

**VIROME AND METAGENOMES ONLINE:
OPTIMIZING FUNCTIONALITY BY LEVERAGING METADATA**

by

Barbra D. Ferrell

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Spring 2015

© 2015 Barbra D. Ferrell
All Rights Reserved

ProQuest Number: 1596849

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 1596849

Published by ProQuest LLC (2015). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

**VIROME AND METAGENOMES ONLINE:
OPTIMIZING FUNCTIONALITY BY LEVERAGING METADATA**

by

Barbra D. Ferrell

Approved: _____
Shawn Polson, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Errol L. Lloyd, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I would like to express my appreciation for the support of my committee chair, Dr. Shawn Polson, who provided direction and encouragement at each increment in this process. Through his patience and consistent guidance, I continued to move forward despite feelings of uncertainty and incompetency. I admire his broad and comprehensive knowledge in so many areas of bioinformatics, and the ways in which he leads so many students.

I would also like to thank my additional committee members, Dr. Cathy Wu and Dr. K. Eric Wommack, for their enthusiastic acceptance of my project and their examples of professionalism and academic achievement. I appreciate VIROME team members Jaysheel Bhavsar and Dan Nasko not only for their hours of guidance and teaching but for their willingness to include me as part of a cohesive group.

Katie Lakofsky has been a great peer and friend since we met early in my program, and has been such a role model of a mother, professional, and successful graduate student.

Thank you to my husband, Robert Ferrell, and family for believing in the value of this achievement. I am grateful for the encouragement I received, the excitement we share now, and the anticipation of what our future holds.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
ABSTRACT	x

Chapter

1	INTRODUCTION	1
	Viral Metagenomics	1
	VIROME and MgOl for Viral Metagenome Exploration	4
	Project Goals	5
2	EVALUATION OF VIROME, MGOL, AND RELEVANT SEQUENCE AND METADATA STANDARDS	7
	The VIROME Pipeline	8
	Metagenomes On-line	11
	International Nucleotide Sequence Database Collaboration	13
	Genomic Standards Consortium	15
	Environment Ontology	18
3	AIM 1: FRAMEWORK TO RETROFIT VIROME AND METAGENOMES ONLINE TO REFLECT METADATA STANDARDS ..	21
	Design Modifications to the Current Database Schema	25
	Retrofit Existing Libraries to Fit New Schema	34
	Implement Database Structure Changes	38
4	AIM 2: IMPLEMENT IMPROVED VIROME INTERFACE TO COLLECT COMPLIANT METADATA FOR FUTURE SUBMISSIONS....	41
	Revise VIROME Library Submission Views.....	41
	Readability.....	42
	View organization	43
	Improve readability of submission form and underlying code.....	45
	Automatic Assignment of EnvO Terms and Other Classifications.....	45

Creating a Dynamic and Adaptive Submission Process.....	49
Dynamic question display	50
Dynamic view display	50
Autofill potentially duplicative fields.....	53
Validation of descriptors	54
Future considerations.....	57
 5 AIM 3: DESIGN OPPORTUNITIES TO BETTER LEVERAGE METADATA IN VIROME AND METAGENOMES ONLINE	 58
Design Modifications to VIROME and MgOl Library Pages	58
Design Outputs to Provide Useful Annotation and Facilitate Batch Submissions.....	61
Design Opportunities to Compare, Search, and Group by Environment and Metadata in VIROME and MgOl	65
Provide an overview of the VIROME libraries by environment.....	66
VIROME: Browse View	68
VIROME: Search View.....	72
VIROME: Compare View	77
VIROME: Interaction Between Pages.....	81
 6 CONCLUSION	86
Impact	86
Future Considerations.....	89
 REFERENCES	91
 Appendix	
A METAGENOMES ONLINE SCHEMA	99
B METAGENOMES ONLINE TERM DEFINITIONS.....	110
C VIROME SUBMISSION FORM FLOWCHART	116
D VIROME SUBMISSION FORM FIELD DEPENDENCY FLOWCHARTS	117
E REVISED VIROME AND MGOL LIBRARY PAGES	120
F VIROME BROWSE VIEW OPTIONS	122
G VIROME SEARCH VIEW WORKFLOW.....	125
H ADDITIONAL VIROME CART OR LAB BENCH FUNCTIONALITY ...	126

LIST OF TABLES

Table 1	Detailed Metagenomes Online database schema. Schema is based upon MlXS table of library descriptors, including field ID, definition, example, field type, syntax, MlXS-compliant category, and preferred units, if applicable.	100
Table 2	Detailed Metagenomes Online database field-specific term definitions. Definitions are extracted from the Environment Ontology and include associated environmental measurement criteria, if applicable.	111

LIST OF FIGURES

Figure 1	MgOl database fields classified in MlXS-compatible sections.....	25
Figure 2	Summary of recommended modifications to MgOl database schema	28
Figure 3	Logic used to assign EnvO terms to retrofit existing MgOl libraries to the new schema.....	36
Figure 4	Modifications to VIROME submission form view organization. Changes include organizing information to read from left to right, enlarging fonts for readability, and grouping fields by their relevant MlXS-compliant categories.	42
Figure 5	Broad organization of VIROME submission form views. Views are arranged by MlXS-compliant categories and include a summary step.	43
Figure 6	Logic used to assign EnvO terms and other classifications to VIROME submissions	49
Figure 7	Logic relevant to physio-chemical modifiers validation, including association with related environmental measurements and classification as an extreme environment.....	55
Figure 8	Sample of VIROME submission form text file output. Several text outputs gather relevant fields in a format compatible with GenBank submissions to the BioProject or BioSample databases or to the Sequence Read Archive.....	63
Figure 9	Sample of VIROME submission form printable file output. The proposed output is formatted to match the unified VIROME and MgOl library views for consistency and ease-of-use across multiple tools.....	65
Figure 10	Proposed dynamic VIROME home view. Graphs and tables will be flexible based upon user-driven criteria. Home-screen displays libraries by one environmental descriptor (environment) in figure 10a, users are able to select a different environmental descriptor in figure 10b, and both the graphic and tabular displays are modified according to that selection in figure 10c.	67

Figure 11	Proposed dynamic VIROME Browse view. Rather than static filters, libraries are sorted based upon user selection of available environment or library descriptors.....	69
Figure 12	Proposed VIROME Browse view using horizontal stacked bar chart for library statistics. This proposed view conserves more vertical space per library than the current pie chart format, allowing the user to evaluate more libraries per screen view	72
Figure 13	Proposed dynamic VIROME Search view. Search criteria can be added or removed to narrow or broaden search results	74
Figure 14	Proposed VIROME detailed search results. Proposed view includes multiple environment descriptors in ORF table	75
Figure 15	Proposed VIROME search view results table header. Modifications suggested to provide clarity and a uniform user navigation experience..	77
Figure 16	Proposed library selection tool to provide functionality of selecting one or multiple libraries and transferring them to another VIROME function.....	83
Figure 17	Proposed VIROME “cart” or “lab bench” function. Selected libraries are maintained in a user-specific container to quickly retrieve user-selected libraries and move them to another VIROME function	85
Figure 18	The revised VIROME submission form flowchart shows form organization by MIxS-compliant categories, conditionally displayed fields, and fields dependent on previous selections.....	116
Figure 19	VIROME revised submission form flowchart for fields dependent upon investigation type selection.	117
Figure 20	VIROME revised submission form flowchart for fields dependent upon environmental fields selections.	118
Figure 21	VIROME revised submission form flowchart for display of environmental measurement fields dependent upon environmental package selection.....	119

Figure 22	Proposed VIROME individual library information view. Displays all associated fields grouped by MIxS-compliant categories, provides supplemental graphics for environmental location and classification, and provides additional environmental groupings for BLAST result hits. A uniform display with MgOl provides consistency and improves ease-of-use across tools.	120
Figure 23	Proposed MgOl individual library information view. Displays all associated fields grouped by MIxS-compliant categories, provides supplemental graphics for environmental location and classification, and provides additional environmental groupings for BLAST result hits. A uniform display with VIROME provides consistency and improves ease-of-use across tools.	121
Figure 24	Proposed VIROME Browse view with flexible sorting available through stacked criteria.	122
Figure 25	Potential VIROME Browse view using tiled selections. Figure 25a shows manageable number of tiles when grouping by environment. Figure 25b demonstrates that grouping by other descriptors such as ecosystem make a tiled display cumbersome.	123
Figure 26	Potential VIROME Browse view using dropdown menu selection. Figure 26a displays tabs and dropdown menus to allow for user selection. Figure 26b demonstrates user selection through drop down menu navigation. Figure 26c shows libraries grouped according to user selection.	124
Figure 27	Proposed VIROME Search view panels show progression of user-specified flexible search criteria.	125
Figure 28	Proposed VIROME cart function from individual library page. Users have the ability to select particular libraries and user them in subsequent comparison exploration.	126
Figure 29	Proposed VIROME cart page with functionality to move libraries to browse, search, or compare views	127

ABSTRACT

Metagenomics has become a dominant tool for profiling the composition of microbial and viral communities, allowing inferences of taxonomic or functional composition through comparison of environmental sequences to reference databases. The power of this approach is limited when environmental proteins show no homology to reference sequences or only show homology to proteins with no known function, which may account for as much as 70% of sequences among viral samples. The Viral Informatics Resource for Metagenomic Exploration (VIROME, <http://virome.dbi.udel.edu>) was developed to provide functional, taxonomic, and environmental homology evidence for viral metagenomes, and to provide visualization capabilities and useful binning and comparison tools. Environmental context is provided through comparison against the Metagenomes Online (MgOl, <http://metagenomesonline.org>) database of predicted proteins identified from 258 microbial and viral metagenomes. MgOl libraries are manually curated with environmental metadata, providing a framework for the sequence homology results increasing the proportion of a metagenome to which meaningful context can be ascribed. This project significantly built upon the utility of VIROME and MgOl by improving the quality and consistency of the associated metadata. Metadata associated with MgOl libraries has been extensively expanded in alignment with standards such as Minimum Information about any (x) Sequence (MIxS) and Environment Ontology (EnvO). An improved VIROME sample submission portal was also designed which allows users to organize their metagenome's or viral genome's metadata in a

MIxS-compliant format. Users have the option to export this metadata in an output format which is compatible with Genbank BioSample submissions. Environmental metadata is further leveraged within each library through new visualizations that enhance a metagenome sequences' environmental context, and throughout VIROME through new search and comparison features allowing exploration of metagenomes with similar environmental profiles or protein homology. Through updates to the MgOl database, the VIROME library submission process, and subsequent library exploration, VIROME is able to leverage environmental annotation to provide flexible, user-driven grouping and comparison and facilitate relevant insights into sequence significance and viral community diversity.

Chapter 1

INTRODUCTION

Viral Metagenomics

The characterization of organisms which inhabit an environment or microbiome allows for deeper understanding of that environment. Profiling the composition of microbial and viral communities can currently be achieved under two main methods (Hamady et al., 2012): targeted gene studies employing small-unit ribosomal RNA (16S rRNA for Archaea and Bacteria or 18S rRNA for eukaryotes) and metagenomic studies. Studies analyzing rRNA utilize highly conserved regions of gene sequencing with species-specific phylogenetic markers to identify species or lineages within a sample. Metagenomic approaches involve shotgun sequencing the genomic content of an environmental sample in order to gain insight about both community diversity and function. Analysis and interpretation of metagenomic data involves comparison of environmental sequences to reference databases composed of sequences from known organisms, allowing taxonomic or functional composition to be inferred (Cole et al., 2009). While both targeted and random sequencing approaches can be informative, they are limited in that many environmental proteins show no homology to reference sequences in the databases, or are homologous to proteins with no known function (Angly et al., 2005; Bench et al., 2007; Breitbart et al., 2002, 2003; Cole et al., 2009; Edwards & Rohwer, 2005; Rosario et al., 2009).

Microbial organisms have been the subject of many metagenomic studies due to their abundance and impact on their environment. Microbial organisms dominate

life in many ways, living in extreme environments such as deep sea vents and leading other life forms in compound recycling, nutrient sequestration, and biomass (Wooley et al., 2010). However, viruses are the most abundant life forms in marine ecosystems, exceeding bacterial abundances by at least one order of magnitude (Suttle, 2005, 2007; Weynberg et al., 2014; K E Wommack & Colwell, 2000). Viruses can infect organisms in all other domains (bacterial, fungal, plant, and animal) (Rohwer et al., 2009; Weynberg et al., 2014) and therefore greatly influence the evolution and population dynamics of their hosts (Suttle, 2005; Weynberg et al., 2014). Viruses mediate microbial host communities through infection and lysis of community members (Bench et al., 2007; Muhling et al., 2005; Thingstad, 2000), and can also alter the phenotypes of host cells through genetic exchange (specific or generalized transduction) or through lysogeny (Bench et al., 2007). Despite their importance, community viral profiling is difficult because there are no universally conserved genetic elements similar to 16S or 18S rRNA (Fierer et al., 2007; Schmidt et al., 2014), culturing is difficult and not representative of environmental diversity (Angly et al., 2005; Staley & Konopka, 1985), viral classification is polythetic and based on phenotypic characteristics (Angly et al., 2005; Lauber & Gorbalenya, 2012; Rohwer & Edwards, 2002; van Regenmortel et al., 2000), and reference databases currently underrepresent viral diversity (Edwards & Rohwer, 2005). The issue is compounded further in environmental samples where most identified proteins are novel (K. E. Wommack et al., 2011; K. E. Wommack et al., 2012; K. E. Wommack et al., 2009).

Multiple databases and bioinformatics tools have been developed to address these shortcomings and to allow for greater exploration and/or visualization of metagenome sequences. UniFrac (Lozupone & Knight, 2005) evaluates microbial

communities based on phylogenetic information. MG-RAST (MetaGenomics Rapid Annotations using Subsystem Technology (Meyer et al., 2008)) provides automatic functional assignment of sequences, phylogenetic and functional summaries, and tools for comparative metagenomics. Phymm (Brady & Salzberg, 2009) phylogenetically classifies short sequence fragments. MetaMine (Bohnebeck et al., 2008) looks for gene patterns in an ecological context through a BLAST search of a user-defined environmentally interesting gene against the Microbial Ecological Genomics Database (MEGDB). Visualization functions have been built into tools and pipelines to allow further exploration of datasets. MEGAN (Huson et al., 2007; Huson et al., 2009) is a metagenome analyzer, evaluating DNA sequences in a BLAST against reference databases and using NCBI taxonomy to summarize and order results. METAREP (Goll et al., 2010) analyzes and compares annotated metagenomics datasets, providing BLAST hit information and corresponding GO (gene ontology) or protein domain IDs. Krona (Ondov et al., 2011) is a visualization tool which allows for exploration of relative abundances. However, many of these reference databases are microbial-centric and therefore miss viral sequences, including those with known functions. Several tools were developed to address this shortcoming, focusing on the exploration of viral sequences. The ACLAME database (A CLAssification of Mobile genetic Elements (Leplae et al., 2004)) is a viral-centric database collection that provides classification of prokaryotic mobile genetic elements (including plasmids and phages). PHACCS (PHAge Communities from Contig Spectrum (Angly et al., 2005)) mathematically models structures of viral communities in order to make predictions about diversity. The Phage Proteomic Tree (Rohwer & Edwards, 2002) describes a genome-based taxonomical system for phage, placing phages relative to near

neighbors and all members in order to predict aspects of phase biology and highlighting genetic markers. Metavir (Roux et al., 2011) and Metavir 2 (Roux et al., 2014) are dedicated to the analysis of viral metagenomes, providing new ways to compare datasets and analyze assembled virome sequences. Finally, additional tools recognize the importance of a sample's environmental context and use environmental annotation increase the exploration of a metagenome. MetaLook (Lombardot et al., 2007) considers a metagenome's corresponding habitat and associated metadata, allowing users to view DNA sequences on a map, organize sequences according to various environmental metrics into environmental containers, and find similar sequences to either georeference database sequences or user-imported sequences through queries against those environmental containers. RAMMCAP (Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline (Li, 2009)) was developed to reduce computational time, explore whole datasets and make use of novel sequences, and to provide new way to compare metagenomes from various environments. Additional tools continue to emerge, including One Codex (<https://beta.onecodex.com>) and Phantome (PHage ANnotation TOols and MEthods, www.phantome.org, though this tool has lost funding resources).

VIROME and MgOI for Viral Metagenome Exploration

The Viral Informatics Resource for Metagenome Exploration, <http://virome.dbi.udel.edu>) is a web-based tool for viral metagenome analysis and exploration (K. E. Wommack et al., 2012). VIROME was developed for metagenomic exploration in response to the needs for web-based analysis, viral-centric tools, environmental context, user-driven grouping and comparison of sequences or libraries, and visualization of analysis results. VIROME utilizes Metagenomes Online (MgOI,

<http://metagenomesonline.org>), an environmental protein database of viral and microbial shotgun metagenome libraries, to provide environmental homology evidence when no homologs can be identified in other annotated reference databases. The power of MgOl, and of its contribution to the VIROME analysis pipeline, is in the environmental context provided with each library and the ability of the user to leverage this metadata in their exploration of analysis results.

This project evaluates the environmental annotation captured in the MgOl database against existing and developing standards and ontologies. Metagenome sequence submissions are influenced by standards established by the International Nucleotide Sequence Database Collaboration (INSDC) and the Genomic Standards Consortium (GSC). Environmental annotation is standardized by the Environment Ontology (EnvO). After evaluation of VIROME and MgOl content and processes against these standards and ontology, updates are recommended to address the modification of the existing MgOl database, the collection of relevant environmental context and metadata as new libraries are submitted to VIROME, and the addition of opportunities to leverage this metadata through user-driven grouping, comparison, and visualization tools.

Project Goals

This project evaluates relevant metagenome and environmental descriptor standards and ontologies, and incorporates three aims to update VIROME and MgOl in ecologically and scientifically relevant ways.

Aim 1 devises a framework to retrofit the current Metagenomes On-line database to annotate existing libraries according to applicable metadata standards. The MgOl database schema is modified in accordance with applicable standards and

ontologies, existing libraries are retrofitted to the revised schema, and the new database structure is implemented.

Aim 2 implements a new VIROME user interface to collect standard-complaint metadata with future pipeline submissions. This aim is accomplished by revising the VIROME user submission pages, automatically assigning relevant environmental descriptors and classifications, and creating a dynamic and adaptive submission tool to facilitate a comprehensive and accurate library annotation.

Aim 3 leverages that environmental metadata to improve the user's exploration of metagenomes, providing new environmental context and comparison features to explore metagenomes with similar environmental protein homology. The VIROME web app is explored for updates and interoperability between its browse, search, and compare tools.

Chapter 2

EVALUATION OF VIROME, MGOL, AND RELEVANT SEQUENCE AND METADATA STANDARDS

The VIROME analysis pipeline is a unique tool for the analysis of viral metagenomes, providing functional, taxonomic, and environmental homology evidence missed by many tools built upon microbial-centric databases. The Metagenomes Online environmental protein database provides environmental evidence of viral and microbial sequences not captured in other reference databases. VIROME's web-based interface, user-driven tools for grouping and comparison, and visualization tools for data exploration create an opportunity for an intuitive and comprehensive metagenome analysis. Where other tools may provide high-level phylogenetic or taxonomic insight, VIROME's connection to well-annotated reference databases provide deep exploration of each library and each sequence. Leveraging the functionality of VIROME and MgOl is dependent on aligning the tool and database with applicable and relevant sequence quality standards, genome and/or metagenome standards, and controlled ontologies. While niche tools have emerged in metagenomic exploration, VIROME's value and longevity will be preserved through its adherence to globally accepted guidance and in its continued ability to produce comprehensive metagenome analysis with functional, taxonomic, and environmental homology evidence. Sequence and metadata standards provided through the International Nucleotide Sequence Database Collaboration (INSDC), Genomic Standards

Consortium (GSC), and Environment Ontology (EnvO) are most pertinent to the targeted VIROME and MgOl improvements of this project.

The VIROME Pipeline

The Viral Informatics Resource for Metagenome Exploration (VIROME; <http://virome.dbi.udel.edu>) is a web-based tool for viral metagenome analysis and exploration (K. E. Wommack et al., 2012). The underlying bioinformatics pipeline emphasizes the classification of viral metagenome sequences (predicted open-reading frames) based on homology search results against known and environmental sequences. The VIROME bioinformatics pipeline consecutively executes sequence quality screening, sequence analysis, and characterization of sequence homology evidence. Functional and taxonomic homology insight is gained through BLASTP results of a homology search against the UniProt Reference Clusters, UniRef 100 database (Suzek et al., 2007). Environmental homology insight is uniquely gained through a BLASTP homology search against a custom database containing predicted peptide sequences from environmental metagenome libraries, Metagenomes Online (MgOl).

Each submission to VIROME is first subject to quality screening. Submissions to VIROME include a nucleotide sequence file in fastq or fasta & qual sequencing format. Each sequence within the file is trimmed for quality, the UniVec database (Kitts, et al., 2011) is used to screen reads for the presence of contaminating vector sequences, and each sequence is trimmed of contaminating linker, adapter, and barcode sequences. The CD-Hit 454 algorithm (Niu et al., 2010) is used to screen 454 sequence libraries for the presence of artificial duplicate sequences known to arise from the 454 library construction protocol. The presence of ribosomal RNA homologs

within the sequence libraries is detected through comparison against a dereplicated collection of ~30,000 ribosomal RNA genes (5S, 16S, 18S, and 23S) from SILVA (Pruesse et al., 2007) and ENSEMBL (Flicek et al., 2012) representing multiple and diverse taxonomy. The presence of tRNAs is determined using tRNAscan-SE (Lowe & Eddy, 1997), and MetaGene Annotator (Noguchi et al., 2008) is used to predict protein-coding open reading frames (ORFs). The predicted ORFs are analyzed using BLASTP against the UniRef 100 and MgOl databases.

Sequence analysis is executed with two parallel BLASTP searches against UniRef 100 and MgOl, respectively. The UniRef 100 peptide database contains clusters of identical peptides within the UniProt Knowledgebase and selected UniParc records, each cluster combining identical sequences and sub-fragments into a single UniRef entry with a representative protein (Suzek et al., 2007; “UniProt knowledgebase,” n.d.). Predicted ORFs from VIROME with a significant hit to a UniRef 100 protein can be assigned taxonomic characterization based on the taxonomic origin of the top UniRef 100 BLAST hit. A VIROME relational database maintains connections between UniRef 100 sequences and six annotation databases (SEED (Overbeek et al., 2005), ACLAME (Leplae et al., 2004), COG (Tatusov et al., 2000), GO (Ashburner et al., 2000), KEGG (Kanehisa et al., 2008), and PHAGE-SEED (<http://phantome.org>) that provide additional functional hierarchical descriptions of peptides. Therefore, where the UniRef 100 homolog also occurs in any of those annotated protein databases, functional characterization of the original VIROME peptide sequence can also be inferred. A significant portion of the VIROME peptide sequences may not show significant homology to known proteins in the UniRef 100 database. To address this issue, the MgOl peptide database was created by

members of the VIROME team, currently with over 56 million predicted peptide sequences from 258 metagenome libraries. Predicted ORFs from the VIROME file of peptide sequences can be environmentally characterized based on homology search results against the MgOl collection of environmental proteins. Environmental descriptions and metadata and associated with each MgOl library provide context for the target sequence(s), allowing users to explore the types of environments in which homologous sequences have been found.

The final step in the VIROME pipeline is to provide characterization of viral ORFs. Since typically less than one-third of viral metagenome library ORFs have a significant hit to a known protein in the UniRef 100 database, a VIROME classification scheme of seven categories was developed to allow investigators to describe viral community diversity. Those ORFs showing significant homology to a known protein within UniRef 100 are classified as either a “Functional protein” or an “Unassigned protein”. VIROME functional proteins must have at least one UniRef 100 homolog with meaningful functional information associated with it in at least one of the six associated annotated databases: the homolog has a GO annotation, belongs to a SEED sub-system, has a KEGG Orthology, has a MEGO annotation, or belongs to a cluster of orthologous groups. VIROME unassigned proteins have at least one UniRef 100 homolog, but that homolog has no meaningful functional annotation associated with it in the six associated annotated databases. Those ORFs showing significant homology only to environmental peptides within MgOl are considered environmental proteins, and are classified as “Viral only”, “Microbial only”, “Top viral hit”, or “Top microbial hit”, based on whether the homologous MgOl peptides originated from viral metagenome libraries, the homologous MgOl peptides originated

from microbial metagenome libraries, the top MgOl hit originated from a viral metagenome library with additional homologous proteins within microbial libraries, or the the top MgOl hit originated from a microbial metagenome library with additional homologous proteins within viral libraries, respectively. Finally, the VIROME predicted ORFs showing no homology to a protein within UniRef 100 or MgOl databases are classified as “ORFans”.

Metagenomes On-line

The Metagenomes On-line resource (<http://metagenomesonline.org>) was developed in response to a need for a protein database to provide environmental evidence of and contextual information about the vast majority of viral metagenomic sequences without significant homologous hits in the UniRef 100 and associated annotation databases. Designed specifically for the VIROME metagenome annotation pipeline but also available as a download for custom analysis, MgOl was originally populated with 137 metagenome libraries. Viral metagenome libraries made up approximately one-third of the total libraries, and were contributed through individual research and through data mining of long read viral genome sequences. Bacterial libraries were added from the CAMERA website (Community CyberInfrastructure for Advanced Microbial Ecology Research and Analysis (Seshadri et al., 2007)) where proteins were identified and from the Venter Institute’s Global Ocean Survey (GOS (Rusch et al., 2007)). Of the initial 137, nine libraries are described as “Eukaryotic” (particles > 1 μm in size), eighty-nine are described as “Microbial” (particles 0.22-1 μm in size), thirty-eight are described as “Viral” (particles < 0.22 μm in size), and one library is described as “Microbial/Eukaryotic” (collected from a 0.22-5 μm fraction). The MgOl database grows in number as publically available libraries with identified

proteins are added to CAMERA and GOS, but that growth is limited as many submitters have not already identified proteins. Submissions of libraries to VIROME for analysis are natural candidates for additions to MgOl once data has been published and becomes public, but not all such libraries are selected for inclusion.

Recognizing that the significance of MgOl's contribution to metagenome analysis and interpretation lay in its ability to provide environmental annotation data that provides VIROME users with additional levels of viral metagenome peptide classification, each library was curated with common-language terms describing a number of the original metagenome sample's environmental features. The interpretation of sequence significance can only be as comprehensive as the metadata that sheds light on its context. Especially with MgOl, where its goal is to provide information on proteins not contained in public databases, environmental context such as the sample collection site, descriptors such as ecosystem or environmental feature, and associated metadata such as sample temperature, pH, and elevation becomes critical to interpretation and significance of sequence analysis. Because of the rich metadata associated with each sequence within MgOl, each predicted ORF passed through the VIROME pipeline may be able to extract biological relevance such as the predominant ecosystems in which the peptide occurs and whether the peptide occurs only in viruses, only in microbes, or in both. The VIROME team developed a form that collects valuable environmental metadata at the time of sequence submission, so that once data is published and publically available, the library and its comprehensive metadata are available for inclusion into MgOl. At the time of MgOl's and VIROME's development a readily available and widely accepted system of environmental annotation did not exist. Consequently, the MgOl database and VIROME submission

form reflect a unique environmental description system. Environments were classified with terms reflecting “genesis” (sample origin, including natural, anthropogenic, or experimental), “sphere” (environmental sector, including aquatic, terrestrial, or organismal), “ecosystem” (selected from a list of anticipated ecosystems or from those requested by submitters), “physical substrate” (physical form of the sample, including sediment, soil, solid substrate, tissue/humor, or water), and “physio-chemical modifiers” (modifiers creating a unique or extreme environment within an ecosystem, such as but not limited to acidic, high temperature, and hypersaline). However, the VIROME team continues to scrutinize its annotation. Text parsing of homolog sequence descriptions can be inaccurate and unreliable, which has guided the development of controlled vocabularies such as the Gene Ontology (GO) and prompted VIROME to classify viral ORFs into VIROME categories according to stringent criterion (*i.e.*, annotation within the GO, KEGG, SEED, PHAGE-SEED, COG, or ACLAME databases). Similarly, common-language environmental descriptions can be variable and misinterpreted across tools, driving VIROME to reevaluate its annotation system and criterion.

Reevaluation of the VIROME and MgOl annotation system involves a review of applicable genomic or metagenomic standards established through the INSDC and GSC, and through investigation of the rapidly evolving Environment Ontology.

International Nucleotide Sequence Database Collaboration

The International Nucleotide Sequence Database Collaboration (INSDC (Cochrane et al., 2010, 2011)) consists of three databases: DNA Data Bank of Japan (DDBJ, www.ddbj.nig.ac.jp (Kaminuma et al., 2011)), the European Nucleotide Archive at the European Molecular Biology Laboratory’s European Bioinformatics

Institute (EMBL-EBI, www.ebi.ac.uk (Leinonen, et al., 2011)), and GenBank at the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/genbank (Benson et al., 2013)). The collaboration developed in response to a need to capture, preserve, and present sequences and their annotation as sequencing technologies emerged and produced massive volumes of data. Consistent data exchange and global access are enabled through standard data formats and annotation conventions. While submission tools and presentation tools are developed and maintained independently among DDBJ, EMBL, and GenBank, searching for an accession number provides the same information and data regardless of the institution. Efforts of member institutions and other advocates of data sharing led to the widely accepted expectation that new sequences described in a publication include INSDC accession numbers. This ‘mandatory submission’ concept is not a strict INSDC standard, but has been widely accepted through the support of INSDC member institutions, INSDC partners, and major life science journal publishers advocating best practices in public data dissemination.

INSDC maintains several collaborative instruments to promote data sharing. The first is the INSDC Feature Table Document (http://www.insdc.org/files/feature_table.html), which describes functional annotation conventions and has become a critical resource for the development of annotation systems based on feature key and qualifier definitions. The second collaborative instrument is the unified accessioning system which universally refers to a given sequence regardless of the query site. A third collaborative instrument is the metadata model that supports the Sequence Read Archive (SRA (Leinonen et al., 2011)), consisting of XML objects defined by individual schema that capture descriptors

related to common (general), study, sample, experiment, run, analysis, and submission metadata. A fourth collaborative instrument is the INSDC status convention (http://www.insdc.org/insdc_status.html) which allows for consistent record availability across partners, allowing for designations such as fully public data, data held confidential until publication, and data suppressed when updated, improved data is available. Finally, the INSDC developed the BioProjects database (Barrett et al., 2012) to gather top-level information that relates multiple studies and records.

VIROME and MgOl already adhere to many of the INSDC collaborative instruments, employing the annotation conventions described in the Feature Table Document, requesting accession numbers from VIROME submitters as part of metadata collection, asking for detailed metadata under many categories represented in the SRA XML documents, and recognizing data as publically available or confidential until publication. Possible considerations as part of this update include the addition of BioProject and BioSample information under the INSDC BioProjects database.

Genomic Standards Consortium

The Genomic Standards Consortium (<http://gensc.org> (Field et al., 2011)) is an international organization formed in 2005 to promote mechanisms that standardize genome descriptions and genomic data exchange and integration. Their mission includes the implementation of new genomic standards, the development of methods to capture and exchange metadata, and the harmonization of metadata collection and analysis efforts. The GSC supports many projects, including the release of standards relevant to the description of genomes, metagenomes, and marker gene sequences. The GSC's interest in creating a genome collection and gaining the most value from that collection by allowing for diverse comparative analysis emphasized the need to

describe each collection as accurately and thoroughly as possible. Community interest supported the call for accurate and comprehensive descriptions for three main reasons. First, interest in testing hypotheses about genomic features through a comparative evo- or eco-genomic approach called for complete information about the nucleic acid sequence source or the environment, respectively. Second, the need to use high-level descriptors of genomes for user- or tool-driven grouping, sorting, and searching emphasized the importance of the collection of standardized descriptors. Finally, the increasing number of environmental shotgun sequencing metagenomic studies demands improved descriptions of genomes and their sources to best interpret the data. For these reasons, the GSC invested efforts into the release of several standards calling for minimum information to be reported with genome, metagenome, and marker gene sequence analyses.

The GSC released two standards in 2008, Minimum Information about a Genome Sequence (MIGS) and Minimum Information about a Metagenome Sequence (MIMS) (Field, 2008). Acknowledging the exponential growth in the quantity of sequencing data, the GSC used the release of these first two standards to promote the standardization of formatting information about genome and metagenome sequences. Both MIGS and MIMS built on reporting information already suggested by the INSDC guidance, and proposed the capture of additional information deemed “minimum” to specific applications and interests. Information was collected in several categories including investigation (general information about the project), environment (sampling site information and environment description), nucleic acid sequence source (primarily information about the genome), and sequencing (methods and assembly methods).

The GSC followed with two additional standards released in 2011, Minimum Information about a MARKer gene Sequence (MIMARKS) and Minimum Information about any (x) Sequence (MIxS) specifications (Yilmaz, Gilbert, et al., 2011; Yilmaz, Kottmann, et al., 2011). MIMARKS provided a standard for reporting marker gene sequences, completing the GSC's intent to release standards relevant to the reporting of genome, metagenome, and marker gene sequences. In addition, the GSC introduced an "environmental package" concept, releasing general environment descriptors with specific sets of associated metadata either required or recommended based on that package. Environmental packages were intended to standardize sets of measurements and observations applicable to particular habitats, and were designed to be utilized across all GSC checklists. Environmental package descriptors were contributed by other working groups most familiar with relevant reporting information: the Human Microbiome Project (www.hmpdacc.org (Turnbaugh et al., 2007)) established packages for host-associated and human-associated environments, the Terragenome Consortium (Vogel et al., 2009) contributed sediment and soil packages, the water package was developed by a collaboration among International Census of Marine Microbes (ICoMM, www.icomm.mbl.edu), Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites (MIRADA-LTERS, <http://amarallab.mbl.edu/index.html>), and the Max Planck Institute for Marine Microbiology (<http://www.mpi-bremen.de/en/>), and the MIMARKS working group developed the remaining packages. Finally, the GSC released the MIxS standard to promote an overall framework for reporting sequence data. MIxS includes the specific checklists from MIGS, MIMS, and MIMARKS,

allows for possible introduction of additional checklists, and allows for the annotation of data using environmental packages.

The INSDC recognized the GSC as an authority for the MIxS standard. Various INSDC members accept MIxS metadata and provide web forms to gather and validate MIxS-compliant metadata fields. The Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata. Other tools have been developed to help users to gather and submit MIxS-compliant data, including MetaBar (Hankeln et al., 2010) and CDinFusion (Hankeln et al., 2011). The Quantitative Insights Into Microbial Ecology (QIIME (Caporaso et al., 2011)) tool generates and validates MIxS-compliant templates. The MIxS single point-of-entry to GSC standards and the underlying MIGS, MIMS, and MIMARKS standards are widely accepted and clear starting points for guiding the collection and display of compliant metadata in VIROME and MgOl. While VIROME and MgOl collected a wide range of metadata at their start, aligning their environmental descriptors with GSC standards remains a key target for improvement. Specifically, the incorporation of environmental packages and the gathering of associated metadata is an area of focus. Part of the environmental package concept is the use of specific and standardized terms to describe the environment, as provided by the Environment Ontology (EnvO (Buttigieg, Morrison, Smith, Mungall, & Lewis, 2013)).

Environment Ontology

As the call for more comprehensive annotation of environmental sources of genome and metagenome sequences increased, the need for a standardized system of describing those environments grew in response. The Environment Ontology (www.environmentontology.org) is a community-led project which seeks to describe

the environment of any organism or biological sample, regardless of application. EnvO is meant to be used across various disciplines, including museum and tissue collections as well as metagenomic samples. Employing a concise and standardized environment description enables the integration and grouping, sorting, and binning of environmental data.

A comprehensive description of an environment using the Environment Ontology will include at least one term from each of three hierarchies, appreciating the many layers of an environment that influence and shape a sample's origin. EnvO's biome, environmental feature, and environmental material classes allow for the non-redundant description of an environment while capturing its qualities in complementary scopes. Biome terms identify the ecosystem from which a sample was collected. Biomes are community-centric environments, defined by the presence of communities that have adapted to it, such as coniferous forest biome or tundra biome. Environmental features are considered to be single entities or features that strongly influence an environment, without any specific reference to ecological communities as in a biome. Examples of environmental features include a coral reef or a prairie. Finally, an environmental material is a mass, volume, or portion of an environmental system, including terms such as air, water, or soil. Additional hierarchies exist within EnvO, including habitats, but a comprehensive annotation is currently recommended by including a term from the biome, feature, and material hierarchies.

EnvO has become widely accepted thanks to its interoperability with ontologies compliant with Open Biomedical and Biological Ontologies (OBO) Foundry principles and its alignment to the Basic Formal Ontology (BFO). The GSC's MIXS standards employs the use of EnvO biome, feature, and material terms to

annotate an environment. EnvO has also been adopted by other organizations and tools and lists many of these on its website at www.environmentontology.org/users, including additional ontologies, the National Center for Biomedical Ontology (NCBO, www.bioontology.org (Rubin et al., 2006)), the International Census of Marine Microbes (ICoMM), and the ISA software suite (Rocca-Serra et al., 2010).

VIROME and MgOl recognized the importance of environmental annotation, but in response to a lack of any comprehensive ontology at the time of their development, created their own unique set of environmental descriptors. While connections can be drawn between VIROME's existing genesis, sphere, ecosystem, and physical substrate descriptors to more universally accepted EnvO terms, the collection of EnvO terms at the time of submission is key to the development of a more visible, interactive, and compatible tool and database.

Evaluation of VIROME, MgOl, the INSDC, applicable GSC standards, and the Environment Ontology contributes to the three aims of this project: retrofitting the current MgOl database, collecting compliant metadata with future VIROME submissions, and leveraging the metadata in scientifically relevant and novel ways.

Chapter 3

AIM 1: FRAMEWORK TO RETROFIT VIROME AND METAGENOMES ONLINE TO REFLECT METADATA STANDARDS

This project explores various standards, practices, and ontologies appropriate to environmental annotation of sequencing data in order to update VIROME and MgOl in ecologically and scientifically relevant ways. After evaluating relevant standards, this project pursued three aims. Aim 1 designs a framework to retrofit the current VIROME tool and Metagenomes On-line database to reflect applicable metadata standards. This aim is achieved through designing modifications to the current MgOl schema, retrofitting existing libraries to fit the new schema, and implementing those database structure changes.

The VIROME tool and MgOl database were implemented to help users gain insight into viral metagenome datasets, recognizing the importance of comprehensive sequence annotation in order to allow full analysis and well-supported conclusions. As such, VIROME and MgOl align well with applicable INSDC, GSC, and EnvO standards. Users are asked for accession numbers for data that have already been submitted to INSDC, and honor privacy settings such as public or confidential data. Users are prompted for metadata to describe the overall project, environment, sequence source, and sequencing strategies, in following with the MIxS recommendations. VIROME collects appropriate environmental metadata such as temperature, elevation, salinity, and pH, and already prompts users for more information than is considered mandatory by MIxS environmental package guidelines.

Environment Ontology terms were used to manually annotate many of the original VIROME libraries, though EnvO has evolved and changed since that time without those changes being reflected in the MgOl database. As such, existing EnvO annotations are no longer informative and are in need of update.

Recommended changes or updates to the current schema are based upon alignment with INSDC, GSC, and Environment Ontology standards. Since VIROME's release, INSDC has developed BioProject and BioSample databases which should be reflected in the MgOl database. Compliant MIxS annotation also includes the selection of an environmental package and the reporting of associated metadata, and the adoption of this practice is integral to VIROME and MgOl updates. Finally, EnvO biome, feature, and material terms must be included in library descriptions.

Such revisions to VIROME and MgOl do have an associated cost. There are drawbacks to the evolving EnvO framework, and there is a great investment of personnel time and effort required to revise VIROME and MgOl. EnvO has changed significantly since VIROME's development in its overall hierarchy and structure, in its lists of current terms and definitions, and in its annotation recommendations. There are still limitations in the EnvO framework that indicate that future work may still be needed to keep VIROME and MgOl current as EnvO changes. One limitation is that the EnvO structure is not parallel or uniform, as is the case with other defined ontologies. Layers or tiers in the environment ontology have highly variable levels of specificity, so that using a identifying a subset of EnvO for VIROME and MgOl classification purposes is not appropriate and users are required to have a good understanding of the EnvO structure and annotation guidelines in order to correctly

describe their sequences. In addition, EnvO's method of capturing extreme environments or physical-chemical modifiers is incomplete compared to VIROME and MgOl's current system for capturing this information. After reviewing EnvO, the closest terms to physio-chemical modifiers such as acidic, alkaline, arid, high temperature, hypersaline, etc., are in the habitat hierarchy. The habitat list, however, does not capture all relevant modifiers. Other branches of the EnvO ontology can be employed to capture a relevant modifier (such as "arid", a "condition" and part of EnvO's comparatively underdeveloped environmental condition hierarchy), or other ontologies can be used to describe a modifier (such as "anoxic", a quality and part of the separate Phenotypic Quality Ontology (PATO, http://wiki.obofoundry.org/wiki/index.php/PATO:Main_Page)). Such physio-chemical modifiers have been used in VIROME and MgOl for grouping and comparative purposes but are not currently represented consistently in EnvO, nor are they part of the recommended minimum annotation. Finally, the personnel time and investment warranted by comprehensive tool and database revisions are a potentially significant drawback to updating VIROME and MgOl. Modifications to the current database schema will involve retrofitting existing libraries and requires extensive validation to maintain database integrity. Revisions to the VIROME submission form organization and content must maintain the form's ease of use while capturing comprehensive metadata in order to best leverage the tool's capabilities as sequencing and analyses tool demands evolve.

Despite potential pitfalls, the advantages of moving VIROME and MgOl toward INSDC, MIxS, and EnvO standards are significant. With relatively minor changes to MgOl's overall content, VIROME has the potential for increased validity,

visibility, collaboration, and user preference. First, VIROME and MgOl already prompt users for detailed metadata such that the database is already populated with descriptors such as sample location, environment descriptors, and associated measurements (*e.g.*, altitude, temperature, and pH). The assignment of EnvO terms to existing libraries can be made based upon current descriptors without the need to mine additional information from literature or other sources. Second, the widespread acceptance of INSDC and MIxS standards has primarily been voluntary and community-driven. Well-reviewed public journals expect INSDC accession number(s) as part of public data dissemination (Cochrane et al., 2011), and INSDC embraces the MIxS standard to set annotation requirements for sequences. Therefore, the adoption of those standards lends credibility to VIROME's submission process, and has the potential to increase the validity and longevity of VIROME. Additionally, there may be increased visibility of and potential collaboration with VIROME as the GSC and EnvO websites maintain lists of current adopters. Not only are those lists a valuable opportunity for the VIROME tool to be explored by potential users, but cooperative data exchange is enabled through the use of common and standardized descriptors. Annotating MgOl libraries in accordance with MIxS and describing environments in alignment with EnvO recommendations makes the MgOl database and user-driven analysis easier to leverage across research interests. Finally, creating opportunities to make use of enhanced metadata collection can create user preference for VIROME, not only in its ability to group, sort, and explore data based on common environmental occurrences but in the tool's capacity to guide appropriate metadata collection and provide it in a format that encourages users to submit their data to other public databases.

Design Modifications to the Current Database Schema

The current MgOl database schema can be divided into four general categories based on the MIxS areas of scope – investigation, environment, nucleic acid sequence source, and sequencing (Figure 1).

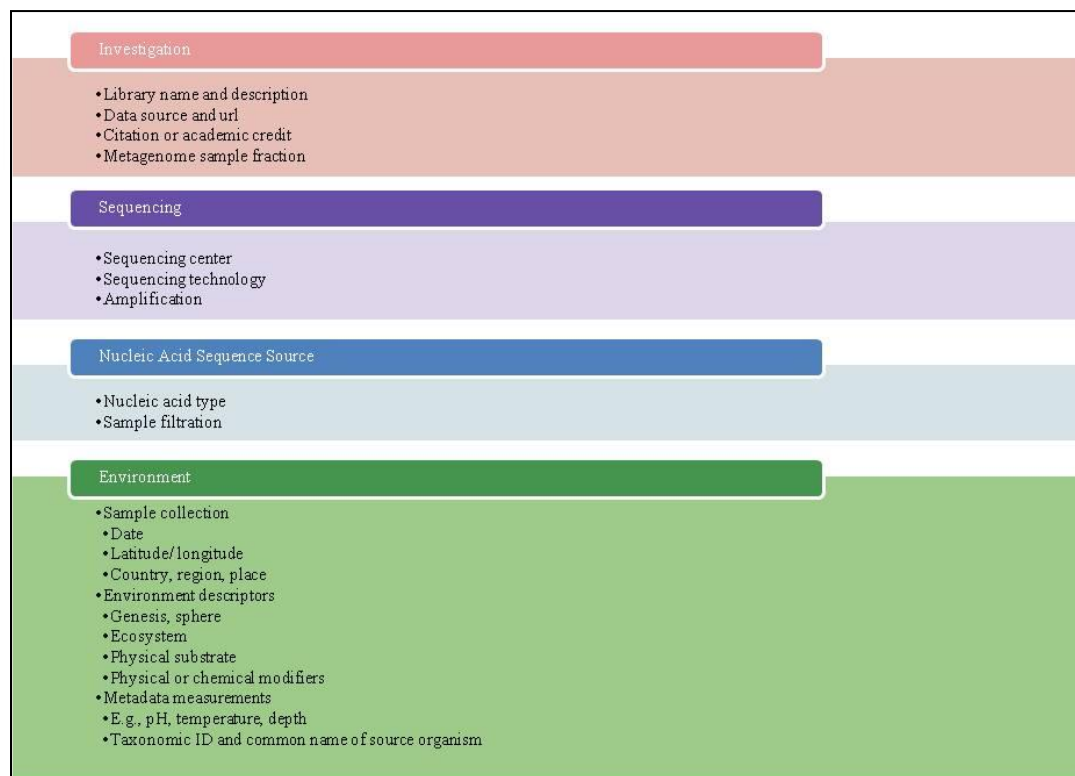


Figure 1 MgOl database fields classified in MIxS-compatible sections

Investigation captures general information about the library, including the library name and description, the library type (for VIROME purposes, the fraction of the metagenome sample that was analyzed – viral, microbial, or eukaryotic based on

filter size), an assigned unique prefix and MgOl ID, the NCBI project and accession number if applicable, and the data source, URL, and citation. Environment holds information regarding the sample date; sample site latitude, longitude, ID, country, region, and place; environment descriptors including genesis, sphere, ecosystem, physical substrate, and physio-chemical modifiers; host information including the taxonomic and common names, if applicable; associated environmental measurements such as altitude, depth, pH, salinity, or temperature; and in some cases VIROME-assigned values for the depth zone or climate zone. Nucleic acid sequence source includes the nucleic acid type (*e.g.*, DNA, double stranded DNA, single stranded DNA, or RNA). Finally, sequencing information includes the sequencing type and center, sequencing release date, amplification type, lower/upper filter sizes if applicable, and in some cases the average read length or GC percentage (calculated by VIROME based on the submitted sequence file).

Evaluation of the MIxS guidance revealed that a while VIROME and MgOl capture most of the mandatory data fields, a great deal of additional data is outlined in MIxS and considered optional. This update does not include the development of an exhaustive metadata tool. Much of the metadata is not relevant to VIROME and the analysis it provides, the metadata is not mandatory for a MIxS-compliant submission, and the quantity of metadata could be overwhelming to a user, preventing them from submitting their sequences through VIROME. Evaluation of the EnvO hierarchies revealed an inconsistent data structure, with varying levels of specificity among tiers within the hierarchy. With a goal of useful grouping, binning, and sorting based on environment metadata or descriptors, this inconsistent data structure is problematic. Not only would child terms need to be traced back to an appropriate parent term for

grouping, but the selection of that appropriate parent term is unclear since it cannot be generally selected based on a given tier in the EnvO hierarchy. Accurate selection of appropriate parent binning terms would require manual annotation of each submitted library if users were permitted to navigate the entire EnvO hierarchy. In addition, for a user unfamiliar with EnvO terminology, the selection of terms from the biome, feature, and material hierarchies may not only be overwhelming in terms of learning, but may also lead to inaccuracies in the final annotation. For example, searching or browsing within EnvO (<http://www.environmentontology.org/Browse-EnvO>) for “ocean” displays the EnvO feature term “ocean” (ENVO:00000015) with child terms including “ocean water.” A user may decide to select “ocean water” (ENVO:00002151) as a more descriptive term, but that term is actually a part of the environmental material hierarchy. Users may easily and inadvertently annotate their sequences inappropriately, not choosing at least one term from each of the biome, feature, and material hierarchies according to MIxS guidance. Based on these observations, recommendations for modifications to the database include the continued use of VIROME- and MgOl-specific environment terms such as genesis, sphere, ecosystem, physical substrate, and physio-chemical modifiers. The VIROME and MgOl environment descriptors are easy to interpret, are comprehensive enough to allow for assignment of standardized EnvO terms, and allow users to make their selections within the VIROME submission form without redirection to an EnvO source or explanation. Based on user selection within those fields, EnvO terms will be automatically designated. In this way, a specific set of MIxS-compliant and high level EnvO terms can be assigned that facilitate useful and appropriate binning, sorting, and searching. Users will have the option to specify additional terms for a more

comprehensive annotation, but tool-driven comparisons can be made against the common subset of assigned EnvO terms.

Based on a review of VIROME and MgOI's current approach and relevant metadata standards, the following database schema modifications are recommended which involve creating, modifying, or removing fields (Figure 2).

Modified Fields		Removed Fields		
Expand controlled vocabulary:		• BiomeEnvO	• EnvO matter	• shortname
• Ecosystem		• Biome description	• EnvO food	• FASTA_nt
• Physical substrate		• BiomeEnvO number	• EnvO annotator	• FASTA_pep
• Physio-chemical modifiers		• EnvO geographic feature	• URL	• chlorophyll
• Sequencing		• EnvO habitat feature	• citation PDF	• altitude zone
		• EnvO mesoscopic geo feature	• NCBI source	• release date

Created Fields			
Investigation	Sequencing	Nucleic Acid Sequence Source	
• Investigation type	• Assembly details	• Number of replicons	• Isolation or growth conditions
• BioProject	• Sequencing details	• Pathogenicity	• Sample collection, processing, and size
• BioSample	• Fragment size	• Biotic relationship	
		• Propagation	

Environment			
Sample Collection	Env Description	Env Metadata	EnvO Terms
• Pooling description	• Environmental package	• Elevation • Conductivity • Density • Humidity • Subject identifier • Built environment	assigned ID, name & user-selected ID, name for: • Biome • Environmental feature • Environmental material • Habitat

Figure 2 Summary of recommended modifications to MgOI database schema

Fields for general investigation information were created in accordance with the INSDC's latest guidance, and contain the investigation type (specifying a genome or metagenome submission), BioProject number, and BioSample number. New environmental fields must minimally include the environmental package and EnvO terms for biome, feature, and material in alignment with MIxS guidance. Those recommended fields are environmental package, EnvO ID number and name for biome, feature, and material, EnvO ID number and name for habitat (to standardize the reporting of physio-chemical modifiers), EnvO ID number and name as selected by the user for biome, feature, material, and habitat (to be discussed in more detail with submission form updates), and additional metadata fields (elevation, conductivity, density, humidity, subject identifier, and additional fields specific to and required for the built environment package). The additional metadata fields were selected based on evaluation of the MIxS environmental package guidelines, and were included either because the fields were mandatory (those fields related to a built environment) or conditional for an environmental package (elevation, density, humidity, and subject identifier). New fields for nucleic acid sequence source information include several fields related to bacterial genomes (number of replicons, pathogenicity, biotic relationship, and isolation or growth conditions), viral genomes (pathogenicity, propagation, and isolation or growth conditions), or metagenomes (sample collection, processing, and size). These fields were selected for addition since they were either mandatory or conditional based on the investigation type according to the MIxS standard. New fields for sample collection and sequencing information include free text fields for pooling description, assembly details, and sequencing details so that pertinent information can be captured and shared on the library

summary page. Finally, new fields for sequencing include the fragment size based on VIROME team recommendations as this is critical information for subsequent metagenome assembly.

The decision to include or exclude MIxS fields in the updating schema was handled with a great deal of consideration. Initial discussion of this thesis project included the suggestion to create a more universal submission tool. The team considered creating a universal submission tool that would be was easy and intuitive to use, help the user annotate with EnvO terms, and provide an output file compatible for INSDC database submission. These factors would ideally increase traffic to the VIROME portal and therefore also increase relevant viral genome or metagenome submissions. This submission tool would accommodate all MIxS submission types (*i.e.*, eukaryotic, plasmid, or marker genes submissions) regardless of their contribution to VIROME's mission and intent to provide a viral-centric tool for metagenomic exploration. Therefore, fields common across all environmental packages were considered for inclusion (organism count, oxygenation status of the sample, perturbation, sample storage duration, location, and temperature, and the sample volume or weight used for DNA extraction). However, review of the draft submission form prompted additional discussion. The additional common fields doubled the list of requested metadata in many environmental packages and made that step in the submission form appear much more arduous, while providing no gain toward VIROME analysis or MgOL's contribution of environmental occurrence and corresponding environmentally-focused metadata. Initially, consideration was given to listing only those most relevant fields on a first metadata page and providing a link the submitter could choose to enter "fully MIxS-compliant" metadata. However, the

common fields under consideration were in no way the only fields that would appear in a fully MIxS-compliant list. For example, the water environmental package has 109 environmental metadata fields in the MIxS guidance document. The VIROME team decided to prompt only those metadata fields already in the MgOl database (altitude, pressure, depth, total depth, pH, salinity, temperature, biomass, chlorophyll, DIC, DIP, DOC, DO, NO₃, taxonomic name, and common name) and the additional fields found to be most capable of contribution to the analysis (elevation, conductivity, density, and humidity), most relevant to the environmental package (subject ID), or mandatory according to MIxS (those fields related to the built environment environmental package). Additional MIxS-compliant fields did not add further value to VIROME or MgOl, but represented a significant resource cost to draft a submission form that captured those fields in an easy and meaningful way. The VIROME team also concluded that pursuit of a universal submission tool was not appropriate for this update version as it would provide nothing to our target analyses.

Particular fields were modified in order to align them with MIxS and other relevant standards. The ecosystem field is expanded to include additional terms for user selection, based on an evaluation of EnvO biome terms and the development of a more comprehensive ecosystem subset. Terms for forest and tundra are added to the ecosystem subset, and the term for soil is removed (“soil” is representative of a physical substrate or environmental material rather than an ecosystem). While “subterranean” was initially removed, the term was ultimately included in the ecosystem subset. EnvO considers “subterrestrial” to be a habitat descriptor, identifying the characteristics of a below-ground environment as particular environmental qualities below a surface ecosystem. However, as subterranean samples

may have little resemblance to their surface-level ecosystems, VIROME kept this term as an ecosystem and will reach out to EnvO to request that “subterranean” be defined within the biome and feature hierarchies. Additionally, the subset of terms for physical substrate is expanded to include all first tier terms in EnvO’s environmental material hierarchy. Since that first tier is comprehensive, including all of VIROME’s and MgOl’s original list of physical substrates plus many additional terms, the entire first tier was selected for inclusion in the update so that criteria for creating a subset were not needed. In addition, the physio-chemical modifier field is modified by removing multi-modifier terms (for example, “Acidic High Temperature” is removed from the field’s subset), by becoming a select-multiple field rather than a select-one (so that the user can select both “Acidic” and “High Temperature” to recreate that removed field, or can create any other combination of modifiers). Initially, the subset of modifiers was expanded to include all terms currently listed in EnvO’s habitat hierarchy (including terms such as arboreal, aquatic, and endolithic). However, the physio-chemical modifiers in VIROME are intended to identify particular or qualities that may characterize an environment as extreme. While the previous list of modifiers mapped well to the EnvO hierarchy, the final recommendation of this update is to maintain a subset of extreme environment modifiers rather than create a list inclusive of the EnvO habitat hierarchy. Finally, the sequencing field is conditionally modified to a select-multiple field dependent on assembly. If data is unassembled, Illumina and Ion-torrent technologies are removed from the list as their unassembled reads are not suitable for the VIROME pipeline, and only one sequencing type may be selected from the drop-down list. If data is assembled, Illumina and Ion-Torrent are active

terms in the drop-down list, and multiple sequencing types may be selected to accommodate hybrid assemblies built across multiple sequencing platforms.

Fields recommended for removal from VIROME and MgOl include those related to the original EnvO terms, unused or empty fields, duplicative fields, and fields related to the original INSDC project terms. While EnvO terms are part of updates to these resources, the original field names and contents are outdated based on the evolution of and changes to EnvO since VIROME and MgOl were developed (EnvO biome, biome description, biome number, geographic feature, habitat feature, mesoscopic geographic feature, matter, and food). Several fields in MgOl remain unused including url, an attached PDF version of a citation, and altitude zone. While the url and PDF citation fields are not related to environment descriptor updates, they are removed during these changes as part of routine evaluation and maintenance. The altitude zone is an environment-related field, but it is also one that may have little contribution to VIROME's analysis and MgOl's contribution. The altitude zone historically was left incomplete by individual submitters and required manual curation, and is difficult to automatically assign since it is not an objective classification based only on elevation. Based on these considerations, the altitude zone is removed from the MgOl schema as part of this update. In addition, two chlorophyll measurement fields are present in the current database. One of those duplicative fields is removed as part of this update. Finally, while the contents of fields related to original INSDC terms (project and accession) need to be moved to newly created fields for BioProject, those original fields will be removed during this update.

While the previous schema for the MgOl database was documented using simple Navicat database analysis and in-field enumeration/comments, the new

database structure and its adherence to public guidelines requires a more detailed database checklist containing the structured comment name, full name of the item, definition, expected value, corresponding section, and value syntax (Appendix A). The checklist also provides definitions or criteria for field enumeration values (*e.g.*, physio-chemical modifiers such as “acidic” where pH is less than 3, or “haline” where salinity is greater than 2 M) (Appendix B). All database fields were renamed to match the submission form field names, clarifying the relationship between the two fields.

Retrofit Existing Libraries to Fit New Schema

Existing MgOl libraries must be updated to fit the new schema. While some fields are not considered mandatory for VIROME or MgOl purposes (new fields related to the nucleic acid sequence source) and may be left blank, other fields must be completed. Rules and/or logic constructed to guide this process not only reduce personnel time that may be required to review each individual library, but allow for a more subjective and comprehensive validation of the update. Guidance below is provided for investigation, environment, nucleic acid sequence source, and sequencing fields.

Modifications to investigation fields include the addition of an investigation type field and the transition from NCBI parent project identifiers to the INSDC’s recent BioProject and BioSample database identifiers. To date, all MgOl libraries are metagenome samples. Their identification as “viral”, “microbial”, “eukaryotic”, or “microbial/eukaryotic” is assigned on user selection in a library type field, equivalent to the metagenomic fraction of the sample based on sample filtering at the viral (<0.22 µm), microbial (0.2-3 µm) or eukaryotic (>3 µm) level. Therefore, all libraries in MgOl can be assigned as “metagenome” libraries. The investigation type field is

added as part of this database update to allow for submission of bacterial or viral genomes to VIROME, or for the addition of those genomes to the MgOl database for a more comprehensive analysis of environmental occurrence of submitted sequences.

The greatest effort in retrofitting existing libraries is required to update environmental fields. With MgOl's comprehensive environmental annotation, it is possible to assign specific EnvO terms when manual review of each library is able to consider the library name, geographic place, ecosystem, physio-chemical modifiers, site ID, and the full range of recorded environmental parameters. However, the intent is to assign one general term from each of the biome, feature, and material EnvO hierarchies to allow for broader grouping and searching within MgOl. Therefore, logic connecting the existing MgOl fields to the new EnvO fields was developed (Figure 3). The ecosystem field is evaluated in order to assign both the representative biome and feature terms. Only when "Agricultural" is selected as an ecosystem is additional information needed to assign a biome. In that case, the genesis field must be considered and used to assign biome. The physical substrate field is used to assign an EnvO material. The sphere field in VIROME is evaluated, and only if the library is designated as "Organismal" does sphere play a role in assigning EnvO terms – in that case, any feature assignment would be overridden and classified as "organic material" and any conflicting material assignment would be overridden and classified as "organic matter". Finally, the physio-chemical modifiers would be evaluated and assigned one or more EnvO habitat terms. In cases where VIROME curators felt more specificity would benefit the library description, additional and more specific terms could be added in `envo_user` fields. For example, libraries from a whalefall metagenome analysis would be automatically annotated with EnvO terms

ENVO:01000033 oceanic pelagic zone (biome), ENVO:01000062 organic matter (feature), and ENVO:01000155 organic matter (material); but could be manually annotated by adding term ENVO:01000140 whale fall (feature). Note that metadata fields for the taxonomic and common name of the sample's host would capture the whale species, highlighting that a comprehensive annotation is essential for drawing accurate and meaningful conclusions.

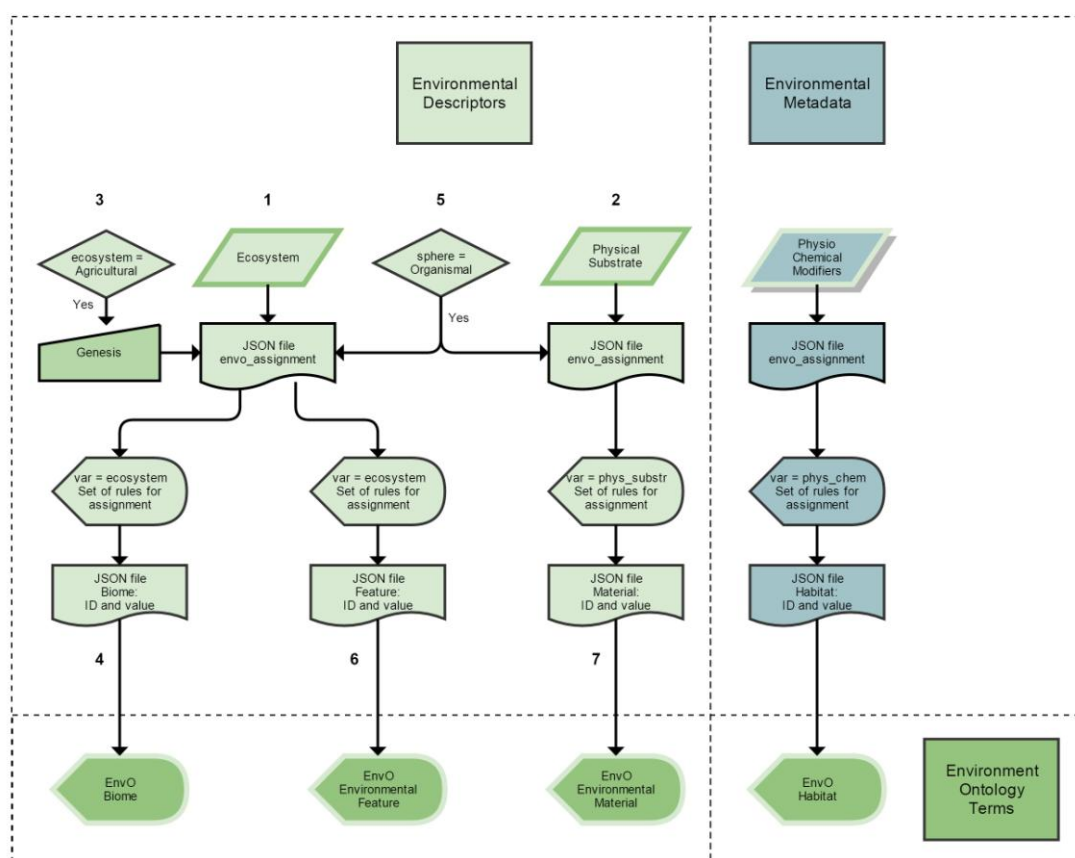


Figure 3 Logic used to assign EnvO terms to retrofit existing MgOl libraries to the new schema

Two examples demonstrate the need for objective logic in the assignment of EnvO terms. First, in a library submitted from Octopus Hot Springs in Yellowstone National Park, manual evaluation and a search on the site name may lead to an environmental feature annotation as ENVO:00002119 “alkaline hot spring”. However, that assignment would mean that the library lost its classification as a “spring” and therefore lost its association with other springs for grouping and searching purposes. This disassociation is a product of the variable tier structure within EnvO – aside from selecting specific target terms there is no appropriate way to select one level within an EnvO hierarchy and trace all child terms back to that level for grouping. Instead, this library would be annotated as ENVO:00000027 “spring” and could be additionally annotated by the submitter or curator as an alkaline hot spring in the EnvO_user fields. In a second example, a library submitted from the Caribbean Sea may be classified by a user considering the place name as environmental material ENVO:00002149 “sea water”. That assignment would differentiate it from other water samples the library would be considered a singleton for grouping purposes. Instead, this library would be annotated based on its physical substrate, water, as ENVO:00002006 “water” and could be additionally annotated by the submitter or curator as saline water, coastal water, and/or sea water, for greater specificity and clarity.

Additional retrofitting of existing libraries is not required. Other new fields include several for nucleic acid sequence source information, and sequencing fragment size. Since all existing libraries are metagenomes, the only relevant nucleic acid source fields are sample collection methods, sample processing, and sample size. While those fields can provide valuable information on a particular library, they also are not likely to drive grouping and searching functions, and can be left blank. For

existing MgOl libraries, the fragment size field will also be left blank as that field is not relevant to environmental annotation updates and would require significant resources to accurately annotate.

Implement Database Structure Changes

Implementing the proposed database schema first involves adding or removing the appropriate fields. This step will remove the outdated EnvO fields and their contents. One possibility is to add the new EnvO fields first and transfer the current descriptors to the user-entered EnvO terms fields. However, many of the current descriptors are inactive or retired. Keeping the retired descriptors is not considered necessary since many libraries were annotated by a member of the VIROME team and not by the original author, and since the goal of this project is to update the VIROME tool and MgOl database with current environment descriptors in accordance with standards and ontologies. To identify retired terms, script can dissect the environment ontology's available OBO file, extract key:value pairs of EnvO numbers and descriptors, and search the current EnvO number keys against that list to identify those numbers that do not appear in the list (retired EnvO numbers). Those retired terms could be eliminated from the transfer and all current terms added. In some cases, an alternative term is proposed by EnvO as a replacement, but those terms are not necessarily appropriate matches and require manual evaluation (*e.g.*, the retired "marine polar biome" may consider "polar desert biome" as a replacement). In addition, the existing feature and matter EnvO descriptors in MgOl are numbers only without the corresponding name. The same key:value pair list from the EnvO OBO file could be utilized to extract the corresponding name descriptor and import it into the updated MgOl database in the appropriate user-entered EnvO term field. Rather

than transferring existing EnvO terms, the second alternative is to eliminate those EnvO annotations from the implementation plan. This option would ensure that no retired terms are transferred into the updated database, but does eliminate annotation that may have come from the original author in a paper or other corresponding documentation. Despite the fact that the user-entered EnvO terms will not be used for grouping, sorting, or searching purposes, the loss of author-entered annotations or a more comprehensive library description is not preferred. However, the evaluation of EnvO terms has the potential to be a labor-intensive component of the database transfer. Therefore, the VIROME team elected not to transfer any of the existing EnvO annotations after considering personnel efforts, the potential for capturing outdated EnvO terms, and the lack of gain in the VIROME analysis. However, the automatic assignment of EnvO terms to each library based on the terms selected for ecosystem, physical substrate, sphere, and physio/chemical modifiers will populate the new EnvO fields, and those assigned terms will be utilized for grouping, sorting, and searching within VIROME. Note that the existing physio/chemical modifier terms must be broken into their component parts first (“Acidic High Temperature” must become two terms, “acidic” and “high temperature”), and be stored in the database in a multiple-value field.

Implementation of the new MgOI schema was handled by creating a new MySQL table using Navicat (www.navicat.com). All field types were matched to the updated submission form field parameters (*e.g.*, length of varchar fields. Submission form updates are discussed in Aim 2). Fields were not identified as enumeration fields with a particular list of acceptable terms since those lists are defined in the submission form scripts. Creating an enumeration field in MySQL only creates an additional

source of error as descriptors' list of terms are updated (*e.g.*, as additional ecosystem terms are added to the appropriate drop-down field in the submission form) since changes must be reflected in multiple places. Migration of existing libraries to the new table will be handled after initial validation of particular test libraries and new submissions.

The second component of the database schema implementation involves updating database statistics scripts. Statistics scripts retrieve and present library information from BLAST results, and must be fluid to accommodate a dynamic database. In addition, statistics scripts must allow for data exploration based on many different criteria across each of the MIxS metadata categories of investigation, environment, nucleic acid sequence source, and sequencing. For example, one must be able to search for a metagenome sample (investigation type), refine by selecting for a water sample (environmental package or environmental material), and further refine by selecting for a coral reef sample (ecosystem, biome, or environmental feature). Ideally, users would also be able to search by particular metadata, adding parameters such as pH or temperature to their search criteria. Full exploration of the data would allow users to search based on almost any of the available fields, including nucleic acid type or sequencing technology. Note that script updates must include the ability to extract single terms from a multiple term field, such as physio-chemical modifiers. Select-multiple fields after all recommended changes are physio-chemical modifiers, EnvO habitat, user-entered EnvO terms, and sequencing method.

Chapter 4

AIM 2: IMPLEMENT IMPROVED VIROME INTERFACE TO COLLECT COMPLIANT METADATA FOR FUTURE SUBMISSIONS

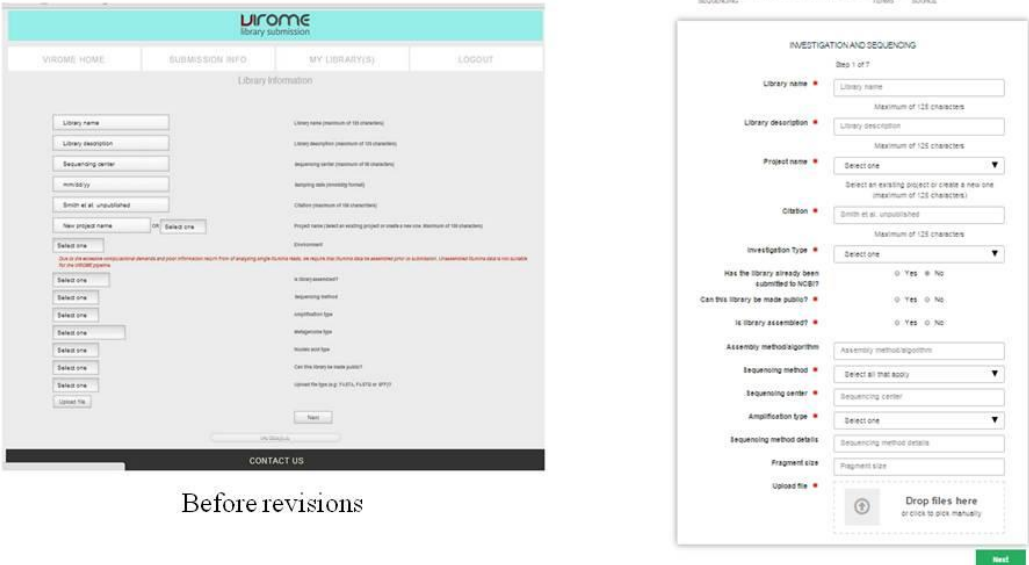
Updating VIROME and MgOl in alignment with applicable standards, practices, and ontologies must address the adaptation of the current systems, the advancement of current tools in compliant ways, and the expansion of the tools' capabilities to best leverage the updates. Aim 2 designs and implements a new VIROME interface and submission form in order to collect compliant metadata for future submissions. This aim is accomplished through revising the VIROME submission organization, drafting scripts to automatically assign EnvO terms and other classifications appropriate for grouping and sorting, and developing a dynamic and adaptive process to facilitate a comprehensive submission.

Revise VIROME Library Submission Views

Recommended updates to the MgOl database schema include the addition of new fields, the classification of fields into MIxS-compliant categories, and the environmental annotation of libraries using Environment Ontology terms. The VIROME library submission form must also adapt to these changes. Primary goals of the revision are to 1) improve the readability and appeal of the submission form by organizing prompts and fields to read from left to right, 2) organize fields into views corresponding to the MIxS categories of investigation, environment, nucleic acid sequence source, and sequencing, and 3) improve the readability of the submission form and underlying code.

Readability

The first component of the revision is to improve the visual appeal of the submission form by organizing user prompts/questions and the corresponding entry fields. An almost immediate (within 500 ms) assessment of a website's visual appeal is used to judge the value of the site's information or credibility (Technologies, 2007). This assessment is often made by moving through page content from left to right and top to bottom (George, 2005). Prompts were reorganized to read from left to right, with the description/question on the left and user entry field on the right (Figure 4).



Before revisions

After revisions

Figure 4 Modifications to VIROME submission form view organization. Changes include organizing information to read from left to right, enlarging fonts for readability, and grouping fields by their relevant MIxS-compliant categories.

View organization

The VIROME submission form was originally segmented into various views in order to limit the amount of information presented to a user on a single screen. This format is maintained through the revision, but the information is regrouped so that fields are displayed by the MIxS-compliant categories of investigation, sequencing, nucleic acid sequence source, and environment. The form was divided into six steps with a seventh step that displays all fields as a summary for review (Figure 5). For a detailed overview of the submission form, see Appendix C.

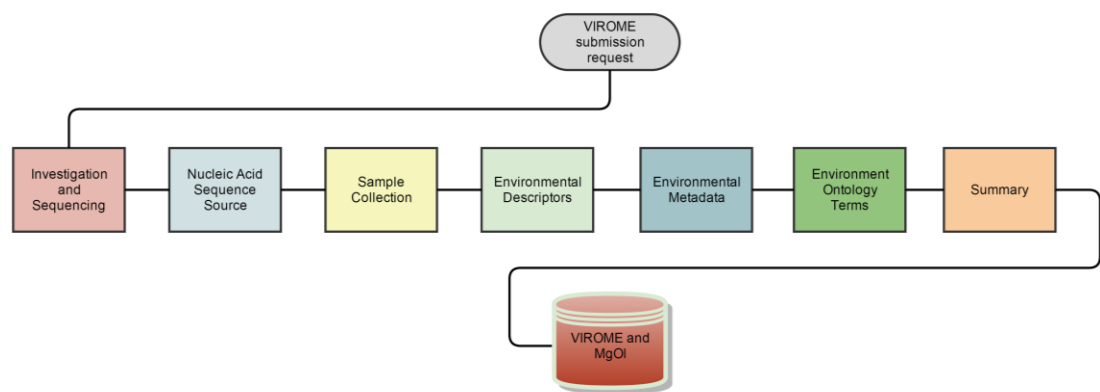


Figure 5 Broad organization of VIROME submission form views. Views are arranged by MIxS-compliant categories and include a summary step.

Each page requires validation of user responses before progression to the next page. For example, required fields cannot be left blank when selecting “Next”. While this means users do not have the ability to preview the entire form, the VIROME team felt that intermediate validity checks during the submission process would ensure that

all mandatory fields were completed and would create a faster submission after the user's final review and approval. Each page requires both "Next" and "Back" buttons so that the user can navigate through sequential pages once each page is complete. In addition, page "Edit" functionality was added to the summary page so that the user can go directly to a particular section and edit information as necessary.

Additional revisions addressed upgrades to the arrangement and order of prompts within the form to improve flow, simplify the overall process, and enhance submitter experience. The project name field was simplified into a single drop-down menu listing the option to create a new project and any existing projects owned by the user; when the 'create' option was selected a text box will appear to capture the new name. Sequencing information is collected on the first page with Investigation fields so that any files unsuitable for the VIROME pipeline can be identified early in the process. The prompt for assembly of the submission was moved to the top of the sequencing section so that assembled libraries can select one or more sequencing technologies, and that unassembled libraries can select only one sequencing technology and cannot select either Illumina or ion torrent reads which are currently inappropriate for the VIROME pipeline because of the massive computational demands and poor information return from the analysis of single reads. The prompt relevant to pooling of samples was relocated to the top of the environmental section, so that fields relevant to pooling would conditionally appear downstream along with appropriate instructions to describe a single, representative sample from the pool. The decision was made to only collect one set of environmental data and associated metadata instead of possible multiple sets when dealing with a pooled submission, as a more exhaustive description of a pooled sample requires an exponential increase in the

amount of metadata (*i.e.* duplication of nearly the entire form for each pooled sample. This amount of complexity would not only be burdensome for implementation, but would be very taxing on the end-user.

Improve readability of submission form and underlying code

Final organization revisions targeted appeal of the viewable web submission form and the readability of the underlying code. Font sizes were increased throughout the submission form to improve the submitter's experience. Arrows were added to drop-down menu displays to differentiate their appearance from free text fields. Help blocks were created to hold helpful information such as an example entry or the amount of available characters, so that the information was removed from the prompt displays and the form appearance simplified. The form's final summary page was broken into sections and labeled according to each page of the submission form (*e.g.*, investigation or environment), and an alternating color pattern was used for each line of the display to assist the submitter in reading across the page. Within the underlying code, additional JSON files were used to store rules for page transitions, including "Back" and "Next" button functions and the associated page view changes, reducing the code length and improving readability. The progress bar was modified to be more informative, showing the label of each page within the submission form and the submitter's current position instead of the previously used percent of completion.

Automatic Assignment of EnvO Terms and Other Classifications

A primary goal of the submission form reorganization was to provide automatic assignment of EnvO terms to each library based upon user selections of easily interpretable environmental descriptors such as ecosystem, physical substrate,

and physio-chemical modifiers. Underlying rules were scripted in a JSON files, creating arrays of appropriate terms based upon user selections. Subsets of broad EnvO terms were selected corresponding to the appropriate ecosystem (EnvO biome and environmental feature), physical substrate (EnvO environmental material), and physio-chemical modifiers (EnvO habitat) to allow for useful grouping, sorting, and searching, since the complexity of overlapping EnvO hierarchies does not lend itself well to the automated retrieval of consistent or appropriate parent terms. This assignment must be handled during the submission process to allow for review by the submitter. Similar to the logic used to assign EnvO terms to existing MgOI libraries, the ecosystem field is evaluated in order to assign both the representative biome and feature terms. When “Agricultural” is selected as an ecosystem, an additional form field is dynamically displayed and must be completed in order to assign the appropriate terms. The physical substrate field is used to assign an EnvO material. The sphere field in VIROME is evaluated, and only if the library is designated as “Organismal” does sphere play a role in assigning EnvO terms – in that case, any feature assignment would be overridden and classified as “organic material” and any conflicting material assignment would be overridden and classified as “organic matter”. Finally, the physio-chemical modifiers are evaluated and one or more EnvO habitat terms are assigned (Figure 6).

Each assigned EnvO term is displayed on the submission form view and the submitter is asked whether the assigned term is accepted. Submitters have the option of selecting “Yes – I accept this descriptor”, “Yes – and I would like to add to this descriptor”, or “No – this descriptor is incorrect”. In each “No” instance, additional `envo_user` fields are displayed and the submitter may enter their own EnvO numbers

and names. The VIROME team recognized the importance of allowing the submitter to select “Yes – and I would like to add to this descriptor” and refine their annotation by adding additional or more specific terms, but the automatically assigned EnvO terms remain the criteria used for subsequent grouping and sorting as it represents a reduced set of controlled terms increasing the utility of these activities. However, we also want to capture instances in which the automatic assignment process failed to accurately annotate a library. When submitters select “No – this descriptor is incorrect”, a message accompanying the submission is sent to a curator triggering a manual review of the library. The `envo_user` fields are evaluated, appropriate EnvO assignments for grouping and sorting can be made, and corrections to the corresponding JSON file can be made or additional options within the form’s prompts can be provided (*i.e.*, the inclusion of another ecosystem in the available list).

Additional classifications can be derived to provide supplementary annotation with no additional input from the submitter. Based on form field entries, automated processes were established to provide the user with the sample site’s latitude zone and Köppen-Geiger climate zone (related to environment), the depth zone (related to environmental measurements), and the average read length and GC content of the library (related to sequencing). These assignments have been made to some of the existing MgOl libraries as part of manual curation, but have been automated as part of the public submission form.

Wladimir Köppen first published a global climate classification map in 1900, which was updated with the help of Rudolf Geiger in 1961 (Kottek et al., 2006). The Köppen -Geiger remains one of the most widely cited climate classification systems, and the development of new climate classifications has not been widely pursued (Peel

et al., 2007). Several updated Köppen -Geiger climate classification systems are available based on more recent climate measurements. Specifically, world climate classification maps based on recent data sets from the Climatic Research Unit (CRU) and the Global Precipitation Climatology Centre (GPCC) are available, with accompanying data files accessible for download (Kottek et al., 2006). The text file of latitude/longitude coordinates available through <http://koeppen-geiger.vu-wien.ac.at/present.htm> is used to automatically assign the Köppen -Geiger climate classification of the sample site based on the latitude and longitude coordinates provided by the submitter.

Zones are also assigned to annotate latitude and depth. Latitude zones include arctic (high latitudes, greater than 66.5°), temperate (latitudes 23-66.5°), and tropical (latitudes from the equator to 23°), as supported by the Environmental Literacy Council (<http://enviroliteracy.org/article.php/680.html>). Oceanic depth zones are assigned based on the depth measurement provided by the user. Depth zones (Yancey, 2011) include epipelagic (0 to -200 m), mesopelagic (-200 to -1,000 m), bathypelagic (-1,000 to -4,000 m), abyssopelagic (-4,000 to -6,000 m), and hadopelagic (greater than -6,000 m).

Finally, automated scripts process the library sequences to provide average read length and GC content. These values are determined during the submission process to provide an overall description of the library.

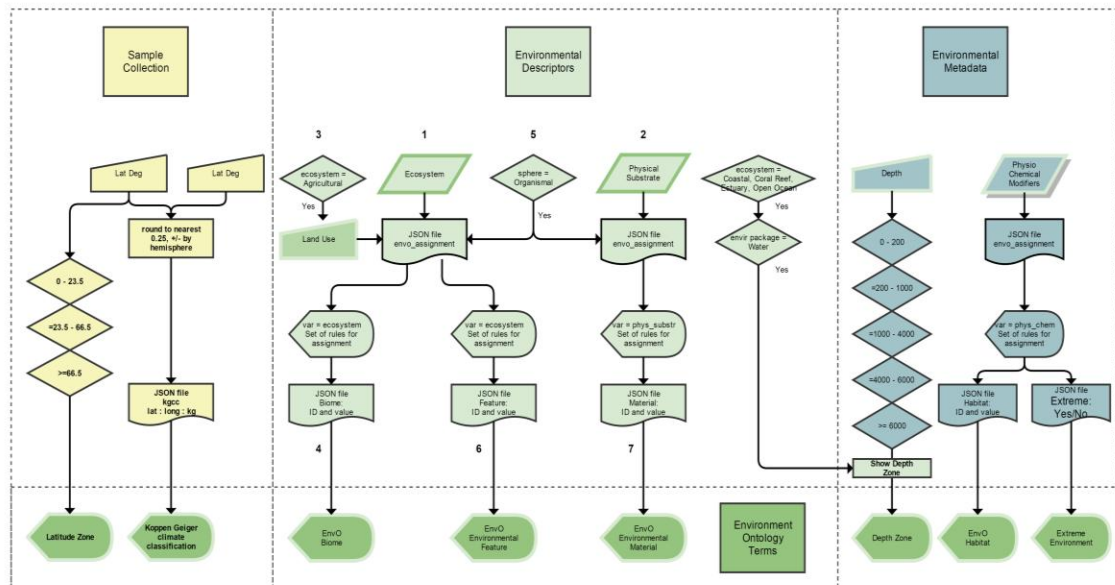


Figure 6 Logic used to assign EnvO terms and other classifications to VIROME submissions

Creating a Dynamic and Adaptive Submission Process

The new submission form is dynamic and adaptive, allowing user responses to guide progression through the form and determine subsequent prompts. The selections of particular responses will trigger additional questions to display, determine entire views of questions, autofill related fields, or mandate corresponding validity requirements.

Dynamic question display

In certain cases, the selection of a particular response will trigger additional questions to be displayed. For example, answering “Yes” to “Has this library been submitted to NCBI?” will show additional questions prompting the submitter for the corresponding NCBI BioProject, BioSample, and accession numbers. Similarly, only when a submitter has indicated that their library is from a collection of pooled samples do fields appear requesting the number of samples and a description of pooling.

Dynamic view display

In other cases, unique subsets of metadata fields must be answered based on a previous response. For example, the identification of the library’s “investigation type” as a metagenome, bacterial genome, or viral genome must influence the questions displayed downstream in the nucleic acid sequence source section. Similarly, the selection of “environmental package” determines the fields displayed for additional environmental measurements. In all scenarios, groups of questions must be hidden or shown, both on the appropriate submission form page and on the summary. For flowcharts of these dependencies, see Appendix D.

Two methodologies exist to accommodate the display of subsets of metadata as required by the parent selections (investigation type and environmental package). Initially, individual groups of all appropriate metadata prompts were built for each possible selection and uniquely identified as a class. The selection of a particular parent response would show its corresponding class and hide all others. This system greatly increased the length of the submission form underlying code, since metadata fields were repeated in as many classes as they were associated (for example, submitters are prompted for temperature in each of the environmental packages. The

code prompting for and gathering temperature data were repeated fourteen times, so that temperature appeared within each environmental package's class of metadata). Since each metadata field requires a unique label, this approach required a complex naming system incorporating both the dictating response and the corresponding metadata (since temperature occurs in each environmental package, a unique name was needed in each instance., *e.g.*, `air_temperature` or `water_temperature`). This approach made validation of the display/hide logic straightforward since entire, contiguous sections of the raw code could be verified against the web form. However, there were unsatisfactory consequences both in cases where the submitter modified a parent selection and in the import of final submission data to the underlying database structure. Under the first condition, when a submitter edited the parent selection after navigating through and completing the form, the downstream metadata fields would change based on the new parent selection (*e.g.*, environmental measurements applicable to soil would need to be displayed when the original environmental package selection "water" was edited and changed to "soil"). The submitter would be required to enter all new corresponding metadata fields, even those common to both parent selections. Since each metadata field would be uniquely identified and present only in its appropriate class (*e.g.*, `water_temperature` and `soil_temperature`), the original temperature entry would be hidden, and the submitter would be required to enter the temperature value again. Submitters attentive to comprehensive annotation could be frustrated by required duplicative entries and possibly disregard VIROME's utility. In the second condition, submission of data to the VIROME pipeline and MgOl database would be confounded by the reduction of so many repetitive fields to a single, stored metadata field. For example, only a single temperature field exists in the

MgOl database. Automated scripts would need to reduce each environmental package's temperature field to a single entry. Additionally, there is only one valid temperature measurement for each library submission based on its single environmental package. However, in cases where a submitter edited the environmental package, the original temperature entry would remain in the submission form and be hidden from view while the new parent's temperature entry would be displayed. In such a case, automated scripts would need to select the appropriate temperature value to be stored in the MgOl database (while one would assume the two temperature entries were equal there could be a discrepancy).

After consideration, the second methodology was adopted to display appropriate metadata based on parent selections (investigative type and environmental package). Metadata prompts were created only once and uniquely identified. Underlying rules were scripted in JSON files, creating arrays of metadata prompts to show or hide based on the parent selection. This approach reduced the total length of those submission form sections by 64% (reducing 106 duplicative submission form rows as part of 3 investigative type classes and 11 environmental package classes to 38 unique form rows uniquely identified). Storing the logic in separate JSON files reduced the length of the functioning code as well by removing the various show and hide displays, increasing readability and workflow. This system simplified the field names to a single metadata name (*e.g.*, temperature) instead of requiring a unique name for each class iteration of the entry. While the complexity of the dynamic display logic was increased versus checking the display of contiguous sections of raw code in option 1, the display of appropriate metadata fields was easily handled through rules specified in the JSON files. Most improved under this second approach is the

management of environmental metadata measurements after editing the parent selection, environmental package. When a submitter edits the parent selection after navigating through and completing the form, the combination of the downstream metadata fields would change accordingly. However, any common fields would still show the original submitted entries. In an example, a submitter inadvertently selected “water” and proceeded to enter appropriate environmental measurements such as temperature. From the summary page, the submitter realized their mistake and selected “Edit”, correcting the environmental package to “soil”, and was subsequently redirected to complete the environmental measurements corresponding to the new “soil” selection. Since temperature is a common metadata field between water and soil, the single block of code corresponding to the prompt/capture of temperature data is again displayed, including the submitter’s original entry. Finally, managing the identification of coding blocks in this way reduces the complexity of importing submissions to the VIROME pipeline and MgOI database, since all metadata fields occur only once in the submission form coding.

Autofill potentially duplicative fields

The VIROME submission form will autofill particular fields based upon a previous selection, reducing the number of potentially redundant prompts. The MIXS-compliant environmental package term is automatically assigned when applicable based on the selection of particular terms (organismal from the substrate list, or air, sediment, soil, and water from the physical substrate list), or is available from a reduced subset (the selection of “organic matter” as a physical substrate limits the environmental package options to miscellaneous and host-, human-, and plant-associated). Initial suggestions were to hide the environmental package field and

dynamically display it only when the user was required to make a selection (*e.g.*, when a user selected air as a physical substrate, air would be assigned as the environmental package, and the environmental package field would not display. Instead, when a user selected rock as a physical substrate, additional information was needed to determine environmental package and the field would display) However, the environmental package field is consistently displayed and allows for real-time automatic assignment based on user selections. It was important to keep term assignments transparent, particularly for the environmental package term that dictates the flow of subsequent form pages and the collection of appropriate metadata.

The submission form will automatically calculate pressure of a water sample when no pressure measurement is provided. This assignment is conditional based upon the selection of water as the environmental package, and is calculated using depth and salinity measurements (both required fields).

Validation of descriptors

The physio-chemical modifiers had unique requirements for submission form validation (Figure 7). Given that the physio-chemical modifier field is used as an indicator of extreme environmental conditions, users are required to enter a corresponding appropriate measurement. For example, when acidic is selected, a corresponding pH measurement is required and that value must match the defined pH range for acidic. In addition, incompatible modifiers cannot be selected together (*e.g.*, acidic and alkaline cannot both be selected; cold temperature and hot temperature cannot both be selected).

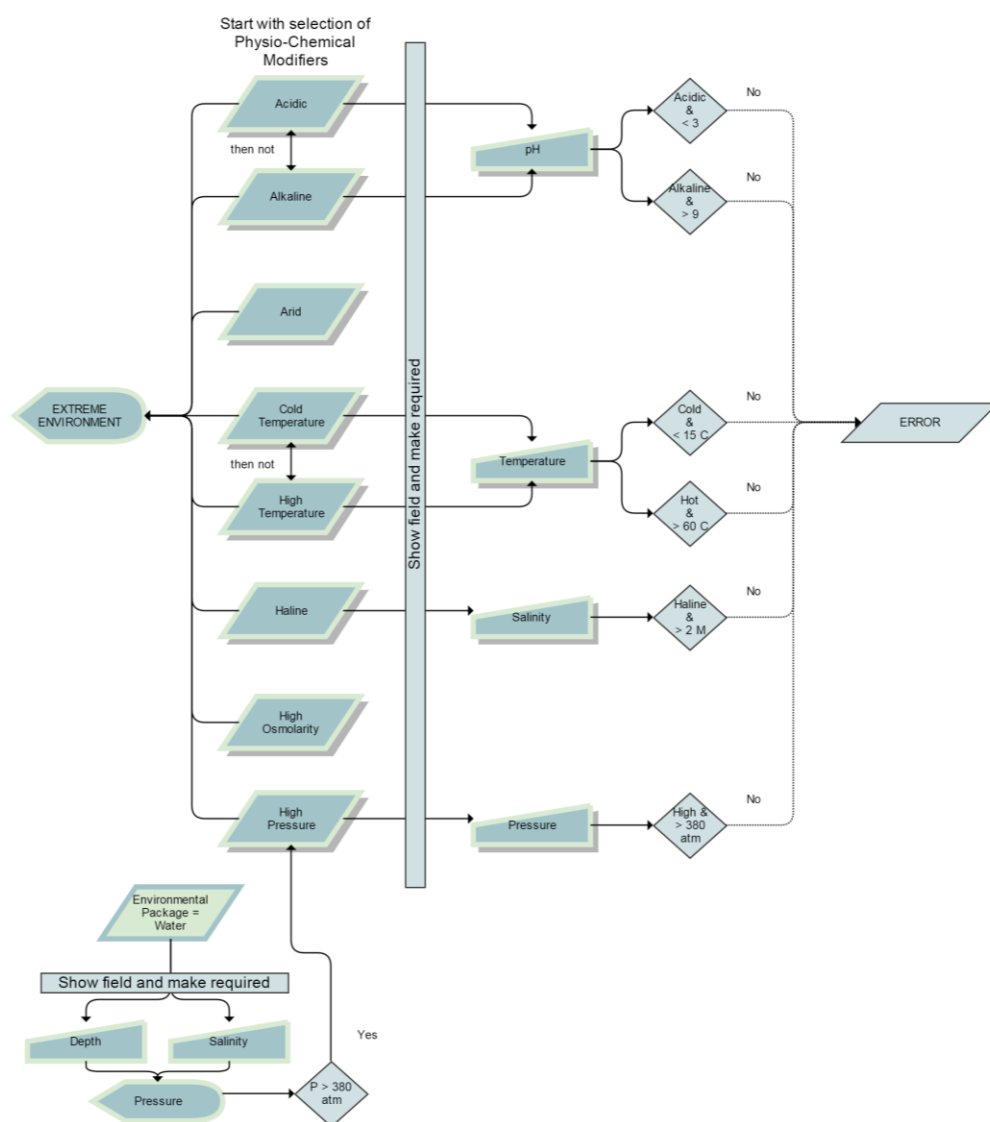


Figure 7 Logic relevant to physio-chemical modifiers validation, including association with related environmental measurements and classification as an extreme environment.

Consideration was given to the automatic assignment of physio-chemical modifiers based upon corresponding measurements (*e.g.*, the assignment of acidic based upon a pH measurement less than 3). However, the physio-chemical modifier terms are mapped to the EnvO habitat hierarchy, which defines a habitat as “a spatial region having environmental qualities which may sustain an organism or a community of organisms” (www.environmentontology.org). Habitats have spatial stability, and specificity to a species or population. Therefore, we decided not to assign habitat values to libraries based on a single, potentially atypical measurement. Emphasizing an accurate annotation, we were careful to rashly or incorrectly assign values or descriptors to a library. An exception is made with pressure measurements. Pressure has more temporal stability than temperature which may fluctuate greatly over a day, and is related to stable environmental conditions such as altitude or depth. Therefore, pressure measurements greater than or equal to 380 atm are automatically designated as high pressure per the EnvO definition of a high pressure habitat.

Finally, the previous submission form allowed the user to designate an environment as extreme, generally supported by the selection of one or more physio-chemical modifiers. The revised submission form now automatically assigns the extreme designation based on user selection from the physio-chemical modifier list. Extreme environments are those which are considered extreme habitats under EnvO (acidic, alkaline, cold temperature, haline, high osmolarity, high pressure, and high temperature) and arid environments (not defined as a habitat under EnvO but considered relevant by the team).

Future considerations

Considerations for future submission form updates include the addition of EnvO visualization tools for the submitter. While this was part of the original project proposal, evaluation of the overlapping EnvO hierarchies and input from the VIROME team suggested that this visualization be excluded from current revision goals. Users unfamiliar with EnvO could become quickly confused or overwhelmed by a visual depiction of the assigned term through its parent terms to the hierarchy root, or could be prompted to select additional terms for refinement that were part of a separate hierarchy and therefore inappropriate for a biome, feature, or material annotation. Assessing a visual aid also creates an additional submitter step beyond the acceptance of a single term, and potentially detracts from the tool's simplicity and accessibility. Any confusion or perceived difficulty could sway the user to withdraw their submission. The VIROME team was confident that the addition of standardized environmental annotation is currently best achieved through the simplest and most automated assignment and acceptance of EnvO terms.

Chapter 5

AIM 3: DESIGN OPPORTUNITIES TO BETTER LEVERAGE METADATA IN VIROME AND METAGENOMES ONLINE

The drivers behind this VIROME and MgOl update and behind many of the projects within INSDC and GSC are the abilities to test hypotheses using evolutionary and ecologically guided genomic approaches, and to group, sort, and search underlying data in meaningful and relevant ways. Collecting and leveraging environmental metadata is key to improving our understanding and interpretation of metagenomic sequencing, particularly as it relates to viral fractions of those environmental samples. Collecting accurate and comprehensive metadata translates into additional work and effort from submitters, which demands clear benefit and increased utility of the dataset in order to ensure a tool's value and sustainability. This project aims to enrich users' experiences with VIROME and MgOl with leveraged metadata by displaying all metadata with the library information in each tool, providing immediate output from the submission process designed to simplify a GenBank submission, and designing various approaches to practical grouping, sorting, and searching based on metadata.

Design Modifications to VIROME and MgOl Library Pages

The effort required on the part of the submitter to provide comprehensive metadata mandates that the metadata be effectively utilized and clearly displayed. The current VIROME individual library pages show a limited subset of the library's annotation metadata. Most of such page's content is designed to display the functional,

taxonomical, and environmental classifications of the sequences identified by BLAST results against the UniRef 100 and MgOl databases. However, the annotation of the query library is unavailable in that VIROME display. The current MgOl individual library pages show a more complete description of the individual library, using that display to provide comprehensive annotation rather than BLAST results. Revisions are still required to address the database schema modifications made as a part of this project. Recommendations for revisions to the individual library pages within VIROME and MgOl include the addition, removal, and grouping of fields to complement the new MgOl database schema, the inclusion of visualizations to increase understanding of environmental classifications, and the development of a unified format to display information in both sites.

Individual VIROME and MgOl library pages must reflect the new MgOl database schema, and should display all collected library information. The comprehensive annotation of libraries promotes VIROME's utility, allowing for improved interpretation of environmental BLAST sequence hits. However, that utility is only available when the environmental metadata is visible, and the user's evaluation of the functionality and effectiveness of the VIROME and MgOl sites may begin with browsing various library pages. In addition, the effort of VIROME submitters to provide a comprehensive annotation must be rewarded by making that information available on the library's display. Not displaying these metadata may give users/submitters the impression that the information is not relevant and providing it was a waste of their time. Instead, these library pages are an opportunity to confirm the importance of an inclusive annotation by providing all of the library's descriptors and metadata. Therefore, the VIROME and MgOl library pages must display all fields

in the MgOl database. Organization becomes critical, since overwhelming a user with text and poorly arranged information may deter their exploration of VIROME and MgOl. Consequently, this project recommends that data is organized according to the MIxS annotation sections (investigation, environment, nucleic acid sequence source, and sequencing) and in the same manner as the VIROME submission form. Consistency across the VIROME submission and final display will contribute to positive user experiences by making data more familiar and easier to locate and utilize.

The inclusion of graphics or visualization tools to support a library's annotation helps to break up large sections of text, make the site more visually appealing, and explains the relevance of environmental metadata. While some metadata require no visualization (*e.g.* temperature measurement or sequencing center), other descriptors could be improved by providing visual context. The current VIROME and MgOl library pages display a map identifying the sample collection site. Recommendations for these pages are to keep that map, and add additional displays to provide hierarchical context to the EnvO terms assigned to each library. Given the complexity of the EnvO hierarchies, some users may need an explanation of terms or benefit from a path of the term to the hierarchy's root. Several options are publically available from the National Center for Biomedical Ontology (NCBO) through their NCBO BioPortal. Code templates are available to add four widgets including a "visualization" display or an "NCBO tree widget" that provides a root to term tree. The tree widget is not recommended for individual libraries pages given the amount of text already displayed on the library pages. While this tool could be useful in other parts of the site, and could be added as a future improvement to the VIROME

submission process to improve a submitter's ability to accurately classify their library, a visualization or graphic is more appropriate on the library page. Revisions to the visualization code would be required to automatically generate the graphic based on the library's classifications.

Finally, this project recommends that the VIROME and MgOl library pages be designed so that the main content and appearance are identical. Identical pages can reduce some of the required coding (though the pages are currently scripted in different formats), builds a common work experience between the two tools, and creates consistency across environments which can improve users' experiences. While sidebars and page functionality must still be unique within VIROME or MgOl, the overall description of the library can be common (Appendix E).

Design Outputs to Provide Useful Annotation and Facilitate Batch Submissions

VIROME and MgOl both benefit from the availability of comprehensively annotated libraries. VIROME's breadth and functionality increase as UniRef continues to grow. MgOl has historically expanded as well-annotated environmental libraries become publically available. The VIROME team advocates for data sharing, not just for increased utility of their tool and database, but for the inherent global benefit of unrestricted data access promoted and maintained by the INSDC. Therefore, this project recommends providing user outputs at the completion of the submission process to facilitate the submission of data to INSDC. This output consists of a printable file with library information for easy reference when creating a submission to another organization or tool, and a text file containing submission entries that can be uploaded directly to NCBI's GenBank for submission.

The metadata text file contains the library's submission information, and is designed to be used in place of the submission file template available for the batch deposit of metagenome or environmental samples to NCBI's BioSample database (Figure 8). Genbank's existing submission templates provide another opportunity to leverage standardized metadata collection tools and ontologies in VIROME and MgOI to maximize their performance, functionality, and utility. Using the BioSample submission template as a guide increases the utility of the text file output in several ways, including the conservation of VIROME team effort, the promotion of data sharing, and the potential increase in VIROME user loyalty. First, VIROME team effort is conserved since time is spent on adapting an already existing template, rather than on developing a new and compatible format. The Genbank submission templates are flexible to allow custom fields, so that all descriptors and metadata collected in the VIROME submission process can be captured and used to annotate the library. Second, creating a text file output promotes data sharing since the file can be used to reduce the length of the submission process to another agency. Many tools or databases allow for batch submissions using a text file rather than a submission form, so that the proposed text file could be uploaded to the tool in place of going through an additional submission form or process. Finally, creating a text file output based on parameters in the Genbank submission template has the potential to increase VIROME user loyalty. VIROME not only provides analysis and examination of their data, but also has the potential to reduce the effort involved in submission to another INSDC or other databases. An INSDC submission, while valuable, is also often considered an arduous task by an investigator whose primary drive may be research and exploration. Reducing the workload by allowing an investigator to go through a submission

process once and receive two beneficial outcomes is an advantage that may trigger users to return to VIROME more often.

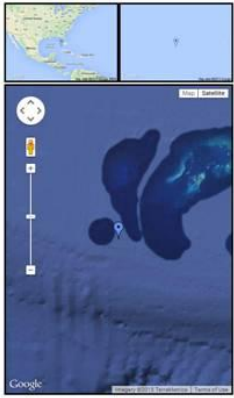
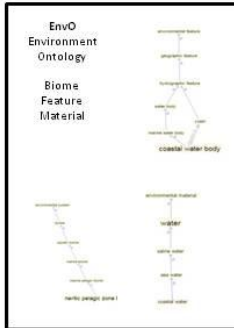
[illegible]

Figure 8 Sample of VIROME submission form text file output. Several text outputs gather relevant fields in a format compatible with GenBank submissions to the BioProject or BioSample databases or to the Sequence Read Archive.

A second printable file output will also be provided, containing the library's submission information, and is designed to mirror the VIROME and MgOl library pages (Figure 9). The printed output is intended to provide a summary of the submission for the user and a quick reference for library information when the user submits to other tools or drafts summaries of data. The consistent format across the printed output, VIROME, and MgOl reduces team effort in developing and maintaining multiple formats, provides a uniform user experience serving to reduce the time and effort required to find information when using each tool, and improves

the user's impression of tool reliability since the same representation of the data is given across tools and applications. First and significantly, using the same design to present the library data saves VIROME team members time and effort in maintaining multiple formats. While different programming languages may be used to generate the same layout in the printed output, VIROME, and MgOl, the common end goal provides a clear description of and limit to the end result, eliminating the need to develop unique arrangements or graphics for each setting. Second, the consistent format saves users time and effort, limiting their investment in locating the same information across multiple tools. A submitter will be able to refer to their printed output, and see the same information presented on the screen when their library is available on VIROME and MgOl. Finally, presenting the same fields in each application translates into confidence in their value, since one cannot question why a field or term is present in one function and not another when the tools are so closely related. While the printed output is not intended to be used directly as a submission or summary of the library information in another resource, the potential exists for the output to be used as a figure or table in a paper or other documentation, providing opportunities for additional VIROME exposure and citation.

Library Information	
General Information	Sample Site Details
Prefix	Pooled
Project	Sample Collection Date
Data Source	Coordinates
Station	Latitude Zone
SubProject	W-O Climate Zone
SubSample	Country
Library accession	Region
Investigation type	Place
	Site ID
Sequencing	
Assembly	Environment Descriptor
Assembly method	Genetic
Sequencing method	Species
Sequencing center	Biome
Sequencing details	Feature
Amplification	Material
Fragment size	Habitat
	Extreme
Sequence Source	Environment Metadata
Investigation Type	Environmental Package
Metagenome method	Altitude
Nucleic acid type	Pressure
Altitude lower bound	Depth
Altitude upper bound	Depth zone
Sample collection, processing, and size	Total depth
	Elevation
	Humidity
Library Statistics and Annotation	pH
Reads	Salinity
ORFs	Temperature
Complete ORFs	Biome
Missing both ends	Chlorophyll
Missing Start (5')	Conductivity
Missing Stop (3')	Density
Average read length	DIC, DIP, DOC, DO, H ₂ O ₂
Coding density	Taxonomic name
GC percentage	Contig name
ORFs	Subject ID
Function Annotated	Organization
Taxonomy Annotated	

Citation: Wenzel, F.B., J. Bhattacharya, J. Chen, M. Dumas, D. Goh, M. Hwang, S. J. Kim, and S. J. Kim. 2012.
 VIROME: a unified operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* 6: 42-49.
<http://www.virome.org>

Figure 9 Sample of VIROME submission form printable file output. The proposed output is formatted to match the unified VIROME and MgOl library views for consistency and ease-of-use across multiple tools.

Design Opportunities to Compare, Search, and Group by Environment and Metadata in VIROME and MgOl

The significant contribution of the VIROME tool and MgOl database is their potential to provide information about the environmental occurrences of viral sequences, which are so underrepresented in other databases. Leveraging that environmental data to provide the best experience and most meaningful exploration of sequencing data must include the abilities to group, search, and compare based on

environmental metadata. Those functionalities currently exist in VIROME within the “browse”, “search”, and “compare” pages and can be maximized by expanding the list of parameters available for each function, adding parameters to the results given by each function, and creating interaction and transition between functions.

The user-driven approach to VIROME’s functionality and exploration of metagenome sequences is one of the tool’s strengths and advantages. Expanding the list of parameters available for establishing group, search, and compare features is an update with a high return value and functionality across features. Incorporating additional parameters must be handled in different ways across the browse, search, and compare pages.

Provide an overview of the VIROME libraries by environment

VIROME’s home page currently shows an overview of MgOl libraries (“Bird’s eye view”), grouped by the previously described “environment” term. This snapshot can guide deeper exploration of libraries using browse, search, and compare functions, and is a valuable opportunity to intrigue a user by leveraging a breadth of environmental parameters. Creating a flexible view, allowing the user to group by any one of several descriptors, provides dynamic means to explore the extensive underlying database. A drop-down menu or other mechanism would allow the user to modify the display and corresponding table (Figure 10), and introduce them to the flexibility of VIROME and the many available mechanisms with which to explore a dataset.

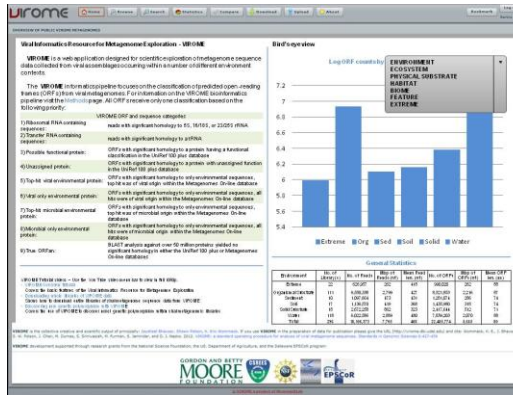


Figure 10a

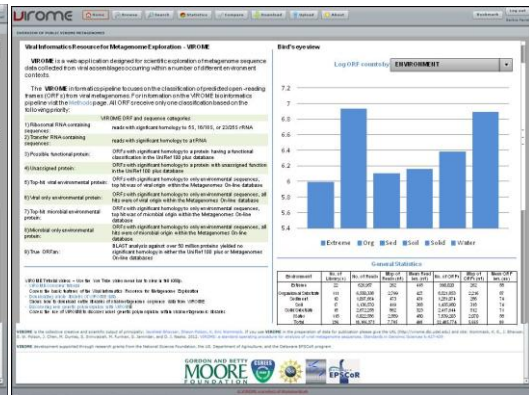


Figure 10b

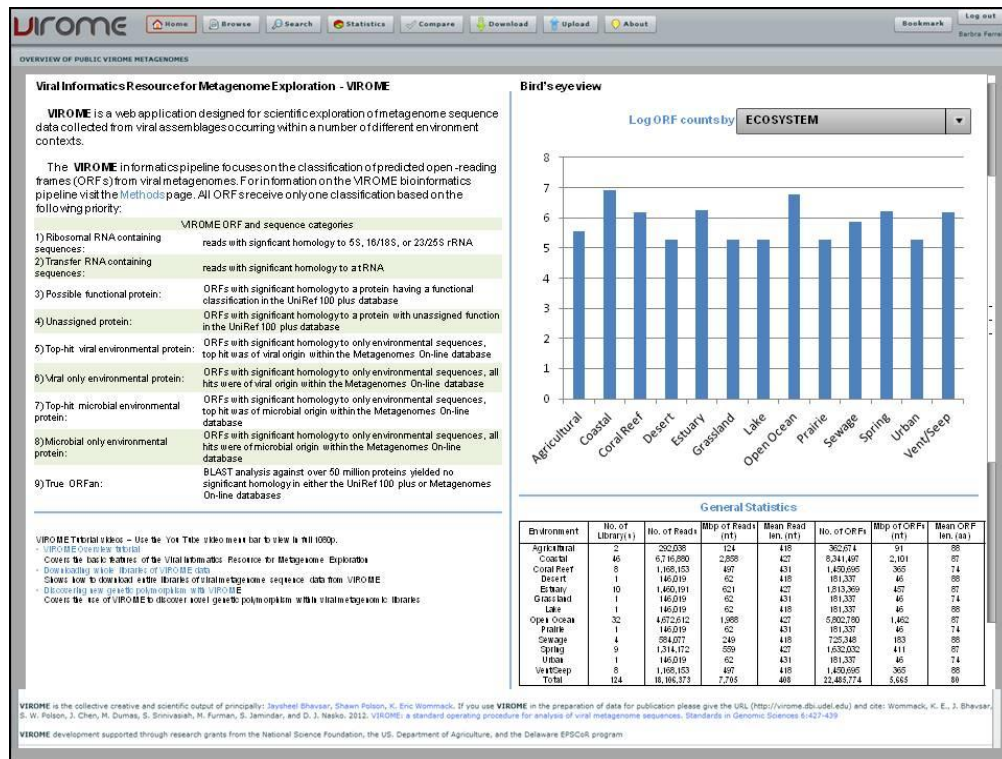


Figure 10c

Figure 10 Proposed dynamic VIROME home view. Graphs and tables will be flexible based upon user-driven criteria. Home-screen displays libraries by one environmental descriptor (environment) in figure 10a, users are able to select a different environmental descriptor in figure 10b, and both the graphic and tabular displays are modified according to that selection in figure 10c.

VIROME: Browse View

VIROME's Browse view allows the user to view a list of available libraries by environment, with no flexibility in this presentation. The "environment" term currently in place will no longer have an exact match in the new schema, though the most closely related term is "environmental package". Here, recommendations include the ability to group by one of several environmental descriptors, to select multiple parameters within that environmental descriptor, and to refine a group by stacking multiple descriptors. First, additional primary environmental descriptors would be added, including ecosystem, biome, feature, material, environmental package, and habitat. While results for ecosystem, biome, and feature are closely related, there will be some differences based on whether a library originated from an organismal sample (assigning a different environmental feature), and there may be user preference for searching based on a particular term given their familiarity with the Environment Ontology. Those primary descriptors could be added as terms in a drop-down menu, as tabs across or within the Browse view, or by another method proposed by the VIROME team for its functionality (Figure 11. For additional mechanisms, see Appendix F).

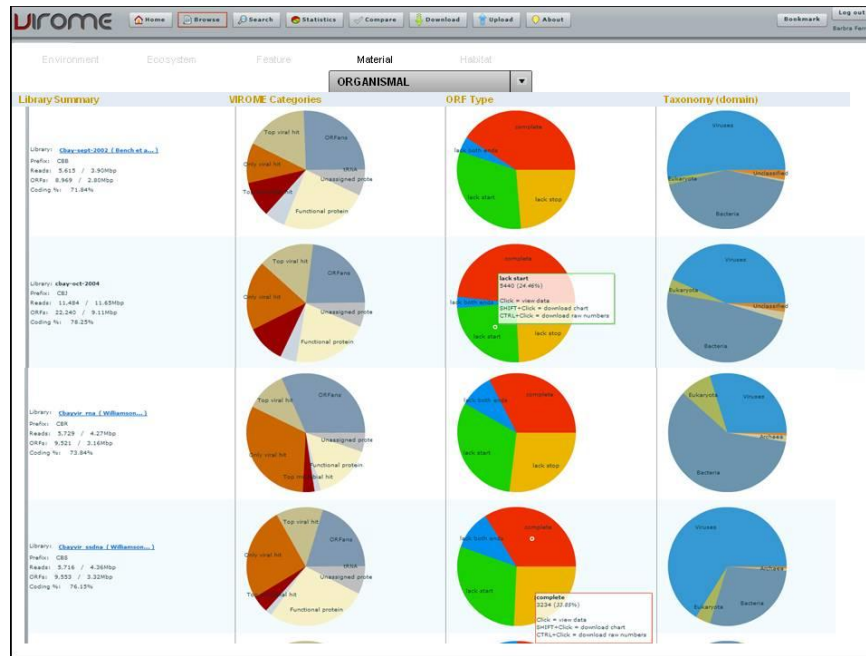


Figure 11 Proposed dynamic VIROME Browse view. Rather than static filters, libraries are sorted based upon user selection of available environment or library descriptors

A list of available terms within the primary descriptor may be best provided as a drop-down menu. While much of the metadata collected for each library is environmental, there is also the potential for adding primary descriptors related to the investigation, nucleic acid sequence source, or sequencing categories. A user may find the need to explore based on investigation type (*e.g.*, metagenome or viral genome), by nucleic acid type (*e.g.*, double stranded DNA or single stranded DNA), or by sequencing technology, and those additional primary descriptors could be added to the Browse view. Second, the option would be added to select multiple terms within the primary descriptor list. For example, a user may want to group libraries by ecosystem

(primary descriptor), and view libraries described as “open ocean” or “coastal”. This option may not change the presentation of available terms (*i.e.*, the functionality of a drop-down menu may still be a good fit for providing applicable terms within the primary descriptor), but does change a normally “select-one” option to a “select-multiple”. Note that additional complications arise in primary descriptors that allow for multiple assignments to the same library (*e.g.* physio/chemical modifiers, which can have multiple terms assigned to the same library in a single database field). This type of grouping must allow for the display of all libraries with a particular term (whether that term is single or one of multiple), and must allow for the user to group using “OR” or “AND” operators (*e.g.*, one must be able to retrieve either all libraries that are either “acidic” or “high temperature”, or all libraries that are both “acidic” and “high temperature”). Finally, browsing or grouping is optimized when multiple descriptors can be stacked to refine a given list of libraries. The previous recommendations allow for choosing one of multiple primary descriptors, and one or multiple terms within that primary descriptor list. This final recommendation allows for adding secondary descriptors to refine a search. Adding this functionality may guide the layout of the Browse view, given that selecting multiple primary descriptors as tabs along the top of the view is complex and prone to user error given that it may be difficult to identify selected descriptors. Instead, using a drop-down menu to select a primary descriptor and adding additional drop-down menus as layers are added to the grouping may make for a cleaner presentation of the grouping path and parameters. This functionality allows a user to select libraries within a particular ecosystem, and to then refine the list by selecting a particular substrate, a particular metagenome fraction, and a particular nucleic acid type. Note that adding

environmental metadata measurements such as pH or temperature allows for more dynamic exploration of libraries, but has unique considerations. Options must clearly be presented for selecting an exact value or a range of values, and for grouping only those libraries within that value or range or for also displaying those libraries with no value provided. Many of the environmental metadata parameters are not required for submission to VIROME or INSDC, and removing those libraries with no entry for environmental measurements may severely limit the available library set. That restriction may either be of great use to an investigator or may be an unwelcome consequence of setting absolute values in a search, and the option to select the strictness of the group is a valuable tool within VIROME.

Additional recommendations to VIROME's Browse view include a change in the graphical representations of each library's VIROME categories, ORF categories, and taxonomies. Currently, those library snapshots are presented in pie charts, which take up significant vertical space on the site and limit a user's view to two libraries. Displaying that information in a stacked bar chart may allow the information to be shown in the same horizontal space with a smaller vertical footprint, allowing libraries to be displayed in a table format (Figure 12).

among other parameters, the environmental descriptors genesis, sphere, ecosystem, physical substrate and region. The ORF tab displays a heat map with hit e-values in various databases, including “META” for Metagenomes Online (for consistency and appropriate reference to MgOl, this header should identify “MGOL” rather than “META.” While MgOl supplies valuable environmental metadata, headers refer to the name of the various databases). When the user hovers over a “META” block, the library name appears, which may give some indication of the environmental origin of the sample. When the user clicks on the “META” block for additional detail, a table is generated which lists, among other parameters, the library’s genesis, sphere, and ecosystem.

Currently, the ability to view environmental parameters of the library or sequence is several layers deep in the search, and there are no opportunities to either refine the search by particular environmental parameters or to display particular environmental metadata. On VIROME’s Search view, recommendations include the ability to refine a search by one or more environmental descriptors and the ability to view those environmental descriptors in multiple levels of the presented results. First, environmental descriptors should be added as search criteria, including ecosystem, biome, feature, material, environmental package, and habitat. Those primary descriptors could be added as terms in a drop-down menu or as additional, optional line items in the search menu (Figure 13. For additional figures, see Appendix G). A list of available terms within each primary descriptor is likely best provided as a drop-down menu, as in the current Search view orientation. Just as in the Browse view, the potential exists to add primary descriptors related to the investigation, nucleic acid sequence source, or sequencing categories. Additional functionality may allow the

user the option to select multiple terms within the primary descriptor list, again changing a possible “select-one” option to a “select-multiple”. Note that, similar to the Browse view, additional complications arise with primary descriptors such as physio-chemical modifiers that allow for multiple terms to be assigned to the same library. The search parameters must be clear to the user, so that the option to choose one term “OR”/”AND” another is plainly communicated. Second, environmental descriptors should be added to results tables in multiple levels (Figure 14). The option to add/remove fields from the visible table display is still ideal. Recommendations include displaying ecosystem and environmental material by default, with the option to display additional descriptors.

Figure 13 Proposed dynamic VIROME Search view. Search criteria can be added or removed to narrow or broaden search results

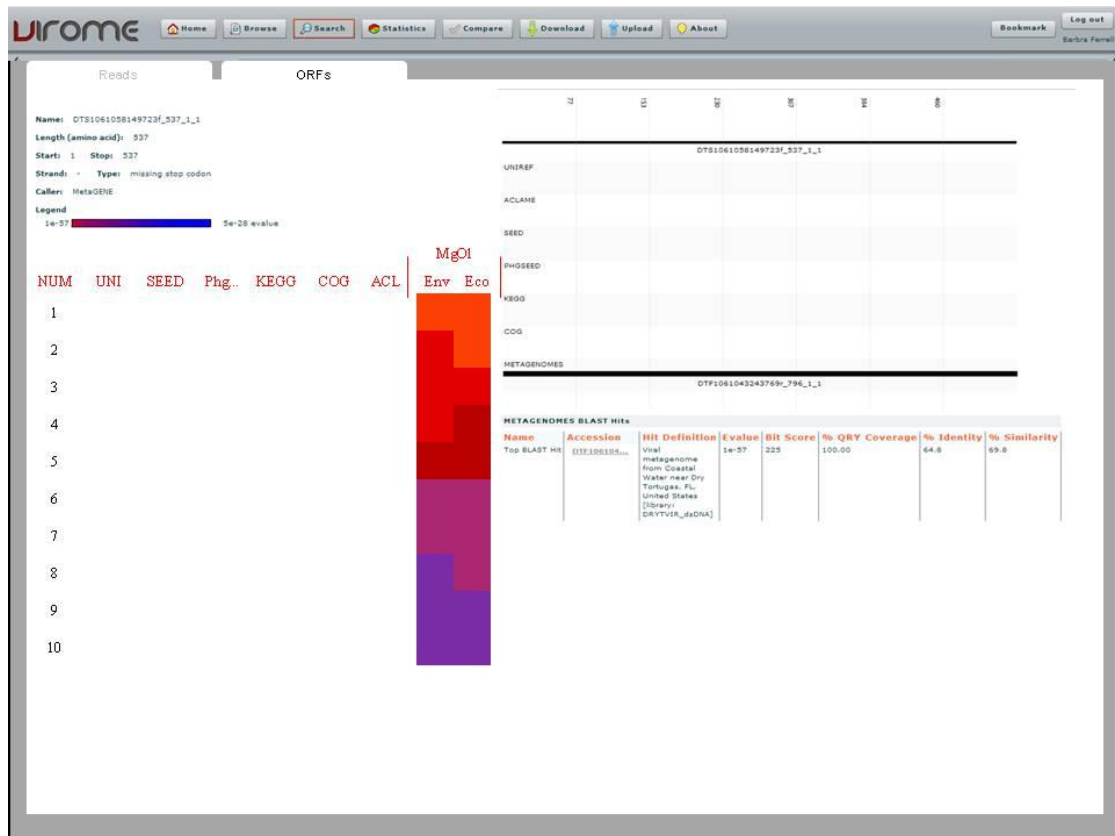


Figure 14 Proposed VIROME detailed search results. Proposed view includes multiple environment descriptors in ORF table

Additional recommendations to VIROME's Search view include a change in the presentation of the two search types, identification of required and optional terms for a search, and consistency of the library name. First, the two search types are not clear on the current view layout. For example, the search type is identified in a vertical header down the side of the view, a layout only used on this particular view in the VIROME tool. Additionally, the other search type is listed at the top of the current search type's menu, making that other search type appear as a header or title rather than as a link to select an alternative search method (when ready to submit a database

search, users may inadvertently click the “BLAST search” link and be redirected to a different view). Recommendations include using consistent header formats to identify the search type and moving the link for the alternative search type (Figure 13, above) or using an alternative method such as a primary drop-down menu. Second, required and optional search fields should be clearly identified. The current layout lists multiple search parameters, but it is unclear whether certain parameters are recommended or required. Recommendations include building additional search parameters in a similar manner to adding additional group parameters on the Browse view. Consistency is an advantage, both in user familiarity with various functions within VIROME and in reduction of team effort in creating multiple presentations of the same function. Finally, consistency in the library description is recommended to make exploration of different VIROME functions more fluid and contribute to an overall positive user experience. Library names can be very similar (*e.g.*, several libraries differ only in their identification as DSDNA or SSDNA), so that the additional of the unique prefix in the list of library names may be a helpful way to distinguish libraries from each other. The unique prefix is listed on the Browse view for each library, so that those users who have explored data there may have a list of libraries to search. That list will be much easier to navigate when a user can select by three letter prefix rather than a longer text field. The unique prefix should also appear each time the library name is mentioned, such as when the cursor hovers over the heat map and library information appears. Finally, the key to the table of sequence blast results should be modified for clarity (Figure 15). Currently, it potentially reads as if the top BLAST hit is identified with an “X”. This issue can be addressed by spacing the key apart from the link to

close the window, or by moving the “X” to close the window to the right side for consistency with the other VIROME views.

	Seq ID	Seq description	E-value	Bit Score	Genes	Species	Eco-system
1	0111041043243766_196_1_1	Unlabeled metagenome from Coastal Water near Dry Tortugas, FL, United States (Library: 0111041043243766)	1e-57	225	Natural	Aquatic	Coastal
2	GAY_JCVI_REF_110590428825	Microbial metagenome from Coastal Water near Northeast of Colon, Panama (Library: 05019)	4.9999999e-99	223	Natural	Aquatic	Coastal

BLAST results Sequence Details
 = top BLAST hit

Figure 15 Proposed VIROME search view results table header. Modifications suggested to provide clarity and a uniform user navigation experience

VIROME: Compare View

VIROME’s Compare view allows the user to select multiple libraries from a list sorted by environment, with no flexibility in this presentation. Selected libraries are compared according to one or more selected metrics, including NCBI taxonomy and the annotated databases which accompany UniRef 100 – ACLAME, COG, GO, KEGG, SEED and PhageSEED. Environmental metrics of library, library type, ecosystem, and EnvO terms are displayed, but are currently inactive and not available for comparison metrics. Output format options include a tab delimited file or a Biological Observation Matrix suitable for QIIME analysis, each with either raw numbers or data normalized by the size of the largest library. The output file displays a row ID, a unique term or combination of terms within the comparison metric, and the number of ORFs within each selected library categorized according to the selected metric. For example, a comparison of libraries based on KEGG metrics may display metric terms such as "F1_Cellular Processes", "F2_Cell Communication", and

"F3_Adherens Junction", with a count of the ORFs within each library classified with that combination of terms. Currently, the list of libraries available for comparison is sorted by environment, but there is no functionality to limit the list of libraries by environment or by any other criteria. The selection of each library for comparison is done individually (there is no functionality to select multiple libraries), and the list resets to the top upon each selection. In addition, the results of each search metric are presented individually. While these lists are still extensive since they list individual classifications as well as all existing combinations, there is no opportunity to build more complex matrices. Also, the list of environmental metrics is limited and inactive. On VIROME's Compare view, recommendations include the ability to refine the original list of available libraries by one of several descriptors, to select additional environmental descriptors as comparison metrics, and to select multiple environmental metrics within the same comparison. First, the list of available libraries should be either limited to or sorted by several available environment descriptors, including ecosystem, biome, feature, material, and environmental package. Additional parameters could be added grouping based upon descriptors related to the investigation, nucleic acid sequence source, or sequencing categories. Note that in the comparison tool application, grouping by habitat is potentially problematic, as either a single library can belong to multiple groups (*e.g.* "acidic" and "high temperature") or each unique combination of habitat terms becomes exclusive and the user may miss certain preferred libraries (*e.g.*, an "acidic" and "high temperature" library would belong only to the "acidic high temperature" group, and not to either the "acidic" or "high temperature" groups). In the current display, with a comprehensive list of libraries only grouped by environment and not restricted to a particular environment,

users are able to see contiguous groups which may promote further investigation. However, this presentation becomes cumbersome as the number of libraries in VIROME and MgOl grows. Second, additional environmental descriptors should be added to the list of available comparison metrics. Minimally, this list should include ecosystem, biome, feature, material, environmental package, and habitat. In this case, habitat and the potential for multiple descriptors is not an issue, and the descriptor is suitable as a comparison metric and for the comparison tool's current output. Additional environmental metadata measurements, such as pH or temperature, could be added as comparison metrics. Again, these measurements have unique considerations, including options for comparing based on an exact value or a range of values, and for noting those sequences with no value provided. Since many of the environmental metadata parameters may not be provided with a library, these fields may not be particularly useful for comparison purposes. Finally, update recommendations include the ability to select multiple environmental metrics within the same comparison. The current comparison output is a separate file for each selected metric. This output format is appropriate for comparisons based on annotated database (*i.e.*, ACLAME, GO, or KEGG) functional classifications, since a particular sequence may have multiple functional assignments (*i.e.*, a multiple-select descriptor) and the comparison output lists each unique combination of classifications as one occurrence. Selecting multiple database functional classifications in a single output file creates a very complex and potentially not useful matrix. However, since most of the environmental descriptors are select-one fields (*e.g.*, ecosystem, biome, feature, material, and environmental package), these descriptors could be used as multiple metrics within a single comparison. For example, a comparison based on ecosystem

and material would count the number of ORFs classified as 1) coastal and sediment, 2) coastal and water, 3) stream and sediment, and 4) stream and water. Note that some combinations of metrics are closely related that few, if any, additional combinations of terms would be generated. For example, a comparison based on ecosystem and biome would never create multiple ecosystem-biome combinations, since one and only one biome is automatically assigned based on the single ecosystem selection during the submission process. In contrast, a comparison based on ecosystem and feature would possibly create multiple ecosystem-biome combinations, since one particular feature term or the term “organic matter” can be automatically assigned based on the ecosystem and sphere selections during the submission process. However, the Compare output does not allow the user the flexibility to display additional columns or parameters as is available within the Search output – doing so would only create additional metric combinations and therefore additional rows in the output (*e.g.*, since not all “coastal” ecosystem sequences are necessarily all “water” environmental material sequences, multiple rows must be created to count sequences within selected libraries for each ecosystem and environmental material combination). Rather than consider the rules involved in creating meaningful outputs built on multiple environmental metrics, it may be in the best interest of the VIROME team to choose not to implement this last recommendation and continue to allow only one comparison metric per output. Note that if the user has grouped the original library list by environmental descriptor and selected only libraries with a common descriptor, that the user has already allowed for this parameter to be considered as part of their comparison matrix.

Additional recommendations to VIROME's Compare view include a change in the format of each library's name or description in the available list, and a change in the library selection process. Each library in the available list should be identified by both the prefix and the library name for consistency across VIROME views and to facilitate the identification and selection of preferred libraries. The library selection process could also be improved by accommodating multiple selections. This function could be accommodated in several ways, including by holding the list static after a selection is made so that the user does not have to scroll through the list to their last selection place, by adding a multiple select functionality using the SHIFT or CTRL/APPLE keys for contiguous or non-contiguous list items, by selection boxes next to each library in the list, or by another method selected by the VIROME team.

VIROME: Interaction Between Pages

The final recommendation to leverage environmental metadata in VIROME and MgOl is to design interaction between the VIROME Browse, Search, and Compare views. Suggested updates on each view include the ability to select by primary environmental descriptors, select multiple terms within those primary descriptors, and to add secondary descriptors for refinement. The differentiation across views lies in the steps following the primary and secondary descriptor selection, in the user's decision to view, search against, or compare those appropriate libraries. Therefore, determining a common procedure across views for the grouping, sorting, or searching of libraries may begin to allow for transferring those selected libraries across functions. Once libraries are selected, the user should be able to take those libraries to any of the Browse, Search, and Compare views without navigating and narrowing the list of libraries again, saving a user considerable time and effort. While

this elimination of redundancy (*i.e.*, trimming a list to view libraries in Browse, then trimming a list to select libraries in Compare) may not be noticeably smoother to new users, the current process is likely to be viewed cumbersome. This view interaction has several considerations, associated with potential view reorganization and the method library transfer. View reorganization involves either keeping the Browse, Search, and Compare features as distinct views, or combining the views and adding search and compare buttons to a browse-like view. This project recommends that the three views remain separate since each function not only is unique but provides complex and multi-layered results. Each function should still be highlighted separately in the menu at the top of the VIROME view so that users are always prompted to continue their investigation or evaluate the data in a different way. The library transfer action could be implemented through direct links among the views, or through the addition of libraries to a “cart” or “lab bench” holding function. Direct links on each view would allow the user to trim a list of available libraries, select several, and either continue in the current process or immediately change course and take the libraries to a different function. For example, the Browse view would allow the user to trim a list of available libraries, select several of interest, and click a link to import those libraries directly into the Compare view (Figure 16). This process is relatively short-term in that it does not identify “favorite” libraries, but only transfers the currently marked libraries to another function. Implementation involves adding buttons to each of the Browse, Search, and Compare views to allow libraries to move between them (consider if it is feasible for a user to begin selecting libraries in Compare and then choose to browse their libraries for additional features before continuing).

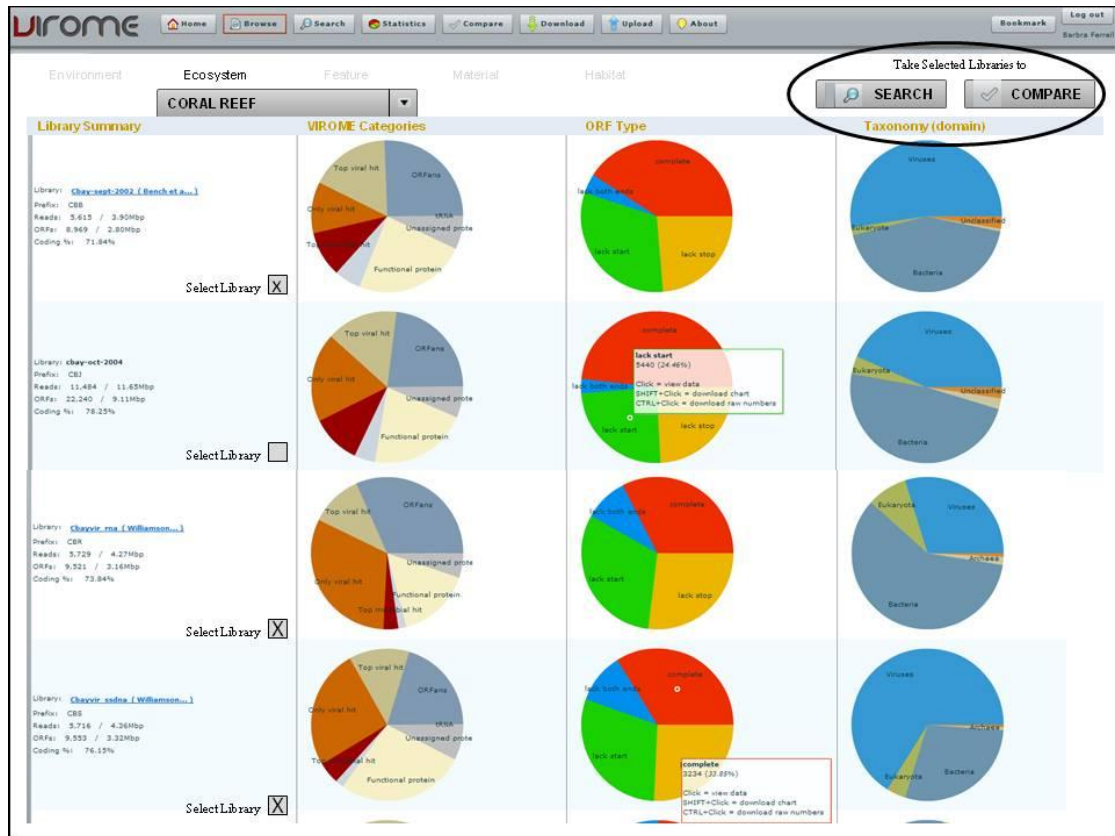


Figure 16 Proposed library selection tool to provide functionality of selecting one or multiple libraries and transferring them to another VIROME function

The second potential process is the creation of a “cart” or “lab bench” feature, which would hold selected libraries and allow users to take their “lab bench” to any of the Browse, Search, or Compare views (Figure 17). The lab bench creates an intermediate view, but also has the potential to create a longer-term selection since the lab bench list would not necessarily clear when the user took those libraries to a particular function. Implementation involves adding a button to each of the Browse, Search, and Compare views to allow libraries to be selected for the lab bench, creating a “cart” or “lab bench” feature or view, and creating links within the lab bench to take

libraries to the Browse, Search, or Compare view. Note that with any method of library transfer, the ability to select particular libraries must be implemented. This may be handled through the addition of selection boxes next to the list of libraries, or through moving selected libraries to an adjacent list (as in currently utilized in the Compare view). For consistency, the same selection method should be employed across the site. Since two adjacent lists are not appropriate for the Browse view, the recommendation is to use selection boxes adjacent to the library list. Developing the lab bench strategy would also permit the addition of a library to the lab bench from its individual library view (additional views available in Appendix H). Giving the users additional instances to note or flag a library of interest creates more cause to explore all of VIROME's functions and potential.

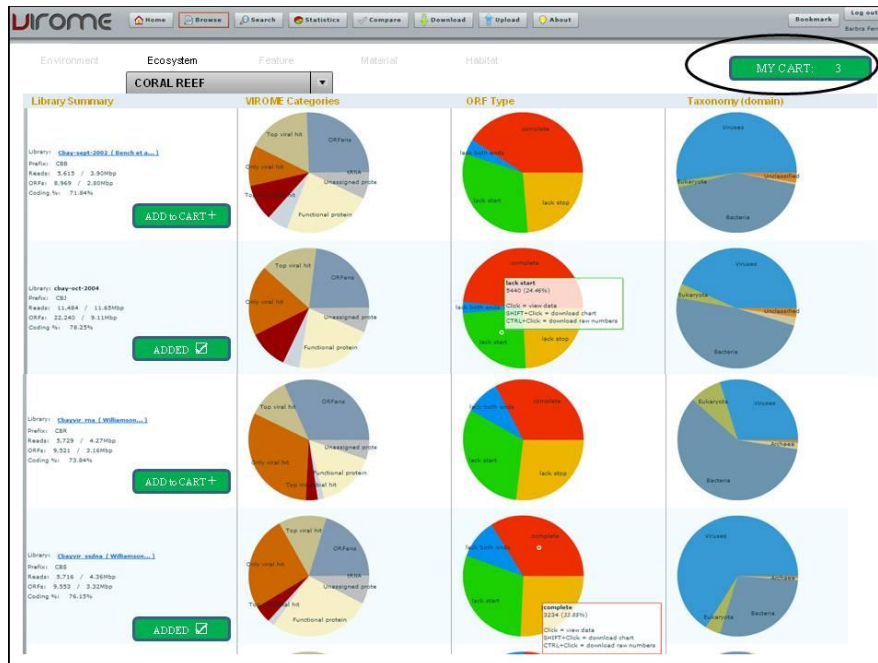


Figure 17 Proposed VIROME “cart” or “lab bench” function. Selected libraries are maintained in a user-specific container to quickly retrieve user-selected libraries and move them to another VIROME function

Chapter 6

CONCLUSION

Impact

The shotgun metagenomic approach remains an ideal tool for profiling environmental composition and diversity. The power of this method is limited, particularly for viral samples, when as much as 70% of sequences show homology to reference database sequences or only show homology to proteins with no known function. Environmental context provides insight to viral metagenomes when homologous sequences are not identified in reference databases with known functional or taxonomic information. The Viral Informatics Resource for Metagenomic Exploration (VIROME) was developed to provide functional, taxonomic, and environmental homology evidence for viral metagenomes, and to provide visualization capabilities and useful binning and comparison tools. The Metagenomes Online (MgOl) database of environmental peptides provides environmental context and gives insight to viral metagenomes when homologous sequences are not identified in reference databases. The 258 microbial and viral metagenomes in MgOl have been manually curated with environmental metadata, which provides a providing a framework for the sequence homology results and increases the proportion of a metagenome to which meaningful context can be ascribed.

This project significantly enhanced the function and value of VIROME and MgOl by enriching the quality and consistency of the associated metadata. Modifications and updates to the VIROME pipeline and MgOl database annotation were designed to align environmental metadata with relevant standards and to leverage

that metadata to increase the value and depth of metagenomic analysis. Targeted improvements were directed at retrofitting the current MgOl database to better reflect metadata standards, implementing a new VIROME library submission interface to capture appropriate metadata on new submissions, and leveraging that metadata to improve users' experiences and depth of analyses in VIROME and MgOl.

Retrofitting the current MgOl database in order to align its environmental metadata with applicable standards was accomplished by designing modification to the current MgOl database schema, manually annotating the current libraries to fit the new schema, and implementing those database structure changes. A review of relevant genome and metagenome submission and annotation standards set forth by the INSDC and the GSC's MIxS standard and of applicable annotation guidance through EnvO influenced revisions to the MgOl database, including the modification, removal, or addition of particular fields. Current libraries were annotated with EnvO terms using automated scripts in order to fit the libraries to the new schema. Finally, a new database was implemented to accommodate the revised schema and libraries. A detailed schema table provides field IDs, the associated MIxS field, field names, definitions, examples, MIxS-compliant category, examples, expected values, proper syntax, and preferred units, if applicable. A second supporting table provides EnvO definitions for appropriate environmental descriptors, and defines criteria for assigning particular terms based on environmental measurements such as pH, temperature, salinity, or pressure. The alignment of MgOl's environmental context to applicable standards and ontologies makes the database more robust, ensuring its longevity and increasing its compatibility with user-driven queries and investigation through the VIROME pipeline.

The implementation of a new VIROME library submission interface designed to collect MIxS-compliant metadata on future submissions was achieved through revisions to the submission form's organization, the automatic assignment of EnvO terms and other classifications, and the creation of a dynamic and adaptive process to facilitate user experience. Modifications improve the form's readability and appearance, and metadata are organized in a MIxS-compliant format. The selection of common environmental descriptors such as ecosystem guide the automatic assignment of EnvO-compliant biome, environmental feature, and environmental material terms, and of additional classifications including Koppen-Geiger climate classification, latitude zone, and depth zone, if applicable. The dynamic and adaptive form guides the submission process, conditionally displaying particular fields or groups of metadata fields based on previous user selection. Extensive validation is in place to provide an accurate and comprehensive metadata set able to provide environmental context to the metagenome submission. MgOl is expanded further as libraries are submitted to VIROME, such that the collection of such comprehensive environmental metadata at the time of library submission facilitates the MIxS-compliant growth of the database as libraries are added to MgOl and available for future analysis.

Additional opportunities were designed to leverage the improved environmental annotation through modification of the VIROME and MgOl library pages, the design of user outputs which provide useful annotation and facilitate submission to other tools, and the design of additional opportunities to compare, search, and group within VIROME by metadata. Improved VIROME and MgOl library pages display all library metadata reinforcing the importance of a complete and comprehensive annotation, group metadata by MIxS-compliant categories, and

provide a uniform format across tools that serves to facilitate user experience and promote exploration of the tools and libraries. Metadata is exported in text file and printable output formats. Multiple text file outputs are compatible with GenBank BioProject, BioSample, or Sequence Read Archive submissions, and reinforce data sharing across resources, create an opportunity to publicize VIROME, and potentially create a preference for VIROME analyses which facilitate INSDC-compliant submissions. The printable output is provided when VIROME analysis is complete and available, and is available in the same uniform format in which metadata is presented in VIROME and MgOl library pages. Environmental context provides new ways to explore a library or sequence, organizing database hits by environment or by environmental measurements for further query or investigation. Proposed changes to VIROME's browse, search, and compare views create flexibility to group, sort, and compare by various environmental descriptors or conditions, and allow users to identify and select particular libraries to explore in those views.

Through updates to the MgOl database, the VIROME library submission process, and subsequent library exploration, VIROME and MgOl capture MIXS-compliant metadata in alignment with relevant standards and ontologies. VIROME and MgOl utilize improved environmental context and visualizations to provide flexible user-driven grouping and comparison and facilitate deeper examination of a dataset, more relevant insights into its significance, and continued study of viral community diversity.

Future Considerations

The number of freely available tools provides potential users with multiple options, and user preference can be determined based on analysis speed, visualization

options, and depth of analysis. Many tools provide instant taxonomic or functional assessment, without the links to supporting information or sequence-specific annotation which make VIROME so powerful. However, the instant assessment is appealing to many users, and would be a valuable addition to the VIROME suite of tools. The incorporation of a rapid assessment, with more thorough investigation available at the completion of the full VIROME pipeline, may draw additional users. In addition, emerging visualization tools capitalize on new technologies such as HTML5, rendering flexible graphical interfaces for both web sites and local applications. Concurrent updates to VIROME and MgOl are incorporating HTML5 for more widely available and user-friendly ways to explore metagenomes. VIROME team members continue to evaluate new and emerging tools, looking for ways to advance VIROME's functionality and relevance.

REFERENCES

- Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., ... Rohwer, F. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, 6, 41. doi:10.1186/1471-2105-6-41
- Ashburner, M., Ball, C. a, Blake, J. a, Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29. doi:10.1038/75556
- Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., ... Ostell, J. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Research*, 40(December 2011), 57–63. doi:10.1093/nar/gkr1163
- Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K., & Wommack, K. E. (2007). Metagenomic characterization of Chesapeake Bay virioplankton. *Applied and Environmental Microbiology*, 73(23), 7629–7641. doi:10.1128/AEM.00938-07
- Benson, D. a., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(November 2012), 36–42. doi:10.1093/nar/gks1195
- Bohnebeck, U., Lombardot, T., Kottmann, R., & Glöckner, F. O. (2008). MetaMine--a tool to detect and analyse gene patterns in their environmental context. *BMC Bioinformatics*, 9, 459. doi:10.1186/1471-2105-9-459
- Brady, A., & Salzberg, S. (2009). Classification with Interpolated Markov Models. *Nature Methods*, 6(9), 673–676. doi:10.1038/nmeth.1358
- Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., & Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology*, 185(20), 6220–6223. doi:10.1128/JB.185.20.6220-6223.2003

- Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., ... Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 14250–14255. doi:10.1073/pnas.202488399
- Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal file:///C:/Users/L%20I%20N/Desktop/ISMAEL/Maestria/cienciometria/ontologia cienciometria.pdf of Biomedical Semantics*, 4, 43. doi:10.1186/2041-1480-4-43
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Walters, W. a. (2011). NIH Public Access, 7(5), 335–336. doi:10.1038/nmeth.f.303.QIIME
- Cochrane, G., Karsch-mizrachi, I., & Nakamura, Y. (2010). The International Nucleotide Sequence Database Collaboration.pdf, 1–4. doi:10.1093/nar/gks1084
- Cochrane, G., Karsch-mizrachi, I., & Nakamura, Y. (2011). The International Nucleotide Sequence Database Collaboration.pdf, 39(November 2010), 15–18. doi:10.1093/nar/gks1084
- Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., ... Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(November 2008), 141–145. doi:10.1093/nar/gkn879
- Edwards, R. A., & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, 3(6), 504–510. Retrieved from <http://www.nature.com/nrmicro/journal/v3/n6/abs/nrmicro1163.html>
- Field, D. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, 26(5), 541–547. doi:10.1038/1360
- Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., ... Wooley, J. (2011). The Genomic Standards Consortium. *PLoS Biology*, 9(6), 8–10. doi:10.1371/journal.pbio.1001088
- Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., ... Jackson, R. B. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*, 73(21), 7059–7066. doi:10.1128/AEM.00358-07

- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., ... Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Research*, 40, 1–7. doi:10.1093/nar/gkr991
- George, C. A. (2005). Usability testing and design of a library website: an iterative approach. *OCLC Systems & Services*, 21, 167–180. doi:10.1108/10650750510612371
- Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. a., & Yooseph, S. (2010). METAREP: JCVI metagenomics reports-an open source tool for high-performance comparative metagenomics. *Bioinformatics*, 26(20), 2631–2632. doi:10.1093/bioinformatics/btq455
- Hamady, M., Knight, R., Stern, A., Mick, E., & Tirosh, I. (2012). Tools , techniques , and challenges Microbial community profiling for human microbiome projects : Tools , techniques , and challenges, (303), 1141–1152. doi:10.1101/gr.085464.108
- Hankeln, W., Buttigieg, P. L., Fink, D., Kottmann, R., Yilmaz, P., & Glöckner, F. O. (2010). MetaBar - a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics*, 11, 358. doi:10.1186/1471-2105-11-358
- Hankeln, W., Wendel, N. J., Gerken, J., Waldmann, J., Buttigieg, P. L., Kostadinov, I., ... Glöckner, F. O. (2011). Cdiffusion - Submission-Ready, On-Line integration of sequence and contextual data. *PLoS ONE*, 6(9), 1–7. doi:10.1371/journal.pone.0024797
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17, 377–386. doi:10.1101/gr.5969107
- Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., & Schuster, S. C. (2009). Methods for comparative metagenomics. *BMC Bioinformatics*, 10 Suppl 1, S12. doi:10.1186/1471-2105-10-S1-S12
- Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., ... Nakamura, Y. (2011). DDBJ progress report. *Nucleic Acids Research*, 39(September), 1–6. doi:10.1093/nar/gkq1041
- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., ... Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(December 2007), 480–484. doi:10.1093/nar/gkm882

- Kitts, P. A., Madden, T. L., Sicotte, H., & Black, L. (2011). UniVec Database. Retrieved from <http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3), 259–263. doi:10.1127/0941-2948/2006/0130
- Lauber, C., & Gorbalenya, a. E. (2012). Toward Genetics-Based Virus Taxonomy: Comparative Analysis of a Genetics-Based Classification and the Taxonomy of Picornaviruses. *Journal of Virology*, 86, 3905–3915. doi:10.1128/JVI.07174-11
- Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., ... Cochrane, G. (2011). The European nucleotide archive. *Nucleic Acids Research*, 39, 1–4. doi:10.1093/nar/gkq967
- Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, 39(November 2010), 2010–2012. doi:10.1093/nar/gkq1019
- Leplae, R., Hebrant, A., Wodak, S. J., & Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Research*, 32, D45–D49. doi:10.1093/nar/gkh084
- Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, 10, 359. doi:10.1186/1471-2105-10-359
- Lombardot, T., Kottmann, R., Giuliani, G., de Bono, A., Addor, N., & Glöckner, F. O. (2007). MetaLook: a 3D visualisation software for marine ecological genomics. *BMC Bioinformatics*, 8, 406. doi:10.1186/1471-2105-8-406
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964. doi:10.1093/nar/25.5.955
- Lozupone, C., & Knight, R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12), 8228–8235. doi:10.1128/AEM.71.12.8228
- Meyer, F., Paarmann, D., D’Souza, M., & Etal. (2008). The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional

- analysis of metagenomes. *BMC Bioinformatics*, 9, 386. doi:10.1186/1471-2105-9-386
- Muhling, M., Fuller, N. J., Millard, A., Somerfield, P. J., Marie, D., Willson, W. H., ... Mann, N. H. (2005). Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environmental Microbiology*, 7, 499–508.
- Niu, B., Fu, L., Sun, S., & Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, 11, 187.
- Noguchi, H., Taniguchi, T., & Itoh, T. (2008). Meta gene annotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Research*, 15, 387–396. doi:10.1093/dnares/dsn027
- Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1), 385. doi:10.1186/1471-2105-12-385
- Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., ... Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), 5691–5702. doi:10.1093/nar/gki866
- Peel, B. L., Finlayson, B. L., & McMahon, T. a. (2007). Updated world map of the Koppen-Geiger climate classification.pdf. *Hydrology and Earth System Sciences*, 11, 1633–1644. Retrieved from <http://www.hydrol-earth-syst-sci.net/11/1633/2007/hess-11-1633-2007.pdf>
- Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21), 7188–7196. doi:10.1093/nar/gkm864
- Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., ... Sansone, S. A. (2010). ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18), 2354–2356. doi:10.1093/bioinformatics/btq415
- Rohwer, F., & Edwards, R. (2002). The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, 184(16), 4529–4535. doi:10.1128/JB.184.16.4529-4535.2002

- Rohwer, F., Prangishvili, D., & Lindell, D. (2009). Role of viruses in the environment. *Environmental Microbiology*, *11*(11), 2771–2774.
- Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y., & Breitbart, M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology*, *11*, 2806–2820. doi:10.1111/j.1462-2920.2009.01964.x
- Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., & Enault, F. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics*, *27*, 3074–3075.
- Roux, S., Tournayre, J., Mahul, A., Debroas, D., & Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics*, *15*, 76. doi:10.1186/1471-2105-15-76
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., ... Musen, M. a. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *Omics : A Journal of Integrative Biology*, *10*(2), 185–198. doi:10.1089/omi.2006.10.185
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ... Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, *5*(3), 0398–0431. doi:10.1371/journal.pbio.0050077
- Schmidt, H. F., Sakowski, E. G., Williamson, S. J., Polson, S. W., & Wommack, K. E. (2014). Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *The ISME Journal*, *8*, 103–14. doi:10.1038/ismej.2013.124
- Seshadri, R., Kravitz, S. a., Smarr, L., Gilna, P., & Frazier, M. (2007). CAMERA: A community resource for metagenomics. *PLoS Biology*, *5*(3), 0394–0397. doi:10.1371/journal.pbio.0050075
- Staley, J. T., & Konopka, A. (1985). Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology*, *39*(October), 321–346.
- Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*, 256–361.
- Suttle, C. A. (2007). Marine viruses - major players in the global ecosystem. *Nature Review Microbiology*, *5*, 801–812.

- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10), 1282–1288. doi:10.1093/bioinformatics/btm098
- Tatusov, R. L., Galperin, M. Y., Natale, D. a, & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1), 33–36. doi:10.1093/nar/28.1.33
- Technologies, E., Lindgaard, G., & Oriented, H. (2007). Aesthetics , Visual Appeal , Usability and User Satisfaction : What Do the User ' s Eyes Tell the User ' s Brain ? *Society*, 5, 1–14. Retrieved from <http://www.doaj.org/doaj?func=abstract&id=235801>
- Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, 45, 1320–1328.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164), 804–810. doi:10.1038/nature06244.The
- UniProt knowledgebase. (n.d.). Retrieved from <http://www.uniprot.org/help/uniref>
- Van Regenmortel, M. H., Fauquet, C. M., Bishop, D. H., Carstens, E. B., Estes, M. K., Lemon, S. M., ... Wickner, R. B. (2000). *Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses*. Academic Press.
- Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., van Elsas, J. D., ... Philippot, L. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Micro*, 7(4), 252. Retrieved from <http://dx.doi.org/10.1038/nrmicro2119>
- Weynberg, K. D., Wood-Charlson, E. M., Suttle, C. a., & van Oppen, M. J. H. (2014). Generating viral metagenomes from the coral holobiont. *Frontiers in Microbiology*, 5(May), 1–11. doi:10.3389/fmicb.2014.00206
- Wommack, K. E., Bench, S. R., Bhavsar, J., Mead, D., & Hanson, T. E. (2009). Isolation independent methods of characterizing phage communities 2: characterizing a metagenome. *Methods in Molecular Biology*, 502, 279–289.

- Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., ... Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, 6, 427–439. doi:10.4056/sigs.2945050
- Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews : MMBR*, 64(1), 69–114. doi:10.1128/MMBR.64.1.69-114.2000
- Wommack, K. E., Srinivasiah, S., Liles, M., Bhavsar, J., Bench, S., Williamson, K. E., & Polson, S. W. (2011). Metagenomic contrasts of viruses in soil and aquatic environments. In Fj. Bruijn (Ed.), *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*. New York: John E. Wiley & Sons.
- Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2). doi:10.1371/journal.pcbi.1000667
- Yancey, P. (MarineBio). (2011). The Deep Sea ~ Ocean biology, Marine life, Sea creatures, Marine conservation... ~ MarineBio.org. Retrieved from <http://marinebio.org/oceans/deep/>
- Yilmaz, P., Gilbert, J. a., Knight, R., Amaral-Zettler, L., Karsch-Mizrachi, I., Cochrane, G., ... Field, D. (2011). The genomic standards consortium: bringing standards to life for microbial ecology, 1565–1567. doi:10.1038/ismej.2011.39
- Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., ... Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nature Biotechnology*, 29(5), 415–420. doi:10.1038/nbt.1823

Appendix A
METAGENOMES ONLINE SCHEMA

Table 1 Detailed Metagenomes Online database schema. Schema is based upon MlXS table of library descriptors, including field ID, definition, example, field type, syntax, MlXS-compliant category, and preferred units, if applicable.

MgOI field	VIROME field	MlXS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
mgol_id			MgOI ID	unique numeric MgOI identifier	1	submission number	auto	investigation	int(6)	
prefix			library prefix	Unique 3 letter combination identifying the library	AUV	3 letter combination	auto	investigation	varchar(6)	
name	ls_name		library name or description	Name of the library within which the sequences were organized		text	manual entry	investigation	varchar(125)	
desc	ls_desc		library description	Description of the library		text	manual entry	investigation	varchar(125)	
project	ls_projectselect	project_name	project name	Description of the overall project within which the library sequencing was conducted		enumeration	selection	investigation	['list of user projects']	
project	ls_project	project_name	project name	Description of the overall project within which sequencing was conducted		text	manual entry	investigation	varchar(125)	
data_src			data source	Description of the data source, whether through VIROME submission or data mining source	CAMERA	text	curator or automatic through VIROME submission	investigation	varchar(50)	
citation	ls_citation		citation or reference	Reference for project within which sequencing was organized		text	manual entry	investigation	varchar(125)	
intel_prop	ls_intel_prop		intellectual property	Statement of intellectual property			manual entry	investigation	text	
investype	ls_investype	investigation_type	investigation type	Submission type, or source of sequencing	Metagenome	text	selection	investigation	varchar(30)	
ncbi	ls_ncbi	submitted_to_insd	submitted to NCBI	Library submitted to NCBI	Yes	text	radio	investigation	varchar(10)	
bipproject	ls_bioproject		BioProject	NCBI BioProject ID		text	manual entry	investigation	varchar(30)	
biosample	ls_biosample		BioSample	NCBI BioSample ID		text	manual entry	investigation	varchar(30)	
accession	ls_accession		accession	NCBI Library accession number		text	manual entry	investigation	varchar(30)	
publish	ls_publish		public data	Data can be made publically available within VIROME	Yes	text	radio	investigation	varchar(10)	
assemble	ls_assemble		assembled	Is the data assembled		text	radio	sequencing	varchar(10)	
assemblymethod	ls_assemblymethod	assembly_name, assembly_method	assembly	Description of assembly methods		text	manual entry	sequencing	varchar(200)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
seqmethod	ls_seqmethod	seq_meth	sequencing method	Sequencing methodology used	Sanger	text	multiple selection	sequencing	varchar(30)	
seqcenter	ls_seqcenter		sequencing center	Name of center at which sequencing was conducted		text	manual entry	sequencing	varchar(50)	
amplification	ls_amplification	nucI_acid_amp	amplification type	Amplification methodology used	Linker Amplification	text	selection	sequencing	varchar(30)	
sequencingmethod	ls_sequencingmethod	seq_meth	sequencing details	Sequencing methodology details		text	manual entry	sequencing	varchar(255)	
fragmentsize	ls_fragmentsize		fragment size	Number of base pairs; size of fragments put into sequencing including any adapters		numeric	manual entry	sequencing	integer	base pairs
replicons	ls_replicons	num_replicons	number of replicons	Number of replicons in the genome of a bacterium		for bacteria: chromosomes		nucleic acid sequence source		
pathogenicity	ls_pathogenicity	pathogenicity	pathogenicity	To what is the entity pathogenic	human, animal, plant, fungi, bacteria	text	selection	nucleic acid sequence source	varchar(30)	
propagation	ls_propagation	propagation	propagation	Propagation mechanism specific to different taxa		text	selection	nucleic acid sequence source	varchar(30)	
bioticrel	ls_bioticrel	biotic_relationship	biotic relationship	Is it free-living or in a host and if the latter what type of relationship is observed		text	selection	nucleic acid sequence source	varchar(30)	
growthcond	ls_growthcond	isol_growth_cond	isolation or growth conditions	Publication reference in the form of pubmed ID (PMID), digital object identifier (DOI), or url for isolation and growth condition specifications of the organism/material		PMID, DOI, or URL	manual entry	nucleic acid sequence source	varchar(125)	
acidtype	ls_acidtype		nucleic acid type	Type and strand structure of nucleic acid	DNA	text	selection	nucleic acid sequence source	varchar(50)	
filter_lowerbound	ls_filter_lowerbound		lower filter range (µm)	Lower range (µm) of filters used prior to sequencing		numeric	manual entry	nucleic acid sequence source	float	µm
filter_upperbound	ls_filter_upperbound		upper filter range (µm)	Upper range (µm) of filters used prior to sequencing		numeric	manual entry	nucleic acid sequence source	float	µm
metatype	ls_metatype		library type	For metagenome submissions, a description of the sample fraction submitted for sequencing	Viral	text	auto	investigation	varchar(30)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
sampcoll	ls_sampcoll	samp_collect_device	sample collection device or method	The method or device employed for collecting the sample	biopsy, niskin bottle, push core	text	manual entry	nucleic acid sequence source	varchar(50)	
sampproc	ls_sampproc	samp_mat_process	sample material processing	Any processing applied to the sample during or after retrieving the sample from the environment	filtering of seawater, storing samples in ethanol	text	manual entry	nucleic acid sequence source	varchar(50)	
sampsize	ls_sampsize	samp_size	amount or size of sample collected	Amount of size of sample (volume, mass, or area) that was collected		text	manual entry	nucleic acid sequence source	varchar(50)	
pooling	ls_pooling		pooling	Submitted library from a collection of pooled samples	Yes	text	radio	environment	varchar(10)	
samples	ls_samples		number of sites	Number of sites from which sample was collected		integer	manual entry	environment	int(4)	
pool_sec	ls_pool_desc		description of pooling	Description of pooling methodology, including whether pooling was done among sites, times, depth, etc		date	manual entry	environment	date	
sampdate	ls_sampdate	collection_date	sample/sampling date	Date upon which sample was collected		text	manual entry	environment	varchar(20)	
latitude	ls_latitude	lat_lon	latitude degrees	Latitude decimal degrees coordinate of the location at which sample was collected		decimal degrees	manual entry	environment	float	
latitude	ls_lat_loc	lat_lon	latitude hemisphere	Latitude hemisphere coordinate of the location at which sample was collected	N	decimal degrees	radio	environment	float	
longitude	ls_longitude	lat_lon	longitude degrees	Longitude decimal degrees coordinate of the location at which sample was collected		decimal degrees	manual entry	environment	float	
longitude	ls_long_loc	lat_lon	longitude hemisphere	Longitude hemisphere coordinate of the location at which sample was collected	E	decimal degrees	radio	environment	float	
country	ls_country	geo_loc_name	geographic country	Country in which sample was collected		text	manual entry	environment	varchar(50)	
region	ls_region	geo_loc_name	geographic region	Region within the country in which sample was collected		text	manual entry	environment	varchar(150)	
geog_place_name	ls_place	geo_loc_name	geographic place name	Specific place name in which sample was collected		text	manual entry	environment	varchar(150)	
identifier	ls_identifier		site identifier	Description of or unique identifier for sample collection site(s)		text	manual entry	environment	varchar(100)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
genesis	ls_genesis		sample genesis	Origin of sample		text	selection	environment	varchar(30)	
sphere	ls_sphere		sphere	Sphere of sample origin		text	selection	environment	varchar(30)	
ecosystem	ls_ecosystem		ecosystem	Representative ecosystem description of sample collection site		text	selection	environment	varchar(30)	
landuse	ls_landuse		agricultural land use	For Agricultural ecosystems, use of land		text	selection	environment	varchar(30)	
phys_subst	ls_phys_subst		physical substrate	Description of the sample substrate		text	selection	environment	varchar(30)	
envpackage	ls_envpackage	env_package	environmental package	From MIxS, a description of the environmental type for reporting of measurements and observations		text	selection	environment	varchar(30)	
physio_chem_mods	ls_physio_chem_mods		physio-chemical modifiers	Conditions, physical or chemical, that modify the environment and potentially make it extreme		text	multiple selection	environment	varchar(100)	
altitude	ls_altitude	altitude	altitude in m	Vertical distance between Earth's surface above sea level and the sampled position in the air		measurement value	manual entry	environment	float	m
depth_zone	ls_depth	depth	sample depth	Vertical distance below local surface		measurement value	manual entry	environment	float	m
totaldepth	ls_totaldepth		water depth	Measurement of total depth of water column		measurement value	manual entry	environment	float	m
elevation	ls_elevation	elev	elevation in meters	The elevation of the sampling site as measured by the vertical distance from mean sea level		measurement value	manual entry	environment	float	m
humidity	ls_humidity	humidity	humidity	Amount of water vapor in the air at the time of sampling		measurement value	manual entry	environment	float	gram per cubic meter
ph	ls_ph	pH	pH	Measurement of pH		measurement value	manual entry	environment	float	
pressure	ls_pressure	pressure	pressure	Pressure to which the sample is subject		measurement value	manual entry	environment	float	atm
salinity	ls_salinity	salinity	salinity psu	Measurement of salinity		measurement value	manual entry	environment	float	%
temperature	ls_temperature	temp	temperature degrees celsius	Temperature of the sample at the time of sampling		measurement value	manual entry	environment	float	C

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
biomass	ls_biomass	biomass	biomass $\mu\text{g/kg}$	Amount of biomass. For MIxS, should include the name for the part of biomass measured, e.g. microbial, total. Can include multiple measurements		measurement value	manual entry	environment	float	$\mu\text{g/kg}$
chlorophyll	ls_chlorophyll	chlorophyll	chlorophyll $\mu\text{g/kg}$	Concentration of chlorophyll		measurement value	manual entry	environment	float	$\mu\text{g/kg}$
conductivity	ls_conductivity	conduc	conductivity	Electrical conductivity of water			manual entry	environment	float	millSiemens per centimeter
density	ls_density	density	density	Density of sample			manual entry	environment	float	gram per cubic meter
dic	ls_dic	diss_inorg_carb	dissolved inorganic carbon $\mu\text{mol/kg}$	Concentration of dissolved inorganic carbon		measurement value	manual entry	environment	float	$\mu\text{mol/kg}$
dip	ls_dip	diss_inorg_phosp	dissolved inorganic phosphorus nmol/kg	Concentration of dissolved inorganic phosphorus		measurement value	manual entry	environment	float	nmol/kg
doc	ls_doc	diss_org_carb	dissolved organic carbon $\mu\text{mol/kg}$	Concentration of dissolved organic carbon		measurement value	manual entry	environment	float	$\mu\text{mol/kg}$
do	ls_do2	diss_oxygen	dissolved oxygen nmol/kg	Concentration of dissolved oxygen		measurement value	manual entry	environment	float	nmol/kg
no3	ls_no3	nitrate	nitrate nmol/kg	Concentration of nitrate		measurement value	manual entry	environment	float	nmol/kg
taxid	ls_taxid	host_taxid	host taxonomic name	Taxonomic name of the host		text	manual entry	environment	text	
host_cn	ls_host_cn	host_common_name	host common name	Common name of the host	human	text	manual entry	environment	varchar(100)	
subjectId	ls_subjectId	host_subject_id	host subject ID	A unique identifier by which each subject can referred to, de-identified		text	manual entry	environment	varchar(30)	
orgsubstrate	ls_orgsubstrate	host_substrate	organism substrate	If from a host-associated sample, the substrate of the host (different from the substrate of the sample, organic material)	water	text	manual entry	environment	varchar(200)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
abshumidity	ls_abshumidity	abs_air_humidity	absolute humidity	Actual mass of water vapor - mh20 - present in the air water vapor mixture		measurement value	manual entry	environment	float	
relhumidity	ls_relhumidity	rel_air_humidity	relative humidity	Partial vapor and air pressure, density of the vapor and air, or by the actual mass of the vapor and air		measurement value	manual entry	environment	float	
occupancytype	ls_occupancytype	build_occup_type	building occupancy type	The primary function for which a building or discrete part of a building is intended to be used		text	selection	environment	varchar(30)	
setting	ls_setting	building_setting	building setting	A location (geography) where a building is set		text	selection	environment	varchar(30)	
carbondioxide	ls_carbondioxide	carb_dioxide	carbon dioxide	Carbon dioxide (gas) concentration at the time of sampling		measurement value	manual entry	environment	float	
filter	ls_filter	filter_type	filter type	A device which removed solid particles or airborne molecular contaminants		text	selection	environment	varchar(50)	
hvac	ls_hvac	heat_cool_type	heating and cooling system	Methods of conditioning or heating a room or building		text	selection	environment	varchar(30)	
space	ls_space	indoor_space	indoor space purpose	A distinguishable space within a structure, the purpose for which discrete areas of a building is used		text	selection	environment	varchar(30)	
light	ls_light	light_type	light type	Application of light to achieve some practical or aesthetic effect. Lighting includes the user of both artificial light sources such as lamps and light fixtures, as well as natural illumination by capturing daylight. Can also includes absence of light		text	selection	environment	varchar(30)	
occupancy	ls_occupancy	occup_samp	occupancy	Number of occupants at time of sampling		measurement value	manual entry	environment	integer	number of occupants
occdensity	ls_occdensity	occupant_dens_samp	occupancy density	Number of occupants per square footage at time of sampling		measurement value	manual entry	environment	integer	number of occupants per square footage

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
typicoccdensity	ls_typicoccdensity	typ_occupant_dens	typical occupancy density	Typical number of occupants per square footage		measurement value	manual entry	environment	integer	number of occupants per square footage
state	ls_state	space_typ_state	space typical state	Space typical state		text	selection	environment	varchar(30)	
ventilation	ls_ventilation	ventilation_type	ventilation type	Ventilation system used in the sampled premises		text		environment	varchar(50)	
misc	ls_misc	misc_param	miscellaneous	Any other measurement performed or parameter collected, that is not listed here						
envo_biome_id	ls_envo_biome_id	env_biome	EnvO biome term ID number	Biome ID number according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Sphere and Ecosystem		EnvO	auto	environment	varchar(200)	
envo_biome_name	ls_envo_biome_name	env_biome	EnvO biome term ID name	Biome name according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Sphere and Ecosystem		EnvO	auto	environment	varchar(200)	
biome_acc	ls_biome_acc			Acceptance of assigned biome term		text	selection	environment	varchar(30)	
envo_biome_id_user	ls_envo_biome_id_user	env_biome	EnvO biome term ID number - user assigned	Biome ID number according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_biome_name_user	ls_envo_biome_name_user	env_biome	EnvO biome term ID name - user assigned	Biome name according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_feature_id	ls_envo_feature_id	env_feature	EnvO environmental feature term ID number	Environmental feature ID number according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Sphere and Ecosystem.		EnvO	auto	environment	varchar(200)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
envo_feature_name	ls_envo_feature_name	env_feature	EnvO environmental feature term ID name	Environmental feature name according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Sphere and Ecosystem		EnvO	auto	environment	varchar(200)	
feature_acc	ls_feature_acc			Acceptance of assigned feature term		text	selection	environment	varchar(30)	
envo_feature_id_user	ls_envo_feature_id_user	env_feature	EnvO environmental feature term ID number - user assigned	Environmental feature ID number according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_feature_name_user	ls_envo_feature_name_user	env_feature	EnvO environmental feature term ID name - user assigned	Environmental feature name according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_material_id	ls_envo_material_id	env_material	EnvO environmental material term ID number	Environmental material ID number according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Physical Substrate.		EnvO	auto	environment	varchar(200)	
envo_material_name	ls_envo_material_name	env_material	EnvO environmental material term ID name	Environmental material name according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Physical Substrate.		EnvO	auto	environment	varchar(200)	
material_acc	ls_material_acc			Acceptance of assigned material term		text	selection	environment	varchar(30)	
envo_material_id_user	ls_envo_material_id_user	env_material	EnvO environmental material term ID number - user assigned	Environmental material ID number according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_material_name_user	ls_envo_material_name_user	env_material	EnvO environmental material term ID name - user assigned	Environmental material name according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
envo_habitat_id	ls_envo_habitat_id		EnvO habitat term ID number	Habitat ID number according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Physio-Chem Modifiers.		EnvO	auto	environment	varchar(200)	
envo_habitat_name	ls_envo_habitat_name		EnvO habitat term ID number	Habitat name according to the Environment Ontology (EnvO). This term is automatically assigned based on selected descriptors for Physio-Chem Modifiers.		EnvO	auto	environment	varchar(200)	
habitat_acc	ls_habitat_acc			Acceptance of assigned habitat term		text	selection	environment	varchar(30)	
envo_habitat_id_user	ls_envo_habitat_id_user		EnvO habitat term ID number - user assigned	Habitat ID number according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
envo_habitat_name_user	ls_envo_habitat_name_user		EnvO habitat term ID name - user assigned	Habitat name according to the Environment Ontology (EnvO). This term is entered by the user.		EnvO	manual entry	environment	varchar(200)	
extreme			extreme environment	Identification of the environment as extreme or not extreme. Appropriate for child terms of "extreme habitat" in EnvO: acidic, alkaline, cold temperature, haline, high osmolarity, high pressure, high temperature		binary	auto	environment	binary(1)	
koppengeiger	ls_koppengeiger		Koppen Gieger climate zone	Assignment of Koppen-Geiger climate zone based on sample collection location	Csc	text	auto	environment	varchar(100)	
latitudezone	ls_latitudezone		latitude zone	Latitude zone of sample collection site, based on given latitude coordinates		text	auto	environment	varchar(30)	
depthzone	ls_depthzone		depth zone	Depth zone of a sample collection site, based on depth value provided		text	auto	environment	varchar(30)	

Table 1 continued

MgOI field	VIROME field	MIxS field	Item	Definition	Example	Expected value	Assignment	Section	Value syntax	Preferred units
avg_read_len			average read length	Average length of sequencing reads			auto	sequencing	float	
GC_pct			GC percentage	Percentage of GC content			auto	sequencing	float	
virome									binary(1)	
qryDb			query database	Indication of a query of this library against the MgOI database (compared to addition to database based on literature mining)			auto	investigation	tinyint(1)	
comments			comments	Miscellaneous and additional comments			manual entry	investigation	text	
deleted			deletion of library from MgOI database	Indication of the deletion of this library from the publically viewable MgOI database			auto	investigation	tinyint(1)	

Appendix B

METAGENOMES ONLINE TERM DEFINITIONS

Table 2 Detailed Metagenomes Online database field-specific term definitions. Definitions are extracted from the Environment Ontology and include associated environmental measurement criteria, if applicable.

VIROME field	Selection	Definition	Definition source			Criteria
			EnvO hierarchy	EnvO ID	EnvO name	
Ecosystem	Agricultural	none specific	feature	ENVO:00000077	agricultural feature	
LandUse	Cropland	An anthropogenic terrestrial biome which is primarily used for agricultural activity and which contains no village or larger human settlement	biome	ENVO:01000245	cropland biome	
LandUse	Rangeland	An anthropogenic terrestrial biome which is primarily used for the rearing and grazing of livestock.	biome	ENVO:01000247	rangeland biome	
Ecosystem	Coastal	Coastal water is a marine water body bordering a coast.	feature	ENVO:02000049	coastal water body	
Ecosystem	Coral Reef	Aragonite structures produced by living organisms, found in shallow, marine waters with little nutrients in the water.	feature	ENVO:00000150	coral reef	
Ecosystem	Desert	A region rendered barren or partially barren by environmental extremes, especially by low rainfall.	feature	ENVO:00000097	desert	
Ecosystem	Estuary	An area of water bordered by land on three sides.	feature	ENVO:00000032	bay	
Ecosystem	Forest	An area with a high density of trees. A small forest may be called a wood.	feature	ENVO:00000111	forest	
Ecosystem	Grassland	An area in which grasses (Graminae) are a significant component of the vegetation.	feature	ENVO:00000106	grassland	
Ecosystem	Industrial	A feature that has been constructed by deliberate human effort.	feature	ENVO:00000070	constructed feature	
Ecosystem	Lake	A body of water or other liquid of considerable size contained on a body of land.	feature	ENVO:00000020	lake	
Ecosystem	Mangrove	A swamp formed of trees and shrubs that grow in saline coastal habitats in the tropics and subtropics.	feature	ENVO:00000057	mangrove swamp	
Ecosystem	Open Ocean	Continuous saline-water bodies that surround the continents and fill the Earth's great depressions.	feature	ENVO:00000015	ocean	

Table 2 continued

VIROME field	Selection	Definition	Definition source			Criteria
			EnvO hierarchy	EnvO ID	EnvO name	
Ecosystem	Prairie	An area of land of low topographic relief that historically supported grasses and herbs, with few trees, and having generally a mesic (moderate or temperate) climate. Dominated by tall grasses (contrast steppe).	feature	ENVO:00000260	prairie	
Ecosystem	Rain Forest	Land having a cover of trees, shrubs, or both.	feature	ENVO:00000109	woodland	
Ecosystem	Salt Marsh	A wetland, featuring grasses, rushes, reeds, typhas, sedges, and other herbaceous plants (possibly with low-growing woody plants) in a context of shallow water.	feature	ENVO:00000035	marsh	
Ecosystem	Sewage	none specific	feature	ENVO:00002272	waste treatment plant	
Ecosystem	Spring	A point where groundwater or steam flows out of the ground, and is thus where the aquifer surface meets the ground surface or where there is a fissure.	feature	ENVO:00000027	spring	
Ecosystem	Stream	Linear body of water flowing on the Earth's surface.	feature	ENVO:00000023	stream	
Ecosystem	Subterranean	A habitat that is below the surface of the earth.	habitat	ENVO:00000572	subterrestrial habitat	
Ecosystem	Swamp	A wetland that features permanent inundation of large areas of land by shallow bodies of water, generally with a substantial number of hummocks, or dry-land protrusions.	feature	ENVO:00000233	swamp	
Ecosystem	Tundra	Treeless, level, or gently rolling plains characteristic of arctic or subarctic regions, having a permanently frozen subsoil, and usually supporting low growing vegetation such as lichens, mosses, and stunted shrubs.	feature	ENVO:00000112	tundra	
Ecosystem	Urban	Place or area with clustered or scattered buildings and a permanent human population.	feature	ENVO:00000062	populated place	
Ecosystem	Woodland	Land having a cover of trees, shrubs, or both.	feature	ENVO:00000109	woodland	
Ecosystem	Vent/Seep	A seep is a spring in which water has filtered through permeable earth to the surface.	feature	ENVO:01000262	seep	

Table 2 continued

VIROME field	Selection	Definition	Definition source			Criteria
			EnvO hierarchy	EnvO ID	EnvO name	
Physical Substrate	Aerosol	Airborne solid particles (also called dust or particulate matter (PM) or liquid droplets.	material	ENVO:00010505	aerosol	
Physical Substrate	Air	The mixture of gases (roughly (by molar content/volume: 78% nitrogen, 20.95% oxygen, 0.93% argon, 0.038% carbon dioxide, trace amounts of other gases, and a variable amount (average around 1%) of water vapor) that surrounds the planet Earth.	material	ENVO:0002005	air	
Physical Substrate	Anthropogenic	Anthropogenic material in or on which organisms may live.	material	ENVO:0010001	anthropogenic environmental material	
Physical Substrate	Clay	A group of hydrous aluminum phyllosilicates (phyllosilicates being a subgroup of silicate minerals) minerals (see clay minerals), that are typically less than 2 micrometres in diameter. Clay consists of a variety of phyllosilicate minerals rich in silicon and aluminum oxides and hydroxides which include variable amounts of structural water.	material	ENVO:00002982	clay	
Physical Substrate	Dust	Minute solid particles with diameters less than 500 micrometers. Occurs in and may be deposited from the atmosphere.	material	ENVO:00002008	dust	
Physical Substrate	Emulsion	A mixture of two immiscible (unblendable) substances. One substance (the dispersed phase) is dispersed in the other (the continuous phase).	material	ENVO:00010506	emulsion	
Physical Substrate	Foam	none specific	material	ENVO:00005738	foam	
Physical Substrate	Gravel	Gravel is an environmental material which is composed of pieces of rock that are at least 2 millimeters (2mm) in its largest dimension and no more than 75 millimeters.	material	ENVO:01000018	gravel	
Physical Substrate	Ice	Ice is water frozen into a solid state. It can appear transparent or opaque bluish-white color, depending on the presence of impurities or air inclusions. The addition of other materials such as soil may further alter its appearance.	material	ENVO:01000277	ice	

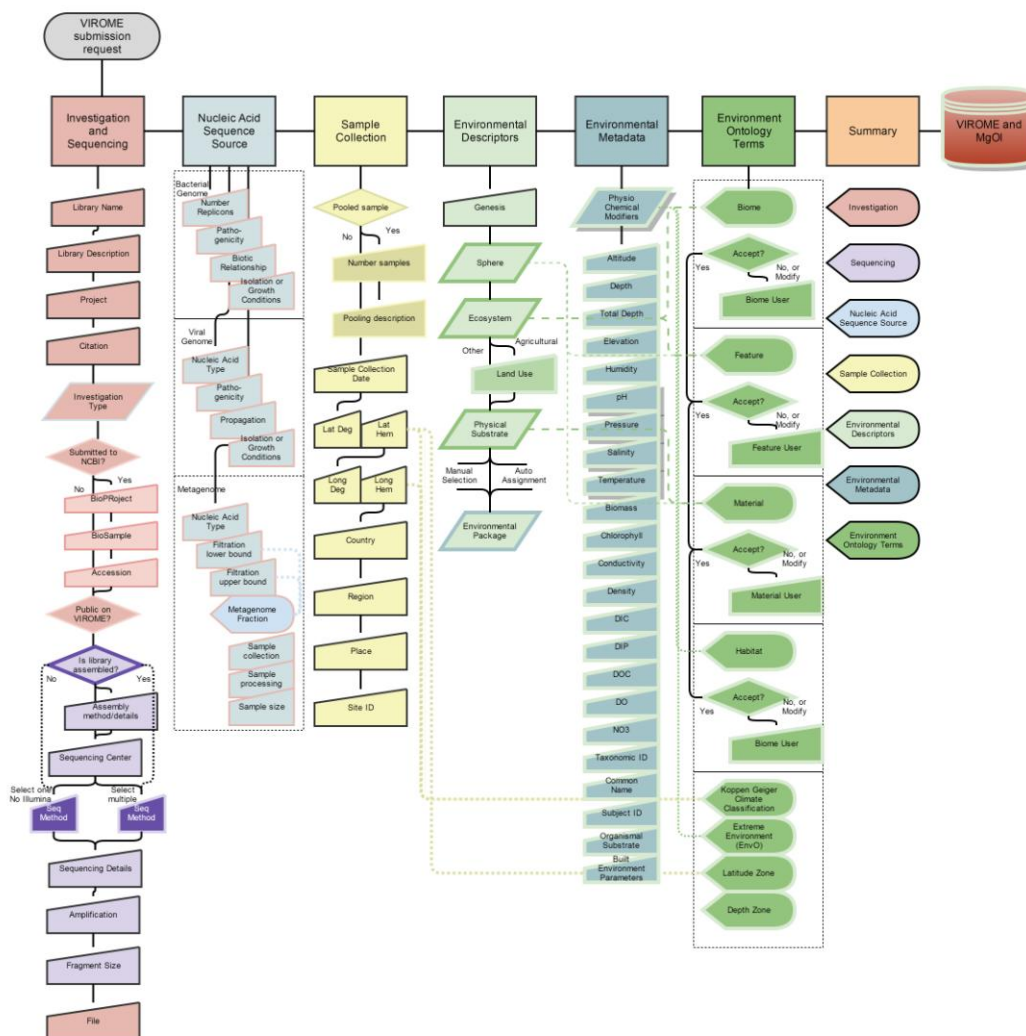
Table 2 continued

VIROME field	Selection	Definition	Definition source			Criteria
			EnvO hierarchy	EnvO ID	EnvO name	
Physical Substrate	Lava	Lava is an environmental material which is primarily composed of molten rock.	material	ENVO:01000231	lava	
Physical Substrate	Mineral	A mineral is an environmental material which is naturally occurring, solid and stable at room temperature, representable by a chemical formula, usually abiogenic, and has an ordered atomic structure.	material	ENVO:01000256	mineral	
Physical Substrate	Mud	A liquid or semi-liquid mixture of water and some combination of soil, silt, and clay.	material	ENVO:01000001	mud	
Physical Substrate	Oil	A viscous liquid state at ambient temperature or slightly warmer, and is both hydrophobic and lipophilic.	material	ENVO:00002985	oil	
Physical Substrate	Organic Material	Environmental material characteristic of, pertaining to, or derived from living organisms.	material	ENVO:01000155	organic material	
Physical Substrate	Rock	A rock is a naturally occurring solid aggregate of one or more minerals or mineraloids.	material	ENVO:00001995	rock	
Physical Substrate	Sand	A naturally occurring granular material composed of finely divided rock and mineral particles.	material	ENVO:01000017	sand	
Physical Substrate	Scree	Broken rock that appears at the bottom of crags, mountain cliffs, or valley shoulders.	material	ENVO:00000194	scree	
Physical Substrate	Scum	A layer of impurities that accumulates at the surface of a liquid (especially water or molten metal).	material	ENVO:00003930	scum	
Physical Substrate	Sediment	Sediment is an environmental substance comprised of any particulate matter that can be transported by fluid flow and which eventually is deposited as a layer of solid particles on the bed or bottom of a body of water or other liquid.	material	ENVO:00002007	sediment	
Physical Substrate	Silt	Silt is a granular material of a size somewhere between sand and clay whose mineral origin is quartz and feldspar.	material	ENVO:01000016	silt	
Physical Substrate	Soil	Any material within 2 m from the Earth's surface that is in contact with the atmosphere, with the exclusion of living organisms, areas with continuous ice not covered by other material, and water bodies deeper than 2 m.	material	ENVO:00001998	soil	
Physical Substrate	Vapor	A vapour is an environmental material in the gas phase at a temperature lower than its critical point.	material	ENVO:01000264	vapour	
Physical Substrate	Water	The liquid form of dihydrogen monoxide.	material	ENVO:00002006	water	

Table 2 continued

VIROME field	Selection	Definition	Definition source			Criteria
			EnvO hierarchy	EnvO ID	EnvO name	
Physio-Chem Modifiers	Acid	A habitat in which the pH is <ph3. Inhabited by acidophilic organisms.	habitat	ENVO:00002021	acid habitat	<3
Physio-Chem Modifiers	Alkaline	A habitat in which the pH is >ph9. Inhabited by alkaliphilic organisms.	habitat	ENVO:00002022	alkaline habitat	>9
Physio-Chem Modifiers	Arid	An arid condition is an environmental condition in which annual precipitation is less than half of annual potential evapotranspiration.	condition	ENVO:01000230	arid	<0.5 average precipitation
Physio-Chem Modifiers	Cold Temperature	A habitat characterized by an average temperature of 15 deg C or lower. Inhabited by psychrophilic (cryophilic) organisms.	habitat	ENVO:00002026	cold temperature habitat	<15 C average
Physio-Chem Modifiers	Haline	A habitat characterized by a concentration of salt at least 2M. Inhabited by halophilic organisms.	habitat	ENVO:00002024	haline habitat	>2 M
Physio-Chem Modifiers	High Osmolarity	A habitat characterized by a high osmolarity, typically the result of a high concentration of sugars. Inhabited by osmophilic organisms.	habitat	ENVO:00002028	high osmolarity habitat	
Physio-Chem Modifiers	High Pressure	A habitat characterized by high gas or liquid pressure, inhabited by barophilic (piezophilic) organisms.	habitat	ENVO:00002023	high pressure habitat	>380 atm
Physio-Chem Modifiers	High Temperature	A habitat characterized by an average temperature of at least 60 deg C. Inhabited by thermophilic organisms.	habitat	ENVO:00002025	high temperature habitat	>60 C average

VIROME SUBMISSION FORM FLOWCHART



116

Appendix D

VIROME SUBMISSION FORM FIELD DEPENDENCY FLOWCHARTS

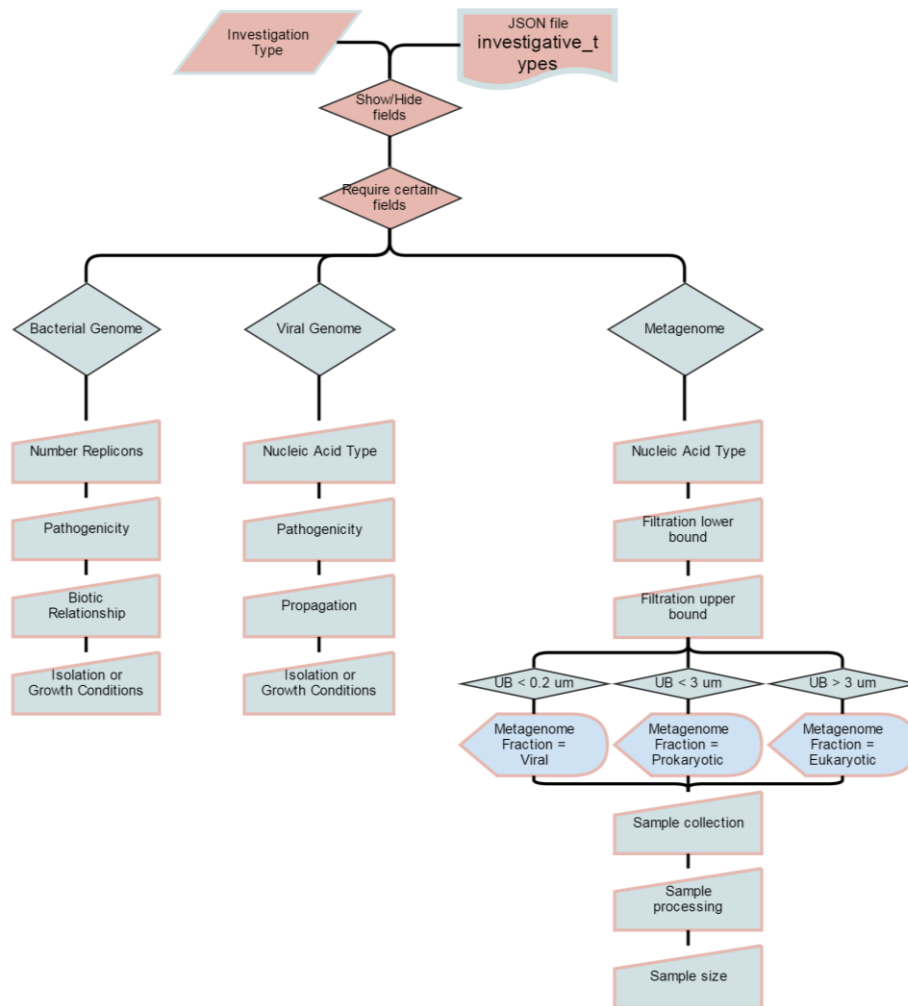


Figure 19 VIROME revised submission form flowchart for fields dependent upon investigation type selection.

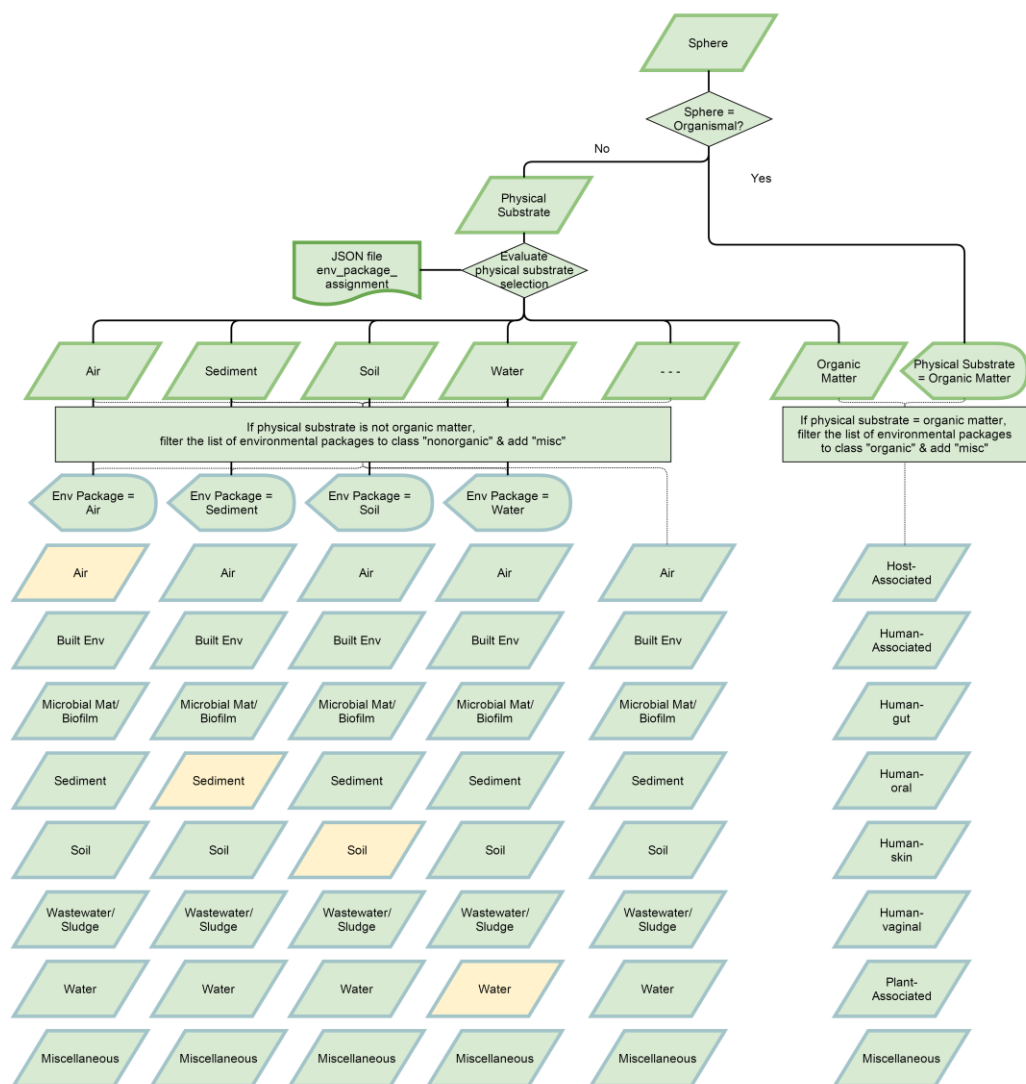


Figure 20 VIROME revised submission form flowchart for fields dependent upon environmental fields selections.

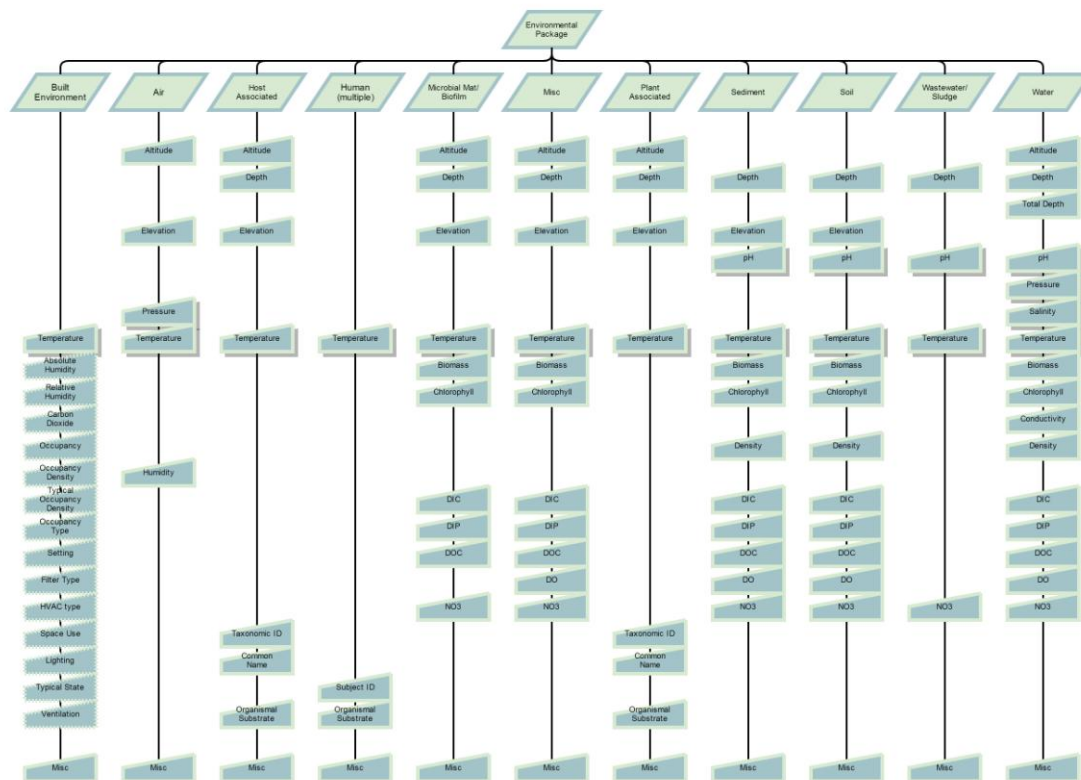


Figure 21 VIROME revised submission form flowchart for display of environmental measurement fields dependent upon environmental package selection.

Appendix E

REVISED VIROME AND MGOL LIBRARY PAGES

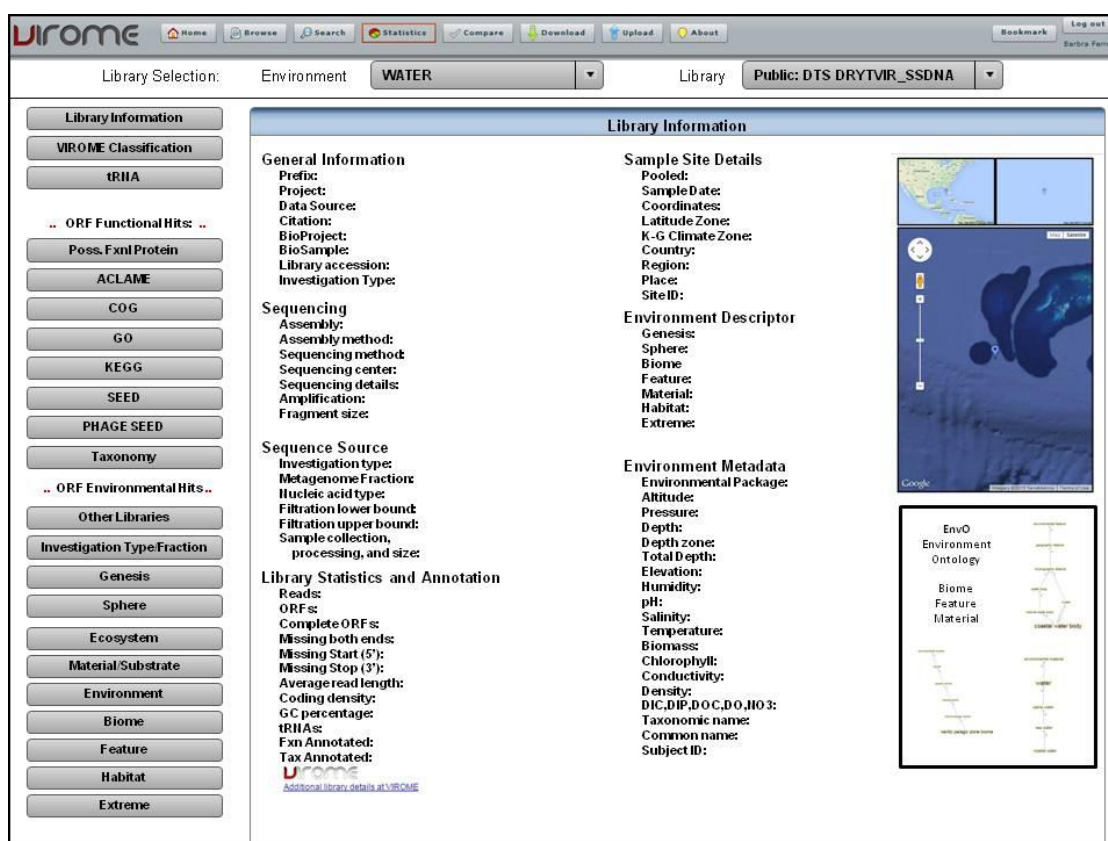


Figure 22 Proposed VIROME individual library information view. Displays all associated fields grouped by MIxS-compliant categories, provides supplemental graphics for environmental location and classification, and provides additional environmental groupings for BLAST result hits. A uniform display with MgOl provides consistency and improves ease-of-use across tools.

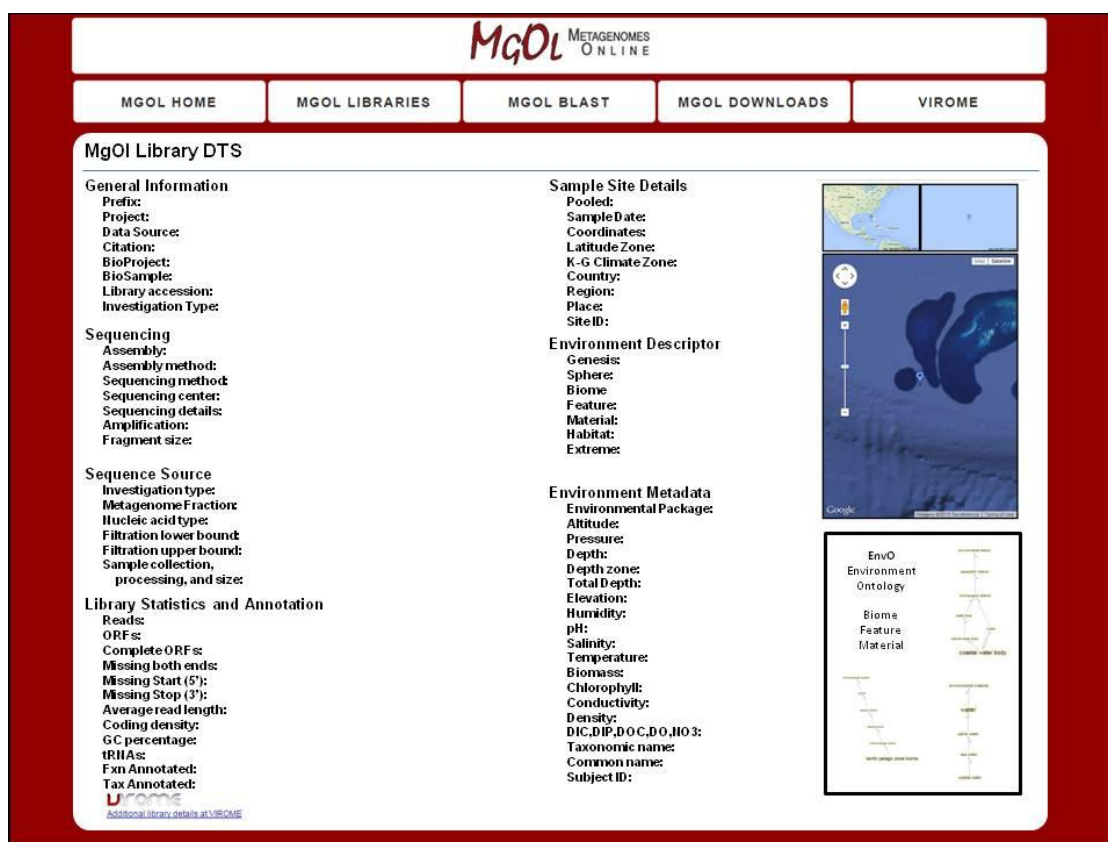


Figure 23 Proposed MgOL individual library information view. Displays all associated fields grouped by MIxS-compliant categories, provides supplemental graphics for environmental location and classification, and provides additional environmental groupings for BLAST result hits. A uniform display with VIROME provides consistency and improves ease-of-use across tools.

Appendix F

VIROME BROWSE VIEW OPTIONS

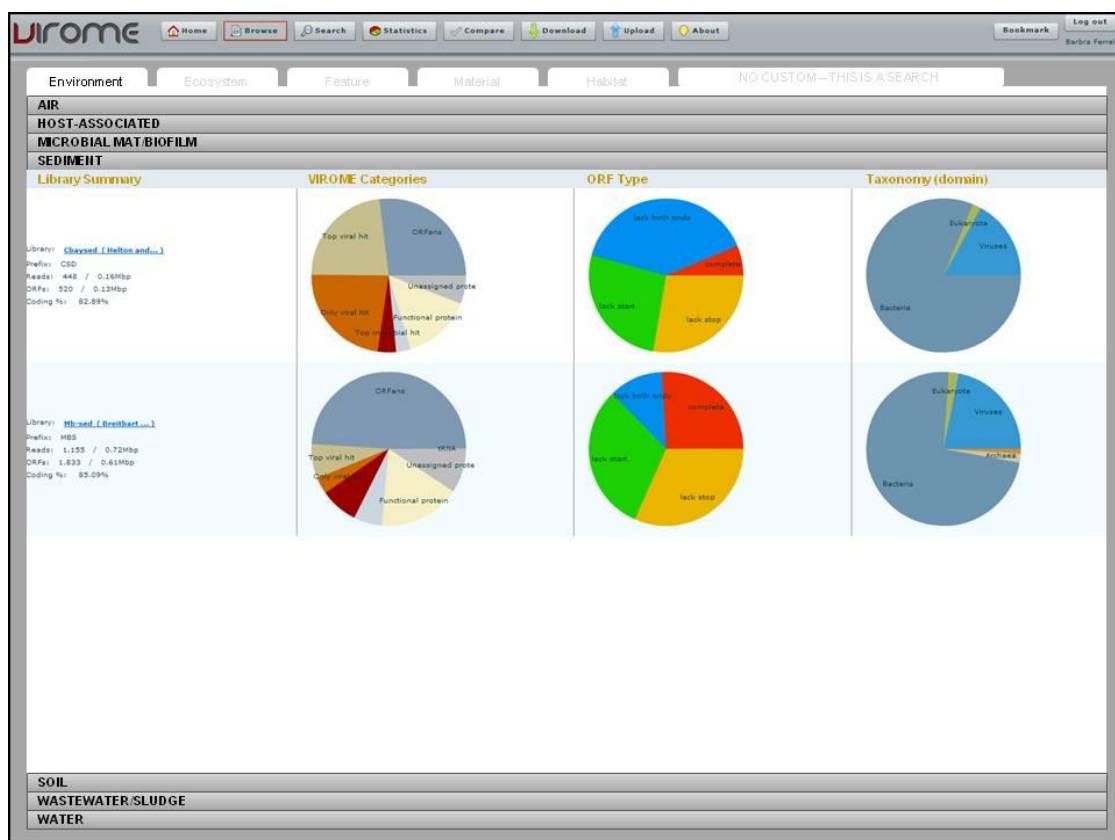


Figure 24 Proposed VIROME Browse view with flexible sorting available through stacked criteria.

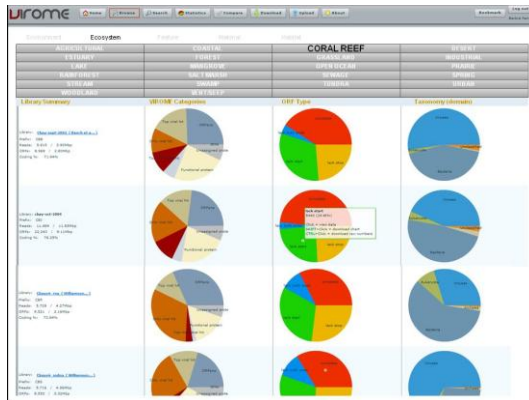


Figure 25a

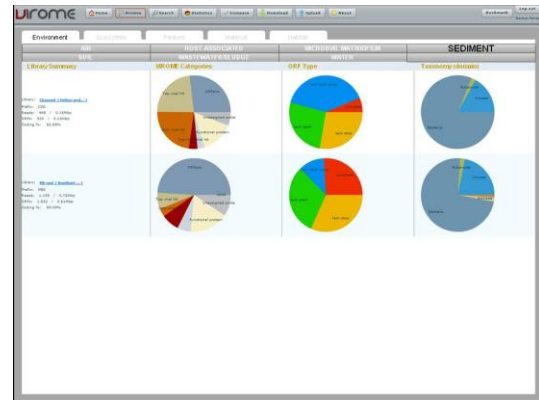


Figure 25b

Figure 25 Potential VIROME Browse view using tiled selections. Figure 25a shows manageable number of tiles when grouping by environment. Figure 25b demonstrates that grouping by other descriptors such as ecosystem make a tiled display cumbersome.

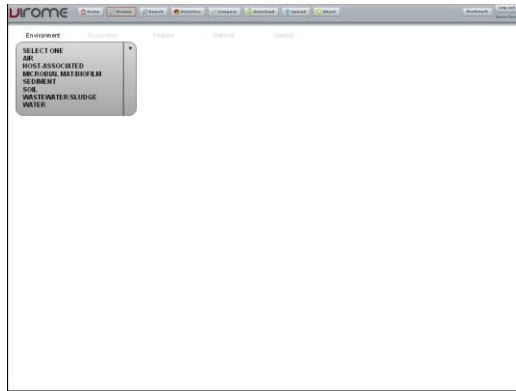


Figure 26a

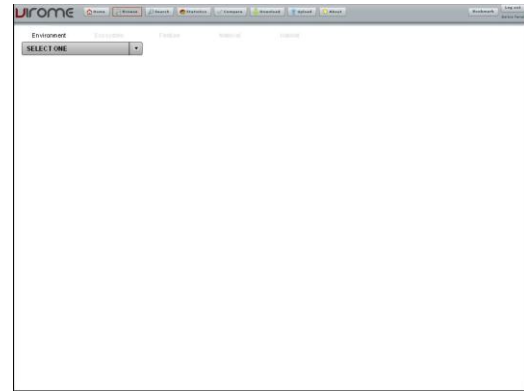


Figure 26b

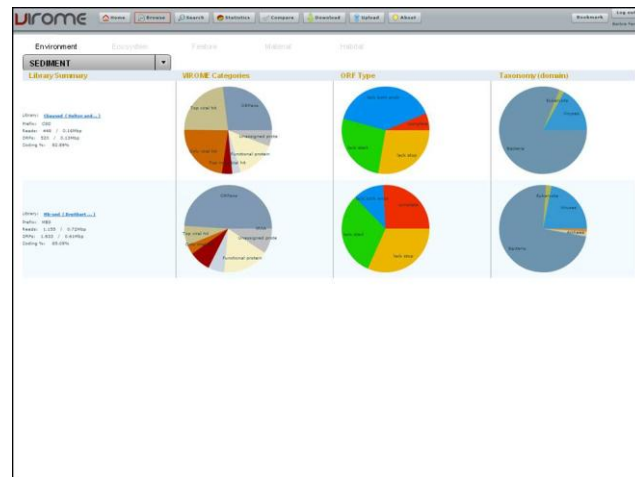


Figure 26c

Figure 26 Potential VIROME Browse view using dropdown menu selection. Figure 26a displays tabs and dropdown menus to allow for user selection. Figure 26b demonstrates user selection through drop down menu navigation. Figure 26c shows libraries grouped according to user selection.

Appendix G

VIROME SEARCH VIEW WORKFLOW

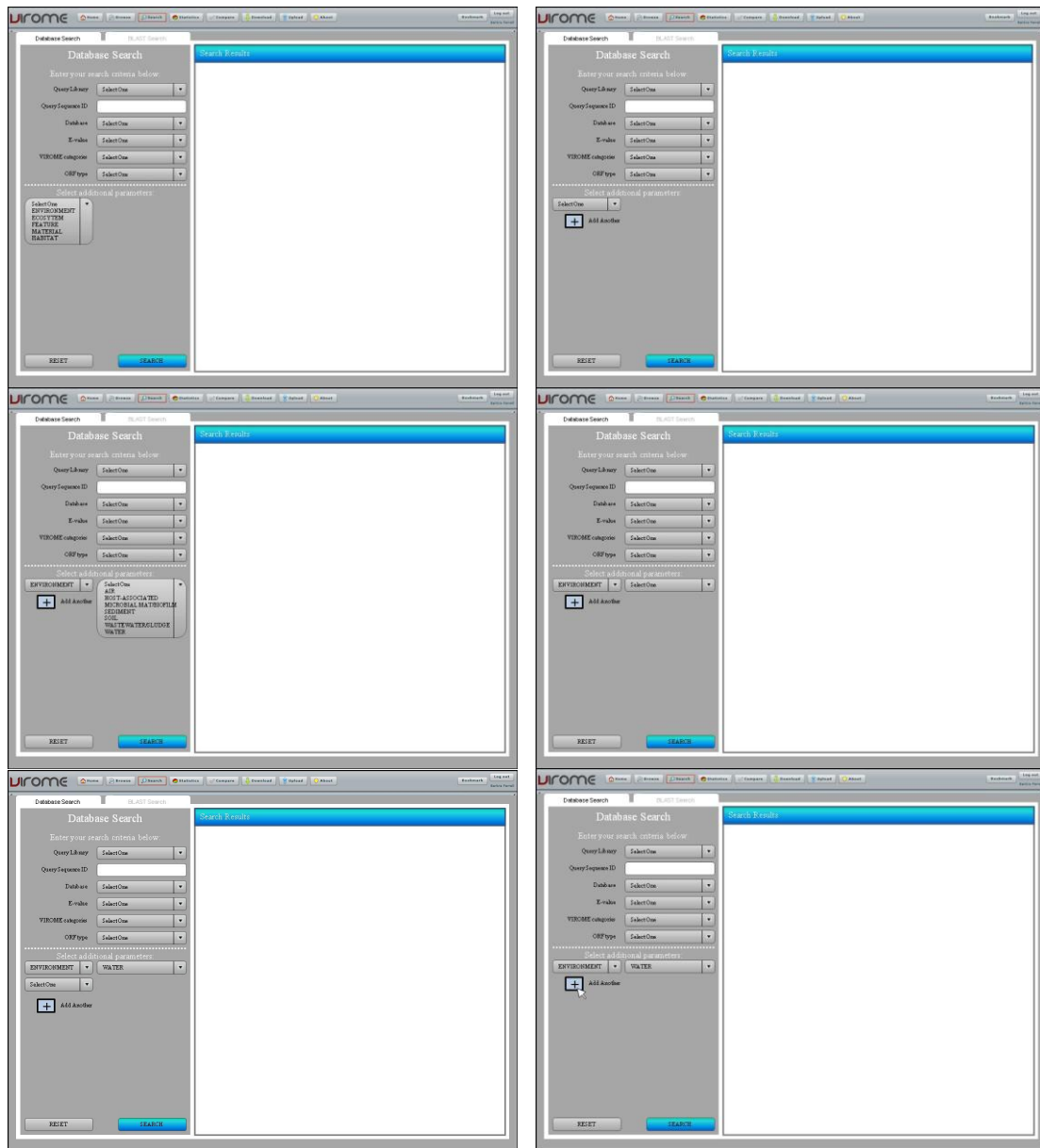


Figure 27 Proposed VIROME Search view panels show progression of user-specified flexible search criteria.

Appendix H

ADDITIONAL VIROME CART OR LAB BENCH FUNCTIONALITY

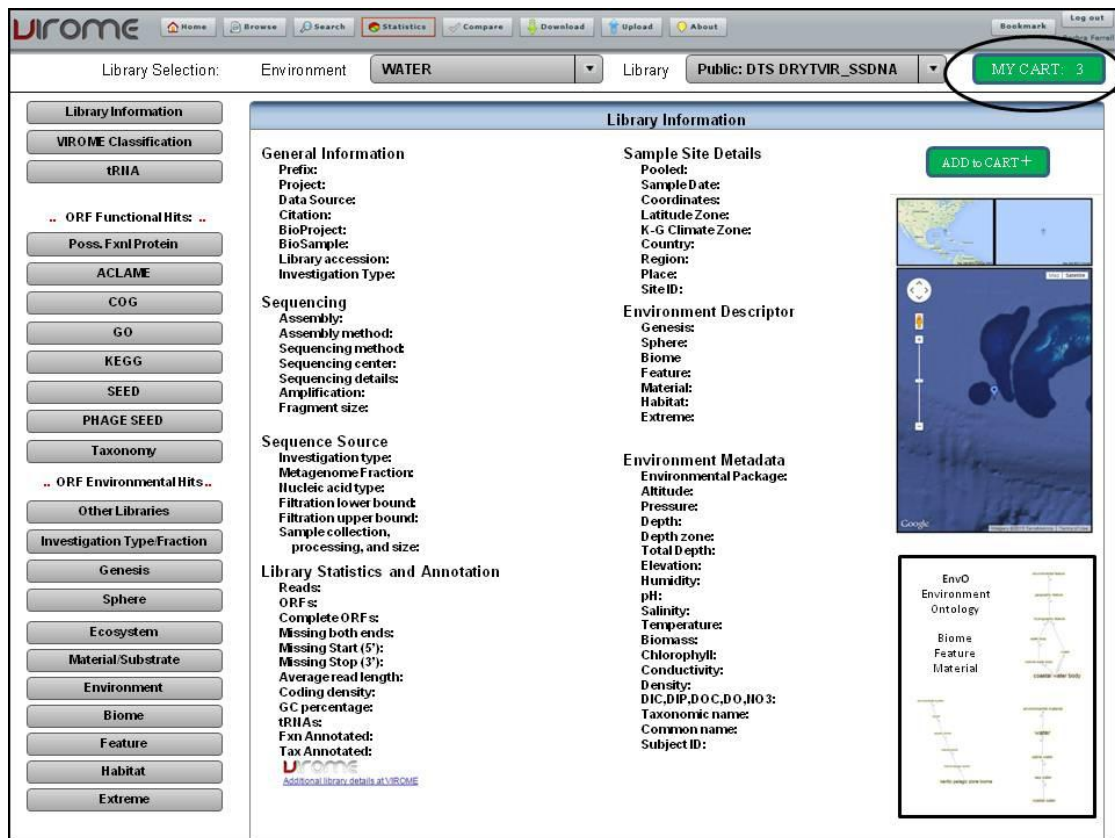


Figure 28 Proposed VIROME cart function from individual library page. Users have the ability to select particular libraries and use them in subsequent comparison exploration.

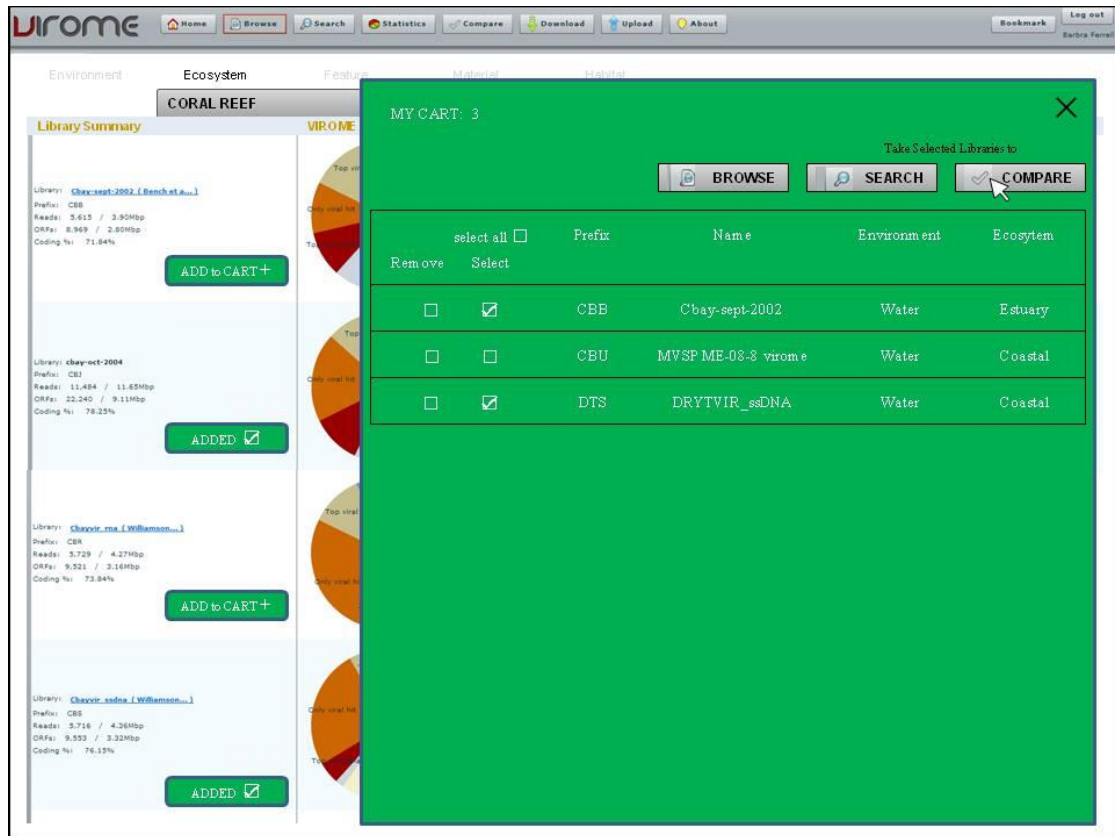


Figure 29 Proposed VIROME cart page with functionality to move libraries to browse, search, or compare views