Lightweight data engineering, tools, and approaches to facilitate data reuse and data science

...

Sean Davis, MD, PhD
National Cancer Institute, National Institutes of Health
AIDR 2019, Carnegie Mellon University

https://seandavi.github.io
@seandavis12

http://bit.ly/SD-AIDR2019

Data Engineering: a Defini

https://en.wikipedia.org/wiki/data_engineering

Background and motivation

…

NIH STRATEGIC PLAN FOR DATA SCIENCE

Introduction

As articulated in the National Institutes of Health (NIH)-Wide Strategic Plan[1] and the Department of Health and Human Services (HHS) Strategic Plan,[2] our nation and the world stand at a unique moment of opportunity in biomedical research, and data science is an integral contributor. Understanding basic biological mechanisms through NIH-funded research depends upon vast amounts of data and has propelled biomedicine into the sphere of "Big Data" along with other sectors of the national and global economies. Reflecting today's highly integrated biomedical research landscape, NIH defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

approach will move toward a common architecture, infrastructure, and set of tools upon which individual Institutes and Centers (ICs) and scientific communities will build and tailor for specific needs. A Software as a Service (SaaS) framework, in which software licensing and delivery are provided and hosted by centralized resources, will greatly facilitate access to, analysis and curation of, and sharing of

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
•Optimize data storage and security •Connect NIH data systems	•Modernize data repository ecosystem •Support storage and sharing of individual datasets •Better integrate clinical and observational data into biomedical data science	•Support useful, generalizable, and accessible tools and workflows •Broaden utility of and access to specialized tools •Improve discovery and cataloging resources	•Enhance the NIH data-science workforce •Expand the national research workforce •Engage a broader community	•Develop policies for a FAIR data ecosystem •Enhance stewardship

https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

FAIR data standards

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

https://www.force11.org/group/fairgroup/fairprinciples

Is FAIR data enough?

- FAIR data standards and guidelines are general; implementation can be challenging.
- FAIR data are not always fit-for-use data.
- FAIR data standards are somewhat context-dependent and are most easily applied within a homogeneous community of data users and providers.
- FAIR data are not always usable data.
- FAIR data are not always useful data.
- [OFF TOPIC]: expense, maintenance, metrics, deprecation, augmentation and correction, provenance over time, ownership and IP

Health

In the context of healthcare, there is a virtual imperative that we learn to value and use our data ethically and effectively. The promise of a learning healthcare system depends on this charge....

Big Biological data

- Acquisition
- Storage
- Distribution
- Analysis
- Integration

PLOS | BIOLOGY

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

1 Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Carl R. Woese Institute for Genomic Biology & Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **4** School of Library and Information Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **5** Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **6** Carl R. Woese Institute for Genomic Biology, Department of Entomology, and Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* mschatz@cshl.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)

CrossMark
click for updates

OPEN ACCESS

Citation: Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Published: July 7, 2015

Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.

https://doi.org/10.1371/journal.pbio.1002195

Comparison of four domains of big data in 2025

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

https://doi.org/10.1371/journal.pbio.1002195.t001

Steel threads: Software engineering constructs for defining, designing and developing software system architecture

Issue title: Special Supplement Issue in Section A and B: Selected Papers from the ISCA International Conference on S

Guest editors

Article type: I

Disconnected and Distributed

Authors: Alko

Affiliations: [a UAE | [b] Mathematics, Engineering Process and C ment, Winona State e University, Winona, MN, U

Correspondence: [*] Corresponding author: Wan D. Bae, Statistics and Computer Science, University of Wisconsin-Stout, 231F Jarvis Hall Science Wing Menomonie, WI 54051, USA. E-mail: baew@uwstout.edu.

Abstract: A steel thread is a software engineering construct that identifies the most important execution paths, including software and hardware elements, through a computer system, while meeting business objectives and demonstrating executable architecture. Steel threads are often used in the context of defining software system architecture. Although there have been references to steel

Baby steps: Data reuse reframed as reproducible research

...

Genomics and Bioconductor as a use case....

OPINION

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] **and Roger D. Peng**[b]

[a]*Associate Professor of Biostatistics and Oncology and* [b]*Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD*

Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research. Consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hy-

been some very public failings of reproducibility across a range of disciplines from cancer genomics (3) to economics (4), and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Popular press articles have raised questions about the reproducibility of all scientific research (5),

computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate

Leek, Jeffrey T., and Roger D. Peng. 2015. https://doi.org/10.1073/pnas.1421412111

Reproducibility, the ability to recompute results, and replicability, the chances other experimenters will achieve a consistent result, are two foundational characteristics of successful scientific research...of late there has been a crisis of confidence among researchers worried about the rate at which studies are either reproducible or replicable. In order to maintain the integrity of science research and maintain the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely employs software tools.

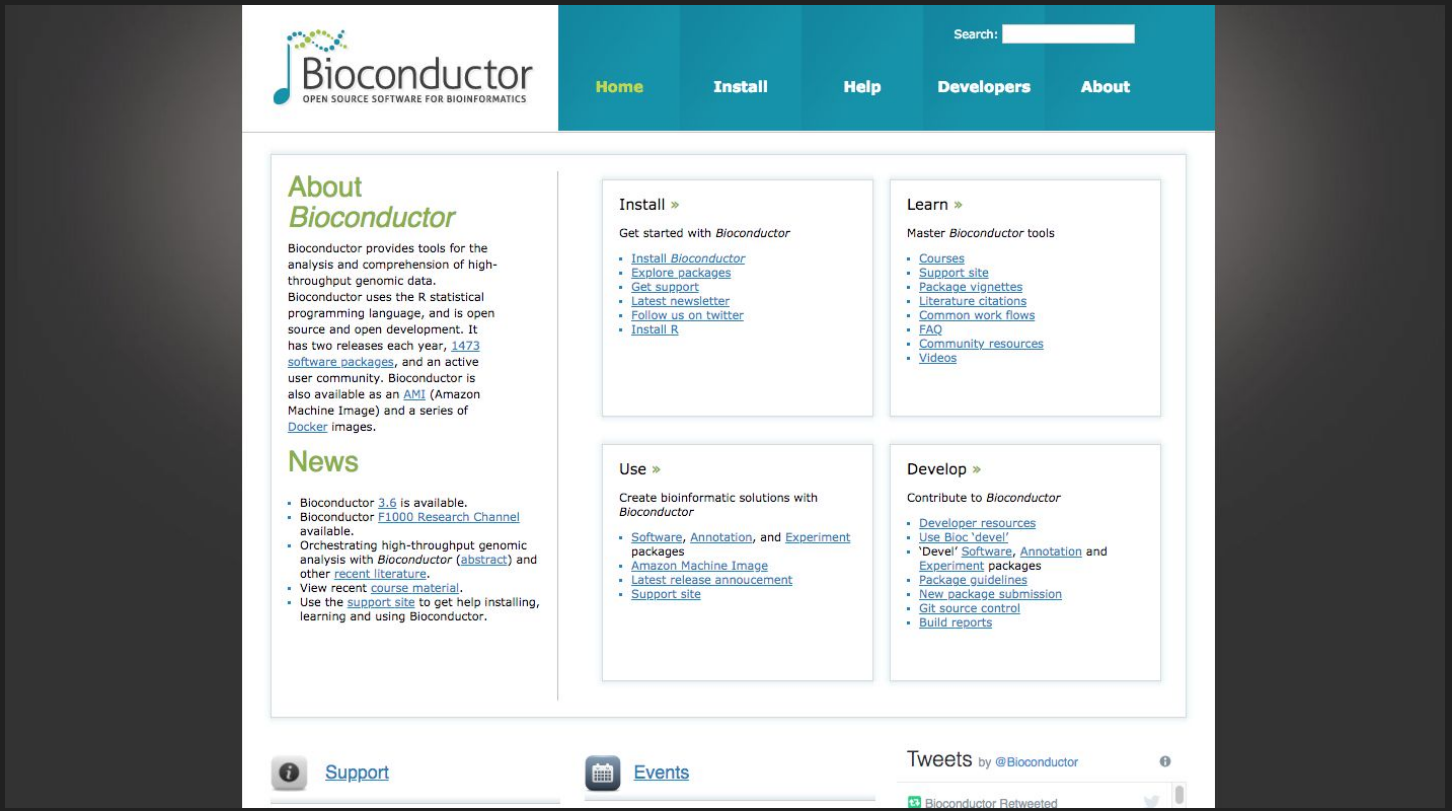
software

encodes

knowledge

Bioconductor is a large, *NIH-funded* open source software *community* dedicated to the *analysis and comprehension* of high throughput *biological data*.

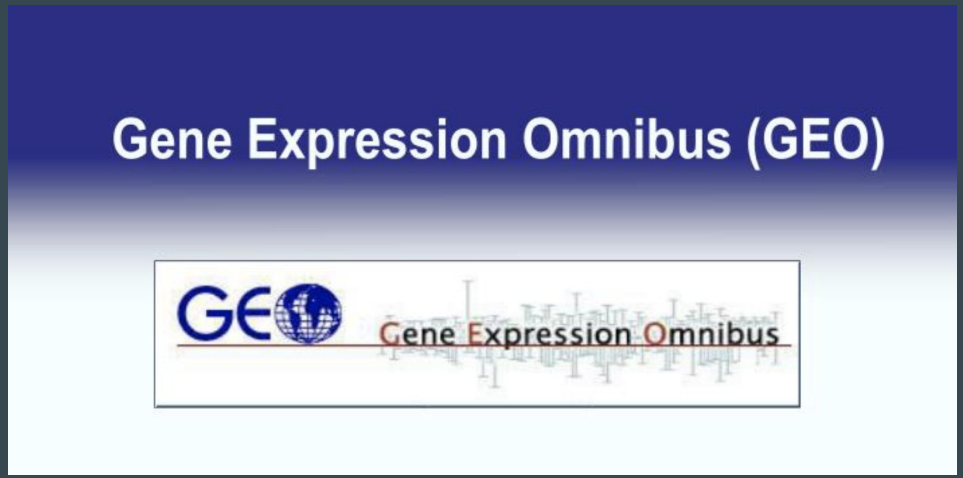
Bioconductor: Education, Training, and Community

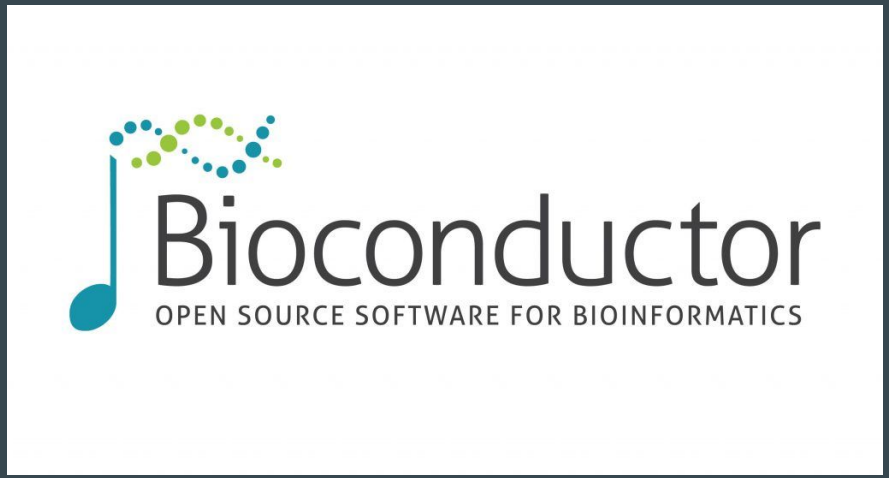


Core infrastructure

Community contribution

Building Bridges

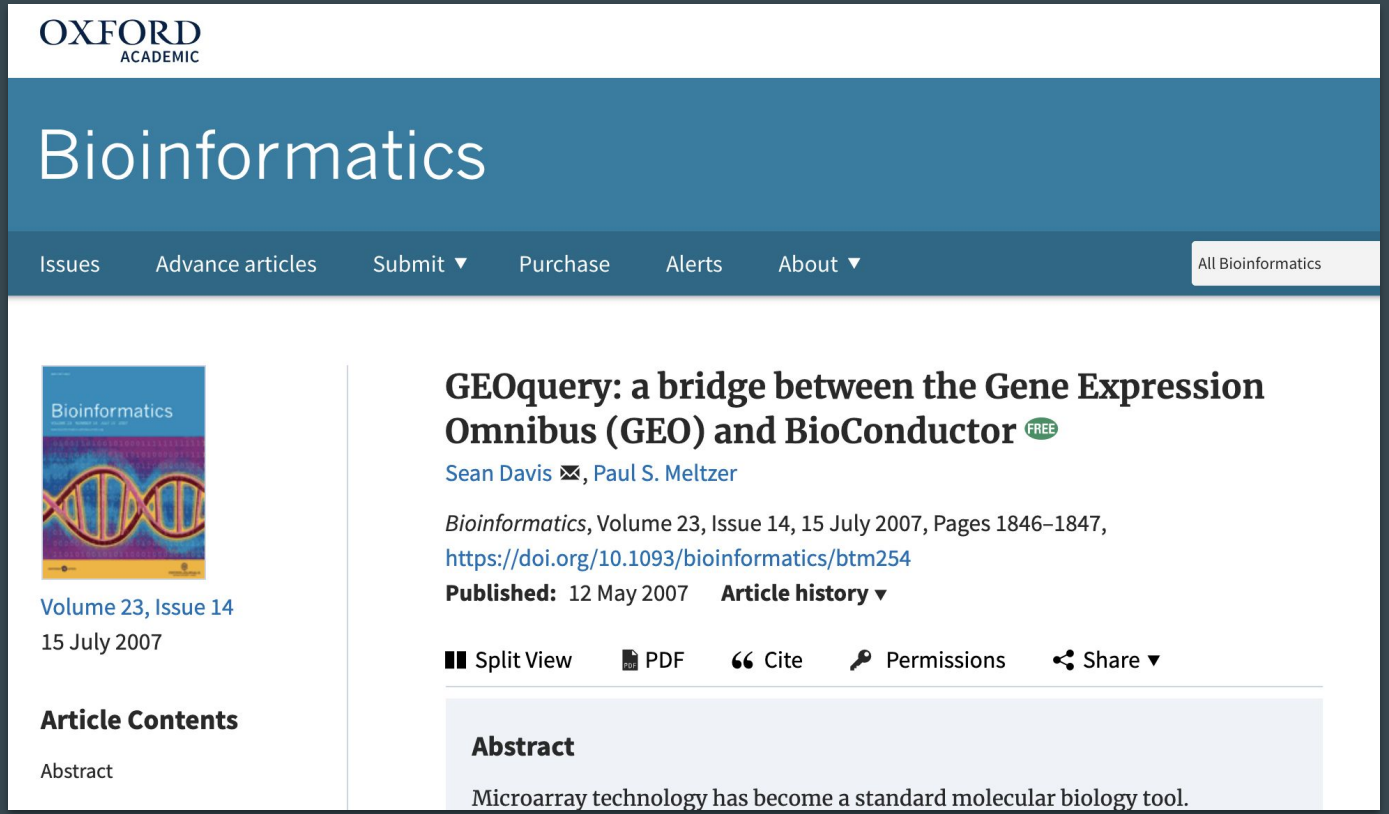




112,837 datasets, 3,031,406 samples publicly accessible

~300,000 users, 1000 active developers, >1700 interop software tools and analysis packages

Building bridges

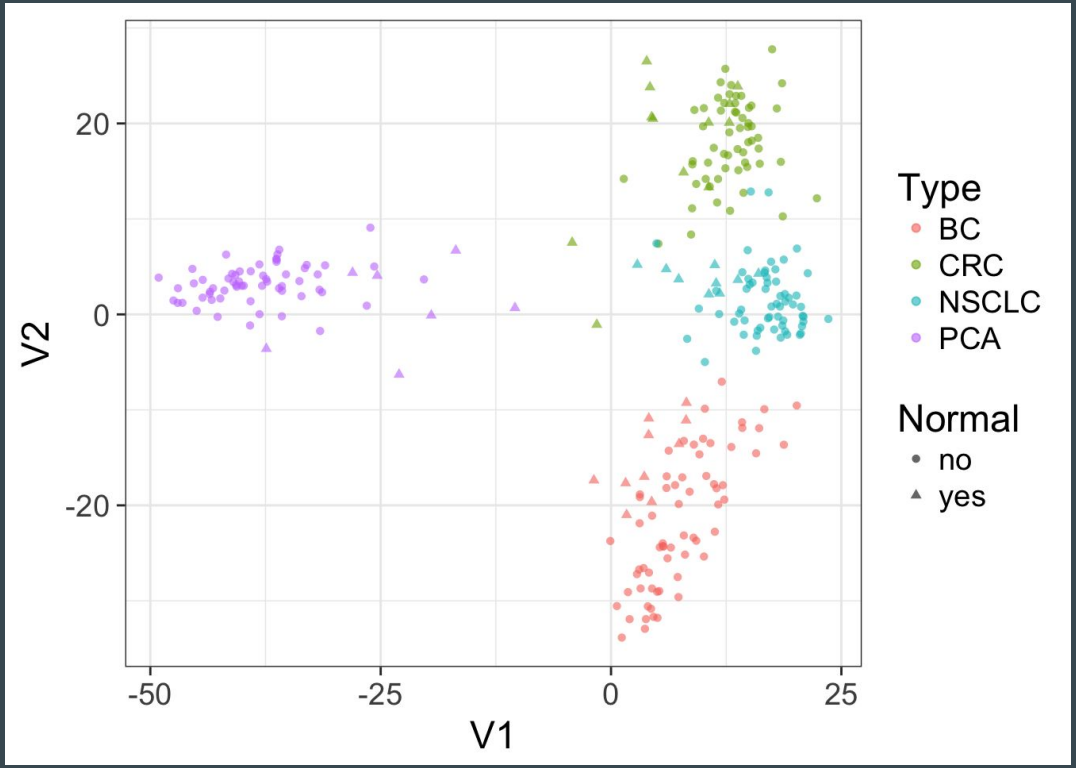


From this....

GSM48681.txt — Edited

```
#calRatio = Calibrated ratio, (or normalized ratio) [float]
#rQuality = Ratio measurement quality (including all intensity measurement
quality, plus the signal-to-noise-ratio requirement). [float]
#confLevel = Confidence Level, typically set at 99% [float]
#lowerLimit = Lower Limit of confidence interval for cal. Ratio [float]
#upperLimit = Upper Limit of confidence interval for cal. Ratio [float]
#M = Magnitude parameter for channel compensation [float]
#CV = Coefficient of variation of the intensity [float]
#Flag = User flagged. 0 - deleted spot by user, 1 - good spot [Integer]
#PRE_VALUE =
#UNF_VALUE = log ratio (log2 of PRE_VALUE)
!sample_table_begin
ID_REF  VALUE   SR_T. Int.  SR_Mean SR_S.Dev    SR_SNR  SR_iQuality SR_bkMean
SR_bkDev    SG_T. Int.  SG_Mean SG_S.Dev    SG_SNR  SG_iQuality SG_bkMean
SG_bkDev    unionArea   propArea    calRatio    rQuality    confLevel   lowerLimit
    upperLimit M    CV  Flag    PRE_VALUE   UNF_VALUE
1   0.2701103   63618   569.2   215.5   34.9    1.0000  380.4   11.9    60066
496.9   202.6   19.8    1.0000  399.6   18.4    67  0.8590  1.2059  1.0 99.0
0.1717  5.8245  0.9498  0.9819  1   1.2059  0.2701103
2   6.3459203   7195    41.6    10.6    2.6 0.2729  381.6   11.6    6754    0.5
20.3    0.0 0.2729  396.8   17.5    17  0.2179  81.3415 0.0 99.0    0.1717  5.8245
```

To this...



http://bit.ly/2W3I8dK

Filter gene expression by variance to find most informative genes.

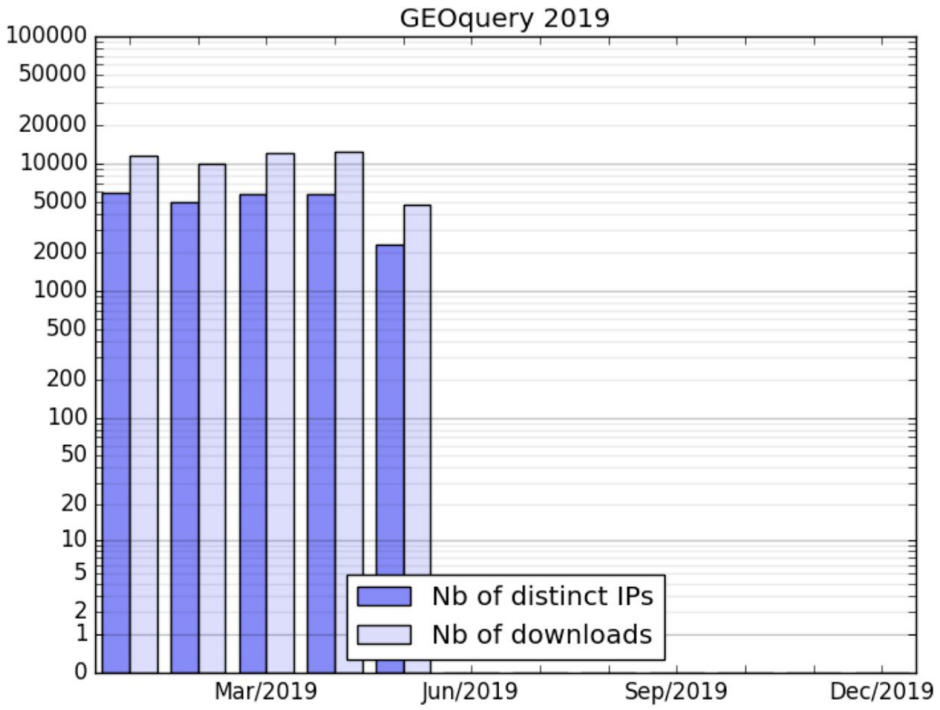
```
sds = apply(exprs(gse),1,sd)
dat = exprs(gse)[order(sds,decreasing = TRUE)[1:500],]
```

Perform multidimensional scaling and prepare for plotting.

```
mdsvals = cmdscale(dist(t(dat)))
mdsvals = as.data.frame(mdsvals)
mdsvals$Type=factor(pData(gse)[,'cancer type:ch1'])
mdsvals$Normal = factor(pData(gse)[,'normal:ch1'])
```

And do the plot.

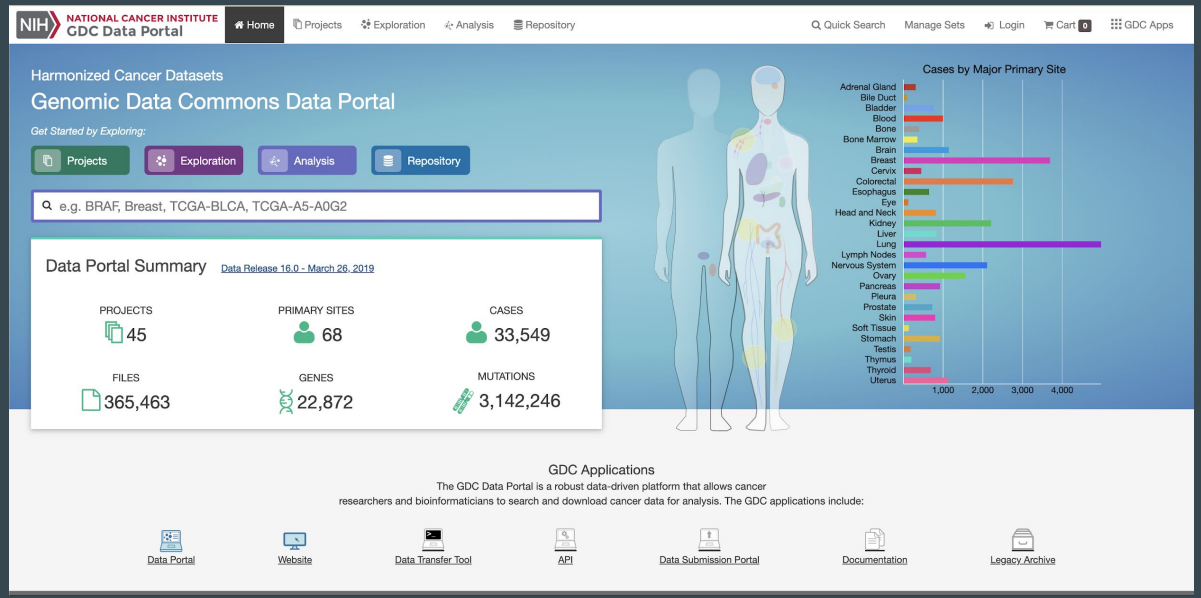
```
library(ggplot2)
ggplot(mdsvals, aes(x=V1,y=V2,shape=Normal,color=Type)) +
    geom_point(size=4, alpha=0.6) + theme(text=element_text(size = 18))
```

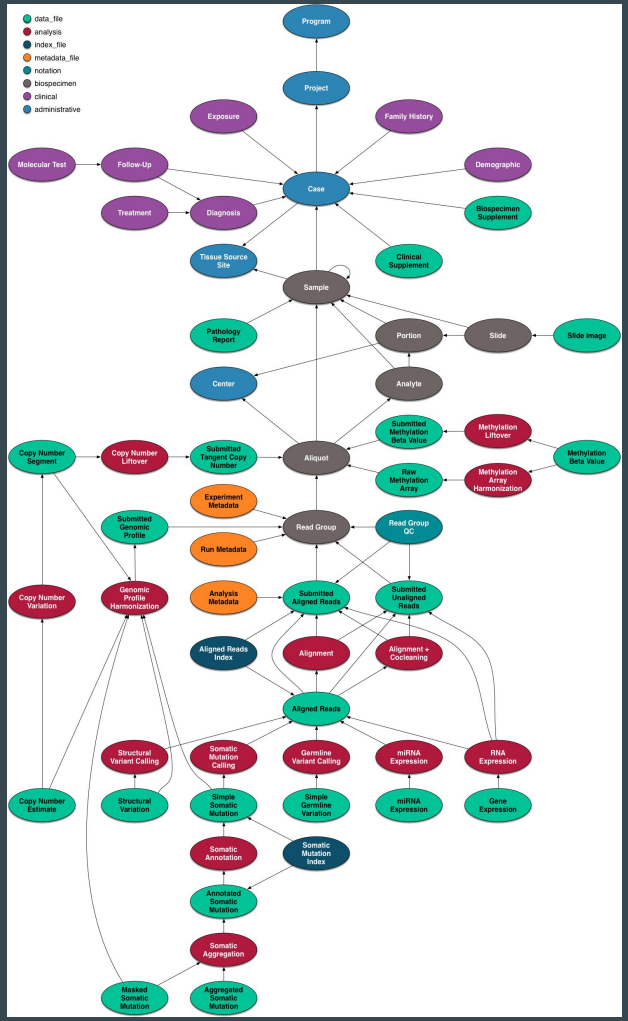


Month	Nb of distinct IPs	Nb of downloads
Jan/2019	5874	11654
Feb/2019	4982	9962
Mar/2019	5741	12199
Apr/2019	5766	12502
May/2019	2323	4725
Jun/2019	0	0
Jul/2019	0	0
Aug/2019	0	0
Sep/2019	0	0
Oct/2019	0	0
Nov/2019	0	0
Dec/2019	0	0
2019	**20937**	**51042**

GEOquery_2019_stats.tab

Understand, expose, and adapt to *your* data ecosystem and community





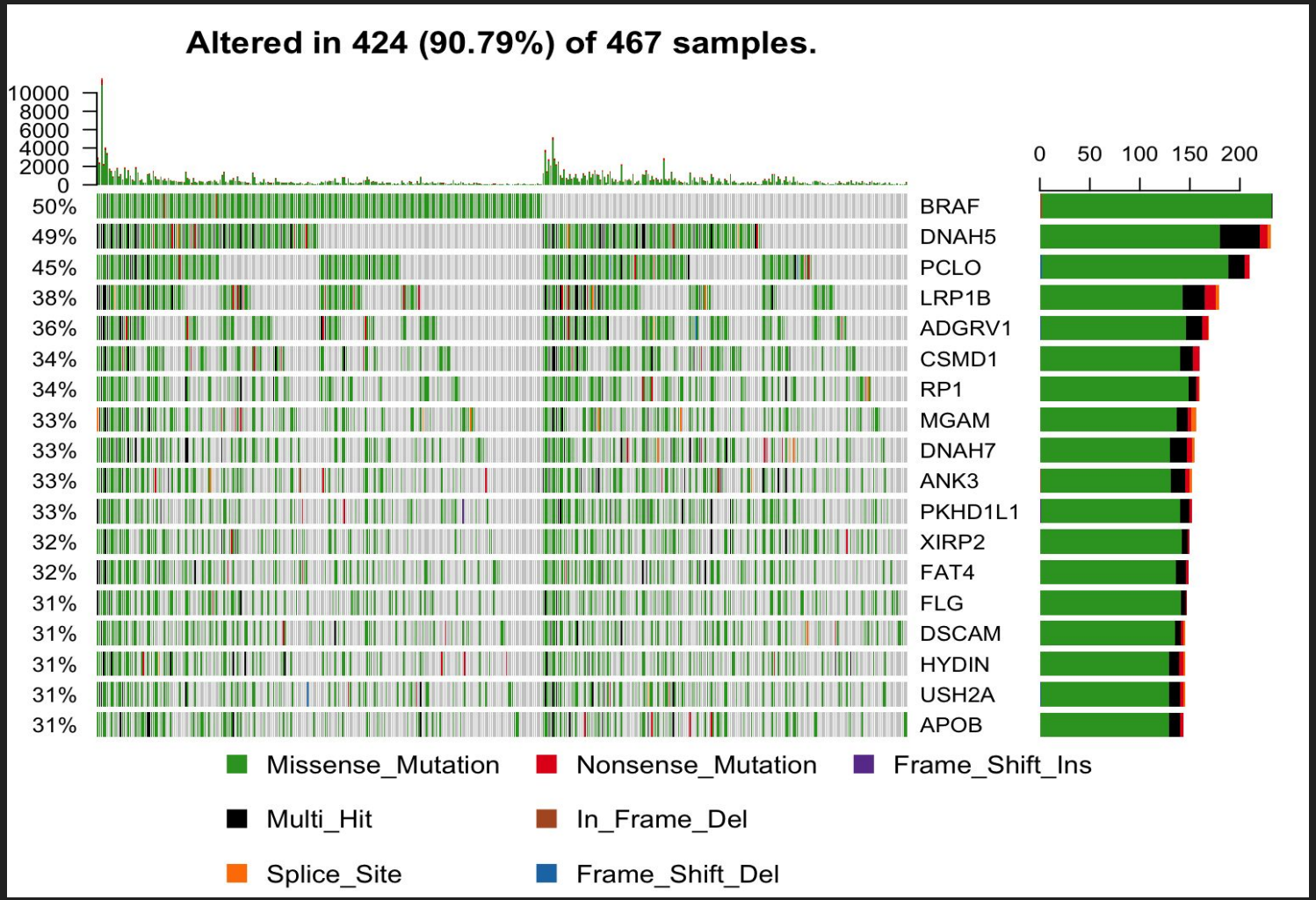
https://gdc.cancer.gov/files/public/image/Data_Model_Oobleck.png

Use the GenomicDataCommons package to find and download variants from the TCGA cutaneous melanoma dataset.

```
library(GenomicDataCommons)
fnames = files() %>%
    GenomicDataCommons::filter(~ cases.project.project_id=='TCGA-SKCM' &
        data_type=='Masked Somatic Mutation' &
        data_format=='MAF' &
        analysis.workflow_type=='MuTect2 Variant Aggregation and Masking') %>%
    ids() %>%
    gdcdata()
```

And now take those data directly to maftools for analysis and visualization.

```
library(maftools)
melanoma = read.maf(maf = fnames[1])
```

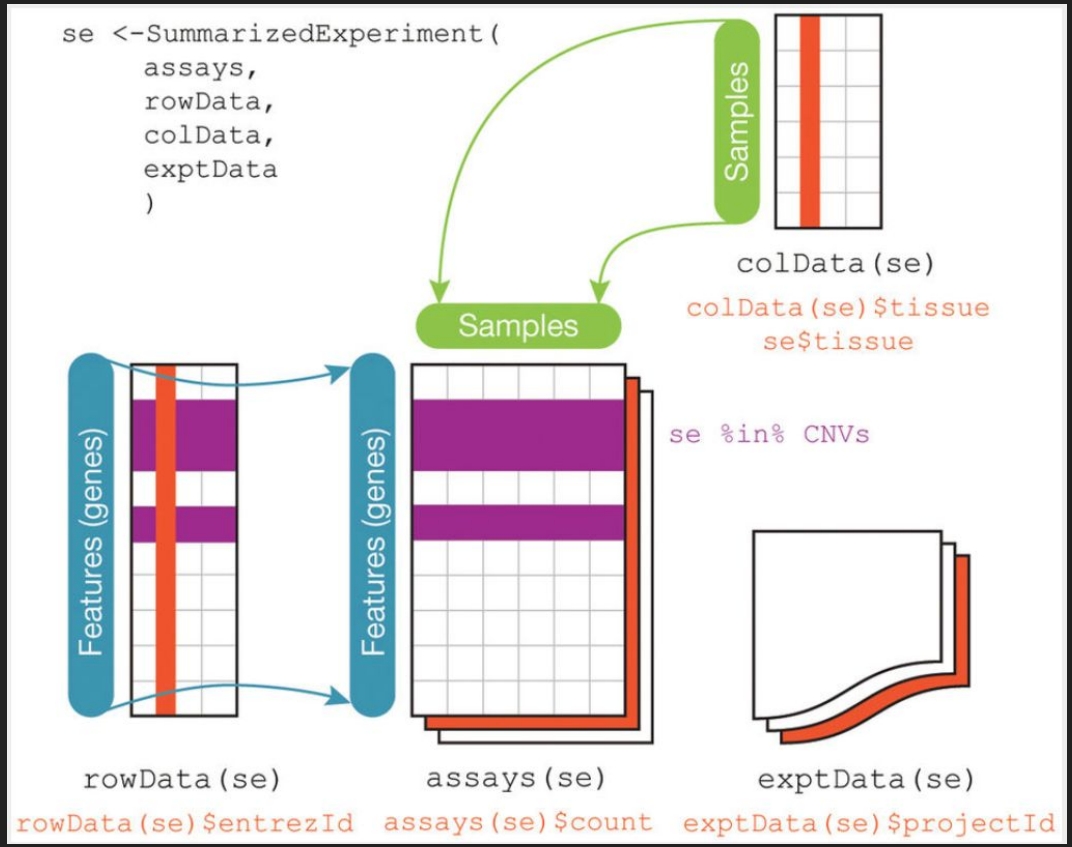

Altered in 424 (90.79%) of 467 samples.
10000
8000
6000
4000
2000
0
0
50
100
150
200
50%
49%
45%
38%
36%
34%
34%
33%
33%
33%
33%
32%
32%
31%
31%
31%
31%
31%
BRAF
DNAH5
PCLO
LRP1B
ADGRV1
CSMD1
RP1
MGAM
DNAH7
ANK3
PKHD1L1
XIRP2
FAT4
FLG
DSCAM
HYDIN
USH2A
APOB
Missense_Mutation
Nonsense_Mutation
Frame_Shift_Ins
Multi_Hit
In_Frame_Del
Splice_Site
Frame_Shift_Del

Baby steps: Reusable components

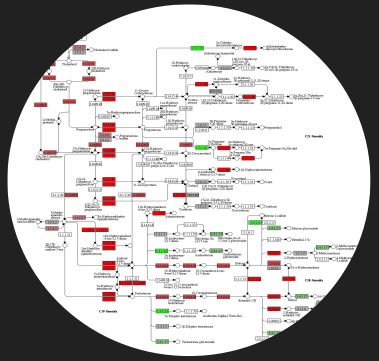
...

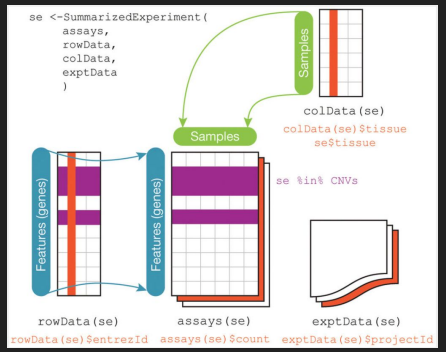
Where necessary, build new software or tools

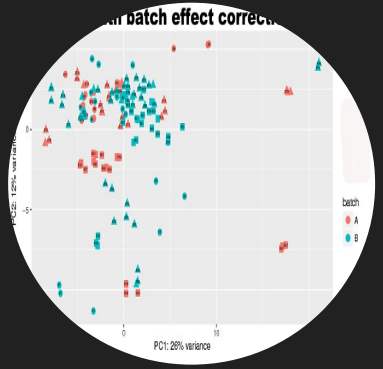
- Recognize complexity in high-throughput biological data
- Version control everything
- Continuous testing and integration
- Text-based workflow (no GUI)
- Literate programming approaches and documentation
- Education on tooling
- Numerous mechanisms for FAIR data sharing

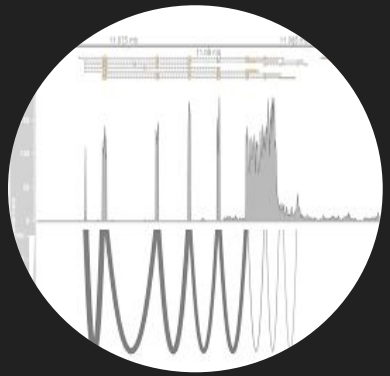


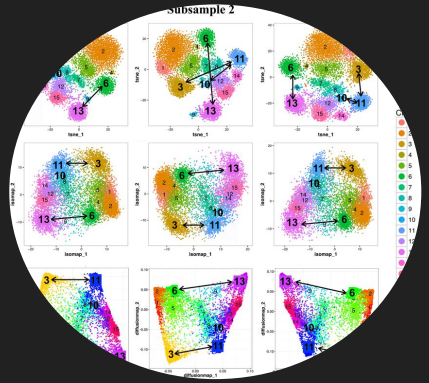
Core value: reuse and interoperability

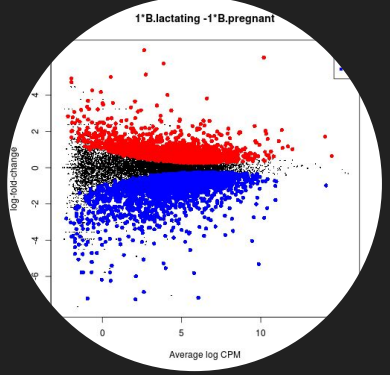






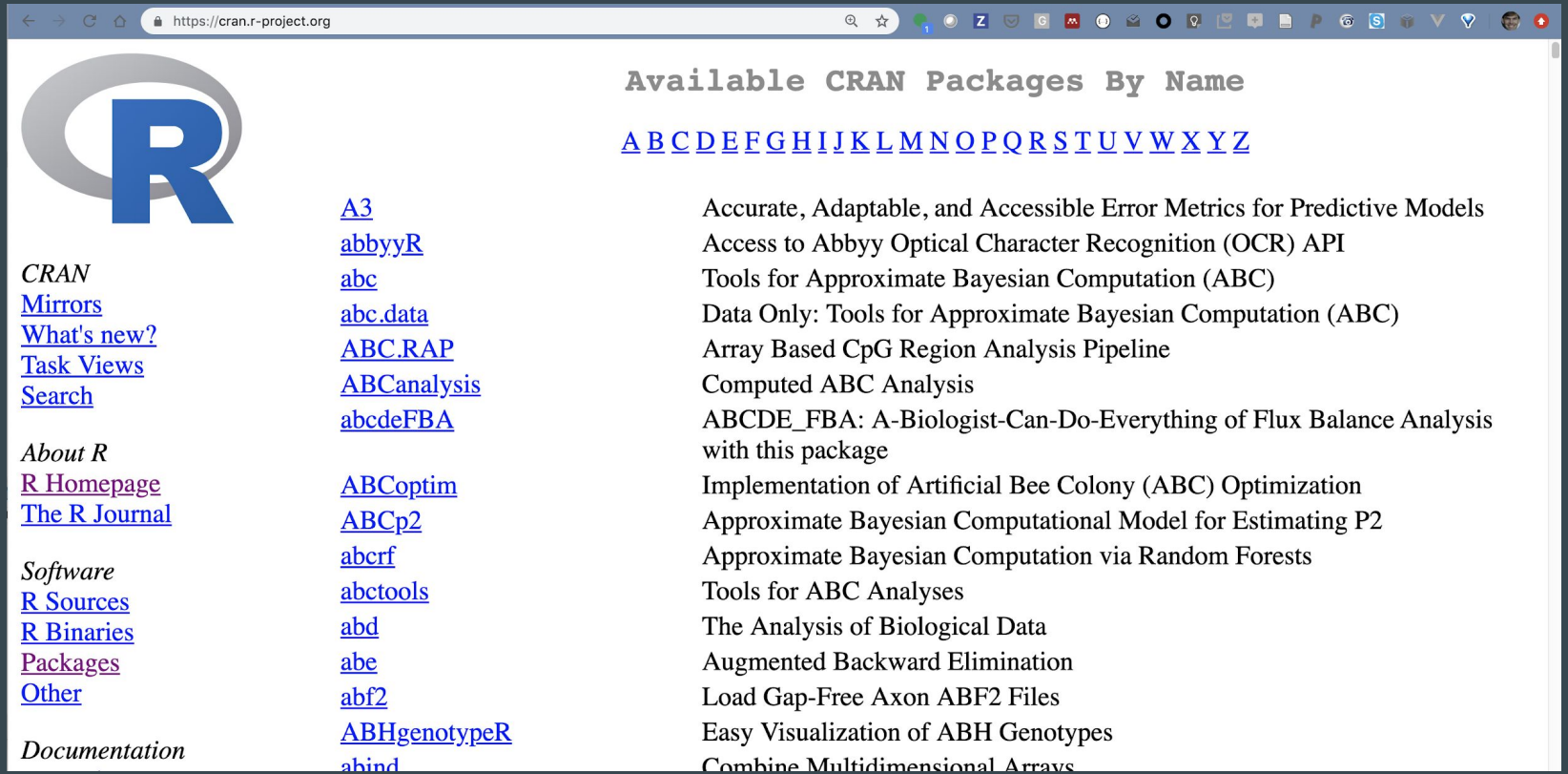






Where possible reuse existing tools, leverage existing communities

The R package: the unit of sharing in the R community



Published, versioned, documented, tested, measured

AnnotationHub

platforms all | rank 52 / 1741 | posts 7 / 0.7 / 3 / 2 | in Bioc 6 years

build warnings | updated before release

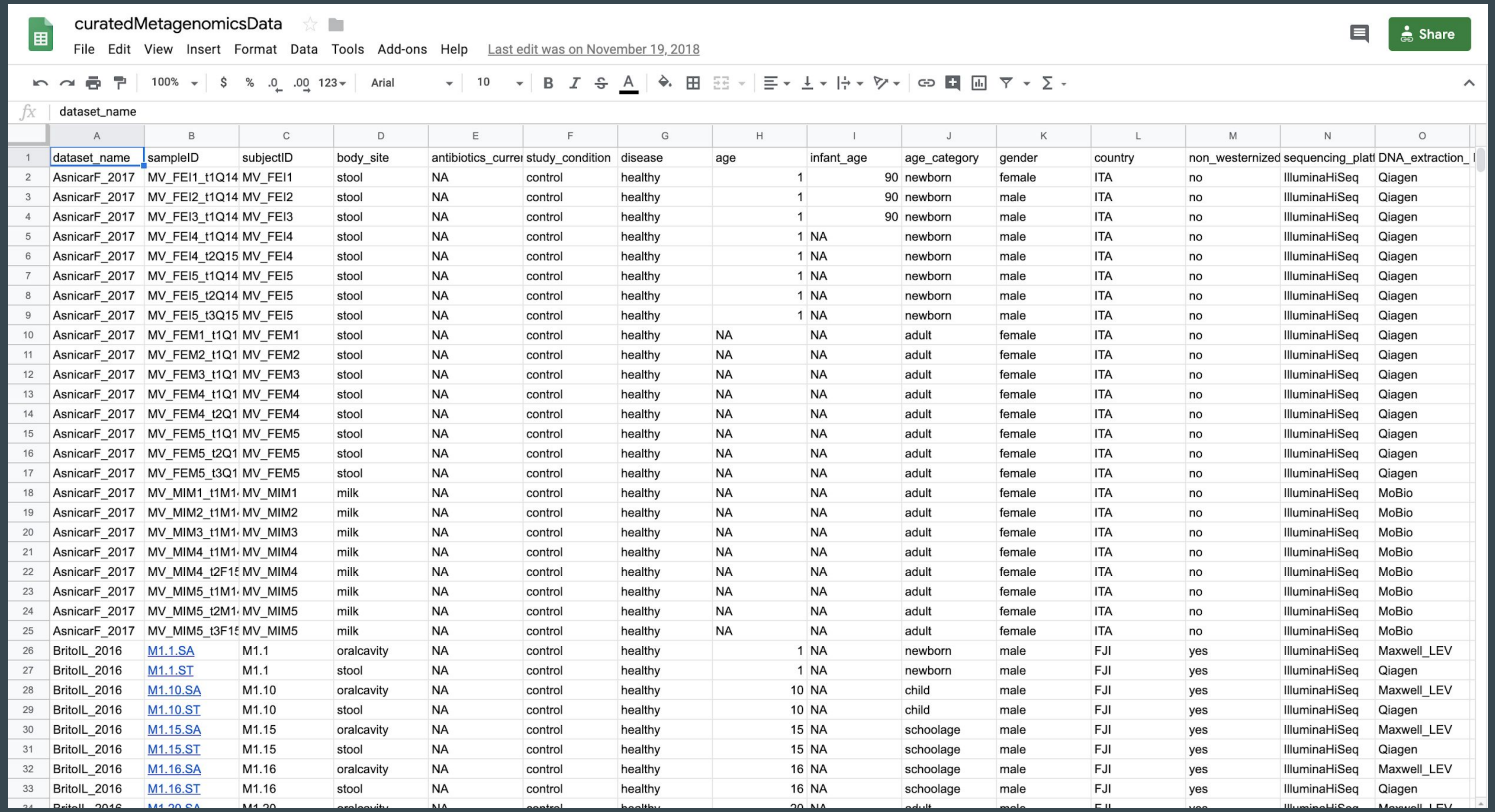
DOI: 10.18129/B9.bioc.AnnotationHub

Client to access AnnotationHub resources

Bioconductor version: Release (3.9)

This package provides a client for the Bioconductor AnnotationHub web resource. The AnnotationHub web resource provides a central location where genomic files (e.g., VCF, bed, wig) and other resources from standard locations (e.g., UCSC, Ensembl) can be discovered. The resource includes metadata about each resource, e.g., a textual description, tags, and date of modification. The client creates and manages a local cache of files retrieved by the user, helping with quick and reproducible access.

Maximize the impact (and reward) of curation efforts



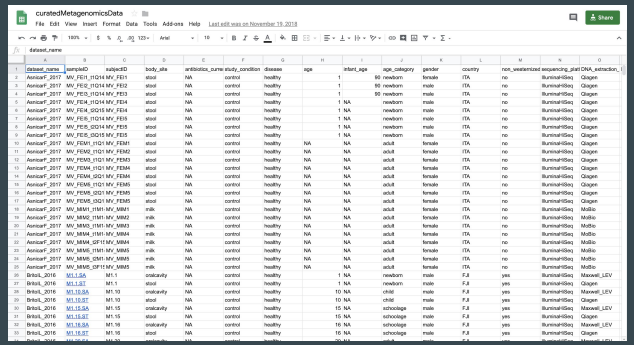
curatedMetagenomicsData

File Edit View Insert Format Data Tools Add-ons Help Last edit was on November 19, 2018

Share

dataset_name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	dataset_name	sampleID	subjectID	body_site	antibiotics_curre	study_condition	disease	age	infant_age	age_category	gender	country	non_westernized	sequencing_platf	DNA_extraction_
2	AsnicarF_2017	MV_FEI1_t1Q14	MV_FEI1	stool	NA	control	healthy	1	90	newborn	female	ITA	no	IlluminaHiSeq	Qiagen
3	AsnicarF_2017	MV_FEI2_t1Q14	MV_FEI2	stool	NA	control	healthy	1	90	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
4	AsnicarF_2017	MV_FEI3_t1Q14	MV_FEI3	stool	NA	control	healthy	1	90	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
5	AsnicarF_2017	MV_FEI4_t1Q14	MV_FEI4	stool	NA	control	healthy	1	NA	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
6	AsnicarF_2017	MV_FEI4_t2Q15	MV_FEI4	stool	NA	control	healthy	1	NA	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
7	AsnicarF_2017	MV_FEI5_t1Q14	MV_FEI5	stool	NA	control	healthy	1	NA	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
8	AsnicarF_2017	MV_FEI5_t2Q14	MV_FEI5	stool	NA	control	healthy	1	NA	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
9	AsnicarF_2017	MV_FEI5_t3Q15	MV_FEI5	stool	NA	control	healthy	1	NA	newborn	male	ITA	no	IlluminaHiSeq	Qiagen
10	AsnicarF_2017	MV_FEM1_t1Q1	MV_FEM1	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
11	AsnicarF_2017	MV_FEM2_t1Q1	MV_FEM2	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
12	AsnicarF_2017	MV_FEM3_t1Q1	MV_FEM3	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
13	AsnicarF_2017	MV_FEM4_t1Q1	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
14	AsnicarF_2017	MV_FEM4_t2Q1	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
15	AsnicarF_2017	MV_FEM5_t1Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
16	AsnicarF_2017	MV_FEM5_t2Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
17	AsnicarF_2017	MV_FEM5_t3Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
18	AsnicarF_2017	MV_MIM1_t1M1	MV_MIM1	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
19	AsnicarF_2017	MV_MIM2_t1M1	MV_MIM2	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
20	AsnicarF_2017	MV_MIM3_t1M1	MV_MIM3	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
21	AsnicarF_2017	MV_MIM4_t1M1	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
22	AsnicarF_2017	MV_MIM4_t2F1	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
23	AsnicarF_2017	MV_MIM5_t1M1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
24	AsnicarF_2017	MV_MIM5_t2M1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
25	AsnicarF_2017	MV_MIM5_t3F1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
26	BritoIL_2016	M1.1.SA	M1.1	oralcavity	NA	control	healthy	1	NA	newborn	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
27	BritoIL_2016	M1.1.ST	M1.1	stool	NA	control	healthy	1	NA	newborn	male	FJI	yes	IlluminaHiSeq	Qiagen
28	BritoIL_2016	M1.10.SA	M1.10	oralcavity	NA	control	healthy	10	NA	child	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
29	BritoIL_2016	M1.10.ST	M1.10	stool	NA	control	healthy	10	NA	child	male	FJI	yes	IlluminaHiSeq	Qiagen
30	BritoIL_2016	M1.15.SA	M1.15	oralcavity	NA	control	healthy	15	NA	schoolage	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
31	BritoIL_2016	M1.15.ST	M1.15	stool	NA	control	healthy	15	NA	schoolage	male	FJI	yes	IlluminaHiSeq	Qiagen
32	BritoIL_2016	M1.16.SA	M1.16	oralcavity	NA	control	healthy	16	NA	schoolage	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
33	BritoIL_2016	M1.16.ST	M1.16	stool	NA	control	healthy	16	NA	schoolage	male	FJI	yes	IlluminaHiSeq	Qiagen


curatedMetagenomicsData

File Edit View Insert Format Data Tools Add-ons Help Last edit was on November 19, 2018

Share

dataset_name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	dataset_name	sampleID	subjectID	body_site	antibiotics_curre	study_condition	disease	age	infant_age	age_category	gender	country	non_westernized	sequencing_platf	DNA_extraction_
2	AsnicarF_2017	MV_FE01_1Y0	MV_FE01	stool	NA	control	healthy		1	90 newborn	female	ITA	no	IlluminaHiSeq	Qiagen
3	AsnicarF_2017	MV_FE02_1Y0	MV_FE02	stool	NA	control	healthy		1	90 newborn	male	ITA	no	IlluminaHiSeq	Qiagen
4	AsnicarF_2017	MV_FE03_1Y0	MV_FE03	stool	NA	control	healthy		1	90 newborn	male	ITA	no	IlluminaHiSeq	Qiagen
5	AsnicarF_2017	MV_FE04_1Y0	MV_FE04	stool	NA	control	healthy		1	NA newborn	male	ITA	no	IlluminaHiSeq	Qiagen
6	AsnicarF_2017	MV_FE04_QG1	MV_FE04	stool	NA	control	healthy		1	NA newborn	male	ITA	no	IlluminaHiSeq	Qiagen
7	AsnicarF_2017	MV_FE05_1Y0	MV_FE05	stool	NA	control	healthy		1	NA newborn	male	ITA	no	IlluminaHiSeq	Qiagen
8	AsnicarF_2017	MV_FE05_QG1	MV_FE05	stool	NA	control	healthy		1	NA newborn	male	ITA	no	IlluminaHiSeq	Qiagen
9	AsnicarF_2017	MV_FE05_QG1	MV_FE05	stool	NA	control	healthy		1	NA newborn	male	ITA	no	IlluminaHiSeq	Qiagen
10	AsnicarF_2017	MV_FEM1_1Y0	MV_FEM1	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
11	AsnicarF_2017	MV_FEM2_1Y0	MV_FEM2	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
12	AsnicarF_2017	MV_FEM3_1Y0	MV_FEM3	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
13	AsnicarF_2017	MV_FEM4_1Y0	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
14	AsnicarF_2017	MV_FEM4_QG1	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
15	AsnicarF_2017	MV_FEM5_1Y0	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
16	AsnicarF_2017	MV_FEM5_QG1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
17	AsnicarF_2017	MV_FEM5_QG1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	Qiagen
18	AsnicarF_2017	MV_MIM1_1Y0	MV_MIM1	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
19	AsnicarF_2017	MV_MIM2_1Y0	MV_MIM2	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
20	AsnicarF_2017	MV_MIM3_1Y0	MV_MIM3	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
21	AsnicarF_2017	MV_MIM4_1Y0	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
22	AsnicarF_2017	MV_MIM4_GF1	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
23	AsnicarF_2017	MV_MIM5_1Y0	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
24	AsnicarF_2017	MV_MIM5_GR1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
25	AsnicarF_2017	MV_MIM5_GF1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	IlluminaHiSeq	MoBio
26	BritoIL_2016	M1.1.GA	M1.1	oralcavity	NA	control	healthy		1	NA newborn	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
27	BritoIL_2016	M1.1.ST	M1.1	stool	NA	control	healthy		1	NA newborn	male	FJI	yes	IlluminaHiSeq	Qiagen
28	BritoIL_2016	M1.10.SA	M1.10	oralcavity	NA	control	healthy		10	NA child	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
29	BritoIL_2016	M1.10.ST	M1.10	stool	NA	control	healthy		10	NA child	male	FJI	yes	IlluminaHiSeq	Qiagen
30	BritoIL_2016	M1.15.SA	M1.15	oralcavity	NA	control	healthy		15	NA schoolage	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
31	BritoIL_2016	M1.15.ST	M1.15	stool	NA	control	healthy		15	NA schoolage	male	FJI	yes	IlluminaHiSeq	Qiagen
32	BritoIL_2016	M1.16.SA	M1.16	oralcavity	NA	control	healthy		16	NA schoolage	male	FJI	yes	IlluminaHiSeq	Maxwell_LEV
33	BritoIL_2016	M1.16.ST	M1.16	stool	NA	control	healthy		16	NA schoolage	male	FJI	yes	IlluminaHiSeq	Qiagen

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Correspondence | Published: 31 October 2017

Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata & Levi Waldron

Nature Methods **14**, 1023–1024 (2017) | Download Citation

Publish the how....

...

New Results

Comment on this paper

recount-brain: a curated repository of human brain RNA-seq datasets metadata

Ashkaun Razmara, Shannon E. Ellis, Dustin J. Sokolowski, Sean Davis, Michael D. Wilson, Jeffrey T. Leek, Andrew E. Jaffe, Leonardo Collado-Torres

doi: https://doi.org/10.1101/618025

F1000Research
Open for Science

Check for updates

METHOD ARTICLE

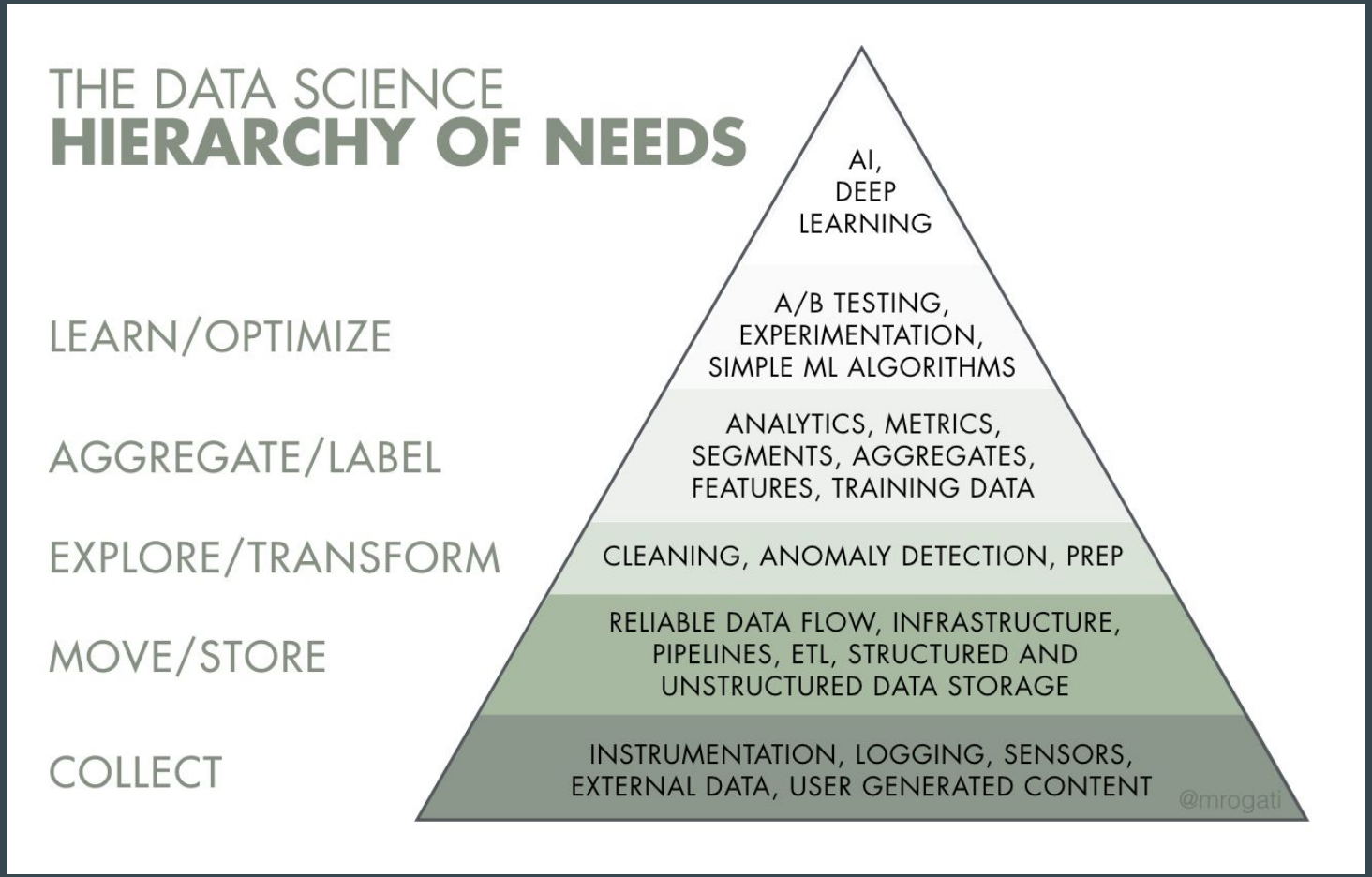
Orchestrating a community-developed computational workshop and accompanying training materials [version 1; peer review: 2 approved]

Sean Davis [1], Marcel Ramos[2,3], Lori Shepherd[3], Nitesh Turaga[3], Ludwig Geistlinger[2], Martin T. Morgan[3], Benjamin Haibe-Kains[4,5], Levi Waldron [2]

Author details

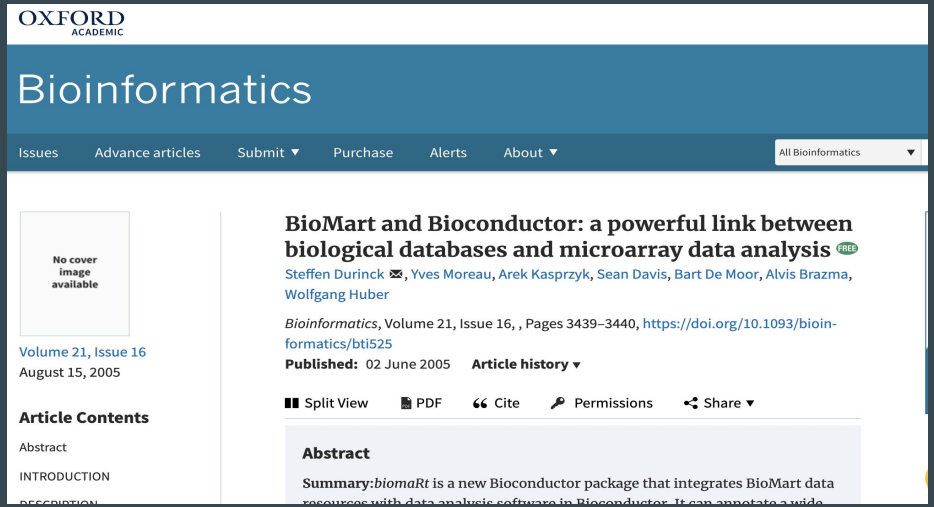
Addtional thoughts

...



Source: Monica Rogati https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Metadata are data



Stakeholders or Data Communities


GEO
Gene Expression Omnibus

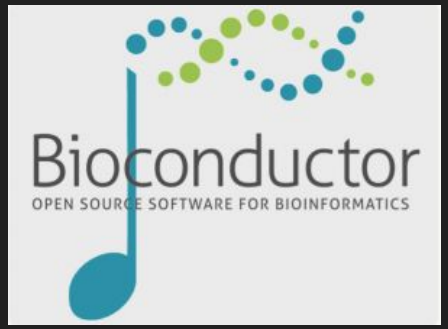

PHARMACODB
MINE MULTIPLE CANCER PHARMACOGENOMIC DATASETS


NCBI


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal

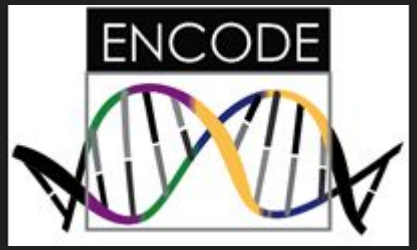

ICGC


DGIdb
THE DRUG GENE INTERACTION DATABASE


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS


e!Ensembl


db GaP
genotypes and phenotypes


ENCODE

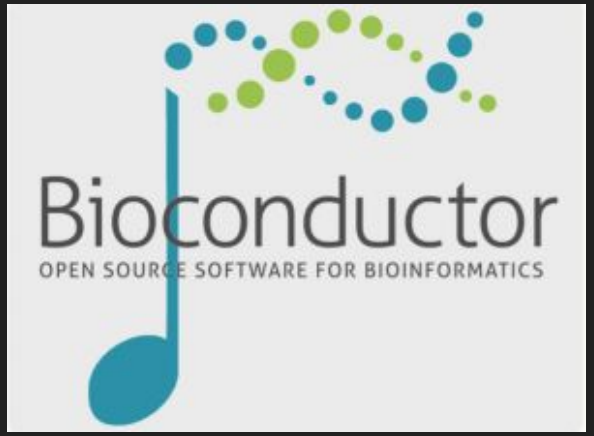

CIViC
CLINICAL INTERPRETATIONS OF
VARIANTS IN CANCER


THE CANCER
IMAGING ARCHIVE

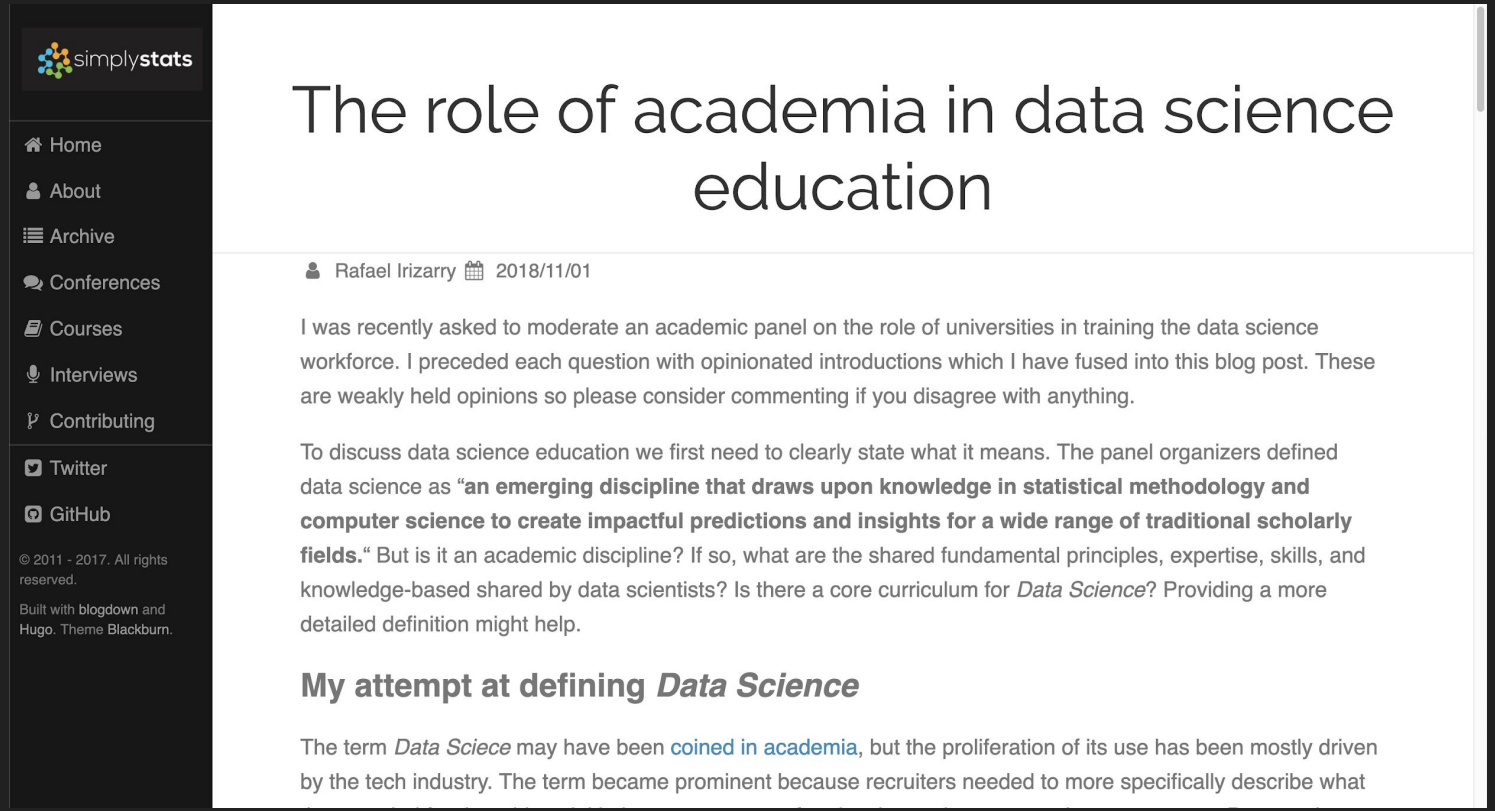

Early Detection Research Network
Biomarkers: the key to early detection


CCLE
Cancer Cell Line
Encyclopedia

Commodity big data technologies



Education: we are all learning and teaching….



We, not *they*

Technical Advisory Group

Vince Carey, Brigham & Women's, Harvard Medical School, USA.

Aedin Culhane, Dana-Farber Cancer Institute, Harvard School of Public Health, USA.

Sean Davis, National Cancer Institute, USA.

Robert Gentleman, Computational Biology, 23andMe, USA.

Kasper Daniel Hansen, Bloomberg School of Public Health, Johns Hopkins University.

Wolfgang Huber European Molecular Biology Laboratory, Heidelberg, Germany.

Rafael Irizarry Dana-Farber Cancer Institute, USA

Michael Lawrence, Genentech Research and Early Development, USA.

Levi Waldron, CUNY School of Public Health at Hunter College, New York, NY.

Questions?

https://bioconductor.org

https://seandavi.github.io