Bioconductor

Software for orchestrating high-throughput biological data analysis

Sean Davis, MD, PhD
@seandavis12
https://seandavi.github.io

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

software

encodes

knowledge

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.

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

Bioconductor by the numbers

- Project started in 2002
- Built on and in R, the open source software platform for data science
  - An estimated 2,000,000 users worldwide
  - 14,376 contributed software libraries (like plugins)
  - Runs on laptops to the largest computer systems in the world
- More than 50,000 unique downloads per month
- More than 22,000 PubmedCentral citations
- Bioconductor Release: more than 2000 individual, biomedical and ‘omic data science software packages
- Receiving submissions of 3-6 new packages *per week*
- Hundreds of active developers

4 a4Base a4Classif a4Core a4Preproc a4Reporting ABAEnrichment ABarray ABSSeq acde aCGH ACME ADaCGH2 adSplit affxparser affy affycomp AffyCompatible affyContam affycoretools AffyExpress affylLM affyio affylmGUI affyPara affypdnn affyPLM affyQCReport
ffyRNADegradation AGDEX agilp AgiMicroRna AIMS ALDEx2 AllelicImbalance alpine alsace altcdfenvs AMOUNTAIN amplican ampliQueso AnalysisPageServer anamiR Anaquin AneuFinder ANF annaffy annmap annotate AnnotationDbi AnnotationFilter AnnotationForge
nnotationFuncs AnnotationHub AnnotationHubData annotationTools annotatr anota anota2seq antiProfiles apComplex apeglm aroma ArrayExpress ArrayExpressHTS arrayMvout arrayQuality arrayQualityMetrics ArrayTools ArrayTV ARRmNormalization ASAFE ASEB
SGSCA ASpli ASSET ASSIGN ATACseqQC attract AUCell BaalChIP BAC bacon BADER BadRegionFinder BAGS ballgown bamsignals banocc basecallQC BaseSpaceR Basic4Cseq BASiCS BasicSTARRseq BatchQC BayesKnockdown BayesPeak baySeq BBCAnalyz
CRANK bcSeq beachmat beadarray beadarraySNP BeadDataPackR BEARscc BEAT BEclear bgafun BgeeDB BGmix bgx BHC BicARE BiFET BiGGR bigmelon bigmemoryExtras bioassayR Biobase biobroom bioCancer BiocCaseStudies BiocCheck BiocFileCache
iocGenerics biocGraph BiocInstaller BioCor BiocParallel BiocSklearn BiocStyle biocViews BiocWorkflowTools bioDist biomaRt biomformat BioMVCClass biomvRCNS BioNet BioQC BioSeqClass biosigner Biostrings biosvd biotmle biovizBase BiRewire birta birte BiSeq Bi
ima BLMA bnbc BPRMeth BRAIN BrainStars branchpointer bridge BridgeDbR BrowserViz BrowserVizDemo BSgenome bsseq BubbleTree BufferedMatrix BufferedMatrixMethods BUMHMM bumphunter BUS CAFE CAGEr CALIB CAMERA canceR cancerclass CancerIn
ancerMutationAnalysis CancerSubtypes CAnD caOmicsV Cardinal casper CATALYST Category categoryCompare CausalR cbaf ccmap CCPROMISE ccrepe cellbaseR cellGrowth cellHTS2 cellity CellMapper CellNOptR cellscape cellTree CEMiTool CexoR CFAssay CG
GHbase CGHcall cghMCR CGHnormaliter CGHregions ChAMP CHARGE charm ChemmineOB ChemmineR Chicago chimera chimeraviz ChIPanalyser ChIPComp chipenrich ChIPexoQual ChIPpeakAnno ChIPQC ChIPseeker chipseq ChIPseqR ChIPSeqSpike ChIPsim
hIPXpress chopsticks chroGPS chromDraw ChromHeatMap chromPlot chromstaR chromswitch chromVAR CHRONOS CINdex cisPath ClassifyR cleanUpdTSeq cleaver clippda clipper Clomial Clonality clonotypeR clst clstutils clustComp clusterExperiment ClusterJudge
usterProfiler clusterSeq ClusterSignificance clusterStab CMA cn CNAnorm CNEr CNORdt CNORfeeder CNORfuzzy CNORode CNPBayes CNTools cnvGSA CNVPanelizer CNVrd2 CNVtools cobindR CoCiteStats codelink CODEX coexnet CoGAPS cogena coGPS COH
oMET COMPASS compcodeR compEpiTools CompGO ComplexHeatmap CONFESS ConsensusClusterPlus consensusOV consensusSeekeR contiBAIT conumee convert copa copynumber CopywriteR CoRegNet Cormotif CorMut CORREP coseq cosmiq COSNet Coun
ovEB CoverageView covRNA cpvSNP cqn CRImage CRISPRseek crisprseekplus CrispRVariants crlmm crossmeta CSAR csaw CSSP ctc CTDquerier ctsGE cummeRbund customProDB CVE cycle cydar cytofkit cytolib CytoML dada2 dagLogo daMA DaMiRseq DAPAR D
ASC DBChIP dcGSA DChIPRep ddCt debrowser DECIPHER DEComplexDisease DeconRNASeq DEDS DeepBlueR deepSNV DEFormats DEGraph DEGreport DEGseq DelayedArray DelayedMatrixStats deltaGseg DeMAND DEP derfinder derfinderHelper derfinderPlo
ESeq DESeq2 destiny DEsubs DEXSeq dexus DFP DiffBind diffGeneAnalysis diffHic DiffLogo diffloop diffuStats diggit Director DirichletMultinomial discordant dks DMCHMM DMRcaller DMRcate DMRforPairs DMRScan DNABarcodes DNAcopy DNAshapeR doppelgangl
OQTL Doscheda DOSE drawProteins DRIMSeq DriverNet DropletUtils DrugVsDisease dSimer DSS DTA dualKS DupChecker dupRadar dyebias DynDoc EasyqpcR easyRNASeq EBarrays EBcoexpress EBImage EBSEA EBSeq EBSeqHMM ecolitk EDASeq EDDA edge
dgeR eegc EGAD EGSEA eiR eisa ELBOW ELMER EMDomics EmpiricalBrownsMethod ENCODExplorer ENmix EnrichedHeatmap EnrichmentBrowser ensembldb ensemblVEP ENVISIONQuery EpiDISH epigenomix epiNEM epivizr epivizrChart epivizrData epivizrServe
pivizrStandalone erccdashboard erma esATAC esetVis eudysbiome EventPointer ExiMiR exomeCopy exomePeak ExperimentHub ExperimentHubData explorase ExpressionAtlas ExpressionView fabia facopy factDesign FamAgg farms fastLiquidAssociation fastseg fCCA
lrame FEM ffpe FGNet fgsea FindMyFriends FISHalyseR FitHiC flagme flipflop flowAI flowBeads flowBin flowcatchR flowCHIC flowCL flowClean flowClust flowCore flowCyBar flowDensity flowFit flowFP flowMap flowMatch flowMeans flowMerge flowPeaks flowPloidy flow
owQ flowQB FlowRepositoryR FlowSOM flowStats flowTime flowTrans flowType flowUtils flowViz flowVS flowWorkspace fmcsR focalCall FourCSeq FRGEpistasis frma frmaTools FunChIP FunciSNP funtooNorm GA4GHclient GA4GHshiny gaga gage gaggle gaia GApred
arfield gaucho gcapc gcatest gCMAP gCMAPWeb gCrisprTools gcrma GDCRNATools gdsfmt geecc GEM genArise genbankr GeneAnswers geneAttribution GeneBreak geneClassifiers GeneExpressionSignature genefilter genefu GeneGA GeneGeneInteR GeneMeta
eneNetworkBuilder GeneOverlap geneplast geneplotter geneRecommender GeneRegionScan geneRxCluster GeneSelectMMD GeneSelector GENESIS geNetClassifier GeneticsDesign GeneticsPed geneXtendeR GENIE3 genoCN GenoGAM genomation GenomeGraph
enomeInfoDb genomeIntervals genomes GenomicAlignments GenomicDataCommons GenomicFeatures GenomicFiles GenomicInteractions GenomicRanges GenomicScores GenomicTuples Genominator genoset genotypeeval genphen GenRank GenVisR GEOmetadb
EOquery GEOsubmission gep2pep gespeR GEWIST GGBase ggbio ggcyto GGtools ggtree girafe GISPA GLAD Glimma GlobalAncova globalSeq globaltest gmapR GMRP GOexpress GOfuncR GOFunction GoogleGenomics GOpro goProfiles GOSemSim goseq GOSim
oSTAG GOstats GOsummaries GOTHiC goTools gpls gprege gQTLBase gQTLstats graph GraphAlignment GraphAT graphite GraphPAC GRENITS GreyListChIP GRmetrics groHMM GRridge GSALightning GSAR GSCA GSEABase GSEAlm gsean GSReg GSRI GSVA g
UIDEseq Guitar Gviz gwascat GWASTools h5vc hapFabia Harman Harshlight HDF5Array HDTD heatmaps Heatplus HelloRanges HELP HEM hiAnnotator HIBAG HiCcompare hicrep hierGWAS HilbertCurve HilbertVis HilbertVisGUI hiReadsProcessor HiTC hmdbQuery
MMcopy hopach hpar HTqPCR HTSanalyzeR HTSeqGenie htSeqTools HTSFilter HybridMTest hyperdraw hypergraph iASeq iBBiG ibh iBMQ iCARE Icens iCheck iChip iClusterPlus iCOBRA ideal IdeoViz idiogram IdMappingAnalysis IdMappingRetrieval iGC IHW illumina
nageHTS IMAS Imetagene ImmuneSpaceR immunoClust IMPCdata ImpulseDE ImpulseDE2 impute InPAS INPower INSPEcT intansv InteractionSet interactiveDisplay interactiveDisplayBase IntEREst InterMineR IntramiRExploreR inveRsion IONiseR iPAC IPO IPPD IRa
sSpatialFeatures iSeq isobar IsoformSwitchAnalyzeR IsoGeneGUI ISoLDE isomiRs ITALICS iterativeBMA iterativeBMAsurv iterClust IVAS ivygapSE IWTomics JASPAR2018 joda JunctionSeq karyoploteR KCsmart kebabs KEGGgraph KEGGlincs keggorthology KEGGpr
EGGREST kimod lapmix LBE ldblock LEA LedPred les lfa limma limmaGUI LINC LineagePulse Linnorm LiquidAssociation lmdme LMGene LOBSTAHS loci2path logicFS logitT Logolas lol LOLA LowMACA LPE LPEadj lpNet lpsymphony lumi LVSmiRNA LymphoSeq M3C
3Drop maanova macat maCorrPlot made4 MADSEQ maftools MAGeCKFlute maigesPack MAIT makecdfenv MANOR manta MantelCorr maPKL maPredictDSC mapscape marray maSigPro maskBAD MassArray massiR MassSpecWavelet MAST matchBox MatrixRider r
axContrastProjection MBAmethyl MBASED MBCB mBPCR MBttest mcaGUI MCbiclust MCRestimate mCSEA mdgsa mdqc MEAL MeasurementError MEDIPS MEDME MEIGOR MergeMaid Mergeomics MeSHDbi meshes meshr messina metaArray Metab metabomxtr
etaboSignal metaCCA MetaCyto metagene metagenomeFeatures metagenomeSeq metahdep metaMS MetaNeighbor metaSeq metaseqR metavizr MetCirc methimpute methInheritSim MethPed MethTargetedNGS methVisual methyAnalysis MethylAid methylInheritance
ethylKit MethylMix methylMnM methylPipe MethylSeekR methylumi methyvim mfa Mfuzz MGFM MGFR mgsa MiChip microbiome microRNA MIGSA mimager MIMOSA MineICA minet minfi MinimumDistance MiPP MIRA MiRaGE miRBaseConverter miRcomp mirIntegrat
iRLAB miRmine miRNAmeConverter miRNApath miRNAtap miRsponge Mirsynergy missMethyl mitoODE MLInterfaces MLP MLSeq MMDiff2 MmPalateMiRNA MODA mogsa monocle MoonlightR MoPS mosaics motifbreakR motifcounter MotifDb motifmatchr motifRG
otifStack MotIV MPFE mpra mQTL msa msgbsR MSGFgui MSGFplus msmsEDA msmsTests MSnbase MSnID msPurity MSstats Mulcom MultiAssayExperiment multiClust MultiDataSet MultiMed multiMiR multiOmicsViz multiscan multtest muscle MutationalPatterns MVC
vGST MWASTools mygene myvariant mzID mzR NADfinder NanoStringDiff NanoStringQCPro NarrowPeaks ncdfFlow NCIgraph ndexr nem netbenchmark netbiov nethet NetPathMiner netprioR netReg netresponse NetSAM networkBMA NGScopy nnNorm NOISeq nond
ormalize450K NormqPCR normr npGSEA NTW nucleoSim nucleR nudge NuPoP occugene OCplus odseq OGSA oligo oligoClasses OLIN OLINgui omicade4 OmicCircos omicplotR omicRexposome OmicsMarkeR omicsPrint Onassis oncomix OncoScore OncoSimulR
neSENSE ontoCAT ontoProc openCyto openPrimeR openPrimeRui OperaMate oposSOM oppar OPWeight OrderedList Organism OrganismDbi OSAT Oscope OTUbase OutlierD PAA PADOG paircompviz pandaR panelcn PAnnBuilder panp PANR PanVizGenerator PAP
arglms parody Path2PPI pathifier PathNet PathoStat pathprint pathRender pathVar pathview PathwaySplice paxtoolsr Pbase pbcmc pcaExplorer pcaGoPromoter pcaMethods PCAN pcot2 PCpheno pcxn pdInfoBuilder PECA pepStat pepXMLTab PGA pgca PGSEA
harmacoGx phenoDist phenopath phenoTest PhenStat philr phosphonormalizer phyloseq Pi piano pickgene PICS Pigengene PING pint pkgDepTools plateCore plethy plgem plier PLPE plrs plw pmm podkat pogos polyester Polyfit POST PPInfer ppiStats pqsfinder prada
REDA predictionet preprocessCore Prize proBAMr PROcess procoil ProCoNA proFIA profileScoreDist progeny pRoloc pRolocGUI PROMISE PROPER PROPS Prostar prot2D proteinProfiles ProteomicsAnnotationHubData proteoQC ProtGenerics PSEA psichomics PSIC
sygenet2r puma PureCN pvac pvca Pviz PWMEnrich pwOmics qcmetrics QDNAseq qpcrNorm qpgraph qrqc qsea QUALIFIER quantro quantsmooth QuartPAC QuasR QuaternaryProd QUBIC qusage qvalue R3CPET r3Cseq R453Plus1Toolbox R4RNA RaggedExperime
ma RamiGO ramwas randPack RankProd RareVariantVis Rariant RbcBook1 RBGL RBioinf rBiopaxParser RBM Rbowtie Rbowtie2 rbsurv Rcade RCAS RCASPAR rcellminer rCGH Rchemcpp RchyOptimyx RcisTarget Rcpi RCy3 RCyjs RDAVIDWebService rDGIdb Rdis
DRToolbox ReactomePA readat ReadqPCR reb recount recoup RedeR REDseq RefNet RefPlus regioneR regionReport regsplice REMP Repitools ReportingTools ReQON restfulSE rexposome rfPred rGADEM RGalaxy RGMQL RGraph2js Rgraphviz rGREAT RGSEA rg
ndf5 rhdf5client Rhdf5lib Rhtslib rHVDM RiboProfiling riboSeqR RImmPort Ringo RIPSeeker Risa RITAN RIVER RJMCMCNucleosomes RLMM Rmagpie RMassBank rMAT RmiR RNAinteract RNAither RNAprobR rnaseqcomp rnaSeqMap RNASeqPower RnaSeqSample
nBeads Rnits roar ROC Roleswitch rols roma ROntoTools ropls ROTS RPA RProtoBufLib RpsiXML rpx Rqc rqt rqubic rRDP RRHO Rsamtools rsbml rSFFreader Rsubread RSVSim rTANDEM RTCA RTCGA RTCGAToolbox RTN RTNduals RTNsurvival RTopper rtracklay
treemix rTRM rTRMui runibic RUVcorr RUVnormalize RUVSeq RVS S4Vectors safe sagenhaft SAGx samExploreR sampleClassifier SamSPECTRAL sangerseqR SANTA sapFinder savR SBMLR SC3 Scale4C SCAN scater scDD scde scFeatureFilter scfind SclSI scmap
Cnorm scone scoreInvHap scPipe scran scsR segmentSeq SELEX SemDist semisup SEPA seq2pathway SeqArray seqbias seqCAT seqCNA seqcombo SeqGSEA seqLogo seqPattern seqplots SeqSQC seqTools SeqVarTools sevenbridges SGSeq shinyMethyl shinyTA

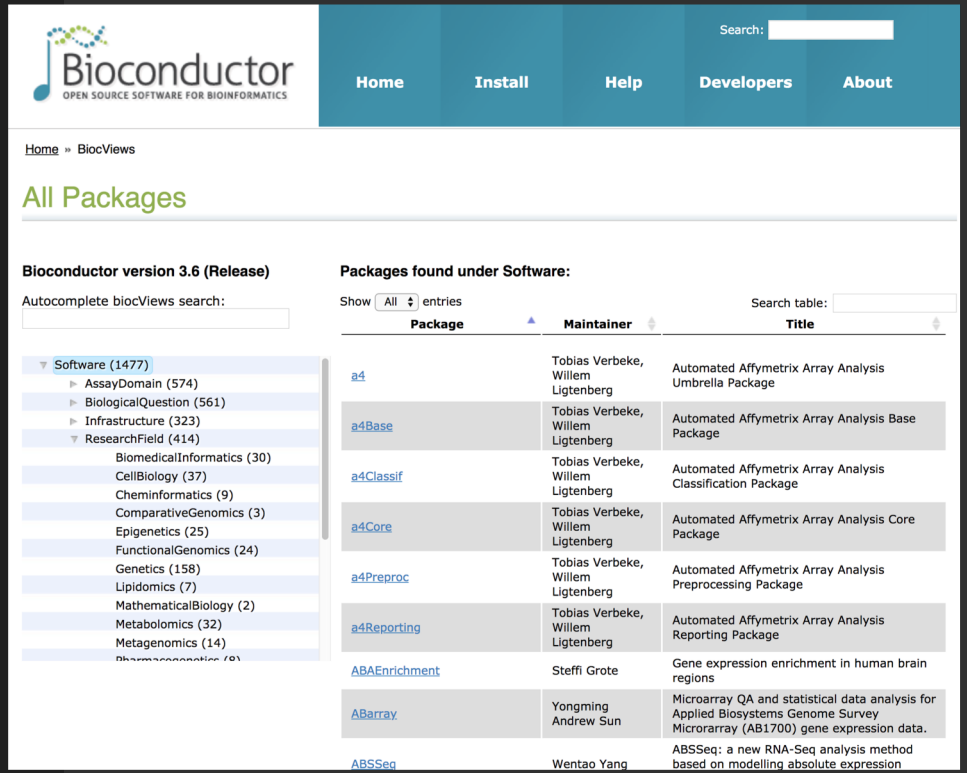
Capabilities

Bioconductor supports all phases of the biomedical data science workflow

- Specialized data import and export
- Data management leveraging FAIR data principles
- Data integration and interpretation, including access to millions of “public” datasets
- Context-specific analysis and statistics
- Visualization and publication-quality plotting
- Efficiency and reproducibility (human in the loop), even at scale
- Reproducible, literate reporting

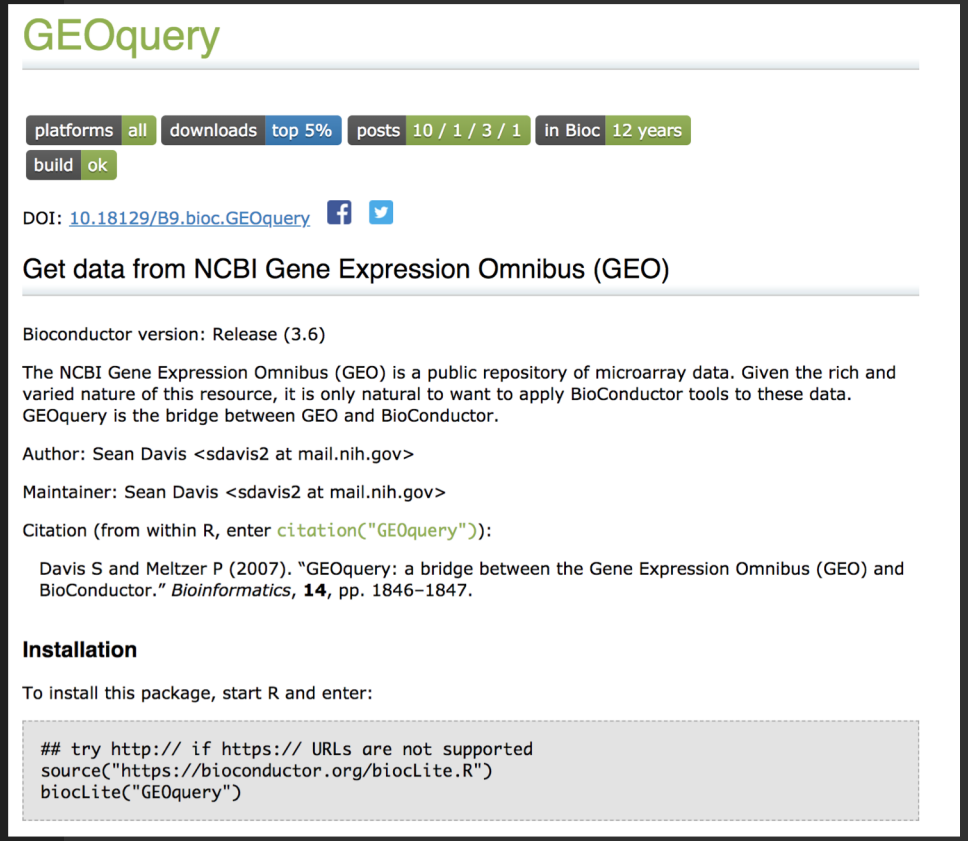
Qualities

- *Discoverable*
- Installable
- Reliable
- Documented
- Supported
- Integrated
- Connected
- Scalable
- State-of-the-art
- Community-driven



Qualities

- Discoverable
- *Installable*
- Reliable
- Documented
- Supported
- Integrated
- Connected
- Scalable
- State-of-the-art
- Community-driven



GEOquery

platforms all | downloads top 5% | posts 10 / 1 / 3 / 1 | in Bioc 12 years | build ok

DOI: 10.18129/B9.bioc.GEOquery

Get data from NCBI Gene Expression Omnibus (GEO)

Bioconductor version: Release (3.6)

The NCBI Gene Expression Omnibus (GEO) is a public repository of microarray data. Given the rich and varied nature of this resource, it is only natural to want to apply BioConductor tools to these data. GEOquery is the bridge between GEO and BioConductor.

Author: Sean Davis <sdavis2 at mail.nih.gov>

Maintainer: Sean Davis <sdavis2 at mail.nih.gov>

Citation (from within R, enter `citation("GEOquery")`):

Davis S and Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, **14**, pp. 1846–1847.

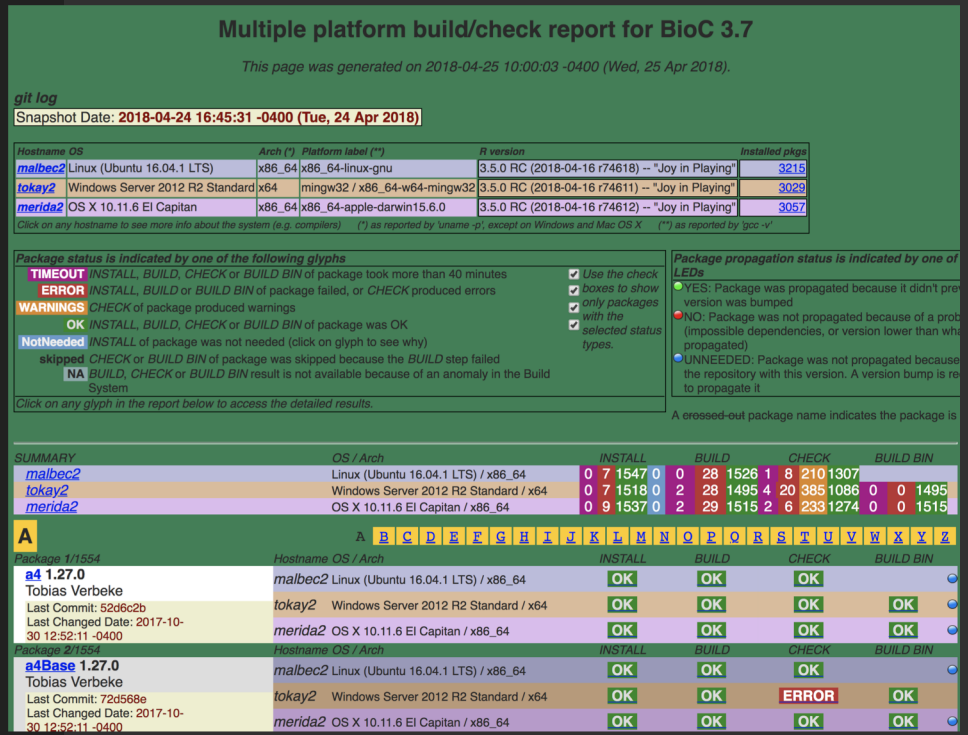
Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("GEOquery")
```

Qualities

- Discoverable
- Installable
- *Reliable*
- Documented
- Supported
- Integrated
- Connected
- Scalable
- State-of-the-art
- Community-driven



Multiple platform build/check report for BioC 3.7

This page was generated on 2018-04-25 10:00:03 -0400 (Wed, 25 Apr 2018).

git log

Snapshot Date: **2018-04-24 16:45:31 -0400 (Tue, 24 Apr 2018)**

Hostname	OS	Arch (*)	Platform label (**)	R version	Installed pkgs
malbec2	Linux (Ubuntu 16.04.1 LTS)	x86_64	x86_64-linux-gnu	3.5.0 RC (2018-04-16 r74618) -- "Joy in Playing"	3215
tokay2	Windows Server 2012 R2 Standard	x64	mingw32 / x86_64-w64-mingw32	3.5.0 RC (2018-04-16 r74611) -- "Joy in Playing"	3029
merida2	OS X 10.11.6 El Capitan	x86_64	x86_64-apple-darwin15.6.0	3.5.0 RC (2018-04-16 r74612) -- "Joy in Playing"	3057

Click on any hostname to see more info about the system (e.g. compilers) (*) as reported by 'uname -p', except on Windows and Mac OS X (**) as reported by 'gcc -v'

Package status is indicated by one of the following glyphs

TIMEOUT *INSTALL, BUILD, CHECK* or *BUILD BIN* of package took more than 40 minutes
ERROR *INSTALL, BUILD* or *BUILD BIN* of package failed, or *CHECK* produced errors
WARNINGS *CHECK* of package produced warnings
OK *INSTALL, BUILD, CHECK* or *BUILD BIN* of package was OK
NotNeeded *INSTALL* of package was not needed (click on glyph to see why)
skipped *CHECK* or *BUILD BIN* of package was skipped because the *BUILD* step failed
NA *BUILD, CHECK* or *BUILD BIN* result is not available because of an anomaly in the Build System

Use the check boxes to show only packages with the selected status types.

Click on any glyph in the report below to access the detailed results.

Package propagation status is indicated by one of LEDs

YES: Package was propagated because it didn't pre version was bumped
NO: Package was not propagated because of a prob (impossible dependencies, or version lower than wh propagated)
UNNEEDED: Package was not propagated because the repository with this version. A version bump is re to propagate it

A ~~crossed-out~~ package name indicates the package is

SUMMARY	OS / Arch	INSTALL				BUILD				CHECK				BUILD BIN			
malbec2	Linux (Ubuntu 16.04.1 LTS) / x86_64	0	7	1547	0	0	28	1526	1	8	210	1307					
tokay2	Windows Server 2012 R2 Standard / x64	0	7	1518	0	2	28	1495	4	20	385	1086	0	0	1495		
merida2	OS X 10.11.6 El Capitan / x86_64	0	9	1537	0	2	29	1515	2	6	233	1274	0	0	1515		

A 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

Package 1/1554	Hostname	OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
a4 1.27.0 Tobias Verbeke Last Commit: 52d6c2b Last Changed Date: 2017-10-30 12:52:11 -0400	malbec2	Linux (Ubuntu 16.04.1 LTS) / x86_64	OK	OK	OK	
	tokay2	Windows Server 2012 R2 Standard / x64	OK	OK	OK	OK
	merida2	OS X 10.11.6 El Capitan / x86_64	OK	OK	OK	OK
Package 2/1554	**Hostname**	**OS / Arch**	**INSTALL**	**BUILD**	**CHECK**	**BUILD BIN**
a4Base 1.27.0 Tobias Verbeke Last Commit: 72d568e Last Changed Date: 2017-10-30 12:52:11 -0400	malbec2	Linux (Ubuntu 16.04.1 LTS) / x86_64	OK	OK	OK	
	tokay2	Windows Server 2012 R2 Standard / x64	OK	OK	ERROR	OK
	merida2	OS X 10.11.6 El Capitan / x86_64	OK	OK	OK	OK

Qualities

- Discoverable
- Installable
- Reliable
- *Documented*
- Supported
- Integrated
- Connected
- Scalable
- State-of-the-art
- Community-driven

The GenomicDataCommons Package

Sean Davis & Martin Morgan

Monday, October 30, 2017

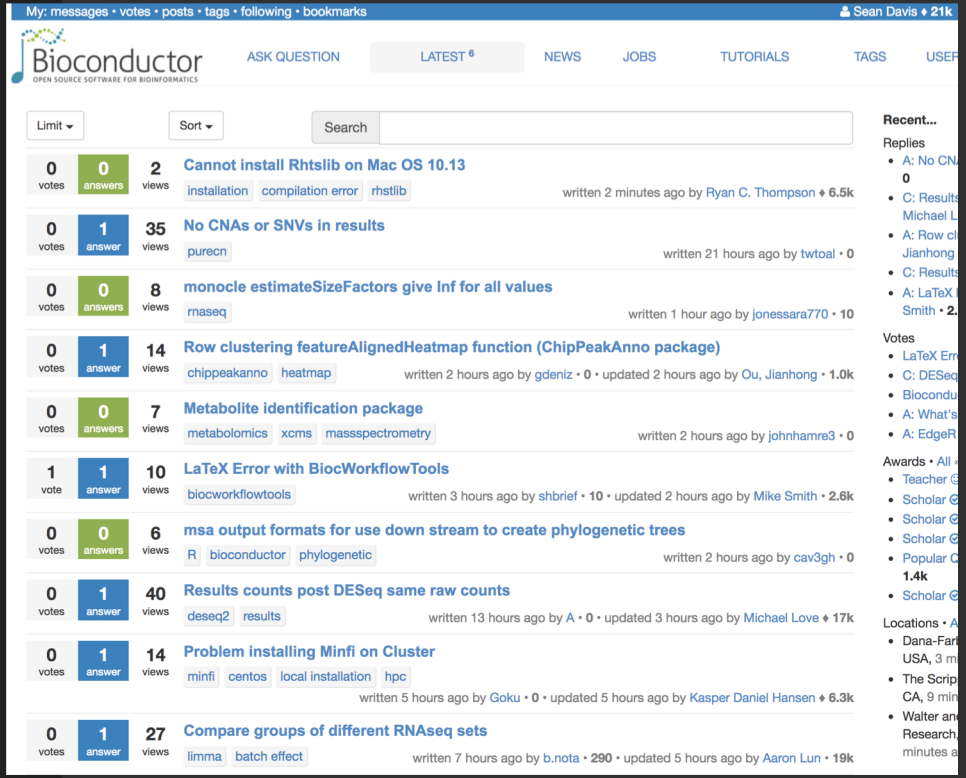
Abstract

The National Cancer Institute (NCI) has established the Genomic Data Commons (GDC). The GDC provides the cancer research community with an open and unified repository for sharing and accessing data across numerous cancer studies and projects via a high-performance data transfer and query infrastructure. The *GenomicDataCommons* Bioconductor package provides basic infrastructure for querying, accessing, and mining genomic datasets available from the GDC. We expect that the Bioconductor developer and the larger bioinformatics communities will build on the *GenomicDataCommons* package to add higher-level functionality and expose cancer genomics data to the plethora of state-of-the-art bioinformatics methods available in Bioconductor.

Contents

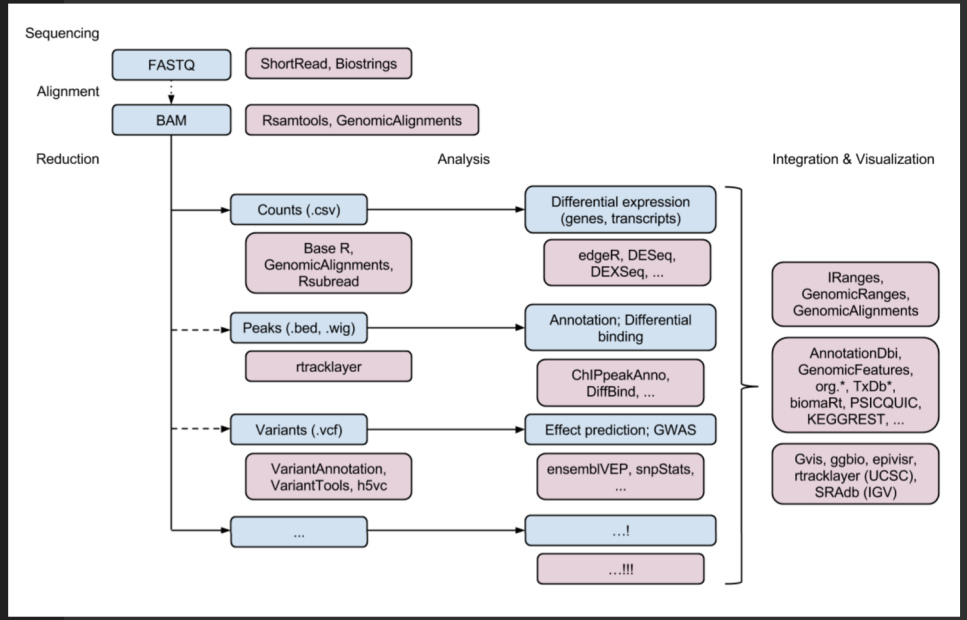
Qualities

- Discoverable
- Installable
- Reliable
- Documented
- *Supported*
- Integrated
- Connected
- Scalable
- State-of-the-art
- Community-driven



Qualities

- Discoverable
- Installable
- Reliable
- Documented
- Supported
- *Integrated*
- Connected
- Scalable
- State-of-the-art
- Community-driven

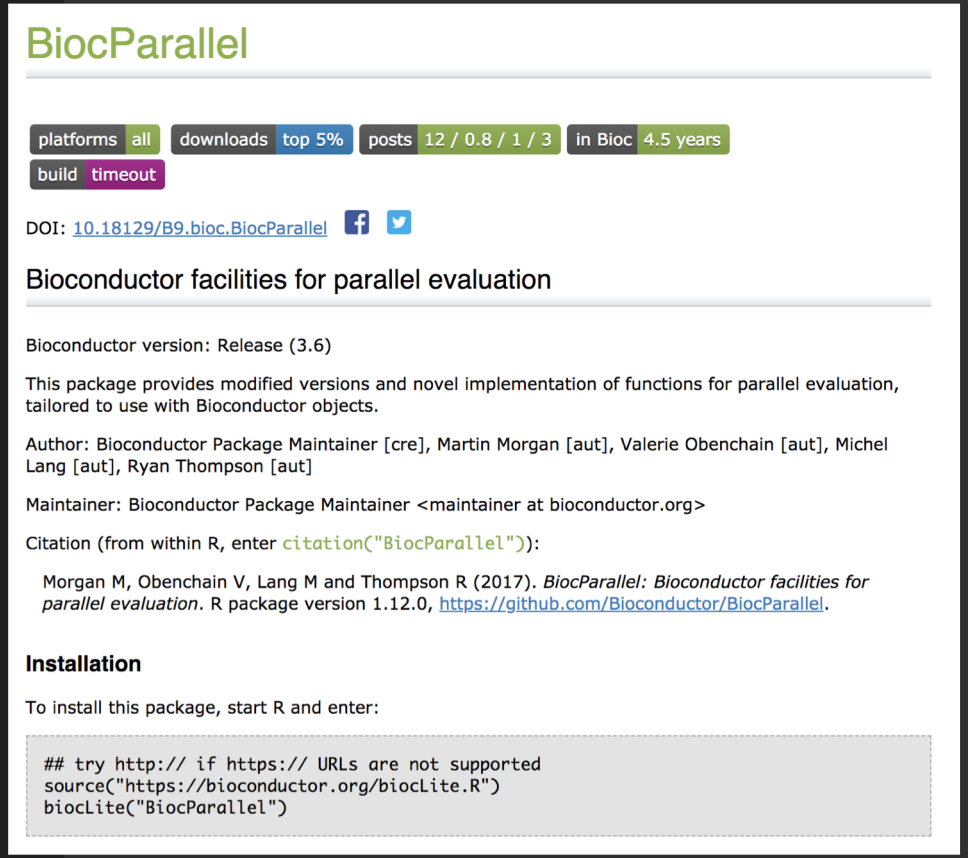


Qualities

- Discoverable
- Installable
- Reliable
- Documented
- Supported
- Integrated
- *Connected*
- Scalable
- State-of-the-art
- Community-driven

Qualities

- Discoverable
- Installable
- Reliable
- Documented
- Supported
- Integrated
- Connected
- *Scalable*
- State-of-the-art
- Community-driven



BiocParallel

platforms all | downloads top 5% | posts 12 / 0.8 / 1 / 3 | in Bioc 4.5 years | build timeout

DOI: 10.18129/B9.bioc.BiocParallel

Bioconductor facilities for parallel evaluation

Bioconductor version: Release (3.6)

This package provides modified versions and novel implementation of functions for parallel evaluation, tailored to use with Bioconductor objects.

Author: Bioconductor Package Maintainer [cre], Martin Morgan [aut], Valerie Obenchain [aut], Michel Lang [aut], Ryan Thompson [aut]

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("BiocParallel")`):

Morgan M, Obenchain V, Lang M and Thompson R (2017). *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.12.0, https://github.com/Bioconductor/BiocParallel.

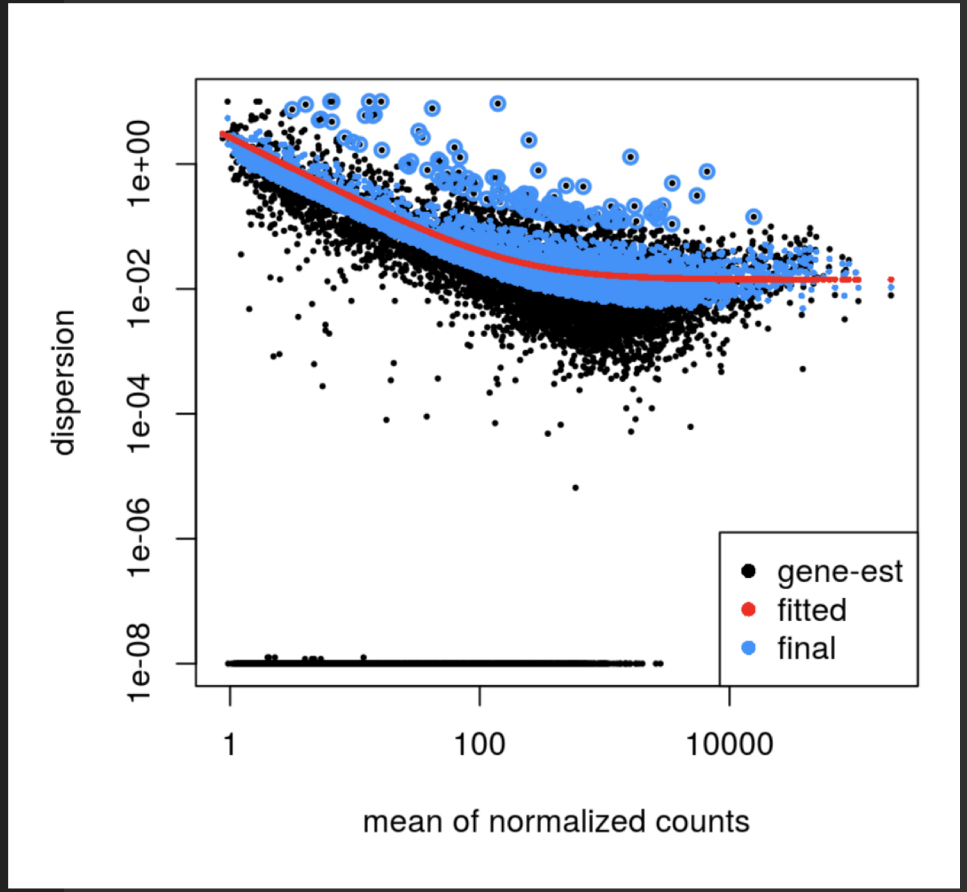
Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("BiocParallel")
```

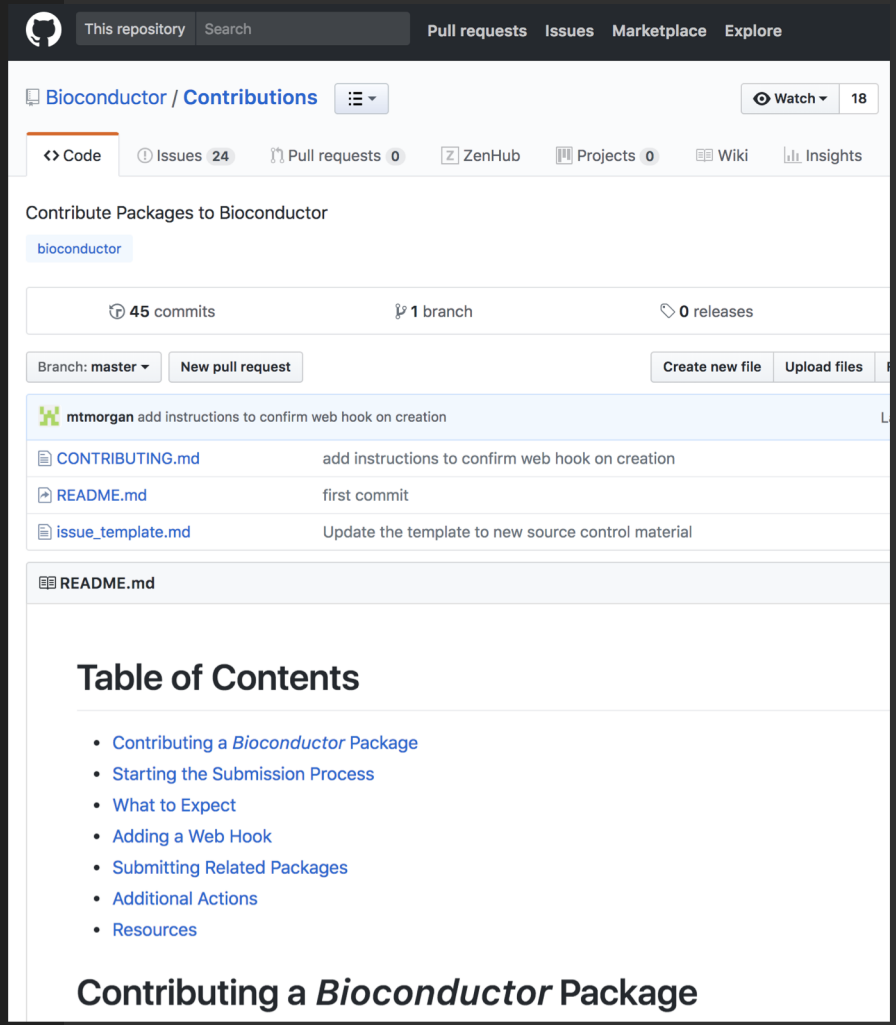
Qualities

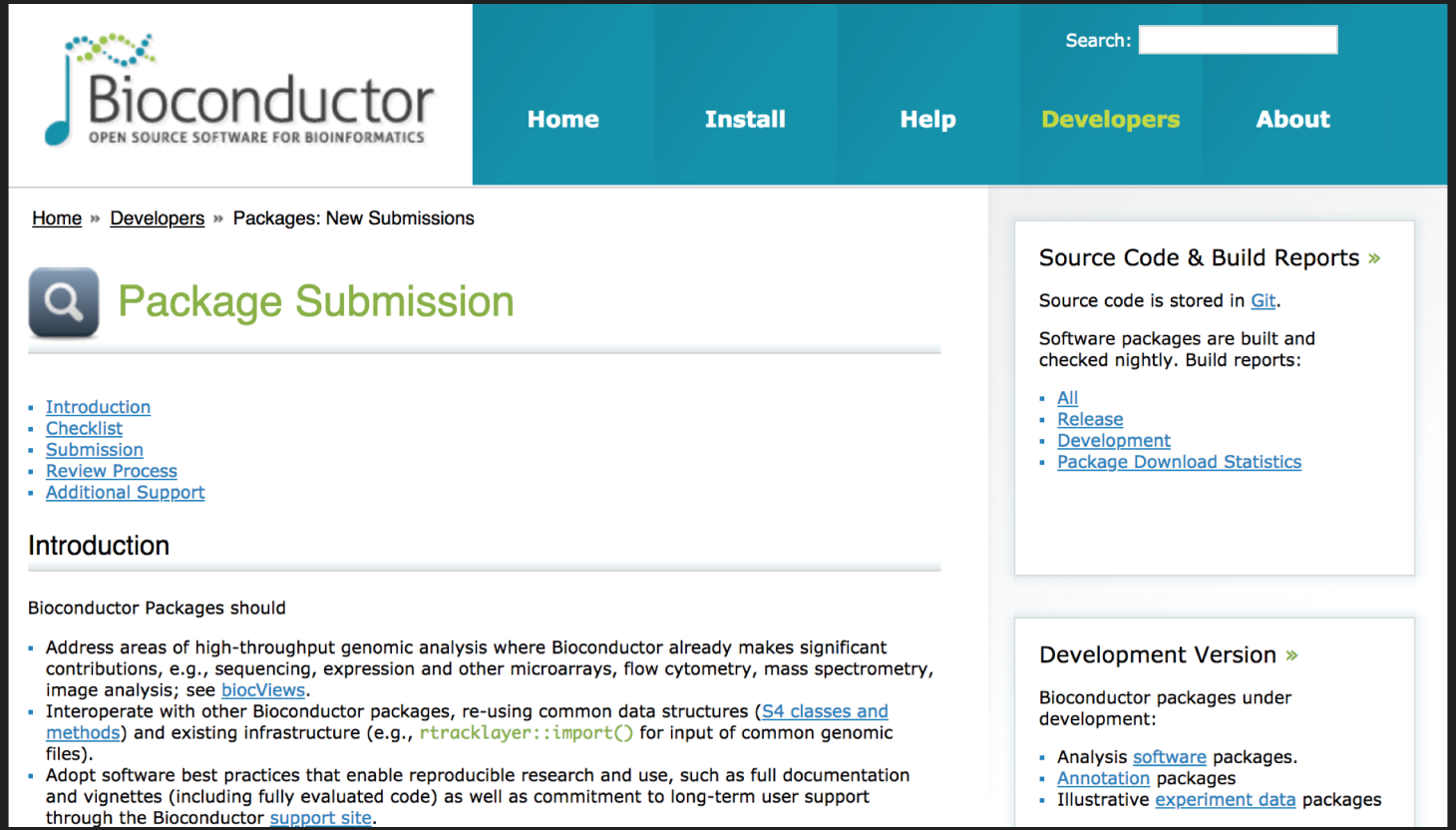
- Discoverable
- Installable
- Reliable
- Documented
- Supported
- Integrated
- Connected
- Scalable
- *State-of-the-art*
- Community-driven



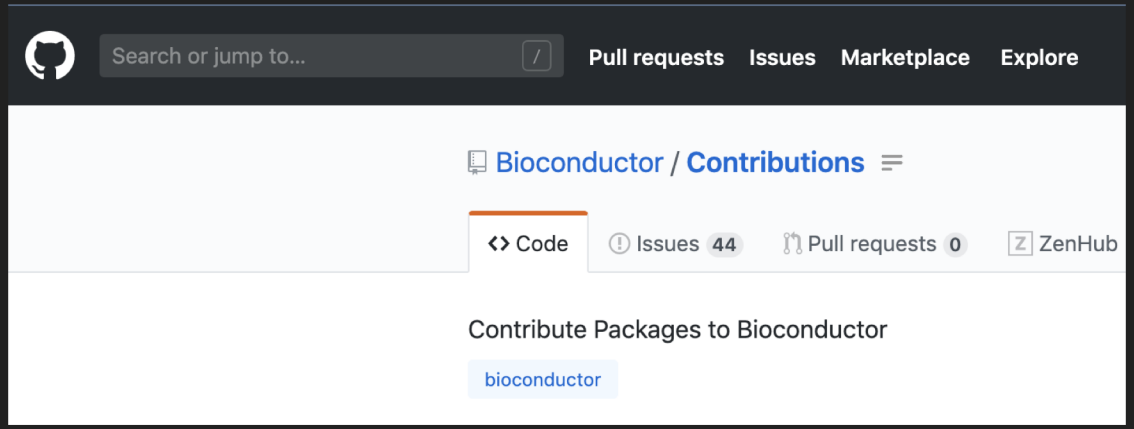
Qualities

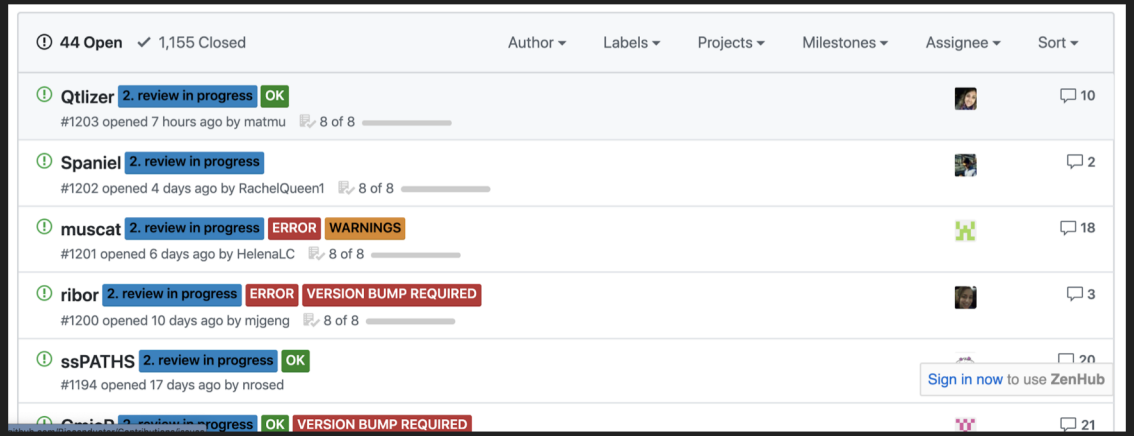
- Discoverable
- Installable
- Reliable
- Documented
- Supported
- Integrated
- Connected
- Scalable
- State-of-the-art
- *Community-driven*





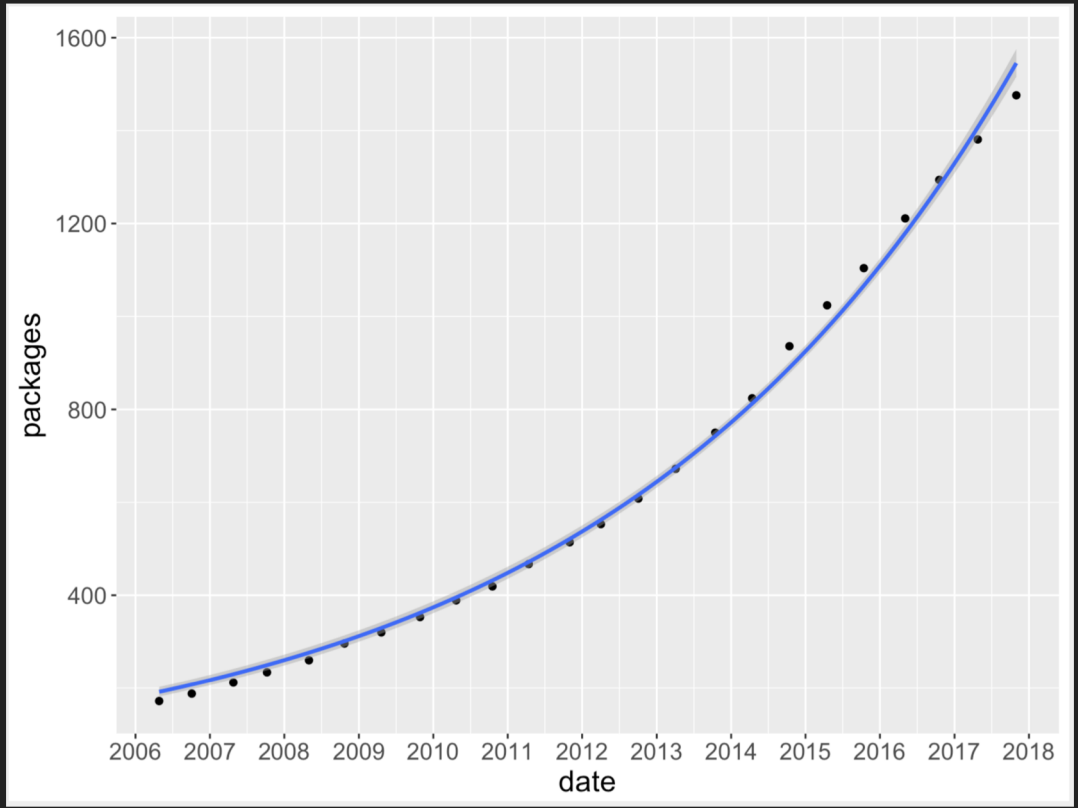
Core value: open and engaged





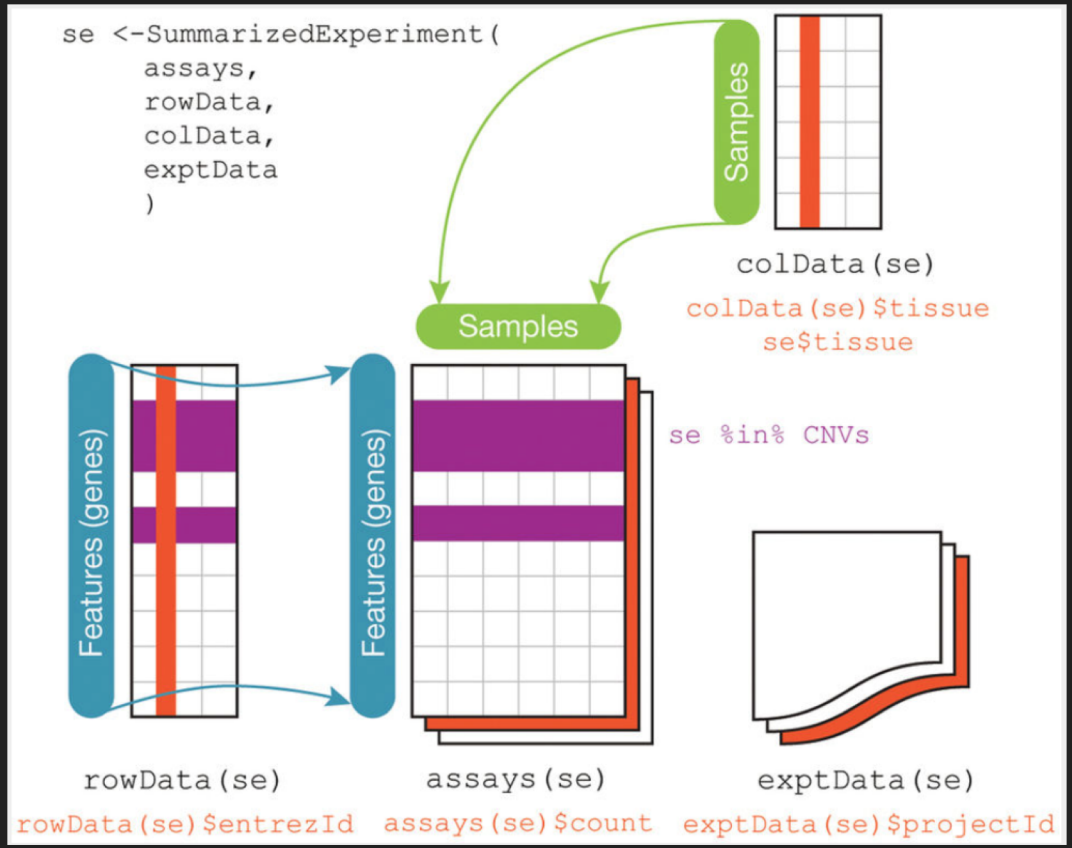
https://github.com/Bioconductor/Contributions

Bioconductor Contributions



Core Value: Reproducible research

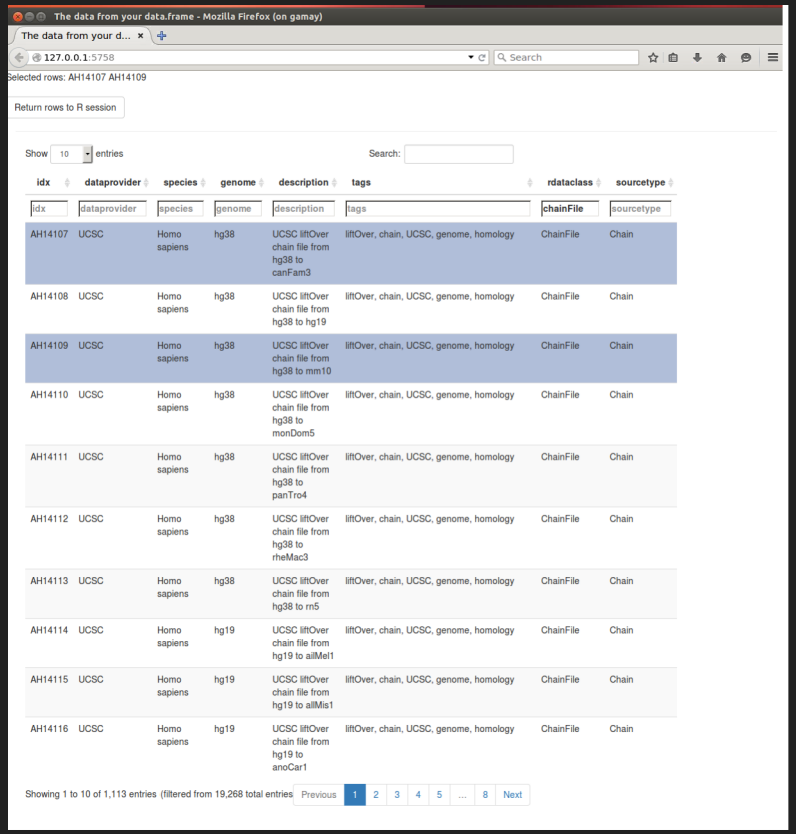
- Recognize complexity in high-throughput biological data
- Version control everything
- Continuous testing and integration
- Text-based workflow (no GUI)
- Literate programming approaches and documentation
- Education on tooling
- Numerous mechanisms for FAIR data sharing



Not just analysis packages

- AnnotationHub: More than 30,000 curated public reference datasets
  - ENCODE
  - UCSC tracks
  - Organism databases for dozens of species
- ExperimentHub: User-submitted, curated data, code, documentation
  - CuratedTCGAData
  - CuratedMetagenomicsData
- Microarray annotation resources (more than 100 platforms, standardized)
- API access to dozens of cancer and biological databases

For free: versioning, information provenance, FAIR, marketing, interoperability across the project



AnnotationHub

```
dm <- query(ah, c("ChainFile", "UCSC", "Drosophila melanogaster"))
dm
```

```
## AnnotationHub with 45 records
## # snapshotDate(): 2019-05-20
## # $dataprovider: UCSC
## # $species: Drosophila melanogaster
## # $rdataclass: ChainFile
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintainer, rdatadateadded, preparerclass,
## #   tags, rdatapath, sourceurl, sourcetype
## # retrieve records with, e.g., 'object[["AH15102"]]'
##
##            title
##   AH15102 | dm3ToAnoGam1.over.chain.gz
##   AH15103 | dm3ToApiMel3.over.chain.gz
##   AH15104 | dm3ToDm2.over.chain.gz
##   AH15105 | dm3ToDm6.over.chain.gz
##   AH15106 | dm3ToDp3.over.chain.gz
##   ...       ...
##   AH15142 | dm2ToDroVir3.over.chain.gz
##   AH15143 | dm2ToDroWil1.over.chain.gz
##   AH15144 | dm2ToDroYak1.over.chain.gz
##   AH15145 | dm2ToDroYak2.over.chain.gz
##   AH15146 | dm1ToDm2.over.chain.gz
```

```
## DataFrame with 6 rows and 5 columns
##                          title dataprovider              species
##                    <character>  <character>            <character>
## AH15102 dm3ToAnoGam1.over.chain.gz         UCSC Drosophila melanogaster
## AH15103 dm3ToApiMel3.over.chain.gz         UCSC Drosophila melanogaster
## AH15104     dm3ToDm2.over.chain.gz         UCSC Drosophila melanogaster
## AH15105     dm3ToDm6.over.chain.gz         UCSC Drosophila melanogaster
## AH15106     dm3ToDp3.over.chain.gz         UCSC Drosophila melanogaster
## AH15107     dm3ToDp4.over.chain.gz         UCSC Drosophila melanogaster
##         taxonomyid      genome
##          <integer> <character>
## AH15102       7227         dm3
## AH15103       7227         dm3
## AH15104       7227         dm3
## AH15105       7227         dm3
## AH15106       7227         dm3
## AH15107       7227         dm3
```

```
eh
```

```
## ExperimentHub with 2223 records
## # snapshotDate(): 2019-04-29
## # $dataprovider: Eli and Edythe L. Broad Institute of Harvard and MIT, NA...
## # $species: Homo sapiens, Mus musculus, Saccharomyces cerevisiae, human g...
## # $rdataclass: ExpressionSet, SummarizedExperiment, SummarizedBenchmark, ...
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintain
## #   tags, rdatapath, sourceurl,
## # retrieve records with, e.g., '
##
##            title
##   EH1    | RNA-Sequencing and cl
##   EH166  | ERR188297
##   EH167  | ERR188088
##   EH168  | ERR188204
##   EH169  | ERR188317
##   ...      ...
##   EH2539 | LAML_GISTIC_Peaks-201
##   EH2540 | LAML_GISTIC_Threshold
##   EH2541 | SKCM_GISTIC_AllByGene
##   EH2542 | SKCM_GISTIC_Peaks-201
##   EH2543 | SKCM_GISTIC_Threshold
```

```
head(unique(eh$dataprovider))
```

```
## [1] "GEO"
## [2] "GEUVADIS"
## [3] "Allen Brain Atlas"
## [4] "ArrayExpress"
## [5] "Department of Psychology, Abdul Haq Campus, Federal Urdu University for
Arts, Science and Technology, Karachi, Pakistan. shahiq_psy@yahoo.com"
## [6] "Department of Chemical and Biological Engineering, Chalmers University
of Technology, SE-412 96 Gothenburg, Sweden."
```

In the same way, you can also see data from different species inside the hub by looking at the contents of species like this:

```
head(unique(eh$species))
```

```
## [1] "Homo sapiens"            "Mus musculus"
## [3] "Mus musculus (E18 mice)" NA
## [5] "Rattus norvegicus"       "human gut metagenome"
```

ExperimentHub

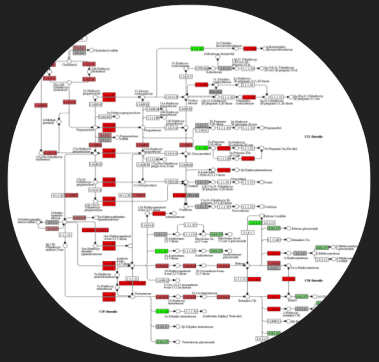
Creating An ExperimentHub Package

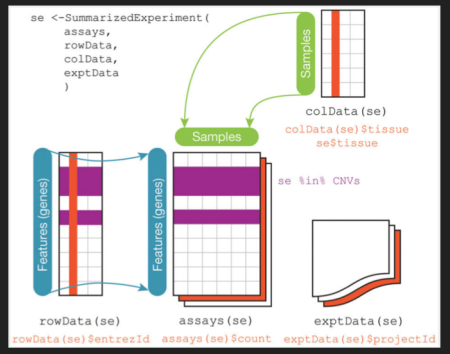
Valerie Obenchain and Lori Shepherd

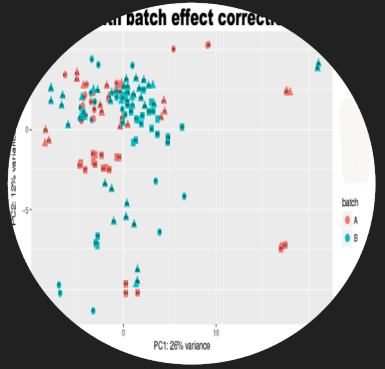
Modified: November 2017. Compiled: 02 May 2019

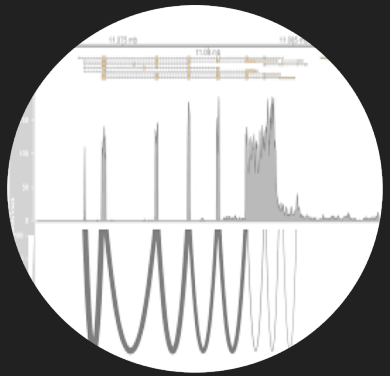
Contents

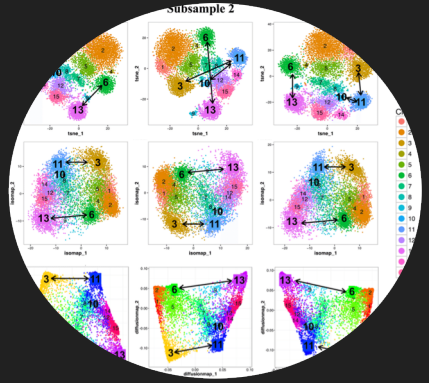
Core value: reuse and interoperability

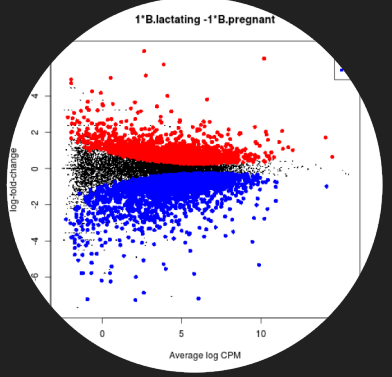


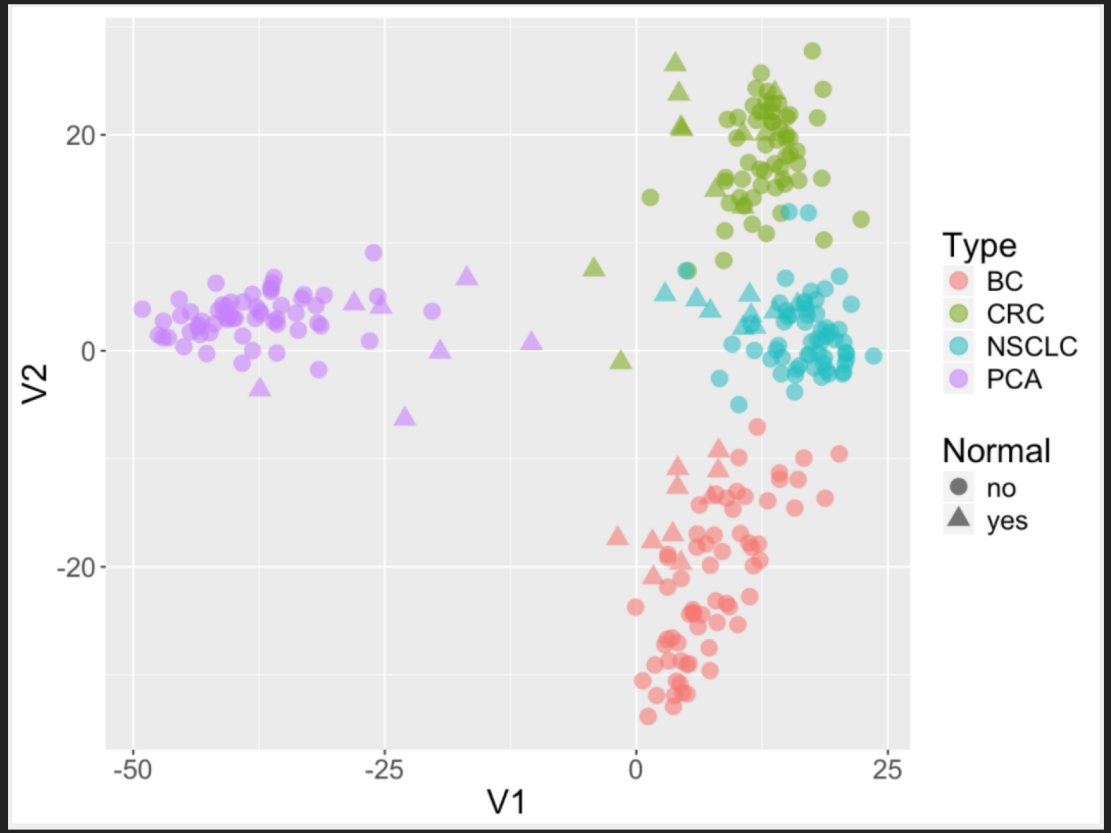












Example: PCA plot from public data (10 lines of code)

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103512

Filter gene expression by variance to find most informative genes.

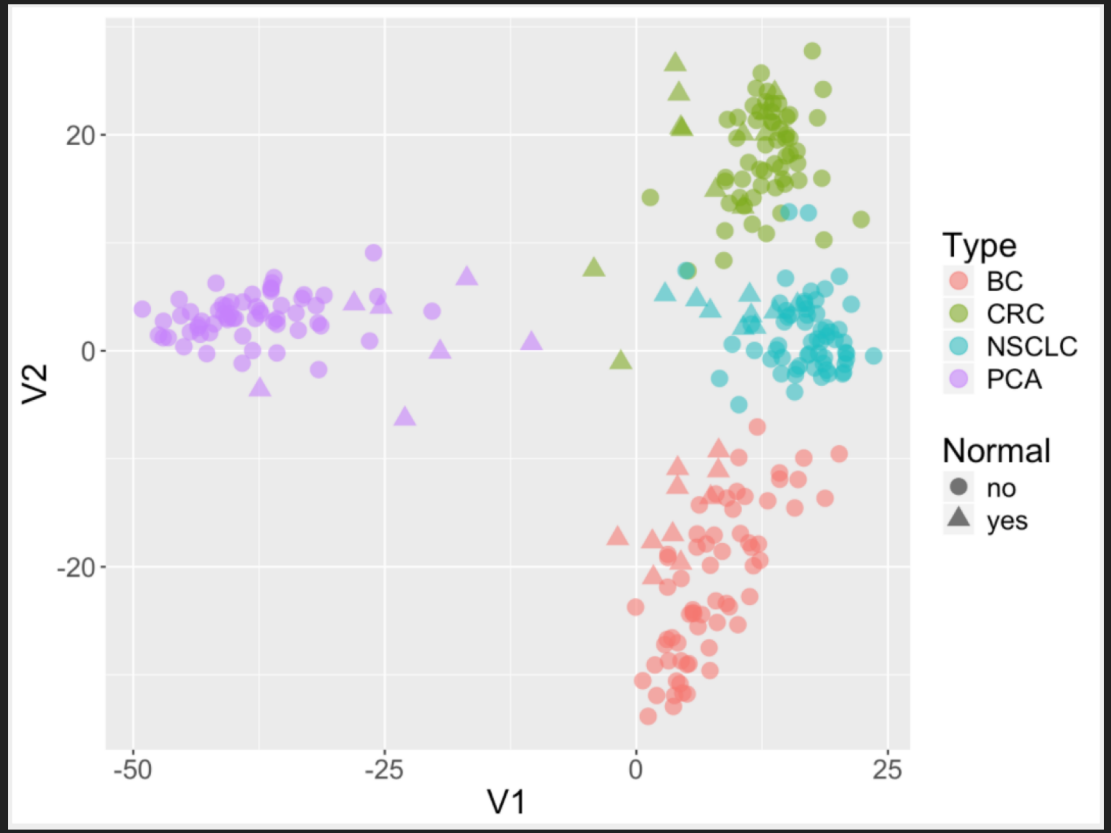
```
sds = apply(exprs(gse),1,sd)
dat = exprs(gse)[order(sds,decreasing = TRUE)[1:500],]
```

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'])
```

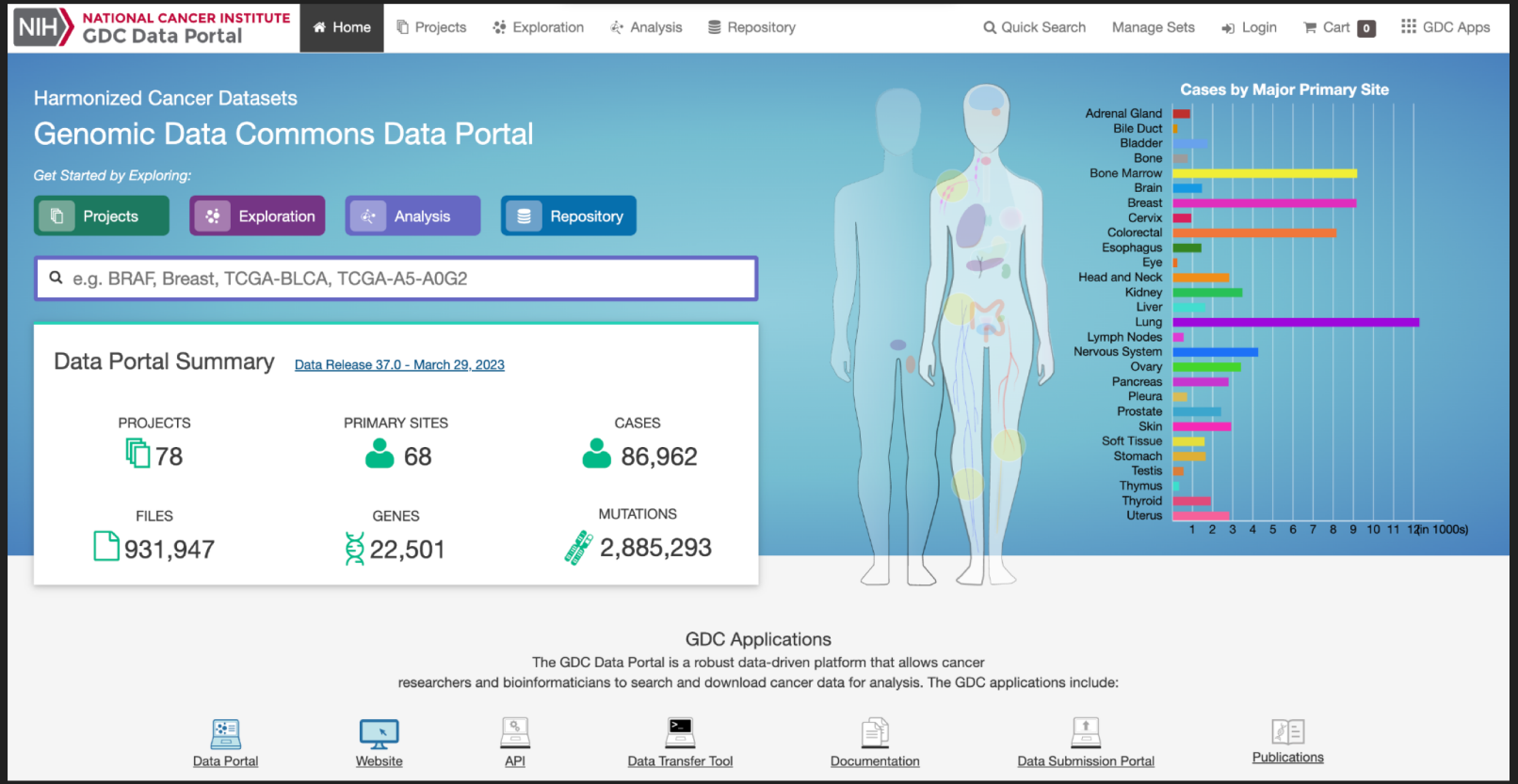
And do the plot.

```
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))
```

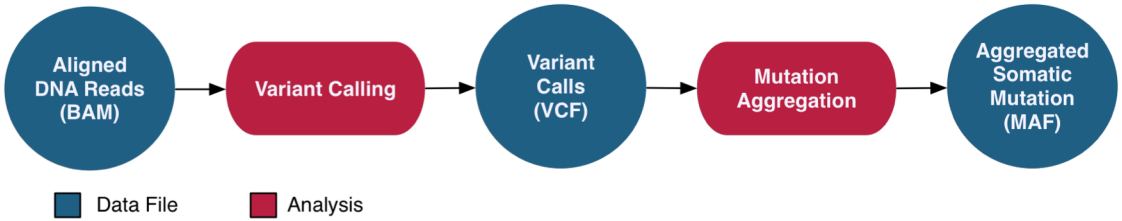


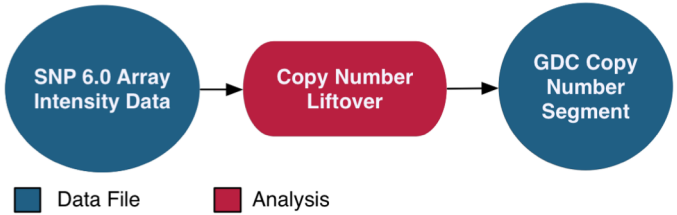
Example: PCA plot from public data (10 lines of code)

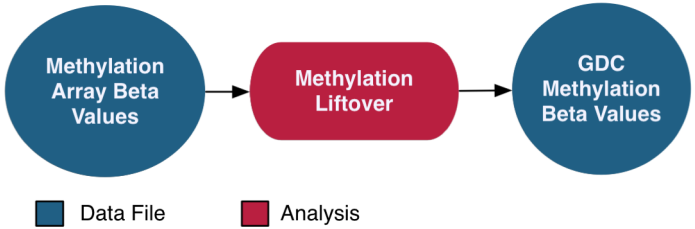
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103512



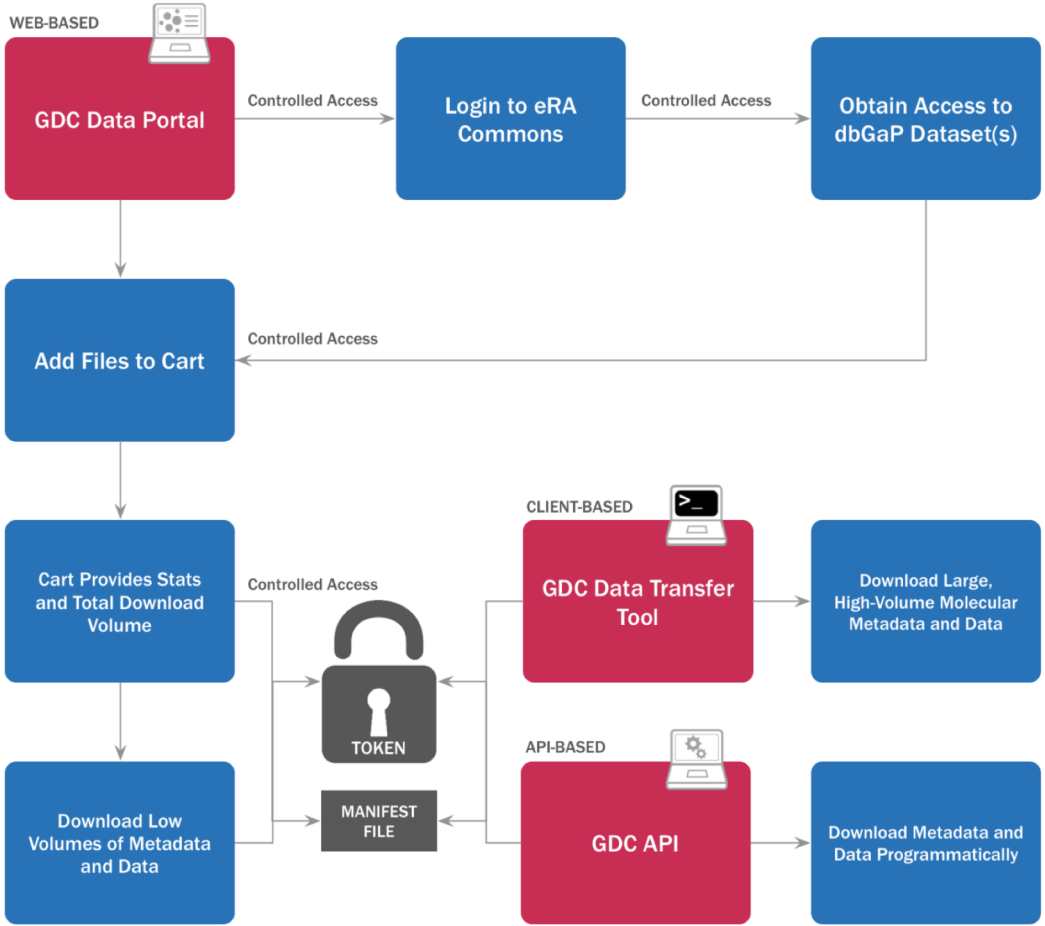
GDC Data Portal





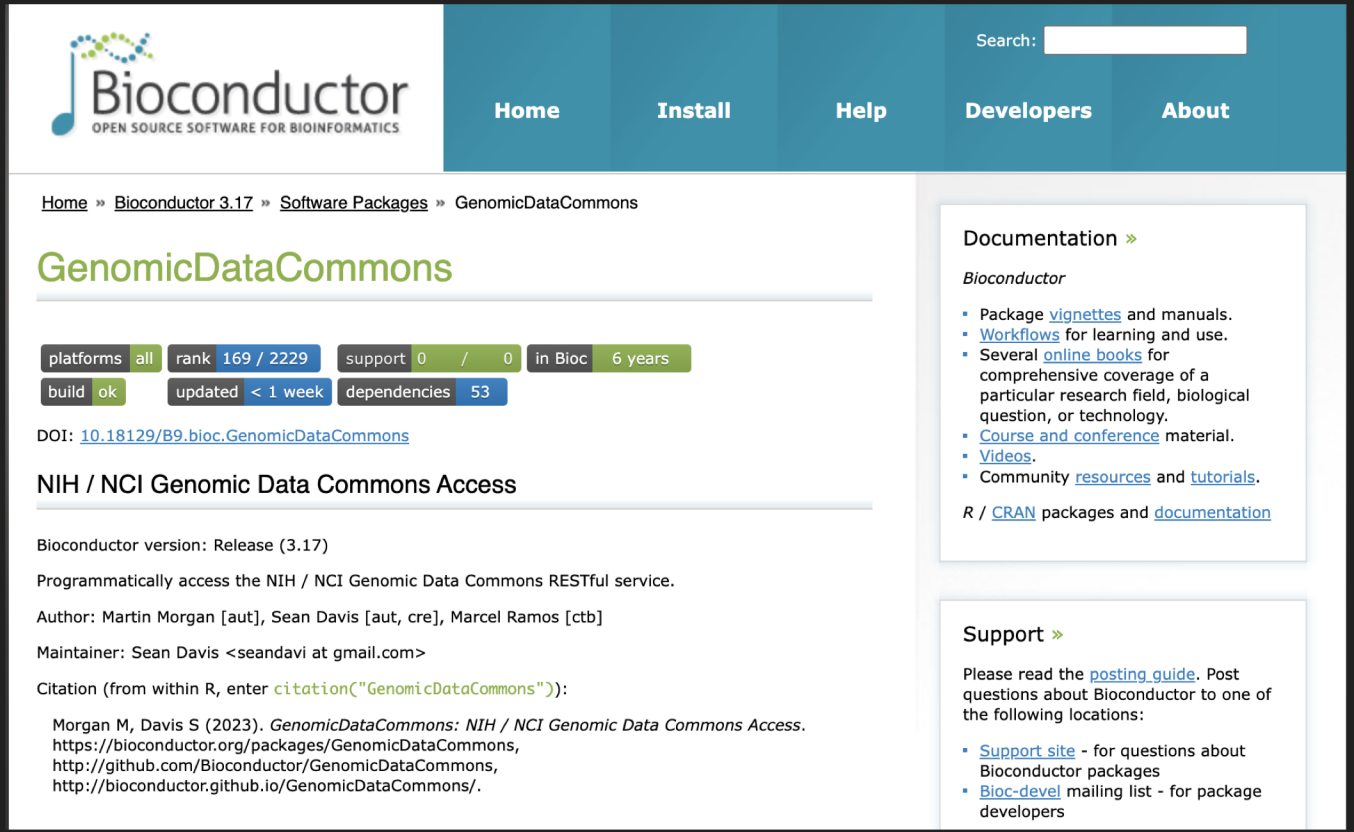


https://gdc.cancer.gov/about-data/data-harmonization-and-generation/genomic-data-harmonization-0#Pipelines


WEB-BASED
GDC Data Portal
Controlled Access
Login to eRA Commons
Controlled Access
Obtain Access to dbGaP Dataset(s)
Add Files to Cart
Controlled Access
Cart Provides Stats and Total Download Volume
Controlled Access
CLIENT-BASED
GDC Data Transfer Tool
Download Large, High-Volume Molecular Metadata and Data
TOKEN
API-BASED
Download Low Volumes of Metadata and Data
MANIFEST FILE
GDC API
Download Metadata and Data Programmatically

GDC Data Model

Complex, flexible, and growing….



GenomicDataCommons

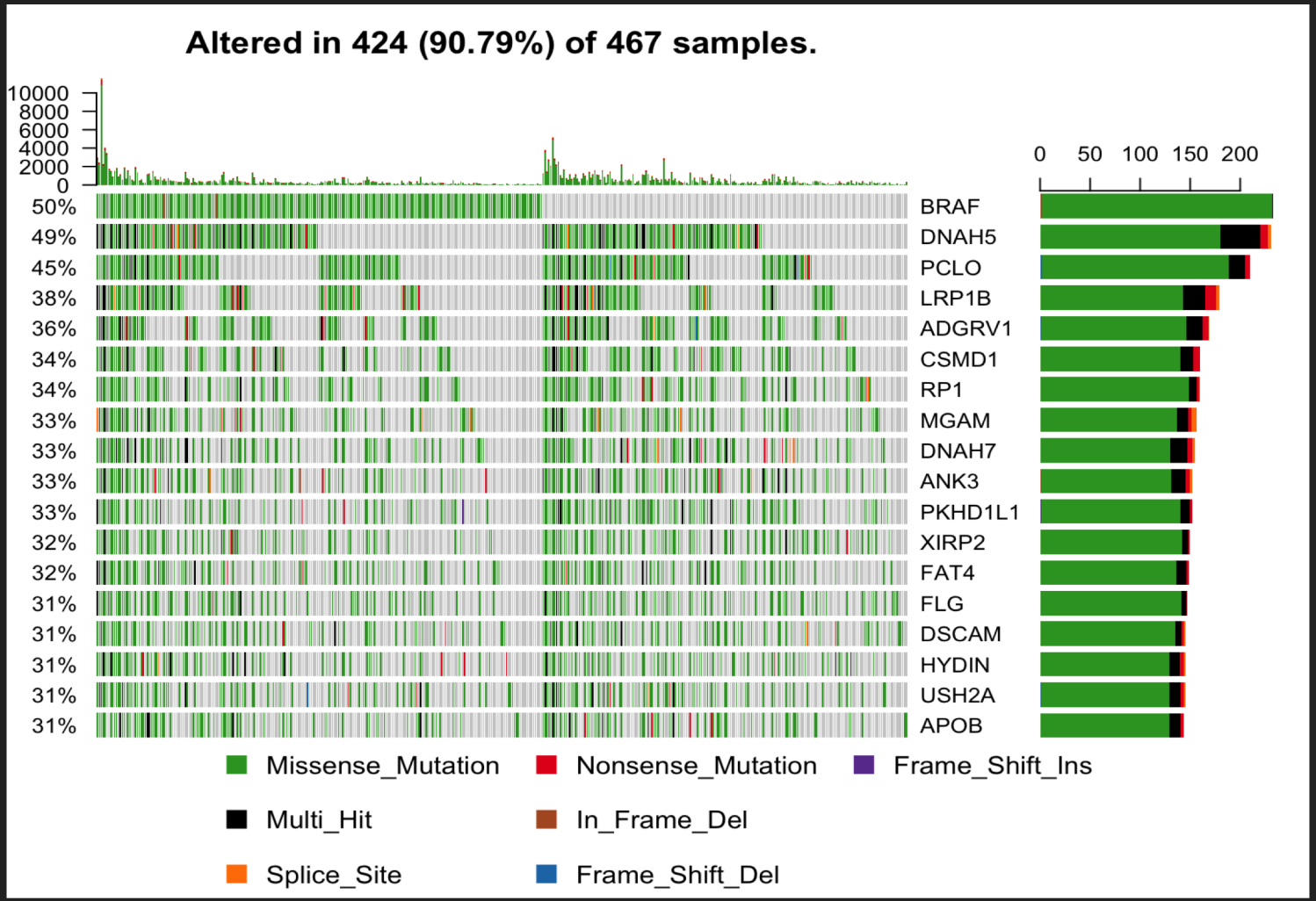
Goal: Explore somatic variants seen in the TCGA cutaneous melanoma cohort in a reproducible, reusable way building on Bioconductor tools.

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])
```

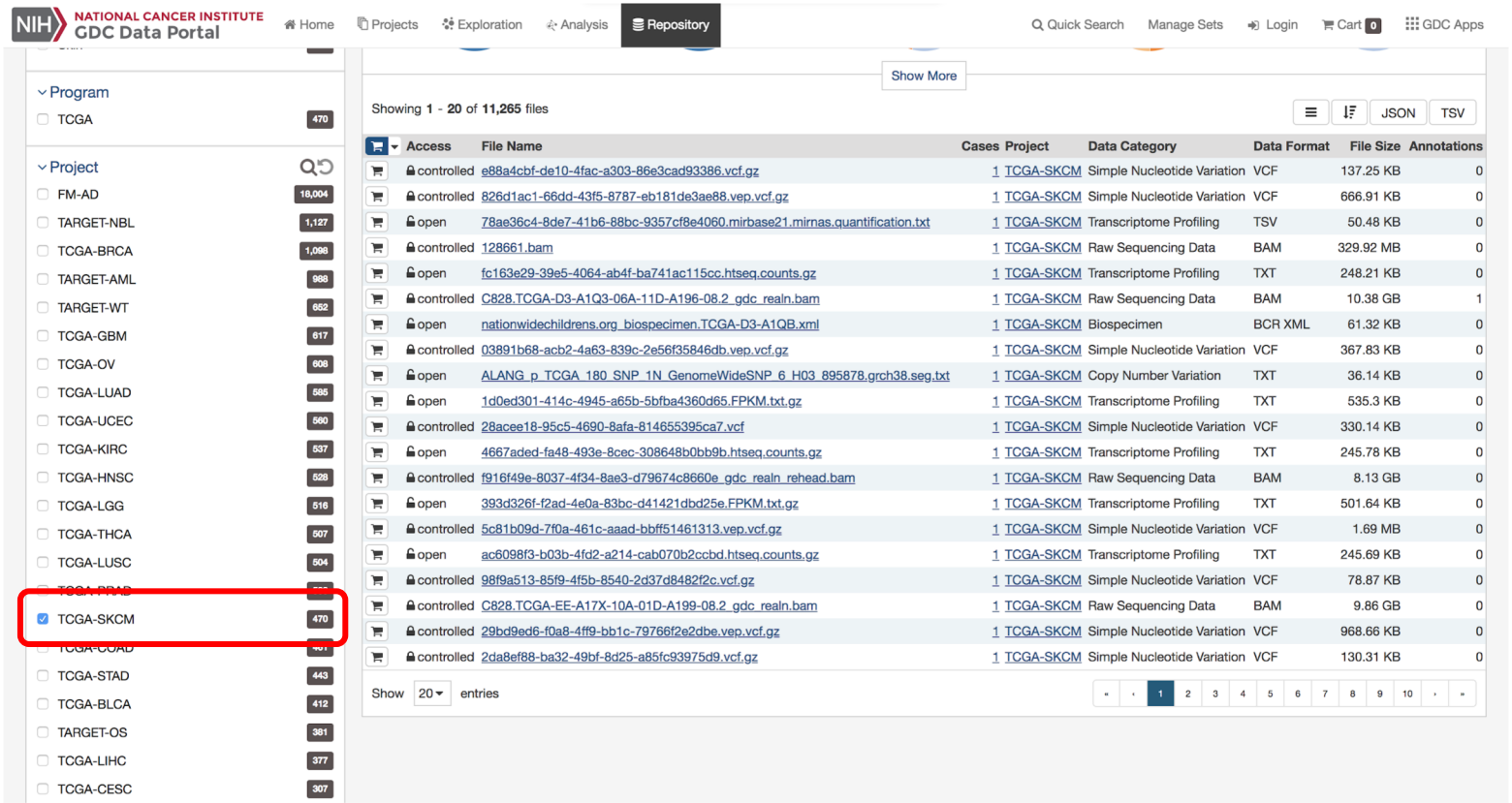

Altered in 424 (90.79%) of 467 samples.
10000
8000
6000
4000
2000
0
0
50
100
150
200
50%
49%
45%
38%
36%
34%
34%
33%
33%
33%
33%
32%
32%
31%
31%
31%
31%
BRAF
DNAH5
PCLO
LRP1B
ADGRV1
CSMD1
RP1
MGAM
DNAH7
ANK3
PKHD1L1
XIRP2
FAT4
FLG
DSCAM
HYDIN
USH2A
APOB
Missense_Mutation
Nonsense_Mutation
Frame_Shift_Ins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

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])
```



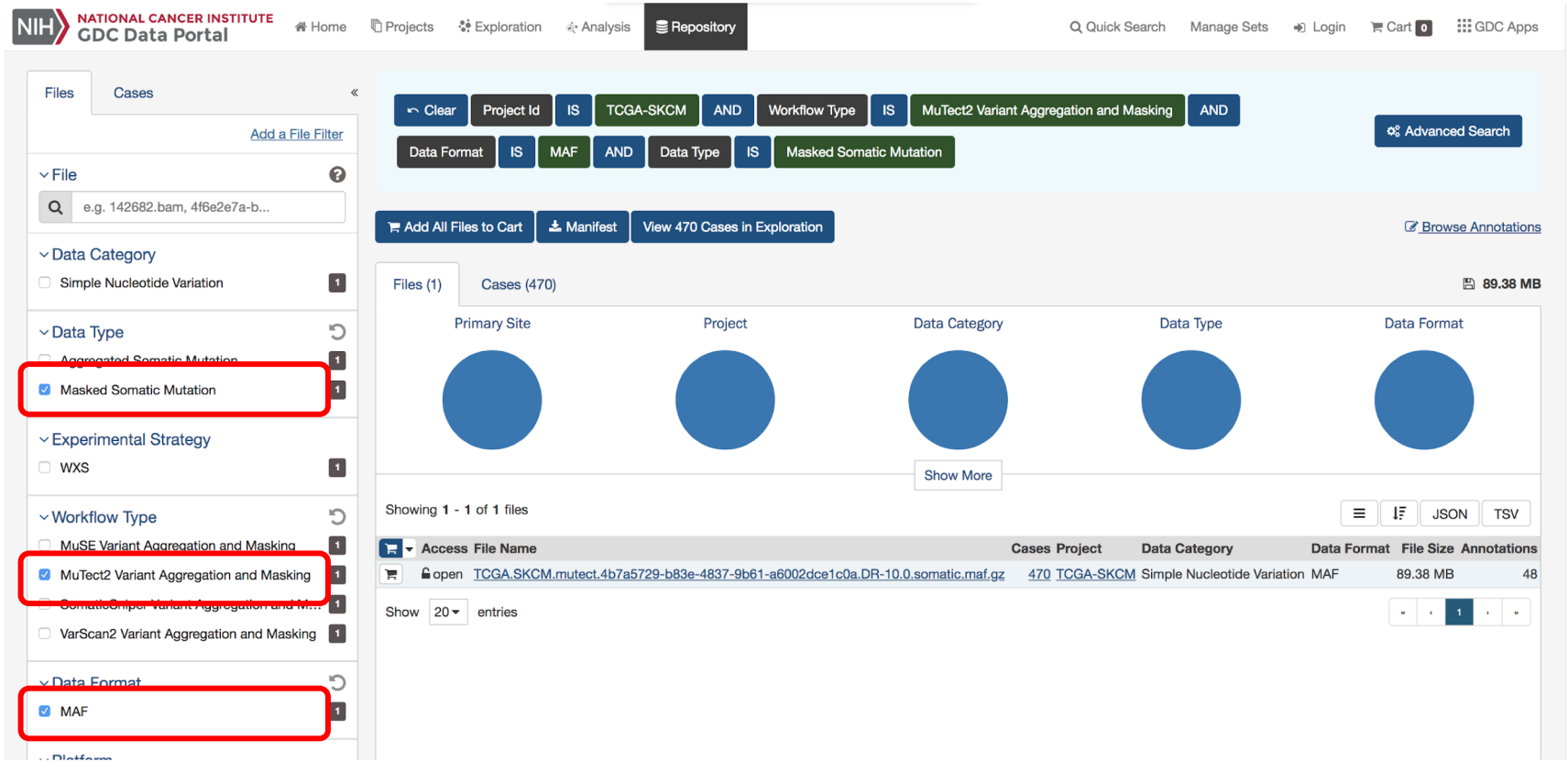
Choose patients based on project

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])
```



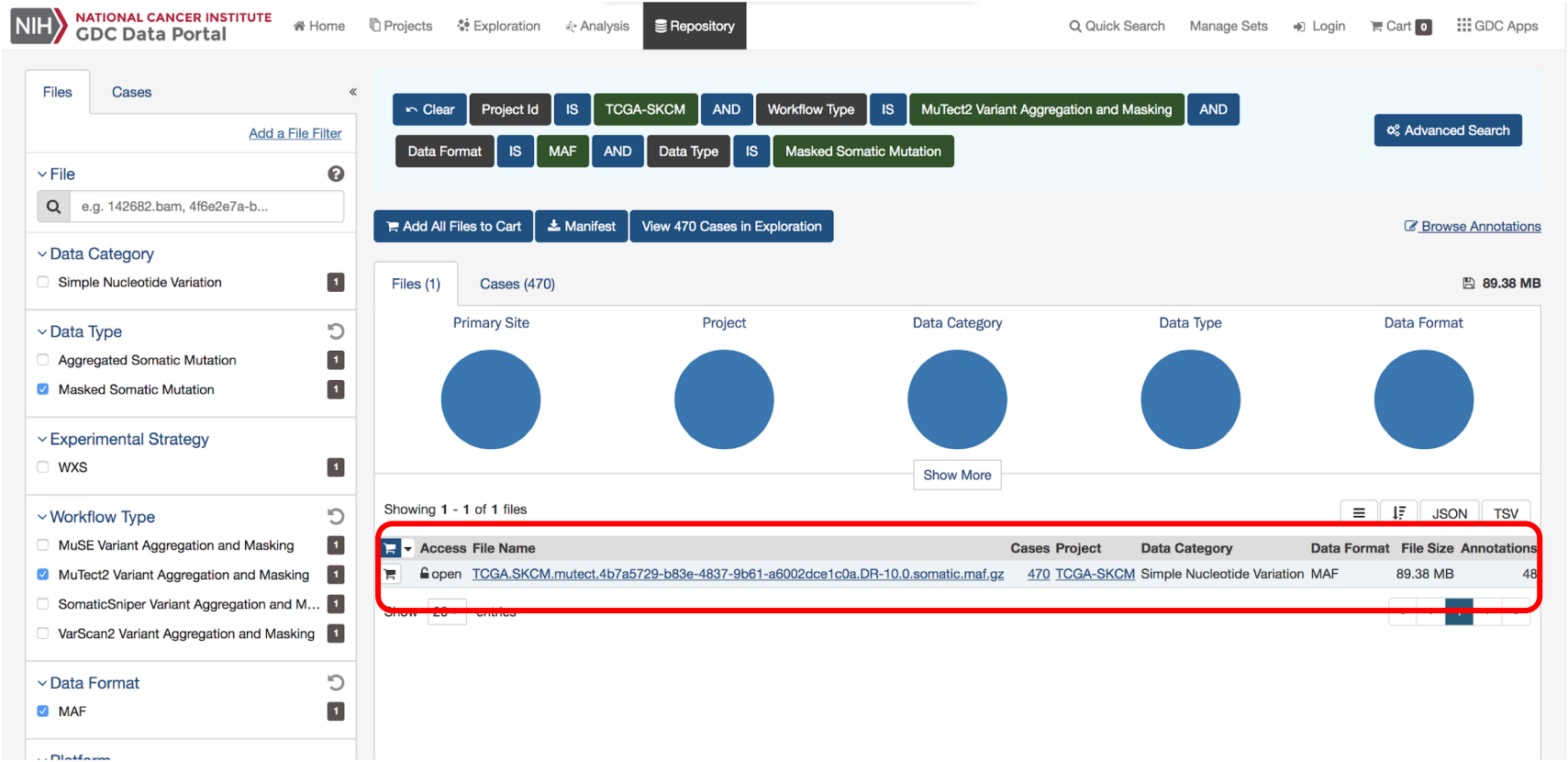
Choose Data Type and Workflow to select files

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])
```



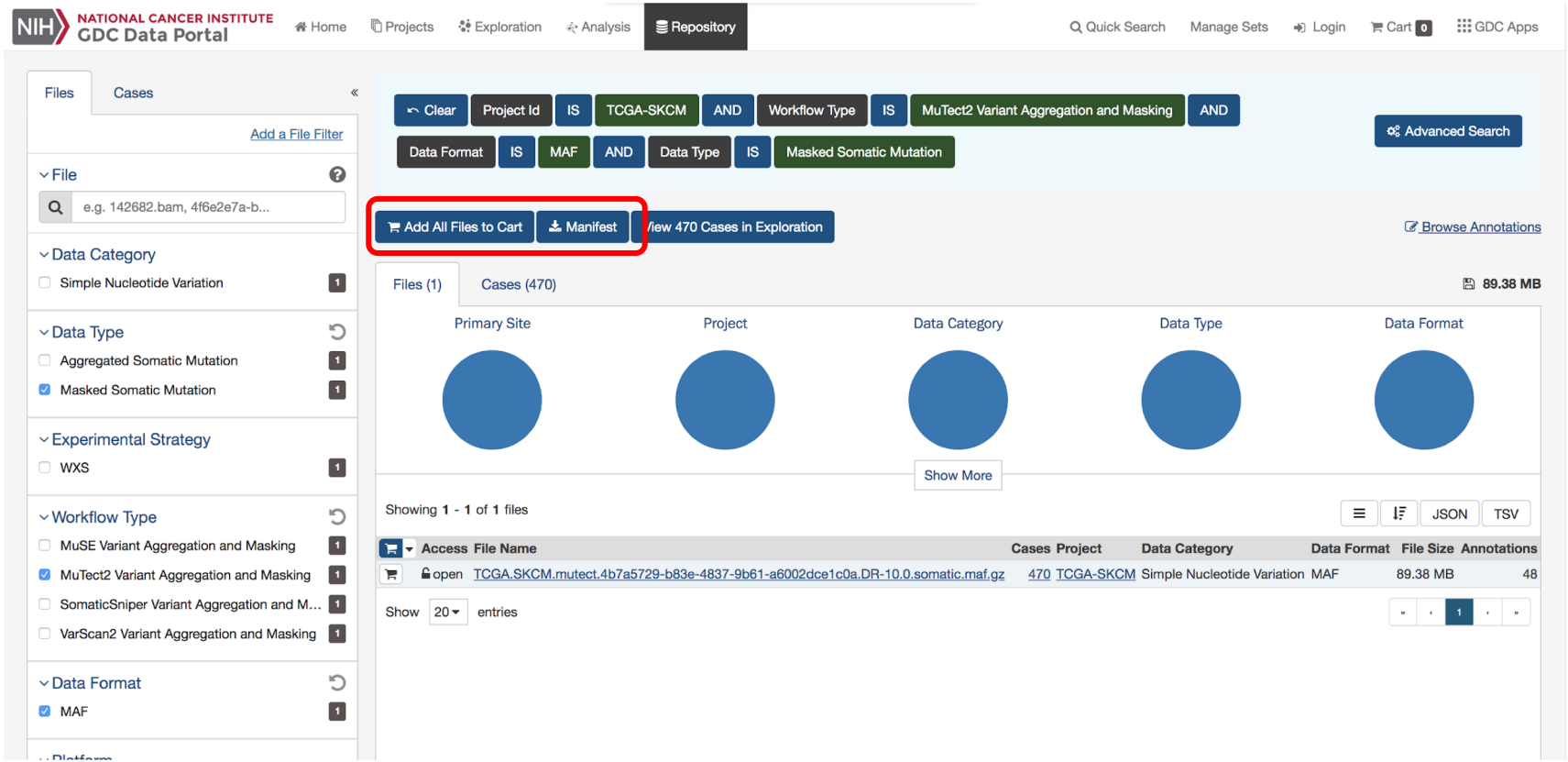
Files, like all entities in the GDC, have an associated UUID

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])
```

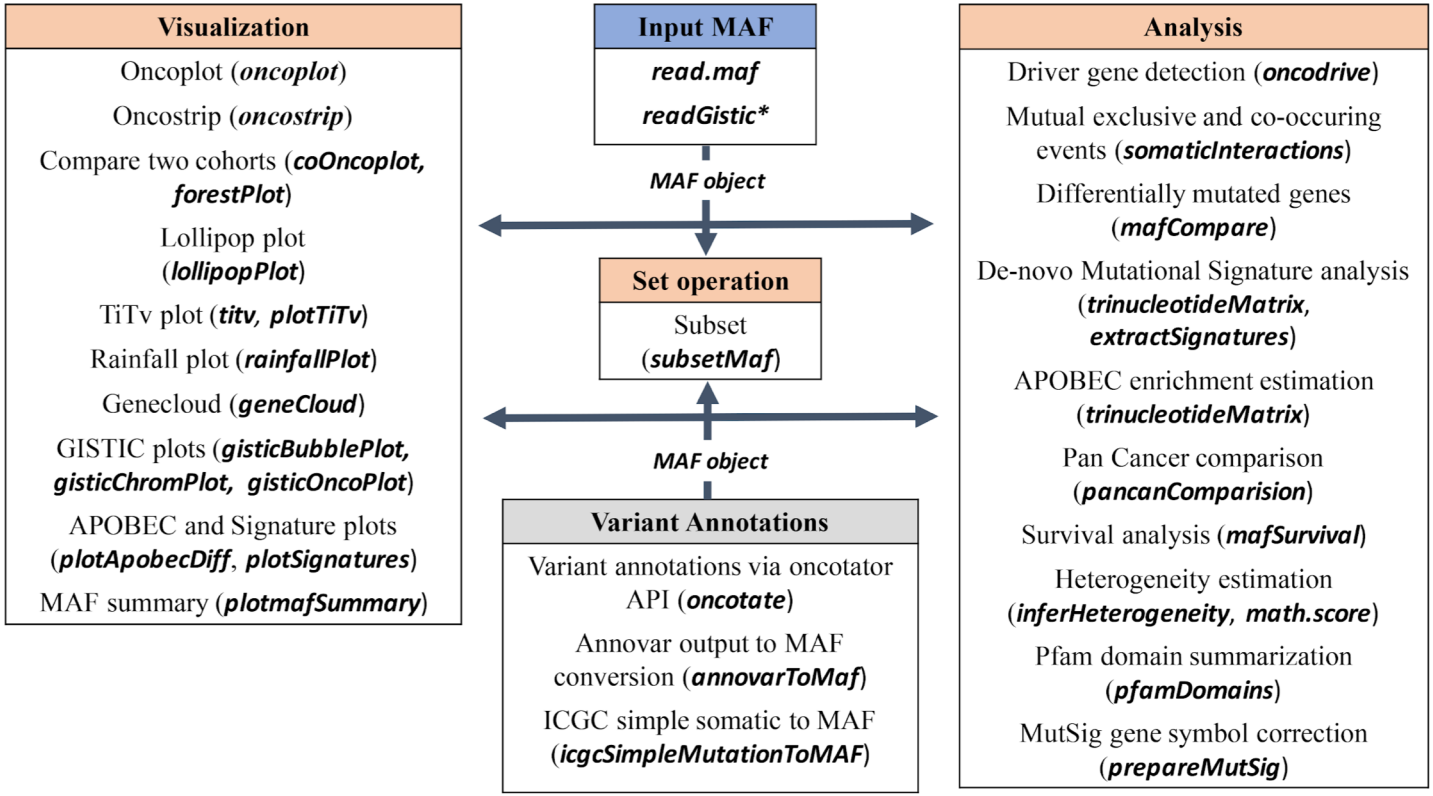

NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Quick Search
Manage Sets
Login
Cart 0
GDC Apps
Files
Cases
Add a File Filter
File
e.g. 142682.bam, 4f6e2e7a-b...
Data Category
Simple Nucleotide Variation 1
Data Type
Aggregated Somatic Mutation 1
Masked Somatic Mutation 1
Experimental Strategy
WXS 1
Workflow Type
MuSE Variant Aggregation and Masking 1
MuTect2 Variant Aggregation and Masking 1
SomaticSniper Variant Aggregation and M... 1
VarScan2 Variant Aggregation and Masking 1
Data Format
MAF 1
Clear
Project Id IS TCGA-SKCM AND Workflow Type IS MuTect2 Variant Aggregation and Masking AND
Data Format IS MAF AND Data Type IS Masked Somatic Mutation
Advanced Search
Add All Files to Cart
Manifest
View 470 Cases in Exploration
Browse Annotations
Files (1)
Cases (470)
89.38 MB
Primary Site
Project
Data Category
Data Type
Data Format
Show More
Showing 1 - 1 of 1 files
JSON
TSV
Access File Name Cases Project Data Category Data Format File Size Annotations
open TCGA.SKCM.mutect.4b7a5729-b83e-4837-9b61-a6002dce1c0a.DR-10.0.somatic.maf.gz 470 TCGA-SKCM Simple Nucleotide Variation MAF 89.38 MB 48
Show 20 entries

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])
```



GDC Programmatic access example:

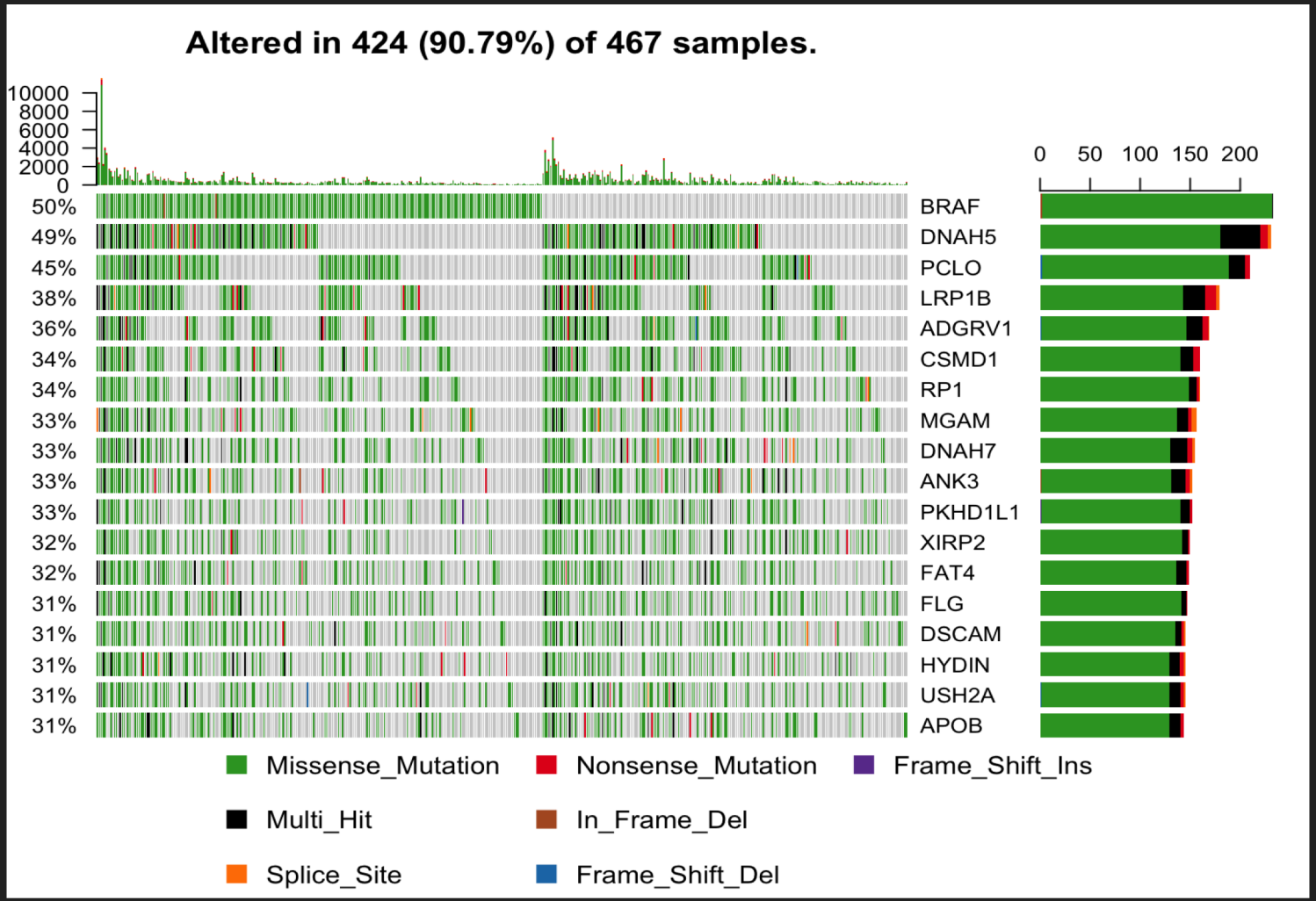
Somatic profiles from TCGA melanoma samples (8 lines of code)

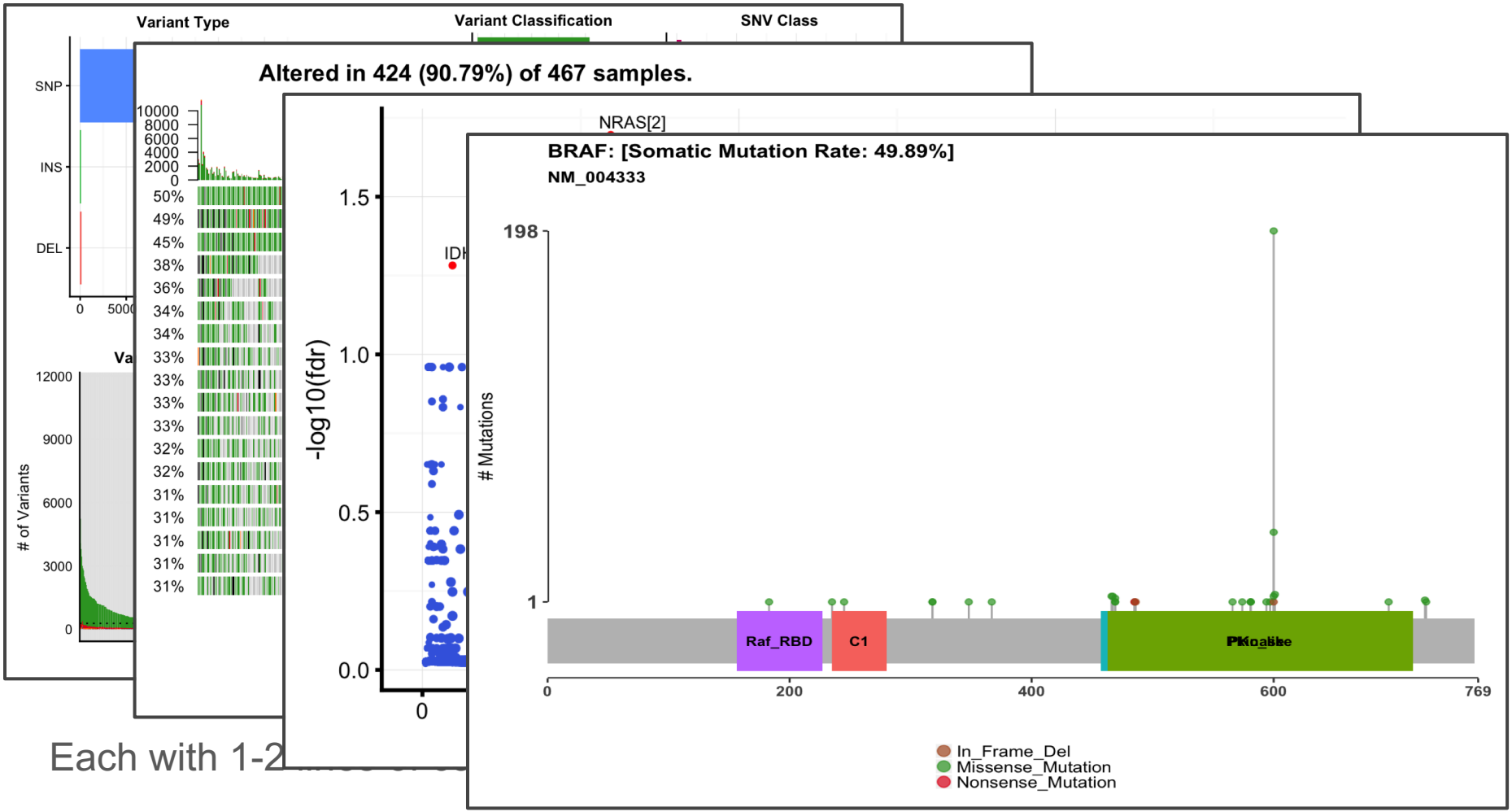
Use the GenomicDataCommons package to find and download variants from the TCGA cutaneous melanoma dataset.

```
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%
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


Variant Type
Variant Classification
SNV Class
SNP
INS
DEL
Altered in 424 (90.79%) of 467 samples.
NRAS[2]
-log10(fdr)
Mutations
BRAF: [Somatic Mutation Rate: 49.89%]
NM_004333
198
1
Raf_RBD
C1
0
200
400
600
769
In_Frame_Del
Missense_Mutation
Nonsense_Mutation

Each with 1-2

Check for updates

SOFTWARE TOOL ARTICLE

REVISED *TCGA Workflow*: Analyze cancer genomics and epigenomics data using Bioconductor packages [version 2; referees: 1 approved, 2 approved with reservations]

Tiago C. Silva[1,2*], Antonio Colaprico[3,4*], Catharina Olsen[3,4*], Fulvio D'Angelo[5,6*], Gianluca Bontempi[3-5], Michele Ceccarelli[7], Houtan Noushmehr [1,8]

* Equal contributors

Author details

Grant information

This article is included in the Bioconductor gateway .

Abstract

Biotechnological advances in sequencing have led to an explosion of publicly available data via large international consortia such as The Cancer Genome Atlas (TCGA), The Encyclopedia of DNA Elements (ENCODE), and The NIH Roadmap Epigenomics Mapping Consortium (Roadmap). These projects have provided unprecedented opportunities to interrogate the epigenome of cultured cancer cell lines as well as normal and tumor tissues with high genomic resolution. The Bioconductor project offers more than 1,000 open-source software and statistical packages to analyze high-throughput genomic data. However, most packages are designed for specific data types (e.g. expression, epigenetics, genomics) and there is no one comprehensive tool that provides a complete integrative analysis of the resources and data provided by all three public projects. A need to create an integration of these different analyses was recently proposed. In this workflow, we provide a series of biologically focused integrative analyses of different molecular data. We describe how to download, process and prepare TCGA data and by harnessing several key Bioconductor packages, we describe how to extract biologically meaningful genomic and epigenomic data. Using Roadmap and ENCODE data, we provide a work plan to identify biologically relevant functional epigenomic elements associated with cancer. To illustrate our workflow, we analyzed two types of brain tumors: low-grade glioma (LGG) versus high-grade glioma (glioblastoma multiform or GBM). This workflow introduces the following Bioconductor packages: AnnotationHub, ChIPSeeker, ComplexHeatmap, pathview, ELMER, GAIA, MINET,

Open Peer Review

Referee Status: ? ✓ ?

Version(s)	Invited Referees 1	2	3
REVISED Version 2 published 28 Dec 2016			? read report
Version 1 published 29 Jun 2016	? read report	✓ read report	

1 **Kyle Ellrott**, Oregon Health & Science University, USA
2 **Elena Papaleo**, Danish Cancer Society Research Center, Denmark
3 **Charlotte Soneson** , University of Zurich (UZH), Switzerland

All reports (3), Responses and comments (3)

Comments on this article

All comments (0)

Add a Comment


GEO
Gene Expression Omnibus

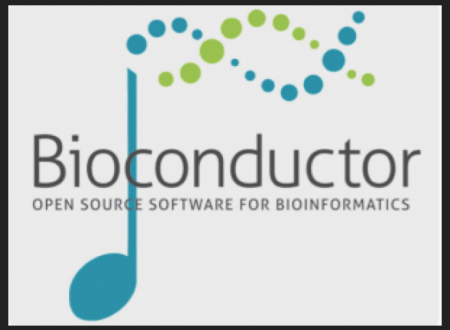

PHARMACODB
MINE MULTIPLE CANCER PHARMACOGENOMIC DATASETS


NCBI


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal

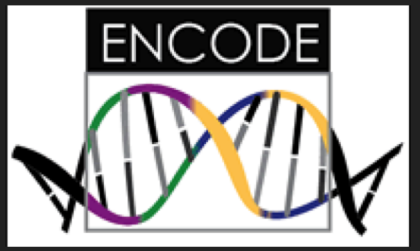

ICGC


DGIdb
THE DRUG GENE INTERACTION DATABASE

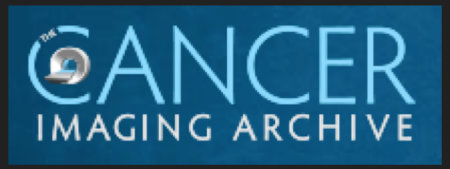

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS


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

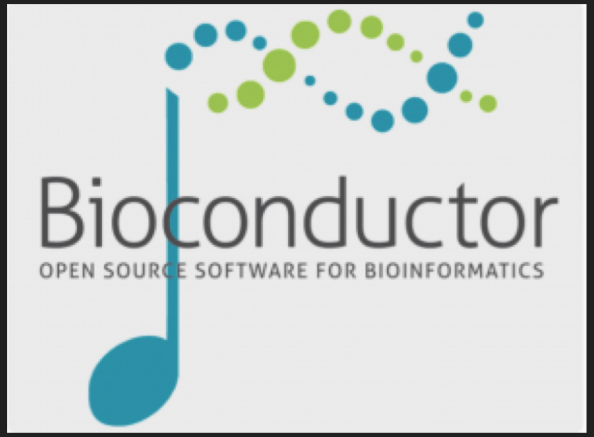

CCLE
Cancer Cell Line
Encyclopedia


amazon
web services


Microsoft Azure


APACHE
Spark


SQL
Google Cloud Platform


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

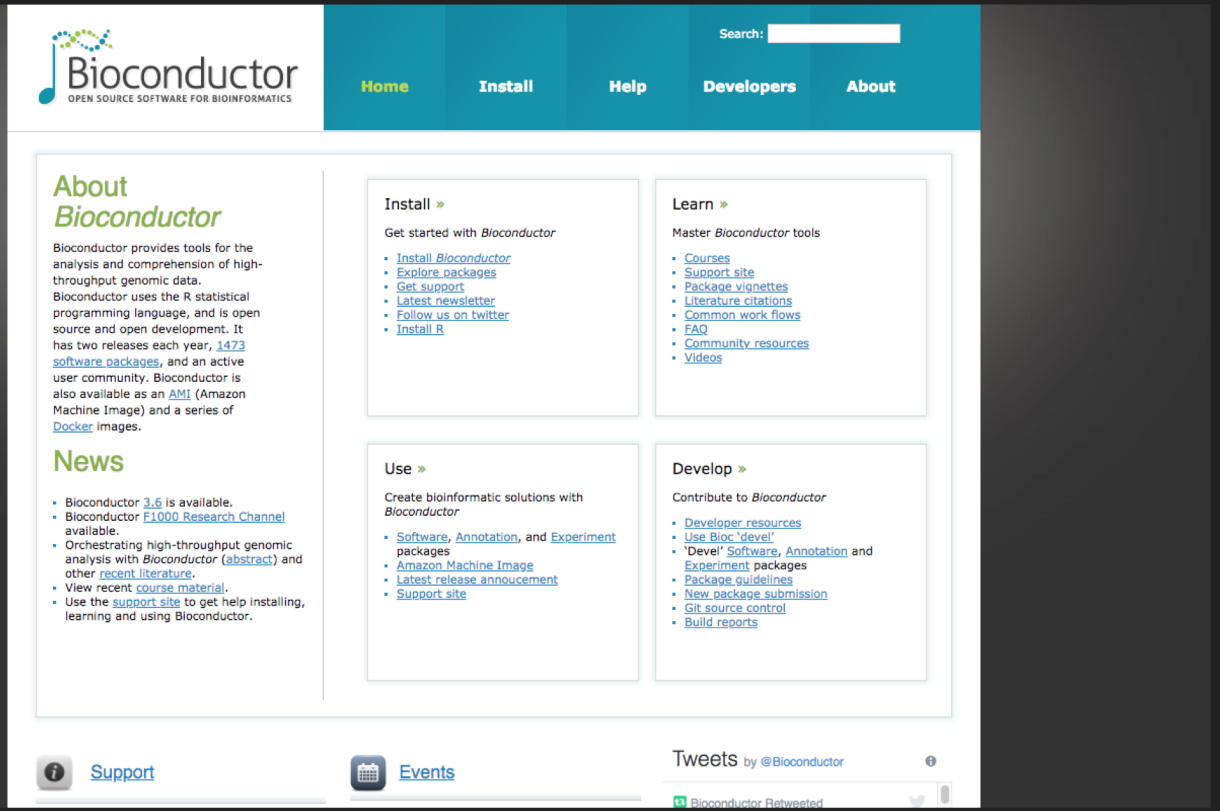

R


TensorFlow


GitHub


rmarkdown
Shiny
tidyr
knitr
ggplot2

Bioconductor: Education, Training, and Community



Core infrastructure

Community contribution

Questions?

https://bioconductor.org

https://seandavi.github.io