

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](https://twitter.com/seandavis12)

<http://bit.ly/SD-AIDR2019>



301 Redirect
?????

Data Engineering: a Definition

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 Plan¹ 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.”

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

- Optimize data storage and security
- Connect NIH data systems

Modernized Data Ecosystem

- Modernize data repository ecosystem
- Support storage and sharing of individual datasets
- Better integrate clinical and observational data into biomedical data science

Data Management, Analytics, and Tools

- Support useful, generalizable, and accessible tools and workflows
- Broaden utility of and access to specialized tools
- Improve discovery and cataloging resources

Workforce Development

- Enhance the NIH data-science workforce
- Expand the national research workforce
- Engage a broader community

Stewardship and Sustainability

- Develop policies for a FAIR data ecosystem
- Enhance stewardship

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

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens¹, Skylar Y. Lee¹, Faraz Faghri², Roy H. Campbell², Chengxiang Zhai³, Miles J. Efron⁴, Ravishankar Iyer¹, 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@csil.edu (MCS); sinhas@illinois.edu (SS); generobi@illinois.edu (GER)



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

Copyright: © 2015 Stephens et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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.

<u>Data Phase</u>	<u>Astronomy</u>	<u>Twitter</u>	<u>YouTube</u>	<u>Genomics</u>
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 Software Engineering

Guest editors:

Article type: I

Authors: Alko

Affiliations: [a]

Mathematics,

Process and C

Winona State

Winona, MN, U

Disconnected and Distributed

UAE | [b]
Engineering
ment,
e University,

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

Jeffrey T. Leek^{a,1} and Roger D. Peng^b

^aAssociate Professor of Biostatistics and Oncology and ^bAssociate 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

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.

aBase a2Classif a4Core a4Preproc a4Reporting ABAEnrichment ABaray ABSSeq acde aCGH ACME ADaCGH adSplit affxparser affy affycomp AffyCompatible affyContam affcoretools AfflyExpress affyILM affyio affymGUI affyPara affypdnn affyPLM affyQCReport
fTyRNA Degradation AGDEX agilip AgiMicroRna AIMS ALDex2 Allelicimbalance alpine alsae altcdtfcns AMOUNTAIN ampican amplico Queso AnamiR Anaquin AneuFinder ANF annaffy annmap annotate AnnotationDbi AnnotationFilter AnnotationForge
AnnotationFuncs AnnotationHubData AnnotationTools annotat anotata anotat2seq anntProfiles apComplex apeglm aprima ArrayExpress ArrayExpressHTS arrayMvout arrayQuality arrayQualityMetrics ArrayTools ArrayTV ARRMNormalization ASAFE ASEB ASGS
ASpli ASSET ASSIGN ATACseqQC attract AUCell BaalChIP BAC bacon BADER BadRegionFinder BAGS ballgown bamsignals banocc basecallQC BaseSpaceR Basic4Cseq BASICS BasicSTARRseq BatchQC BayesKnockdown BayesPeak baySeq BBCAnalyzer BCRANK
beachmat beadarray beadarraySNP BeadDataPackR BEARSeC BEAT BEclear bgafun BgeeBD BGmix bxg BHC BiCARE BiFET BiGGR bigmelnom bigmemory Extras bioassayR Biobase biobroom bioCanCer BioCaseStudies Biocheck BioFileCache BioGenerics biocProcr
biocInstaller BioCor BiocParallel BiocSkeLearn BiocStyle biocViews BiocWorkflowTools biotDist biomaRt biomformatr BioMVCClass biomvCNBS BioNet BioQC BioSeqClass biosigner Biostlings biosvd biotmlc biotvzBase BiReWire biarta birte BioSeq BitSeq BitSeq2 blma bnbc BPFC
BRAIN BrainStars branchpointer bridge BridgeDBr BrowserViz BrowserVizDemo BSgenome bsseq BubbleTree BufferedMatrix BufferedMatrixMethods BUMHMM bumphunter BUS CAFE CAGER CALIB CAMERA cancerR cancerclass CancerInSilico CancerMutationAnalysis
cancerSubtypes CanD caOmicsV Cardinal casper CATALYST Category categoryCompare CausalR cbaf cemap CCPROMISE ccrepe cellbaseR cellGrowth cellHTS2 cellity CellMapper CellNOptR cellscape cellTree CEMITool CexoR CFAssay CGEN CGHbase CGHcall cghl
CGHnormalizer CGHregions CHAMP CHARGE charm ChemmineOB ChemmineR Chicago chimera chimeraviz ChIPanalyze ChIPcanon ChIPComp chipenrich ChIPQC ChIPAnno ChIPQC ChIPseeker chipseq ChIPseqR ChIPSeqSpike ChIPsim ChIPXpress chopsticks chrom
chromDraw ChromHeatMap chromPlot chromstatR chromswitch chromVAR CHRONOS CNldex cisPath ClassifyR cleanUpTSeq cleaver clipdda clipper Clonal Clonality clontyper Rclst clustils clustComp clusterExperiment ClusterJudge clusterProfiler clusterSeq
clusterSignificance clusterStar CMA cn CNAnorm CNEr CNORdt CNORfeeder CNORfuzzy CNORode CNPBayes CNTools cnvGSA CNVPanalyzer CNVrd2 CNVtools cobindR CoCiteStats codellink coxnet CoGPars cogena coGPS COHCAP coMET COMPASS
commodore compEpiTools CompGo ComplexHeatmap CONFESS ConsensusClusterPlus consensusOV consensusSeeker contIBAIT conueur convert copya copynumber Copywriter CoRegNet CODEX corNet CORREP coseq cosmic COSnet CountCloud covEB
coverageView covRNA cpvSNP cqcn CRImage CRISPRseek crisprseekplus CrisprVariants crimm crossmeta CSAR csaw CSSP cst CTDebugger ctsGe cummeRbund customProd CBVE cycle cydar cytokit cytobil CytoML dada2 dagLogo dAma DAmIRseq DAPAR DART DA
DBChIP dcGSA DchipRep ddCT debrowser DECIPHER DEComplexDisease DeconRNaseq DEDS DeepBlueR deepSNV DEFformats DEGraph DEGreport DEGsseq DelayedArray DelayedMatrixStats deltaGseq DeMAND DEP derfinder definderHelper definderPlot DESeq
DESeq2 destiny DESubs DESeqSeq dexuss DFP DiffBind diffGeneAnalysis diffHiCl DiffLogo difflo diffuStats diggit Director DirichletMultinomial discordant dks DMCHMM DMRcaller DMRcate DMRforPairs DMRScan DNABarcodes DNACopy DNashaper doppelgangr DOQT
doscheda DOSe drawProteins DRIMSeq DriverNet DropletUtils DrugVsDisease dSimer DSS DTA dualKS DupChecker dupRadAr dyebias DynDom EasyqcR easyRNAseq EBarrays EBoxexpress EBMage EBSEA EBSeq EBSeqHMM ecolitk EDASeq EDDA edge edgeR ege
EGAD EGSEA eiR eisa ELBOW ELMER EMDomics EmpiricalBrowsnMethod ENCODEExplorer ENmix EnrichedHeatmap EnrichmentBrowser ensembldb ensemblVEP ENVISIONQuery EpIDISH epigenomix epiNEM epivizr epivizrChart epivizrData epivizrServer epivizrStand
epochdashboard erma esATAC estVis eudydbiom EventPointer ExMiR exomeCopy exomePeP ExperimentHub Exp experimentHubData explore ExpressionAtlas ExpressionView fabia facpa facidDesign FamAgg farms fastLiquidAssociation fastest fCAC fCI fdrframe FEM fir
Genet fgsea FindMyFriends FIShtyser FiTHic flagma fliplow flowAI flowBeads flowBin flowcatchR flowCHIC flowCL flowClean flowClust flowFlow flowCyBar flowDensity flowFit flowFP flowMap flowMatch flowMerge flowPeaks flowPlots flowFlow flowQ flowOb
flowRepositoryR FlowSOM flowStats flowTime flowTrans flowType flowUtils flowViz flowVS flowWorkspace fmcsR focalCall FourCseq FRGEpistasis frma frmTools FunChIP FunciSNP funtooNorm GA4GHclient GA4GHshiny gaga gage gaggle gaia GAPrediction garfield ga
gcap gcastest gCMAP gCMAPWeb gCrispTools gcrma GDCRNATools gdsfmt geecc GEM genArise genbankR GeneAnswers geneAttribution GeneBreak geneClassifiers GeneExpressionSignature genefilter genufu GeneGA GeneGeneInteR GeneMeta GeneNetworkBuilder
geneOverlap geneplast genetplot geneRecommender geneRegionScan geneRxCluster GeneSelectMMD GeneSelector GENESIS geNetClassifier GeneticsDesign GeneticsPed geneXTender GENIE3 genoCN GenoGam genomation GenomeGraphs GenomInfoDb
genomelIntervals genomes GenomicAlignments GenomicDataCommons GenomicFeatures GenomicFS GenomicInteractions GenomicRanges GenomicScores GenomicTypes Genominator geneset genotypeeval ghenphen GenRank GenVisR GEOmetadb GOquery
GESubmission gep2pep gesper GEWIST GGBase ggbio ggcyto GGtools ggtree girafe GISPA GLAD Glimma GlobalAncova globalSeq globalltest grmrR GMRP GOexpress GOfuncR GOFunction GoogleGenomics GoPro goProfiles GOSemSim goseq GOSim goSTAG GO
GOsummaries GOWatch goTools gpIs gapre ggQLbase ggQLtStats graph GraphAligner GraphAT graphite GraphPAC GRENTIS GreyListChIP GRmetrics groHMM GRridge GSALighting GSAR GSQA GSEABase GSEAdm gsean HSRC GSVA gtrellis GUIDESge i
igswat gwascat GWASTools h5vc hapFlake Harman Harshlight HDF5Array HDT heatmaps Heatplus HelloRanges HELP HEM hiAnnotator HiBAG HiCompare hicrep hierGWAS HilbertCurve HilbertViz HiRes Processor HITC hmbdQuery HMMcopy hopack hp
htPCR HTSAnalyzer HTSeqGenie htSeqTools HTSfilter HybridMTest hyperdraw hypergraph iASeq iBBiG ibh iBMQ ICARE lcms iCheck iChip iClusterPlus iCOBRA ideal IdeoViz idogram IdMappingAnalysis IdMappingRetrieval IGC IIHW illuminaIo imageHTS IMAS Imeta
ImmuneSpaceR immunoChIP IMPCData ImpulseDE ImpulseDE2 impute InPAS InPower INSPECT Intansy InteractionSet interactiveDisplay InteractiveInterMine InterMineR IntramixREplorer InveRSion InOniseR iPAC IPO IPPD IRGenes IrisSpatialFeatures IQseq i
ioswitchSwitchAnalyzer IsoGeneGUI ISOLED isomiRs ITALICS iterativeBMA iterativeMBAsurv iterClust IVAS ivygapSE IWTools JASPAR2018 jda JunctionSeq karyoploter KcSmart kebabs KEGGgraph KEGGLines keggorthology KEGGprofile KEGGSTEST kimes lapmx L
block LEA LedPred les Ifa limma limmaGUI LINC LineagePulse Linnom LiquidAssociation lmdme LMGene LOBSTASH loci2path logicFS logitL Logolas lol LOLA LowMACA LPE LPeadJ IpNet Ipsymphony lumi LVSMiRNA LympHoSeq M3C M3D M3Drop maanova macat
macarrPlot made4 MADSEQ mafTools MAGeckFlute maigesPack MAIT makecdfenv MANOR manta MantelConr mAPLK mapPredictDSC mapscape marray masSigPro maskABand MassArray massIR massScan massSpecWavelet MAST matchBox MatrixIndex matter MaxContrastProject
MBAmethyl MBASED MBCB mBCPR MBtest mckaGUI MBciolut MCRestimate mCSEA mdgsa mdmc MEAL MeasurementError MEDIPS MEDME MEIGOR MergeMail Mergeoemics MeSHDbi meshes meshr messina metaArray Metab metaboxMr MetaboSignal metataCCA
metaCyto metagene metagenomeFeatures metagenomeSeq metahdp metaMS MetaNeighbor metaSeq metaseqR metavariz MetCirc methimpupe methInheritSim Meth Ped MethTargetedNGS methVisual methyAnalysis MethylAid methylInheritance methylKit MethylMix meth
methyNPipe methyNSearch methylumi methylvim mfa Mfuzz MGFm MGFR mgsa MiChip microbiome microRNA MicroMica mimager MIMOSA MineLCa minet minfi MinimumDistance MIPP MIRa MiRaGe miRBase converter miRcomp miIntegrator miLAB miRmine miRNANameCom
miRNAPath miRNApat miRysponge Mirsyngery missMethyl mitoODE MLInterfaces MLP MLSeq seq MMDiff2 MmPalateMiRNA MODA mogsa monoca moonlightR MoPS mosaics motifbreakR motifconvar MotifDb motifStack MotiV MPFE mptra mql mTL msa ml
MSGgui MSGFplus msmsEDA msmsTests MSnbase MSnId msPurity MSstats Mulcom MultiAssayExperiment multiClust MultiDataSet MultiMed multiMir multiOmicsViz multiscan multitest muscle MutationalPatterns MVCClass mvGST MWASTools mygene myvariant mzId
NAfinder NanoStringDiff NanoStringQCPro NarrowPeaks ncdiflow NClgraph ndexr nem netbenchmark netbiov nethet NetPathMiner netprioR netReg netresponse NetSam networkBMA NGScsq noNorm NOISeq nondetects normalize450K NormqPCR normr npGSEA NT
ncleoSNP nucleR nudge NuPoP ocugene OCPuls odseq OGSA oligo oligoClasses OLIN OLingui omicade4 OmicCircos omicplotR omicRespomex OmicsMarker omicsPrint Onassis onco mix Oncocore OncoSimulr onseNSE ontoCAT ontoProc openCyto openPrimer
openPrimeRUI OperaMeta oposSom oppar OPWeight OrderedList Organism OrganismDbi OSCA Oscope OTUbase OutlierD PAA PADOG paircompviz pandar panelan PannBuilder pamp PANR PanVizGenerator PAPI parglims parody Path2PI pathifier PathNet PathoStat
pathprint pathRender pathVar pathway PathwaySplice paxtools Phase pbcmc pcaExplorer pcaGoPromoter pcaMethods PCAN poot2 PCpheno pcxn pdinfoBuilder PECA pepStat pepXMLTab PGA pgca PGSEA PharmacGx phenoDist phenopath phenoTest PhenStat philr
physophorocalzer phyloseq Pi piano pickgene PLCS Pigengene PING pint pkgDepTools plateCore pletthy plgem plier PLPE prls plw pmn podkat pogos polyester Polyfit POST PPInfer piStats psfinder prada prebs PREDA predictionr preprocessCore Prize probAMr PRO
procoil ProCoNa proFIA profileScoreDist progeny pRoloc pRolocGUI PROMISE PROPER PROPS Prostarr prot2D proteinProfiles ProteomicsAnnotationHubData proteoQC ProtGenerics PSEA psychomics PSICQUIC psyenet2r puma PureCN pvac pvca Pviz PWMEnrich pvz
pmetrics QDNaseq qpcrNorm qpgrah qrqc qsea QUALIFIER quanto quantsmooth QuartPAC QuasR QuaternaryProd QUBIC quusage qvalue R3CPET r3Cset R453Plus1Toolbox R4RNA RaggedExperiment rain Rama RamiGu ramwas randPack RankProd RareVariantVis
rareRatBook1 RBGL RBioinf rBiopaxParser RBm Rbowtie Rbwoltie rbsurv Rcade RCAS RCASPAR rcellminer rCGH RchেমppR RchyOptimyx RcisTarget RcpI RcY3 RcYjs RDADVidWebService rDGlnd Rdspid RDRToolbox ReactomeFA readat ReadqPCR rec recount re
redmer REDseq RefNet RefPlus regionR regionReport regsplice REMR Repitools ReportingTools ReQON restfulse response rFRed rGADeM RGcalay RGMLQ RGraph2Js RGraphviz rGREAT RGSEA rsdep rhdf5 rhdf5Client rhdf5lib Rhitslib rHvDM RiboProfiling riboSeq
RimPort Ringo RIPseeker Risa RITAN RIVER RJMCMCNucleosomes RLMM Rmagpie RMassBank rMAT RmiR RNAinteract RNAtiter RNAProbR rnaseqcomp rnaSeqMap RNASeqPower RnaSeqSampleSize RnBeads Rnits roar ROC Roleswitch rols roma ROntoTools ro
ROTS RPAPRProtoBufLib RpsiXML rpx Rqc rqt rubric rRDP RRHO Rsamttools rsbmI rSFfeader Rsusbread RSvSim RTANDEM RTCA RTCA RTCA RTGAT RTGAToolbox RTN RTNduals RTNsuvival RTopper rtracklayer Rtreemix rTRM rTRMui runibuc RUVCOR RUVnormalize RUVSeq
4Vectors safe sagenhaft SAGx samExploreR sampleClassifier SamSPECTRAL sangerseqR SANFindr savR SBMLR SC3 ScaleCA SCAN scater scDD sode scFeatureFilter schind ScSI smap SCnorm score scoreInvhp scpPipe scanr scsR segmentSeq SELEX Seq
semmap SEPA seq2pathway SeqArray seqbias seqCAT seqc seqA seqcombo seqGSEA seqLogo seqPattern seqplots SeqSQC seqTools SeqVarTools sevenbridges SGseq shinyMethyl shinyTANDEM ShortRead SiCtools sigar SigCheck SigFuge siggenes sights signer Sigs

Bioconductor: Education, Training, and Community

The screenshot shows the Bioconductor website homepage. At the top is the Bioconductor logo with the tagline 'OPEN SOURCE SOFTWARE FOR BIOINFORMATICS'. To the right of the logo is a search bar and a navigation menu with links: Home, Install, Help, Developers, and About. The main content area is divided into four columns. The first column, titled 'About Bioconductor', provides an overview of the project, mentioning its use of R, open-source nature, and availability as an AMI and Docker images. The second column, titled 'Install', lists links for getting started, including installing Bioconductor, exploring packages, getting support, the latest newsletter, following on Twitter, and installing R. The third column, titled 'Learn', lists links for mastering Bioconductor tools, including courses, support site, package vignettes, literature citations, common work flows, FAQ, community resources, and videos. The fourth column, titled 'Develop', lists links for contributing to Bioconductor, including developer resources, using Bioc-devel, devel software, annotation and experiment packages, Amazon Machine Image, latest release announcement, support site, developer resources, use Bioc-devel, devel software, annotation and experiment packages, package guidelines, new package submission, git source control, and build reports. At the bottom of the page, there are links for Support, Events, and a Twitter feed by @Bioconductor.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

About 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, [1473 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker images](#).

News

- Bioconductor [3.6](#) is available.
- Bioconductor [F1000 Research Channel](#) available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* ([abstract](#)) and other [recent literature](#).
- View [recent course material](#).
- Use the [support site](#) to get help installing, learning and using Bioconductor.

Install »

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment packages](#)
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

Develop »

Contribute to *Bioconductor*

- [Developer resources](#)
- [Use Bioc-devel](#)
- [Devel Software](#), [Annotation](#) and [Experiment packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Git source control](#)
- [Build reports](#)

[Support](#) [Events](#)

Tweets by @Bioconductor

Core infrastructure

Community contribution

Building Bridges

Gene Expression Omnibus (GEO)



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



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

Building bridges

OXFORD
ACADEMIC

Bioinformatics

Issues

Advance articles

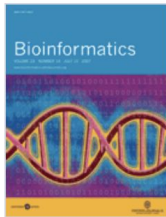
Submit ▼

Purchase

Alerts

About ▼

All Bioinformatics



Volume 23, Issue 14
15 July 2007

Article Contents

Abstract

GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor FREE

Sean Davis ✉, Paul S. Meltzer

Bioinformatics, Volume 23, Issue 14, 15 July 2007, Pages 1846–1847,
<https://doi.org/10.1093/bioinformatics/btm254>

Published: 12 May 2007 **Article history** ▼

■ Split View

📄 PDF

“ Cite

🔑 Permissions

🔗 Share ▼

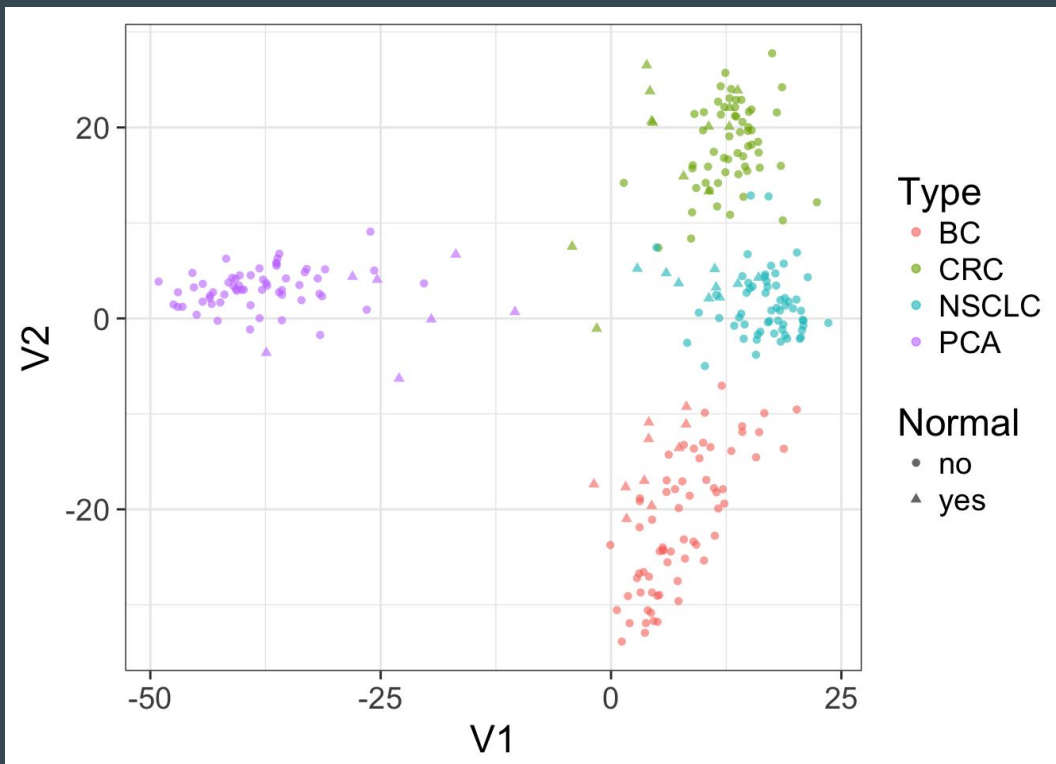
Abstract

Microarray technology has become a standard molecular biology tool.

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/2W3l8dK>

Filter gene expression by variance to find most informative genes.

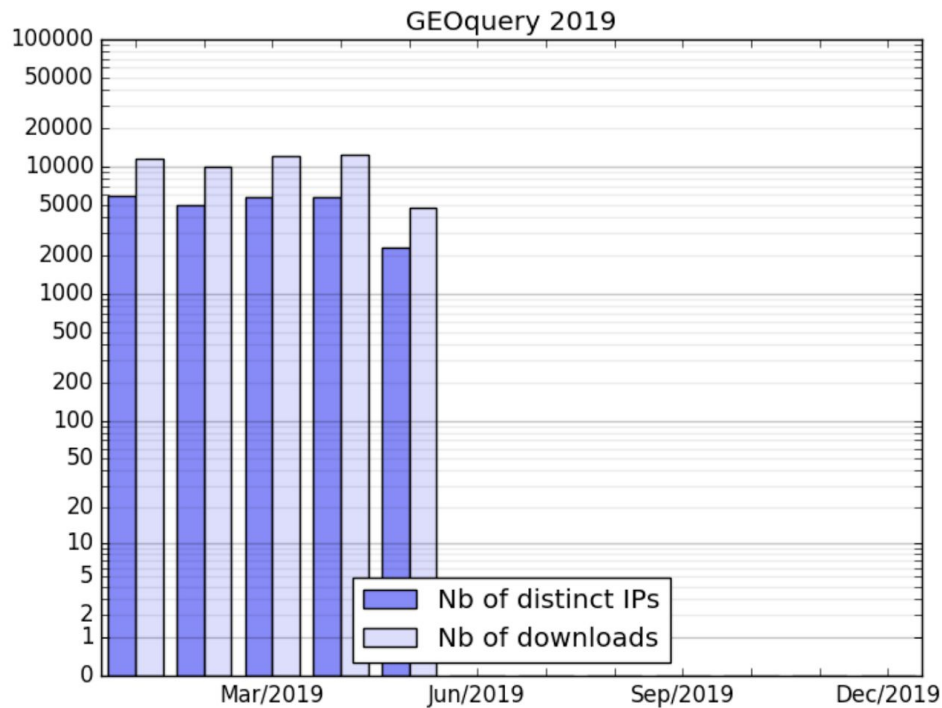
```
sds = apply(exprs(gse), 1, sd)
dat = exprs(gse)[order(sds, decreasing = TRUE)[1:500],]
```

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](#)

NATIONAL CANCER INSTITUTE
GDC Data Portal

[Home](#)
[Projects](#)
[Exploration](#)
[Analysis](#)
[Repository](#)

[Quick Search](#)
[Manage Sets](#)
[Login](#)
[Cart 0](#)
[GDC Apps](#)

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

[Projects](#)
[Exploration](#)
[Analysis](#)
[Repository](#)

Data Portal Summary

[Data Release 16.0 - March 28, 2019](#)

PROJECTS

45

PRIMARY SITES

68

CASES

33,549

FILES

365,463

GENES

22,872

MUTATIONS

3,142,246

Primary Site	Cases (Approximate)
Adrenal Gland	100
Bile Duct	100
Bladder	100
Blood	100
Bone	100
Bone Marrow	100
Brain	100
Breast	3,500
Cervix	100
Colorectal	2,800
Esophagus	100
Eye	100
Head and Neck	100
Kidney	1,200
Liver	100
Lung	4,200
Lymph Nodes	100
Nervous System	2,200
Ovary	100
Pancreas	1,200
Pleura	100
Prostate	1,000
Skin	1,000
Soft Tissue	100
Stomach	1,000
Testis	100
Thyroid	100
Uterus	1,000

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

[Data Portal](#)
[Website](#)
[Data Transfer Tool](#)
[API](#)
[Data Submission Portal](#)
[Documentation](#)
[Legacy Archive](#)



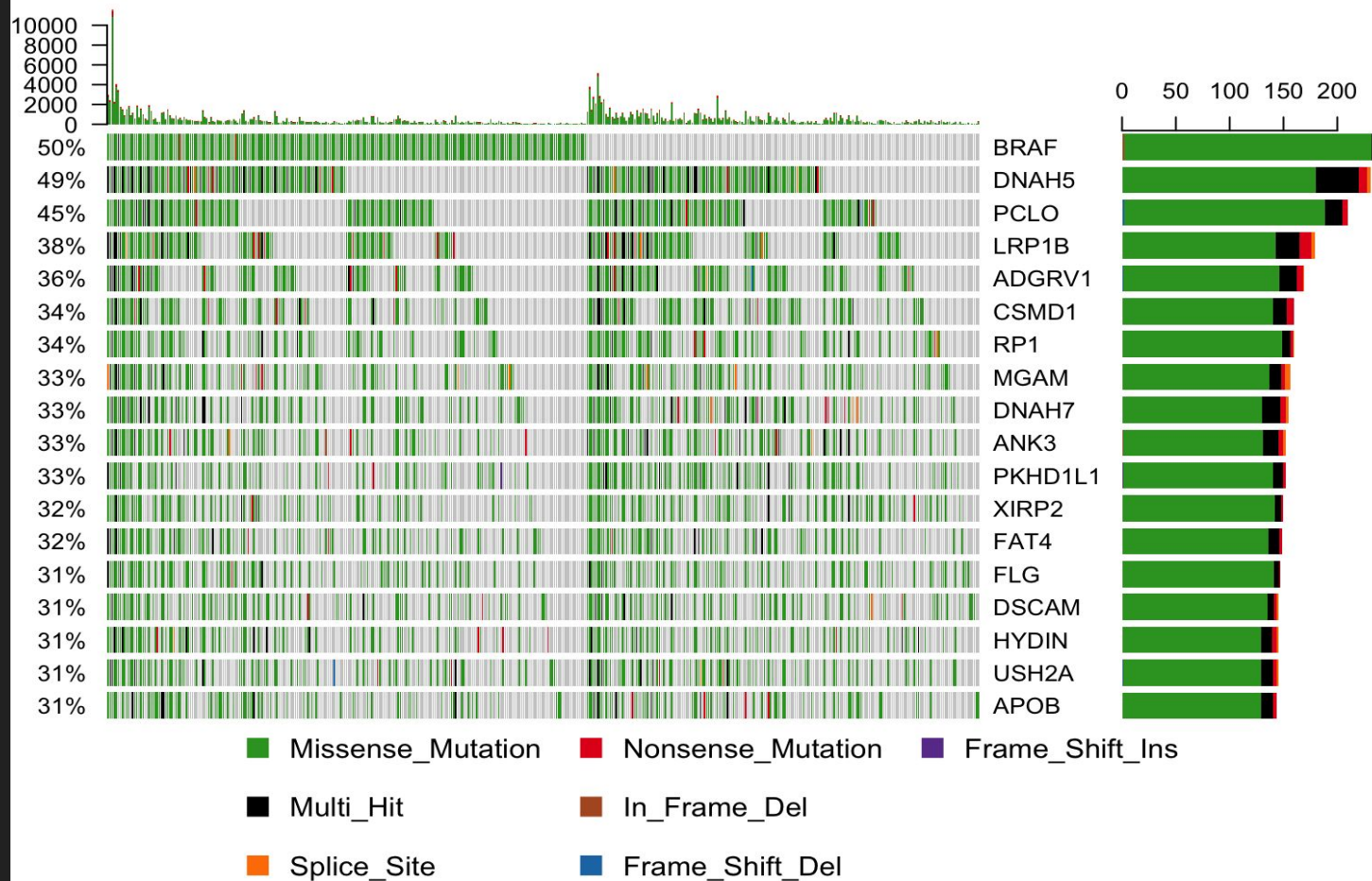
Use the **GenomicDataCommons** package to find and download variants from the TCGA cutaneous melanoma dataset.

```
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.

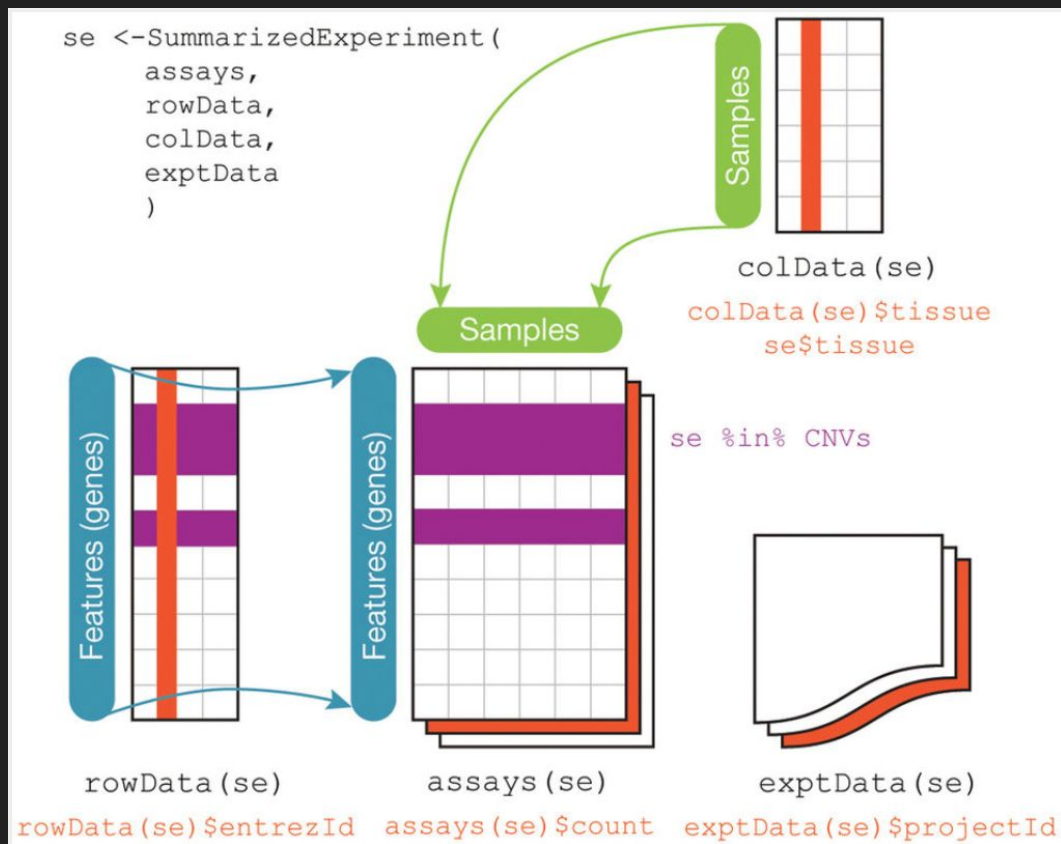


Baby steps: Reusable components

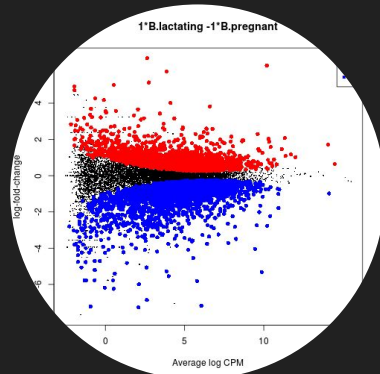
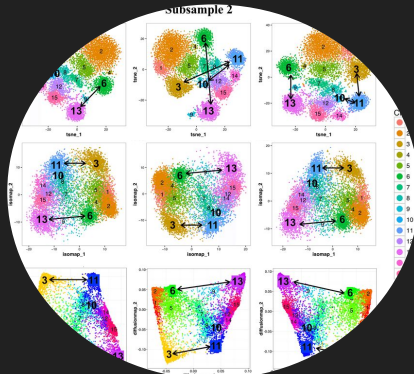
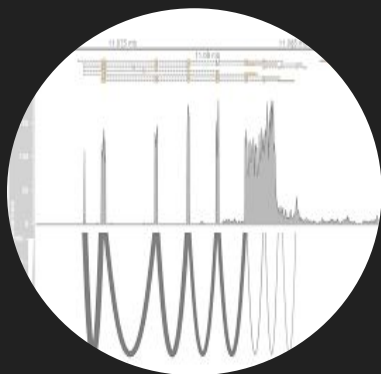
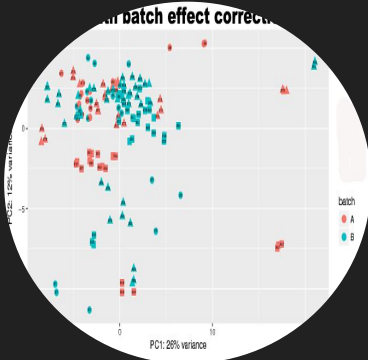
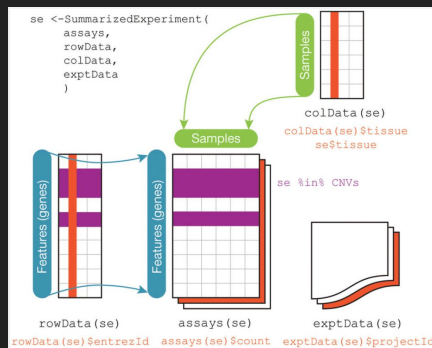
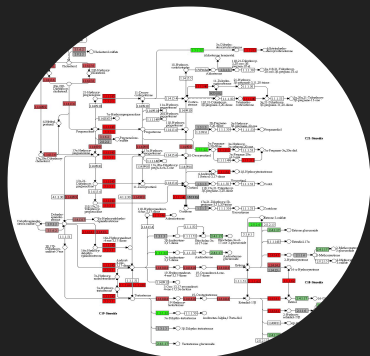
...

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



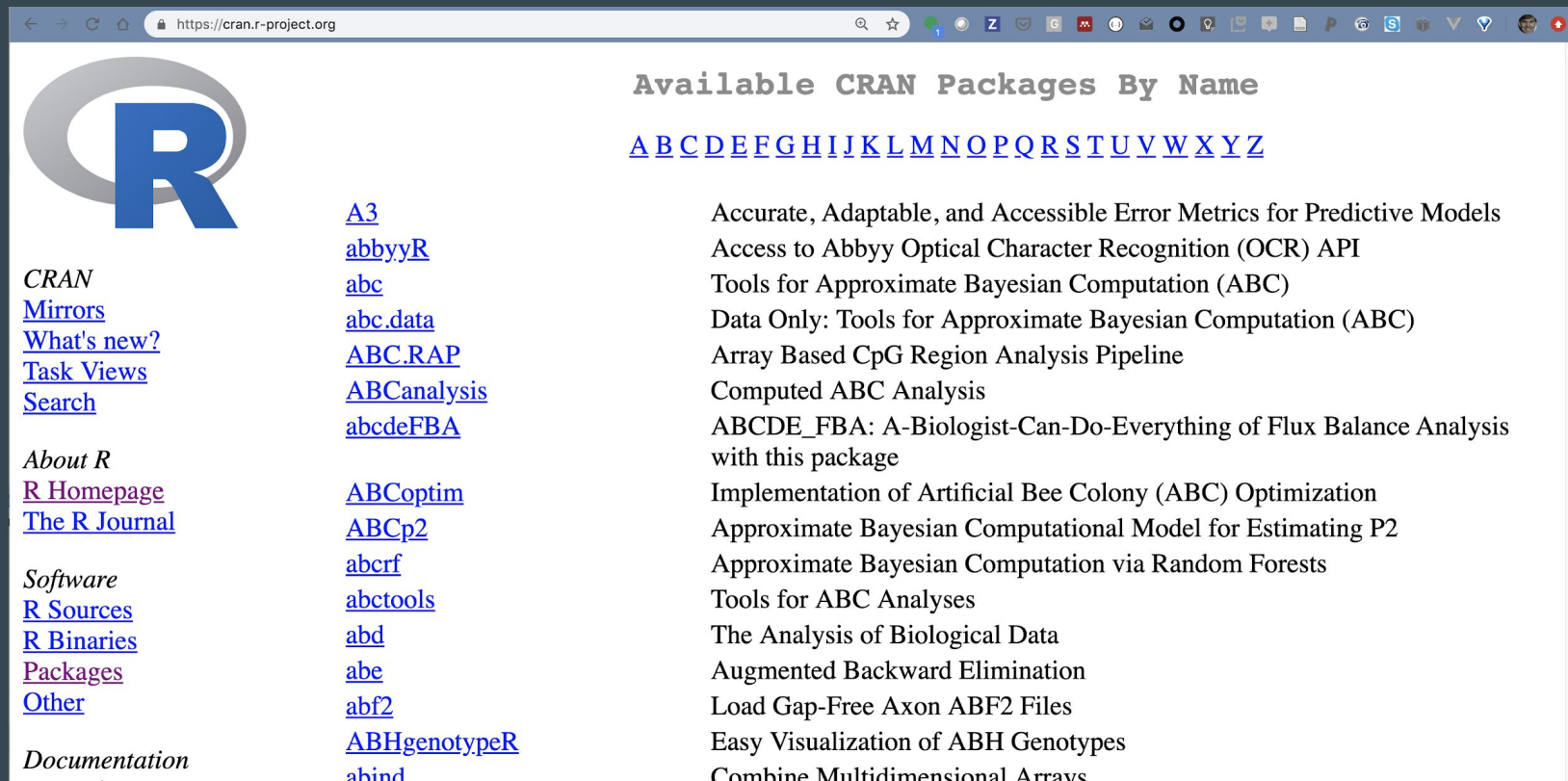
Core value: reuse and interoperability



Where possible reuse existing tools, leverage existing communities



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



<https://cran.r-project.org>



Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

<i>CRAN</i>	A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
Mirrors	abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
What's new?	abc	Tools for Approximate Bayesian Computation (ABC)
Task Views	abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
Search	ABC.RAP	Array Based CpG Region Analysis Pipeline
	ABCanalysis	Computed ABC Analysis
<i>About R</i>	abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
R Homepage	ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
The R Journal	ABCp2	Approximate Bayesian Computational Model for Estimating P2
	abcrf	Approximate Bayesian Computation via Random Forests
<i>Software</i>	abctools	Tools for ABC Analyses
R Sources	abd	The Analysis of Biological Data
R Binaries	abe	Augmented Backward Elimination
Packages	abf2	Load Gap-Free Axon ABF2 Files
Other	ABHgenotypeR	Easy Visualization of ABH Genotypes
<i>Documentation</i>	abind	Combine Multidimensional Arrays

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](https://doi.org/10.18129/B9.bioc.AnnotationHub)



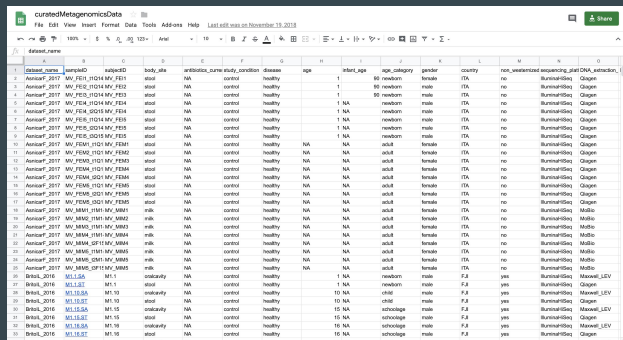
Client to access AnnotationHub resources

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																
100% 0.00 123 Arial 10 B I U A																
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_curren	study_condition	disease	age	infant_age	age_category	gender	country	non_westernized	sequencing_plat	DNA_extraction_	
2	AsnicarF_2017	MV_FEI1_11Q14	MV_FEI1	stool	NA	control	healthy		1	90 newborn	female	ITA	no	illuminaHiSeq	Qiagen	
3	AsnicarF_2017	MV_FEI2_11Q14	MV_FEI2	stool	NA	control	healthy		1	90 newborn	male	ITA	no	illuminaHiSeq	Qiagen	
4	AsnicarF_2017	MV_FEI3_11Q14	MV_FEI3	stool	NA	control	healthy		1	90 newborn	male	ITA	no	illuminaHiSeq	Qiagen	
5	AsnicarF_2017	MV_FEI4_11Q14	MV_FEI4	stool	NA	control	healthy		1 NA	newborn	male	ITA	no	illuminaHiSeq	Qiagen	
6	AsnicarF_2017	MV_FEI4_12Q15	MV_FEI4	stool	NA	control	healthy		1 NA	newborn	male	ITA	no	illuminaHiSeq	Qiagen	
7	AsnicarF_2017	MV_FEI5_11Q14	MV_FEI5	stool	NA	control	healthy		1 NA	newborn	male	ITA	no	illuminaHiSeq	Qiagen	
8	AsnicarF_2017	MV_FEI5_12Q14	MV_FEI5	stool	NA	control	healthy		1 NA	newborn	male	ITA	no	illuminaHiSeq	Qiagen	
9	AsnicarF_2017	MV_FEI5_13Q15	MV_FEI5	stool	NA	control	healthy		1 NA	newborn	male	ITA	no	illuminaHiSeq	Qiagen	
10	AsnicarF_2017	MV_FEM1_11Q1	MV_FEM1	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
11	AsnicarF_2017	MV_FEM2_11Q1	MV_FEM2	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
12	AsnicarF_2017	MV_FEM3_11Q1	MV_FEM3	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
13	AsnicarF_2017	MV_FEM4_11Q1	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
14	AsnicarF_2017	MV_FEM4_12Q1	MV_FEM4	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
15	AsnicarF_2017	MV_FEM5_11Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
16	AsnicarF_2017	MV_FEM5_12Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
17	AsnicarF_2017	MV_FEM5_13Q1	MV_FEM5	stool	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	Qiagen	
18	AsnicarF_2017	MV_MIM1_11M1	MV_MIM1	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
19	AsnicarF_2017	MV_MIM2_11M1	MV_MIM2	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
20	AsnicarF_2017	MV_MIM3_11M1	MV_MIM3	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
21	AsnicarF_2017	MV_MIM4_11M1	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
22	AsnicarF_2017	MV_MIM4_12F1	MV_MIM4	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
23	AsnicarF_2017	MV_MIM5_11M1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
24	AsnicarF_2017	MV_MIM5_12M1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
25	AsnicarF_2017	MV_MIM5_13F1	MV_MIM5	milk	NA	control	healthy	NA	NA	adult	female	ITA	no	illuminaHiSeq	MoBio	
26	Britoll_2016	M1.1.SA	M1.1	oralcavity	NA	control	healthy		1 NA	newborn	male	FJI	yes	illuminaHiSeq	Maxwell_LEV	
27	Britoll_2016	M1.1.ST	M1.1	stool	NA	control	healthy		1 NA	newborn	male	FJI	yes	illuminaHiSeq	Qiagen	
28	Britoll_2016	M1.10.SA	M1.10	oralcavity	NA	control	healthy		10 NA	child	male	FJI	yes	illuminaHiSeq	Maxwell_LEV	
29	Britoll_2016	M1.10.ST	M1.10	stool	NA	control	healthy		10 NA	child	male	FJI	yes	illuminaHiSeq	Qiagen	
30	Britoll_2016	M1.15.SA	M1.15	oralcavity	NA	control	healthy		15 NA	schoolage	male	FJI	yes	illuminaHiSeq	Maxwell_LEV	
31	Britoll_2016	M1.15.ST	M1.15	stool	NA	control	healthy		15 NA	schoolage	male	FJI	yes	illuminaHiSeq	Qiagen	
32	Britoll_2016	M1.16.SA	M1.16	oralcavity	NA	control	healthy		16 NA	schoolage	male	FJI	yes	illuminaHiSeq	Maxwell_LEV	
33	Britoll_2016	M1.16.ST	M1.16	stool	NA	control	healthy		16 NA	schoolage	male	FJI	yes	illuminaHiSeq	Qiagen	
34	Britoll_2016	M1.20.SA	M1.20	oralcavity	NA	control	healthy		20 NA	adult	male	FJI	yes	illuminaHiSeq	Maxwell_LEV	



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>

This article is a preprint and has not been peer-reviewed [what does this mean?].

[Abstract](#)[Full Text](#)[Info/History](#)[Metrics](#)[Preview PDF](#)



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 Shepherd³, Nitesh Turaga³, Ludwig Geistlinger², Martin T. Morgan³, Benjamin Haibe-Kains^{4,5}, Levi Waldron ²



[Author details](#)



METRICS

917



VIEWS

80



DOWNLOADS

Additional thoughts

...

THE DATA SCIENCE HIERARCHY OF NEEDS

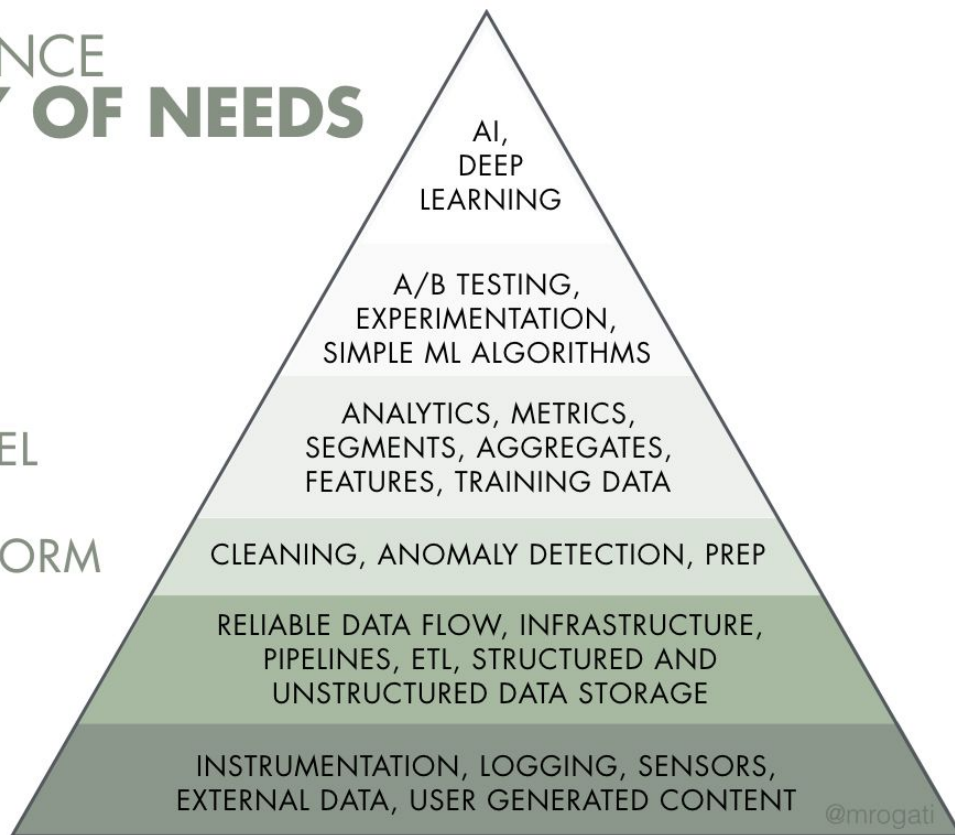
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



Metadata are data

BMC Part of Springer Nature Explore Journals Get Published About BMC Search Login

BMC Bioinformatics

Home About Articles Submission Guidelines

Abstract Background Design Implementation Results and discussion Conclusions Availability and requirements Declarations References

Software | Open Access

SRADB: query and use public next-generation sequencing data from within R

Yuelin Zhu, Robert M Stephens, Paul S Meltzer and Sean R Davis

BMC Bioinformatics 2013 14:19
<https://doi.org/10.1186/1471-2105-14-19> © Zhu et al.; licensee BioMed Central Ltd. 2013
Received: 28 September 2012 | Accepted: 11 January 2013 | Published: 17 January 2013

Abstract

Background

The Sequence Read Archive (SRA) is the largest public repository of sequencing data from the next generation of sequencing platforms including Illumina (Genome Analyzer, HiSeq, MiSeq, etc), Roche 454 GS

Download PDF

Export citations

Section

Sequence analysis (applications)

Metrics

Article accesses: 10691

Citations: 34 more information

Altmetric Attention Score: 21

OXFORD ACADEMIC Sign In Register

Bioinformatics

Issues Advance articles Submit Purchase Alerts About

GEOMETADB: powerful alternative search engine for the Gene Expression Omnibus

Yuelin Zhu, Sean Davis, Robert Stephens, Paul S. Meltzer, Yidong Chen

Bioinformatics, Volume 24, Issue 23, 1 December 2008, Pages 2798–2800, <https://doi.org/10.1093/bioinformatics/btn520>
Published: 07 November 2008 Article history

Split View PDF Cite Permissions Share

Abstract

The NCBI Gene Expression Omnibus (GEO) represents the largest public repository of microarray data. However, finding data in GEO can be challenging. We have

1 INTRODUCTION

OXFORD ACADEMIC

Bioinformatics

Issues Advance articles Submit Purchase Alerts About All Bioinformatics

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, Wolfgang Huber

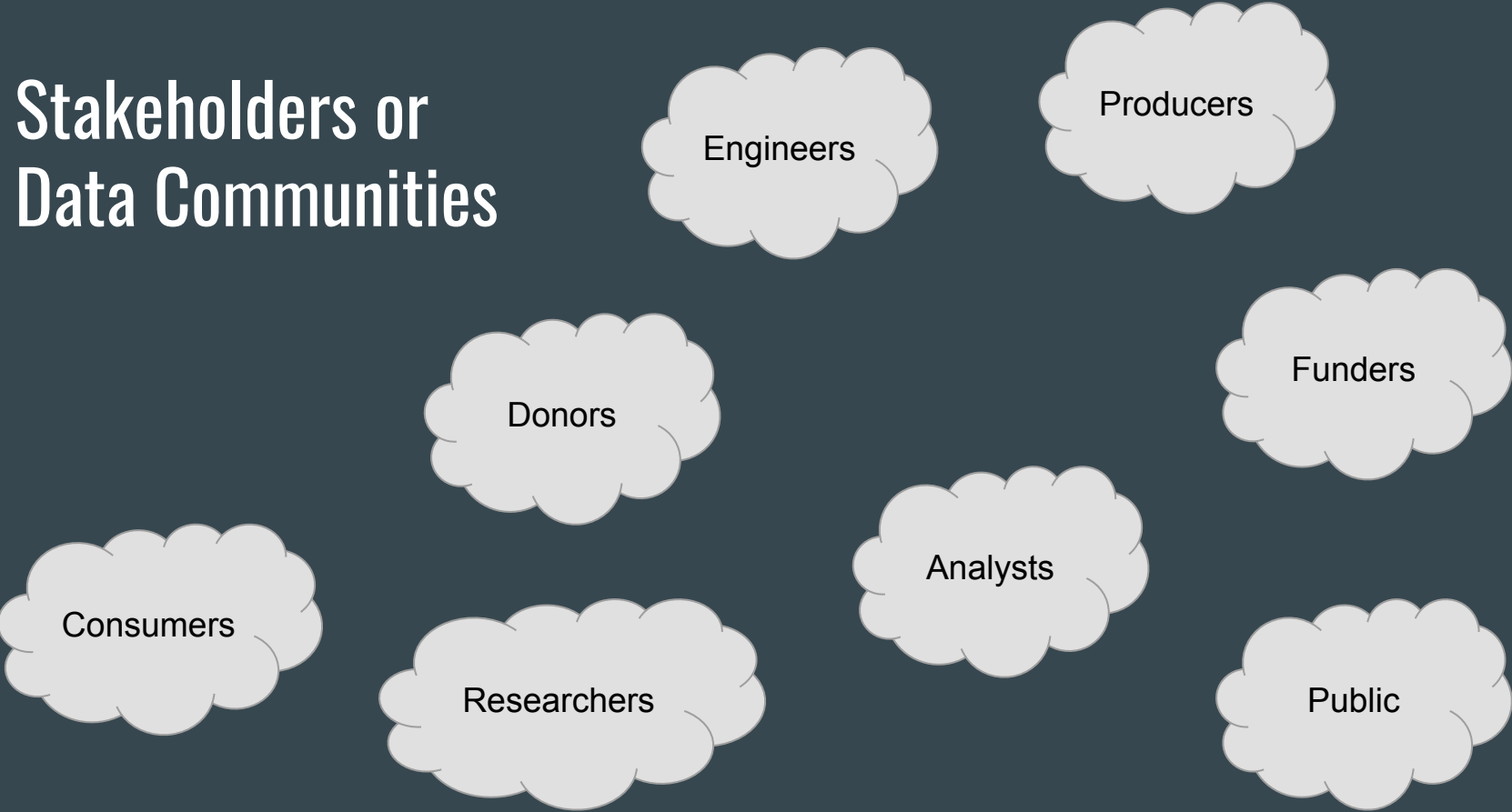
Bioinformatics, Volume 21, Issue 16, , Pages 3439–3440, <https://doi.org/10.1093/bioinformatics/bti525>
Published: 02 June 2005 Article history

Split View PDF Cite Permissions Share

Abstract

Summary: *biomaRt* is a new Bioconductor package that integrates BioMart data resources with data analysis software in Bioconductor. It can annotate a wide

Stakeholders or Data Communities



Engineers

Producers

Donors

Funders

Consumers

Researchers

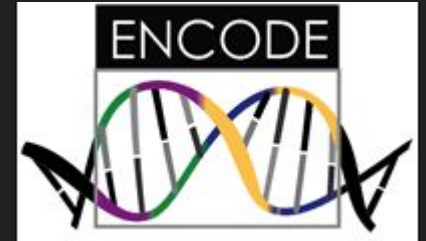
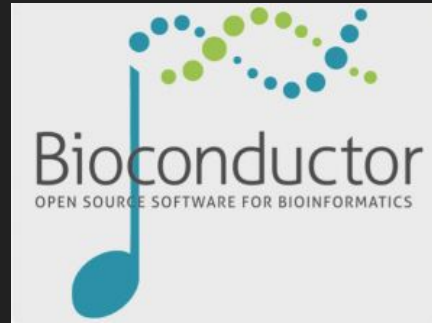
Analysts

Public

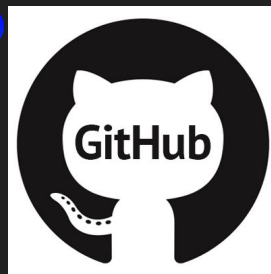
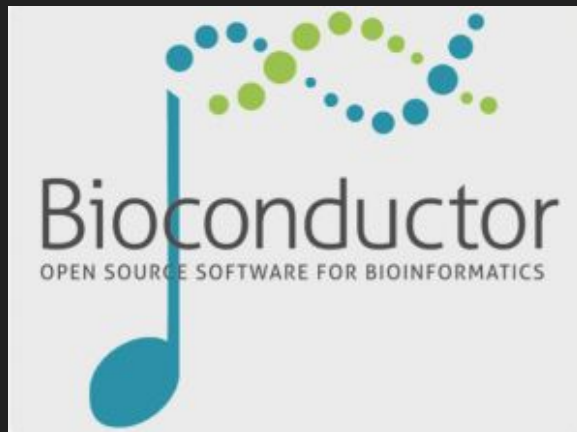
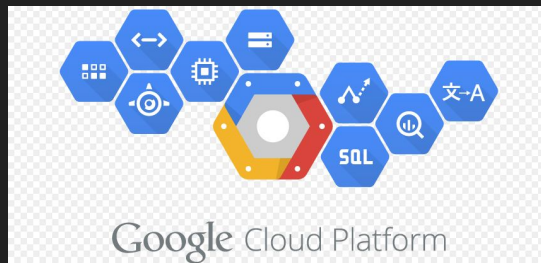


PHARMACODB

MINE MULTIPLE CANCER PHARMACOGENOMIC DATASETS



Commodity big data technologies



Education: we are all learning and teaching....



- Home
- About
- Archive
- Conferences
- Courses
- Interviews
- Contributing

- Twitter
- GitHub

© 2011 - 2017. All rights reserved.
Built with [blogdown](#) and [Hugo](#). Theme [Blackburn](#).

The role of academia in data science education

 Rafael Irizarry  2018/11/01

I was recently asked to moderate an academic panel on the role of universities in training the data science workforce. I preceded each question with opinionated introductions which I have fused into this blog post. These are weakly held opinions so please consider commenting if you disagree with anything.

To discuss data science education we first need to clearly state what it means. The panel organizers defined data science as **“an emerging discipline that draws upon knowledge in statistical methodology and computer science to create impactful predictions and insights for a wide range of traditional scholarly fields.”** But is it an academic discipline? If so, what are the shared fundamental principles, expertise, skills, and knowledge-based shared by data scientists? Is there a core curriculum for *Data Science*? Providing a more detailed definition might help.

My attempt at defining *Data Science*

The term *Data Sciece* may have been [coined in academia](#), but the proliferation of its use has been mostly driven by the tech industry. The term became prominent because recruiters needed to more specifically describe what

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>

