C-DEBI Data Management and Integration

John F. Heidelberg
University of Southern California
Who we are

DMI Director: John Heidelberg (USC)

Web Master: Matthew Janicak (USC)

Data Portal Lead: Robert Pockalny (URI)

Bioinformatics Specialist: Benjamin Tully (USC)
C-DEBI’s major science objective include to coordinate, integrate, support, and extend the science associated with subseafloor science.
DMI Purpose

- Our larger community expresses the critical needs as:
  - All the data is in public repositories
    - Data saved into perpetuity
    - GenBank, BCO-DMO, IODBdb, MGDS, PRIDEdb, etc.
  - Data and publications are locatable
    - BCO-DMO
    - C-DEBI search page
  - Enabling complex analysis
    - C-DEBI Server
    - Bioinformatics helpdesk and workshops
    - iMicrobe
    - protocols.io
DMI Purpose

• All the data is in public repositories
  – Data saved into perpetuity
  – GenBank, BCO-DMO, IODBdb, MGDS, PRIDEdb, etc.

• Data and publications are locatable
  – BCO-DMO
  – C-DEBI search page

• Enabling complex analysis
  – C-DEBI Server
  – Bioinformatics helpdesk and workshops
  – iMicrobe
  – protocols.io
Data Capture and Exchange

• Every awardee agrees to the C-DEBI Data Management Philosophy and Policy
  – Data are made available as soon as possible
  – Students have a "standard" moratorium until dissertation acceptance or for two years, which ever comes first
  – Data must be deposited in a public data repository
  – All C-DEBI awardees get email inquires about the data status
  – Failure to follow results in no further C-DEBI funds
DMI Purpose

- All the data is in public repositories
  - Data saved into perpetuity
  - GenBank, BCO-DMO, IODBdb, MGDS PRIDEdb, etc.
- Data and publications are locatable
  - BCO-DMO
  - C-DEBI search page
- Enabling complex analysis
  - C-DEBI Server
  - Bioinformatics helpdesk and workshops
  - iMicrobe
  - protocols.io
Data Capture and Exchange

• All C-DEBI data must be either deposited in or linked to the C-DEBI Program within BCO-DMO. [http://www.bco-dmo.org/program/554979](http://www.bco-dmo.org/program/554979)
  – Links to data in public repositories (GenBank, IODP)
  – Accepts most “odd” data types
  – Accepts ALL data (not just published data)

• This allows the community to have access to all C-DEBI generated data in one portal
Data Capture and Exchange

• Currently 50 C-DEBI Projects in BCO-DMO

Data deposition status

- Fully Compliant
- Some Compliance
- Noncompliant
Data Capture and Exchange

- C-DEBI Search tool
  - C-DEBI-contributed publications
  - Links to BCO-DMO for associated datasets
  - Links to protocols.io for associated protocols
    - With this grant cycle we will be requiring protocols deposited.
    - We are working to get past projects to update on a voluntary basis.
  - Expanding to include abstracts, presentations, education and outreach materials, and technical materials
C-DEBI Peer-Reviewed Publications

Search Parameters

Query:  
Type:  
Publication Type:  

Filter Results
Publication Type

All | Book Chapter | Ebook | Film | Journal Article | Magazine Article | Proceedings | Published Abstract | Report | Thesis | Web Article

Sort Results
Order

Date Desc  | Date Asc  | Title Asc  | Title Desc

Published: December 8, 2016

GENOME ANNOUNCEMENTS

Draft Genome Sequence of Antarctic Methanogen Enriched from Dry Valley Permafrost

Authors: Joy Buongiorno, Jordan T. Bird, Kirill Krivushin, Victoria Oshurkova, Victoria Shcherbakova, Elizaveta M. Rivkina, Karen G. Lloyd, Tatiana A. Vishrivetskaya
C-DEBI Contribution Number: 339

Published: November 21, 2016

PROCEDINGS OF THE NATIONAL ACADEMY OF SCIENCES

An oligotrophic deep subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers

Authors: Maggie C. Y. Lau, Thomas L. Kieft, Olukayode Kuloyo, Borja Linage-Alvarez, Esta van Heerden, Melody R. Lindsay, Cara Magnabosco, Wei Wang, Jessica B. Wiggins, Ling Guo, David H. Perlman, Saw Kyin, Henry H. Shwe, Rachel L. Harris, Youmi Oh, Min Joo Yi, Roland Purttscht, Greg F. Slater, Shuhei Ono, Siwen Wei, Long Li, Barbara Sherwood Lollar, Tuilis C. Oinstit
C-DEBI Contribution Number: 340
Project Data

Potential phosphorus uptake mechanisms of the deep sedimentary biosphere
Project Coordinators: Adina Paytan, Laura Zinke, Delphine Defforey, Benjamin J. Tully, Jason B. Sylvan, Barbara J. Cade-Menun, Brandi Kiel Reese

Phosphorus cycling in ridge-flank sediments
Project Coordinators: Adina Paytan, Delphine Defforey

Investigating microbial activities driving organic matter transformations in the deep subsurface
Project Coordinators: Adrienne Hoarfrost, Carol Arnosti
Protocols

Search Parameters

Query:
Search ...

Type:
Protocols

Submit

Sort Results

Order

Date Desc  Date Asc  Title Asc  Title Desc

- Protocols

Basic Illumina Sequence Quality Control
Authors: Benjamin J. Tully

- Protocols

Detecting 16S rRNA Gene Fragments from a Metagenome to Assemble Full-Length 16S Sequences
Authors: Benjamin J. Tully

- Protocols

Detection of protein-synthesizing microorganisms in the environment via bioorthogonal non-canonical amino acid tagging (BONCAT)
Authors: Roland Hatzénichler, Victoria J. Orphan

Posted: November 11, 2016

Posted: November 11, 2016

Posted: November 11, 2016
Search Results

Potential phosphorus uptake mechanisms of the deep sedimentary biosphere
Project Coordinators: Adina Paytan, Laura Zinke, Delphine Defoeroy, Benjamin J. Tully, Jason B. Sylvan, Barbara J. Cade-Menun, Brandi Kiel Reese

Deep phylogenetic and metagenomic analysis of microbial diversity associated with ferromanganese nodules collected from the South Pacific Gyre
Project Coordinators: John F. Heidelberg, Benjamin J. Tully

C-DEBI Newsletter - April 1, 2015

[...] from senior graduate students, postdoctoral researchers, fellows, and newly appointed assistant professors, due May 01, 2015. Arts & Entertainment Ben Tully will always be C-DEBI’s King of the Nerds If you watched this season’s TBS show “King of the Nerds”, you already know that C-DEBI Postdoc Ben Tully, a.k.a. “Marine Biologist” did not win the title. However, he made [...]
Search Results

Search Parameters
Query: nitrate
Type: All

Filter Results
Type: All, Awards, Newsletters, Posts, Project Data, Publications

Sort Results
Order: Date Desc, Date Asc, Title Asc, Title Desc

Project Data

Microbiology and biogeochemistry of Juan de Fuca Ridge flank borehole fluids
Project Coordinators: Michael S. Rappé

[...] viruses, and a wide range of associated biogeochemical measurements including dissolved gases, particulate and dissolved organic carbon, sulfate, nitrate, and others. [...] 

Project Data

North Atlantic Meridional Circulation during the Last Glacial Maximum: Density Structure and Pre-formed Nitrate: Phase I
Project Coordinators: Arthur J. Spivack, Steven L. D'Hondt, Robert Pockalny

[...] bathymetry, seafloor backscatter, and CHIRP sub-bottom seismic data. The team would analyze the composition (chloride, dissolved oxygen, and nitrate concentrations) of pore waters in the recovered sediments shipped to detect the relict signal of deep water chemistry during the last glacial maximum. These measurements will allow the researchers to directly test the influence of [...] 

Project Data

IODP Expedition 336- Mid-Atlantic Ridge Microbiology
Project Coordinators: Carolyn Buchwald, Scott D. Warkel, Wiebke Ziebis

[...] modes of nitrogen transformations have not been widely investigated. We used the N and O isotopes of poroswater nitrate from the North Pond site in the oligotrophic North Atlantic (North Pond) to provide constraints on the extent of both nitrate production (via nitrification) and consumption (via denitrification). Nitrate accumulates far above bottom seawater concentrations [...]
Microbiology and biogeochemistry of Juan de Fuca Ridge flank borehole fluids

Project Coordinators: Michael S. Rappé

Posted: November 11, 2016

Hydrothermally heated fluids circulate everywhere within the permeable basement rock of the upper ocean crust, providing warm temperatures and chemical gradients that support a deep subsurface marine biosphere. The volume of oceanic lithosphere habitable by microbial life is thought to be a substantial portion of the Earth’s crust – extending thousands of meters below the seafloor. During expeditions from 2008 to 2014 we repeatedly sampled basalt-hosted, deep subsurface crustal fluids from four different boreholes drilled along the Juan de Fuca Ridge flank in the Northeast Pacific Ocean using pumps and samplers capable of collecting whole water and filtered particulates in situ. The instrumented boreholes, sitting at 2500 m depth, penetrate ~100 to 250 m of bottom sediments and another ~48 to 300 m of igneous basement where they tap into hot (up to 65 degrees C), anoxic fluid within Earth’s largest deep subsurface aquifer. Nearby bottom seawater and sediments were sampled as controls. Associated data sets include small subunit ribosomal RNA and functional gene amplicon DNA sequences, metagenome sequences, single cell genome sequences, direct counts of microbial cells and viruses, and a wide range of associated biogeochemical measurements including dissolved gases, particulates and dissolved organic carbon, sulfite, nitrate, and others.

Source: http://www.bco-dmo.org/project/638568

Dataset Collections

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th># of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan de Fuca Ridge Flank SSU rRNA gene sequences</td>
<td>4</td>
</tr>
</tbody>
</table>

Related Items

Awards

- Graduate Fellowships
  - Award Dates: December 1, 2012 — December 31, 2014
  - Awarded to: [Specific individual or organization]

- Ridge flank crustal fluid microbial community genomics, expression, and phylogeny
  - Awarded to: Sean P. Jungbluth (University of Hawaii)
  - Advisor: Michael S. Rappé (University of Hawaii)

Research Grants

- Metagenomics, metatranscriptomes, and single-cell genomics of microbial communities inhabiting Juan de Fuca Ridge flank borehole fluids
  - PI: Michael S. Rappé (University of Hawaii)
  - Co-PI: Sean P. Jungbluth (University of Hawaii)
DMI Purpose

• All the data is in public repositories
  – Data saved into perpetuity
  – GenBank, BCO-DMO, IODBdb, MGDS, PRIDEdb, etc.
• Data and publications are locatable
  – BCO-DMO
  – C-DEBI search page
• Enabling complex analysis
  – C-DEBI Server
  – Bioinformatics helpdesk and workshops
  – iMicrobe
  – protocols.io
Enabling complex analysis

- **C-DEBI Server “Kuat”**
  - Dell PowerEdge 9200
  - 96 cores
  - Xenon 2.3 GHz Processors
  - 1 Tb Ram
  - RAID storage – Max Capacity 48 Tb
    - Primarily a Genome/Metagenome Assembler but used for modeling
    - Currently 18 C-DEBI researches have account

- **Bioinformatics Helpdesk - Ben Tully**
  - Available to answer bioinformatics questions for the larger C-DEBI community
  - Offer letters of support and proposal planning
  - Direct analysis of research results
Enabling complex analysis

- Data Analysis & Informatics Support (DAIS) Case Study.

DNA extraction

HiSeq paired end (150 bp)

Generate Hypotheses

get 16S rRNA reads using Meta-RNA

assemble near full-length 16S rRNA

assign taxonomy

Preformed by PI Dr. Jason Sylvan

Perform in Collaboration with B. Tully and the DIAS

unassembled reads

bin into pan-genomes

construct metabolic models

test hypotheses

assemble all reads

$\text{HS}_2\text{O}_3 \rightarrow \text{SO}_4^{2-}$

$\text{soxB}$

search for your favorite functions

Generate Hypotheses

HS₂O₃ → SO₄²⁻

soxB

search for your favorite functions

Data Management and Integration

C-DEBI Site Visit
Enabling complex analysis

• Workshops
  – C-DEBI Introduction to Bioinformatics
    • Lead by Ben Tully
    • Held in 2015 and 2016 the weekend before the AGU meeting
    • Hosted at Stanford
    • The basics of getting started in genome/metagenome analysis
  – ECO-GEO Bioinformatics Workshop.
    • Ben Tully was a co-organizer of a Bioinformatics workshop from the ECO-GEO RCN
    • Full workshop is on protocols.io
Stephanie Carr, 2015 Workshop Participant: The C-DEBI two day genomic workshop delivered a comprehensive and efficient introduction to metagenomic analyses, providing me the ability to complete my work independently. Two thumbs up!

Methanosaeta Bin

![Map of Antarctica with a star indicating a location](image)

- 97 m below seafloor
- 25 m below seafloor
- 14 m below seafloor

![Genome circular diagram](image)
Enabling complex analysis

• Enabling data synthesis through collaborations
  – C-DEBI wants to enable scientist to get the most from ‘Omics data without “reinventing the wheel”
    • iMicrobe
    • Protocols.io
  – Allow data to be useful for train at small primarily undergraduate teaching institutions.
    • Test case in development with Dr. Michael McCormick, Hamilton College
Enabling complex analysis

• Collaborating with an established Cyberinfrastructure for Microbes – iMicrobe and Dr. Bonnie L. Hurwitz,
  – Microbial ecology is often a data intensive science.
  – Bioinformatics allow scientists to integrate diverse and large-scale datasets ranging from DNA to satellite images.
  – C-DEBI, on it’s own, cannot build and maintain the cyberinfrastructures needed.
  – Rather we see the value of collaborations with the iMicrobe project
    • To work with their computational infrastructure to analyze our data.
    • iMicrobe already develops methods to find, tag and reuse data and analyze metagenomic samples.
    • This collaboration will promote large-scale studies and a better understanding of the deep biosphere.
Enabling complex analysis

Project: Metagenomic analysis of microbial CpG methylation in Delaware estuarine riverbank sediment

Project Group: C-DEBI

Project Name: Metagenomic analysis of microbial CpG methylation in Delaware estuarine riverbank sediment

URL: http://lod.bco-dmo.org/id/project/628237

Project Groups (1)
- C-DEBI

Domain (0)
Unknown

Investigators (3)
- J Biddle
- A Marsh
- Dr Adam Marsh (University of Delaware (UDel))

Institution
N/A

Project Code
628237

Project Type
Enabling complex analysis

Project Group: C-DEBI

Map data ©2017 | Terms of Use | Report a map error
Enabling complex analysis

• Through iMicrobe, we already have access to a large (and growing) toolbox.
• Therefore for C-DEBI it is important to teach the use of these tools, not for us to reinvent them...

<table>
<thead>
<tr>
<th>Tools</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABySS</td>
<td>De novo sequence assembly</td>
</tr>
<tr>
<td>ALLPATHS-LG</td>
<td>De novo sequence assembly</td>
</tr>
<tr>
<td>khmer</td>
<td>Probabilistic de Bruijn graphs</td>
</tr>
<tr>
<td>Meta-IDA</td>
<td>De Bruijn graph multiple alignments</td>
</tr>
<tr>
<td>MetaVelvet</td>
<td>De Bruijn graph coverage and connectivity</td>
</tr>
<tr>
<td>Newbler</td>
<td>De novo assembly based on read overlap</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>Single-genome assembler tuned for metagenomics</td>
</tr>
<tr>
<td>SPA</td>
<td>Short peptide assembly for metagenomes</td>
</tr>
<tr>
<td>Velvet</td>
<td>De Bruijn graph coverage and connectivity</td>
</tr>
</tbody>
</table>
Enabling complex analysis

- C-DEBI scientists are constantly improving existing methods and protocols
  - Hard to publish the improvements in a timely fashion.
- C-DEBI funds high risk projects that do not always work.
  - Publication of protocols that do not work is very difficult
- protocols.io solves both these problems with a free protocol repository.
  - Currently, this is voluntary submission, but this grants cycle will have mandatory protocol deposition
Enabling complex analysis
Enabling complex analysis

• There is a huge amount of untapped data in all ‘omic studies.

• Analysis of this data is a largely untapped recourse for independent research projects for undergrads at small colleges

• A test case is under development with Test case in development with Dr. Michael McCormick from Hamilton College
  – Dr. McCormick attended the 2016 Bioinformatic Workshop
  – Working currently with Ben Tully to determine how to make data analysis of deposited ‘omics studies viable as projects for undergraduates.
Questions?