Software for Metagenomics and Metadata Standards

Iddo Friedberg
CAMERA / CALIT2
University of California San Diego

camera
Marine Microbial Ecology

it2

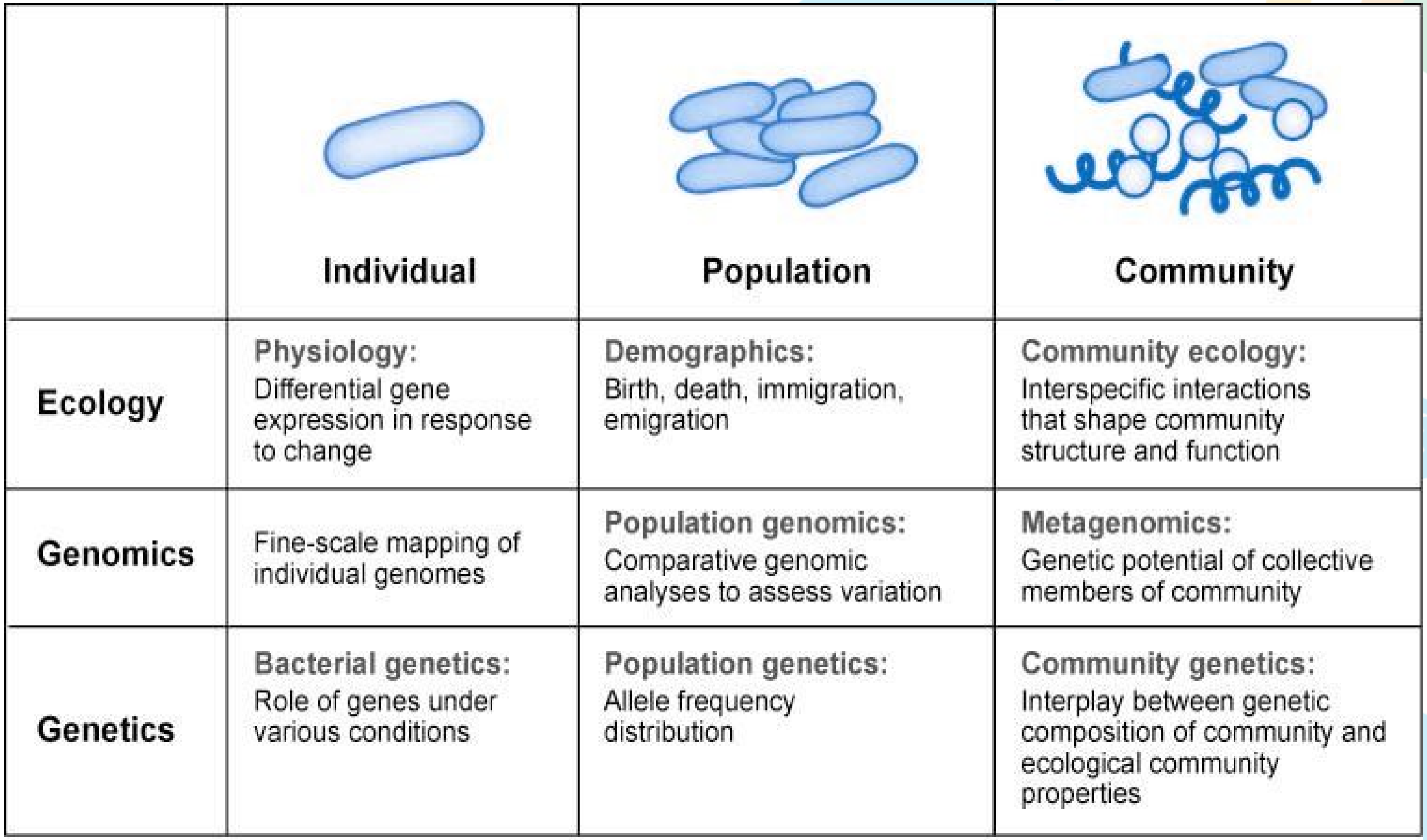
Gordon and Betty
MOORE
FOUNDATION

UCSD

What is Metagenomics?

- Culture-free approach to study microbial communities
  - < 1% of microbes can be cultured
  - DNA directly isolated from environmental sample and sequenced
- Examining genomic content of organisms in community/environment to better understand:
  - Diversity of organisms
  - Their roles and interactions in the ecosystem

Metagenomics: Appy Genomics to Populations & Communities



	Individual	Population	Community
Ecology	**Physiology:** Differential gene expression in response to change	**Demographics:** Birth, death, immigration, emigration	**Community ecology:** Interspecific interactions that shape community structure and function
Genomics	Fine-scale mapping of individual genomes	**Population genomics:** Comparative genomic analyses to assess variation	**Metagenomics:** Genetic potential of collective members of community
Genetics	**Bacterial genetics:** Role of genes under various conditions	**Population genetics:** Allele frequency distribution	**Community genetics:** Interplay between genetic composition of community and ecological community properties

AR Little AEF, et al. 2008.
Annu. Rev. Microbiol. 62:375–401

What Can We Learn?

- **Taxonomic content**: Taxon diversity in a habitat (using taxonomic markers)
- **Functional content**: biological functions, qualitative and quantitative profiles
- **Coping with the environment**: differences in functional content between habitats
- **Decompose the biotic / abiotic elements in a habitat**: metadata analysis
- …

Some Metagenomic Studies

- **Bacterial rhodopsins** (Beja *et al.* 2000)
- **Acid mine drainage study** (Tyson *et al.* 2004)
- **Sargasso sea study** (Venter *et al.* 2004)
- **Wisconsin soil study** (DeLong *et al.* 2006)
- **Termite Hindgut** (Warnecke *et al.* 2007)
- **Human Obesity** (Turnbaugh *et al.* 2006)

… and many, many others

camera Marine Microbial Ecology
it2
Gordon and Betty MOORE FOUNDATION
UCSD

What is CAMERA?

- Community Cyberinfrastructure for Advanced Marine Microbial Research and Analysis
- 7 year $24.5 mil grant from the Moore foundation
- Goal: build a community computational resource for researchers in metagenomics
- ”Cyberinfrastrcture”: hardware, software & data

camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

What is CAMERA? Hardware

What is CAMERA? Hardware

512 CPU, 200 TB, 5 TFlops

What is CAMERA? Data

- Metagenomic sequence data are:
  - Voluminous
  - Noisy
  - Partial
- At the very least they should be:
  - Standardized for processing
  - Associated with Metadata

Why Metadata?

- Microbial communities are affected by and affect their habitats
- Therefore habitat data, in addition to sequence data, is crucial for an **environmental** genomic picture
- Also, sample condition data is needed for reproducibility

© 2001 National Geographic Society. All rights reserved. nationalgeographic.com

Why Metadata?

- Microbial communities are affected by and affect their habitat
- Sequence information + metadata = whole picture

- Habitat Type
- Geographic Location (large area)
- Sample Location (smaller area)
- Country
- Filter Size
- Latitude (exact location)
- Longitude (exact location)
- Depth
- Wat. Dep.
- Chlorophyll
- Oxygen
- Fluor.
- Salin.
- Temp
- Trans.
- BioMass
- Inorg. Carbon
- Inorg. Phospate
- Org. Carbon
- Nitr.
- # Pooled
- # Sampled
- Volume
- Coll. Date

camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Why Metadata?

- Microbial communities are affected by and affect their habitat
- Sequence information + metadata = whole picture

Habitat

Altitude

pH

Country

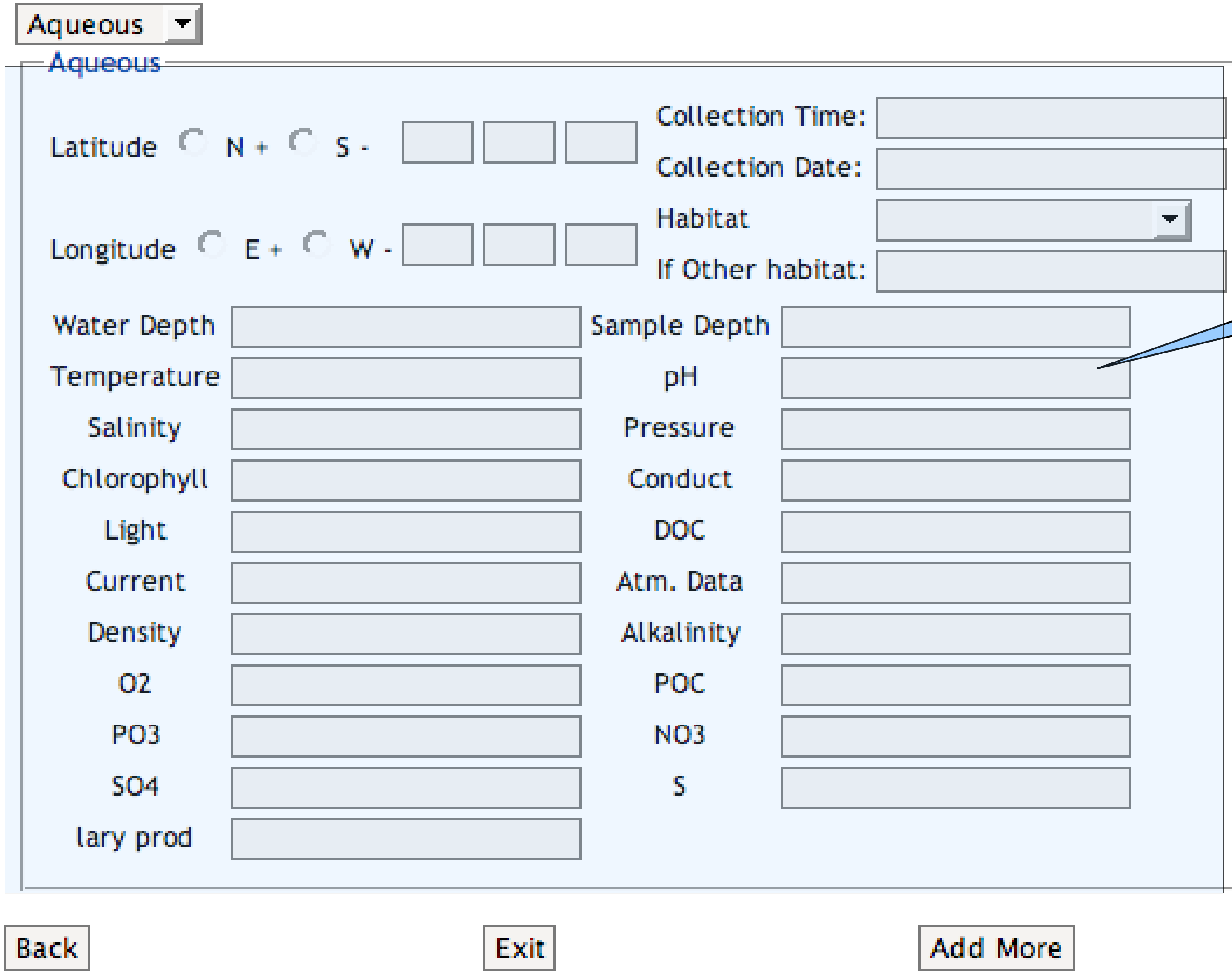
Temperature

Filter size

Data Standards and Data Acquisition

- Minimal Information for (Meta)Genomic Sequences: MIGS/MIMS
- A Metadata standard, developed by the Genomics Standards Consortium
  - Controlled vocabularies e.g. EnvO, PATO, CABRI
  - Common language: GCDML
- Submissions shall comply with a MIMS/MIGS core, but any metadata can be entered via keywords and free text
- Different metadata submission forms for different habitats: (water, soil, air, hosts)

Standards Compliance: MIMS/MIGS



Aqueous ▼

Aqueous

Latitude ○ N + ○ S -		Collection Time:	
		Collection Date:	
Longitude ○ E + ○ W -		Habitat	▼
		If Other habitat:	
Water Depth		Sample Depth	
Temperature		pH	
Salinity		Pressure	
Chlorophyll		Conduct	
Light		DOC	
Current		Atm. Data	
Density		Alkalinity	
O2		POC	
PO3		NO3	
SO4		S	
lary prod			

MIMS

GSC

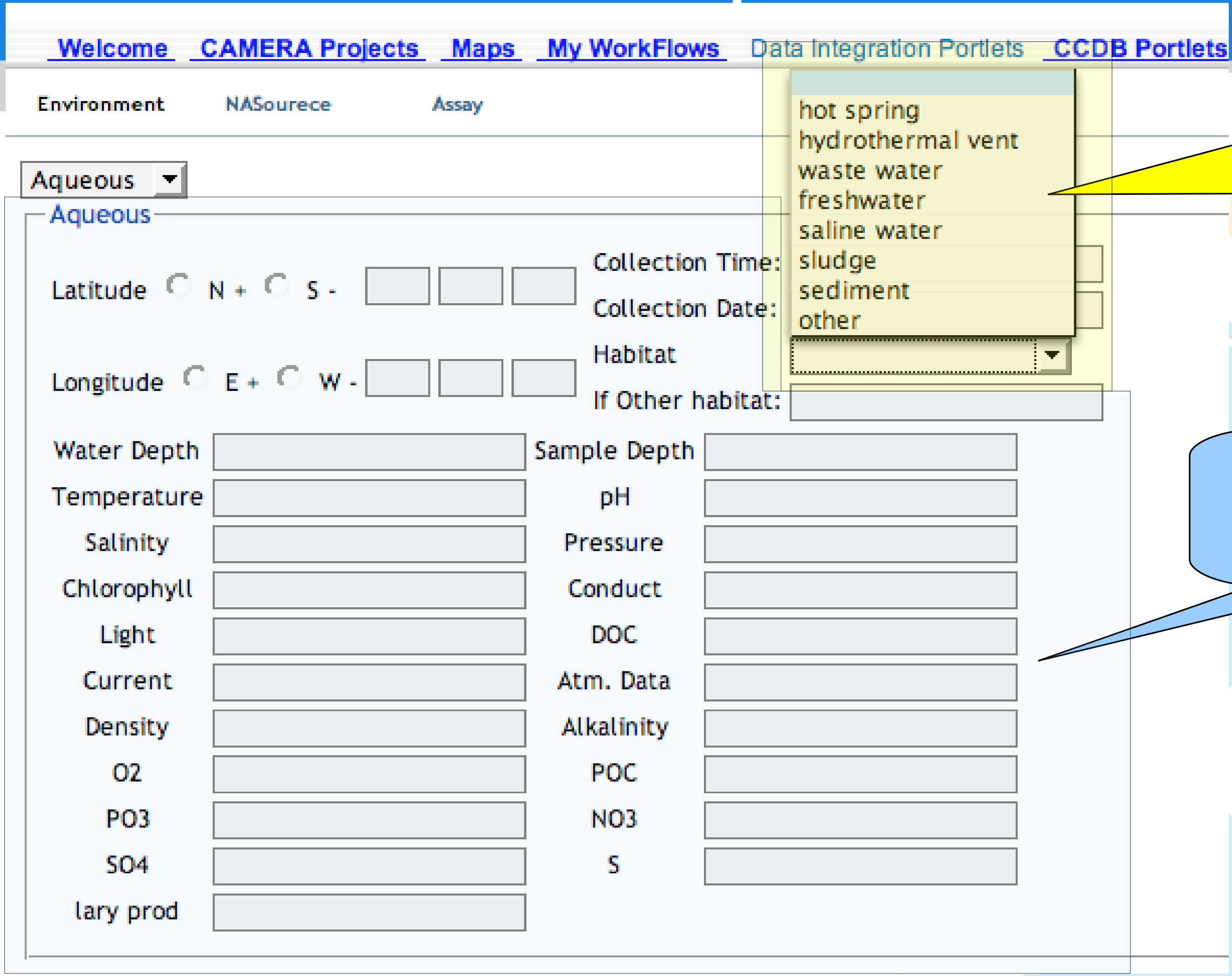
camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Standards Compliance: MIMS/MIGS



Welcome | CAMERA Projects | Maps | My WorkFlows | Data Integration Portlets | CCDB Portlets

Environment | NASourece | Assay

Aqueous ▾

Aqueous

Latitude ○ N + ○ S -

Longitude ○ E + ○ W -

Collection Time:

Collection Date:

Habitat ▾

- hot spring
- hydrothermal vent
- waste water
- freshwater
- saline water
- sludge
- sediment
- other

If Other habitat:

Water Depth		Sample Depth	
Temperature		pH	
Salinity		Pressure	
Chlorophyll		Conduct	
Light		DOC	
Current		Atm. Data	
Density		Alkalinity	
O2		POC	
PO3		NO3	
SO4		S	
lary prod			

Environmental Ontology

EnvO

MIMS

GSC

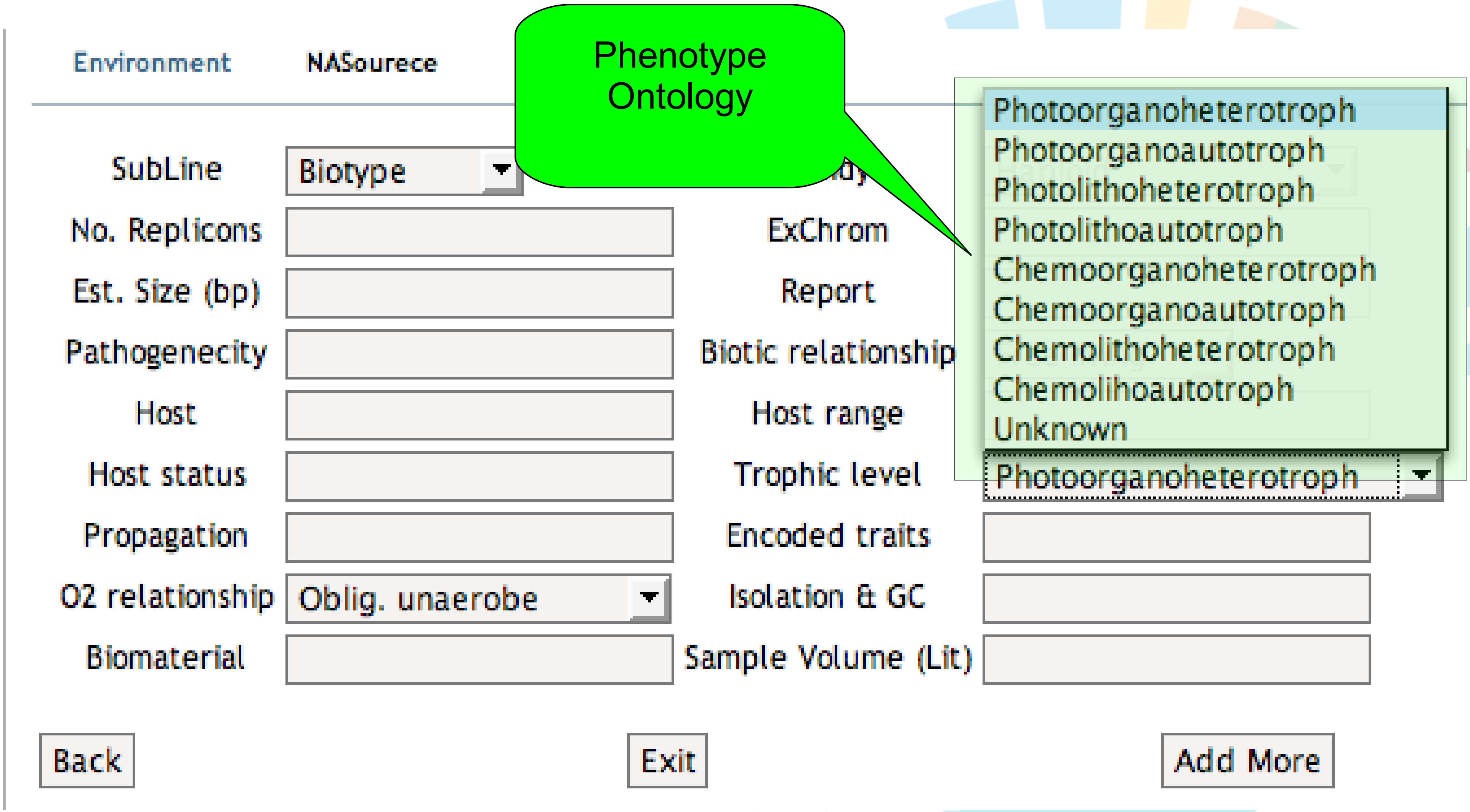
camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Standards Compliance: PATO



Phenotype Ontology

SubLine	Biotype ▾		
No. Replicons		ExChrom	
Est. Size (bp)		Report	
Pathogenecity		Biotic relationship	
Host		Host range	
Host status		Trophic level	Photoorganoheterotroph ▾
Propagation		Encoded traits	
O2 relationship	Oblig. unaerobe ▾	Isolation & GC	
Biomaterial		Sample Volume (Lit)	

- Photoorganoheterotroph
- Photoorganoautotroph
- Photolithoheterotroph
- Photolithoautotroph
- Chemoorganoheterotroph
- Chemoorganoautotroph
- Chemolithoheterotroph
- Chemolihoautotroph
- Unknown

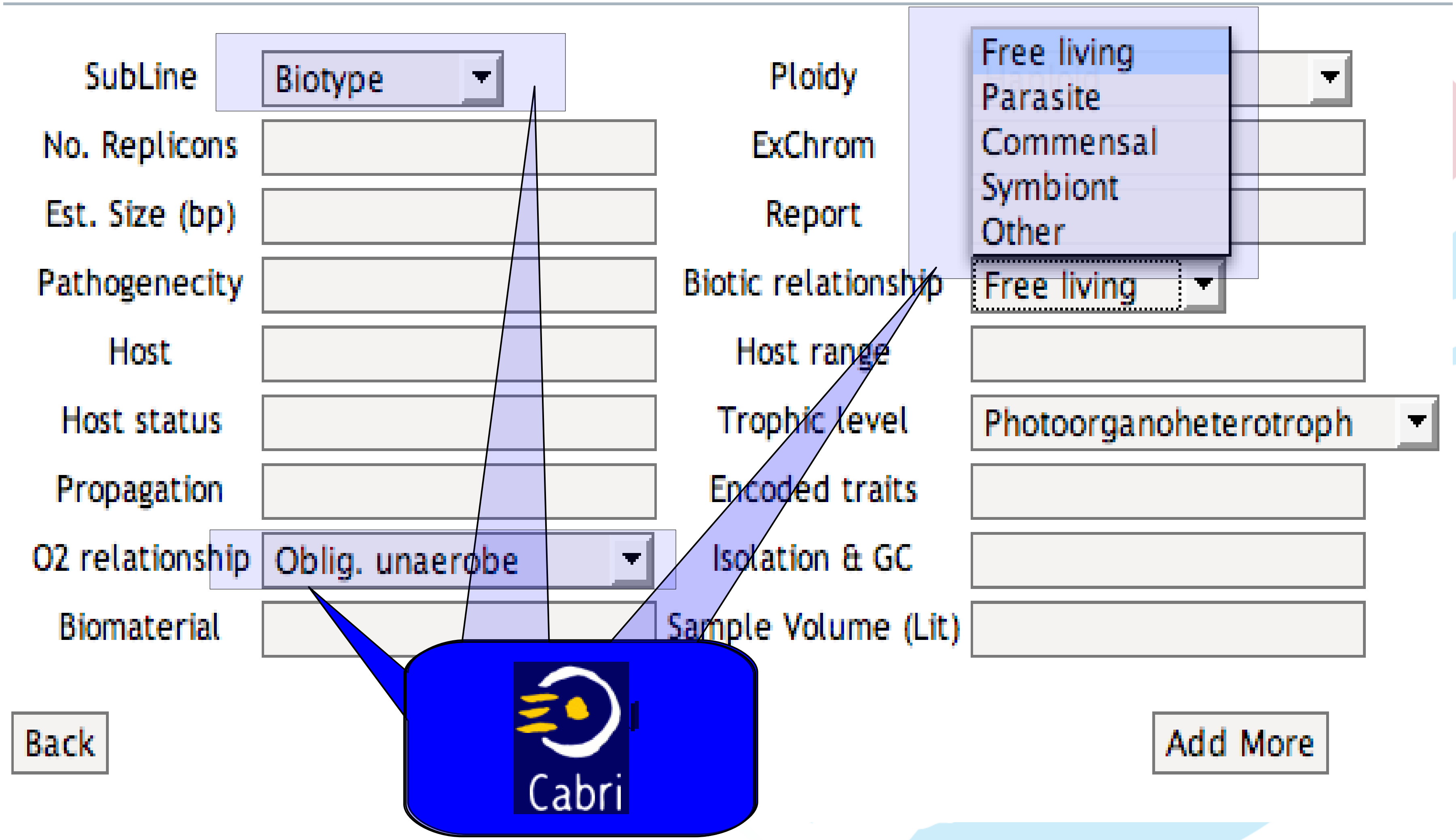
camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Standards Compliance: Controlled Vocabularies



SubLine	Biotype	Ploidy	
No. Replicons		ExChrom	
Est. Size (bp)		Report	
Pathogenecity		Biotic relationship	Free living
Host		Host range	
Host status		Trophic level	Photoorganoheterotroph
Propagation		Encoded traits	
O2 relationship	Oblig. unaerobe	Isolation & GC	
Biomaterial		Sample Volume (Lit)	

Things to Think About

- Visualization:
  - How do we look at ”disembodied” sequence data?
  - ”Fragment recruitment” track
  - Visualization of sequence data <--> metadata associations
- Database: association of metadata and sequence data; queries by metadata
-

Would you Like to Know More?

- The Genomic Standards Consortium. MIMS & GCDML: http://gensc.org
- BioMIST (soon): http://sourceforge.net/projects/biomist/
- CAMERA: http://camera.calit2.net
- Me: http://iddo-friedberg.net

camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Would you Like to Know More?

- The Geno nsortium. MIMS & G http://gens
- BioMIST (http://sour ts/biomist/
- CAMERA: t2.net
- Me: http://

camera Marine Microbial Ecology

it2

Gordon and Betty MOORE FOUNDATION

UCSD

Thanks

- CAMERA data acquisition team:
  - Brian Fox
  - Shulei Sun
  - Jing Chen
  - Laurence Bohannan
  - Jeffrey Grethe
- Genomic Standards Consortium
  - Dawn Field
  - George Garrity
  - Renzo Kottman

camera Marine Microbial Ecology

it²

Gordon and Betty MOORE FOUNDATION

UCSD