Response to Three Key Questions from One Person from One IC

BD2K All Hands Meeting – Panel Session

John J. McGowan, Ph.D.
Deputy Director Science Management, NIAID
November 30, 2016
Examples
Knowledge Gained From Large Data Sets

- Machine learning technologies applied to large data sets for AMR/TB - 4000+ drug susceptible and resistant strains with other public data sets and able to predict antimicrobial resistance for high priority pathogens.

- Analyzing 1400 genomes of Ebola virus recently revealed a naturally occurring genetic mutation that increases transmission.

- Predictive signatures/models of disease from integrating large diverse data sets - predicting risks for infection, severity of disease, response to therapeutics.

- Studying gut microbiome have begun establish and confirm both the role and function of the microbiome in allergy, asthma, AMR, Clostridium difficile, HIV, malaria and responses to vaccines, new data, new knowledge that may be clinically actionable data after validation.
The Human Microbiome Project
Funded by the NIH Common Fund, FY2007-2016
Phase I: 242 "Healthy Human Subjects"

Phase II "Integrative HMP"

Explore biological properties from both the microbiome and host from three different cohort studies of microbiome-associated conditions using multiple "omics" technologies.

- Pregnancy and Pre-term birth
- Inflammatory bowel disease
- Prediabetes

Data available at http://hmp2.org/resources/data_browser.php
(why we built it, what it does, how, & for whom)

**Extend HMP Investment**

- Allow users to download data and install tools locally to perform analysis
- Enable efficiencies and effective work toward sustainable solutions that require less maintenance, technology and infrastructure
- Engage the research community to establish standard and reproducible approaches to data analysis and data sharing
MEET NEPHELE
Microbiome analysis without boundaries.

Free Promotional Access
Submit your email address to receive single-use promotional codes.

Alternatively, launch Nephele using your own AMR account — learn more

Analysis Engine
Use the Nephele cloud to analyze your data. Learn more

3D Microbiome
Visualize data from the Human Microbiome Project. Learn more

User Guide
Find information for making use of the Nephele platform. Learn more

About
Learn about the Nephele project & the human microbiome. Learn more

https://nephele.niaid.nih.gov/#home

Break free from data analysis constraints
- No setup required—get right to on-demand data analysis on the cloud
- No need for dedicated hardware for computation and storage, or for specialists to perform analysis
The Immune Tolerance Network

immunetolerance.org

Accelerating clinical development of immune tolerance therapies and biomarkers

CREATE AN ACCOUNT

Creating an account for ITN TrialShare is free and simple. Just click the button below to get started.

WHAT'S NEW – NOV 28th, 2016

Peanut Allergy Reduced by Early Exposure

Data and analyses for the follow-up study have been published. **Effect of Avoidance on Peanut Allergy after Early Peanut Consumption (EAPC)** (Gibson Lack et al, 2016) and are available in the LEAP-01 study.

"Using Data for the Public Good"

ITN TrialShare has won the National Academy of Sciences Data and Information Challenge. The Academy’s Board on Research Data and Information held the challenge themed “Using Data for the Public Good,” created to increase awareness of issues surrounding collection and use of scientific research data including accessibility and reproducibility, as well as promoting opportunities for maximum societal benefit from the available data.

Data Updates

TrialShare is currently populated with clinical and mechanistic assay data from 25 clinical trials, including 3,200 patients and 72,000 study visits.


SIGN IN

Username

Password

Forgot password?

I agree to the **Terms of Use**.

CREATE AN ACCOUNT

Your privacy is important to us and we encourage you to read our [Privacy Policy](https://www.immunetolerance.org/privacy-policy).

ABOUT US

The Immune Tolerance Network (ITN) is an international clinical research consortium sponsored by NIADDK, part of the National Institutes of Health. ITN’s mission is to advance the clinical application of immune tolerance therapies and biomarker development by performing high quality clinical trials of emerging therapies integrated with mechanistic-based research with a focus on allergy & asthma, organ transplantation, and autoimmune disease including Type 1 diabetes.

ITN TrialShare shares information about ITN’s clinical studies and specimen bio-repository with the scientific research community. Data and analysis code underlying ITN-published manuscripts are publicly available with the goal of promoting transparency, reproducibility, and scientific collaboration.

Sample Sharing: ITN also has thousands of biological specimens from clinically well-phenotyped study participants that are available for investigators to request for their own research. An inventory of the available samples and information about submitting a request for specimens are available on ITN TrialShare.

MOST VIEWED STUDIES

- **RAVE-I ND021A**: 16,050 views
- **ALATE I ND27A**: 9,331 views
- **LEAP I ND020D**: 3,145 views
- **R2-RAFA I ND010A**: 3,066 views
- **HALT-M I ND23A**: 2,022 views
- **Calsequestrin A**: 1,947 views
- **Shapiro I ND005C**: 1,772 views
- **FACTOR I ND607ST**: 1,454 views
- **WISP-1 I ND2293T**: 984 views
- **GPAC I ND25AD**: 176 views
- **Other studies**: 99 views

ASSAY SAMPLES

- **Specific Antibody**: 1056 views
- **Flow Cytometry**: 3169 views
- **Alloantibody**: 449 views
- **Autoantibody**: 359 views
- **Cytokine**: 335 views
- **Histology**: 195 views
- **Gene Expression**: 175 views
- **HLA typing**: 96 views

Samples in log scale

Updated on 10 June 2013 at 10:29 PST
Bioinformatics and Systems Biology

Computer applications have become an essential part of biomedical research to analyze the often huge amounts of data. NIAID develops and applies bioinformatics tools for sequencing and alignment, structural analysis and prediction, genome annotation, and simulations and 3D modeling.

Bioinformatics

- Bioinformatics and Computational Biosciences Branch
- Data Publishing and Format
- Epitopes and MHC/HLA
- Flow Cytometry
- Gene Expression and Transcriptome Analysis
- Genomics and DNA Analysis
- NIAID Microbiome Program
- Multiple Autoimmune Disease Genetics Consortium (MADGC)
- Multi-Resource Portals
- Mutation, Recombination and SNP
- Phylogenetics and Ontology
- Proteomics and Protein Analysis
- Systems Biology
- Software Applications
- Publications and Findings

Bioinformatics Multiresource Portals

- Bioinformatics Resource Centers Pathogen Portal (BRC)
  - Eukaryotic Pathogen Database Resources (EuPathDB)
  - Influenza Research Database (IRD)
  - Pathsystems Resource Integration Center (PATRIC)
  - VectorBase
  - Virus Pathogen Resource (VPR)

- University of Texas Medical Branch at Galveston
- Genomic Sequencing Centers for Infectious Diseases
  - The Broad Institute
  - Institute for Genome Sciences at the University of Maryland School of Medicine
  - J. Craig Venter Institute (JCVI)

- HIV Databases
  - HIV Databases/HIV-1 Resistance Mutation Database
  - HIV Databases/HIV Database Tools
  - HIV Databases/HIV Molecular Immunology Database
  - HIV Databases/HIV Sequence Database

- Immune Epitope Database and Analysis Resource (IEDB)
- Immunology Database and Analysis Portal (ImmPort)
  - Beyond Data Storage - ImmPort Offers a Platform for Future Research Advances

- Simmune
- Structural Genomics Centers for Infectious Diseases
  - Center for Structural Genomics of Infectious Diseases (CSGID)
  - Seattle Structural Genomics Center for Infectious Diseases (SSGCD)

- Systems Approach to Immunology
- Systems Biology for Infectious Diseases Research
  - Systems Biology for Enteropathogens – Battelle Pacific Northwest National Laboratory
  - Systems Influenza – Institute for Systems Biology
  - TB Systems Biology – Stanford University and the Broad Institute
  - Systems Virology – University of Washington

https://www.niaid.nih.gov/research/bioinformatics-multiresource-portals
Phylogenetic Trees
IRD uses PhyML (Guindon and Gascuel, 2003) and offers multiple evolutionary models to infer phylogenies. Decoration options let you color tree leaves by metadata. Export image and legend, or download trees as Newick or PhyloXML files for other viewing software.

Key Highlights:
- Quick Tree produces an accurate topology for > 1,000 sequences. Custom Tree lets you choose an evolutionary model and parameter values.
- ModelCompare (Custom Tree) suggests an evolutionary model to best fit the data.
- Specify metadata for the leaves (strain name, etc.)

https://www.flpdb.org/brc/home.spg?decorator=influenza
The PATRIC workspace and associated services will be down for maintenance. The work to be completed within a few hours.

Comparative Analyses

Comparative Pathway Tool
Supports comparative pathways in which annotated metabolic pathways across closely related or diverse groups of genomes are aligned and visualized using interactive KEGG maps and databases. The heatmap view is an interactive visualization tool that provides an overview of the distribution of genes across the set of EC numbers within a selected pathway.

Disease View
Presents infectious disease, virulence, and outbreak data associated with various taxa. Includes an interactive graph presenting relationships between pathogens, genes, and diseases and a disease map, which geo-locates reports of associated diseases around the globe in real-time. **Note:** Links to Mycobacterium Disease View example.

Protein Family Sorter
Compares protein families across closely related or diverse groups of genomes, visualizes them using interactive heatmaps, and generates multiple sequence alignments and phylogenetic trees for individual families. The heatmap view is an interactive visualization tool that provides an overview of the distribution of proteins across a selected set of genomes.

Transcriptomics Gene List and Heatmap
Mine selected transcriptomics datasets and filter based on Log Ratio or Z-score cut-off, up/down regulation, or gene functions. The associated Heatmap is an interactive visualization providing an overview of the distribution of gene regulation within a selected set of experiments or comparisons, enabling discovery of genes with similar expression patterns. **Note:** Links to example Gene List from GEO Accession: GSE68939.

Specialized Searches

Antibiotic Resistance Genes Search
We have integrated and mapped known antibiotic resistance genes from ARDB and CARD. Find these genes for organisms of interest based on taxonomy, source, special property class, and keywords.

BLAST
Provides analysis by all BLAST flavors (blastn, blastp, blastx, tblastn, tblastx) and protein sequences from PATRIC, RefSeq, and protein databases. PATRIC also offers plasmid-specific BLAST databases containing genomic sequences and proteins (annotated by PATRIC and RefSeq) from all of the plasmids in the PATRIC database.

Correlated Genes
View gene expression profiles for all available expression profiles (positive and negative) across all available data sets. Search for genes via correlation cutoff. **Note:** Links to example Correlated Genes Tab for VBMycTub87468_2933.

Feature Finder
Locates specific genomic features(s) based on taxonomy (e.g., genus or species), genomic feature type(s), and annotation types. Currently, PATRIC supports and corresponds to PATRIC features.

Gene Page Transcriptomics
Summarizes all comparisons in which a gene is differentially expressed and associated metadata such as Strain, Gene Modification, and Experimental Condition. Expression value based filters and metadata graphical summaries enable quick discovery of top experimental conditions and gene manipulations related to the differential expression of the gene.

ID Mapping
Quickly maps PATRIC identifiers to those from other prominent external databases, such as GenBank, RefSeq, UniProt, etc. Alternatively, researchers can start with a list of external database identifiers and map them to PATRIC ID numbers by using the PATRIC features.

Specialty Gene Search
Specialty Genes refers to genes that are of particular interest to infection and disease, such as virulence factors, resistance genes, antibiotic resistance genes, drug targets, and human homologs. Use this tool to search for specialty genes of interest.

Transcriptomics Experiment and Comparison Lists
Browse through all publicly available Experimental datasets in PATRIC with the Experiment List of studies or the Comparison List of studies. Compare compositions of two samples or experimental conditions. Filter the lists via manually curated Metadata (such as Organism, Strain, Gene Modification, and Experimental Condition) to quickly find datasets of interest.
The PATRIC workspace and associated services will be offline due to infrastructure maintenance. We expect the work to be completed within a few hours.

Specialty Gene Search
Specialty genes refer to those that are of particular interest to infectious disease researchers, such as virulence factors, antibiotic resistance genes, drug targets, and human homologs. Use this tool to search for specialty genes of interest.

Transcriptomics Experiment and Comparison Tools
Browse through all publicly available Experimental datasets in PATRIC with the Experiment UX or the Comparison of Curated pair-wise comparisons of two samples or experimental conditions. Filter the lists via manually curated Metadata (such as Organism, Strain, Gene Modification, and Experimental Condition) to quickly find datasets of interest.

Differential Expression Data Upload
This service allows you to upload differential expression data to your workspace and compare it with other expression data in PATRIC. This tool supports differential expression data in the form of log ratios, generated by comparing to samples/conditions/time points.

Specialized Searches

**Antibiotic Resistance Genes Search**
We have integrated and mapped antibiotic resistance genes from ARDB and CARD. Find those genes for organisms of interest based on taxonomy, source, special property classes, and keywords.

**BLAST**
BLAST provides analysis by all BLAST flavors (blastn, blastx, blast2seq, tblastn, tblastx) against genes and protein sequences from PATRIC, RefSeq, and third party annotations. PATRIC also offers plasmid-specific BLAST databases containing genomic sequences and proteins (annotated by PATRIC and RefSeq) from all of the plasmids in the PATRIC database.

**Page Gene Transcriptions**
Summarizes all comparisons in which a gene is differentially expressed and associated metadata such as Strain, Gene Modification, and Experimental Condition. Expression value based filters and metadata graphical summaries enable quick discovery of top experimental conditions and gene manipulations related to the differential expression of the gene. Link to example Gene Page for YVdWh76450.2003.

**Pipeline and Services**

**Genome Assembly**
The Genome Assembly Service App can be used to perform an automatic genome assembly using the latest computational tools. Single or multiple assemblers can be invoked to compute high quality genome annotations. Note: You must have a PATRIC user account and be logged in to use this service.

**MG-RAST**
MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data. The server provides web based upload, quality control, automated annotation and analysis for samples and datasets.

**Phylogeny Viewer**
Allows exploration of phylogenetic relationships between species and genus-level coloring schemes. PATRIC’s phylogeny viewer also supports custom groups of PATRIC genome groups, to be used as a basis for analysis in other PATRIC tools.

**Comparison Region Viewer**
Allows comparison of genomic regions of interest across closely related genomes. Detect differences in translation start sites, potential frame-shifts, or missing genes; visually identify group proteins with similar functions. Link to RNA polymerase factor RpoD (YVdWh76450.2003.84.2003) in Mycobacterium tuberculosis H37Rv and closely related genomes.

**Overview Browsers**

**Blat**
BLAT provides analysis by all BLAST flavors (blastn, blastx, blast2seq, tblastn, tblastx) against genes and protein sequences from PATRIC, RefSeq, and third party annotations. PATRIC also offers plasmid-specific BLAST databases containing genomic sequences and proteins (annotated by PATRIC and RefSeq) from all of the plasmids in the PATRIC database.

**Correlated Genes**
View genes with correlated expression profiles (positive and negative) across all available datasets. See gene functions and filters via correlation cutoff. Note: Links to example Correlated Genes Tab for YVdWh76450.2003.

**Feature Finder**
Locates specific genomic features(s) based on taxonomy (e.g., genus or species), genomic feature type (e.g., gene, CDSD, RNA, etc.), keyword, sequence status, and/or annotation type. Currently, PATRIC supports over 40 genomic feature types.

**Pathogen Portal**
Supports probing and analyzing high-throughput antibiotic resistance data using PATRIC genome as references, providing alignment and quantitative expression data, records provenance information, batch analysis for multiple samples, and results sharing and publishing. Note: Links to Pathogen Portal.

**RNA-Seq Pipeline**
RNA-Seq Pipeline provides a streamlined pipeline for analysis of RNA-seq data. Supports alignment and profiling of RNA-seq data using RAST (Rapid Annotation using Subsystem Technology) to provide high quality genome annotations. Note: Must have a PATRIC user account and be logged in to use this service.

**Rapid Annotation**
Rapid Annotation Service (RAS) is a fully-automated pipeline for the analysis of large genome datasets.

**Transcriptomics Experiment and Comparison Tools**
Browse through all publicly available Experimental datasets in PATRIC with the Experiment UX or the Comparison of Curated pair-wise comparisons of two samples or experimental conditions. Filter the lists via manually curated Metadata (such as Organism, Strain, Gene Modification, and Experimental Condition) to quickly find datasets of interest.

**Transcriptomics Gene List and Heatmap**
View selected transcriptomics datasets and filter based on Log ratio or z-score cutoff, gene regulation, or gene functions. The associated Heatmap is an interactive visualization providing an overview of the distribution of gene regulation within a set of experiments or comparisons. Mouseover discovery of genes with similar expression patterns. Note: Links to example Gene List from QPAC Accession: QPAC80001.
ImmPort’s Mission

ImmPort is a long-term, sustainable data warehouse for the purpose of promoting re-use of immunological data generated by NIADDK and DMD funded investigators.

ImmPort supports analysis of flow cytometry results and HLA genetic associations.

What is ImmPort

ImmPort Galaxy

The ImmPort flow cytometry analysis tool suite moved to ImmPort Galaxy! All your ImmPort flow cytometry analysis tools, plus many more, are now accessible only at immportgalaxy.org. Please contact us if you have any questions at Disc_Helpdesk@niaid.nih.gov

Open ImmPort

- Browse and search for shared study data
- Cytokine and cell interaction literature mining: ImmuneExpresso
- Example R and Python analysis code
- Cytokine registry
- Cell Ontology Visualizer

Data Release

June 20, 2016 - ImmPort Data Release 19 is out with 11 new studies. 6 studies from Clinical Trials in Organ Transplantation (CTOT) project contain parsed clinical lab test data, adverse events data and rejection and graft loss assessments - see studies SDY479, SDY557, SDY571, SDY670 and SDY689. Another transplantation study of interest, SDY788, submitted by Dr. Julie Yabu of Stanford University School of Medicine is a retrospective immune profile evaluation of sensitized kidney transplant candidates with flow cytometry, CyTOF and array assay data. Data Release notes can be accessed at Data Release 19.

Data Summary

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies</td>
<td>222</td>
</tr>
<tr>
<td>Subjects</td>
<td>37140</td>
</tr>
<tr>
<td>Experiments</td>
<td>1011</td>
</tr>
<tr>
<td>Total Results</td>
<td>940340</td>
</tr>
<tr>
<td>ELISA Results</td>
<td>210620</td>
</tr>
<tr>
<td>ELISPOT Results</td>
<td>30525</td>
</tr>
<tr>
<td>Flow Cytometry Results</td>
<td>90466</td>
</tr>
<tr>
<td>Gene Expression Results</td>
<td>14</td>
</tr>
<tr>
<td>HAI Results</td>
<td>43473</td>
</tr>
<tr>
<td>Total Studies</td>
<td>8722</td>
</tr>
</tbody>
</table>
Welcome to ImmPort Galaxy!

This version of ImmPort Galaxy is a beta-release. Don't take anything on this page for granted and expect unexpected "features". Thank you for your understanding!

ImmPort Galaxy is your resource for analyzing flow cytometry data. Create a login and get started!

This guide can help you get things going. Get data from ImmPort, upload your own or use one of the test datasets published in Shared Data. Explore available tools on the left panel or use one of the available published workflows in Shared Data.

If there is a tool you would like to use, or if you'd like to contribute a tool please do contact us here.

Behold the power of Parallel Coordinate Graphs

Compare markers and samples in one interactive look
Erosion Over the Last Decade

Unsolicited

Solicited to

- Build a portfolio in a new area
- Establish a pipeline
- Address clinical priorities
- Address research resources and infrastructure needs
ICs Responsibility - BD2K Responsibility
Ensure the Best Science

Extramural

Intramural
Debate on Kit Science

Research Technologies Branch

Areas of Expertise

- Light microscopy (confocal, multiphoton, colocalization, TIRF, FRET, high-resolution 3D imaging, laser microcapture, correlative techniques, and post-collection imaging processing)
- Electron microscopy (high-resolution scanning and transmission, cryoimmobilization/viewing, and immunolocalization of selected antigens)
- Flow cytometry (up to 13-color sorting, up to 16-color analysis, BSL-3 sorting and analysis, multispectral imaging cytometry, and multiplex bead array assays)
- Custom antibodies (hybridoma expansion, purification, and labeling)
- Protein chemistry (peptide synthesis, protein sequencing, mass spectrometry, protein identification, protein separation, and assay development)
- Genomics (Agilent Sureprint custom spotted, Illumina BeadChip, and Affymetrix microarrays: microarray design; Illumina and Roche next-generation DNA sequencing and Q-PCR)
- Bioinformatics and biostatistics (experiment design, data management, statistical analysis, exploratory analysis, data mining, and database integration)
Models for Capacity Building in Institutions

- **Core Facility Model:** Centralized, cross-cutting big data “centers of excellence” within an IC, which can provide shared services across research domains and lines of business
  - **Pro:** Can be more efficient and cost-effective, as well as accommodate/help to develop complementary skill sets of individuals in the core
  - **Con:** Can be less agile and requires more oversight to manage and prioritize activities

- **Nested Capability Model:** Distributed model of data scientists embedded into various units of the organization
  - **Pro:** Agility coupled with direct access to embedded domain experts can provide efficient results to certain types of needs/questions
  - **Con:** May be less equipped to handle complex or cross-cutting questions due to limited resources

- **Outsourcing Model:** Hire experts to do big data analysis work
  - **Pro:** Can provide faster access to expertise that doesn’t have to be in-house
  - **Con:** Costly, with potential for vendor/technology lock-in, along with high switching costs to eventually change vendors or take big data initiatives on in-house
Challenge for Data-Intensive Science

- Is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Guiding principles to make data:
  - Findable,
  - Accessible,
  - Interoperable, and
  - Re-usable.

https://www.force11.org/fairprinciples
Other Related Challenges

- Availability of online content is not enough to claim success of an open data policy – what does?
- What seem like straightforward requirements can be hard for researchers to understand or difficult to achieve.
- Open data is the result of strategic, legal, scientific, community and operational issues with scientific, health care and cultural institutions.
- Open access as an alternative for open data; copyrighted material, licensed, patented, non-commercial, non-transformative.
Some Specific Challenges

- Setting realistic expectations for helping to solve Big Data challenges by:
  - NIH (BD2K, NLM, NCBI, CIT)
  - IC’s (NIAID)

- Incentives to share data types that will allow re-use of data-data of knowledge value

- Data integration of diverse data sets and generation predictive models and signatures-computational tools needed in modeling, statistical analysis, and unmet need is the next step which is validation of the data sets for potential use in clinical decisions/clinical application

- Security and sustainability or not over time

- Prioritizing in limited budgets, diverse data sets and use of the data
Needed from BD2K

- Leadership
- Incentives for the Research Community and their ICs to share genomic and other data types that will allow re-use of data-data of knowledge value and integration of the diverse sets by supporting pilots, innovate for the future keep clear from what IC should support.
- Creating a community of practice and knowledge sharing across IC’s, scientific, engineering and analytical forums
- Training-workforce development and investment at NIAID intramural and extramural communities. Training computational scientists and “biologist (many categories) who understand both the computational tools and data analysis and the questions and hypothesis to be asked with data sets. This also includes training in statistics and modeling.
What is on the Immediate Horizon?
Thank You