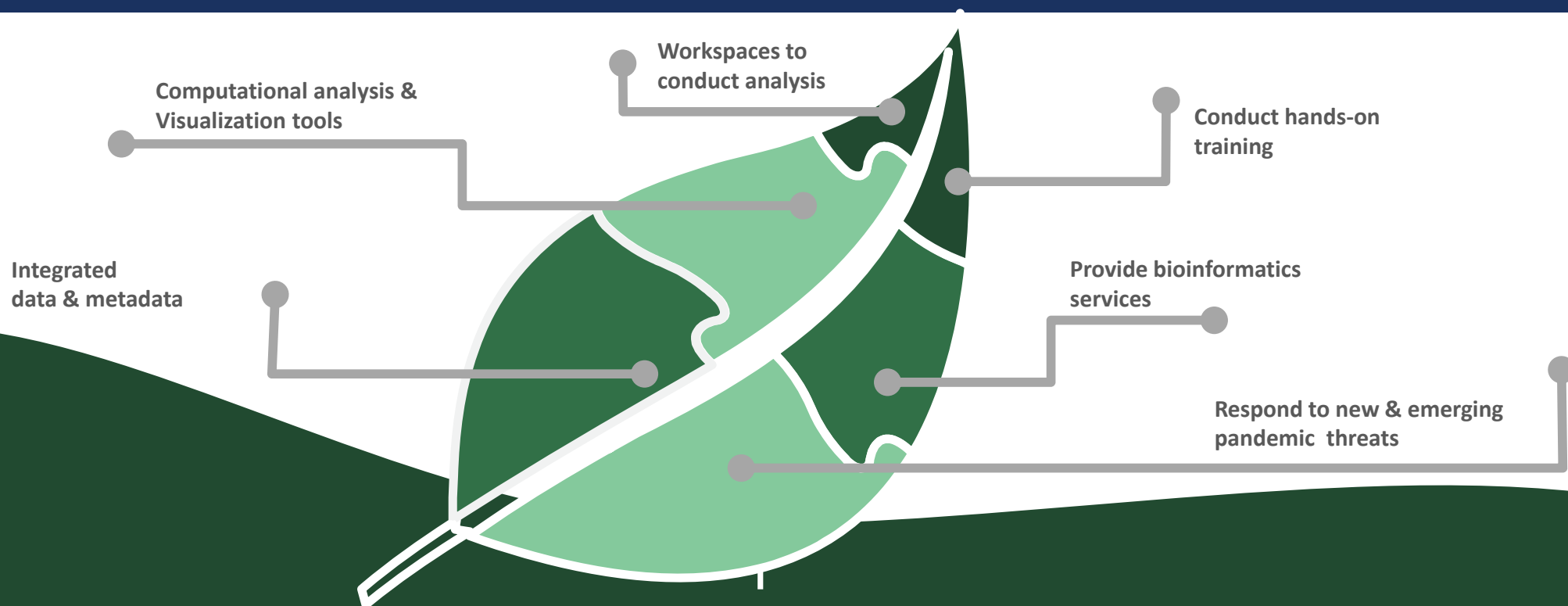

BIOINFORMATICS RESOURCE CENTERS

A knowledgebase of Integrated Data, Metadata, & Analysis
Tools to Support Data-enabled Infectious Disease Research

Ishwar Chandramouliswaran
Office of Data Science & Emerging Technologies
NIAID, NIH, DHHS



KEY CAPABILITIES



The BRCs provide integrated bioinformatics resources in support of basic & applied infectious disease research

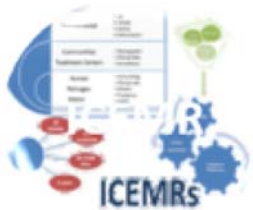
THE CENTERS ORGANIZED BY PATHOGEN GROUPS

Institute name	PI name	BRC	Pathogen
University of Notre Dame	Mary Ann McDowell	VectorBase	Invertebrate vectors of human pathogens
University of Pennsylvania	David Roos	EuPathDB/FungiDB	Eukaryotic/fungal pathogens
Northrop Grumman	Richard Scheuermann	ViPR/IRD	Viruses/Influenza
University of Chicago	Rick Stevens	PATRIC	Bacteria

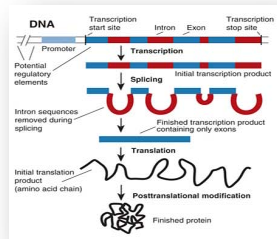


COLLABORATE TO INTEGRATE DATA ACROSS NIAID PROGRAMS

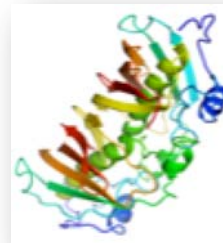
Sequencing



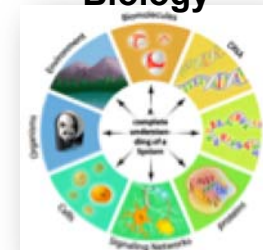
Functional Genomics



Structural



Systems Biology



Bioinformatics

Genomic Research Resources

Genomic/Omics Data Sets, Databases, Bioinformatics Tools, Biomarkers, 3D Structures, Protein Clones, Predictive Models

To address key questions in
microbiology and infectious
disease

To identify new targets and develop
new strategies for vaccines,
diagnostics and therapeutics

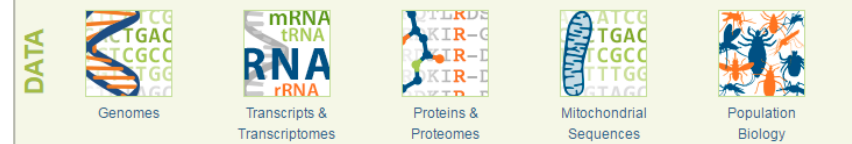


VECTORBASE BRC FOR INVERTEBRATE VECTORS OF HUMAN PATHOGENS

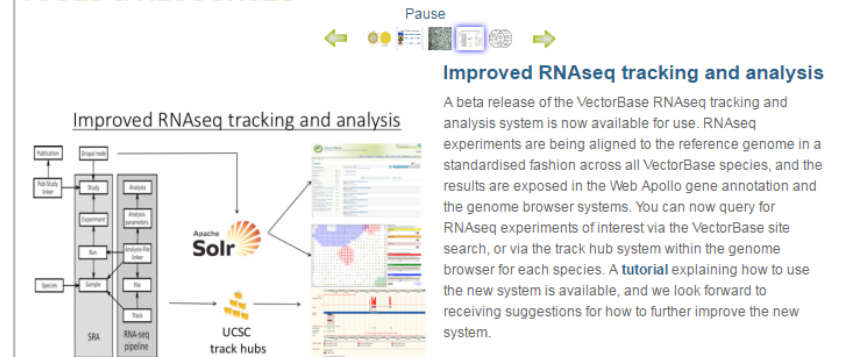
- **An unique and integrated bioinformatics resource for vector community**
 - Genome of **35 organisms** with annotated gene features, genome variation, transcript and protein expression
 - **Population** genetics data integrated with variation and **insecticide-resistance** phenotypes, field-associated samples from **surveillance** studies, and pathogen **transmission** data
 - Transcriptomes (RNA-seq) and gene expression data
 - Pathways
- **Data access, analysis and visualization tools**
 - Data-centric search system
 - **Computational tools** made accessible through a Galaxy instance implemented in VB
 - Expression map
 - Ontology browser

Welcome to VectorBase!

VectorBase is a NIAID Bioinformatics Resource Center providing genomic, phenotypic and population-centric data to the scientific community for invertebrate vectors of human pathogens.



TOOLS & RESOURCES



SUBMIT YOUR DATA TO VECTORBASE

VectorBase is not just for accessing data and using tools - we

Tweets by @VectorBase



EUPATHDB

BRC FOR INVERTEBRATE VECTORS OF HUMAN PATHOGENS

An unique resource for protozoan & fungal pathogens



- A comprehensive collection genomic and functional **genomic data, isolate data and phylogenomics**
 - genome sequences and annotation, strand-specific RNA-seq data, splice junction predictions, phosphoproteomic data, high-throughput phenotyping data, single nucleotide polymorphism, and expression quantitative trait loci data.
- Sophisticated **search strategies** using metadata
- Computational **analysis** on Galaxy

EuPathDB provides a unified entry point for the Eukaryotic Pathogen Bioinformatics Resource Center where you can ask questions that leverage data across organisms using orthology:
<http://EuPathDB.org>

OrthoMCL allows you to identify phylogenetically conserved or unique proteins among genomes from all branches of life:

<http://OrthoMCL.org>

Organism specific websites allow you search diverse data types from multiple species:

AmoebaDB Amoebozoa specific resource including *Entamoeba* - *Acanthamoeba* coming soon:

<http://AmoebaDB.org>

CryptoDB *Cryptosporidium* specific resource:
<http://CryptoDB.org>

GiardiaDB *Giardia* specific resource includes assemblages A, B and E:

<http://GiardiaDB.org>

MicrosporidiaDB Includes multiple microsporidia species:

<http://MicrosporidiaDB.org>

PiroplasmaDB *Babesia* and *Theileria* specific resource:

<http://PiroplasmaDB.org>

PlasmoDB Resource for *Plasmodium* species:
<http://PlasmoDB.org>

ToxoDB *Toxoplasma*, *Neospora* and *Eimeria* resource:

<http://ToxoDB.org>

TrichDB Resource for *Trichomonas vaginalis*:

<http://TrichDB.org>

TriTrypDB Resource for kinetoplastida including *Trypanosoma* spp. and *Leishmania* spp.:

<http://TriTrypDB.org>



VIRUS PATHOGEN RESOURCE (ViPR) INFLUENZA RESEARCH DATABASE (IRD)

- **A comprehensive collection of influenza and multiple virus family related data.**
 - sequences and annotations, immune epitopes, 3D protein structures, HTP omics data, clinical phenotype, surveillance, serology, host factor data, and curated antiviral drug data.
- **A suite of analytical and visualization tools**
 - phylogenetic tree, sequence variation determination, metadata-driven Comparative Analysis Tool for Sequences (meta-CATS), short peptide identification, PCR primer design, Sequence Feature and Phenotypic Variant Type (PVT) annotation, HA clade classification, HA subtype numbering conversion, surveillance data visualization, protein structure visualization, and host factor data enrichment analysis
- **Personal workbench spaces for data storage, sharing and analysis**
- **An infrastructure for antiviral drug data management and analysis.**
 - Drug interaction site curation
 - Integration of host factor data with drug target
 - Anti-viral drug resistance risk assessment tool

The image shows two screenshots of web interfaces. The top screenshot is the Influenza Research Database (IRD) homepage, featuring a world map with red and green markers indicating surveillance data. The bottom screenshot is the Virus Pathogen Resource (ViPR) homepage, which displays a search bar, a list of data types (Genomes, Genes & proteins, Immune epitopes, 3D protein structures, Host Factor Data, Antiviral Drugs), and a section for analyzing data online (Sequence Alignment, Phylogenetic Tree, Sequence Variation (SNP), Metadata-driven Comparative Analysis, BLAST). It also includes a 'Save to Workbench' section with options to store and share data, integrate data with ViPR, and create a custom search alert.

IRD Influenza Research Database

Surveillance
Human, avian and non-human mammalian surveillance data can be searched based upon location and various host characteristics for download and display on a map with bird flyway overlay. Create a report showing counts of surveillance records grouped by user-driven criteria.

Key Highlights:

- Human, avian and non-human mammalian surveillance data
- User-driven serotype comparison
- Display data on Google Map with flu prevalence color coded and bird flyways overlaid
- View results by metadata matrix report

ViPR Virus Pathogen Resource

Search
Search our comprehensive database for:
• Genomes
• Genes & proteins
• Immune epitopes
• 3D protein structures
• Host Factor Data
• Antiviral Drugs
Browse All Search Types

Analyze
Analyze data online:
• Sequence Alignment
• Phylogenetic Tree
• Sequence Variation (SNP)
• Metadata-driven Comparative Analysis
• BLAST
Browse All Tools

Save to Workbench
Sign up for a workbench to:
• Store and share data
• Combine working sets
• Integrate your data with ViPR data
• Store and share analyses
• Custom search alert
Sign In

PATHOGEN RESOURCE INTEGRATION CENTER (PATRIC) BRC FOR BACTERIAL SPECIES



Integrated Data Collections

Genomes, Transcriptomes (uArray and RNAseq), Proteomics, Metabolomics, Curated gene collections, Protein-Protein Interactions, Tn-Seq

AMR panel data, Disease & isolation phenotype metadata,, Virulence, Drug Targets, Biochemical Pathways

DATA WORKSPACES SERVICES HELP

Data Types
Antibiotic
Resistance
Genomes
Genomic
Features
Pathways
Protein Families
Specialty Genes
Transcriptomics

Download Data
FTP Server

Specialty Data Collections
PATRIC Collaborations
PATRIC DBPs
NIAID Clinical
Proteomics
NIAID Genome
Sequencing
NIAID Structural
Genomics
NIAID Systems Biology
NIAID Functional
Genomics

Personal Workspace for analysis of User data and “Virtual Integration” of user data in the context of all the public datasets

PATRIC / home / Special Collections / NIAID Systems Biology Centers / Omics4TB

Name	Size	Owner	Members	Created
Parent folder				
MTB TF fitness		PATRIC	Public	5/6/17, 12:02 PM
MTB TF overexpression		PATRIC	Public	5/6/17, 12:12 PM
MTB TRIP relative fitness profiling		PATRIC	Public	3/25/18, 11:29 AM
MTB hedges		PATRIC	Public	5/6/17, 12:09 PM
		PATRIC	Public	5/6/17, 12:11 PM
		PATRIC	Public	3/25/18, 9:29 AM

Antibiotic Resistance /
Susceptibility Phenotype
Predictions

Computational Analysis Services

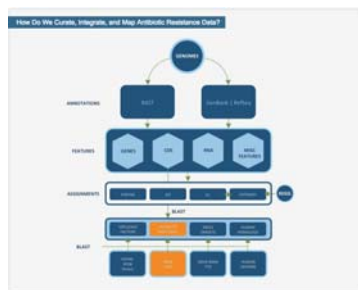
SERVICES HELP All Data Types Find a gene genome

Genomes
Assembly
Annotation
BLAST
Similar Genome Finder
Variation Analysis
Tn-Seq Analysis
Phylogenetic Tree
Metagenome Binning

Protein Tools
Protein Family Sorter
Protein Comparison

Metabonomics
Comparative Pathway
Model Reconstruction

Data
ID Mapper



What do we mean by...

Antibiotics: Antibiotics are a type of antimicrobial drugs used in the treatment and prevention of bacterial infections. PATRIC provides basic information about commonly used antibiotics, including their chemical and physical properties, pharmacology, and mechanism of action. In addition, each antibiotic is linked to other relevant data available in PATRIC, such as AMR phenotypes for genome, AMR genes, and AMR regions. Below are some examples.

- [Amikacin](#)
- [Gentamicin](#)
- [Streptomycin](#)

AMR Genes: AMR genes refer to the genes implicated in or associated with the resistance to one or more antibiotics. The resistance may result from the presence or absence of a gene or specific mutations acquired spontaneously or through evolution over time. We integrate and map known antibiotic resistance genes from the following sources:

- [AMR](#)
- [SAR](#)
- [NDAR](#)
- [PATRIC AMR](#)

[View all AMR genes](#)

AMR Phenotypes: AMR phenotypes refer to the resistance or susceptibility of a given organism to one or more antibiotics. PATRIC collects AMR phenotype data generated using antimicrobial susceptibility testing methods (AST) from published studies and collaborators. In addition, we also provide predicted AMR phenotypes using machine learning classifiers. See AMR phenotype data for selected genes:

- [Mycobacterium](#)
- [Staphylococcus](#)
- [Streptococcus](#)

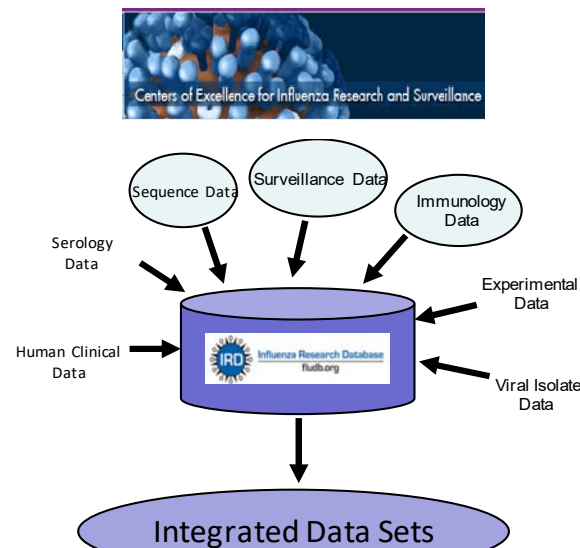
AMR Regions: AMR regions refer to the small genomic regions implicated in or associated with the resistance to one or more antibiotics. The AMR regions are computationally predicted using machine learning classifiers used to predict AMR phenotypes. They may map to existing genes or intergenic regions and may help identify new AMR genes or understand AMR mechanisms.

[View all AMR regions](#)

Summary of Omics4TB Datasets

Experiment	PMID	Data Source
The DNA-binding network of Mycobacterium tuberculosis	25081035	• BioProject PRJNA255884 Raw sequence data for ChIP-seq experiments (BAM format) • MTH Network Portal ChIP-seq, Sorted and Indexed ChIP-seq files
Mapping and manipulating the Mycobacterium tuberculosis transcriptome using a transcription factor overexpression-derived regulatory network	25389855	• GEO Accession: GSE59008 Transcription factor overexpression data from RNA-seq experiments • PATRIC: Raw reads as fastq files • MTH Network Portal TPOE Sample information for each experiment
A high-resolution network model for global gene regulation in Mycobacterium tuberculosis	25230956	• MTH Network Portal TPOE Data sets and additional information
A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis	25077515	• PATRIC: ChIP-seq ChIP-seq experiment information page • PATRIC: TPOE Expression TPOE Experiment information page • GEO Accession: GSE59008 Transcription factor overexpression data from RNA-seq experiments • MTH Network Portal TPOE Sample information for each experiment • PATRIC: TPOE Expression TPOE Experiment information page
Network analysis identifies Rv0226 and Rv0228 as regulators of isoniazid tolerance in Mycobacterium tuberculosis	25173154	• GEO Accession: GSE57349 Expression profiling data
AMR-18-regulated molecular network orchestrates cell fate in the innate and adaptive immune response to Mycobacterium tuberculosis	26048045	• PATRIC: Proteomics Proteomics
Abscissa Proteome Composition and Dynamics during dormancy and resuscitation of Mycobacterium tuberculosis		• PATRIC: Proteomics Proteomics
MTH Transcriptional Response Induction Phenotype (TRIP) Screen		• PATRIC: TRIP Response

RESPOND TO NEEDS OF ID RESEARCHERS



- Clinical & Epidemiological Data
- PopBio Data & Tools

Zika Virus

SEARCH DATA ANALYZE & VISUALIZE WORKBENCH VIRUS FAMILIES HELP

Zika Virus

Taxonomy: Group IV ((+ssRNA); Flaviviridae; Flavivirus; Zika virus most similar to St. Louis encephalitis virus

Virion: enveloped, spherical, ~50 nm in diameter with an electron dense core of ~30 nm

Genome: ~10.8 kb positive-sense, single-stranded RNA

Proteome: single polyprotein, co- & post-translationally cleaved into 11 mature proteins

Infection: initiated by binding of the viral E protein to a host receptor

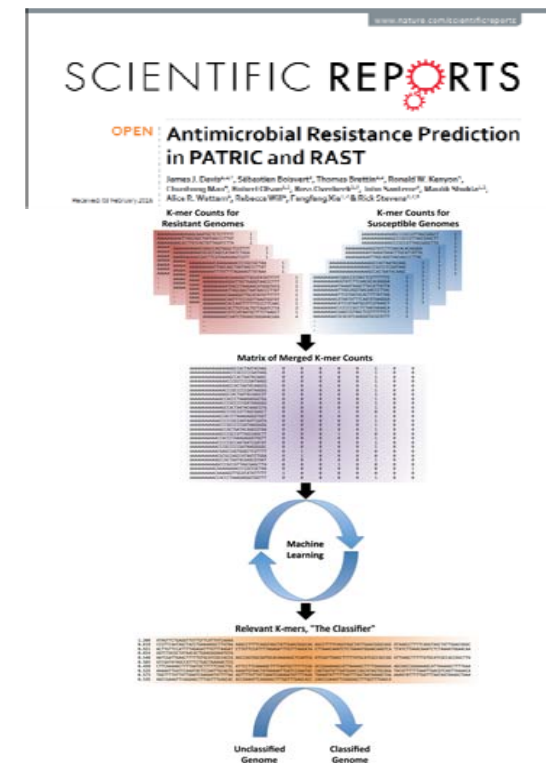
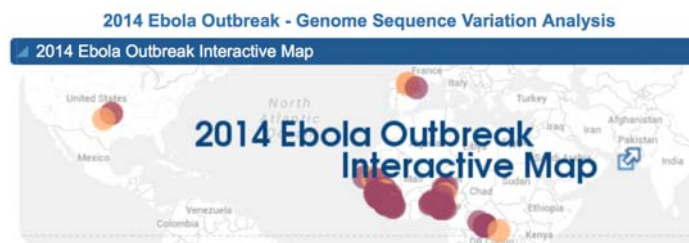
RNA Transcript: 5' methyl cap, no 3' poly-A tail

Transmission: primarily through the bite of an infected Aedes mosquito; rarely from mother to child; infrequently through blood transfusion and sexual contact

Epidemiology: sporadic outbreaks in Africa, Southeast Asia, and the Pacific Islands before 2015; major outbreaks have been occurring in the Americas since May 2015. Currently, 30 countries have active transmission of the Zika virus while numerous others have reported travel-associated cases.

Clinical: the most common symptoms of Zika virus infection are fever, rash, joint pain, and conjunctivitis; a link to neurological syndromes, including Guillain-Barre syndrome, and birth defects is strongly suspected.

References: David M. Kriple, Peter M. Howley, et al. Fields' Virology, 5th ed. 2007 Lippincott Williams & Wilkins, USA; <http://www.cdc.gov/zika>

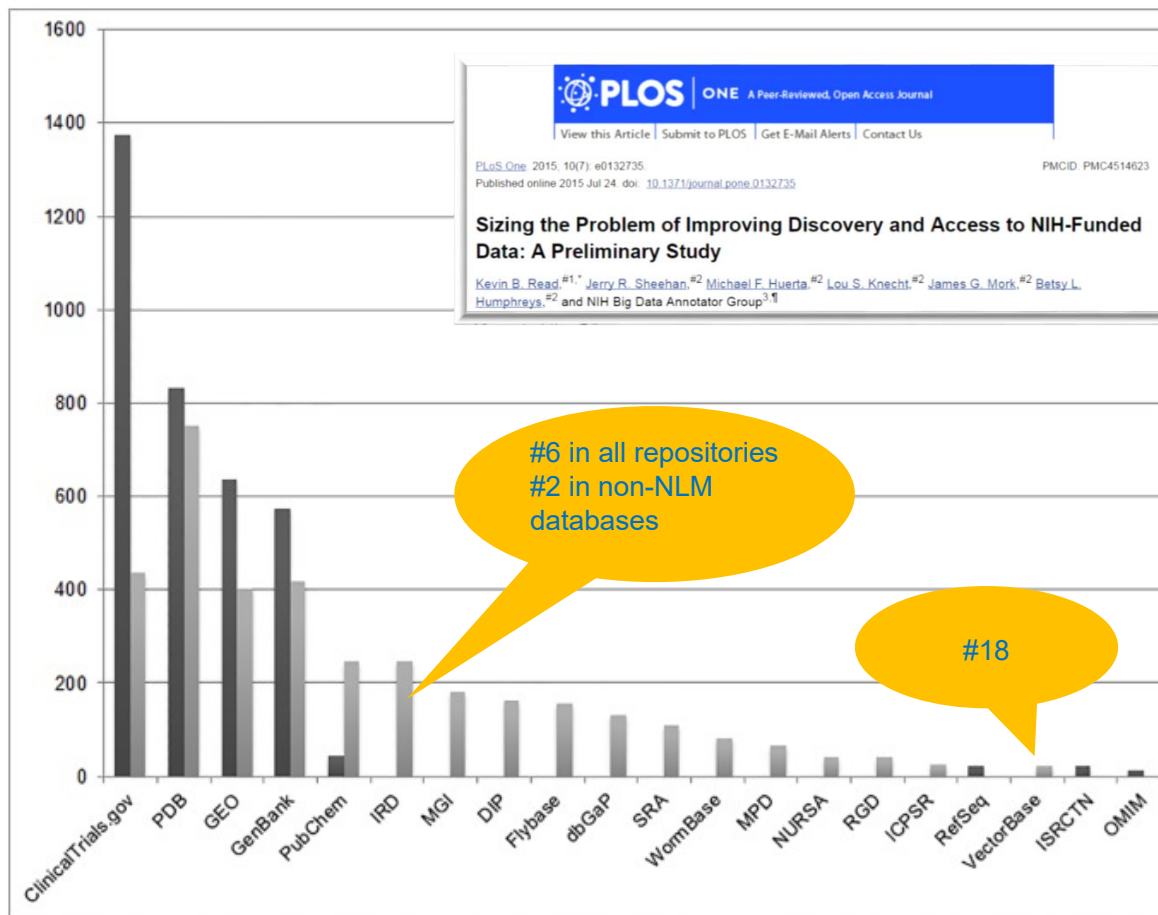


- AMR data collection
- AMR prediction tools

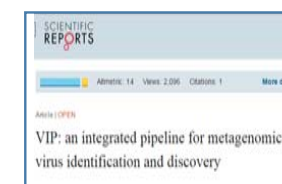
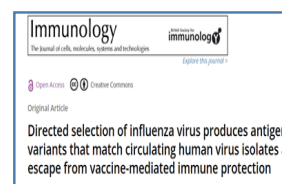
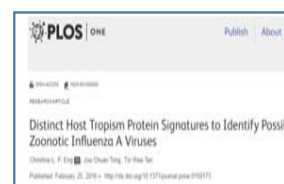
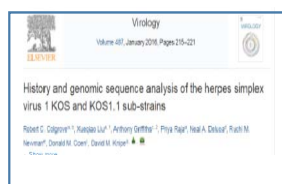
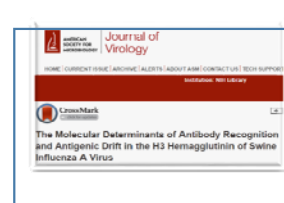
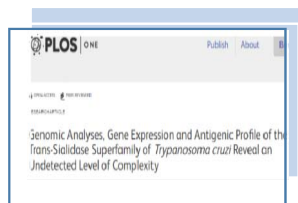
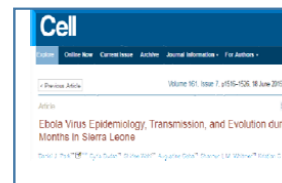
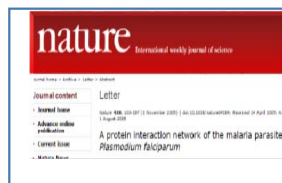
HANDS-ON TRAINING WORKSHOPS GLOBALLY



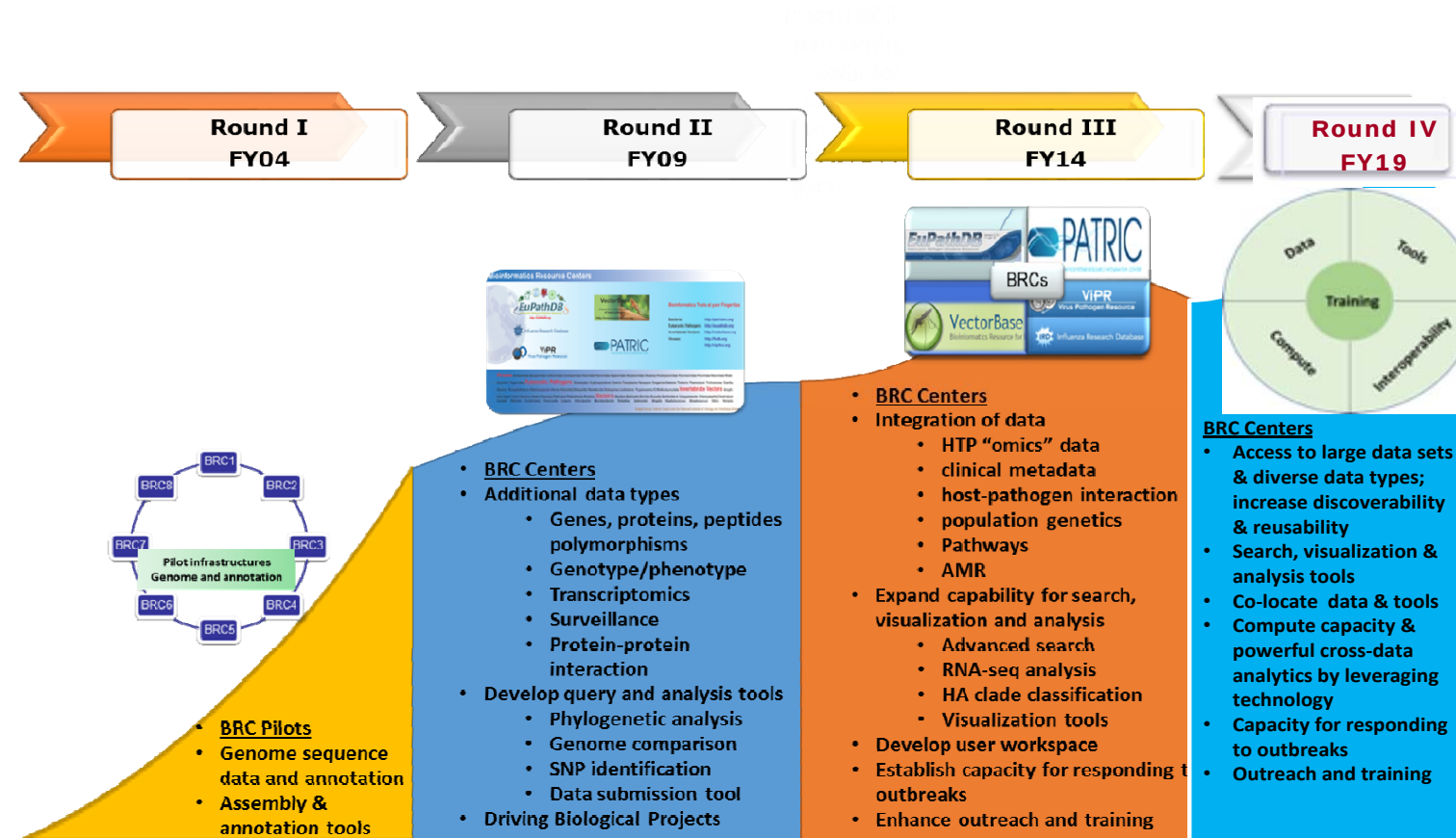
TO INCREASE USAGE & IMPACT



SCIENCE ENABLED BY THE BIOINFORMATICS RESOURCE CENTERS



EVOLUTION OF THE RESOURCES



THE NIAID DATA CHALLENGE

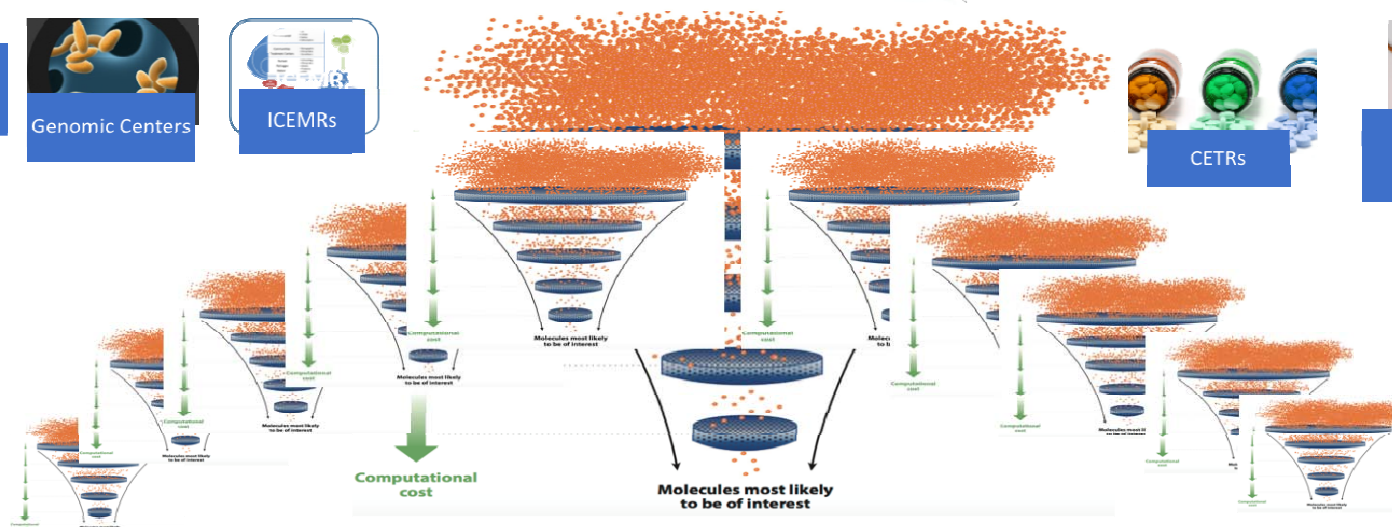
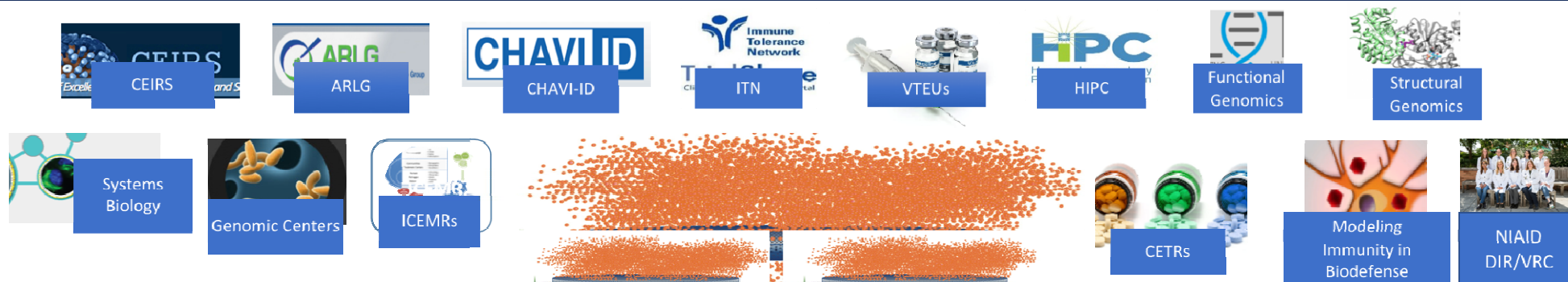


Image Courtesy: Dr. Emily Erbelding (modified)

NEW MODELS OF DATA-DRIVEN HYPOTHESIS GENERATION

Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing

Rachel Sparks,¹ William W. Lau,² and John S. Tsang^{1,*}

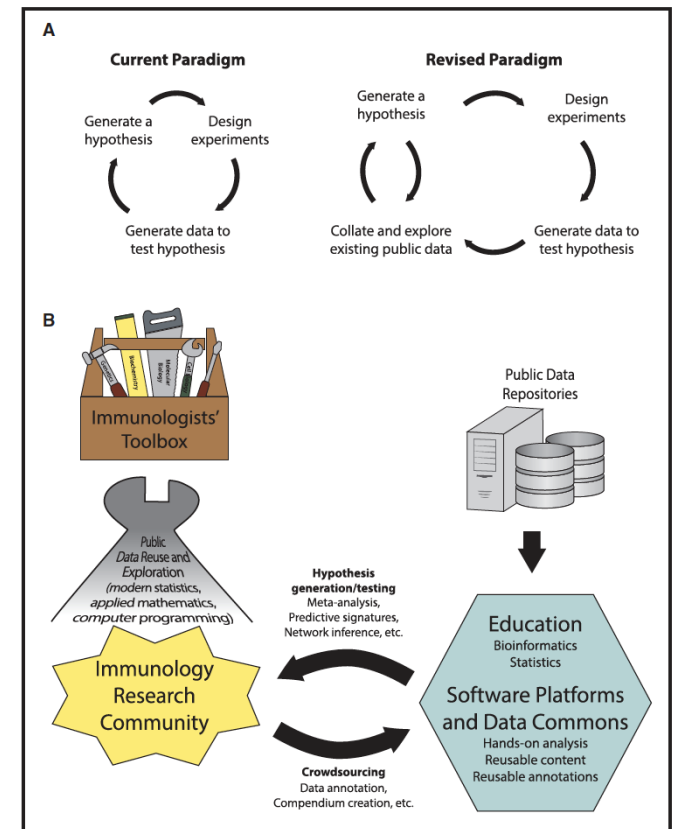
¹Systems Genomics and Bioinformatics Unit, Laboratory of Systems Biology, National Institutes of Allergy and Infectious Diseases, National Institutes of Health

²Office of Intramural Research, Center for Information Technology, National Institutes of Health

*Correspondence: john.tsang@nih.gov

<http://dx.doi.org/10.1016/j.immuni.2016.12.008>

New technologies have been propelling dramatic increases in the volume and diversity of large-scale public data, which can potentially be reused to answer questions beyond those originally envisioned. However, this often requires computational and statistical skills beyond the reach of most bench scientists. The development of educational and accessible computational tools is thus critical, as are crowdsourcing efforts that utilize the community's expertise to curate public data for hypothesis generation and testing. Here we review the history of public-data reuse and argue for greater incorporation of computational and statistical sciences into the biomedical education curriculum and the development of biologist-friendly crowdsourcing tools. Finally, we provide a resource list for the reuse of public data and highlight an illustrative crowdsourcing exercise to explore public gene-expression data of human autoimmune diseases and corresponding mouse models. Through education, tool development, and community engagement, immunologists will be poised to transform public data into biological insights.



FAIR GUIDELINES

Findable
Accessible
Interoperable
Reusable

Image Courtesy: Dr. Patti Brennan



[Comment](#) | [OPEN](#) | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons [✉](#)

Scientific Data **3**, Article number: 160018 (2016) | [Download Citation](#) [↓](#)

REVISIT DATA MANAGEMENT & SHARING PLANS

Resource Sharing Plan

In compliance with current NIH requirements, we herewith outline a plan to share materials and our management of intellectual property. We will adhere to the NIH Grant Policy on Sharing of Unique Research Resources including the Sharing of Biomedical Research Resources Principles and Guidelines for Recipients http://ottd.od.nih.gov/policy/rt_guide_final.html. We will make genetic and phenotypic data upon request. Material more restrictive terms than in the Simple Letter A requirements. Material such as microbial strains Repository or other public repositories. The BEI arrangement would be followed in such cases. It requires a patent, we will ensure that the technical available to the research community in accordance Principles and Guidelines document.

The following Table outlines the data types and to along with timelines for deposition.

Data and/or Reagent Type	Public Repository for Raw and Finished Data
RNA sequencing (transcriptomics)	GEO / NCBI
DNA Sequencing (clinical isolates)	NCBI SRA
Genome Assemblies and associated metadata	NCBI Assembly and/or NIAID BRIC or other approved repository
Model Organisms / Strains	Stock center / BEI Resources Repository / or other approved repository
Other data types (e.g., metabolomics)	NIAID BRICs or other repositories (e.g., PATRIC)

Data and Resource Sharing Plan

UNIVERSITY OF MAASTRICHT, UNIVERSITY OF MAASTRICHT

In the original proposal the following table was included to sum from the proposed work. This table is included here for reference data types.

Item	Description	Data Type
14.1	Whole cell and/or CD acquisition	CM
14.2	Whole cell acquisition	CM, 16S, MET
14.3	Whole cell of severe disease	CM
14.4	Whole cell of severe disease	CM, 16S, MET
14.5	CD acquisition	CM, 16S
14.6	Genetic association with CD	CM, 16S, 16S
14.7	Microbiome/CD acquisition	16S
14.8	Baseline colonization/CD	CM
14.9	CD colonization acquisition	CM
14.10	Microbiome negative colonization	CM, 16S, 16S
14.11	Microbiome colonization acquisition	CM, 16S, 16S

Table 2: Patient privacy and the relationship to the proposed research. "All identifiers other than the collection of de-identified data and can identify cases of CD based on positive laboratory results. "All CD cases data available in the hospital data warehouse. Data types: CM—clinical metadata; 16S—small ribosomal subunit gene sequences; Data analysis types: 16S—sequenced library; 16S—unsequenced library; 16S—sequenced library; 16S—unsequenced library.

To summarize the data types and resources that will be generated:

- 1) Metabolic and sequence data
 - a. 16S-sequenced gene amplicon library sequence
 - b. C. difficile genome sequence data (16S)
 - c. Metagenomic sequence datasets
 - d. Metagenomic sequence datasets
- 2) Metagenomic datasets
 - a. Targeted analysis
 - b. Untargeted analysis
- 3) Clinical metadata
 - a. Patient history
 - b. Fecal swabs
 - c. Difficult isolates
- 4) Fecal specimens
 - a. Fecal swabs
 - b. Fecal swabs

The remainder of this document will detail our plans for making these data and resources available to the general scientific community as per the NIAID/Division of Microbiology and Infectious Diseases (DMID) Data Sharing and Release Guidelines and the [Data Science Data Sharing Policy](#).

Data accessibility/sharing

The omics data in this proposal can be split into 4 categories. These data are all collected with Illumina while protecting the intellectual property essential for commercial development. Research data will be shared according to the most recent NIH policies on data sharing and open access to research data, subject to any exclusions mandated by subsequently enacted laws, regulations, or executive orders. Research data will comply with guidelines for Data sharing, intellectual property rights, and

Resource Sharing Plan

A) Data Sharing Plan: BCM, University of Houston, Texas A&M and UCLA are committed to public dissemination of its scientific results and developing health care products that improve global public health while protecting the intellectual property essential for commercial development. Research data will be shared according to the most recent NIH policies on data sharing and open access to research data, subject to any exclusions mandated by subsequently enacted laws, regulations, or executive orders. Research data will comply with guidelines for Data sharing, intellectual property rights, and

publicly available as soon as possible, if it is not already publicly available in the literature and in peer-reviewed scientific journals. Publication terms shall be in a Simple Letter Intellectual property also which requires a

to research community in accordance with

in the following table below

of Metagenomics, Metadescription,

real has been performed with utilization of

of clinical confidential data. Our clinical

IA-certified clinical diagnostic and research

use of the investigators on this project

arch on multiple entities. Data and specimen

research staff are experienced in following

at Baylor College of Medicine in the staff

a CD case, we are fully confident as our

for clinical and operational compliance with

data will be collected on the appropriate

devices and will be transferred to the local

We will develop a dedicated, customized

database systems (e.g., Flatfile, CDL,

data and that is most compatible with any

of all resources, they will have a dedicated

name of our local repository. This will

we intend of BCIR and keep in secure data

components that support information and

ing. The CDL and TCMC leverage this

ing in the overall computational capacity of

most efficient to duplicate copies of the

cludes all metagenomic sample handling,

performed in a secure, data of all met

mental techniques. The CDL is doing

in the HGSO. To enable further growth,

only managed and monitored data center

ies, and about 1.25 petabyte (PB) of data

in this proposal are committed to

on. Specifically, we propose to

research community by depositing

following the NIAID Data Sharing

generated from local samples

to, we will exercise the strictest

is to the research community. As

are and distribute the research

lers and oral presentations).

all times, including dissemination

identified data formats.

els (e.g., Data Use Agreements)

others interested in the project(s).

locally available via BEI Resources

sity (non-disclosure) Agreements,

gits while exchanging information

community.

any data in the case of vulnerable

rare diseases) where release of

legal consequences. Associated

guidance from Program staff

Examples/reagents available on

Read Archive (SRA) of NCBI in

one per sample) with low-sequenced

pooling—samples—data—sequencing

as quickly as possible, and no later

the NIAID/Division of Microbiology

and Infectious Diseases (DMID) Data

Sharing and Release Guidelines

Whole metagenomic sequencing data (100 samples/1-year) will be submitted in a

less—interval—of—every—6—months—pending—approximately—60—samples—into—one

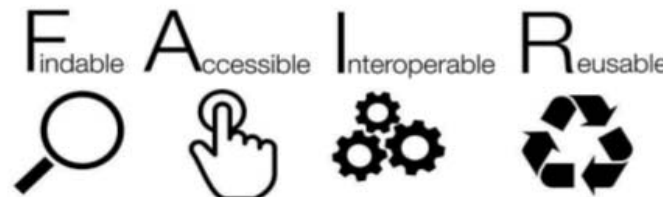
submission—batch. Whole metagenomic sequencing data will be submitted as

quickly as possible, and no later than 45 calendar days of being generated, as per

the NIAID/Division of Microbiology and Infectious Diseases (DMID) Data Sharing and

Release Guidelines

FAIR principles and metrics for evaluation



Michel Dumontier, Ph.D.

Distinguished Professor of Data Science



Maastricht University

Bioprocessing device sharing (e.g., new isolates or reagents) Years 3-6

BEI Resources (NIH) or from PI per DTA

Validated products / reagents are publicly made available per fair policies

* One or more deposition sites will be identified as appropriate based on the NIAID / DMID Data Sharing & Release Guidelines (2016)

* Distribution of data as early as is appropriate based on publication, embargo, and/or IP filing per institutional policies & practices

B. Resource or Data Use Policies & Practices

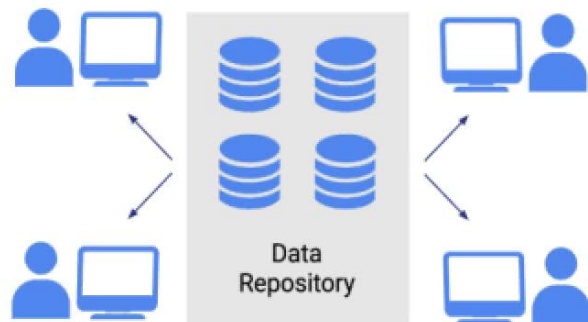
Per CDL general policies, users of [agg](#) released data shall act responsibly to recognize the scientific contribution of the data generators/producers by following normal standards of scientific etiquette and fair use of unpublished data. Such guidelines can be found in Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility (PDF) / Toronto Data Release Workshop (2006) Nature 441:166-170.

@micheldumontier::#DANSLOD:2017-05-01

- Whole metagenomic sequencing data (100 samples/1-year) will be submitted in a less—interval—of—every—6—months—pending—approximately—60—samples—into—one submission—batch. Whole metagenomic sequencing data will be submitted as quickly as possible, and no later than 45 calendar days of being generated, as per the NIAID/Division of Microbiology and Infectious Diseases (DMID) Data Sharing and Release Guidelines

CO-LOCATE DATA & TOOLS ON CLOUD

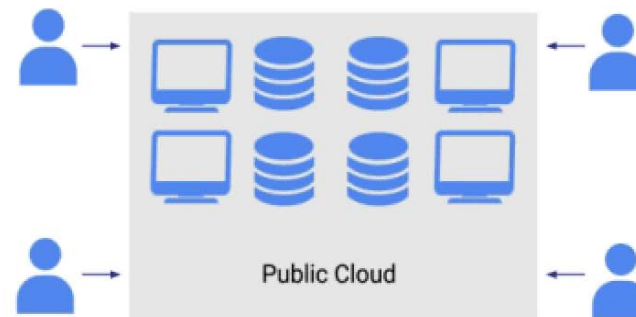
Traditional Way: Bring data to the researchers



Problems

Data sharing = data copying
Limits Access
Largely fixed compute
Individual security implementations

Cloud Way: Bring researchers to the data



Solutions

True data sharing
Democratizes access
Elastic compute and storage
Centralized security implementation

MINT DIGITAL OBJECT IDENTIFIERS



DATA INDEXING

Google Dataset Search Beta

Search for Datasets



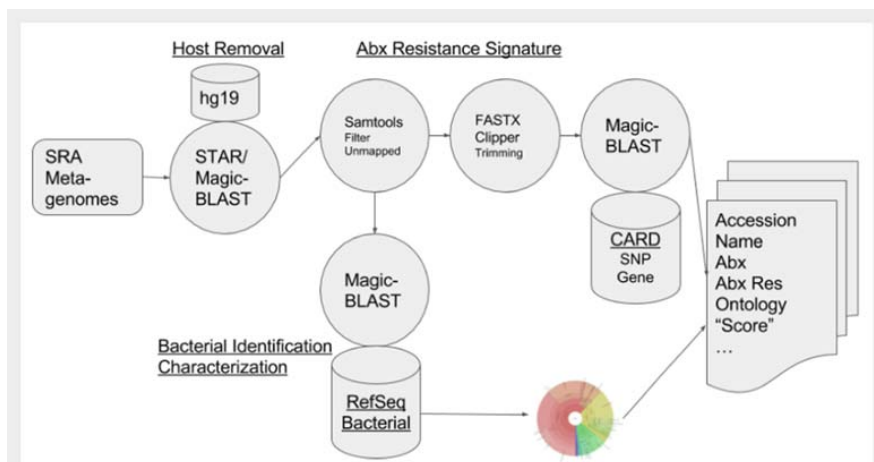
Try [boston education data](#) or [weather site:noaa.gov](#)

CONTAINERIZATIONS OF TOOLS

NIH Data Science Learning and Testing Hackathon

February 23, 2018!

The NIH will host a Learning and Testing Data Science hackathon on February 23rd, 2018 on the main campus in Bethesda, MD. Learners will test alpha and beta code that have been generated in full, collaborative development hackathons for a wide range of scientific problems, including general bioinformatics and genomic analyses in addition to text, image, and sequence processing. This event is for researchers who are in the early stages of their data science journey, including students and postdocs. Other non-scientific developers, mathematicians, or librarians in a similar educational place are also welcome! Learning in this event will be primarily hands-on and self-driven, but will also include short workshops on topics such as Docker and GitHub. Mentors will also be available to assist with questions.



F1000Research
Open for Science

Search

BROWSE GATEWAYS HOW TO PUBLISH ▾ ABOUT ▾ BLOG

Check for updates

SOFTWARE TOOL ARTICLE

NastyBugs: A simple method for extracting antimicrobial resistance information from metagenomes [version 1; referees: 2 approved with reservations]

✉ Hsinyi Tsang¹, Matthew Moss², Greg Fedewa³, Sharif Farag⁴, Daniel Quang⁵, Alexey V. Rakov⁶ 
⁶ ✉ Ben Busby ⁷

 Author details

 Grant information

 This article is included in the **Hackathons** collection.

Abstract

Multidrug resistant bacteria are becoming a major threat to global public health. While there are many possible causes for this, there have so far been few adequate solutions to this problem. One of the major causes is a lack of clinical tools for efficient selection of an antibiotic in a reliable way. NastyBugs is a new program that can identify what type of antimicrobial resistance is most likely present in a metagenomic sample, which will allow for both smarter drug selection by clinicians and faster research in an academic environment.

METRICS

462
VIEWS

185
DOWNLOADS

 Get PDF

 Get XML

 Cite

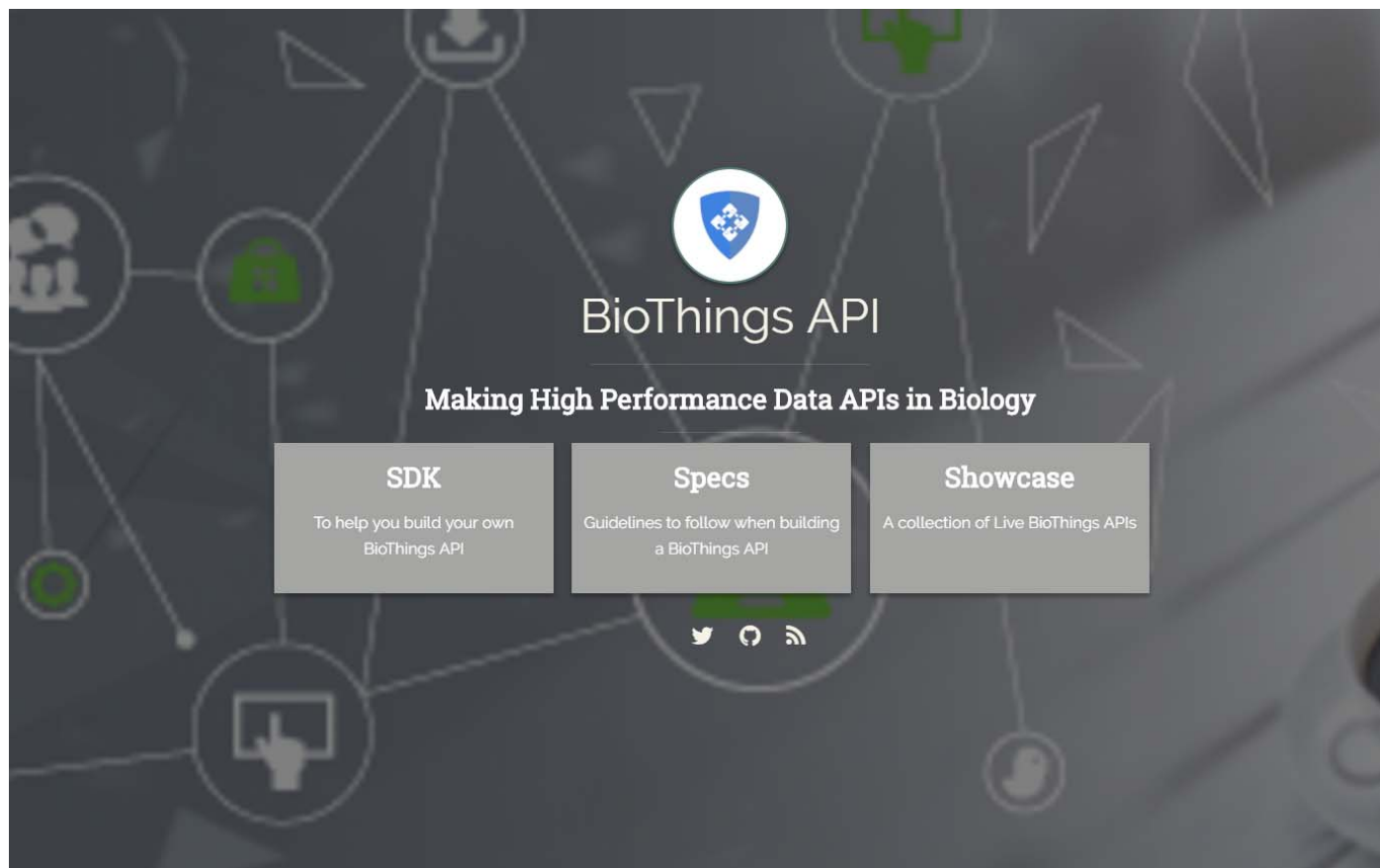
 Export

 Track

 Email

 Share

USE OF APPLICATION PROGRAMMING INTERFACES



Not an endorsement but an example

USE OF SMART-IRB, AUTHN, AUTHZ SOLUTIONS



Supporting single IRB review
Advancing collaborative research



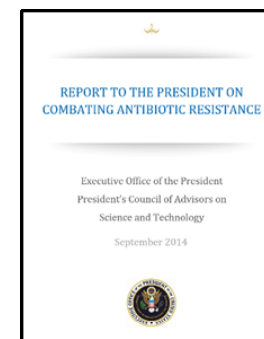
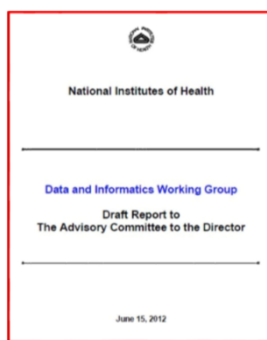
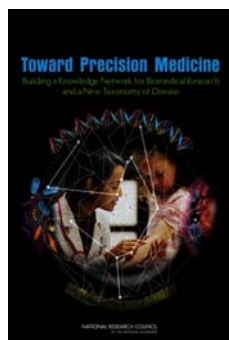
DUOS



Not an endorsement but an example

REALIZE DATA SHARING

- Aligns with multiple federal and NIH Missions:
 - 2011: Precision Medicine Report
 - 2012: Data and Informatics Working Group (DIWG) of the Advisory Committee to the NIH Director
 - 2013: OSTP White House Memo - Increasing Access to Federally Funded Scientific Research Results
 - 2013: NIH Big Data to Knowledge (BD2K) Initiative
 - 2014: National Strategy for Combating Antibiotic-Resistant Bacteria (CARB)
 - 2018: NIH Strategic Plan for Data Science



EXPAND USER BASE - INTERACTIONS OF USERS & CONTRIBUTORS



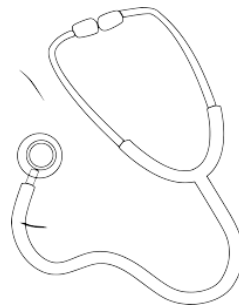
PI / Biologist
web access



**Computational
Research Scientist**
Python, R, SQL



Algorithm Developer
ssh, programmatic
access



ANTICIPATED OUTCOMES – DEMOCRATIZE RESOURCES

Discover & access relevant generated data, tools, & standards

Identify/provide home for research data that needs a home

Enable new paradigm/model of data-driven hypothesis-driven science

Continuous assessment and optimization of data science research resources

Reduce informatics costs in the long run

FUNDING OPPORTUNITY ANNOUNCEMENTS (FOA)

Secondary Analysis of Existing Datasets for Advancing Infectious Disease Research (R21 Clinical Trial Not Allowed)

R21 Exploratory/Developmental Research Grant

New

None

PA-19-068

Informatics Methodology and Secondary Analyses To Explore Shared Immunology Study Data in ImmPort

Program Announcement With Special Receipt, Referral, and or Review Considerations—proposed FY 2020 initiative

Webale

Ishwar Chandramouliswaran
Office of Data Science & Emerging Technologies
NIAID, NIH, DHHS



@ishwartweets
<https://www.linkedin.com/in/ishwarc/>

