BOLD & mBRAVE: Purpose-Built Research Data Platforms in the Biodiversity Domain

SUJEEVAN RATNASINGHAM

UNIVERSITY of GUELPH

Centre for Biodiversity Genomics

international BARCODE OF LIFE

DNA Barcodes

THE ROYAL SOCIETY

Received 29 July 2002
Accepted 30 September 2002
Published online 8 January 2003

Biological identifications through DNA barcodes

Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Species Grid - 1.9 Million Known, 20 Million Estimated

FEATURE

The Insect Apocalypse Is Here

What does it mean for the rest of life on Earth?

Biodiversity in Rapid Decline

Warning of 'ecological Armageddon' after dramatic plunge in insect numbers

Three-quarters of flying insects in nature reserves across Germany have vanished in 25 years, with serious implications for all life on Earth, scientists say

Flying insects caught in a malaise trap, used by entomologists to collect samples. Photograph: Courtesy of Entomologisher Verein Krefeld

theguardian

Insect abundance has fallen by 75% over the last 27 years

Guardian graphic | Source: Hallmann et al, PLOS ONE

Environmental Risks Rise to Global Dominance

Global risks considered the most likely in the next ten years, by category (1=most likely)

A 'global risk' is defined as an uncertain event/condition that can cause significant negative global impact within the next 10 years. Some category definitions have been adapted over time.
Based on surveys of business, government, civil society and thought leaders
Source: WEF - The Global Risks Report 2020

statista

#4
Major biodiversity loss and ecosystem collapse (terrestrial and marine)

Road Map to Address Data Gaps

TRACKING ECOSYSTEMS

TIMELINE

REVEALING SYMBIOMES

Domain Specific Data Platforms

2005

Reference Data Construction

2018

NGS Big-Data Funnel

iBOL Consortium Organizational Structure

international
BARCODE
OF LIFE

phase two

iBOL Consortium – Stakeholders

30+ Nations

The BOLD Platform – Launched 2005

Data Standards

Centralized Databases

Secure Collaboration

Analytical Tools

Supported Scales

10^2

10^3

10^4

10^5

Tripartite Data Standard Implementation

Data Visitation

Beta Diversity - Sorenson

Sorenson	BIOUG:Churchill	BIOUG:WATERTON-NP:4	BIOUG:PRINCEALB-NP:2	BIOUG:WATERTON-NP:2	BIOUG:PRINCEALB-NP:1	BIOUG:PRINCEALB-NP:3	BIOUG:PRINCEALB-NP:4	BIOUG:WATERTON-NP:1	BIOUG:WATERTON-NP:5
BIOUG:Churchill	0	0	0	0	0	0.0159	0	0	0
BIOUG:WATERTON-NP:4	0	0	0	0.04	0	0.0241	0	0.0357	0.1
BIOUG:PRINCEALB-NP:2	0	0	0	0	0	0	0	0	0
BIOUG:WATERTON-NP:2	0	0.04	0	0	0.0476	0.0412	0.0465	0.1714	0.1111
BIOUG:PRINCEALB-NP:1	0	0	0	0.0476	0	0.08	0.0952	0	0.0625
BIOUG:PRINCEALB-NP:3	0.0159	0.0241	0	0.0412	0.08	0	0	0.0194	0.046
BIOUG:PRINCEALB-NP:4	0	0	0	0.0465	0.0952	0	0	0	0.0606
BIOUG:WATERTON-NP:1	0	0.0357	0	0.1714	0	0.0194	0	0	0.0333
BIOUG:WATERTON-NP:5	0	0.1	0	0.1111	0.0625	0.046	0.0606	0.0333	0

BOLD

OTU	Process ID	Sample ID	Average Distance	Max Distance	Taxon	Count	Distance to NN
OTU-1	WALPA2863-12	BIOUG04512-B01	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-1	WALPA2864-12	BIOUG04512-B02	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-1	WALPA2865-12	BIOUG04512-B03	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-1	WALPA2866-12	BIOUG04512-B04	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-1	WALPA2898-12	BIOUG04512-D12	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-1	WALPA3035-12	BIOUG04521-H06	0.053061947	0.15948963	Leptozestis	6	11.946902
OTU-2	WALPA2867-12	BIOUG04512-B05	0	0	Heteroteucha kershawi	1	8.4529505
OTU-3	WALPA2869-12	BIOUG04512-B07	0.32125536	0.5050505	Chrysonoma sp. ANIC26	3	7.7441072

Variable LifeCycles

Mass

Chemical Properties

Isotopes

Application Extensions

Ecology

Morphology

???

Data Flow

e.g. EOL/GBIF
NCBI
Institutional DBs
Regional DBs
National DBs
Thematic DBs
Public API
BOLDSYSTEMS.org

Database Growth

Authoritative Repository

Clarivate Analytics

Products | Specialty | Newsroom | Training | About | Investor Relations | Blog

← Home

Web of Science

- Web of Science Core Collection
- Data Citation Index
- Databases
- Open Access
- Release Notes
- Set up Web of Science
- Highly Cited Researchers
- Training
- Kopernio

Getting started

Data Citation Index

Connecting Data to the Research it Informs

COVERED BY DATA CITATION INDEX CLARIVATE ANALYTICS

The Data Citation Index (DCI) helps researchers start where discoveries begin

The *Data Citation Index* provides a single point of access to quality research data from repositories across disciplines and around the world. In this index, descriptive records are created for data objects and linked to literature articles in the *Web of Science*. As data citation practices increase, the resource aims to provide a clearer picture of the full impact of research output, as well as to act as a significant tool for data attribution and discovery.

Collaborative Networks

2005 – 102 users from 30 institutions

BOLD Collaboration Network - 2015
1K – 10K
10K – 100K
100K+
Across Nations
1000+ Institutions from 94 countries sharing data on BOLD
Within Nations

BOLD Approach

1. Users own the data until publication
2. Users manage and share their data
3. BOLD curators collaborate with users
4. BOLD curators enforce standards

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) WikiPathways: Pathway Editing for the People. PLoS Biol 6(7): e184. doi:10.1371/journal.pbio.0060184

Layered Data Schema

Oraesia Janzen 03

Notioplusia illustratata

Ctenoplusia sp. ANIC1

BOLD:AAF2716

Image Annotation

d Image
on Framework:

ng sizes of organisms
raction
ecies – morpho traits

sourcing
Network (AI)

NGS & *Big Data* Problems

Drinking from the firehose

Google Scholar – "metabarcoding

Year	Top Pubs	NCBI SRA	Dryad/ Zenodo	ENA	N/A
2013	10	0	1	2	**7**
2015	10	4	3	1	**2**
2017	10	3	2	0	**5**

mBRAVE

Storage

–

Analysis

–

Indexing

–

Discovery

–

Standards

mbrave.net

Macrobiome/Microbiome

tttgagtatacaact

ttcgagcatacgact

aacgtccaaaggagt

ttggagcatacgact

aaggtccaaagagt

ttcgagcatacgact

atcgtccaatggagt

aaggtccaaacgagt

aacgtccaaaggagt

tttgagtatacaact

BARS
LINES WITH STEPS
QV Distribution (Reads)
GC Distribution (Reads)
Length Distribution
Original
Post-Filter
Post-Dereplication
Precluster

Read Length vs QV (Point size: Count Reads)
Read Length vs GC Composition (Point size: Count Reads)
1 to 10
10 to 100
100 to 1K
1K to 10K
10K to 100K
2D Histogram counting sequences at (Length, Mean QV); counted unit is sequences.
2D Histogram counting sequences at (Length, Comp GC); counted unit is sequences.

Filter Runs by Groups: FrozenLysis RegularLysis ADD GROUP
CCDBIONS5-014
1050674 Reads
369 BINs matched
Pending OTUs remaining
81% Filtered
22% Dereplicated
Show More
1012487 Reads
372 BINs matched
2595 OTUs remaining
82% Filtered
23% Dereplicated
949291 Reads
388 BINs matched
83% Filtered
24% Dereplicated
926694 Reads
396 BINs matched
82% Filtered
23% Dereplicated
961442 Reads
370 BINs matched
82% Filtered
22% Dereplicated
941911 Reads
365 BINs matched
84% Filtered
22% Dereplicated
RegularLysis
FrozenLysis
GROUP

Positional QV
Positional GC Composition
Positional QV Post-Filter
Positional GC Composition Post-Filter
Mean QV computed for each sequence column in uploaded sequences; counted unit is sequence.
Positional GC composition for original uploaded sequences; counted unit is sequences.
Mean QV computed for each sequence column in filtered and trimmed sequences; counted unit is sequences.
Positional GC composition for filtered sequences; counted unit is sequences.

Analytical Parameters
1 TRIMMING
2 FILTERING
3 OTHER PARAMETERS
1 Trimming
Trim Front: 30bp 1 000bp
Trim End: 30bp 1 000bp
Trim Length: 300bp 720bp 1 000bp
2 Filtering
Min QV: 0qv 30qv 50qv
UPDATE PARAMETERS

Family
Min Seq Count: 1 11 100
UPDATE
309 BINs included of 329 BINs total (93.92%)
157251 Sequences included of 157311 Sequences total (99.96%)
BIN Composition
Chironomidae (20.71%)
Agromyzidae (8.41%)
Cecidomyiidae (7.44%)
Cicadellidae (6.47%)
Phoridae (6.15%)
Ichneumonidae (4.53%)
Syrphidae (3.88%)
Sciaridae (3.56%)
Anthomyiidae (2.91%)
Ceratopogonidae (1.94%)
other (33.98%) - (See Pie 2)

BRAVE
SUJEEVAN RATNASINGHAM
PROJECT LIST
EDIT PROJECT
USER ACCESS
UPLOAD RUN
BATCH UPLOAD RUNS
RUN PROCESSING
ANALYTICS
Home > Project List > Project
Project MBR-MISEQ375 MiSeq Analysis of gradient lysis reactions from varying insect body parts based on 375 insect panel
Project Overview
MiSeq Analysis of gradient lysis reactions from varying insect body parts based on 375 insect panel.
Project created on 07-Nov-2016 . It consists of 10 runs involving COI-5P data.
Project Analytical Parameters
EDIT
1 Trimming
Trim Front: 50bp
Trim End: 0bp
Trim Length: 450bp
2 Filtering
Min QV: 20qv
Min Length: 200bp
Max Bases with Low QV (< 20): 10%
3 Other Parameters
Pre-Clustering Threshold: 1.0%
ID Distance Threshold: 2%
Minimum OTU Size: 1
OTU Threshold: 2.0%

BRAVE

Facilitating Data Uploads

Data Model

BRAVE

international BARCODE OF LIFE

gsc genomic STANDARDS consortium

ENVO

Buttigieg *et al. Journal of Biomedical Semantics* (2016) 7:57
DOI 10.1186/s13326-016-0097-6

Journal of Biomedical Semantics

RESEARCH Open Access

The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation

Pier Luigi Buttigieg[1*], Evangelos Pafilis[2], Suzanna E. Lewis[3], Mark P. Schildhauer[4], Ramona L. Walls[5] and Christopher J. Mungall[3]

15 Phyla
80 Order
394 Species
Trolla
LADE
Pirbadet
Nedre Charlottenlund
Persaunet
Trondheim
ILA
Nidaros Domkirke
Nidelva
Sverresborg Trøndelag Folkemuseum
Sverresborg
St. Olavs hospital HF
NTNU
Tyholttårnet
Brundalen
Moholt
Othilienborg
Sorgenfri
Steinan
Sluppen

Taxonomic Composition

Orthocladius
Chironomidae
Branchiopoda
Lumbricidae
Diplostraca
Arthropoda
Ephemeroptera
Insecta
Monogononta
Diptera
Ploima
Naididae
Baetidae
Annelida
Cnidaria
Clitellata
Hydrozoa
Haplotaxida
Proteobacteria
Chordata
ohaproteobacte
Rotifera
Hydridae
Other

Site Comparison – Beta Diversity

BRAVE

Beta Diversity Results

RDC-RDM Call 2 Funding

Acknowledgements

Paul Hebert
Evgeny Zakharov
Dirk Steinke
Jeremy Dewaard
Alex Borisenko
Dan Janzen & Winnie Hallwachs
Michelle van der Bank
Many others!

Canadian Centre for DNA Barcoding
CBG Collections Team

CANADA FIRST RESEARCH EXCELLENCE FUND
APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE
INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
OntarioGenomicsInstitute
NSERC CRSNG
GenomeCanada
Ontario MINISTRY OF RESEARCH AND INNOVATION