The role of molecular biology in geotechnical engineering

Le rôle de la biologie moléculaire en géotechnique

Stewart D.I. , Fuller S.J.
School of Civil Engineering, University of Leeds, Leeds, UK

Burke I.T., Whittleston R.A., Lockwood C.L.
School of Earth and Environment, University of Leeds, Leeds, UK

Baker A.
School of Biology, University of Leeds, Leeds, UK.

ABSTRACT: This paper reviews techniques from molecular biology for characterising microbial populations that are accessible to Geotechnical or Geo-Environmental Engineers. With reference to data from contaminated land studies, it discusses which techniques it might be appropriate to use in an engineering context, how the data generated can be visualised and interpreted, and the dangers of over interpretation. Finally it reports on the capabilities of the latest high throughput next-generation sequencing platforms, and speculates on what engineering developments may result from this technological advance.

RÉSUMÉ: Ce document passe en revue les techniques de la biologie moléculaire pour la caractérisation des populations microbiennes qui sont accessibles aux ingénieurs géotechniques ou géo-environnementaux. En faisant référence aux données provenant d'études de terres contaminées, il aborde les techniques qu'il pourrait être approprié d'utiliser dans un contexte d'ingénierie, la manière dont les données générées peuvent être visualisées et interprétées, et les dangers d'une sur-interprétation. Enfin, il rend compte de la capacité des plateformes les plus récentes de séquençage à haut débit de prochaine génération, et s'interroge les développements techniques qu'il pourrait résulter de cette avancée technologique.

KEYWORDS: Geo-environment, Molecular Biology, DNA, rRNA, 16S gene sequencing

1 INTRODUCTION

In recent years Geotechnical and Geo-Environmental Engineers have started to exploit soil microorganisms, nature's catalysts, to deliver sustainable engineering solutions to big problems facing society. Such microorganisms obtain energy from catalysing thermodynamically favourable chemical reactions between natural soil constituents, but in the process can also catalyse chemical reactions that are of engineering interest. To date approaches such as monitored natural attenuation and active bioremediation have become well-established for the treatment of soils contaminated with petroleum hydrocarbons and organic solvents. However this field is about to expand rapidly, with techniques such as the reductive precipitation of contaminant metals and radionuclides, microbial induced calcite precipitation to improve soil strength, bacterially mediated phosphate recovery from waste streams and bacterially enhanced carbon capture likely to emerge from a research setting and into engineering practice in the near future.

What all these applications have in common is that they involve managing populations of microorganisms to bring about chemical transformations within an engineering context. Thus, if engineers are to manage these populations effectively, they need to characterise microbial populations to identify whether the necessary organisms are present, or better still to determine the genetic potential of the population to perform particular chemical transformations. In the near future engineers seeking better process control might wish to identify which metabolic pathways are active under particular conditions in order to predict which chemical transformations are about to occur.

To be able to quantify the contributions of microorganisms to a process, and ultimately to control that contribution, it is necessary to first know what organisms are present, secondly how this population changes with the conditions, and thirdly which organisms and conditions are the most important for achieving the desired outcome. Traditional microbiology methods involve culturing, identifying and enumerating the microorganisms present. However these suffer from a number of disadvantages; not all microorganisms can be cultured, the culture conditions selected can favour some species over others, and identification requires a high level of expertise in microbial taxonomy. In contrast methods based on nucleic acids, DNA and RNA the genetic material of all organisms, have become quick, simple and relatively cheap. A modest investment of a few thousand pounds can equip a laboratory for such analyses.

With the exception of some viruses the genome of all organisms is made up of DNA, Watson and Crick's famous double helix, in which two strands that run anti-parallel to one another are held together by H-bonds between complementary bases; A (adenine) always bonds with T (thymine), and G (guanine) with C (cytosine). This complementary base pairing allows each strand to provide the information for synthesis of its complementary strand during DNA replication. In the cell this process is carried out by enzymes called polymerases using building blocks called deoxynucleotide triphosphates (dNTPs) and is essential for cells to replicate. The DNA contains all the genes necessary to specify all structures and functions of the cell. As some processes (such as synthesising proteins) are fundamental to all cells, some genes are very similar in all organisms. Others play much more specialised roles and their presence can be used to infer the presence of specific organisms (see section 4). Fundamental to all the methods to be discussed is the ability to amplify and determine the sequence of specific sections of DNA from environmental samples. This allows inferences to be drawn about the presence or absence of organisms or to gain insights into the populations present and their dynamics.

2 THE POLYMERASE CHAIN REACTION (PCR)

The polymerase chain reaction (PCR) is a technique for replicating a selected section of a DNA fragment. It starts with one or two copies of the target section, and increases that by several orders of magnitude. PCR involves repeatedly heating and cooling the DNA using a piece of equipment known as a thermocycler. There are usually three discrete temperature

steps. The first step is denaturation, which involves the highest temperature in the cycle (typically 94-95°C; Promega, 2012; Roche, 2011a). This separates the strands of double stranded DNA to act a template for DNA synthesis. The second step is annealing, which involves the lowest temperature in the cycle. In this step PCR primers become attached to the template DNA. PCR primers are short fragments of DNA which are designed and synthesised to match to the ends of a target section of DNA, and serve as a starting point for DNA replication. The annealing temperature depends on the properties of the primers being used, but is usually 42–65°C (Brown, 2001). The third step is extension where double stranded DNA is reconstructed base-pair by base-pair from dNTPs in the reaction mixture by the polymerase enzyme acting at the 3' end of the annealed primer (typical temperature 68-72°C). This three step cycle is repeated many times, with the amount of the target DNA fragment doubling (in theory) in each cycle. In practice there is initially exponential amplification, but it levels off with increasing numbers of cycles as the polymerase enzyme loses activity and the reagents (dNTPs and primers) are consumed until, eventually, no further product is produced. If the primers have been appropriately designed and the reaction conditions optimised only the target DNA fragment should be amplified.

The very high amplifications achieved by repeated cycling make PCR a very powerful technique, but users need to be aware of potential artefacts that can arise. The three main ones are contamination, polymerase errors and bias. The highly sensitive nature of PCR means that even low levels of contaminating DNA (from other samples or the laboratory environment) can lead to amplification of products that don't originate from the sample. Thus scrupulous cleanliness and the use of negative controls (where no DNA sample is added to the reaction) are mandatory. Small errors and slight biases in a single amplification cycle can over the course of many cycles lead to gross distortions in the representation of different fragments in the final PCR product. Thus, as a rule, data from an analysis involving a PCR reaction should be treated as qualitative rather than quantitative (the exception being where the more advanced tool of qPCR is used). PCR errors can arise due DNA polymerase errors (the Taq polymerase error rate is $\sim 3 \times 10^{-5}$ per nucleotide per duplication; Acinas et al. 2004), the formation of chimeric molecules, and the formation of heteroduplex molecules. The best way to avoid this type of problem is to avoid unnecessary over-amplification because such errors are cumulative (i.e. use the smallest possible number of PCR amplification cycles compatible with the intended application; Qiu et al. 2001; Acinas et al. 2005). For some purposes it may also be necessary to use a "proof-reading" polymerase enzyme with a far lower intrinsic error rate than Taq. PCR bias arises because of intrinsic differences in the amplification efficiency of different templates (e.g. due to differences in the GC-content). In late stages of amplification self-annealing of the most abundant templates can hinder their further amplification. PCR bias is reduced by using high template concentrations, performing fewer PCR cycles (Polz & Cavanaugh 1998) and by using a thermocycler that ramps quickly between the cycling temperatures (Acinas et al. 2005).

In Geo-Environmental Engineering PCR is most frequently used to characterise microorganisms or microbial populations. PCR is often used to amplify the same section of the same gene of different species of microorganisms in an environmental sample. The 16S rRNA gene is frequently used to identify microbes, and to study diversity, because it is present in all prokaryotic organisms (those without a nucleus, like bacteria). In this case the primers are targeted at two conserved regions of the gene where the base sequence is almost identical between widely different organisms because that part of the molecule encodes a function necessary for life so that the intervening region, which is more divergent, is amplified. The difference in sequence between the divergent regions can be used to infer evolutionary distance. PCR is also used to amplify sections of DNA between genes since these are often very variable in length and/or base sequence even for quite closely related species. An example of this approach is rRNA Intergenic Spacer analysis (RISA), which is often used for community fingerprinting with the aim of identifying when there has been a significant change in the microbial population.

3 CHARACTERISING MICROBIAL POPULATIONS

Simple PCR based techniques can identify that a particular gene is present within a microbial population, and to "fingerprint" bacterial populations to identify significant changes in population over time or under different conditions (section 4), but actually identifying the bacterial species present in a sample requires some method of separating out the individual DNA fragments in an environmental sample, and sequencing them.

The traditional approach to this problem is "cloning and sequencing" (see e.g. Islam et al. 2004). This approach starts with a PCR reaction on environmental DNA using broad specificity primers that target a suitable gene (usually the 16S rRNA gene). The resulting PCR product contains multiple copies of the target gene from all the species in the sample. These double stranded DNA fragments are then ligated to (joined-into) a standard cloning vector (e.g. pGEM-T, TOPO, etc.) to form circular double-stranded DNA molecules called plasmids. This is achieved with standard molecular biology kits available from suppliers such as Promega or Life Technologies. Each plasmid will contain a different DNA fragment from the PCR reaction. The plasmid is then "transformed" (inserted) into specially weakened laboratory strains of E-coli that lack resistance to antibiotics. The standard cloning vectors are designed to confer antibiotic resistance to any cell into which they are inserted, allowing selection of those cells that take up a plasmid. An important feature of the transformation is that it is very inefficient (which is why selection is necessary), and thus it is unlikely that more than one plasmid is inserted into a cell. The cells are then plated out on agar plates containing the antibiotic, so that only cells containing the plasmid grow. If this is done with care then bacterial colonies will grow on the plates so that each colony has grown from a single cell, and will thus contain copies of a single DNA fragment from the environmental sample. These cells can be harvested, and the plasmid they contain sent for gene sequencing. Figure 1 shows the bacterial population of soil from near a lime kiln waste tip determined by this technique (Burke et al. 2012). This showed that the bacterial population of the sample of buried soil was dominated by a single, unidentified species within the Comamonadaceae family of β-proteobacteria. Determination of the geochemical conditions allowed this study to postulate a link between anaerobic respiration of this specie and the

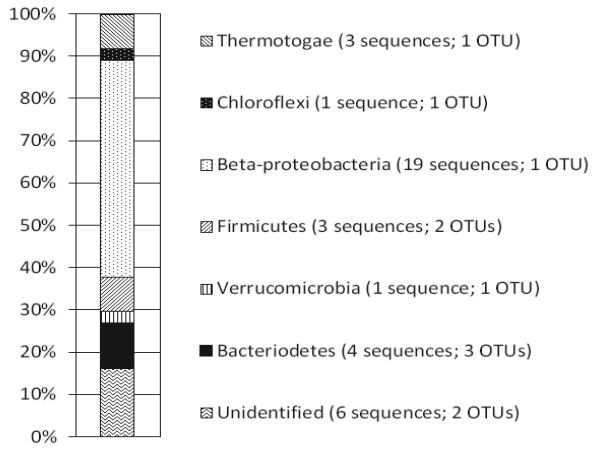

Figure 1. Phylogenetic diversity of 16S rRNA gene sequences extracted from sample. Key shows the number of OTUs within each phylum

reduction of nitrate in the groundwater.

DNA sequencing is relatively expensive, and therefore cheaper techniques for monitoring microbial populations are useful in order to decide whether more detailed analysis is required. There are several ways of "fingerprinting" a microbial population, the most useful of which, due to its simplicity of use and low cost, is probably RISA (Borneman, 1997). RISA exploits differences in the length of bacterial DNA between two genes: the 16S and 23S rRNA genes. This varies between 150 and 1500bp. The analysis uses a PCR that amplifies the intergenic spacer region using primers that target conserved sections of DNA (within the 16S and 23S genes) that flank the region (Cardinale, 2004). The PCR product is then size separated using agarose gel electrophoresis; the patterns in the bands that are visible on the gel image are a fingerprint for the bacterial population. Figure 2 shows a RISA gel image from a study of a soil/groundwater system investigating the effect of different amendments (bicarbonate and acetate) on the microbial population. The gel image shows that microbial populations were significantly different 175 days after amendment.

High-throughput sequencing (or next-generation sequencing) technologies read many thousands of sequences in parallel. There are a variety of "platforms" for high-throughput sequencing (see Metzker 2010 for a review), but for brevity this paper will focus on 454 pyrosequencing as, possibly, the most straight-forward approach to creating a 16S rRNA gene library. It employs an initial PCR reaction on a DNA sample to isolate the gene fragment of interest and attach "adapter" sequences to both ends of that fragment. Unique identifier codes can also be incorporated between the adaptor and the gene fragment at this stage so that several samples can be sequenced at same time and separated during subsequent analysis (potentially offering a cost saving). During pyrosequencing fragments of the template DNA are isolated by attaching them to microscopic DNA capture beads using the adaptor. These beads are suspended inside water droplets in an oil solution in separate picoliter-volume wells on a multi-well plate. A PCR reaction using a luciferase then generates a light signal from each well as individual nucleotides are added to a DNA strand, which can be read in parallel.

The Roche GS FLX Titanium system can sequence fragments of up to 600bp, including the adaptor sequences (Roche, 2011b), which is why it is suited to 16S rRNA library construction. However, because it requires an initial PCR to attach the adaptors, it does not escape from problems associated with PCR errors and bias, although a proof-reading polymerase and a low number of PCR cycles will minimise these effects. Other high-throughput sequencing approaches can directly sequence environmental samples without a PCR reaction to attach an adaptor sequence. These currently yield shorter read lengths than the approach described above, and the post sequencing analysis to identify the gene of interest is more complex, but it should be noted that this is a particularly fast moving area of scientific development, and direct sequencing of environmental samples may become the norm in the near future.

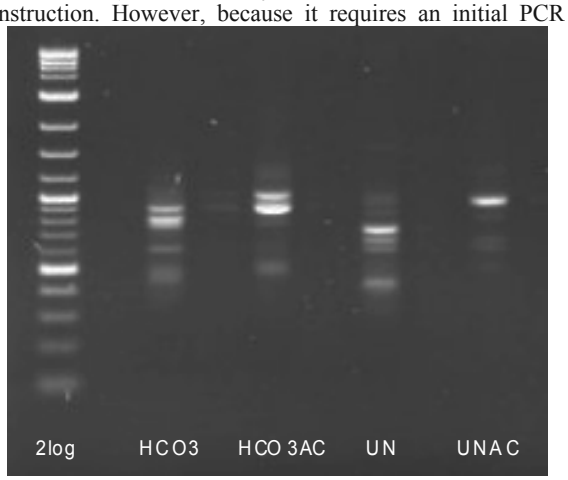

Figure 2. RISA signatures for soil-groundwater incubations with different amendments on day 175. HCO3 → HCO_3^- amended; HCO3AC → HCO_3^- & acetate amended; UN → unamended; UNAC → acetate amended, 2log → NEB 2-log DNA ladder

4 PCR TECHNIQUES FOR IDENTIFYING THE PRESENCE OF A MICROORGANISM

PCR is also used to identify the presence of a particular gene within a bacterial population. An example of this approach is to use a PCR reaction using primers that target the invA gene to identify the presence of Salmonella in a sample as this gene has very high specificity to Salmonella strains (Sunar et al. 2009). Figure 3 shows a gel image of a product from a PCR targeting invA gene of Salmonella (product at 285 bp). In this experiment DNA from an environmental sample was mixed with increasing concentrations of a competitor fragment (length 183bp), so that the number of gene copies could be estimated (so called competitive PCR).

5 IDENTIFYING THAT A SPECIFIC GENE IS BEING EXPRESSED UNDER THE PREVAILING CONDITIONS

DNA contains the genes of an organism but for these genes to perform their actions in the cell they must first be copied into RNA which is then usually converted into proteins. Proteins may be structural, or carry out chemical reactions. This process of copying genes into RNA is called transcription and when a gene is transcribed it is said to be expressed. Some genes are expressed under all or almost all conditions others are only expressed in specific situations, for example in the presence of a particular electron donor or acceptor. To monitor gene expression RT-PCR is commonly used. RT stands for reverse transcription and describes the process of copying RNA into DNA. In most cells information flows one way from DNA to RNA to protein (the central dogma) but enzymes called reverse transcriptases, isolated from some viruses, can copy RNA into DNA. PCR only works on DNA so to amplify sequences derived from RNA isolated from a sample an RT step has to be carried out first. Then PCR is carried out using primers for the gene whose expression is to be tested. This information will generally be qualitative unless conditions are carefully optimised to obtain quantitative data. Quantitative or qPCR generally monitors amplification in real time to ensure that the level of product in different reactions are compared within the exponential phase of the cycle, and by comparison to known standards either a relative or absolute number of gene copies can be calculated. qPCR allows monitoring of 10s of genes but high throughput technologies (RNA-seq: Wang et al. 2009) now allow analysis of all the genes being expressed by an organism at a given time (the transcriptome). However the use of this for environmental samples 'metatranscriptomics' is in its infancy (e.g. Marchetti et al. 2012).

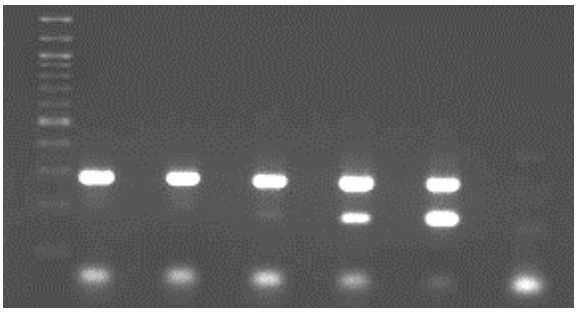

Figure 3: Agarose-TBE gel image showing the presence of the Salmonella invA gene (product at 285 bp). Samples contained an increasing concentration of a competitor fragment (product at 183bp)

6 GENE SEQUENCE DATA ANALYSIS

Sequence data is usually provided in a text file in FASTA format, where there a description line and then the sequence of nucleotides reported as single-letter codes (A,G,C,T). In a Geoenvironmental context, the purpose of sequencing a gene is usually to identify the species from which the sequence came. This is done by comparison with open-access databases such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/), the EMBL nucleotide sequence database (http://www.ebi.ac.uk/embl/), or the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/). These databases are maintained by public bodies in the USA, Europe and Japan collaborating as the International Nucleotide Sequence Database Collaboration (http://www.insdc.org/). Sequences obtained from samples can be compared with sequences in the database using a variety of free, public domain software. BLAST (Basic Local Alignment Search Tool) makes pair-wise comparisons with sequences in the chosen database and reports the statistically most significant matches. SEQMATCH available from the Ribosomal Database Project (http://rdp.cme.msu.edu/index.jsp) performs a similar function, and readily allows the user to restrict the quality of sequences to which matches are reported (e.g. type species, isolates, long read lengths, "good" quality).

CLASSIFIER, which is also available from the Ribosomal Database Project, is a naïve Bayesian Classifier that can place bacterial 16S rRNA sequences within Bergey's Taxonomic Outline of the Prokaryotes (Wang et al. 2007). It is easy to use, and can be used for classifying single rRNA gene sequences or for the analysis of libraries of thousands of sequences.

For some types of analysis it may be necessary to align sequences from the same gene of different species prior to detailed analysis. An alignment is a way of arranging gene sequences to identify regions of similarity that indicate functional, structural, or evolutionary relationships between the sequences (Mount, 2004). There is a variety of open-access software available for aligning gene sequences, two of the more popular of which are ClustalW (Cluster Analysis) and MUSCLE (MUltiple Sequence Comparison by Log-Expectation) both of which are available from the European Bioinformatics Institute website (amongst other sources). Phylogentic relationships between the aligned sequences can be displayed as phylogenetic trees using software such as TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html Page, 1996), or organised into "operational taxonomic units" (OTUs) using software such as MOTHUR (http://www.mothur.org/; Schloss et al., 2009). In this context an OTU is a grouping defined by sequence similarity, which can be set by the user to correspond roughly with phylum, class, order, family, genus, species, as appropriate. Rarefaction analysis (which can also be undertaken by MOTHUR) can characterize the diversity of a clone library using either rarefaction curves or a numerical indicator such as the Shannon Index (Krebs, 1999).

Next generation sequencing can produce 2-3 orders of magnitude more data than traditional approaches based on cloning and sequencing. Thus, while the basic stages in analysis are similar to the traditional approach, the task of applying it to many thousands of sequences in parallel usually requires the use of different software. The RDP project (described above) has a pyrosequencing pipeline that "processes and converts the data to formats suitable for common ecological and statistical packages". Similarly, QIIME (Quantitative Insights Into Microbial Ecology) is an open source software package for analysing high-throughput amplicon sequencing data, such as 16S rRNA gene sequences (http://qiime.org/).

7 DISCUSSION AND CONCLUSIONS

Microbes can be expected to impact most if not all processes occurring in the geo-environment, and geotechnical engineers should be aware of the potential for harnessing microbial metabolism to bring about desired aims. PCR based methodologies permit the detection of the microbes present and how they change with changing conditions. PCR is relatively easy to use in an engineering setting and the availability of reagents in kit form along (with detailed protocols) means that the barriers to adoption are reasonably low. However this is a rapidly moving field and the advent of high throughput deep sequencing technologies have led to the development of 'metagenomics' and 'metatranscriptomics' which investigates the composite genetic potential of an ecological niche. Instrumentation and cost of sample analysis are still relatively high but likely to fall as capacity and technology increase. The sheer volume of data generated poses a significant challenge in terms of bioinformatics and fully exploiting these technologies will require multidisciplinary collaborations between engineers, molecular biologists and informaticians.

8 REFERENCES

Acinas S.G. et al. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430(6999), 551-554

Acinas, S. G. et al. 2005. "PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample." Appl. Environ. Microbiol. 71(12): 8966-8969.

Borneman, J. & Triplett, E.W. 1997. Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. Appl. Environ. Microbiol. 63:2647-2653

Burke, I.T. et al 2012. Biogeochemical reduction processes in a hyper-alkaline affected leachate soil profile. Geomicrobiology Journal 29 (9), 769–779.

Cardinale, M. et al. 2004 Comparison of different primer sets for use in automated ribosomal intergenic spacer analysis of complex bacterial communities. Appl. Environ. Microbiol. 70, 6147-6156.

Krebs, C.J. 1999. Ecological Methodology. Addison-Welsey Educational Publishers Inc, Menlo Park, CA.

Islam, F.S.. et al. 2004. Role of metal-reducing bacteria in arsenic release from Bengal delta sediments. Nature, 430, 6995, 68-71.

Marchetti, A., et al. 2012 Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. PNAS www.pnas.org/cgi/doi/10.1073 /pnas.1118408109

Metzker, M.L. 2010 Sequencing technologies the next generation. Nature Reviews Genetics 11, 31-46.

Mount, D.W. 2004. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Polz, M. F. and C. M. Cavanaugh 1998. "Bias in Template-to-Product Ratios in Multitemplate PCR." Appl. Environ. Microbiol. 64(10): 3724-3730.

Promega 2012. GoTaq® DNA Polymerase Protocol. http://www.promega.com/. Last accessed 4th December 2012

Qiu, X., Wu, L. et al. (2001). "Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning." Appl. Environ. Microbiol. 67(2): 880-887.

Roche 2011a. FastStart Taq DNA Polymerase dNTPpack: Version 7. https://cssportal.roche.com/. Last accessed 4th December 2012.

Roche 2011b. 454 Sequencing System Guidelines for Amplicon Experimental Design. http://my454.com/. Last accessed 10-12-12.

Sunar, N.M. et al. 2009. Enumeration of salmonella in compost material by a non-culture based method. Sardinia 2009: 12th Int. Waste Management and Landfill Symp., 1005-1006.

Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

Wang, Q. et al. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Appl. Environ. Microbiol. 73 5261–5267.

Wang, Z. et al. 2009. RNA-seq a revolutionary tool for transcriptomics. Nature Review Genetics 10 57-63