

# Bioconductor

*Software for orchestrating high-throughput biological data analysis*

*Sean Davis, MD, PhD*

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

<https://seandavi.github.io>

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

**Jeffrey T. Leek<sup>a,1</sup> and Roger D. Peng<sup>b</sup>**

<sup>a</sup>Associate Professor of Biostatistics and Oncology and <sup>b</sup>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

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

a4Base a4Classif a4Core a4Preproc a4Reporting ABAEnrichment ABAEnrichmentArray ABSSeq acde aCGH ACME AdACGH2 adSplit affxparser affycomp AffyCompatible affyContam affycoretools AffyExpress affyILM affyio affyImGUI affyPara affypdnn affyPLM affyQCReport affyRNAdegradation AGDEX aglip AgniMicroRna AIMS ALedX2 AllelicBalance alpine alsace altcdfs AMOUNTAIN amplican ampliQueso AnalysisPackageServer anamiR Anquan AneuFinder ANF anffanfy anmap annotate AnnotationDb2 AnnotationFilter AnnotationForge AnnotationFunctions AnnotationHub AnnotationHubData annotationTools annotat anota anota2seq anProfiles aComplex apegml aroma AromaExpress ArrayExpressArrayExpressHts arrayMvout arrayQuality arrayQualityMetrics ArrayTools ArrayTV ARrnNormalization ASAFE ASEMB GSCSA ASpli ASSET ASSIGN ATACseqQC attract AUCell BaalChIP BAC bacon BADER BadRegionFinder BAGS ballgown bamsignals banoco basecallQC BaseSpaceR Basic4Cseq BASICS BasicSTARRseq BatchQC BayesKnockdown BayesPeak baySeq BBCAnalysis CRANK bcSeq beachmat beadarray beadarraySNP BeadDataPackR BEARscc BEAT BEclear bgafun BgeeDB Bgmix bgx BHC BioCARE BIFET BiGGR bigmelson bigmeryomrExtras bioassayR Biobase bioinformatics bioCancer BiocCaseStudies BiocCheck BiocFileCache biometrics biocGraph BiocInstaller BioCor BiocParallel BiocSlearn BiocStyle biocViews BiocWorkflow biocDist biomaRt biomformat BioMVCClass biomerz BiocNet BioQC BiocSeqClass biosigner BioStings biosvd biotml biotmlBase BiRewire birta birt BiSeq BiomBLMA bmbc BPRM BIOC BRAIN BrainStars branchpointer bridge BridgeDbR BrowserViz BrowserVizDemo BSgenome bsseq BubbleTree BufferedMatrix BufferedMatrixMethods BUMHMM bump hunter BUS CAFE CAGER CALIB CAMERA cancer cancerclass CancerIntegrator MutationAnalysis CancerSubtypes CanD caOmicsV Cardinal casper CATALYST Category categoryCompare CausalR cbaf ccomp CCPROMISE corepe cellbaseR cellGrowth cellHTS2 cellity CellMapper CellNetOPR cellscape cellTree CEMiTool Cexor CFAssay CGHubBase CGHcall cghMCR CGHnormalizer CGHRegions ChAMP CHARGE charm ChemmineOB ChemmineR Chicago chimera chimeraViz CHIPanalyzer CHIPComp chipenrich CHIPexoQal CHIPpeakAnno CHIPQC CHIPreseq chipseq CHIPSeqR CHIPSeqCluster CHIPs CHIPXpress chosticks chroGPS chromDraw ChromHeatMap chromPlot chromStar chromSwitch chromVAR CHRONOS CIndex cisPath Classifier cleanUp4Seq cleaver clipda clipa clonal Clomical Clonality clonotypeR clst clstutils clustComp clustExp clustExperiment ClusterJudge clusterProfiler clusterSeq ClusterSignificance clusterStab CMA cn CNAnorm CNEr CNORdt CNORfeeder CNORfuzzy CNORode CNPBayes CNTools cnvGSA CNVPanelizer CNVrd2 CNVtools cobindR CoCiteStats codelink CODEX coenxnet CoGAPS cogena coGPS COHMM COMPASS compcodeR compEpiTools CompGo ComplexHeatmap CONFESS ConsensusClusterPlus consensusOf consensusSeeker contibAT conumee convert copia copynumber Copywriter CoRegNet Conormif CorMut CORREP coseq cosmi CosNet Count2Vec CoverageView covRNA cpvSNP cqn CRImage CRISPRseq crisprseqTools CrisprVariants clrm crosstata CSAR csaw CSSP ct CTDquerier ctsGE cummeRbund customProDB cve cycle cydar cytokif cytobil CytOml dada2 dagLogo daMA DaMiRseq DAPAR DEASC DBChIP dcGSA DChIPRep ddCt debrowser DECIPHER DEComplexDisease DeconRNASeq DEDS DeepBlueR deepSNV DEFormats DEGraph DEGreport DEGseq DelayedArray DelayedMatrixStats deltaGseq DeMAND DEP derfinder derfinderHelper derfinderPlot DESeq DESeq2 destiny DESubs DEXSeq dexF DiffBind diffGeneAnalysis diffHic DiffLogo diffloop diffuStats diggit Director DirichletMultinomial discordant dks DMCHMM DMRRcall DMRCare DMRRforPairs DMRSscan DNABarcodes DNAccessy DNashaper dopdelganga OQTL Doscheda DOSE drawProteins DRIMSeq DriverNet DropletUtils DrugsDisease dSimer DSS DTA dualks DupChecker dupRadar enrichment DyrnDoc Easyqpcr easyRNASeq EBarrays EBcseqexpress Eblmage EBSEA EBSeq EBSeqHMM ecoliik EDASeq EDPA edgeR edgeR eegc EGAD EGSEA eIR eisa ELBOW ELMER EMDomics EmpiricalBrownsMethod ENCODExplorer ENmix EnrichedHeatmap EnrichmentBrowser ensembleR ensembleVPEP ENVISIONQuery EpiDISH epigenomix epiNEM epiviz epivizChart epivizData epivizServer epivizStandalone ercdashboard era esATAC esetVis eudysbioeme EventPointer ExiMIR exomeCopy exomePeak ExperimentHub ExperimentHubData explore ExpressionAtlas ExpressionView fabia faciao facidFunction FamAgg farms fastLiquidAssociation fastseq FCCA frame FEM fipe FGNset fgshea FindMyFriends FISHalyseR FiHiChI flagma flowBin flowcatchR flowCHIC flowCL flowClean flowCluster flowCore flowCyBar flowDensity flowFit flowFP flowMap flowMatch flowMeans flowMerge flowPeaks flowPlidy flowPlot flowQB FlowRepositoryR FlowSOM flowStats flowTime flowTrans flowType flowUtils flowViz flowVizFlow workspace fmsR focalCall FourCSeq FRGEpistasis frm frcmaTools FunChIP FunciSNP funtootNorm GA4GHclnt GA4GHshiny gaga gage gagle gaia GAgred workflow gauch gpac gteast gCMAP gCMAPWeb gCrisprTools gcma GDCRNATools gdsfmt geec GEM genArise genBankR GeneAnswers geneAttribution GeneBreak geneClassifiers GeneExpressionSignature genefilter genu GeneGA GeneGeneInteR GeneMeta GeneNetworkBuilder GeneOverlap genePlat genePlotter geneRecommender GeneRegionScan geneRxCluster GeneSelectMMD GeneSelector GENESIS geNetClassifier GeneticsDesign GeneticsPed geneXtender GENIE3 genoCN GenoGAM genotation GenomeGraphs genomeInfoDb genomeIntervals genomes GenomicAlignments GenomicDataCommons GenomicFeatures GenomicFiles GenomicInteractions GenomicRanges GenomicScores GenomicTuples Genominator geneset genotypEval genphen GenRank GenVisR GEomat2 GEOquery GEOsubmission gep2pep gespeR GEWIST GGBase ggbio ggcyto GGtools ggtree girafe GISPA GLAD GJlimma GlobalAncova globalSeq globaltest gmapR GMRP GOexpress GOfuncR GOFunction GoogleGenomics Gopro goProfiles GOSemSim goseq GOSim GOSTAD GOSTats GOsummaries GOWHIC goVts gplf gregie gQTLBase gQTLStats graph GraphAlignment GraphAT graphite GraphPAC GRENTIS GreyListChIP GRmetrics groHMM GRgrid GSALightingn GSCA GSQA GSEABase GSEAlm gsean GSReg GSRi GSVA gVIZDESeq Guistar Guiz gwascat GTATools h5Tools haps Fabra Harman Harshlight HDFAssay HDTD heatmaps Heatmap HellowRanges HELP HEM hhAnnotator HIBChIP HiCmetrics groHicr hierGWAS HilbertCurve HilbertVis HilbertViz GUI hiReadsProcessor HiTC hmsb QueryHMMcopy 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 ImageHTS IMAS Imetagene ImmuneSpaceR ImmunoClust lMPCdata ImpulseDE ImpulseDE2 impute InPAS InPower INSPEt intansy InteractionSet interactiveDisplay interactiveDisplayBase InterEST InterMineR IntramirExplorer invaRsion iOniseR iPAC IPO IPPD IRan iSpatialFeatures iSeq isobar IsoformSwitchAnalyzeR IsoGeneGUI IsoLde isomiRs ITALICS iterativeBMA IterativeBMASur iverCluster IVAS IVOBST IWTomics JASPAR2018 joda JunctionSeq karyomplateR Kcsmat kebabs KEGGgraph KEGGlincs kerrythology KEGGP gREGEST kimod lpxse LBE lBlock LEA LedPred les Ifa liemna liimmaGUI LINC LineagePulse linnorma LiquidAssociation lmdme LMGene LOBSTASTS loci2path logicFS logitI Logolas lol LOLA Loyaloma LPE LPEadJ IpNet Ipsyomix gum LumLvsMiri LymphoSeq M3D3Drop maanova macat maCorrPlot made4 MADSEQ maTools MAGeCKFlute maigesPack MAIT makecdfv MANOR manta MantelCorr maPKL maPredictDSC mapscape marray maSigPro maskBAD MassArray massiR MassSpecWavelet MAST matchBox MatrixRider mcbx ContrastProjection MBAMethyl MBASED MBCB mbPCR MBtest mcaGUI MCBiclust MCRestimate mCSEA mdgsa mdqc MEAL MeasurementError MEDIPS MEDME MEIGOR MergeMaid Mergeomics MESHdb meshes meshr messina metaArray Metab metabomxt metaboSignal metaCCA MetaCryo metagene metagenomeFeatures metagenomeSeq metahdep metaMts MetaNeighbor metaSeq metaseqR metaviz MetCirc methupath methInheritSim MethPEd MethTargetedNGS methylVisual methylAnalysis Methyl4Id methylInheritance methylKit MethylMix methylMnM methylPipe MethylSeekR methylumi methylvim mfa Mfuzz MGFm MGFR mgsa MiChip microbiome microRNA MIGSA mimager MIMOSA MineICA minet minfi MinimumDistance MIPP MIRA MiRAGE miRBaseConverter miRcomp mirIntegrator iRLAB mlRmine mlRfNAmeConverter miRNAPath miRNAPath miRNAPath miRSponge Mistryergy missMethyl mitoDE MLIInterfaces MLP MLSeq MMDiff2 MmPalateMiRNA MODA moga monocle MoonlightR MoPS mosaics motifbreakR motifcounter MotifDb motifmatchr motifRG motifStack MotifV MPFE mpra mQTL msa msbgsR MSGFgui MSGFplus msmsEDA msmsTests MSnbase MSnId msPurity MSstats Mulcom MultiAssayExperiment multiClust MultiDataSet MultiDataSeq multiMiR multiMiR multiOmicsViz multiscan multest muscle MutationalPatterns MVCvcST MWASTools mygene myvariant mZID mzR NADfinder NanoStringDiff NanoStringQCPro NarrowPeaks ndcfFlow NClgraph ndexr nem netbenchmark netbioV nethet NetPathMiner netprioR netReg netresponse NetSAM networkBMA NGScopy nnNorm NOISeq nond normalize450K NormQCPro normR nGSEA NtW nucleoS1N nucleoS2 nudge NuPoP occugene OCplus odseq OGSa oligo oligoClasses OLIN OLINguj omicade4 OmicCircos omicplotR omicRexposome OmicsMarkeR omicsPrint Onassis oncomix OncoScore OncoSimulR onSENSE ontoCAT ontoProc openCryo openRPrimeRui OpenMate oposSOM oppar OPWedge OrderedList Organism OrganismDbi OSAT Oscape OTUbase OutlierD PAA PADQC paircomp parzander paneler PannBuilder PannBuilder PannP PanVizGenerator PAPA Rgmlms Rgmls Rhamada Gx phenoDist phenocPath phenoTest Phir phosphonormalizer phyloseq Pi piano pickGene PICS Pigene Gene PING pint pkgDepTools plateCore prethy plgem plier PLPE plrs plw pmm podkat pogos polyester Polyfit POC PPInfer ppiStats pgsfinder prada PREDADA prediction preprocCore Prize probAMR PROcess procoil PRiCoNA proFIA profileScoreDist prgncy pRoloc pRolocGUI PROMISE PROPER PROPS Prostar proT2D proteinProfiles ProteomicsAnnotationHubData proteoQC ProtGenomics PSEA psychomics PSIC pSynge2r puma PureCN pvac pvca Pviz PWMEnrich pvOmics qemetrics QDNaseq qpcrNorm qprghr qrcv qsea QUALIFIER quanto quantsmooth QuartPAC Quasar QuaternaryProd QUbic qvalc qvalue R3CPET r3Cseq R453Plus1Toolbox R4RNA RaggedExperiment Ramigo ramwas randPack RankProd RareVariantVis Rariant RbcBook1 RBGL Rbiofin rBiopaxParser RBM Rbowtie Rbowtie2 rbsurv Rcade RCAS RCASPAR reClminer rCGH Rchempp RchyOptimyx RoisTarget RcpI RcY3 RCy3s RDAVIDWebService rDGlDb Rdis RNATools ReactomePA readat ReadqPCR rec recount recour Reder RedSeq RefNet RefPlus regioneR regionReporter regsplice REMP Repitools ReportingTools ReQon restfulSei rexposome rFRed rGADEM RGalaxy RGMQL RGMQL2s Rgraphviz rGREAT RGSEA reg hdf5 hdf5clnt Rhd5lib Rhtslib rHVDM RiboProfiling riboSeqR RlimmPort Ringo RIPseeker Risa RITAN RIVER RJMCMCNucleosomes RLMM Rmagine RMassBank rMAt RmiR RnAInteract RnAInter RnAProbr rmasseqmap rnaSeqMap RNASeqPower RnaSeqSamples rNBeads Rnlis rolr ROC RoleSwitch rols roma RQntoTools ropls ROTS RPA RProtoBufLib RpsiXML rpx Rqc rqt rubic rRDP RRHQ Rsamtools rsbml rSFFreader Rsubread RSVSim rTANDEM RTCA RTCGA RTCGAToolbox RTN RTNduals RTNsurvival RTopper rTracklay rTremix rTRM rTRMUI rhanic RUVcov RUVnormalize RUVSeq RVS S4Vectors safe sagehant SAGx samExplorer sampleClassifier SamSPECTRAL sangerSeqR SANTA sapFinder saVR SBMLR SC3 Scale4C SCAN scater scDD scde scFeatureFilter scfnd ScSl scmap ScNorm rGene scoreplyR scrPine scrPine SCR sementSeq SFLX SemDist semisun SEPA seq2Pathway SeqArray sebahias seqCAT seqCN seqNBA seqrncbm SeqRSEFA seq\_oon seqPattern seqStats SeqSOC seqTools SeqVarTools sevenbirds SGScan ShinyMethyl shinyV

# 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

**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Home » BioViews

## All Packages

**Bioconductor version 3.6 (Release)**

Autocomplete biocViews search:

- ▼ Software (1477)
  - ▶ AssayDomain (574)
  - ▶ BiologicalQuestion (561)
  - ▶ Infrastructure (323)
  - ▼ ResearchField (414)
    - BiomedicalInformatics (30)
    - CellBiology (37)
    - Cheminformatics (9)
    - ComparativeGenomics (3)
    - Epigenetics (25)
    - FunctionalGenomics (24)
    - Genetics (158)
    - Lipidomics (7)
    - MathematicalBiology (2)
    - Metabolomics (32)
    - Metagenomics (14)
    - Pharmacogenetics (8)

**Packages found under Software:**

Show  entries Search table:

Package	Maintainer	Title
<a href="#">a4</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Umbrella Package
<a href="#">a4Base</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Base Package
<a href="#">a4Classif</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Classification Package
<a href="#">a4Core</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Core Package
<a href="#">a4Preproc</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Preprocessing Package
<a href="#">a4Reporting</a>	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Reporting Package
<a href="#">ABAEnrichment</a>	Steffi Grote	Gene expression enrichment in human brain regions
<a href="#">ABArray</a>	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (AB1700) gene expression data.
<a href="#">ABSesq</a>	Wentao Yang	ABSesq: a new RNA-Seq analysis method based on modelling absolute expression

# 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](https://doi.org/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 r74618) -- "Joy in Playing"	3028
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

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

**Package propagation status is indicated by one of the LEDs**

- YES: Package was propagated because it didn't pre-emptively bump version
- NO: Package was not propagated because of a problem (impossible dependencies, or version lower than what was propagated)
- UNNEEDED: Package was not propagated because the repository with this version. A version bump is required 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 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		

**Package 1/1554**

**a4 1.27.0**  
Tobias Verbeke  
Last Commit: 5266e2b  
Last Changed Date: 2017-10-30 12:52:11 -0400

Hostname	OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
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**

**a4Base 1.27.0**  
Tobias Verbeke  
Last Commit: 72d568e  
Last Changed Date: 2017-10-30 12:52:11 -0400

Hostname	OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
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

---

#### 1 What is the GDC?

#### 2 Quickstart

- 2.1 Installation
- 2.2 Check basic functionality
- 2.3 Find data
- 2.4 Download data
- 2.5 Metadata queries

#### 3 Usage

- 3.1 Querying metadata
  - 3.1.1 Creating a query
  - 3.1.2 Retrieving results
  - 3.1.3 Fields and Values
  - 3.1.4 Facets and aggregation
  - 3.1.5 Filtering
- 3.2 Authentication
- 3.3 Datafile access and download

# Qualities

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

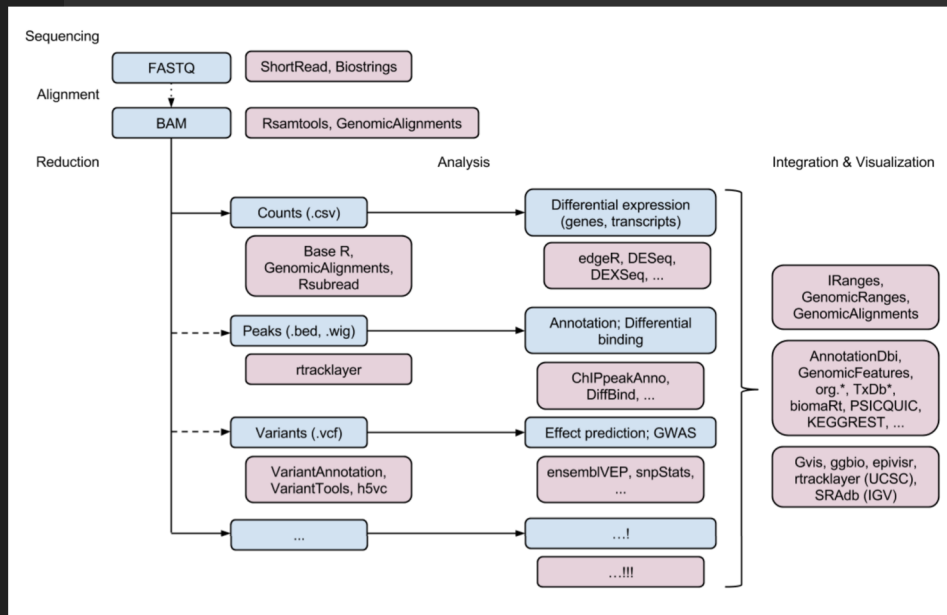
The screenshot shows the Bioconductor website interface. At the top, there is a navigation bar with links for 'My: messages', 'votes', 'posts', 'tags', 'following', and 'bookmarks'. The user 'Sean Davis' is logged in with 21k reputation. The main header includes the Bioconductor logo and navigation links for 'ASK QUESTION', 'LATEST', 'NEWS', 'JOBS', 'TUTORIALS', 'TAGS', and 'USER'. Below the header is a search bar with 'Limit' and 'Sort' dropdowns. The main content area displays a list of questions and answers, each with a title, a grid of vote and answer counts, a view count, tags, and the author's name and reputation. The questions listed are:

- Cannot install Rhtslib on Mac OS 10.13** (0 votes, 0 answers, 2 views) - written 2 minutes ago by Ryan C. Thompson • 6.5k
- No CNAs or SNVs in results** (0 votes, 1 answer, 35 views) - written 21 hours ago by twtoal • 0
- monocle estimateSizeFactors give Inf for all values** (0 votes, 0 answers, 8 views) - written 1 hour ago by jonessara770 • 10
- Row clustering featureAlignedHeatmap function (ChipPeakAnno package)** (0 votes, 1 answer, 14 views) - written 2 hours ago by gdeniz • 0 • updated 2 hours ago by Ou, Jianhong • 1.0k
- Metabolite identification package** (0 votes, 0 answers, 7 views) - written 2 hours ago by johnhamre3 • 0
- LaTeX Error with BiocWorkflowTools** (1 vote, 1 answer, 10 views) - written 3 hours ago by shbrief • 10 • updated 2 hours ago by Mike Smith • 2.6k
- msa output formats for use down stream to create phylogenetic trees** (0 votes, 0 answers, 6 views) - written 2 hours ago by cav3gh • 0
- Results counts post DESeq same raw counts** (0 votes, 1 answer, 40 views) - written 13 hours ago by A • 0 • updated 3 hours ago by Michael Love • 17k
- Problem installing Minfi on Cluster** (0 votes, 1 answer, 14 views) - written 5 hours ago by Goku • 0 • updated 5 hours ago by Kasper Daniel Hansen • 6.3k
- Compare groups of different RNAseq sets** (0 votes, 1 answer, 27 views) - written 7 hours ago by b.nota • 290 • updated 5 hours ago by Aaron Lun • 19k

On the right side, there are sections for 'Recent... Replies', 'Votes', 'Awards', and 'Locations', each with a list of related items and their respective counts.

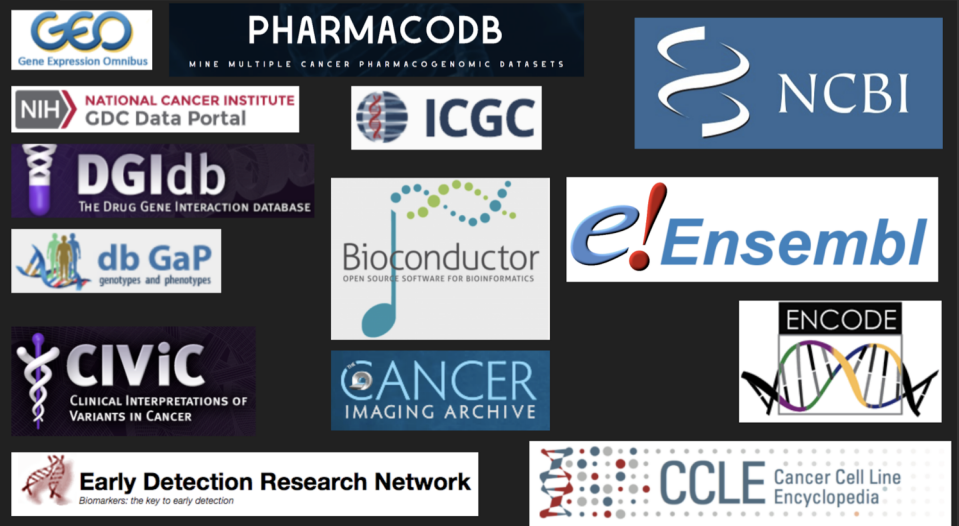
# Qualities

- 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](https://doi.org/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

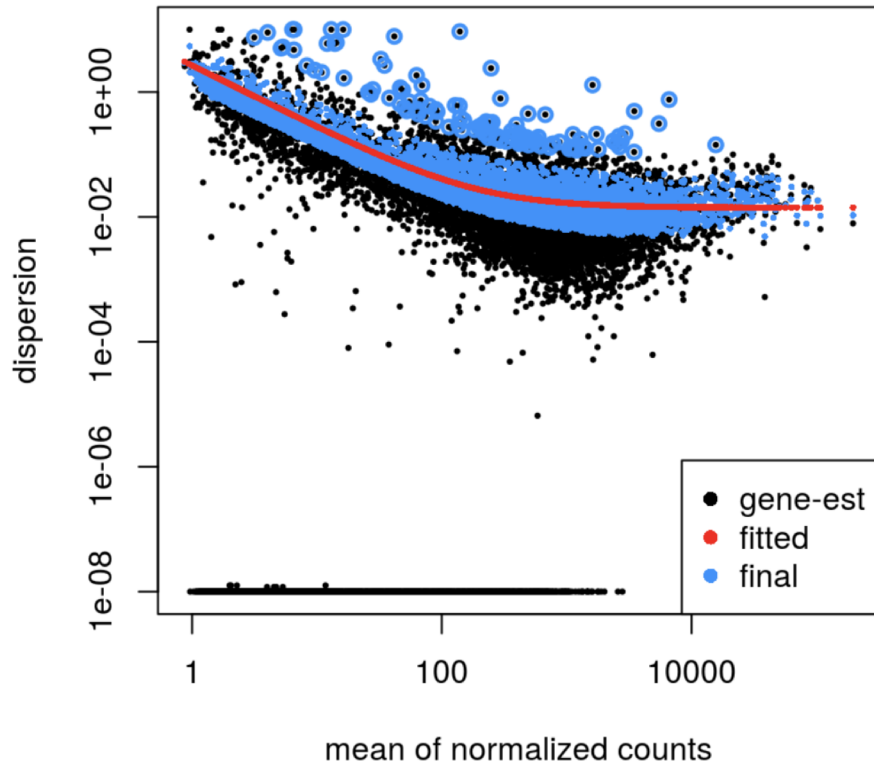
To install this package, start R and enter:

```
## 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*

The screenshot shows the GitHub repository page for Bioconductor. The repository name is 'Bioconductor / Contributions'. The page displays the following information:

- Repository: Bioconductor / Contributions
- Watch: 18
- Code, Issues (24), Pull requests (0), ZenHub, Projects (0), Wiki, Insights
- Contribute Packages to Bioconductor
- bioconductor
- 45 commits, 1 branch, 0 releases
- Branch: master, New pull request, Create new file, Upload files
- mtmorgan add instructions to confirm web hook on creation
- CONTRIBUTING.md: add instructions to confirm web hook on creation
- README.md: first commit
- issue\_template.md: Update the template to new source control material
- README.md
- Table of Contents
  - [Contributing a \*Bioconductor\* Package](#)
  - [Starting the Submission Process](#)
  - [What to Expect](#)
  - [Adding a Web Hook](#)
  - [Submitting Related Packages](#)
  - [Additional Actions](#)
  - [Resources](#)
- Contributing a *Bioconductor* Package

[Home](#) » [Developers](#) » Packages: New Submissions



## Package Submission

- [Introduction](#)
- [Checklist](#)
- [Submission](#)
- [Review Process](#)
- [Additional Support](#)

### Introduction

Bioconductor Packages should

- Address areas of high-throughput genomic analysis where Bioconductor already makes significant contributions, e.g., sequencing, expression and other microarrays, flow cytometry, mass spectrometry, image analysis; see [biocViews](#).
- Interoperate with other Bioconductor packages, re-using common data structures ([S4 classes and methods](#)) and existing infrastructure (e.g., `rtracklayer::import()` for input of common genomic files).
- Adopt software best practices that enable reproducible research and use, such as full documentation and vignettes (including fully evaluated code) as well as commitment to long-term user support through the Bioconductor [support site](#).

### Source Code & Build Reports »

Source code is stored in [Git](#).

Software packages are built and checked nightly. Build reports:

- [All](#)
- [Release](#)
- [Development](#)
- [Package Download Statistics](#)

### Development Version »

Bioconductor packages under development:

- Analysis [software](#) packages.
- [Annotation](#) packages
- Illustrative [experiment data](#) packages

*Core value: open and engaged*



Search or jump to...



Pull requests Issues Marketplace Explore

Bioconductor / Contributions

Code

Issues 44

Pull requests 0

ZenHub

Contribute Packages to Bioconductor

bioconductor

44 Open 1,155 Closed

Author

Labels

Projects

Milestones

Assignee

Sort

Qtlizer 2. review in progress OK

#1203 opened 7 hours ago by matmu 8 of 8



10

Spaniel 2. review in progress

#1202 opened 4 days ago by RachelQueen1 8 of 8



2

muscat 2. review in progress ERROR WARNINGS

#1201 opened 6 days ago by HelenaLC 8 of 8



18

ribor 2. review in progress ERROR VERSION BUMP REQUIRED

#1200 opened 10 days ago by mjgeng 8 of 8



3

ssPATHS 2. review in progress OK

#1194 opened 17 days ago by nrosed



20

Sign in now to use ZenHub

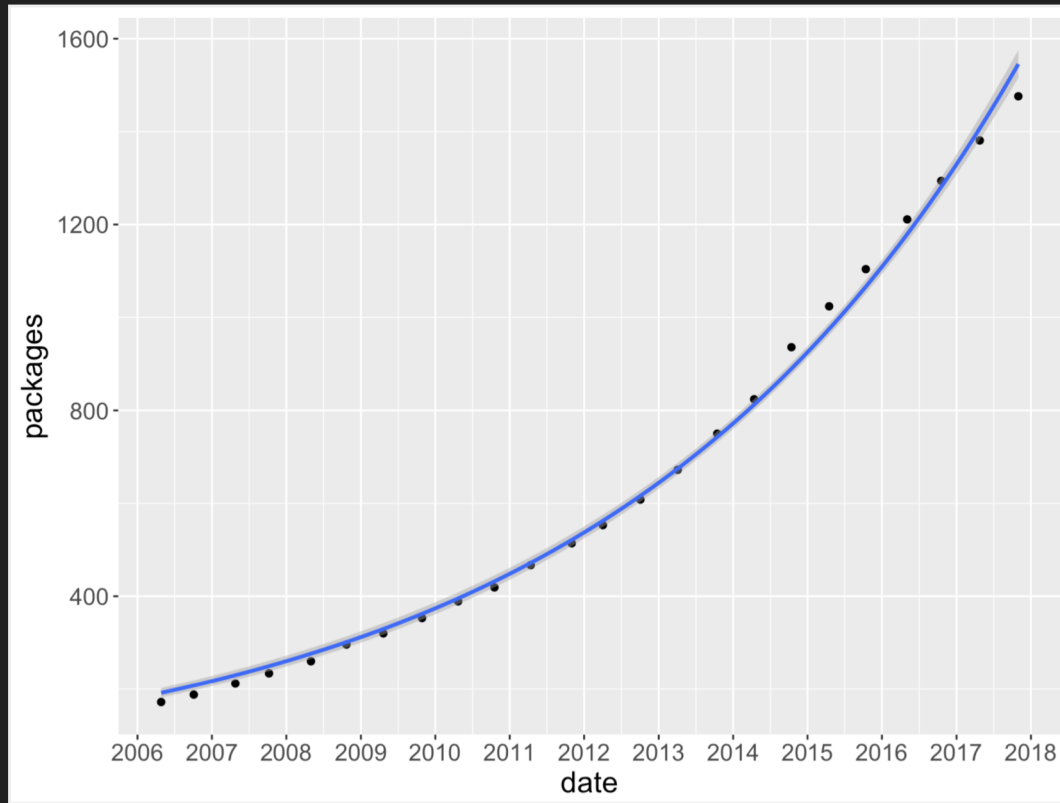
Omnia 2. review in progress OK VERSION BUMP REQUIRED



21

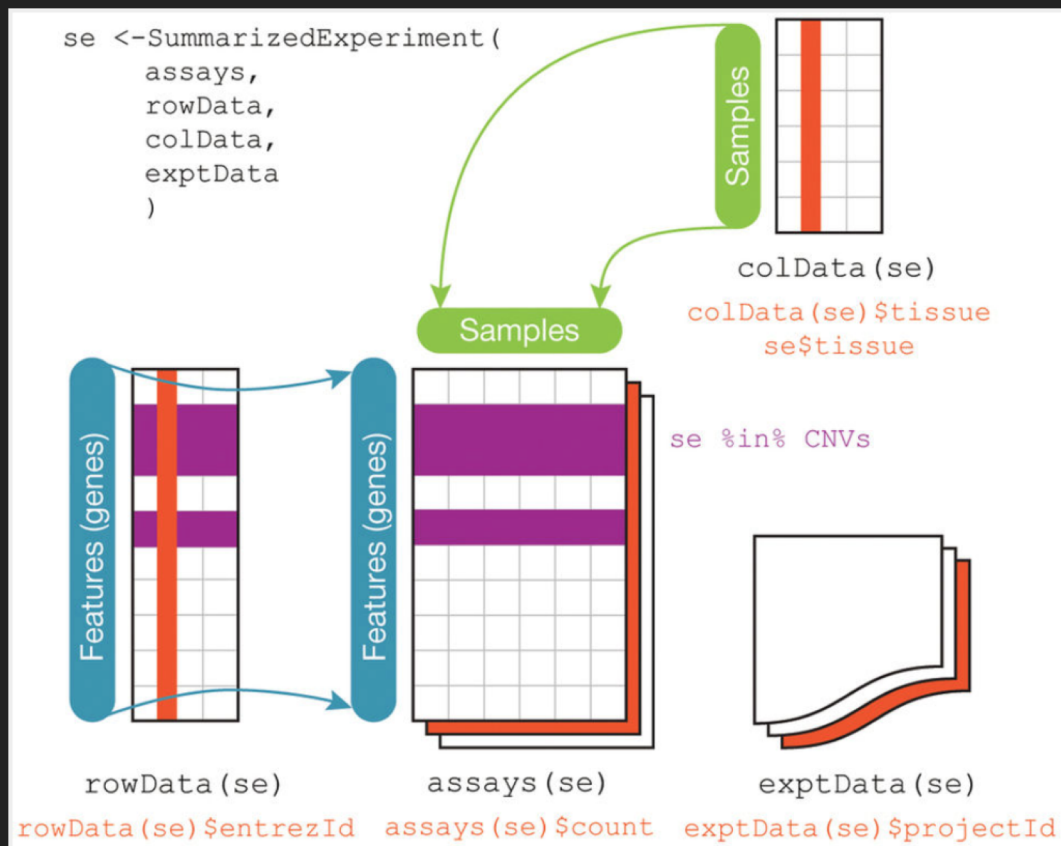
<https://github.com/Bioconductor/Contributions>

# 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

The data from your data.frame - Mozilla Firefox (on gamay)

The data from your d... x

127.0.0.1:5758

Selected rows: AH14107 AH14109

Return rows to R session

Show 10 entries

idx	dataprovder	species	genome	description	tags	rdataclass	sourcetype
AH14107	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to canFam3	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14108	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to hg19	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14109	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to mm10	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14110	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to monDom5	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14111	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to panTro4	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14112	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to rheMac3	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14113	UCSC	Homo sapiens	hg38	UCSC lftOver chain file from hg38 to rn5	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14114	UCSC	Homo sapiens	hg19	UCSC lftOver chain file from hg19 to allMe1	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14115	UCSC	Homo sapiens	hg19	UCSC lftOver chain file from hg19 to allMis1	lftOver, chain, UCSC, genome, homology	ChainFile	Chain
AH14116	UCSC	Homo sapiens	hg19	UCSC lftOver chain file from hg19 to anoCar1	lftOver, chain, UCSC, genome, homology	ChainFile	Chain

Showing 1 to 10 of 1,113 entries (filtered from 19,268 total entries)

AnnotationHub

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

```
## AnnotationHub with 45 records
## # snapshotDate(): 2019-05-20
## # $dataprovder: 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 | dm2ToDroWill.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 dataprovder 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...
## # $rdaclass: ExpressionSet, SummarizedExperiment, SummarizedBenchmark, ...
## # additional mcols(): taxonomyid, genome, description
## # coordinate_1_based, maintainer
## # 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_Threshol
## EH2541 | SKCM_GISTIC_AllByGene
## EH2542 | SKCM_GISTIC_Peaks-201
## EH2543 | SKCM_GISTIC_Threshol
```

```
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*

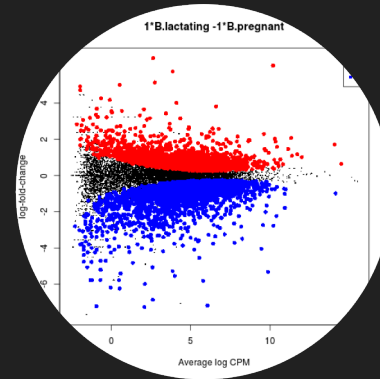
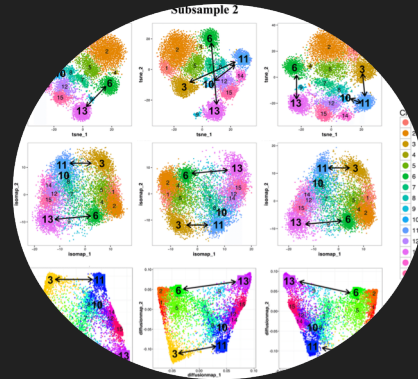
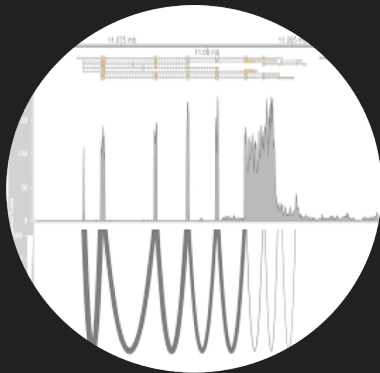
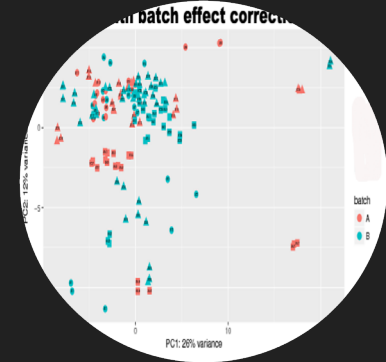
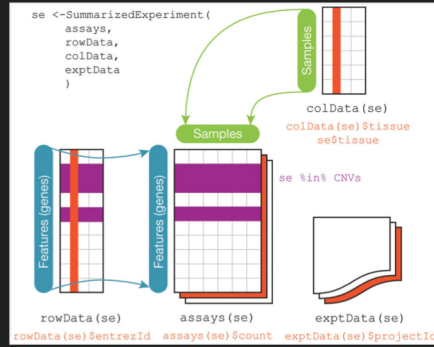
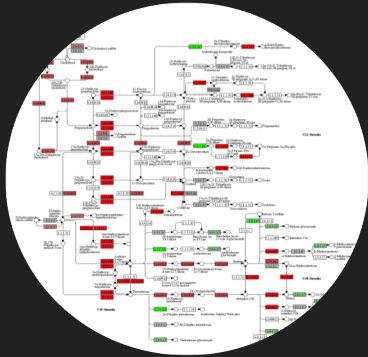
**Modified: November 2017. Compiled: 02 May 2019**

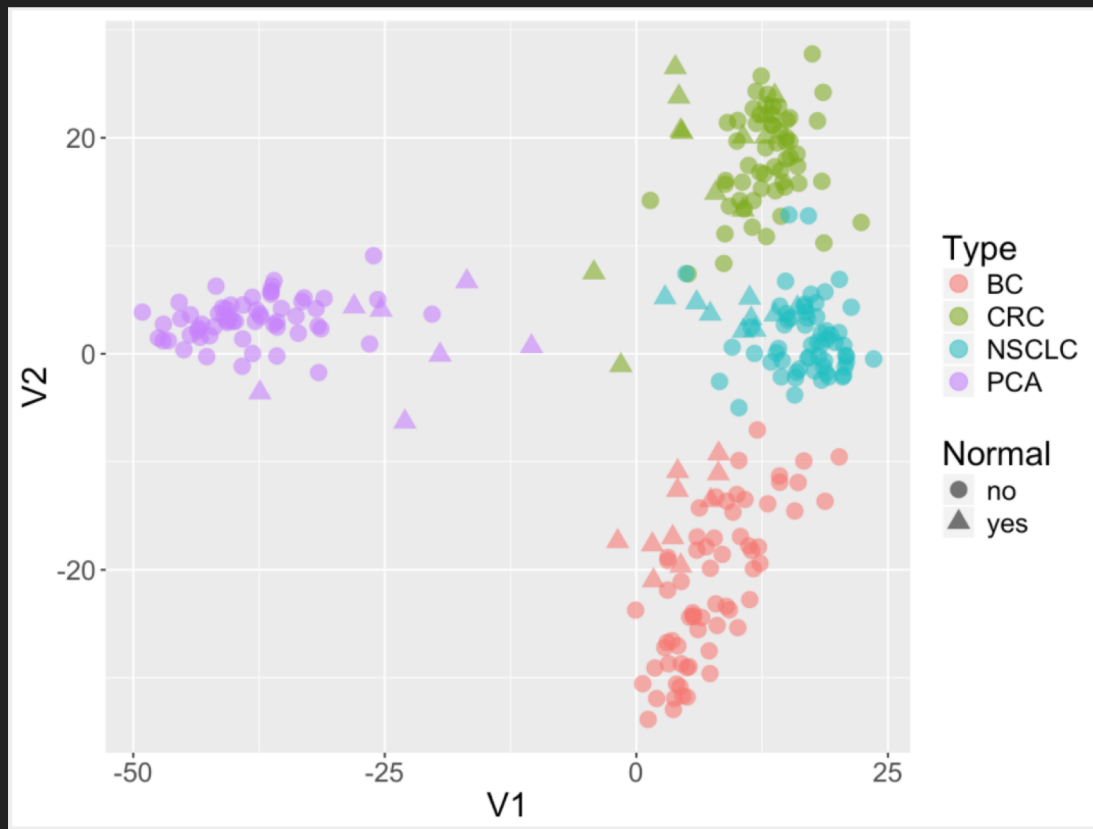
## Contents

---

- 1 Overview
- 2 New resources
  - 2.1 Notify Bioconductor team member
  - 2.2 Building the data experiment package
  - 2.3 Data objects
  - 2.4 Metadata
  - 2.5 Package review
- 3 Add additional resources
- 4 Bug fixes
  - 4.1 Update the resource
  - 4.2 Update the metadata
- 5 Remove resources
- 6 Uploading Data to S3
- 7 Validating
- 8 Example metadata.csv file and more information

# Core value: reuse and interoperability





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

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)

# Harmonized Cancer Datasets Genomic Data Commons Data Portal

Get Started by Exploring:

- Projects
- Exploration
- Analysis
- Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

## Data Portal Summary [Data Release 37.0 - March 29, 2023](#)

PROJECTS

78

PRIMARY SITES

68

CASES

86,962

FILES

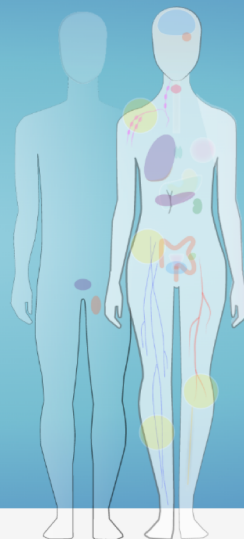
931,947

GENES

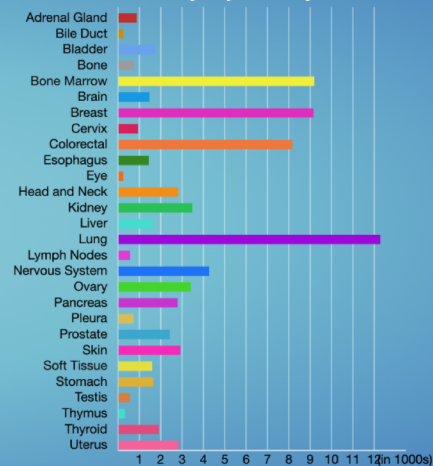
22,501

MUTATIONS

2,885,293



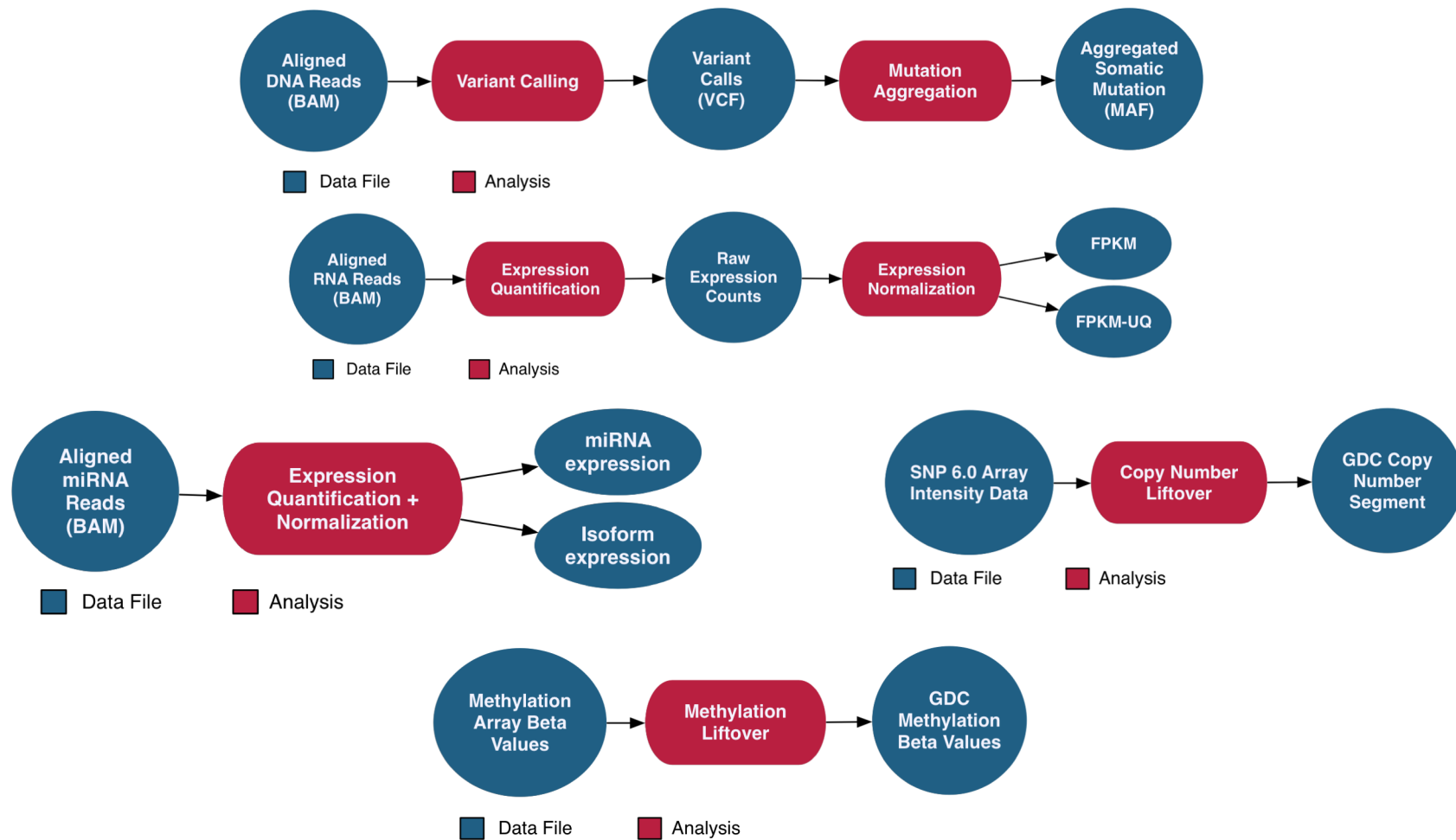
Cases by Major Primary Site



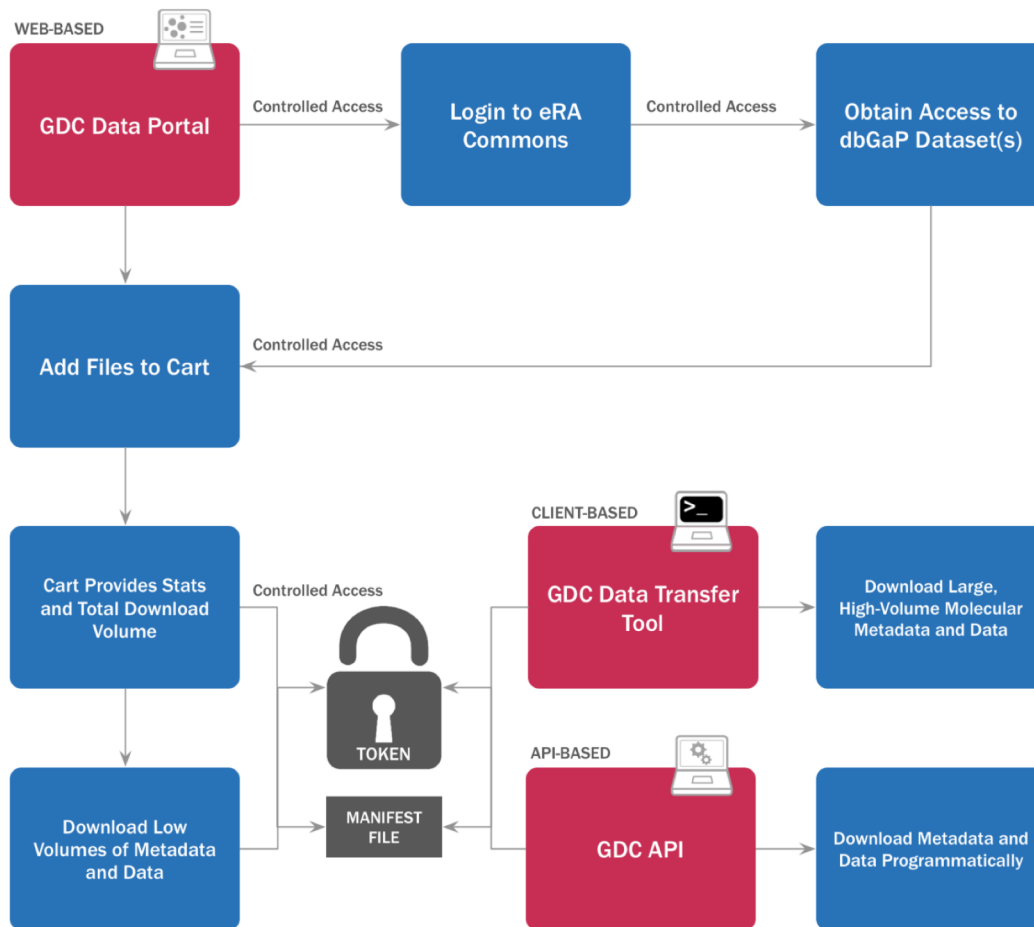
## GDC Applications

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:



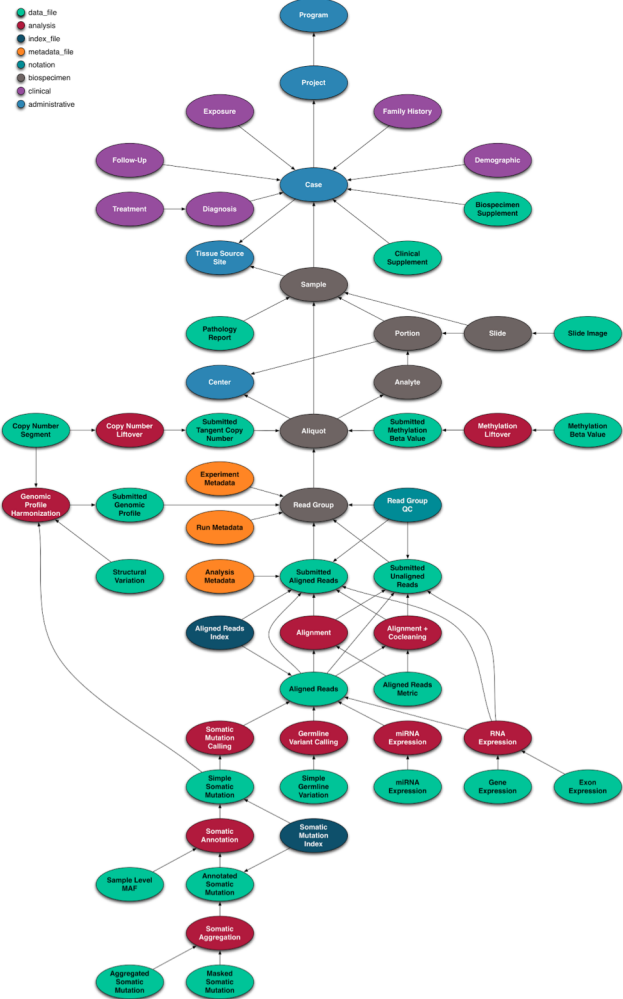






# GDC Data Model

Complex, flexible, and growing....



[Home](#) » [Bioconductor 3.17](#) » [Software Packages](#) » [GenomicDataCommons](#)

## GenomicDataCommons

platforms **all** rank **169 / 2229** support **0 / 0** in Bioc **6 years**  
build **ok** updated **< 1 week** dependencies **53**

DOI: [10.18129/B9.bioc.GenomicDataCommons](#)

### NIH / NCI Genomic Data Commons Access

Bioconductor version: Release (3.17)

Programmatically access the NIH / NCI Genomic Data Commons RESTful service.

Author: Martin Morgan [aut], Sean Davis [aut, cre], Marcel Ramos [ctb]

Maintainer: Sean Davis <seandavi at gmail.com>

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

Morgan M, Davis S (2023). *GenomicDataCommons: NIH / NCI Genomic Data Commons Access*.  
<https://bioconductor.org/packages/GenomicDataCommons>,  
<http://github.com/Bioconductor/GenomicDataCommons>,  
<http://bioconductor.github.io/GenomicDataCommons/>.

### Documentation »

*Bioconductor*

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

### Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

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.



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])
```

Program

TCGA

470

Project

FM-AD

18,004

TARGET-NBL

1,127

TCGA-BRCA

1,088

TARGET-AML

988

TARGET-WT

652

TCGA-GBM

617

TCGA-OV

608

TCGA-LUAD

685

TCGA-UCEC

560

TCGA-KIRC

537

TCGA-HNSC

528

TCGA-LGG

516

TCGA-THCA

507

TCGA-LUSC

504

TCGA-SKCM

470

TCGA-CCAD

443

TCGA-STAD

412

TCGA-BLCA

381

TARGET-OS

377

TCGA-LIHC

307

Show More

Showing 1 - 20 of 11,265 files

JSON TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
controlled	e88a4cbf-de10-4fac-a303-86e3cad93386.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	137.25 KB	0
controlled	826d1ac1-66dd-43f5-8787-eb181de3ae88.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	666.91 KB	0
open	78ae36c4-8de7-41b6-88bc-9357cf8e4060.mirbase21_mirnas.quantification.txt	1	TCGA-SKCM	Transcriptome Profiling	TSV	50.48 KB	0
controlled	128661.bam	1	TCGA-SKCM	Raw Sequencing Data	BAM	329.92 MB	0
open	fc163e29-39e5-4064-ab4f-ba741ac115cc.htseq.counts.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	248.21 KB	0
controlled	C828.TCGA-D3-A1Q3-06A-11D-A196-08.2_gdc_realn.bam	1	TCGA-SKCM	Raw Sequencing Data	BAM	10.38 GB	1
open	nationwidechildrens.org_biospecimen.TCGA-D3-A1QB.xml	1	TCGA-SKCM	Biospecimen	BCR XML	61.32 KB	0
controlled	03891b68-acb2-4a63-839c-2e56f35846db.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	367.83 KB	0
open	ALANG_p.TCGA_180_SNP_1N_GenomeWideSNP_6_H03_895878.grch38_seg.txt	1	TCGA-SKCM	Copy Number Variation	TXT	36.14 KB	0
open	1d0ed301-414c-4945-a65b-5bfb4360d65.FPKM.txt.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	535.3 KB	0
controlled	28acee18-95c5-4690-8afa-814655395ca7.vcf	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	330.14 KB	0
open	4667aded-fa48-493e-8cec-308648b0bb9b.htseq.counts.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	245.78 KB	0
controlled	f916f49e-8037-4f34-8ae3-d79674c8660e_gdc_realn_rehead.bam	1	TCGA-SKCM	Raw Sequencing Data	BAM	8.13 GB	0
open	393d326f-f2ad-4e0a-83bc-d41421dbd25e.FPKM.txt.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	501.64 KB	0
controlled	5c81b09d-7f0a-461c-aaad-bb5f1461313.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	1.69 MB	0
open	ac6098f3-b03b-4fd2-a214-cab070b2ccbd.htseq.counts.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	245.69 KB	0
controlled	98f9a513-85f9-4f5b-8540-2d37d8482f2c.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	78.87 KB	0
controlled	C828.TCGA-EE-A17X-10A-01D-A199-08.2_gdc_realn.bam	1	TCGA-SKCM	Raw Sequencing Data	BAM	9.86 GB	0
controlled	29b9aed6-f0a8-4f9-bb1c-79766f2e2dbe.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	968.66 KB	0
controlled	2da8ef88-ba32-49bf-8d25-a85fc93975d9.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	130.31 KB	0

Show 20 entries

1 2 3 4 5 6 7 8 9 10

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])
```

Files Cases

[Add a File Filter](#)

File

Q 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

Somatic Copy Number 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
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  entries

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 Cases

[Add a File Filter](#)

File

Q 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
<input type="checkbox"/>	open	<a href="#">TCGA.SKCM.mutect.4b7a5729-b83e-4837-9b61-a6002dce1c0a.DR-10.0.somatic.maf.gz</a>	470	<a href="#">TCGA-SKCM</a>	Simple Nucleotide Variation	MAF	89.38 MB	48

Showing 1 - 20 entries

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])
```

Files Cases

[Add a File Filter](#)

File

Q 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](#) /few 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

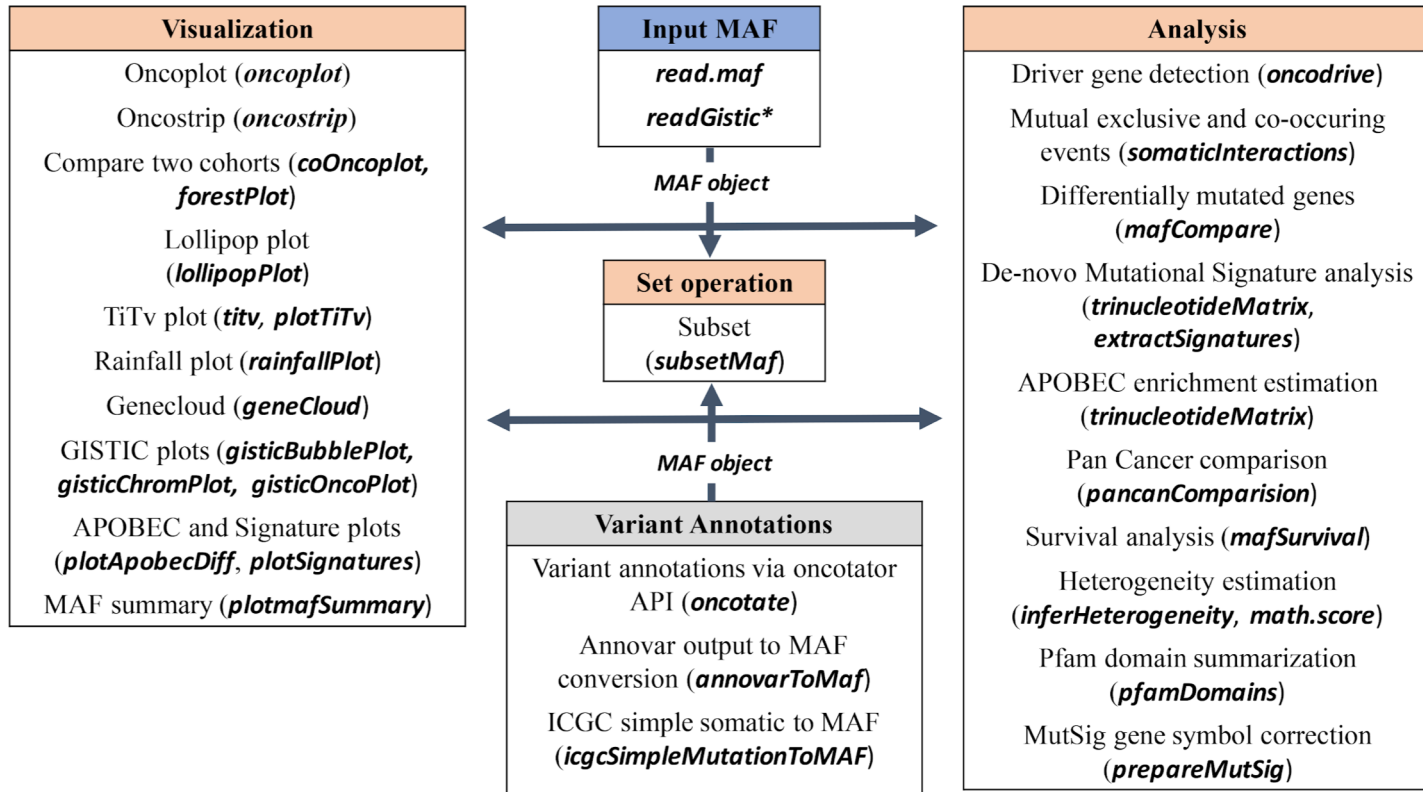
1

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)*



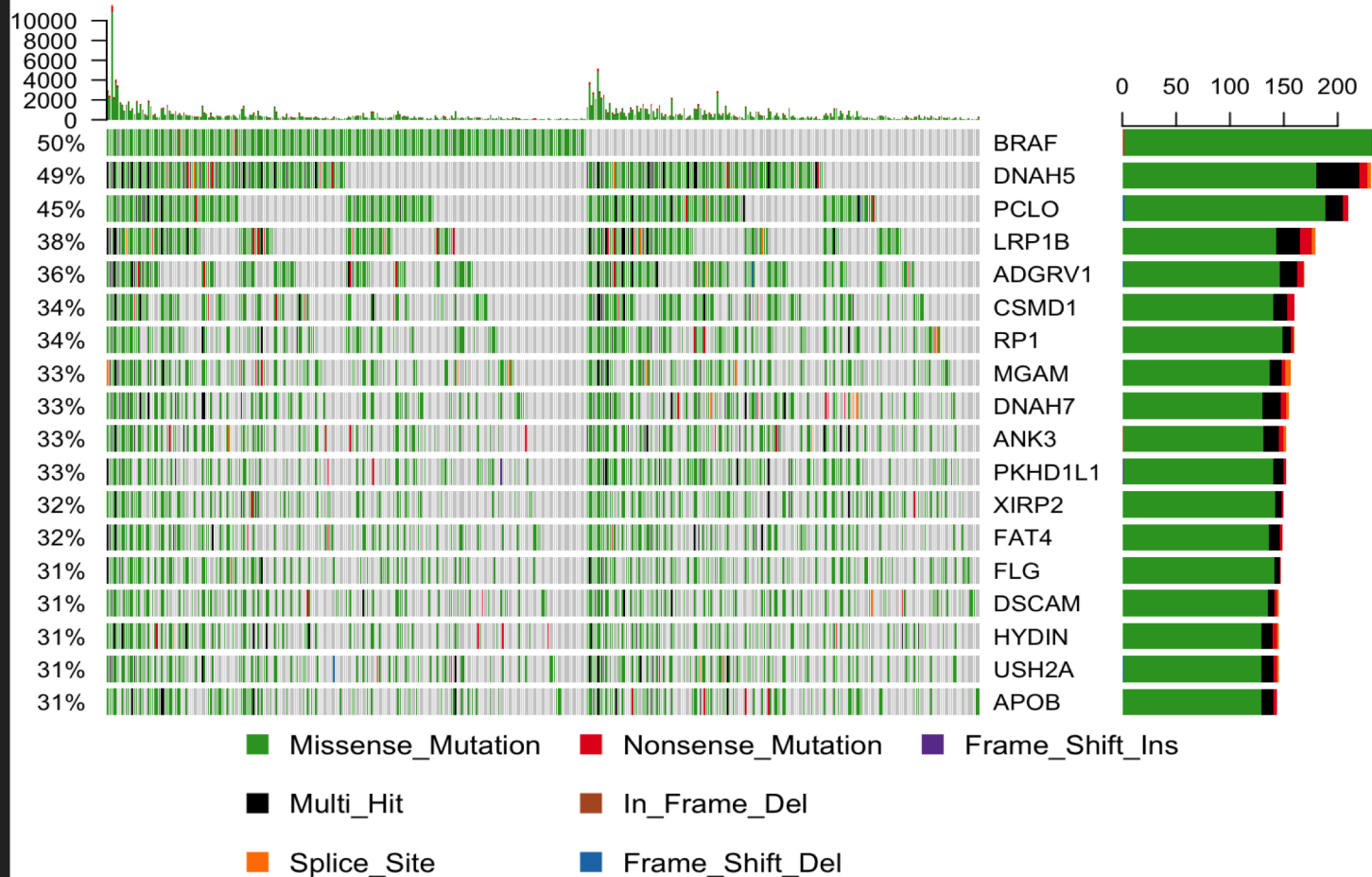
Use the **GenomicDataCommons** package to find and download variants from the TCGA cutaneous melanoma dataset.

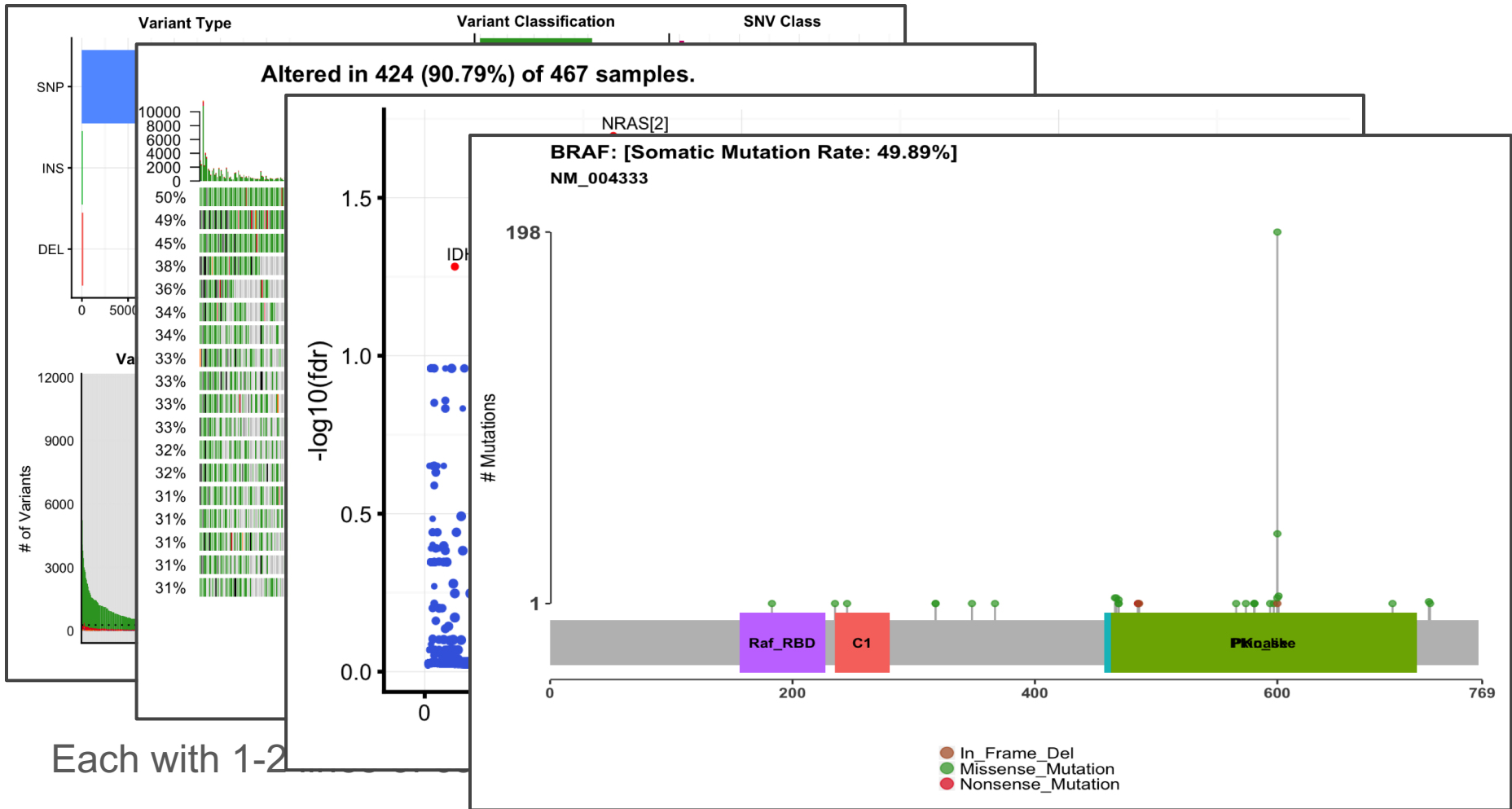
```
library(GenomicDataCommons)
fnames = files() %>%
  GenomicDataCommons::filter(~ cases.project.project_id=='TCGA-SKCM' &
    data_type=='Masked Somatic Mutation' &
    data_format=='MAF' &
    analysis.workflow_type=='MuTect2 Variant Aggregation and Masking') %>%
  ids() %>%
  gdcdata()
```

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.







SOFTWARE TOOL ARTICLE

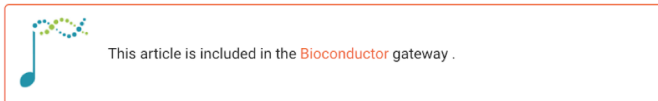
**REVISED** *TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages [version 2; referees: 1 approved, 2 approved with reservations]*

✉ **Tiago C. Silva**<sup>1,2\*</sup>, Antonio Colaprico<sup>3,4\*</sup>, Catharina Olsen<sup>3,4\*</sup>, Fulvio D'Angelo<sup>5,6\*</sup>, Gianluca Bontempi<sup>3,5</sup>, Michele Ceccarelli<sup>7</sup>, ✉ **Houtan Noushmehr** <sup>1,8</sup>

\* Equal contributors

+ Author details

+ Grant information



**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](#),

**METRICS**

10413  
VIEWS

2137  
DOWNLOADS

Get PDF

Get XML

Cite

Export

Track

Email

Share

**Open Peer Review**

Referee Status: ? ✓ ?

	Invited Referees		
Version(s)	1	2	3
<b>REVISED</b> 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 Sonesson** , University of Zurich (UZH), Switzerland

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

**Comments on this article**

All comments (0)

Add a Comment

**Subscribe to Table of Content Alert**

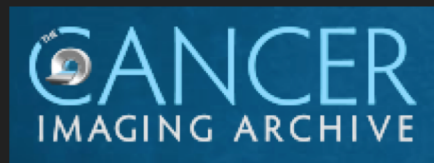
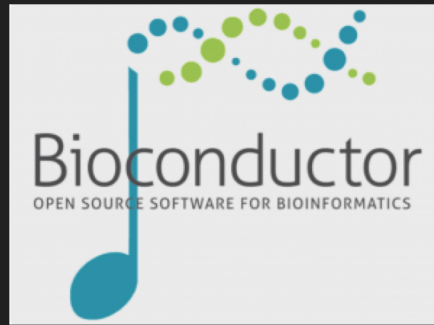
Your email address

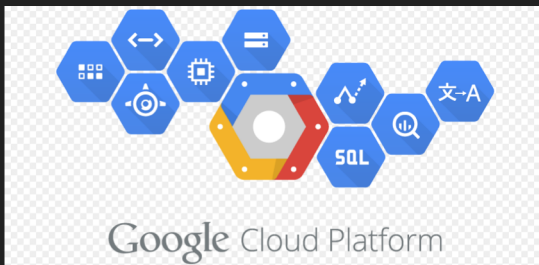
SUBSCRIBE



# PHARMACODB

MINE MULTIPLE CANCER PHARMACOGENOMIC DATASETS





# Bioconductor: Education, Training, and Community

The screenshot shows the Bioconductor website homepage. At the top left is the Bioconductor logo with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right is a search bar and a navigation menu with links for Home, Install, Help, Developers, and About. The main content area is divided into several sections: "About Bioconductor" with a paragraph describing the software and its use of R; "News" with a list of recent updates; "Install" with a list of links for getting started; "Learn" with a list of resources for mastering the tools; "Use" with a list of links for creating bioinformatic solutions; and "Develop" with a list of links for contributing to the project. At the bottom, there are links for Support, Events, and a Twitter feed.

**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 Bloc "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

# Questions?

<https://bioconductor.org>

<https://seandavi.github.io>



