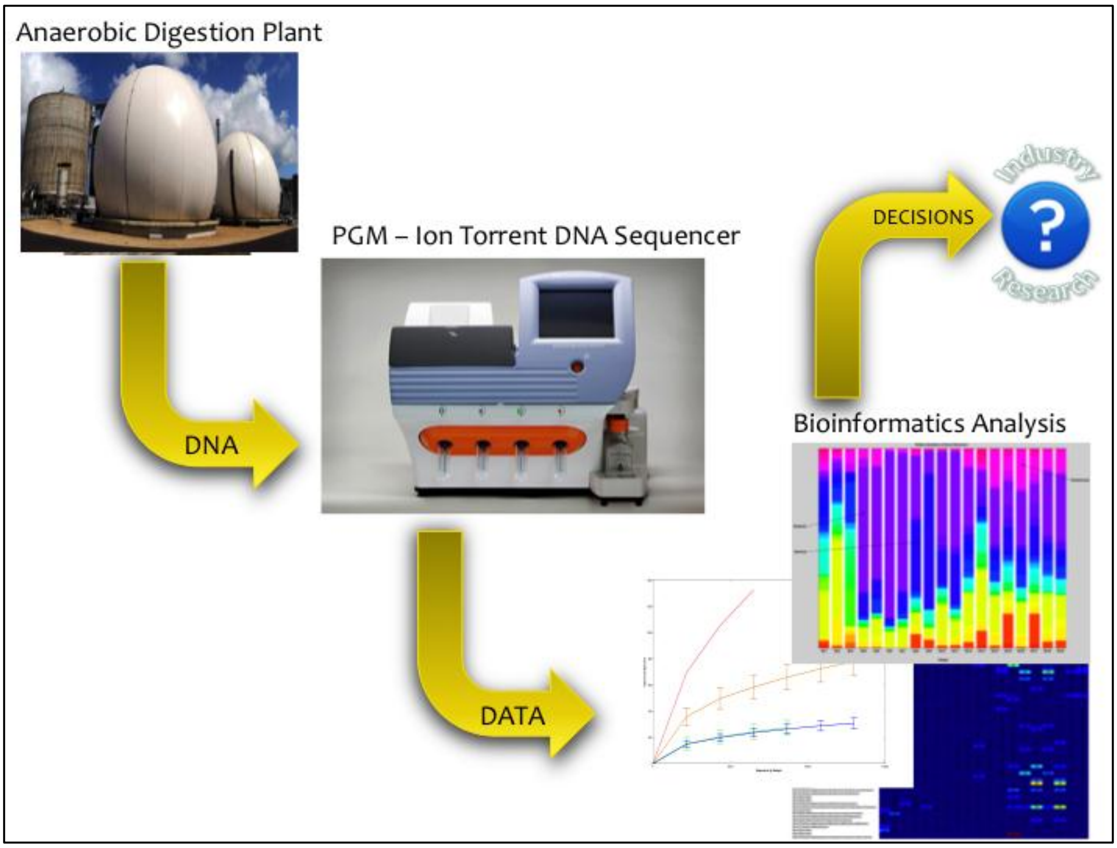
The sequencing data can be used to correlate specific microbial activity with both environmental conditions and process performance. This will lead to the development of a knowledge-base that can aid engineering decisions and, ultimately, facilitate optimisation of digester operation.

It is the development of this knowledge-base that will be the basis for facilitating acceptance of the technology by the industry, in that it:

- Will offer a direct relation between process knowledge and biological data acquired through microbial analysis (NGS).

- Can be tailored as a bespoke service to address site-specific challenges (e.g. operational changes, plant performance) or as a more general system characterisation.
- Negates requirement for operator training or implementation of new technology on-site.
- Will minimise end-user requirements to only sample collection and delivery of process measurement and operational data. Sequencing and bioinformatics tasks will be handled "in-house" to deliver results that are transparent and appropriate to the end-user.

Figure 2 Information flow of proposed service



2.0 Project Objectives

2.1 What did you set out to achieve from the feasibility study and what are your aims for the demonstration

The central aim of the feasibility study was to assess the viability of Next Generation Sequencing, specifically the Ion Torrent platform, for providing a deeper insight into the microbial ecology of full-scale anaerobic digester systems. Further to this, the study aimed to assess the current limitations of the technology in respect to both practical effort in operation of the sequencer and also downstream analysis of the data via bioinformatics tools. Essentially, it is known that NGS will provide much greater sequencing information than Sanger or DGGE/FISH and at a relatively lower cost. However, as Ion Torrent is still an emerging technology, the feasibility study was required to demonstrate any outstanding issues that may inhibit effective and efficient processing of biological samples.

The following deliverables have been completed within the feasibility phase:

- Sampling of Cockle Park Anaerobic Digester: 3 samples taken from primary digester;
- Next Generation Sequencing of samples: Ion Torrent run completed successfully;

- Bioinformatics analysis of sequence data: QIIME analysis presented in this report;
- Assessment of results and engineering interpretation: Presented in this report; and
- AD site selection for Phase 2 demonstration: Contact from four waste companies and potential for further sites to be included.

The demonstration phase of the project aims to develop a more rigorous approach for AD plants to exploit and benefit from NGS technology. Samples will be collected from a range of AD facilities within the waste sector, covering different operating configurations (e.g. one or two phase), feedstocks, scales etc. Process data will be collated from available on-site measurements and supplemented with laboratory chemical analysis where necessary in order to correlate process information with biological data generated by NGS. Bioinformatics and statistical data analysis tools will be used to process this information, with the aim to generate useable knowledge for addressing AD plant performance issues. As well as the diagnostic aspect of the demonstration phase, the data generated will be stored in a database that will act as a repository for information on AD process behaviour, characterised by both biology and physio-chemical information. The demonstration phase will also evaluate the suitability of offering NGS as a service for waste companies, in which the processing, analysis and development of process knowledge (e.g. for optimisation, characterisation or failure diagnosis and identification) are offered based on the needs of the end-user.

A commercial service will be developed to exploit the knowledge and know-how, as well as the data, generated by the sequencing and bioinformatics work. The exact mechanism by which the service will be implemented will be explored within Phase 2 in order to best deliver the right information to the end-user. It is likely that the industries will provide an initial assessment of their requirements based on an identified challenge or scenario (e.g. plant start-up, process failure, introduction of new feedstock). A contract will be made by the University offering to provide NGS of sample(s), data analysis and to collaborate with the industrial partner to meet the expected outcomes through improved biological and process knowledge.

2.2 Meeting the desired outcomes of the 'Driving Innovation' programme

The proposal aims to address the challenges facing the AD sector as follows:

- Given the financial constraints faced by AD operators in the current market, and the limited resources available for development and investment in new technology, the comparatively low cost and rapidity of new sequencing technology is attractive.
- Knowledge and understanding of operational AD systems are limited by the extent, reliability and quality of process measurements available. Optimisation of the plant, either at a coarse scale (standard operating procedure adjustments) or a fine scale (automatic control) may suffer from poor or sparse data. Rapid DNA sequencing coupled with a robust bioinformatics pipeline can ameliorate this issue by providing timely and detailed analysis of the underlying microbiology of the system. Coupled with the process data (environmental and performance information), this is a powerful tool for assessing the health of the AD plant and guiding decision-making through targeted diagnostics.
- With regard to the level of expertise required for performing the sequencing, as well as the level of knowledge required for interpretation of the results, operators will only be required to provide samples of the system at pre-defined frequencies, or at times when the process is performing sub-optimally or below a performance threshold. Sequencing of the samples would be performed by a third-party service (Newcastle University – Environmental Engineering Group at present), but the bioinformatics analysis has the potential to be performed either by third-party or by the AD operators themselves.

Bioinformatics pipelines are becoming more user-friendly as the user-input requirements are reducing and high-performance computing techniques is more accessible, more widespread and less costly.

- The use of the novel techniques for assessing the health and optimisation of AD plants presents a strong opportunity for both technology transfer and for accessing state-of-the-art research and development with partnership between industry and academia. Anaerobic digestion systems are generally multi-disciplinary by nature, but this is often not given much attention outside of research. However, this proposal draws in knowledge from microbial ecology, process engineering, and informatics, with scope to address other fields such as control engineering, plant design and strategy development. By assessing the feasibility, validity and worth of low-cost sequencing / bioinformatics for AD systems, the industry has the potential to achieve significant gains in understanding the technologies capabilities. Furthermore, interest from a broad spectrum of research bodies, investors and stakeholders can be realised.

3.0 State of technology

3.1 Development history of the technology

Next Generation Sequencing technologies are in constant development, having emerged in 2005 via the 454 Life Technologies pyrosequencing platform. The 2nd-generation sequencing platform used in this feasibility study, the **Ion Torrent Personal Genome Machine** (PGM™), has been available on the market since February 2010. Ion Torrent is the first commercial ion semiconductor DNA sequencing system. It uses proprietary semiconductor sensors to perform real-time measurement of the hydrogen ions produced during DNA replication. A high-density array of wells on the ion semiconductor chips provides millions of individual reactors while integrated fluidics allow reagents to flow over the sensor array. This combination of fluidics, micromachining, and semiconductor technology enables the direct translation of genetic information (DNA) to digital information (DNA sequence). Its use in this project represents one of the first applications in the field of microbial ecology and, as such, offers a valuable insight into its benefits and limitations in addressing the challenge of characterising the microbial community structure of AD plants.

3.2 Previous use/evidence to support the use of your technology from other countries, sectors or industries where applicable

Currently, the Ion Torrent technology has 17 citations in academic and cited trade literature (Scopus, 2012), of which 10 relate to the medical, public health and genetic fields, 6 are focused on technological assessment and only 1 citation discusses its use in non-medical microbiology (food industry). However, considering NGS technologies as a whole, then there are over 3000 citations in academic literature covering a range of fields, including microbial ecology and engineered biological systems. Of these, only 4 are concerned with observing or characterising anaerobic microbial communities, typically rumen based organisms.

3.3 Previous tests i.e. desk-based studies, lab-scale, on the ground

The Environmental Engineering group at Newcastle University has extensive experience working with anaerobic digestion and the microbial ecology of these systems. Although the acquisition and operation of Ion Torrent is relatively new, the group have a long-standing background in using molecular tools and techniques for engineered biological systems, including handling and processing of both Sanger sequencing and next generation sequencing data (454 - pyrosequencing).

4.0 Legislation

As it is not anticipated that the technology is to be installed on-site at the AD facilities but offered as a service initially by the University, there are no legislative or regulatory requirements applicable to the NGS. However, there will be a requirement for sampling of the anaerobic reactors prior to sequencing in which collection of material from the site and transportation to the sequencing facility will be performed. All handling and processing and biological materials will be covered by University health and safety policy, including adherence to the COSHH risk assessment guidance. All laboratory work will be performed by staff trained in handling biological materials and chemicals required for operating the Ion Torrent sequencer.

5.0 Detailed technical appraisal of technology

5.1 Theory/process behind the technology

Sampling
In order to test the feasibility of the sequencing technology described above, three samples were collected from the University's Anaerobic Digestion facility at Cockle Park farm in Northumberland (see Table 3) and stored at -20°C prior to sample preparation. The feedstock to the digester comprised a mixture of swine and cattle slurry from the farm.

Table 3 Sample information

	Date	Type	Process temperature (°C)
Sample 1	01/03/2012	Sludge from primary digester	37.5
Sample 2	11/04/2012	Sludge from primary digester	40.6
Sample 3	26/04/2012	Sludge from primary digester	38.5

Figure 3 Cockle Park anaerobic digester

Sample preparation
DNA was extracted from each sample using the Fast DNA Soil Kit (Qbiogene Inc.,). An enzymatic-based 16S rRNA library was then constructed by using the Ion Xpress^{TM} Plus Fragment Library Kit (Life Technologies) and pre-determined primers. The forward (Britschgi T.B., 1994, FEMS Microbiology Ecology 13, 225-232) and reverse primers (Sogin M.L., 2006, PNAS, 103(32), 12115-12120) selected for best coverage of the microbial populations are shown in Table 4, and the calculated coverage according to the RDP 16S rRNA gene sequence database is listed in Table 5.

Table 4 Forward and Reverse (Universal) Bacterial Primers selected for DNA extraction

Annealing Temp	Type	Region	Forward	Forward sequence (5'-3')	Position
52-58	Bacteria (Universal)	V6	926f	AAACTYAAAKGAATTGRCGG	906-926
60-62	Bacteria (Universal)	V6	1046r	CGACAGCCATGCANCACCT	*1064-1046*
58-60	Bacteria (Universal)	V6	1046r-PP	CGACAACCATGCANCACCT	*1064-1046*
62-64	Bacteria (Universal)	V6	1046r-AQ1	CGACGGCCATGCANCACCT	*1064-1046*
60-62	Bacteria (Universal)	V6	1046r-AQ2	CGACGACCATGCANCACCT	*1064-1046*

Table 5 Coverage evaluation of the universal bacterial primers

Bacteria		Archea	
Hits in RDP probe match	**% coverage of 16S rRNA gene sequences in RDP**	**Hits in RDP probe match**	**% coverage of 16S rRNA gene sequences in RDP**
923275/986742	94	12100/13075	93
398885/986742	40	3/13075	0.02
524212/986742	53	1/92463	0.001
16789/986742	1.70	2708/13075	20.71
4835/986742	0.49	0/13075	0.00

The templates for sequencing analysis were prepared using the automated Ion OneTouch System, following the IonSet1^{TM} Template Kit (Life Technologies). Barcode adaptors were added to each sample library to enable identification of sample specific sequences after sequencing, as shown in Table 6.

Table 6 IonSet1 Template Kit Barcodes and Adaptors

	Barcode	**Adaptor**
Sample 1	TACTCACGATA	CTGCTGTACGGCCAAGGCGT
Sample 2	TCGTGTCGCAC	CTGCTGTACGGCCAAGGCGT
Sample 3	TGATGATTGCC	CTGCTGTACGGCCAAGGCGT

The templates were prepared for sequencing using the Ion Sequencing Kit (Life Technologies) protocol and then sequenced with the Ion Torrent Personal Genome Machine^{TM} (Life Technologies).

Raw sequences were compiled as a FastQ file from the standard flowgram file (SFF) generated by the Ion Torrent software (*Torrent Suite 1.5*, Life Technologies). The analysis options were modified to allow for the retention of barcodes and primers in the FastQ file, which are normally trimmed from the sequences by the software. This was required to enable bioinformatics tools to discriminate between the three samples. However, the key sequence (TCAG) used for signal calibration was removed during this step. The final sequences stored in the FastQ file have the following structure:

BARCODE-ADAPTER-PRIMER-TARGET_SEQUENCE-REVERSE_PRIMER

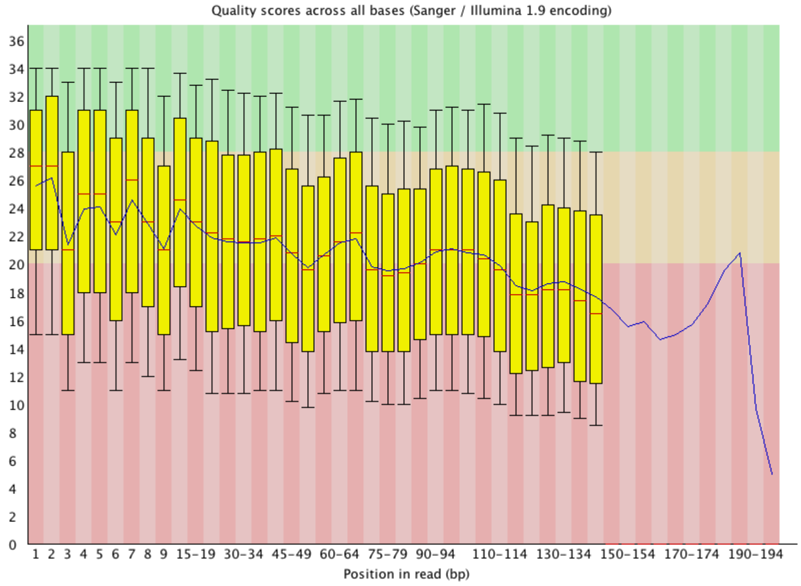
The FastQ file contains both sequence information and per base quality scores, which are required for downstream bioinformatics analysis. A perl language script called *fasta_convert.py* was used to convert the FastQ file to a .fasta file (sequences) and .qual file (quality scores). The script was modified to allow for Ion Torrent encoding as the quality scores produced are in Phred (-10*log10(error)) scale and in the FASTQ file they are encoded in ASCII with an offset of 33 (this was 64 in the original script used for Illumina encoding).

ii. Inputs and outputs including energy balance and mass balance, process flow and/or other technical diagrams

The overall quality of the Ion Torrent run was determined using the FastQC software (Babraham Bioinformatics), which extracts the information from the FastQ file to produce a series of metrics and plots that determine run performance. Figure 4 shows the positional per base quality score statistics for all sequences present in the FastQ file, adjusted to Ion Torrent encoding. The software produces a Box Whisker plot with the following elements:

- *Median Value:* the central red line.
- *Inter-quartile range (25-75%):* the yellow box.
- *10% and 90% points:* the upper and lower whiskers.
- *Mean Quality Score:* blue line.

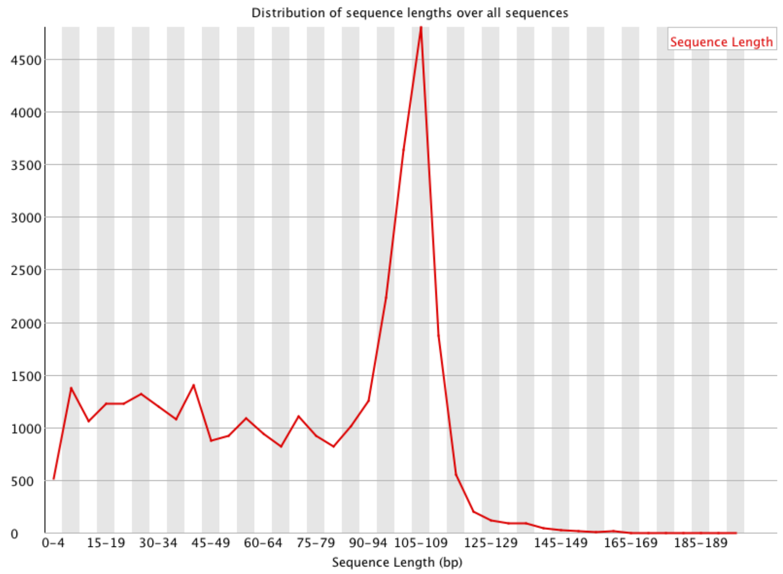
Figure 4 Quality scores statistics across all bases according to read position



The quality score statistics indicate that quality deteriorates with length along the sequence, which has been a phenomenon reported in 454 sequences, for example. Nevertheless, quality is satisfactory up to 100 base pairs.

Figure 5 shows the distribution of sequence lengths and indicates that the majority of sequences are around 100bp, which was the anticipated result. Some shorter fragments are noticeable and may be present due to issues in the PCR amplification step, but these can be excluded from further analysis, as they do not provide meaningful comparisons with library databases.

Figure 5 Sequence length distribution curve



QIIME analysis

Bioinformatics analysis was performed using the QIIME (Quantitative Insights Into Microbial Ecology) open-source software (Qiime V1.5, qiime.org), which is used for comparison and analysis of microbial communities based on high-throughput amplicon sequencing data. The software offers a wide range of analysis tools and the following steps were taken to process the Ion Torrent sequencing data:

- The individual sequences in the fasta file were binned according to sample they were extracted from using their assigned barcodes. A mapping file was used to specify sample name, barcode, adaptor-primer, reverse primer and sample description. The Qiime script **split_libraries.py** was used with the .fasta, .qual and mapping file to perform the binning, with additional parameters set as follows:
  - –b 11: indicates the length of the barcode if different from the default (12).
  - –l 100: specifies a minimum length of read to be retained (100bp). All sequences below this threshold are removed from the sequence file. 100 was selected as ~50bp are accounted by barcode-adaptor-primer.
  - –s 20: specifies the minimum average quality score for each read. If the average score is below 20 for any sequence, it is removed from the sequence file.
- The additional sequence trimming resulted in the following statistics:

Number raw input seqs	41304
Length outside bounds of 100 and 1000:	22598
Num ambiguous bases exceeds limit of 6:	0
Missing Qual Score:	0
Mean qual score below minimum of 20:	4112
Max homopolymer run exceeds limit of 6:	36
Num mismatches in primer exceeds limit of 0:	8959

```
Sequence length details for all sequences passing quality filters:
Raw len min/max/avg          100.0/199.0/134.6
Wrote len min/max/avg        49.0/148.0/83.6

Barcodes corrected/not       44/7
Uncorrected barcodes will not be written to the output fasta file.
Corrected barcodes will be written with the appropriate barcode category.
Corrected but unassigned sequences will not be written unless --retain_unassigned_reads is
enabled.

Total valid barcodes that are not in mapping file      0
Sequences associated with valid barcodes that are not in the mapping file will not be written.

Barcodes in mapping file
Num Samples  3
Sample ct min/max/mean: 1221 / 2241 / 1864.00
SampleSequence Count         Barcode
Ion1.1 2241   TACTCACGATA
Ion1.2 2130   TCGTGTCGCAC
Ion1.3 1221   TGATGATTGCC

Total number seqs written          5592
```

- The clean and trimmed sequence file was then passed to the pick_otus_through_otu_table.py script that performs the following analysis:
  - Sequences were clustered using the *uclust* method into Operational Taxonomic Units (OTUs) at 97% sequence similarity. Clusters at this level of similarity are typically acknowledged as having taxonomic relatedness at the species level;
  - Representative sequences from each OTU were then selected as the first sequence assigned to a cluster, which was the cluster seed defined by the *uclust* method;
  - Taxonomy was assigned to each representative sequence using the Ribosomal Database Project (*RDP*) classifier and the latest Greengenes database (a 16S rRNA gene database);
  - Alignment of the sequences was then performed using the *PyNAST* algorithm and a lanemask file was used to filter this alignment for completely gapped columns and highly variable locations;
  - From the aligned representative sequences, a phylogenetic tree was constructed using the *fasttree* method.
- From the aligned sequences, an alpha rarefaction curve and alpha diversity metrics (equitability, evenness and richness) were calculated using the Qiime script alpha_rarefaction.py.

Results

The sequencing has generated a significant number of reads after filtering and trimming of the sequences, as shown in Table 7. The time required to sequence the DNA and for post-processing of the sequences was **less than two hours**. A full breakdown of the time for sequencing and analysis is shown in Table 8.

Table 7 Sequence and OTU numbers per sample

	No. Sequences	No. OTUs
Sample 1	2241	824
Sample 2	2130	736
Sample 3	1221	514

Table 8 Sequencing and data analysis time requirements

Processing step	Time
Sample Preparation:	
Primer selection and optimisation of working conditions	4 days
Library preparation	2 days
Template preparation	7 hours
Sequencing:	
Cleaning & Initialisation of machine	2 hours
Manual preparation	0.5 hours
Sequencing run (per chip)	1.5 hours
Bioinformatics:	
Pre-processing (Ion Torrent software)	0.5 hours
Denoising (1 sample, 60000 reads)[§]	2 hours
Community analysis (QIIME)	0.5 hours

[§] Using 16-core processor (32 CPUs) with 32GB RAM

The output from the QIIME pipeline generated the relative abundance of the observed organisms present in the three samples. Table 9 shows the distribution of Phyla within the samples and indicates that the largest number of organisms are unclassifiable Bacteria (65%), with Firmicutes, Bacteroidetes and Euryarchaeota being subsequently the most abundant phyla present. Of all the phyla, only Chloroflexi and Planctomycetes are not present in all three samples, being absent from Sample 3.

Table 9 Percentage relative abundance of OTUs classified by Phylum

Taxon	Sample 1	Sample 2	Sample 3
Archaea;Crenarchaeota	0.044622936	0.046948357	0.081900082
Archaea;Euryarchaeota	**7.898259705**	**7.840375587**	**7.616707617**
Archaea;Other	0.089245872	0	0
Bacteria;Actinobacteria	0.490852298	0.845070423	1.064701065
Bacteria;Bacteroidetes	11.28960286	10.04694836	11.05651106
Bacteria;Chloroflexi	0.044622936	0.281690141	0
Bacteria;Firmicutes	13.0745203	14.03755869	14.82391482
Bacteria;Other	**65.28335564**	**65.68075117**	**64.04586405**
Bacteria;Planctomycetes	0.401606426	0.093896714	0
Bacteria;Proteobacteria	0.133868809	0.046948357	0.163800164
Bacteria;Spirochaetes	0.267737617	0.657276995	0.327600328
Bacteria;Synergistetes	0.089245872	0.046948357	0.081900082
Bacteria;Tenericutes	0.089245872	0.093896714	0.327600328
Unassignable;Other	0.223114681	0	0
Unclassified;Other	0.58009817	0.281690141	0.40950041

Table 10 shows the percentage relative abundance of OTUs that have been generally well classified (i.e. to the family or genus taxonomic level) or are of particular interest in the anaerobic environment. The most dominant organism within the group is *Methanoculleus*, which is known to be an indicator of methanogenic activity in anaerobic environments (Barret *et al.*, 2012). Of the other Archaea, *Methanocorpusculum* (Robb, 1995) and *Methanosarcina* (Galagan *et al.*, 2002) are known methanogens.

From the identified bacteria OTUs, Clostridia (sulphite-reducing bacteria) and Planctomycetacia are known anaerobic orders, Staphylococcaceae and Marinilabiaceae are gram-negative anaerobes, and Corynebacterium and Dietzia are gram-positive facultative anaerobes.

Table 10 Percentage relative abundance of OTUs classified by Genus

Taxon	Sample 1	Sample 2	Sample 3
Archaea;Crenarchaeota;Thermoprotei;Other;Other;Other	0.044622936	0.046948357	0.081900082
Archaea;Euryarchaeota;Methanomicrobia;Methanomicrobiales;Methanocorpusculaceae;Methanocorpusculum	0.490852298	0.234741784	0.163800164
Archaea;Euryarchaeota;Methanomicrobia;Methanomicrobiales;Methanomicrobiaceae;Methanoculleus	**3.971441321**	**3.990610329**	**4.914004914**
Archaea;Euryarchaeota;Methanomicrobia;Methanomicrobiales;Methanomicrobiaceae;Other	0.713966979	0.751173709	0.327600328
Archaea;Euryarchaeota;Methanomicrobia;Methanosarcinales;Methanosarcinaceae;Methanosarcina	0.713966979	1.079812207	0.655200655
Archaea;Euryarchaeota;Methanomicrobia;Methanosarcinales;Methanosarcinaceae;Other	0.044622936	0.046948357	0
Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Other;Other	0.267737617	0.14084507	0.081900082
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Corynebacteriaceae;Corynebacterium	0	0	0.081900082
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Dietziaceae;Dietzia	0	0	0.081900082
Bacteria;Actinobacteria;Actinobacteria;Coriobacteriales;Coriobacteriaceae;Other	0.178491745	0.093896714	0.081900082
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Marinilabiaceae;Other	0.044622936	0	0
Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Chryseobacterium	0.044622936	0	0
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Other	0	0	0.081900082
Bacteria;Firmicutes;Bacilli;Lactobacillales;Other;Other	0	0.234741784	0.163800164
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium	0.446229362	0.469483568	0.40950041
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Sarcina	0	0.14084507	0
Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Dorea	0.044622936	0	0
Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Other	0.401606426	0.093896714	0
Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Other;Other	0.044622936	0	0
Bacteria;Spirochaetes;Spirochaetes;Spirochaetales;Spirochaetaceae;Other	0.223114681	0.516431925	0.327600328
Bacteria;Spirochaetes;Spirochaetes;Spirochaetales;Spirochaetaceae;Treponema	0.044622936	0.14084507	0
Bacteria;Synergistetes;Synergistia;Synergistales;Synergistaceae;Other	0.089245872	0.046948357	0.081900082
Bacteria;Tenericutes;Mollicutes;Acholeplasmatales;Acholeplasmataceae;Acholeplasma	0.044622936	0.093896714	0.327600328

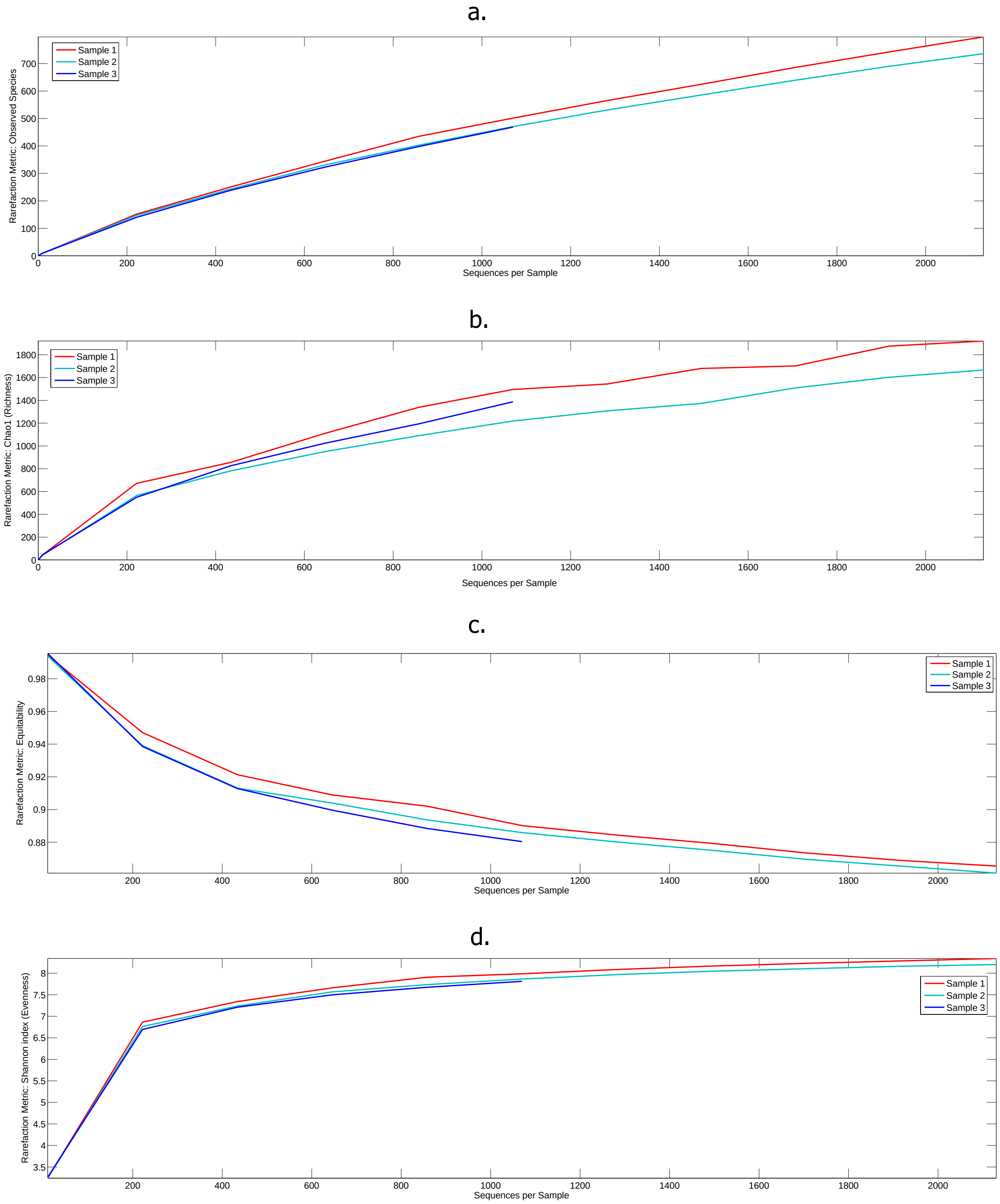
Two representative sequences were selected from the identified bacteria for a manual search using the BLAST nucleotide search algorithm with the 16S rRNA sequences (Bacteria and Archaea) database. The BLAST database allows identification at species level, but it is not as reliable as RDP for pairwise matching. The results are shown in Table 11.

Table 11 Manual BLAST search of 2 sequences isolated from representative sequence data

Family	**Species**	**Blast Score**	**Description**
Synergistaceae	*Aminobacterium colombiense*	158	Strictly anaerobic, amino acid-fermenting, Gram-negative bacterium isolated from an anaerobic lagoon of a dairy wastewater treatment plant. Optimal growth temperature is 37°C
Spirochaetaceae	*Treponema amylovorum* strain HA2P	152	Intermediate-sized, obligately anaerobic, helically coiled, mobile treponeme. Saccharolytic.

The within-sample diversity metrics generated by Qiime were used to develop rarefaction curves (Figure 6), using sub-sampling to generate values up to the maximum sequence length per sample. Figure 6a. shows the number of OTUs present with respect to the number of sequences sampled and is analogous to the information provided in Table 7. It indicates that with greater sampling of Sample 3, the total number of OTUs across all samples would be equivalent. The Chao1 metric shown in Figure 6b. is an estimate of the species richness in a sample, which does not account for abundance but merely their presence. Figure 6c. indicates the equitability or relative abundance, whilst Figure 6d. is the Shannon index that represents evenness of a sample, factoring in the absolute abundance of species. For example a high evenness will be observed if all OTUs in a sample have similar levels of abundance. The graphs show that, although Sample 1 is slightly more diverse and even than preceding samples, the diversity between samples is not significantly different; following the same general trend as sampling increases. This is indicative of a stable microbial community over the measurement time period.

Figure 6 Rarefaction curves for the three samples; a. Observed OTUs, b. Chao1, c. Equitability, d. Shannon Index



5.2 Comparison with 'business as usual'

There is currently no cited utilisation of next generation sequencing for gaining deeper understanding of anaerobic digestions systems at full-scale and subsequent optimisation of the process. The 'business as usual' case focuses on monitoring of process variables such as pH, gas yield, temperature that, at best, can offer some warning of process failure. However, the response to abnormal process variables is reactive and can lead to potential loss of reactor (biomass) viability and subsequent economic issues.

Next generation sequencing should not be considered as an on-line monitoring tool as its use during stable operating periods is limited and would incur unnecessary costs. Instead, the technology will be best utilised as a diagnosis method correlating process variation to biological community structure and activity. NGS can also be used to characterise the process at start-up or during changes in operating conditions or feedstock, to better understand process dynamics under critical operational phases.

5.3 Risk Analysis

Table 12 Risk analysis

Risk	Level	Mitigation
Sequencing errors result in poor datasets	Medium	Repeat sequence analysis with stored sample
Failure to find willing AD sites for phase 2	Medium	CEG have good connections with waste industry with sites having AD systems. Requests to sample from these sites will be submitted

6.0 Economic / Cost Benefit Analysis

6.1 Appraisal of Cost to industry/facilities including capex and opex

Prior to completion of Phase 2, it is difficult to account for the true cost of the proposed service for industry. It is proposed that the delivery of sequencing, personnel, data analysis and generation of a report and/or guidance documentation, will be factored into the service cost. The cost will depend on the agreement entered into between the service provider (Newcastle University) and the AD company, relating to:

- Number of samples to be collected and analysed.
- Contract duration.
- Knowledge delivery and assessment of impact.

Knowledge delivery assumes that the sequencing and analysis can assist plant operators to address the objectives laid out in the contract (e.g. identification of process problems, characterisation of good operating conditions) leading to effective changes to the plant that will yield process and economic benefits.

Indicative costs for a single Ion Torrent run (multiple samples can be analysed in one run) using the 314 chip (£66 per chip) is given in Table 12. A comparison with other NGS technologies is also provided, demonstrating the cost benefits of Ion Torrent:

Table 12 Sequencing costs for Ion Torrent

Platform	Cost	Number of reads	Length of reads	No. of bases	£ Cost per base	£ Cost per base	p Cost per base	£ Cost per sequence
454*	£6,000	1,000,000	400	4.00E+08	1.50E-05	0.00001500	0.0015000	0.0060000
Illumina*	£1,400	60,000,000	100	6.00E+09	2.33E-07	0.00000023	0.0000233	0.0000233
Ion Torrent (excl. Amplicon prep)	£780	1,000,000	100	1.00E+08	7.80E-06	0.00000780	0.0007797	0.0007797
Ion Torrent (excl. Personnel)	£417	1,000,000	100	1.00E+08	4.17E-06	0.00000417	0.0004167	0.0004167
Ion Torrent (incl. Amplicon prep)	£1,211	1,000,000	100	1.00E+08	1.21E-05	0.00001211	0.0012107	0.0012107
Ion Torrent (excl. Personnel)	£485	1,000,000	100	1.00E+08	4.85E-06	0.00000485	0.0004847	0.0004847
* from BBSRC Genome Analysis								
Depreciation costs over 10 yrs			100,000					
Number of years to obsoletion			10					
Annual depreciation			10000					
Monthly depreciation			833.33					
Annual total (20% per project of monthly)			**£167**					
Daily rate for personnel			**£363**					
Cost amplicons (10 samples)								
DNA extraction		£5 / sample	£50					
PCR (in triplicate w/adapter primer)		£1.80 / sample	£18					
Including agar gel electrophoresis								
Sub-Total			**£68**					
2 days for 1 run			£726					
Includes amplicon prep (10 samps)			£68					
Library prep + Sequencing			£250					
Total cost incl. depreciation			**£1,211**					
1 day for 1 run			**£363**					
As above excl. Amplicon prep			£250					
Total cost for preparation and sequencing per run			**£780**					

Additional costs for bioinformatics per run = £500
Total cost of sequencing and data analysis including personnel = **£1,711**

It is expected that partners can provide process measurements to accompany the biological samples (e.g. COD, VFA, methane/biogas composition). Additional cost may be incurred if these measurements have to be performed by a third-party service or contracted to Newcastle University. Discussions with the partners will be held prior to sampling on this issue.

7.0 Phase 2 demonstrations

This section of the report explains what will be achieved as part of the Phase 2 demonstration.

7.1 Aims and objectives

The aim of Phase 2 is to demonstrate the potential of high-throughput next generation sequencing technology coupled with bioinformatics analysis pipelines as a fast and economic indicator of anaerobic digestion plant health at a wider scale than the feasibility study. It is intended to perform this work in collaboration with a number of full-scale AD plants and show how the coupling of existing process and engineering knowledge with new biological community information can result in optimisation of plant performance.

The objectives of Phase 2 are:

- To demonstrate the sequencing technology on a full-scale anaerobic digester.
- To evaluate the data acquired through sequencing in relation to process performance over a significant timescale.
- To demonstrate the potential for plant optimisation through the understanding of microbial community function coupled with monitored process plant parameters (e.g. pH, temperature, gas yield).
- To assess the economic viability of the technology for full-scale AD systems in relation to process benefits.
- To assess and evaluate the commercialisation potential of implementing sequencer technology amongst the wider AD community.
- To assess the viability of implementing the technology on a site-by-site basis or to utilise a centralised sequencing service for processing multiple AD plant samples.

A major benefit of this work is that no disturbance to existing AD sites is required, such as building or implementation of equipment or instrumentation. All analysis will be performed off-site at Newcastle University and end-user sites will have only the obligation of collecting biological samples from the process.

It is intended to use the technology and analysis methods amongst a range of AD systems to demonstrate the potential of NGS for varying plant designs/configurations, operating conditions, feedstocks and digester functionality. As such, the expected demonstrations will vary in their objectives, tailored to suit the individual requirements of the AD plants. It is expected that, after consultation with the participatory stakeholders, the following objectives and benefits will be defined. A number of examples are shown in **Table 13**.

Table 13 Objectives and potential benefits of demonstration phase

Objective	Benefit
Basic characterisation of the AD biology (microbial community structure) over a period of digester operation	Understanding of microbial roles during stable operation. Database of "desired" microbial community
Identification of microbial dynamics during process changes (e.g. change in feedstock, process start-up)	Understanding of microbial sensitivity to process dynamics and identification of population changes
Identification of AD biology during process failures or sub-optimal performance	Ability to characterise community structure relating to process disturbances and to identify potential susceptible or responsible candidate organisms
Characterisation of relationship between	Better identification of relationship between

available process measurements and microbial population	process and biology of AD systems (for monitoring, control and optimisation)

7.2 Methodology for demonstration

Given the site-specific objectives, sampling protocols will be developed with the partners to define the location (e.g. primary digester, effluent), frequency and duration of the sampling regimes. Samples will be collected by the partners and sent to Newcastle University for sequencing. In parallel, available process measurements will be collected by the partners and provided to the University. These measurements are expected to include influent and effluent COD, VFA, biogas/methane production, temperature, pH, CO_2, and hydrogen production and H_2S if possible. It is estimated that about 100 biological samples will be acquired over the 10-month sampling period, but this figure might change given further discussions with the site partners.

Sample storage, preparation and sequencing will be performed in the same manner as Phase 1, with careful monitoring and evaluation of the sequencing runs to ensure successful generation of sequencing data. Data analysis will be performed using the Torrent Suite software for post-processing of the sequence data prior to quantitative and statistical analysis using QIIME. At present denoising of Ion Torrent sequencing data has not been validated, but this is expected to be achieved in the short-term with an adaptation to the AmpliconNoise software (Chris Quince, University of Glasgow) or via the new Acacia software from University of Queensland.

Given the outputs generated from sequencing and data analysis, Newcastle University will work with the partner companies to generate knowledge that demonstrates the importance of the information gained through the technology to enable better engineering decisions to be made for the design and operation of AD systems. This will be achieved by a combination of scientific understanding and knowledge transfer, whereby candidate scenarios are presented to address the objectives defined at the outset of the work. In some cases, the information gained by sequencing may not contribute significantly to enabling system improvements. This may be due to the limitations of NGS or the inability to draw significant conclusions from the biological and process observations from the given sample datasets. In these cases, the difficulties will be documented and proposed methodologies investigated where possible (e.g. greater sampling coverage/frequency, support from other microbial techniques).

In the cases where the knowledge does contribute to addressing the site objective, a report will be generated detailing the steps performed, the results achieved and a cost-benefit analysis of the work undertaken.

All data, information and reports will be archived appropriately to develop an initial knowledge-base on NGS for anaerobic digestion plants in the waste industry.

7.3 Include any statements confirming/evidencing the selection and securing of sites, stakeholders, personnel and contractors.

As part of Phase 1, Newcastle University, with the support of WRAP, endeavoured to engage with industrial partners from the Waste sector, with a view to acquiring stakeholder AD facilities for performing the Phase 2 demonstration work. After presenting an overview of Next Generation Sequencing for the AD industry at the Spring AD Research & Development Forum organised by FERA in April, 2012, Newcastle University have received interest from a number of organisations in participating in Phase 2, as listed in Table 14. It is anticipated that a number of other companies will partake in Phase 2, including CPI in Teeside (with a provisional offer of an AD test facility), but discussions are on-going to formalise these commitments.

Table 14 Companies with a stated interest in Phase 2 participation

Name	Capacity (kWe)	Feedstock	Supplier	Location	Contact	Start	Cost	Input (t/y)	Output
Oxford Renewable Energy Ltd	2100	Cattle slurry & food waste	Agrivert, Biogas Weser-Ems GmbH & Kirk Environmental	Worton Farm, Cassington, Oxon, OX29 4EB.	Karen Moutos	2010	£7.5 million	45,000	Combined heat and power
GWE Biogas	2000	Commercial & industrial food and waste	NES GmBH	Driffield, East Yorkshire, YO25 9DR.	Tom Megginson	2010	£7.5 million	50,000	Combined heat and power
Lower Reule Bioenergy	1300	Manure, crops &food waste	Weltec BioPower GmbH	Gnossall, Staffordshire, ST20 0EA.	Matt Dove	2010	£2.6 million	30,000	Combined heat and power
South Shropshire Biodigester	200	Source-separated food waste	BiogenGreenfinch	Ludlow, Shropshire, SY8 1XE	Becky Arnold	2006	-	5,000	Combined heat and power

8.0 Complete and detailed project timescale

8.1 Project development and implementation plan

Table 15 shows the proposed timescale for the demonstration phase. Month 1 will focus on finalising demonstration AD sites with the project participants and developing sampling plans in accordance with the objectives for individual sites. Site visits will be completed by Month 2 prior to commencing the sampling work, which is expected to last until Month 11. For some sites, sampling will periodically take place over the full 10 month duration, whilst for other sites, sampling may be limited to discrete times within this period.

Sample sequencing will begin in Month 4 and end in Month 11 with a frequency determined by the sampling objectives. In many cases, samples will be pooled together and sequenced on a single chip, using barcodes to discriminate between the individual samples. If more extensive coverage of a sample is required, the number of samples per chip can be reduced.

Data analysis will be performed immediately after sequencing is complete. In the case that a comparison of sites or samples collected at different times is required, bioinformatics may be delayed until all candidate samples have been processed.

In parallel with the bioinformatics work and based on the biological information acquired through sequencing and, where available, process measurements, Newcastle University will work with the participant organisations to develop knowledge. Based on the objectives of the site-specific work an industrial report will be submitted to the stakeholders which will include the findings from this exercise and any recommendations A cost-benefit analysis will also be carried out in the final two months to determine the effectiveness of these recommendations in addressing the challenges or requirements of the end-users.

An assessment of the commercialisation options will be performed from Month 10 to Month 12, to determine the viability of the service in the AD sector, given the demonstration work.

A 6-month interim report will be submitted in Month 7 to WRAP followed by industrial reports in Months 8 and 12, and a final report in Month 12.

Table 15 Project workplan

	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
1. Finalise AD sites for sampling	🟥											
2. Development of sampling protocol	🟥											
3. Site visits	🟥	🟥										
4. Site sampling		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
5. Sequencing of samples				🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	
6. Data analysis (Bioinformatics)					🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
7. Knowledge transfer to end-user								🟦	🟦	🟦	🟦	🟦
8. Service assessment										🟧	🟧	🟧
9. Cost-benefit Analysis											🟧	🟧
10. Reporting							🟧	🟪				🟧

8.2 Operation

Operation of the Ion Torrent Sequencer will be performed by trained staff at Newcastle University and commence in Month 4. Samples will be stored at -20°C until DNA extraction is performed, to preserve sample integrity. Bioinformatics analysis will begin in Month 5 and be performed after each sequencing run, and data can be analysed separately or together, depending on the process requirements.

9.0 Commercialisation of technology

In addressing this part of the report it is appropriate to first clarify the nature of the technology that is being addressed both from a perspective on IP and also on commercialisation. Commercialisation of technology in the usual context is not directly applicable in this particular project.

An important point to clarify is that what is being developed is not NGS technology itself but the knowledge stemming from its application in an environment with which it would not normally be associated – in effect know-how and expertise that would not be patentable IP.

As part of the investigation of routes to commercialisation, it is appropriate to consider how the outputs from the exercise, such as knowledge dissemination, will benefit the end-users and the supply chain. In the case of the application of technology in this project, the real benefit to the end-users is ultimately the know–how or knowledge-base facilitated by information generated through the use of NGS. We consider here the relevant end-users to be both operators and process designers as the latter have the potential to develop a much deeper understanding of their particular processes that may facilitate technical improvements in plant/process development. In the first instance it is expected that the results will be more applicable to the smaller AD plants rather than the larger utility AD plants.

To achieve the greatest impact in moving forward it is therefore proposed that a number of measures are undertaken.

In the event of a successful move to phase 2 of the project the opportunity will be taken with interested partners to publicise the project and its aims to develop initial awareness of the development activity.

As the project develops further, detailed evidence of the application of the technology and its potential process benefits to AD operators (through engagement with operators and process designers in early trials) can enable further awareness to be developed through case study publication.

Ultimately the know-how needs to develop to a point where sufficient knowledge and results are available to:

a) Demonstrate a realisable benefit to be gained from the more detailed insight into the AD process at a microbial population level and what the methodology would need to be to achieve it (e.g. frequency of measurement etc.);
b) Use this information to demonstrate the value to the end user through cost benefit analysis for individual plants (or potentially plant design). On this particular point it cannot be discounted that knowledge generated in the course of the project may also be useful on a more generic basis.

From the basis of the latter and based on the assumption of a successful outcome the University would be in a position to offer more detailed training in the application and understanding of the methodology to AD end-users and other interested parties.

Throughout the awareness programme the University would be looking to engage with appropriate organisations and representative groups to assist in dissemination activities.

From a commercial viewpoint, process knowledge (in both engineering and biological terms) derived from the interpretation of results generated from NGS and the bioinformatics analysis can be provided on a commercial basis, whilst demonstrating real commercial worth to the industry. Phase 2 aims to demonstrate the industrial benefits and develop a commercial model for delivering these knowledge and skills to the industrial partners.

10.0 Key personnel

Dr. Matthew Wade is a Research Associate and Bioinformatician in the School of Civil Engineering and Geosciences at Newcastle University with an MSc in Environmental Engineering and a PhD in Electrical & Electronic Engineering. His background includes work in chemical and process engineering, wastewater treatment systems, artificial intelligence, process monitoring and control, knowledge discovery and data mining. He was the scientific manager for the European FP6 collective project *Agrobiogas* (EC/SES6-CT-030348-2; 2.9M€), aimed at increasing the efficiency of AD plants utilising co-digestion of agricultural wastes. He was also responsible for project coordinating 5 other EC funded projects and has successfully acquired funding for two FP7 projects on the topics of sustainable manufacturing to control and optimisation in the anaerobic digestion process. More recently, he has been active in the bioinformatics field, developing solutions for handling large sequencing datasets and how these tools can help both research and industry in rapidly identifying microbial communities and their influence on engineered systems, particularly in the water and AD sectors.
Role: Project leader. Responsible for sample collection, bioinformatics analysis, managing delivery of knowlege.

Prof. Tom Curtis is a Professor of Environmental Microbiology in the School of Civil Engineering and Geosciences at Newcastle University with a MEng and PhD in Public Health Engineering. He has worked in construction, research and public policy. He has undertaken research into domestic wastewater treatment in the tropics (including anaerobic treatment) and temperate regions. Latterly he has worked with a variety of excellent colleagues to not only apply molecular tools to wastewater treatment to not only give insights and guidance about the operation of existing systems but to investigate and where possible determine and apply new and hopefully universal theoretical descriptions for open engineered biological systems. He has taken a particularly strong interest in the interrelated questions of microbial diversity and how microbial communities form and change. He has been supported in these endeavours by a number of research grants from industry, government and the European Union, most recently a Marie Curie Excellence grant (PI Ian Head) and most recently as a renewed EPSRC Platform grant. He recently completed a three year BBSRC research development fellowship and a RAEng Global research award. Most recently, he has been awarded an ESPRC Dream Fellowship for his research project entitled "Simple Rules for Complex Systems: A Shortcut on the Path to the "In Silico" Sewage Works. He is s a member of the editorial board of Water Research, Applied and Environmental Microbiology, and the International Association of Microbial Ecology Journal.
Role: Advisory – microbial ecology at the interface of engineering systems.

Dr. Jan Dolfing is a CeG Research fellow in the School of Civil Engineering and Geosciences at Newcastle University, with an MSc in Environmental Engineering and a PhD in Microbiology from Wageningen University in the Netherlands. The recurring theme in his work is the use of thermodynamics to determine and study the limits and logic of anaerobic microbial processes. He has more than 30 years experience in anaerobic microbiology, with a special interest in anaerobic bioreactors, syntrophic interactions and the degradation of

halogenated compounds. He was the first to study the microbiology of granular methanogenic sludge, constructed the first defined dechlorinating consortium, was the first to show that micro-organisms can obtain energy from catalyzing reductive dechlorination of halogenated compounds, and was the first to isolate a bacterium that can degrade aromatic compounds in the absence of molecular oxygen. His work is widely cited and respected: seven of his papers have been cited 100+ times. His current EPSRC funded work involves the temperature limits of methanogenic wastewater treatment and the development of energy efficient wastewater treatment systems for the personal care industry (with L'Oreal, France). He is a past member of editorial board of Applied and Environmental Microbiology and of Microbial Ecology, has reviewed manuscripts for a host of other journals including Science, Water Research, and Environmental Science &Technology, and has evaluated research proposals for the EU in Brussels. He has authored seven invited book chapters and 65 papers in refereed international journals with first authorship of 58 % of these.
Role: Advisory – microbial ecology relating to the anaerobic digestion process, development of knowledge based on NGS results.

Dr. Russell Davenport is an RCUK Academic Fellow in Environmental Engineering at Newcastle University. After graduating in microbiology in Leeds, he worked on the design, operation and technical monitoring of anaerobic wastewater treatment plants with Yorkshire Water. He gained an MSc in Industrial Biotechnology and PhD in Environmental Engineering at Newcastle University. His area of expertise is the development and application of molecular tools to understand relationships between microorganisms and the processes that they mediate in natural and especially engineered environments. Much of his work has been in the quantification of specific functional groups and the key processes they mediate in biological treatment systems, and more recently, in the development and validation of methods to drastically improve diversity estimates and taxa assignments of data from next-generation sequencing technologies. Some of this research has been highly cited (Quince *et al.,* 2009, Davenport *et al.,* 2000). He is Principal Investigator on grants worth £1.78M, including a prestigious EPSRC Challenging Engineering award (EP/I025782/1), and has been a Co-Investigator/Co-I Researcher on grants worth > £3M including an EPSRC grant to determine the true temperature limits of anaerobic treatment of wastewaters using cold-adapted microbial communities (EP/G032033/1).
Role: Advisory – molecular tools and relating microbial ecology to engineering principles, key to generating knowledge for end-user requirements.

Dr. Paola Meynet is a Senior Research Associate at Newcastle University. After gaining her MSc in Analytical Chemistry in the University of Turin (Italy), she has undertaken a PhD in photo(electro)chemical disinfection of water and wastewaters in Chemical Engineering at Newcastle University. During her postdoctoral experience, she has successfully developed skills in microbiology and molecular biology to complement her training as a chemist. Her work has been focused on investigating microbial communities in wastewater systems and in contaminated soils undergoing remediation, using quantitative and qualitative molecular microbial ecology tools. As part of her industrial experience, she has been responsible for in situ pilot-scale testing of anaerobic treatment systems for highly concentrated industrial wastewaters, and she has investigated the commercialisation of a quantification assay of dioxygenases genes for the routine monitoring of polyaromatic hydrocarbons (PAHs) biodegradation in contaminated soil sites. Her research has been supported by EPSRC, Marie Curie and KTS funding. She is currently in charge of the installation, operation and protocols development of the Ion Torrent Personal Genome Machine (PGM^{TM}) sequencer for processing of environmental samples at Newcastle University.
Role: Laboratory sequencing and analysis of the anaerobic digestion samples.

11.0 Evaluation and monitoring for the purpose of WRAP reporting

Rigorous evaluation and monitoring will be in place throughout the demonstration phase. All work will be performed in strict agreement with the end-user sites, in which a preliminary document will be developed setting out the aims, objectives and strategy for each site. All sequencing and data analysis steps will be performed in accordance with defined protocols and outcomes monitored for quality. In the case of poor sequencing performance, identification of the problem will be made and a repeat run will be carried out.

An interim report will be produced after 6 months (to be delivered in Month 7) to be provided to WRAP, as well as end-user reports to be submitted based on work carried out on samples collected from AD sites. These will be collated as industry reports to be available in Month 8 and 12.

12.0 Health and Safety

Health and safety related to handling of biological materials, chemical and laboratory practice are covered by the Newcastle University Environmental Engineering Laboratory safety policy (2010), which includes hazard and risk assessments, laboratory rules, vaccination, first-aid, management of safety and out-of-hours working. All staff working in the laboratory are obliged to adhere to this policy and complete any relevant training and documentation.

13.0 Conclusion

Next Generation Sequencing is an emerging technology that has the potential to completely change the way that the anaerobic digestion industry interprets the biology of their systems. As the cost of sequencing is ever decreasing and the speed and performance increases, the potential for utilising NGS in engineered biological systems starts to become a reality. Phase 1 of the project has assessed the feasibility of NGS and the associated bioinformatics data analysis tools for characterising the microbial community of an on-farm biogas plant. The results show that the system was stable over the period of sampling and that the cost and processing time was comparatively lower than other NGS technologies. Phase 2 will broaden the scope of the work and aims to demonstrate the potential of NGS on a variety of AD systems.

14.0 References

Barret, M., Gagnon, N., Morissette, B., Topp, E., Kalmokoff, M., Brooks, S.P.J., Matias, F., Massé, D.I., Masse, L. and Talbot, G. (2012). *Methanoculleus* spp. as a biomarker of methanogenic activity in swine manure storage tanks. *FEMS Microbiology Ecology*, **80**: 427–440.

Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.,G., MacDonald, P., Fitzhugh, W., Calvo, S., Engels, R. et al. (2002). The Genome of *M. Acetivorans* Reveals Extensive Metabolic and Physiological Diversity. *Genome Research*, **12**(4): 532–542.

Kröber, M., Bekel, T., Diaz, N.N., Goesmann, A. and Sebastian, J. (2009). Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *J. Biotech.*, **142**: 38–49.

Robb, F.T. (1995). Archaea: A Laboratory Manual. Cold Spring Harbor Laboratory Press, U.S.

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., Szczepanpwski, R., Tauch, A., Tilker, A., Viehöver, P. and Goessmann, A. (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analyzed by the 454-pyrosequencing technology. *J. Biotech.*, **136**: 77–90.

Sundberg, C., Abu Al-Soud, W., Larsson, M., Svennson, B., Sörensson, S. and Karlsson, A. (2011). 454-pyrosequencing analyses of bacterial and metagenomic Archaea DNA and RNA diversity in 20 full-scale biogas digesters. In *Proceedings of the First International Conference on Biogas Microbiology*. 14–16 September, Leipzig. UFZ Press, Edited by Kleinsteuber, S., Nikolausz, M., Leipzig: 2011: 47.

Wirth, R., Kovács, E., Maróti, G., Bagi, Z., Rákhely, G. and Kovács, K. (2012). Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnology for Biofuels*, **5**: 41.

Waste & Resources
Action Programme

The Old Academy
21 Horse Fair
Banbury, Oxon
OX16 0AH

Tel: 01295 819 900
Fax: 01295 819 911
E-mail: info@wrap.org.uk

Helpline freephone
0808 100 2040

www.wrap.org.uk/diad

recycle