Uncovering the Microbial Diversity of the Alberta Oil Sands through Metagenomics: A Stepping Stone for Enhanced Oil Recovery and Environmental Solutions

Writing Team: Julia Foght[1], Robert Holt[2], Tom Jack[3], Steve Larter[4], Rekha Seshadri[5], Cor van Kruijsdijk[6], Gijs van Rooijen[7] and Gerrit Voordouw[3]

[1]Department of Biological Sciences, University of Alberta; [2]Genome Science Center, B.C. Cancer Agency; [3]Department of Biological Sciences, University of Calgary; [4]Department of Geology and Geophysics, University of Calgary; [5]J. Craig Venter Institute, Rockville, Md; [6]Shell Canada, Calgary, AB; [7]Genome Alberta, Calgary, AB.

Team and deliverables

The multidisciplinary R&D team will include representatives from Industry and Government, as well as leading researchers with strong grant winning and publication track records in the areas of microbial ecology, genomics, bioinformatics, microbiology and petroleum geoscience. This enables the team to deliver the following key project components targeting the genomes of the subsurface Heavy Oil and Tar Sands (HOTS) reservoirs; the oil sand mine tailings ponds and the oil sand sediment components and associated waters of the Athabasca and Peace Rivers. The R&D team plans to deliver:

1. Methods for sample selection, validation and genome isolation from HOTS.
2. Definition of the microbiomes of subsurface HOTS reservoirs; the oil sand mine tailings ponds and the oil sand sediment components and associated waters of the Athabasca and Peace Rivers.
3. Definition of the main biogeochemical processes active in these systems.
4. Identification and elucidation of microbial processes for cleaner recovery of energy and feedstocks from the tar sands and for mediation of environmental damage caused by processing.
5. Definition of industrial processes for cleaner recovery of energy, feedstocks from the tar sands with reduction in greenhouse gas emissions.
6. Definition of the biotechnological potential of the Canadian HOTS province.

1. Preamble and Summary.

One of the biggest challenges of this century is deciding on which course to chart towards our energy future. Our present society is greatly dependent on fossil fuels with oil, gas and coal contributing 40, 24 and 22% of our current energy supply, respectively, with smaller contributions from nuclear energy (6%) and renewable energy forms (8%). The environmental consequences of this heavy reliance on fossil fuels are well known and include significant increases of CO_2 in the Earth's atmosphere contributing to global warming. Hence, governments are under significant pressure to act to halt or reverse environmental consequences of our reliance on fossil fuels. Key steps towards a sustainable energy future will likely be:

- to reduce per capita energy consumption,
- to ensure that renewables contribute a larger fraction of our energy supply and
- to make extraction and use of fossil fuels as green as possible.

The Province of Alberta is uniquely positioned to play a major role in carving our Energy Future. In the past 70 years, exploitation of oil and gas reserves in the Western Canadian Sedimentary Basin (WCSB), the large coal deposits and the Athabasca Oil Sands have led to the emergence of a burgeoning, multi-faceted Energy Industry and an associated entrepreneurial spirit that tackles energy problems worldwide. At several hundred billion barrels of currently recoverable oil and with up to 2 trillion barrels of oil in place, the Alberta oil sands are the largest remaining petroleum resource in North America. However, production of oil from the oil sands requires large inputs of water, natural gas and hydrocarbon solvent diluents contributing significantly to the environmental problems just identified. New technologies that aim to reduce these inputs must be able to deal with the vast scale of oil production from the oil sands, which is ramping up from one to five million barrels per day in the next decade. However, this vast scale also has its advantages, as incremental improvements in technology would allow proportionate reductions of water or natural gas inputs and thus would be highly valuable. The value of incremental improvements increases if they are made regularly through a sustained effort over a significant period of time.

Although microbes are commonly considered to be detrimental to human health, agriculture, etc., only a few species are in fact harmful. Many others are essential to food and beverage production, manufacture of pharmaceuticals, and biotechnology. It is not generally known by the public that microbes also play important roles in the oil industry, for example by causing “souring” of oil wells and fouling of oil production facilities. They were also responsible, over geological time, for conversion of conventional oil to the current heavy oil and oil sands bitumen in northern Alberta, where they continue to impact the industry, both in the subsurface and in tailings ponds, as described below. The wide variety of microbes, in both the natural subsurface and engineered environments of the tar sands, and their biotechnological potential have not been exploited to date. However, recent technological developments and marketplace incentives have made this the key time to explore and examine the microbial interface with the science and industrial base of the oil sands, both to mitigate detrimental activities and to profit from biotechnological opportunities in this sector for immediate and future long-term benefits.

In order to contribute towards this long-term goal representatives from Industry (15), Government (10) and Academia (25) gathered in Calgary for a Genome Alberta-

sponsored workshop on September 27 and 28, 2006 to discuss the potential of biotechnologies for the oil sands. The focus was on how state-of-the-art DNA sequence technologies, such as used for determining the sequence of the human genome, can uncover the properties of microbes living in the oil sands and associated heavy oil deposits and in oil sand extraction-generated environments, such as tailing ponds. Potential deliverables and biotechnologies resulting from determining the oil sands metagenome (the DNA sequence of microorganisms present in the oil sands) are to:

- understand the anaerobic microbial biodegradation processes that shaped the oil sands,
- identify microbes that enhance oil recovery or reduce oil viscosity by upgrading recalcitrant oil sands fractions,
- achieve more efficient water recycling through tailing ponds,
- allow sustained, incremental improvements to production
- reduce environmental impacts through improved process water use and decreased greenhouse gas emissions.

Determining the oil sands metagenome would provide a database for concurrent and subsequent continued development of biotechnology options for the oil sands, as well as for heavy oil reservoirs in Alberta and elsewhere. Funding for such a project can be envisaged through (A) Genome Canada, (B) Genome Alberta and Province of Alberta Agencies (AIF, AERI) or (C) US agencies such as the DOE Joint Genome Institute. Discussions at the workshop and since then have indicated that option (B), involving Genome Alberta, Alberta Agencies and Institutions, as well as other Canadian and International partners is preferred. The intended path forward is described in this document.

2. Why do oil sands metagenomics?

The oil in the oil sands is heavily biodegraded. Light components (low molecular weight alkanes and aromatic hydrocarbons and non hydrocarbons) have been removed over geological time by microbial consortia acting in situ, which react hydrocarbons with water, producing CO_2 and methane. These reactions are still going on today both in the oil sands and in oil sands tailings ponds and provide a strong incentive for determining "who is there, what are they doing and how can we steer their actions to our advantage". One of the problems in characterizing environmental microbial consortia is that many microbes cannot be cultured in the laboratory. This problem can be circumvented by applying genomics techniques, an area referred to as "metagenomics". In metagenomics total DNA is extracted from appropriately chosen environmental samples, propagated in the laboratory by cloning techniques and subjected to large-scale sequence analysis. As an example, use of this approach to analyze microbial communities in the Sargasso Sea uncovered large numbers of genes encoding proteins active in conversion of light into chemical energy. Metagenomics is emerging as a powerful method to study the function and physiology of the microbial biosphere, and is causing us to reevaluate basic precepts of microbial ecology and evolution.

Two recovery processes are currently used to extract bitumen from the oil sands. In surface mining, bitumen is separated from water, sand and associated clay and other minerals to produce bitumen and enormous volumes of tailings waste. The solid wastes in turn are separated from the process water in giant settling basins, where settling of solids

is compromised by the presence of fine particulates (clays). Large amounts of methane (up to ten million liters per day) are generated in these basins by microbial reactions of oil sands hydrocarbons and diluent with water, similar to those that shaped the oil sands. The anaerobic methanogenic bacteria catalyzing this process contribute positively to resource extraction because their activity accelerates settling of fines, improving water recycling. However, they also contribute negatively, because the methane bubbling from the tailings ponds into the atmosphere is an even more potent greenhouse gas than is CO_2.

Production of bitumen from deeper layers(>120m) requires *in situ* viscosity reduction to move the oil. This is achieved by heat through injection of steam, as in steam-assisted gravity drainage (SAGD), injection of solvent (VAPEX) or a combination of the two. Production of a barrel of bitumen through SAGD typically uses two to five barrels of water as steam, currently generated by burning natural gas. The resulting production-associated CO_2 emissions contribute negatively to the environmental footprint of deep oil sands extraction. Although this problem can potentially be solved by injection and storage of produced CO_2 in nearby subsurface reservoirs, alternate ways of viscosity reduction should be considered to reduce energy input and the associated environmental footprint. Also, current production methods recover only 20-50% of the oil in place (a problem similar to that in production from conventional reservoirs) and do not allow effective production from intermediate depth zones that are too deep to mine but too shallow for high-pressure steam injection or from geologically complex reservoirs with water zones and shale barriers in the reservoir. They are also not effective at recovering bitumen from the large resource trapped in Devonian carbonate rocks that underly the tars sands. *In situ* microbial viscosity reduction strategies insitu, targeting microbial removal of heteroatoms such as sulphur and nitrogen or recalcitrant, higher molecular weight bitumen fractions, such as the asphaltenes, may be developed once mechanisms for microbial breakdown of these fractions are known. A production alternative is also to accelerate the existing process of heavy oil formation, which involves microbial reduction of water by hydrocarbons, producing methane, to recover heavy oils as methane on a production timescale of years. This may not be possible in the tar sands perse where degradation has already progressed too far but would be valuable in the Canadian heavy oil belt around Lloydminister for example. Alternatively, what we learn in the tailing ponds may allow us to microbially soften up the bitumen prior to a steamflood, by microbial carbon dioxide production for example or to reinfect steamed reservoirs to recover residual oil as methane after SAGD or CSS (cyclic steam stimulation) processing. An added benefit is that this information should be directly applicable to enhanced recovery of some conventional crude oil resources and the knowledge may even allow us to target the vast residual oil resources in the many Canadian waterflooded oil fields. The possibilities are truly vast!

Apart from these biotechnology applications the oil sands should also be regarded as a unique environment where oil, is converted to CO_2 through natural processes, that include biodegradation, as well as physical and chemical weathering. Studying this process in detail does not require a prohibitive drilling budget. The oil sands are thus an ideal natural laboratory for studying the mechanisms of natural recycling of oil hydrocarbons over geological time. With the deep subsurface biosphere being the bulk of the microbial biosphere of Earth, the tar sand reservoir portal provides a convenient access route to the potentially enormous biotechnological resource of the deep biosphere.

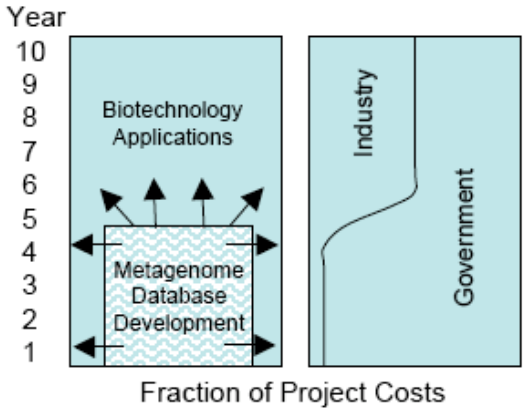
All of these considerations make oil sands metagenomics timely, important and worthwhile!

3. Oil sands metagenomics as part of the larger picture.

Given the very size of the oil sands hydrocarbon resource (on par with presently known reserves of conventional world oil of three trillion barrels), many reports have appeared already on how best to coordinate R&D to develop the resource, contribute to upgrading and solve transport and environmental problems. As an example, the report "Bitumen and very heavy crude upgrading technology" by L. Flint (2004) produced with sponsorship of NRCan and AERI indicates some biotech opportunities for desulfurization and ring cleavage, but without providing much detail. Representatives from Provincial Government Agencies (ASRA, Alberta Innovation and Science and AIF), attending the Genome Alberta Workshop, stressed the need for a more broadly based, long term (10 years) oil sands biotechnology project, which makes effective use of the database generated by the metagenomics project. Hence biotechnology applications should be developed concurrently with and subsequently to the metagenomic data collection, which is estimated to take 4 years (see diagram). A larger biotechnology project should aim to uncover the microbiology, as well as to steer and utilize microbial processes in Canadian HOTS. Broader biotechnology and bioengineering objectives would include:

- Microarray monitoring of HOTS microbes to check progress of biotechnology processes;
- Determining oil-water adhesion potential of HOTS microbes for use in demulsification;
- Stimulation of surfactant production in HOTS under *in situ* conditions;
- Characterization/production of enzymes able to break HOTS resins and asphaltenes;
- Identifying HOTS enzymes with broader anaerobic petroleum degradation specificity;
- Release of V, Ni from HOTS petroporphyrins to prevent upgrading-catalyst poisoning;
- Bio-upgrading by selective N and S removal from HOTS aromatic heterocycles;
- Enhanced *in situ* bioconversion of specific HOTS hydrocarbon suites to methane;
- Preventing methane emission from HOTS tailings ponds by stimulating methane-consuming bacteria at the oxic surface;
- Improving HOTS strains by introducing plasmids, phage resistance or other traits;
- Identifying HOTS genes for generating bioproducts and improving human health.

As indicated in the diagram, metagenome database development would consume the majority of project funds for the first 4 years with a minority of funds reserved for biotechnology applications. The arrows indicate the interdependence of these parts. In years 5-10 virtually all funds for the project are reserved for biotechnology applications. The project would be primarily funded by Government (90%) in years 1-4. The Industry contribution to total funding of the project would rise from 10% initially to 50% in years 6-10.



4. Research Plan

Considerable effort was spent during presentations and discussions at the Genome Alberta workshop on deciding how best to determine the oil sands metagenome. The discussion focused on three aspects: (1) Sampling, (2) Sequencing data generation and (3) Analysis. Based on these discussions, the proposed research plan is as follows:

Sampling. Frozen core samples and produced or well test water samples will be obtained from "pristine" reservoirs from both oil sand and heavy oil (e.g. Athabasca, Peace River, Cold Lake, Lloydminister) reservoirs. Large volumes of material, particularly the water phase, will be collected to allow adequate DNA recovery from low abundance microbes. Tar sand samples will be collected at various depths, from the lower oil-water contact zone in subsurface reservoirs through to atmospheric contacts in mine and outcrop settings. Tailings pond material will be studied as a composite sample from different strata or different ponds to give an overview of diversity. There will also be sampling of laboratory enrichments for, for example, anaerobic hydrocarbon-degrading, aerobic bitumen-degrading and naphthenic acids-degrading consortia. Methods for reproducible and representative DNA recovery from these samples will be optimized.

High throughput cultivation. Since a vast majority of HOTS microorganisms are uncultured, a technique known as gel encapsulation can be combined with massively parallel microbial cultivation under low nutrient flux conditions, to help cultivate new species from the oil sands. Briefly, this procedure involves encapsulation of cells in agarose gel microdroplets, growth in environmental extract media followed by flow cytometry to detect microdroplets containing microcolonies, that are subsequently sorted into microtiter plates for further functional characterization. This technology has been applied previously to samples from various environments, including seawater and soil (K. Zengler, PNAS, 2002). Isolates can be genetically characterized, or enrichment experiments can be devised for simulating microbial recovery (hydrogen and methane) or microbial upgrading (heteroatom removal, phenanthrene degradation, etc.) processes.

Sequence data generation. Genomic libraries will be prepared from isolated DNA using various vector systems. First, small insert libraries (2 to 10kb) will be made from all samples in plasmid vectors, and approximately 1000 random clones from each library will be sequenced using the traditional Sanger method and capillary electrophoresis instruments. These sequences will be used to determine library quality and to estimate the diversity or species represented. From each sample we will also undertake 16S ribosomal RNA sequencing as an independent assessment of species diversity. Additional high molecular weight DNA will be obtained from the samples that show the highest diversity, and large insert libraries (40 to 200 kb), will be constructed in fosmid and/or Bacterial Artificial Chromosome vector systems. The quality and representation of these libraries will be verified and they will be sequenced to a much greater depth (approximately 1 million reads total per sample) in order to capture data from low abundance organisms and to assemble complete composite genome sequences for the most prevalent organisms. Large inserts libraries will also be screened (by PCR) for presence of key genes (e.g., phylogenetic anchors, process-specific genes) and targeted for sequencing. This allows the possibility of recovering complete operons such as for metabolic pathways, and to identify linking clones and expand genome coverage.

The utility to metagenomics of next generation sequencing platforms (such as those offered by the companies Solexa or 454) is not yet clear. While these instruments promise to provide orders of magnitude more sequence data per unit cost, significant drawbacks include the short read lengths (25 to 200bp) and the fact that unlike clone-based Sanger sequencing approaches, the sequencing by synthesis procedure does not establish a clone archive to support sequence finishing and functional studies. However, given the power of these approaches to generate large amounts of sequence data, we will use them to the degree possible to supplement our data sets.

Bioinformatic analysis and data distribution: At the core of this project is the analysis of the sequencing data with the goal of understanding and exploiting the biological processes at work in these environmental communities. We will be assembling a complex dataset, which is a significant challenge to the design parameters of currently available genome assembly and annotation softwares. Shotgun sequence data will be assembled into contigs using the Celera Assembler and/or other assembly software (we will continuously evaluate and use the most appropriate assembler). Resulting contigs and additional binning methods will allow us to link multiple genes from the same genome or bacteria, and may result in draft whole genome sequences from the most abundant organisms. Assembly components will be grouped to estimate the stoichiometry of the organisms observed in the dataset. After assembly, sequences will be annotated by comparison to all publicly available sequence data, capturing homology to known genes, while simultaneously revealing sequence conservation in previously un-annotated regions, suggesting areas of regulatory functions or perhaps even new classes of genes. These similarity searches will be done in conjunction with gene finding algorithms to fully annotate the environmental sequences. This automated analysis will be the starting point for more thorough study and curation of genes, gene families, and even complete genomes for high interest organisms in HOTS environments. We will also leverage existing resources such as high-performance optical networking access via CANARIE/National Lambda Rail (http://www.canarie.ca/about/about.html) to utilize grid computing resources provided by new metagenomics cyberinfrastructure initiatives worldwide.

5. Progress

Culture-based methods to study oilfield microbes have provided some limited information on the composition of microbial consortia. The primary focus has been on understanding the physiological potential of specific groups of microorganisms, such as sulfate-reducing and fermentative bacteria, rather than the composition of the entire community *in situ*. More recently, insights have been gained into the membership of the oil reservoir microbiota through analysis of bacterial 16S rRNA sequences (Magot, M., et al., *Microbiology of petroleum reservoirs.* AVL 2000.; Orphan, V.J., et al., *Culture-dependent and culture-independent characterization.* AEM, 2000; Bonch-Osmolovskaya, E.A., et al., *Radioisotopic, culture-based, and oligonucleotide microchip.*AEM, 2003). Predominant groups comprise various mesophilic and thermophilic Bacteria and Archaea, exhibiting fermentative, iron-reducing, nitrate-reducing, sulfate-reducing, acetogenic and methanogenic metabolisms. Very little is known about the relationship between the activities of these individual groups and the functioning of the ecosystem, and attempts to culture isolates have not always been successful.

Analysis of microrganisms has been performed on both HOTS reservoir core and reservoir water samples. Contamination issues remain a major issue but can be overcome with care and appropriate choice of samples guided by geochemical and geological insights. Generation of supporting geochemical and geological information and understanding will be a significant part of the project as without this intelligent choice of samples and sample quality validation will not be possible. Several such studies have already occurred.

Examination of 16S rRNA gene diversity in production water samples *(Casey Hubert, Thomas Oldenburg, Gerrit Voordouw, Steve Larter (University of Calgary), Ian Head (University of Newcastle), Rekha Seshadri (J. Craig Venter Institute).* Basal mine production waters of Athabasca oil sands were surveyed for 16S rRNA gene diversity and microbial lipids. These waters (0-20 m thick) were located below and directly connected to a 50-80 m oil sands body, with no contact to surface aquifers. The basal waters were discharged 500-2500 m away from the active production area two years in advance of excavation. Waters discharged from six wells around the excavation with a total flow rate of 250 m^3/hour were collected under sterile conditions. Biomass was vacuum filtered, lysed, and DNA was isolated using a modified Marmur method.

A 16S rRNA gene clone library was constructed for each and 384 clones were end-sequenced. To eliminate the possibilities of chimaeras, we used an open-access program (http://foo.maths.uq.edu.au/~huber/bellerophon.pl) to detect chimeric sequences in multiple sequence alignments. Bacterial and archaeal communities exhibited limited diversity. The archaeal phylotypes appeared to be associated primarily with the orders Methanosarcinales and Methanomicrobiales (Fig. 1). The Bacteria were predominantly in the delta (*Geobacter*) and epsilon subdivisions of the Proteobacteria (*Arcobacter, Sulfurospirillum*). The latter organisms are often found in oil fields and thought to oxidize oil organics or sulfide with nitrate or oxygen.

Examination of 16S rRNA gene diversity in tar sands samples *(Ian Head, Steve Larter The University of Newcastle upon Tyne).* As part of the Bacchus2 project, frozen core samples from an Athabasca tar sand reservoir were obtained from a depth of 486-488 m and frozen immediately. The temperature at the sampling depth was 30^{o}C and associated gas contained 99 mol % methane. Nested PCR was used to detect archaeal 16S rRNA gene sequences. Several archaeal lineages were identified in 16S rRNA gene clone libraries including sequences from a novel lineage and sequences associated with environments exhibiting anaerobic oxidation of methane (Head et al, 2003; Fig. 1). A considerable proportion of the 16S rRNA gene fragments recovered from tar sand samples were related to sequences of the ANME-1a group, associated with environments exhibiting anaerobic methane oxidation. Some sequences were associated with Marine Benthic Group D Euryarchaeota. A further very deep-branching group of novel Archaea was also identified. These are quite different from Archaea typically reported in produced water samples from petroleum reservoirs. Metagenome analysis will be enabled and guided by these extensive and growing libraries of 16S rRNA data from the tar sands environment that will allow us to focus our efforts in a cost-effective manner to study this environment.

Metagenomic sequencing of the tar sands *(Rob Holt, Canada's Michael Smith Genome Sciences Centre).* Core samples were obtained from the Poplar Creek region of the Athabasca tar sand region. Tar sand material was removed from a section of the core

corresponding to a depth of 75 meters. Tar sand material was extracted in TE buffer with manual stirring at ~50°C. The aqueous layer was removed and treated with lysozyme, 5% Triton X100 and then Proteinase K to release microbial DNA. DNA was isolated by SCODA (Synchronous Coefficient of Drag Alteration) electrophoresis, with a yield of approximately 1 ng of high molecular weight genomic DNA per gram of tar sand material. DNA was sheared, end polished, adapter ligated and cloned into a medium copy number plasmid. 2,659 clones were end sequenced from this library. After screening out vector, poor quality, viral and eukaryotic sequences 2,050 sequence remained and these were compared by BLAST to the NCBI nr database using alignments criteria of a minimum of 50 bp match length and at least 30% sequence identity over the matched region. While no perfect matches were observed, the most abundant sequences were those with matches (~ 50% identity on average) to *Caulobacter crescentus* (10% of hits), *Novosphingobium aromaticivorans* (8% of hits), *Halomonas elongata* (7% of hits), and *Rhodopseudomonas palustris* (4% of hits). This low similarity to existing genes in the NCBI databases suggests that the rate of discovery of new genes and organisms in this environment is potentially very high.

Analysis of methanogenic microbial communities from oil sands processing tailings *(Tara Penner, University of Alberta, MSc 2005)*. Mature fine tailings (MFT) from two methane-producing settling basins at Syncrude Canada Ltd. were sampled at different depths for total DNA. Archaeal and Bacterial clone libraries were constructed (total ~600 clones) and analyzed by Amplified Ribosomal DNA Restriction Analysis (ARDRA) and unique16S rRNA gene phylotypes were sequenced fully. Most Archaea were related to acetate-utilizing *Methanosaeta* spp., whereas the Bacteria were commonly related to beta-Proteobacteria (e.g., hydrocarbon-degraders), delta-Proteobacteria (e.g., sulfate-reducers) and a few potentially syntrophic clostridia. In contrast, methanogenic enrichment cultures comprised hydrogen-utilizing *Methanocalculus* spp. and the Bacteria were predominantly clostridial homoacetogens. Complementary chemical experiments in microcosms demonstrated that hydrocarbons present in naphtha diluent (used in bitumen extraction) that enters the settling basins does support methanogenesis by MFT (Siddique et al., Env Sci Technol. 2006, and submitted). The results indicate that a complex consortium of microbes exists in the settling basins with intertwined metabolism that results in production of the enormous volumes of methane being emitted every day.

FIGURE 1 LEGEND:
Phylogenetic neighbor-joining tree of representative groups of Euryarchaeota in production water and tar sands samples from the oil-water transition zone. Tar sand sequences are highlighted in yellow while production water sequences are shown in blue.