

Immunological Data Standards and Repositories

R. Burke Squires



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

Share — copy and redistribute the material in any medium or format



Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

Please Note: parts of this work have been borrowed from other individuals and may contain a different license.

Objectives

- Enable you to name one or more immunologically relevant data standards
- Enable you to search for immunologically relevant data repositories
- Enable you to take a step in becoming a reproducible scientist

Outline

- Immunological Data
 - Standards
 - Repositories

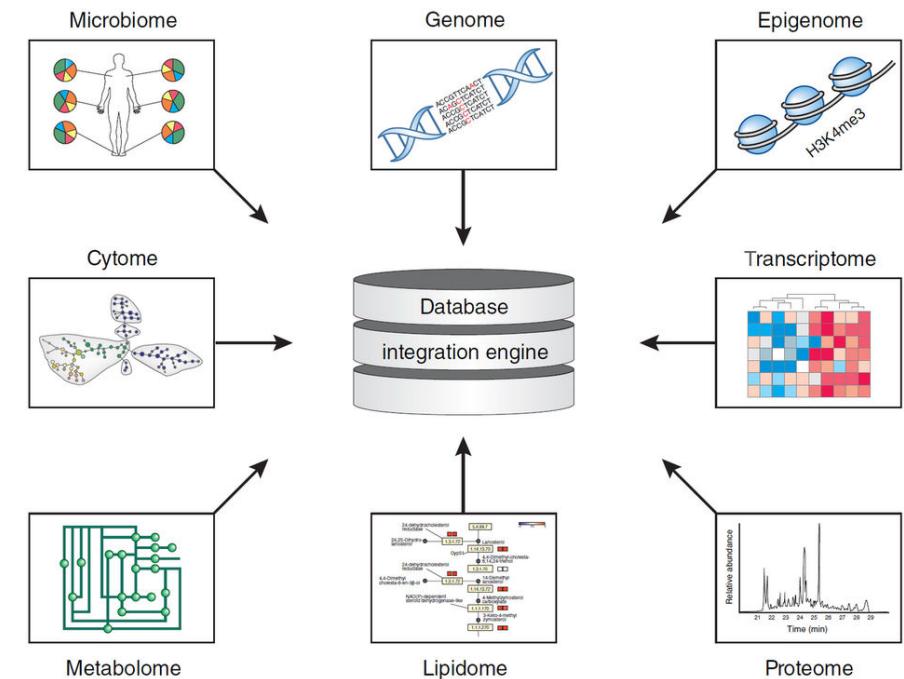


Image: Schultze, J. L. Teaching ‘big data’ analysis to young immunologists. *Nature Immunology* **16**, 902–905 (2015).

Immunological Data

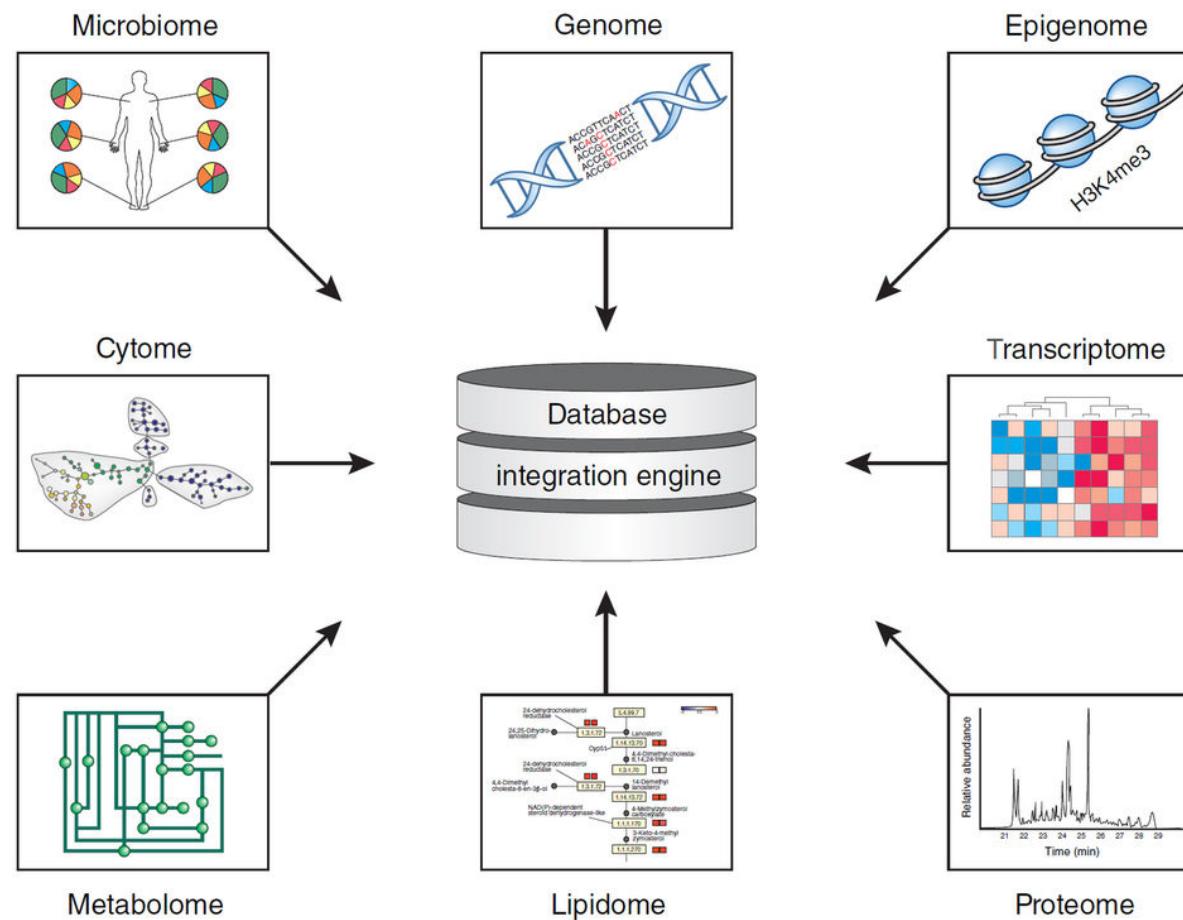


Image: Schultze, J. L. Teaching 'big data' analysis to young immunologists. *Nature Immunology* **16**, 902–905 (2015).

Data Standards

Data Standards

- What do we mean by standards
- Why have them?
- What is meant by minimal information?
- Where can you find data standards?
- What are some immunologically relevant data standards?

Data Standards: Standards – a Definition

- Agreed-upon conventions for doing ‘something’, established by community consensus or an authority
 - e.g. managing a process or delivering a service

Data Standards: Interoperability Standards

- Nuts and Bolts
- Standardization initiated in 1864 by William Sellers
- Only widely adopted a century later!



Data Standards: Interoperability Standards

- Agreed-upon specifications, guidelines or criteria designed to ensure data and any other digital object (such as code, algorithms, workflows, models, software, or journal articles) are FAIR

- FAIR
 - Findable
 - Accessible
 - Interoperable
 - Reusable

www.nature.com/scientificdata

SCIENTIFIC DATA 

OPEN **Comment: The FAIR Guiding Principles for scientific data management and stewardship**

Mark D. Wilkinson *et al.*^{*}

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplary implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data is poorly positioned to support these needs, especially for large-scale research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the long-term curation of valuable digital assets, with the goal that they should be discoverable and re-used for downstream applications. When combined with newly generated data, the outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is again a challenge for the data repository and data manager communities. To clarify around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publications. Importantly, it is our view that the principles apply not only to 'data' in the traditional sense, but also to algorithms, tools, and workflows that relate to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

*Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl). #A full list of authors and their affiliations appears at the end of the paper.

SCIENTIFIC DATA | 3:160018 | DOI:10.1038/sdata.2016.18

1

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).

Data Standards: Interoperability Standards – Fundamentals

- Essential for discovery of data and any other digital object, but also citation and credit
- Enable the operational processes, underlying their (re)use
 - such as exchange, aggregation, integration, comparison
- Optimal interoperability is reached when all processes are automated for both human and machine, this requires
 - metadata: or descriptors for the digital objects to understand what it is, where to find it, how to access it etc.
 - identifiers: unique, resolvable and versionable identifiers are essential elements of the digital word

Data Standards: Interoperability Standards – Realizing Potential

- Identifiers and metadata must be implemented by technical experts in tools, registries, catalogues, databases, services
 - to find, store, manage (e.g., mint, track provenance, version) and aggregate (e.g., interlink and map etc.) these digital objects.
- Implementations are essential to make standards ‘invisible’ to users, such as researchers, who often have little or no familiarity with them

Data Standards: Metadata Standards – Fundamentals

- Metadata - data about the data!
- Descriptors for a digital object that help to understand what it is, where to find it, how to access it etc.
- The type of metadata depends also on the digital object
- The depth and breadth of metadata varies according to their purpose (discovery, citation, credit)
 - e.g. reproducibility requires richer metadata

Data Standards: Metadata Standards – Fundamentals

Sharing starts with good metadata...but this not!

S1Sh.cuo			
	A	B	C
	Group1	Group2	D
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
Day 0			
	Sodium	139	142
	Potassium	3.3	4.8
	Chloride	100	108
	BUN	18	18
	Creatinine	1.2	1.2
	Uric acid	5.5*	6.2*
Day 7			
	Sodium	140	146
	Potassium	3.4	5.1
	Chloride	97	108

Annotations pointing to specific cells:

- Unhelpful document name (points to the header cell)
- Meaningless column titles (points to the Group1 and Group2 headers)
- No units (points to the numerical values in the Day 0 row)
- Special characters can cause text mining errors (points to the asterisk in the Uric acid values)
- Formatting for information that should be in metadata (points to the row numbers 1 through 12)
- Undefined abbreviation (points to the Potassium value in the Day 0 row)

Data Standards: Metadata Standards – Fundamentals

....this is much clearer!

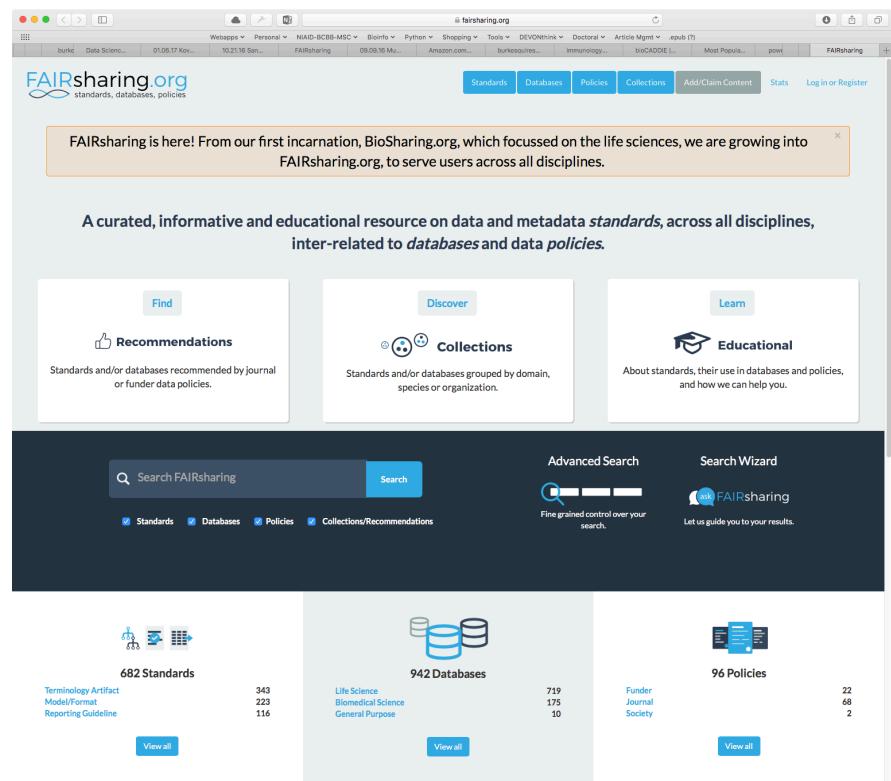
Table_S1_Shanghai_blood.xls						
	A	B	C	D	E	F
1	Parameter	Day	Control	Treated	Units	P
2	Sodium	0	139	142	mEq/l	0.82
3	Sodium	7	140	146	mEq/l	0.70
4	Sodium	14	140	158	mEq/l	0.03
5	Sodium	21	143	160	mEq/l	0.02
6	Potassium	0	3.3	4.8	mEq/l	0.06
7	Potassium	7	3.4	5.1	mEq/l	0.07
8	Potassium	14	3.7	4.7	mEq/l	0.10
9	Potassium	21	3.1	3.6	mEq/l	0.52
10	Chloride	0	100	108	mEq/l	0.56
11	Chloride	7	97	108	mEq/l	0.68
12	Chloride	14	101	106	mEq/l	0.79

Data Standards: Searching for Standards

- Where do we find them?

- Biosharing.org is now FAIRsharing.org

- FAIRsharing.org
 - Standards
 - Databases
 - Policies
 - Collections
 - Add...



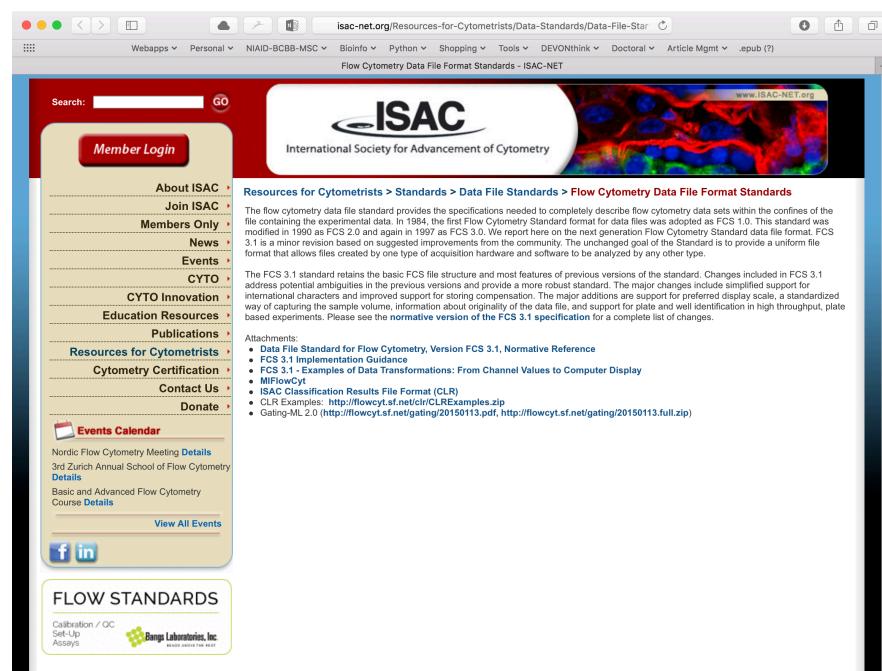
Demo

What Are Some Immunologically Relevant Data Standards?

- Flow Cytometry Data File Format Standards (FCS 3.1)
- ISAC Classification Results File Format (CLR)
- ISAC standard for representing gating descriptions in flow cytometry (Gating-ML)
- Minimum Information for Biological and Biomedical Investigations (MIBBI)
 - Minimum Information About a Bioinformatics investigation (MIABI)
 - Minimum Information about a Flow Cytometry Experiment (MIFlowCyt)
- LINCS Extended Metadata Standard
 - Antibody reagents
 - Cell lines
 - Differentiated cells
 - Embryonic stem cells
 - iPSCs
 - Nucleic acid reagents
 - Other reagents
 - Primary cells
 - Proteins
 - Small molecules

Flow Cytometry Data File Format Standards (FCS)

- The flow cytometry data file standard provides the specifications needed to completely describe flow cytometry data sets within the confines of the file containing the experimental data.
- Goal of the Standard is to provide a uniform file format that allows files created by one type of acquisition hardware and software to be analyzed by any other type.
- FCS Is implemented in
 - FlowRepository
 - The Immunology Database and Analysis Portal - OpenImmport



ISAC Classification Results File Format (CLR)

- CLR has been developed to exchange the results of manual gating and algorithmic classification approaches in a standard way in order to be able to report and process the classification.
- Although it was originally designed for the field of flow cytometry, it is applicable in any domain that needs to capture either soft or unambiguous classifications of virtually any kinds of objects.

<http://flowcyt.sf.net/circlatest.pdf>

CLR – the Classification Results File Format



Classification Results File Format **CLR**

International Society for Advancement of Cytometry
Candidate Recommendation

Document Status

This document is an ISAC Candidate Recommendation. Features and design aspects specified in this document may be changed in the final version of the Recommendation. The CLR file format has not changed since CLR 1.0 version 110318, only the format documentation and examples have been improved.



The work may be used under the terms of the [Creative Commons Attribution-ShareAlike 3.0](http://creativecommons.org/licenses/by-sa/3.0/legalcode) license. You are free to share (copy, distribute, and transmit), and adapt the work under the conditions specified at <http://creativecommons.org/licenses/by-sa/3.0/legalcode>.

Disclaimer of Liability

The International Society for Advancement of Cytometry (ISAC) disclaims liability for any injury, harm, or other damage of any nature whatsoever, to persons or property, whether direct, indirect, consequential, or otherwise, arising from the use of information contained in this Specification, or reliance on this Specification, and users of this Specification, as a condition of use, forgive release ISAC from such liability and waive all claims against ISAC that may in any manner arise out of such liability. ISAC further disclaims all warranties, whether express, implied or statutory, and makes no assurances as to the accuracy or completeness of any information published in the Specification.

In issuing and making this Specification available, ISAC is not undertaking to render professional or other services for or on behalf of any person or entity, nor is ISAC undertaking to perform any duty owed by any person or entity to someone else. Anyone using this document should rely on his or her own independent judgment or, as appropriate, seek the advice of a competent professional in determining the exercise of reasonable care in any given circumstance.

Attention is called to the potential implementation of this Specification may require use of subject matter covered by patent rights. By publication of this Standard, no action is taken with respect to the existence or validity of any patent rights in connection therewith. ISAC shall not be responsible for identifying patents or patent applications for which a license may be required to implement an ISAC standard or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention.

This work is supported by NIH/NIBIB supplemental award to grant no 1R01EB008400.
CLR 1.0, ISAC Candidate Recommendation, version 140903, September 3, 2014.

Version 1.0, 140903

Warning: This is an ISAC Candidate Recommendation

i

<http://isac-net.org/PDFS/65/659585c7-9ef4-4152-829c-859aad25aafb.pdf>

NIH Library of Integrated Network-Based Cellular Signatures (LINCS)

- The Library of Integrated Network-Based Cellular Signatures (LINCS) Program aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents.
- Data portal
 - <http://lincsportal.ccs.miami.edu/dcic-portal/>
- Cells
 - <http://lincsportal.ccs.miami.edu/cells/>

Minimum Information About a Flow Cytometry Experiment (MIFlowCyt)

- Minimum Information about a Flow Cytometry Experiment (MIFlowCyt)
 - Outlines the minimum information required to report about flow cytometry experiments
 - Represents the community consensus
 - 33 coauthors from 19 institutions
 - ISAC Recommendation
 - Required/recommended by Cytometry A and Nature



ORIGINAL ARTICLE

Cytometry
Journal of the International Society for Advancement of Cytometry

MIFlowCyt: The Minimum Information About a Flow Cytometry Experiment

Jamie A. Lee,^{1†} Josef Spidlen,^{2†} Keith Boyce,³ Jennifer Cai,¹ Nicholas Crosbie,⁴ Mark Dolphin,⁵ Jeff Furlong,⁶ Maura Gasparetto,² Michael Goldberg,⁷ Elizabeth M. Goralczyk,⁸ Bill Hyun,⁹ Kirstin Jansen,⁶ Tobias Kollmann,¹⁰ Megan Kong,¹ Robert Leif,¹¹ Shannon McWeeney,^{12,13,14} Thomas D. Moloshok,³ Wayne Moore,¹⁵ Garry Nolan,¹⁶ John Nolan,¹⁷ Janko Nikolic-Zugich,¹⁸ David Parrish,³ Barclay Purcell,¹⁹ Yu Qian,¹ Biruntha Selvaraj,¹⁹ Clayton Smith,² Olga Tchuvatkina,⁷ Anne Wertheimer,²⁰ Peter Wilkinson,²¹ Christopher Wilson,⁶ James Wood,²² Robert Zigon,²³ The International Society for Advancement of Cytometry Data Standards Task Force, Richard H. Scheuermann,^{1,24} Ryan R. Brinkman^{2*}

Abstract
A fundamental tenet of scientific research is that published results are open to independent validation and refutation. Minimum data standards aid data providers, users, and publishers by providing a specification of what is required to unambiguously interpret experimental findings. Here, we present the Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) standard, stating the minimum information required to report flow cytometry (FCM) experiments. We brought together a cross-disciplinary international collaborative group of bioinformaticians, computational statisticians, software developers, instrument manufacturers, and clinical and basic research scientists to develop the standard. The standard was subsequently vetted by the International Society for Advancement of Cytometry (ISAC) Data Standards Task Force, Standards Committee, membership, and Council. The MIFlowCyt standard includes recommendations about descriptions of the specimens and reagents included in the FCM experiment, the configuration of the instrument used to perform the assays, and the data processing approaches used to interpret the primary output data. MIFlowCyt has been adopted as a standard by ISAC, representing the FCM scientific community including scientists as well as software and hardware manufacturers. Adoption of MIFlowCyt by the scientific and publishing communities will facilitate third-party understanding and reuse of FCM data. © 2008 International Society for Advancement of Cytometry

¹Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas 75390

²Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada V5Z 1L3

³Immune Tolerance Network, Pittsburgh, Pennsylvania 15238

⁴Inival Technologies, Mentone 3194, Victoria, Australia

⁵Amgen Inc., Thousand Oaks, California 91320

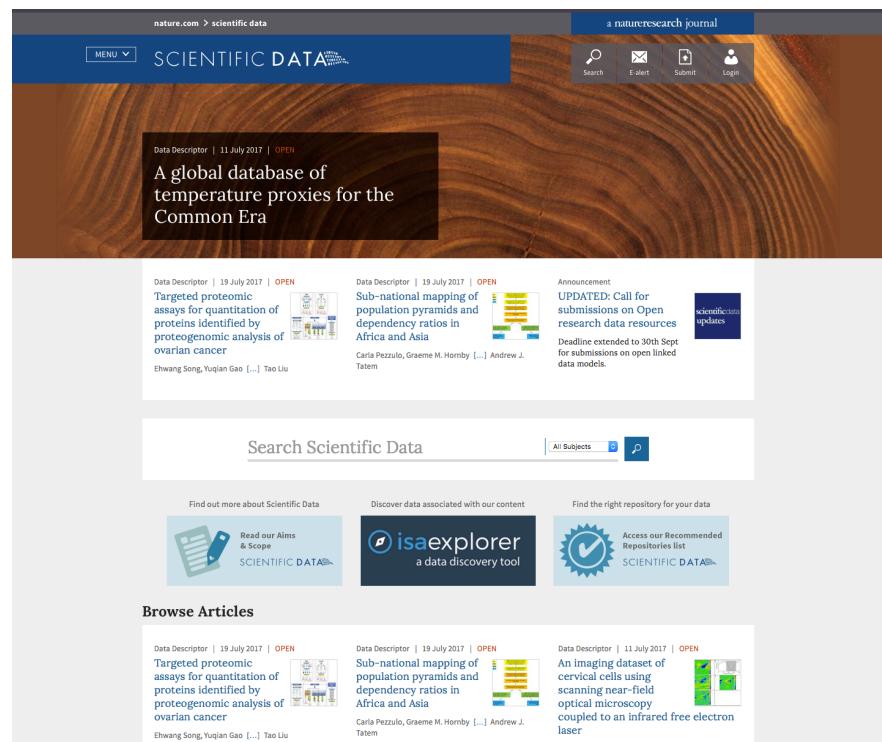
⁶Department of Immunology, University of Washington, Seattle, Washington 98195

⁷BD Biosciences, San Jose, California 95131

Data as a Primary Resource

Data as a Primary Resource

- It gets its own journal – from Nature Publishing Group!
- *Scientific Data*



Not Everyone is Happy With Data Reuse: “Research Parasites”

- On January 21, 2016, the Editors of the New England Journal of Medicine, Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D., published an editorial that characterized [some] scientists who re-analyzed published data sets as "research parasites"

The NEW ENGLAND JOURNAL OF MEDICINE

EDITORIALS



Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

The aerial view of the concept of data sharing is beautiful. What could be better than having new high-quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unnoted? The potential for leveraging existing results for even more benefit pays appropriate increased tribute to the patients who put themselves at risk to generate the data. The moral imperative to honor their collective sacrifice is the trump card that takes it from theory to reality.

However, many of us who have actually conducted clinical research, managed clinical studies and data collection and analysis, and curated data sets have concerns about the details. The first concern is that someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters. Special problems arise if data are to be combined from independent studies and considered comparable. How heterogeneous were the study populations? Were the eligibility criteria the same? Can it be assumed that the differences in study populations, data collection and analysis, and, most importantly, between study-specified and unspecified, can be ignored?

A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."

N ENGL J MED 374;3 NEJM.ORG JANUARY 21, 2016

The New England Journal of Medicine
Downloaded from nejm.org on July 20, 2017. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.

Immunological Repositories

Repositories

Data (meta)Repositories

- DataMed
 - Data repository search engine
- Nucleic Acid Research (NAR)
 - Database Issues
 - Website Issues
- Review Some Data Repositories

Tool (meta)Repositories

- OMICSTools
 - Immunological tools
- LabWorm
 - Immunological tools

DataMed

The screenshot shows the DataMed v2.0 web interface. At the top, there is a navigation bar with links to various categories like Webapps, Personal, NIAID-BCBB-MSC, Bioinfo, Python, Shopping, Tools, DEVONthink, Doctoral, Article Mgmt, .epub (?), and several tabs for specific datasets. The main header includes the DataMed logo (BETA version) and the bioCADDIE logo.

The main content area has a title "Engaging The Community Toward a Data Discovery Index (DataMed v2.0)". Below it is a search bar with options to search for data set or repository, and links to Advanced Search and help.

Three main sections are displayed:

- Statistics:** Shows 66 REPOSITORIES, 15 DATA TYPES, 1,375,977 DATASETS, and 4 PILOT PROJECTS.
- Top 8 Repositories:** A horizontal bar chart showing the number of datasets for each repository. The data is as follows:

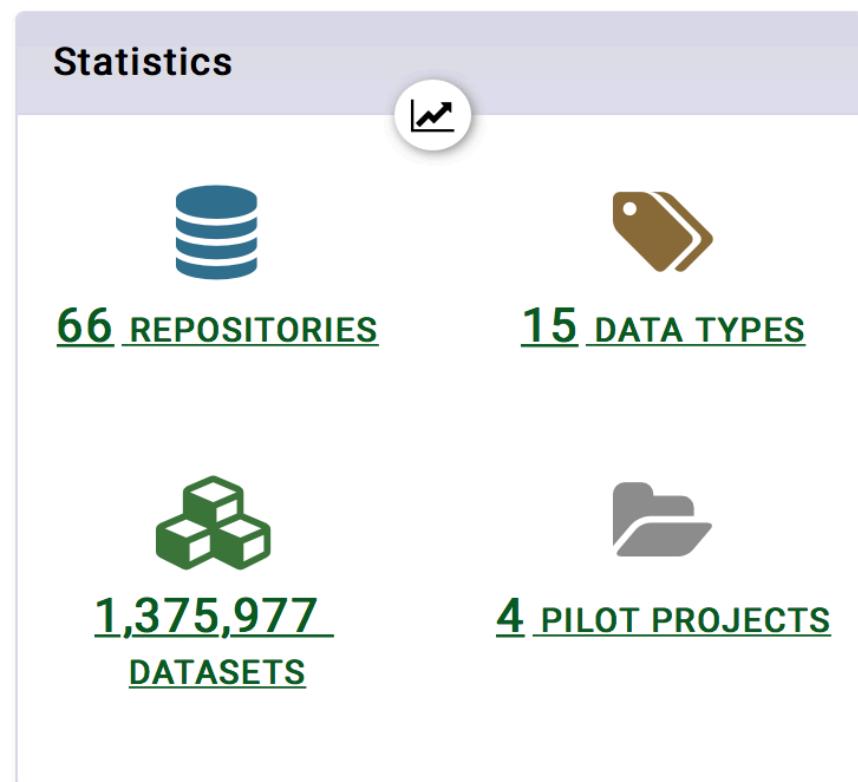
Repository	Number of Datasets
Swiss-Prot	438182
ClinVar	208560
BioProject	155850
PDB	122339
Dryad	82837
OmicsDI	78201
ArrayExpress	68189
Dataverse	60303
- New Features:** A timeline of updates:
 - Feb 28, 2017, v2.0:
 - Increase coverage to more repositories
 - Duplicate datasets display feature
 - Usability enhancements based on user feedback and user interviews
 - User-reported issues resolved
 - Nov 23, 2016, v1.5:
 - Increased coverage to twice the number of repositories
 - Total number of datasets doubled
 - Visualization of results via timeline ...

At the bottom, there is a section for Pilot Projects featuring GWAS Finder, iSEE-DELVE, DataRank, and Data Citation Discovery.

Ohno-Machado, L. et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* **49**, 816–819 (2017).

Repositories: DataMed

- “Immunology”
 - Repositories
 - Dryad (98)
 - BioProject (71)
 - PDB (67)
 - ArrayExpress (55)
 - OmicsDI (53)
 - ImmPort (45)
 - RGD (16)
 - dbGaP (12)
 - GEMMA (4)
 - ProteomeXchange (1)
 - VectorBase (1)
 - Zenodo (1)



Demo

Nucleic Acid Research (NAR) Database Issue

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
 - <http://www.oxfordjournals.org/nar/database/c>
- Cell biology
 - <http://www.oxfordjournals.org/nar/database/c>

PRINT ISSN: 0305-1048
ONLINE ISSN: 1362-4962

Nucleic Acids Research

VOLUME 45 DATABASE ISSUE JANUARY 4 2017
<https://academic.oup.com/nar>



OXFORD
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online



<http://www.oxfordjournals.org/nar/database/>

Nucleic Acid Research (NAR) Database Issues: Immunological Databases

- [AAgAtlas](#)
- [ALPSbase](#)
- [AntigenDB](#)
- [AntiJen](#)
- [BCIpep](#)
- [bNAber](#)
- [dbMHC](#)
- [DIGIT](#)
- [Epitome](#)
- [GPX-Macrophage Expression Atlas](#)
- [HaptentDB](#)
- [HPTAA](#)
- [IEDB](#)
- [IEDB-3D](#)
- [IL2Rgbase](#)
- [IMGT](#)
- [IMGT/GENE-DB](#)
- [IMGT/HLA](#)
- [IMGT/LIGM-DB](#)
- [IMGT/mAb-DB](#)
- [ImmuNet](#)
- [InnateDB](#)
- [Interferon Stimulated Gene Database](#)
- [IPD - Immuno Polymorphism Database](#)
- [IPD-ESTDAB](#)
- [IPD-HPA - Human Platelet Antigens](#)
- [IPD-KIR - Killer-cell Immunoglobulin-like Receptors](#)
- [IPD-MHC](#)
- [MHCBN](#)
- [MHCPEP](#)
- [MPID-T2](#)
- [MUGEN Mouse Database](#)
- [Protegen](#)
- [SAbDab](#)
- [SuperHaptent](#)
- [VBASE2](#)

PRINT ISSN: 2395-1048
ONLINE ISSN: 1362-4962

Nucleic Acids Research

VOLUME 45 DATABASE ISSUE JANUARY 4 2017
<https://academic.oup.com/nar>



OXFORD
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online

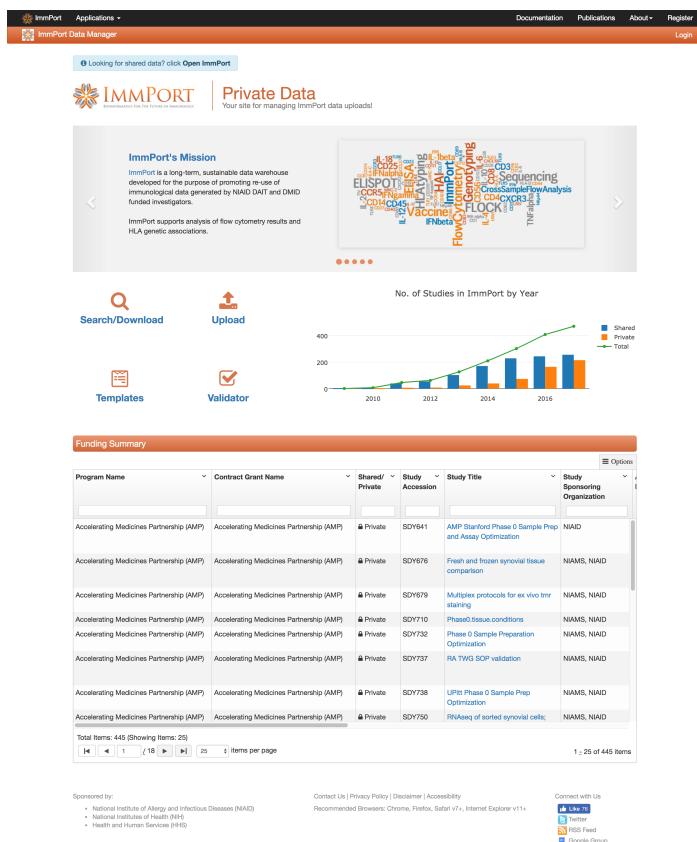


<http://www.oxfordjournals.org/nar/database/cat/14>

Data Repositories

Immunology Database and Analysis Portal (ImmPort)

- “The goals of the ImmPort project are to:
 - Provide an open access platform for research data sharing
 - Create an integrated environment that broadens the usefulness of scientific data and advances hypothesis-driven and hypothesis-generating research
 - Accelerate scientific discovery while extending the value of scientific data in all areas of immunological research
 - Promote rapid availability of important findings, making new discoveries available to the research community for further analysis and interpretation
 - Provide analysis tools to advance research in basic and clinical immunology”
- Private data and pre-release data are stored in private workspaces of investigators at the ImmPort site located at NIAID, <http://immport.niaid.nih.gov>.



<http://immport.niaid.nih.gov>

Demo

FlowRepository

- A database of flow cytometry experiments where you can query and download data collected and annotated according to the MIFlowCyt standard.

The screenshot shows the FlowRepository homepage. At the top, there is a navigation bar with links for "Login", "Help", "Logout", and "Support". Below the navigation bar, there is a search bar with the placeholder "Query FlowRepository" and a "Query" button. To the right of the search bar, there is a link to "Show query fields".

The main content area has several sections:

- Help:** A section containing an open access article about the MIFlowCyt standard, a "Quick start guide", and a "FAQ".
- FlowRepository:** A brief description of the repository as a data deposition place for experimental findings published in peer-reviewed journals in the flow cytometry field.
- Query FlowRepository:** A search interface where users can enter a term to search all publicly available experiments.
- Links:** A section with links to "Browse public datasets", "Browse OMIP datasets", "Referencing FlowRepository", "Browse community datasets", "Quick start guide", "FlowRepository Steering Committee & Advisory Board", and "Submit data".
- Citing FlowRepository:** A section with citation information, including the reference: Spidlen J, Breuer K, Rosenberg C, Kotenka N and Brinkman RR. FlowRepository: A Public Database of Annotated Flow Cytometry Datasets. Association for Advancement of Microscopy Publications. *Cytometry A*. 2012 Sep; 81(9):227-31..
- Supporting journal:** A section featuring the journal "Cytometry" with a thumbnail image of a flow cytometry histogram.

At the bottom of the page, there are links for "Terms of Service", "Privacy Policy", "Help", "Feedback", and "Developers".

Demo

Immune Epitope Database (IEDB)

- Free resource that offers easy searching of experimental data characterizing antibody and T cell epitopes studied in humans, non-human primates, and other animal species.
- Epitopes involved in infectious disease, allergy, autoimmunity, and transplant are included.
- Hosts tools to assist in the prediction and analysis of B cell and T cell epitopes.
- A free resource, funded by a contract from the [National Institute of Allergy and Infectious Diseases](#)

The screenshot shows the IEDB homepage with a search interface and various analysis tools. The search interface includes fields for Epitope, Assay, Antigen, MHC Restriction, Host, and Disease, each with specific filters and search buttons. The analysis tools section on the right includes T Cell Epitope Prediction, B Cell Epitope Prediction, and Epitope Analysis Tools, each with detailed descriptions and links to further resources. The page also features a summary of metrics and a user workshop announcement.

Summary Metrics

Peptidic Epitopes	305,461
Non-Peptidic Epitopes	2,531
T Cell Assays	321,919
B Cell Assays	394,663
MHC Ligand Assays	618,185
Epitope Source Organisms	3,612
Restricting MHC Alleles	743
References	18,622

User Workshop
25-26 October 2017
NIH, Bethesda, MD, USA
Information available at workshop.iedb.org.

Epitope Analysis Resource

- T Cell Epitope Prediction**: Scan an antigen sequence for amino acid patterns indicative of:
 - MHC I Binding
 - MHC II Binding
 - MHC I Processing (Proteasome,TAP)
 - MHC I Immunogenicity
- B Cell Epitope Prediction**: Predict linear B cell epitopes using:
 - Antigen Sequence Properties
 - Predict discontinuous B cell epitopes using antigen structure via:
 - Discotope
 - EliPro
- Epitope Analysis Tools**: Analyze epitope sets of:
 - Population Coverage
 - Conservation Across Antigens
 - Clusters with Similar Sequences

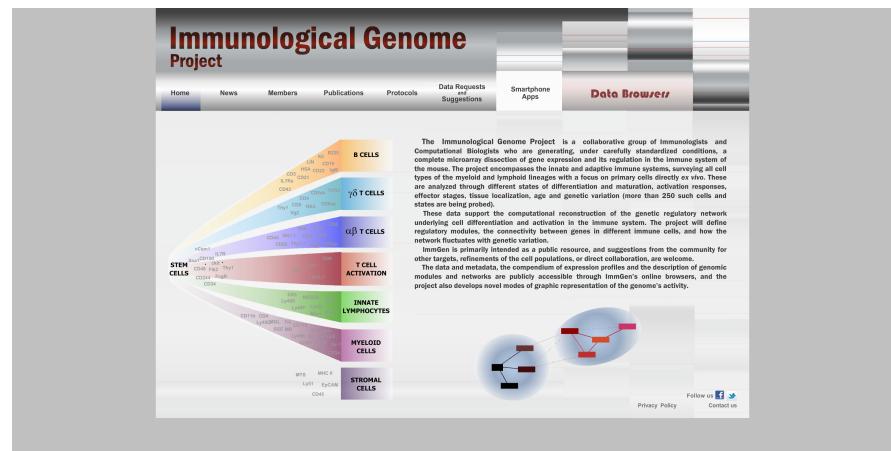
Data Last Updated: July 16, 2017

<http://www.iedb.org/>

Demo

Immunological Genome Project (ImmGen)

- Goal of the project is to computationally reconstruct the gene regulatory network in immune cells
- A gene-expression database for all characterized immune cells in the mouse
- Compendium of microarray data currently include over 250 immunologically relevant cell types, from all lymphoid organs and other tissues which are monitored by immune cells



Demo

VDJ Server

- A free, scalable resource for performing immune repertoire analysis and sharing data.
- Manage, analyze, and archive your data through our web resource.
- You can also download our open source analysis software for local use.

The screenshot displays the VDJ Server web application. At the top left is the logo 'VDJ SERVER'. To its right is a circular icon containing various scientific and analytical symbols like microscopes, test tubes, and DNA helixes. To the right of the icon is the 'WELCOME!' page, which includes fields for 'USERNAME' and 'PASSWORD', and links for 'LOGIN', 'Forgot password?', 'Create Account', 'Send Us Feedback', 'Documentation', and 'VDJServer Wiki'. Below this is a 'NEWS & ANNOUNCEMENTS' section featuring a thumbnail of a video titled 'VDJServer Intro' dated April 25, 2017. Further down are four main functional sections: 'UPLOAD' (with an upward arrow icon), 'ANALYZE' (with a magnifying glass icon), 'PUBLISH' (with a document icon), and 'SHARE' (with a person icon). At the bottom of the page is a 'About VDJServer' section with a brief description of the project's funding and partners.

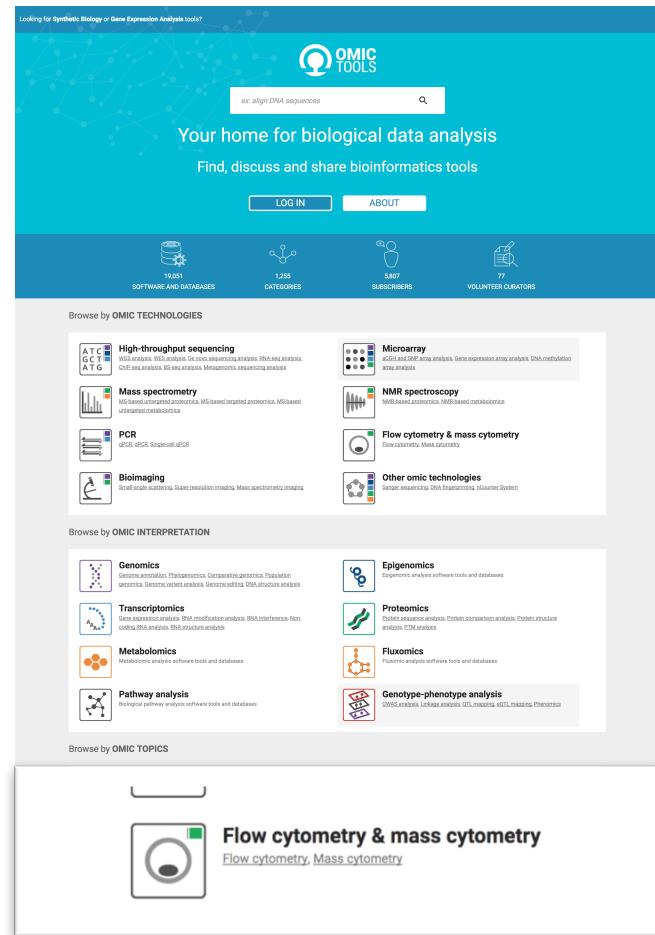
<https://vdjserver.org/>

Demo

Tool (Software) Repositories

OmicsTools

- Your home for biological data analysis
- Find, discuss and share bioinformatics tools
- 19,051 Software And Databases
- 1,255 Categories
- Browse By
 - Omic Technologies
 - Omic Interpretation
 - Omic Topics



OmicsTools

- Omic Technologies
 - Flow cytometry (81 tools)
 - Mass cytometry (15 tools)
- Omic Interpretation
- Omic Topics
 - Immunology

The screenshot shows the homepage of the OmicsTools website. At the top, there is a search bar with placeholder text "Looking for Synthetic Biology or Gene Expression Analysis tools?". Below the search bar is the OmicsTools logo with the tagline "Your home for biological data analysis" and the subtext "Find, discuss and share bioinformatics tools". There are "LOG IN" and "ABOUT" buttons. Below the header, there are four statistics: 19,051 SOFTWARE & DATABASES, 1,258 CATEGORIES, 5,807 SUBSCRIBERS, and 77 VOLUNTEER CURATORS. The main content area is divided into three sections: "Browse by OMIC TECHNOLOGIES" (High-throughput sequencing, Microarray, Mass spectrometry, PCR, Bioimaging, Other omic technologies), "Browse by OMIC INTERPRETATION" (Genomics, Epigenomics, Transcriptomics, Proteomics, Metabolomics, Fluxomics, Pathway analysis, Genotype-phenotype analysis), and "Browse by OMIC TOPICS" (Flow cytometry & mass cytometry). Each section contains a grid of icons and brief descriptions.

Demo

Labworm

- Search and discover the best online tools to assist you in your research
 - The scientist's toolbox
 - Research & productivity
 - Get help for bench work
 - Read and write science
 - Search materials and services
 - The developer's toolbox
 - Be a part of the community
 - Databases & Bioinformatic tools
 - Data & Tools by research field
 - Data & Tools by model organism
 - Health & Disease related data
 - Genomics & Transcriptomics
 - RNA & DNA
 - Proteomics & Metabolomics
 - Protein & Peptide
 - Text mining & Statistics

The screenshot shows the LabWorm website interface. At the top, there is a navigation bar with links for "Ask a question", "Post a tool", and "Sign In / Up". Below the header, a main banner features a blue-toned DNA helix background with the text "Search and discover the best online tools to assist you in your research". A search bar is positioned below the banner, with placeholder text "e.g. Statistical analysis" and a green "Search" button. To the right of the search bar is a "Explore categories" button. On the left side of the page, there is a sidebar titled "Categories" with sections for "New this week", "All time best", "The scientist's toolbox", and "Databases & Bioinformatic tools". Under "The scientist's toolbox", there are links for "Research & productivity", "Get help for bench work", "Read and write science", "Search materials and services", "The developer's toolbox", and "Be a part of the community". Under "Databases & Bioinformatic tools", there are links for "Data & Tools by research field", "Data & Tools by model organism", "Health & Disease related data", "Genomics & Transcriptomics", "RNA & DNA", "Proteomics & Metabolomics", "Protein & Peptide", and "Text mining & Statistics". Below the sidebar, there is a "Recent questions" section showing a post from "Roy Winters" with a "Awesome job with ggsl!!" comment and a "Zakher Bouragaoui" post with a "Demo" link. The main content area displays a list of tools with their logos, names, descriptions, and interaction buttons. The tools listed are TeachEng (Teaching Engine for Genomics), Labstep (A timeline for your lab experiments. Build protocols, attach results & share your data with your lab), Scientist.com (Outsource everything but the genius), and BenchSci.

Labworm - Immunology

- Collections
 - Immunological Tools
 - Margaret Jordan
 - Tools
 - IMGT
 - IMGT/HLA
 - Immgene
 - Immune Epitope Database
 - Kabat Database
 - Immuno Polymorphism Database
 - InnateDB
 - Bcepred
 - INTERFEROME
 - ABCpred
 - Antigenic

The screenshot shows the homepage of the LabWorm website. The header features the LabWorm logo, a search bar with placeholder text "e.g. Statistical analysis", and a "Search" button. Navigation links include "Ask a question", "Post a tool", "Sign In / Up", and "Explore categories". The main content area has a blue background with a DNA helix. A search bar at the top says "Search and discover the best online tools to assist you in your research". Below it is a search input field with placeholder "e.g. Statistical analysis" and a "Search" button. To the right is a "Recent questions" section with a user profile for Roy Winters. The left sidebar contains "Categories" such as "New this week", "All time best", "The scientist's toolbox", "Data & Bioinformatic tools", and "Recommended for you". The right sidebar shows "Recent questions" from Roy Winters and Zaher Bouragaoui, along with a "Ask a tool owner directly" button. The main content area displays several tool cards: TeachEng (Teaching Engine for Genomics), Labstep (A timeline for your lab experiments), Scientist.com (Outsource everything but the genius), and BenchSci.

<https://labworm.com/collection/Margaret.Jordan/immunological-tools>

Demo

GitHub

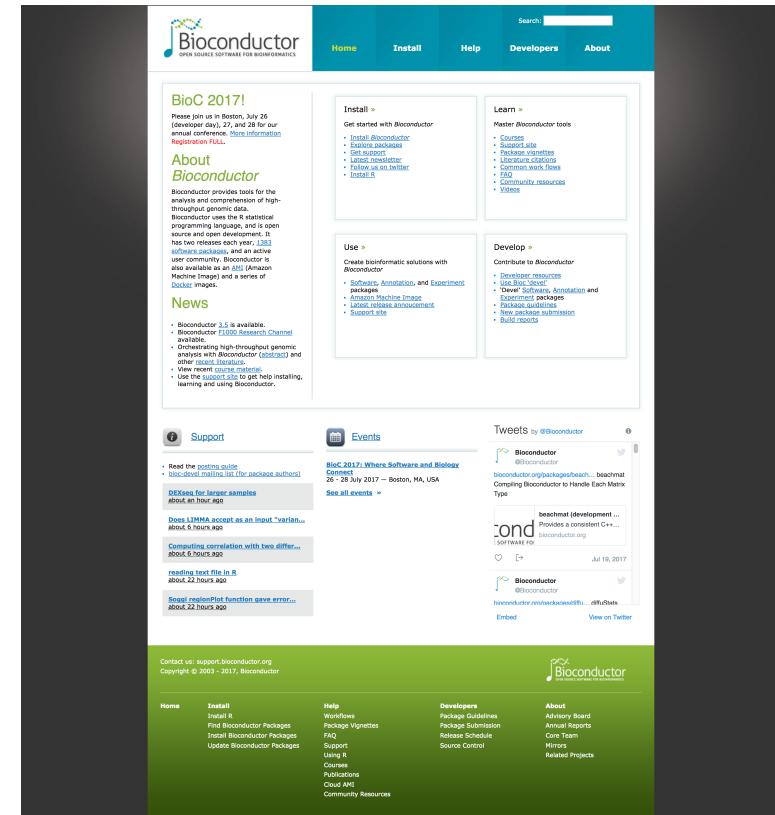
- GitHub is a development platform inspired by the way you work. From open source to business, you can...
- host and review code
- manage projects
- build software alongside millions of other developers.

The screenshot shows the GitHub repository page for 'burkesquires / immunology-informatics'. The repository has 19 commits, 1 branch, 0 releases, and 2 contributors. The code is licensed under CC-BY-SA-4.0. The repository contains files such as 01_intro_command_line, 02_Intro_R, 03_auto_flow_cyto_analysis, 04_RNAseq_analysis, 05_downstream_analysis, 06_immune_data, 07_bio_network, .gitignore, License, and README.md. The README.md file contains information about the Immunology Informatics Tutorials, which are part of the American Association of Immunologists (AAI) Course in Big Data Analysis in Immunology. It encourages users to issue pull requests and improve the materials, noting that much of the material is authored by many talented individuals. The repository also features sections for Immunology Informatics and Big Data Analysis in Immunology.

Demo

Bioconductor

- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.
- Bioconductor uses the R statistical programming language, and is open source and open development.
- It has two releases each year, 1383 software packages, and an active user community.



Demo

Reproducible Research

Becoming a Reproducible Scientist...

- Remember – take it one step at a time
 - Good – Document where you get your data, the steps you use to produce your analysis
 - Better – Script your analysis
 - Best – Scripts your analysis, upload your scripts to GitHub, contribute well document data to a data repository

Additional Resources

Resources

- Big Data 2 Knowledge (BD2K) Lecture series
 - <https://bigdatau.ini.usc.edu/data-science-seminars>
- Winipedia (via wikiwand)
 - Computational immunology
 - http://www.wikiwand.com/en/Computational_immunology
 - Flow cytometry bioinformatics
 - http://www.wikiwand.com/en/Flow_cytometry_bioinformatics
- “Awesome” Resource Lists of all types (many bioinformatics or computational)
 - <https://github.com/sindresorhus/awesome>
- Medical clipart – share alike
 - <http://smart.servier.com>

Resources Used

- BD2K Guide to the Fundamentals of Data Science Series Talks
 - <https://www.youtube.com/channel/UCKIDQOa0JcUd3K9C1TS7FLQ>
 - [Metadata Standards](#), Dr. Susanna-Assunta Sansone from University of Oxford
 - [Introduction to Big Data and the Data Lifecycle](#), Dr. Mark Musen from Stanford University

Papers of Interest

- Genser, B., Cooper, P. J., Yazdanbakhsh, M., Barreto, M. L. & Rodrigues, L. C. A guide to modern statistical analysis of immunological data. *BMC Immunology* 2007 8:1 8, 27 (2007).
 - “This paper will help the immunologist to choose the correct statistical approach for a particular research question.”
- Zhang, G. L., Sun, J., Chitkushev, L. & Brusic, V. Big Data Analytics in Immunology: A Knowledge-Based Approach. *BioMed Research International* 2014, 1–9 (2014).

Objectives

- Enable you to name one or more immunologically relevant data standards
- Enable you to search for immunologically relevant data repositories
- Enable you to take a step in becoming a reproducible scientist

Q & A