White Paper

Discovering the functional potential of microbial communities through whole metagenome shotgun sequencing analysis

A metagenomic survey of Antarctic desert soil microbiomes with CLC Microbial Genomics Module

Winnie Ridderberg, Ph.D. and Jonathan Jacobs, Ph.D.
QIAGEN Bioinformatics – Aarhus, Denmark

Introduction

The vast majority of microbes inhabiting our planet remain uncultivated. It is estimated that under 1% of all microbes have been characterized in culture (1). Metagenomic approaches are the key to unlocking this treasure of hidden biological information. Metagenomics allows us to discover not only which microbes are uncultivable, but also gain insight into their functional potential.

QIAGEN®'s CLC Microbial Genomics Module, an extension to CLC Genomics Workbench designed explicitly for analyzing microbial genomes, provides a range of tools for characterizing microbial communities and analyzing their functional content. In this white paper, we demonstrate the capabilities within CLC Microbial Genomics Module for community profiling, metagenome assembly and functional prediction using previously published data.

Microbial communities in extreme environments are receiving increasing interest. These prokaryotic inhabitants can survive under extraordinary conditions, and insights to their functional capabilities and survival strategies are of potential value for industrial, medical and commercial applications.

Terrestrial Antarctica is among the most extreme environments on the planet, with extreme freezing temperatures, repeated freeze-thaw cycles, UV radiation and limited water, carbon and nitrogen availability. Nevertheless, an astonishing abundance of life has been discovered in these habitats. Studies have shown that Antarctic microbial soil communities are highly specialized and most likely remain in a dormant state, in order to limit energy consumption. Antarctic soils are deficient in organic carbon and contain very few primary producers. Therefore, one would wonder where these microbial communities get their energy and carbon from? This was the central question in a recent study by Mukan Ji and co-workers (2), in which shotgun metagenomics was used to study the functional potential of microbial communities from Antarctic soils. We have used the Antarctic soil metagenome data of Ji et al. to demonstrate the functionalities within the CLC Microbial Genomics Module for functional characterization of microbes.

Materials

Three samples of surface soils from Robinson Ridge, Antarctica, were collected in 2005. Robinson Ridge is an ice-free polar desert in the coastal region of eastern Antarctica. The site is devoid of vascular plants, but harbors limited macrofauna with tardigrades and nematodes. The three samples were collected from the top ten centimeters of soil with a distance of two meters between them.

Metagenomic libraries were constructed using the Nextera™ DNA Library Preparation Kit (Illumina®) and sequenced on a HiSeq® 2000, producing 2x100 bp paired-end reads. Sequencing data were deposited by authors in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession numbers SRR5223441, SRR5223442 and SRR5223443.

For comparison, we downloaded data from an additional eight study sites in Antarctica (Figure 3). The files are accessible in SRA through accessions: Lake Vanda, fresh water (SRA644586, SRA644581, SRA644577); Club Lake, saline water (SRA616639, SRA616635); Rauer Islands, saline water (SRA616912, SRA616908); Deep Lake, saline water (SRA620732, SRA620728); Adelie Basin, marine sediment (SRA557553, SRA557553, SRA557553); Devils Point – Livingston Island, rhizospheric soil (SRA642611, SRA642611); Ace Lake, saline water (SRA741076, SRA741075, SRA741069); subglacial lake, ice (SRA457286). These data are unrelated to the study of Ji et al.

Data analysis

The analysis pipeline is outlined in Figure 1 and consists of two parts. The first is determining the composition of the microbial community in the soil of Robinson Ridge (part A), and the second is investigating the functional potential encoded within the genomes of individual members of the soil microbiome (part B). All tools used are available in CLC Genomics Workbench 12 and CLC Microbial Genomics Module 4, and later versions.

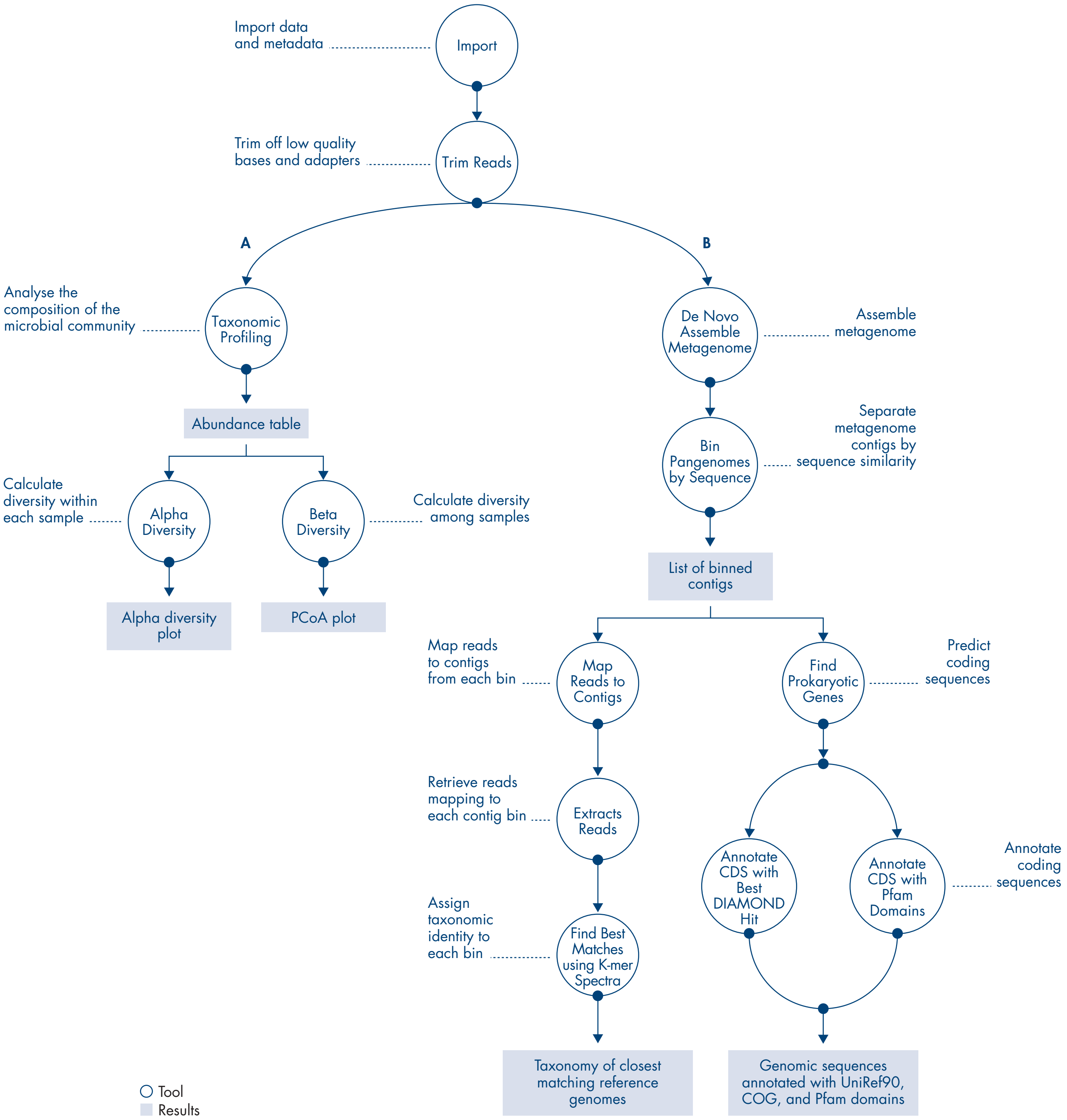
Import data and metadata
Import
Trim off low quality bases and adapters
Trim Reads
A
B
Analyse the composition of the microbial community
Taxonomic Profiling
Abundance table
Calculate diversity within each sample
Alpha Diversity
Beta Diversity
Calculate diversity among samples
Alpha diversity plot
PCoA plot
De Novo Assemble Metagenome
Assemble metagenome
Bin Pangenomes by Sequence
Separate metagenome contigs by sequence similarity
List of binned contigs
Map reads to contigs from each bin
Map Reads to Contigs
Find Prokaryotic Genes
Predict coding sequences
Retrieve reads mapping to each contig bin
Extracts Reads
Annotate CDS with Best DIAMOND Hit
Annotate CDS with Pfam Domains
Annotate coding sequences
Assign taxonomic identity to each bin
Find Best Matches using K-mer Spectra
Taxonomy of closest matching reference genomes
Genomic sequences annotated with UniRef90, COG, and Pfam domains
○ Tool
▨ Results

Figure 1. Tools used for data analysis.

Results

Composition of the microbial soil community at Robinson Ridge

To investigate the microbial diversity of microbiome samples, CLC Microbial Genomics Module offers the tool *Taxonomic Profiling*. This tool maps reads against a database of genome references and reports the abundance of matching reads at different taxonomic levels. The soil community at Robinson Ridge was sampled in triplicates that were processed and sequenced independently. Before analysis (Figure 1, part A), we merged the triplicate sequence files. Before running the *Taxonomic Profiling* tool, a database of reference genomes must be specified, using the built-in function *Download Microbial Reference Database*. This function allows users to either download one of two predefined genome databases optimized on size (to enable running on a machine with 16 GB or 22 GB memory) or to create and download a custom database. In many cases, the predefined databases are sufficient, but for this particular study, we chose to create a custom database to ensure as many of the more unusual microbes expected to reside in Antarctica would be represented in the database. After downloading the database, an index file must be created with the tool *Create Taxonomic Profiling Index*.

The bacterial community at Robinson Ridge was dominated by Actinobacteria (53 %) and Proteobacteria (27 %) (Figure 2). Interestingly, Cyanobacteria were found at very low abundance (0.7 %) as well as Acidiphilium and Rhodospirillaceae (< 0.1%). The absence of large communities of phototrophic bacteria coupled with the deficient availability of organic carbon in the Antarctic soils suggests alternative energy sources must sustain the diverse microbial community observed.

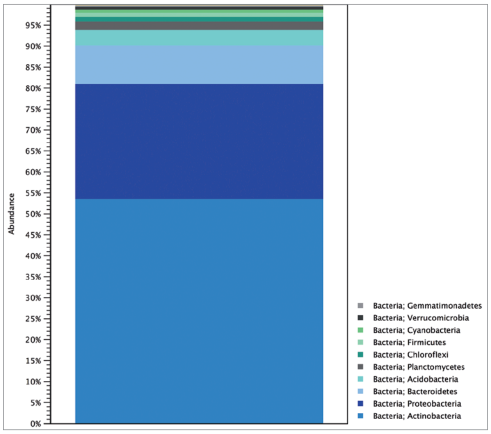

Figure 2. Composition of the bacterial community in desert soil at Robinson Ridge, Antarctica.

For comparison, we downloaded sequencing data sets from eight additional sites in Antarctica (Figure 3). All data were downloaded via the tool *Search for Reads in SRA*, that directly downloads read files and any associated metadata from the Sequence Read Archive at NCBI. To compare the polar desert community profile at Robinson Ridge with different types of communities in Antarctica, we included data from saline water (Club Lake, Ace Lake, Deep Lake, Rauer Islands), freshwater (Lake Vanda), ocean sediment (Adelie Basin), soil (Devils Point) and ice from a subglacial lake.

Geographic Location	Sample Type	SRA
Subglacial lake	Ice	SRA457286, SRA457286 SRA457286
Ace Lake	Saline water	SRA741076, SRA741075 SRA741069
Devils Point, Livingston Island	Rhizospheric soil	SRA642611, SRA642611
Adelie Basin	Marine sediment	SRA557553, SRA557553 SRA557553
Deep Lake	Saline water	SRA620732, SRA620728
Rauer Islands	Saline water	SRA616912, SRA616908
Club Lake	Saline water	SRA616639, SRA616635
Lake Vanda	Fresh water	SRA644586, SRA644581 SRA644577
Robinson Ridge	Polar desert sand	SRA535076

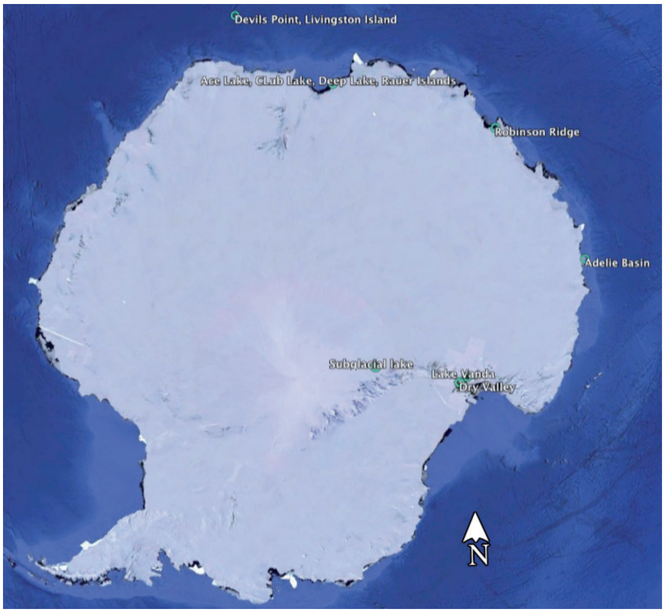
Devils Point, Livingston Island
Ace Lake, CLub Lake, Deep Lake, Rauer Islands
Robinson Ridge
Adelie Basin
Subglacial lake
Lake Vanda
Dry Valley
N

Figure 3. Additional sampling sites included for comparison. Antarctica map: Google Earth Pro (3).

Taxonomic profiling of the additional Antarctic sites showed significant variations in the microbial compositions among the different locations (Figure 4 A). To create the bar chart in Figure 4, abundance tables for each individual sample were merged (*Merge Abundance Tables*). Sites such as Club Lake, Rauer Islands and the subglacial lake were dominated by Proteobacteria. Ace Lake was dominated by Bacteroidetes, and Lake Vanda, the freshwater lake, was dominated by Cyanobacteria. To compare diversity among the different sites, we calculated the beta diversity using the tool *Beta Diversity*. The tool takes an abundance table as input and has several different algorithms available for estimating the diversity among the included samples. In this case, we estimated the beta diversity using the Bray-Curtis measure. The PCoA plot of the Bray-Curtis dissimilarity shows a clear separation of the microbial communities at each location (Figure 4 B). Interestingly, the microbial community in the soil at Robinson Ridge is highly distinct from the other sites sampled.

Functional potential of soil microbiome

To further investigate how the members of the microbial community detected at Robinson Ridge survive in this low-carbon environment without phototrophic residents, we constructed draft genomes and analyzed the functionalities encoded within these (Figure 1 B).

Using the tool *De Novo Assemble Metagenome*, reads were assembled into 15,024 contigs (minimum length 1000 kb). The contigs were separated into individual draft genomes using *Bin Pangenomes by Sequence*, that bin contigs based on sequence similarity and coverage profiles. The contigs were separated into 39 bins containing between 15 and 2901 contigs. The top ten bins, based on the length of the draft genome, were chosen for further investigation (Table 1).

First, we sought to resolve the taxonomic identity of the draft genomes. This was done using our tool *Find Best Matches using K-mer Spectra*. Since this tool requires reads as the input for analysis, we first needed to retrieve the reads belonging to each draft genome assembly. Consequently, we mapped the original reads to each of the draft genomes (*Map Reads to Contigs*) and extracted the reads from the resulting mapping file (*Extract Reads*). The subset of reads retrieved was subsequently used as input for the *Find Best Matches using K-mer Spectra* tool. The tool identifies the closest matching reference among a specified list of reference genomes. Therefore, a database must be provided. The nature of this database depends on the specific study, and in this case, we used the reference database already built for use with the *Taxonomic Profiling* tool as described above. The resulting identification at the phylum level is shown in Table 1, with most draft genomes belonging to the Actinobacteria and just a single genome belonging to the Proteobacteria.

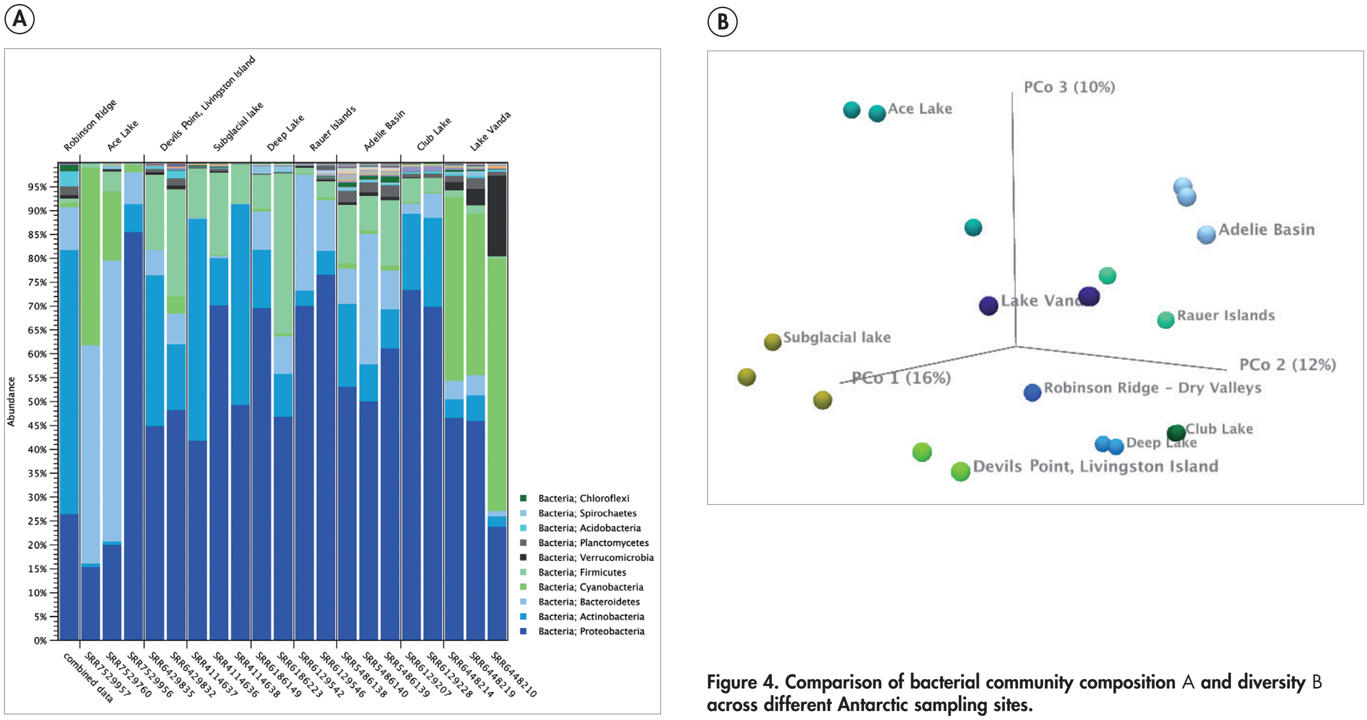

Figure 4. Comparison of bacterial community composition A **and diversity** B **across different Antarctic sampling sites.**

Table 1. Top ten bins obtained by binning metagenomic contigs

Bin	No. contigs	No. nucleotides in contigs	No. reads	Average contig coverage	Phylogeny
Bin00	2901	7057623	1278410	17	Actinobacteria
Bin02	1477	2775636	360314	12	Actinobacteria
Bin03	1227	2056184	177868	8	Actinobacteria
Bin01	897	1601855	248077	15	Actinobacteria
Bin28	728	1459536	181400	12	Actinobacteria
Bin26	693	1357991	154914	11	Actinobacteria
Bin14	718	1283873	175960	13	Actinobacteria
Bin27	601	1202493	128005	10	Actinobacteria
Bin12	682	1048531	116842	11	Proteobacteria; Rhizobiales
Bin19	580	1000287	126402	12	Actinobacteria

Next, we investigated the functional potential encoded within the ten draft genomes. This was done by predicting coding sequences in each genome using the tool *Find Prokaryotic Genes*. The predicted open reading frames were then annotated using three different databases: UniProt Reference Clusters (UniRef90) (4), Pfam (5) and Clusters of Orthologous Genes (COG) (6). To annotate the coding sequences with information from UniRef90 and COG we used the tool *Annotate CDS with Best DIAMOND Hit*. To annotate with information from Pfam, we used the tool *Annotate CDS with Pfam Domains*. All databases are available for download directly from within CLC Genomics Workbench or CLC Microbial Genomics Module.

All ten draft genomes harbored terminal oxidase genes (Table 2) to support aerobic respiration, in line with the habitat being aerated. All genomes also contained a large number of genes to sustain oxidation of organic carbon (data not shown). Surprisingly, genes supporting fixation of CO_2 were widespread among the genomes. Six different pathways for CO_2 fixation are currently known (7), and we detected genes from the Calvin-Benson-Bassham cycle, the reductive TCA cycle and the 3-hydroxypropionate cycle. We detected further genes involved in utilization of CO as a source of energy and carbon (8) and finally, we detected genes supporting H_2 oxidation consistent with the role of H_2 as an energy source (9).

These findings collectively suggest that the bacterial community in the nutrient-poor conditions of an Antarctic desert survive by scavenging CO_2, CO and H_2 from the atmosphere for use as their primary source of energy and carbon.

▷

Table 2. Genes detected in the draft genomes for aerobic respiration and carbon fixation

	Bin00	Bin02	Bin03	Bin01	Bin28	Bin26	Bin14	Bin27	Bin12	Bin19
Terminal oxidase genes										
Heme-Copper oxidases/cytochrome c	x	x	x	x	x	x	x	x	x	x
aa3-type	x	x		x						
cbb3-type	x	x	x			x		x		x
Cytochrome bd oxidases	x	x			x	x				
CO_2 fixation via the Calvin-Benson-Bassham Cycle										
Rubisco (ribulose bisphosphate carboxylase)	x					x	x			
Phosphoglycerate kinase	x									x
Glyceraldehyde-3-phosphate dehydrogenase	x	x								x
Aldolase (fructose-bisphosphate aldolase)	x	x	x		x	x				
Fructose-1,6-bisphosphatase	x	x			x		x			
Phosphoglucoisomerase						x				
CO_2 fixation via the Reductive TCA cycle										
Malate dehydrogenase		x		x	x		x			
Fumarate hydratase	x									
Fumarate reductase	x	x	x				x			
Succinyl-CoA synthetase	x		x	x	x	x				
2-oxoglutrate: ferredoxin oxidoreductase			x		x		x			
Aconitase		x								
CO_2 fixation via the 3-hydroxypropionate cycle										
ATP dependent acetyl-CoA carboxylase	x				x	x			x	
Biotin carboxylase	x			x					x	
Biotin carboxyl carrier protein	x								x	
Carboxyltransferase α	x									
Carboxyltransferase β	x									
ATP dependent propionyl-CoA carboxylase	x	x			x					
Methylmalonyl-CoA epimerase	x	x	x			x				
Methylmalonyl-CoA mutase	x	x	x	x		x	x		x	x
Succinate dehydrogenase			x				x			
Fumarate hydratase	x								x	
H_2 oxidation										
Ni-Fe hydrogenase	x	x		x		x	x	x		
Nickel-depedent hydrogenase		x	x			x				
CO oxidation										
Carbon monoxide dehydrogenase	x				x		x	x		

Conclusions

In this white paper, we demonstrate the utility of CLC Genomics Workbench and CLC Microbial Genomics Module for analyzing the composition and functional potential of metagenomes.

Using these solutions, we show that Actinobacteria and Proteobacteria dominate the microbial community at a polar desert in Antarctica. The community contained very few phototrophic bacteria, such as Cyanobacteria, which often act as primary producers supporting the growth of the remaining community members. However, functional analysis of the draft genomes produced from the metagenomic data indicates that community members could sustain life by using the atmospheric trace gases CO_2, CO and H_2 as their carbon and energy sources.

References

1. Lloyd K., et al (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:e00055.
2. Ji M., et al (2017) Atmospheric trace gases support primary production in Antarctic desert surface soil. Nature 552:400.
3. Google Earth Pro, 7.3.2.5775. Data SIO, NOAA, U.S. Navy, NGA, GEBCO. Image U.S. Geological Survey. Image PGC/NASA. Image Landsat/Copernicus.
4. Suzek B., et al (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926.
5. El-Gebali S., et al (2019) The Pfam protein families database in 2019. Nucleic Acids Research (2019) doi: 10.1093/nar/gky995
6. Tatusov R., et al (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33.
7. Saini R., et al (2011). CO_2 utilizing microbes — A comprehensive review. Biotechnol Adv. 29:949.
8. Quiza L., et al (2014) Land-use influences the distribution and activity of high affinity CO-oxidizing bacteria associated to type I-coxL genotype in soil. Front Microbiol [Internet]. [cited 2019 May 20]. Available from: **https://www.frontiersin.org/articles/10.3389/fmicb.2014.00271/full**
9. Greening C., et al (2016) Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J. 10:761.

To learn more about our bioinformatics solutions, visit these informative sites.

CLC Genomics Workbench
https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/

CLC Microbial Genomics Module
https://www.qiagenbioinformatics.com/products/clc-microbial-genomics-module/

Tutorials and webinars
https://www.qiagenbioinformatics.com/clc-microbial-genomics-module-resources/

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at www.qiagen.com or can be requested from QIAGEN Technical Services or your local distributor.

The CLC Genomics Workbench and CLC Genomics Server are intended for molecular biology applications. These products are not intended for the diagnosis, prevention or treatment of a disease.

Ordering and Technical Support **bioinformaticssales@qiagen.com** | Website **www.qiagenbioinformatics.com**