

Bioconductor

*Software for orchestrating
high-throughput biological data analysis*

*July 19, 2024,
BigCare 2024
UC Irvine*

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

Bioconductor

*Software & Community for orchestrating
high-throughput biological data analysis*

*July 19, 2024,
BigCare 2024
UC Irvine*

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

software

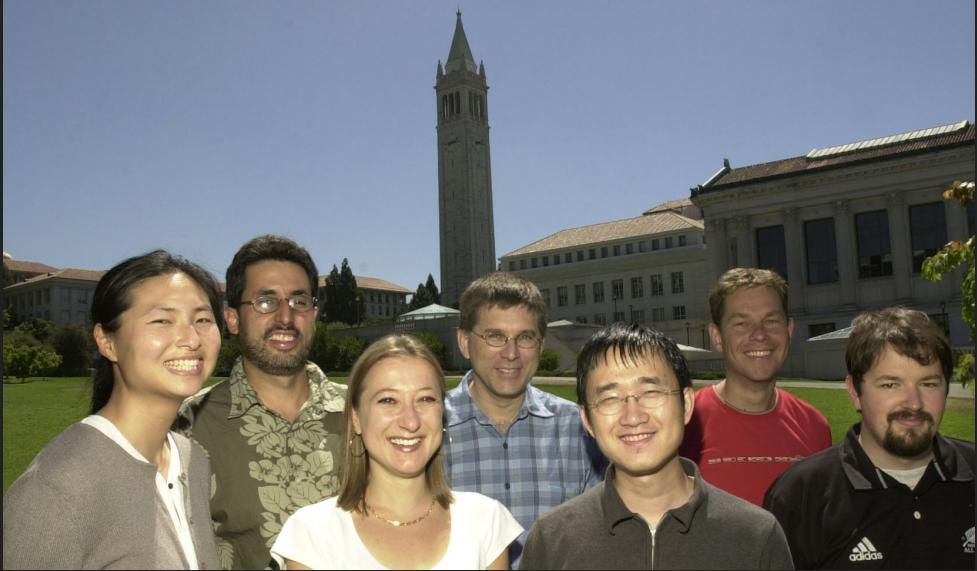
encodes

knowledge

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

2003: Changing the way
genomics is done - Jean Yang,
Rafa Irizarry, Sandrine Dudoit, Rob
Gentleman, Cheng Li, Wolfgang
Huber, Ben Bolstad -- foundations for:

2023+ Broadening
competencies for
inclusive,
collaborative science
worldwide, so more
voices are heard and
insights cultivated



Bioconductor by the numbers

- Project started in 2002
- Approximately 100,000 unique downloads per month
- More than 22,000 PubmedCentral citations
- Bioconductor Release:
 - 2217 Software packages
 - 926 Annotation packages
 - 430 ExperimentHub packages
- Receiving submissions of 3-6 new packages *per week*
- More than 1000 active developers

Just some of the many
Bioconductor available packages!!



a4Base a4Classif a4Core a4Preproc a4Reporting ABAEnrichment ABSSeq acde aCGH ACME ADaCGH2 adSplit affxparser affy affycomp AffyCompatible affyContam affycoretools AffyExpress affyIILM affyio affyImGUI affyPara affyPdnn affyPLM affyQCReport affyRNADegradation AGDEX agilP AgiMicroRna AIMS ALDEx2 AllelicImbalance alpine alsace altccfenvs AMOUNTAIN ampican ampliQueso AnalysisPageServer anamiR Anaquin AneuFinder ANF annaffy anmapl annotate AnnotationDbi AnnotationFilter AnnotationForge annotationFuncs AnnotationHub AnnotationHubData annotationTools annotat anota anota2seq antiProfiles apComplex apegilm aroma ArrayExpress ArrayExpressHTS arrayMvout arrayQuality arrayQualityMetrics ArrayTools ArrayTV ARRMnormalization ASAFEE ASEB ASEG ASGS ASI ASSET ASSIGN ATACseqQC attract AUCell BaalChIP BAC bacn BADER BadRegionFinder BAGS ballgown bamsignals banocc basecalQC BaseSpaceR Basic4Cseq BiASIS BasicSTARSeq BatchQC BayesKnockdown BayesPeak baySeq BBCAnalyzer BCRAN Beachmat beadarrayR beadarraySNP BeadDataPackR BEARscc BEAT BEclear bgafun BGeeDB BGmx bgx BHC BicARE BiFET BiGGR bigmelson bigmemoryExtras bioassayR Biobase biobroom bioCancer BioCaseStudies BiocCheck BiocFileCache BiocGenerics biocGraph BiocInstaller BioCor BiocParallel BiocStyle biocViews BiocWorkflowTools biocDist biomatformat BioMVCClass biomvRCNS BioNet BioQC BioSeqClass biosigner Bioscripts biosvd biotmle biovizBase BiRewire birta birte BiSeq BitSeq blima BLMA bnbc BPBPRAIN BrainStars branchpointer bridge BridgeDbr BrowserViz BrowserVizDemo BSgenome bsseq BubbleTree BufferedMatrix Methods BUMHMM bumphunter BUS CAFE CAGER CALIB CAMERA cancerR cancerclass CancerInSilico CancerMutationAnalysis cancerSubtypes CAnD caOmicsV Cardinal casper CATALYST Category categoryCompare CausalR cbaf ccmpl CCPROMISE crepe cellbaseR cellGrowth cellHTS2 cellity CellMapper CellNOpT rcellscape cellTree CEMITool CexoR CFASy CGEN CGHbase CGHcall cghGhnormalizer CGHregions ChAMP CHARGE charm ChemmineOB ChemmineR Chicago chimera chimeraVz ChIPAnalyses ChIPComp chipenrich ChIPExpoQual ChIPpeak Anna ChIPQC ChIPseeker chipseq ChIPseeker ChIPSpike ChIPsim ChIPxpress chopsticks chroGnDraw ChromHeatMap chromPlot chromstar chromswitch chromVAR CHRONOS CINdex cisPath ClassifyR cleanUpdTSeq cleppda clipper Clomial Clonality clonotypeR cld clstltis clustComp clusterExperiment ClusterJudge clusterProfiler clusterSeq clusterSignificance clusterStab CMA cn CNAnorm CNEr CNORdt CNORfeeder CNORfuzzy CNORode CNPBayes CNTools cnvGSA CNVPanelizer CNVrd2 CNVtools cobindR CoCiteStats codelink CODEX coexnet CoGPS COHCAP coMET COMPASS CompcodeR compEpitools CompGO ComplexHeatmap CONFESS ConsensusClusterPlus consensusOV consensusSeeker R contiBAIT connume convert copa copynumber Copywriter R CoRegNet Cormotif CORREp coseq cosmiq COSNet CountClust covEB coverageView cvnRNA cpvSNP cqN CRISPRseek crispseekplus CrispRVariants crimm crossmeta CSAR csaw CSSP ctc CTQderquer ctsGE cummerBrnd customProDB CVE cycle cydar cytokit cytolib CytoML dada2 dazLogo daMA DaMiRseq DAPAR DART DA BCBI dcGSA DChIPRep ddCt debrowser DECIPIER DEComplexDisease DeconRNASEq DEDS DeepBlueR deepSNV DEFormats DEGraph DEGreport DEGseq DelayedArray Darley DelayedMatrixStats deltaGseg DEMAND DEP derfinder derfinderHelper derfinderPlof DESeq ESeq2 destiny DEsubs DEXSeq dexus DFP DiffBind diffloop diffuStats diggit Director DirichletMultinomial discordant dks DMCHMM DMRCaller DMRCat DMRforPairs DMRCat Scan DNABarcodes DNAcopy DNAshapeR doppelgangR DQOTL Dscheda DOSE drawProteins DRIMSeq DriverNet DropletUls DrugVsDisease dSimer DSS DTA dualKS DupChecker dupRadar dyebias DynDoc EasyqcR easyRNASeq EBarrays EBcoexpress EBImage EBSeq EBSeqHMM ecolitk EDASeq EDDA edge edgeR eeGAD EGSEA eiR elSA ELBOW ELMER EMDomics EmpiricalBrownsMethod ENCODEExplorer ENmix EnrichedHeatmap EnrichmentBrowser ensembleDB ensembleVEP ENVISIONQuery EpiDISH epigenomix EpiNEM epivirZ epivirZChart epivirZData epivirZServer epivirZStandard ccashboard erma esATAC esetV esudysys EventPointer ExiMiR exomeCopy exomePeak ExpressionHub ExpressionAtlas ExpressionView fabia facopy factDesign FamAgg farms fastLiquidAssociation fastseg fCCAC fCI fdrame FEM ffGNet fgsea FindMyFriends FiSHalyseR FitHiC flagme flipflow AF flowBeads flowBin flowCatchR flowCHIC flowCL flowClean flowClust flowCore flowCyBar flowDensity flowFit flowFP flowMap flowMatch flowMeans flowMerge flowPeaks flowPlots flowQ flowW flowRepositoryR FlowSOM flowStats flowTime flowTrans flowType flowUtils flowViz flowVS flowWorkspace fmcsR focalCall FourCSeq FRGEpistasis frmaTools FunChIP FunciSNP funtooNorm GA4GHclient GA4GHshiny gaga gage gagggle gaia GAprediction garfield gacpc gcatest gCMAP gCMAPWeb gCrisprTools grcma GDCRNATools gdsfmt gecg GEM genArise genbankr GeneAnswers geneAttribution GeneBreak geneClassifiers GeneExpressionSignature genefilter genefu GeneGA GeneGeneIntR GeneMeta GeneNetworkBuilder geneOverlap geneplast genePlotter geneRecommender GeneRegionScan geneRxCluster GeneSelectMMD GeneSelector GENESIS geNetClassifier GeneticsDesign GeneticsPed geneXtendeR GENIE3 genoCN GenoGAM genonation GenomeGraphs GenomeInfoDb genomeIntervals genomes GenomicAlignments GenomicDataCommons GenomicFeatures GenomicFiles GenomicInteractions GenomicRanges GenomicScores GenomicTuples Genominator genoset genotypeeval genphen GenVisR GEometadab GEOquery GEoSubmission gerp2Pep gespeR GEWIST GGBase ggbio ggcyto GGTools ggtree girafe GISPA GLAD Glimma GlobalAncova globalSeq globaltest gmapR GMRP GOexpress GOfuncR GOfunction GoogleGenomics GPro protein goProfiles GOSeqSim goSeq GOSim GGOSim goSTAP GOsummaries GOTHiC goTools gpls gprege gQLTBase gQLTstats graph GraphAlignment GraphAT graphite GraphPAC GRENTIS GreyListChIP GRmetrics groHMM GRridge GSALightning GSAR GSCA GSEAlm gsean GSReg GSRI GSVA gtrellis GUIDEseq Gqviz gwascat GWASTools h5vc hapFabia Harman Harshlight HDF5Array HDTD heatmaps Heatplus HelloRanges HELP HEM hiAnnotator HIBAG HiCcompare hicrep hierGWAS HilbertCurve HilbertVis HilbertVisGUI hiReadsProcessor HITC hmdbQuery HMmcopy hapoch hpc TqPCR HTSanalyzeR HTSeqGenie htSeqTools HTSFilter HybridMTTest hyperdraw hypergraph iASeq iBBIG ihb iBMQ ICARE lcons iCheck iChip iClusterPlus iCOBRA ideal IdeoViz idogram IdMappingAnalysis idMappingRetrieval IGC IHW illumina imageHTS IMAS Imeta ImmuneSpaceR immunoClust IMPCdata ImpulseDE2 ImpulseDE2 impute InPAS INPower INSPeCT intansy InteractionSet interactiveDisplayBase IntERest InterMineR IntriAmRExport inveRsion INoiseR iPAC IPO IPPD IRanges IrisSpatialFeatures iSeq formSwitchAnalyzeR IsoGeneGUI ISOle isomiRs ITALICS iterativeBMA iterativeBMAsurv iterClust iVAS ivygapSE IWToomics JASPAR2018 joda JunctionSeq karyoploteR KCsmart kebabs KEGGgraph KEGGLincs keggorthology KEGGPROFILE KEGGREST kimod lapmix Lblock LEA LedPred les Ifa limma limmaGUI LINC LineagePulse Linnorm LiquidAssociation lmdre LMGene LOBSTAHS loci2path logicFS logitT Logolas lol LOLA LowMACA LPE LPEadj lpNet lpsymphony lumi LVSmirNA LymphoSeq M3C M3D M3Drop maanova macat aCorrPlot made4 MADSEQ mathtools MAGeCKFlute maigesPack MAIT makecdfenv MANOR manta MantelCorr mapKL maPredictDSC mapscape marray maSigPro maskBAD MassArray massR MassSpecWavelet MAST matchBox MatrixRider matter MaxContrastProject BAmetryl MBASED MBCB mBPCR MBtest mcaGUI MCbiclust MCReestimate mCSEA mdgs mdqc MEAL MeasurementError MEDIPS MEDME MEIGOR MergeMaid Mergeomics MeSHdb meshes meshr messina metaArray Metab metabomxR MetaboSignal metraCCA metraCyto metagenomeFeatures metagenomeSeq metahdep metaMS MetaNeighbor metaSeq metaseqR metavirz MetCirc methimpute methInheritSmt MetPedMet PedTargetedNGS methViewS view methyAnalysis MethylAid methylInheritance methylKit MethylMix methyLPipe MethylSeekR methylum methylwm mfa Mfuzz MGFM MGFR mgsa MiChip microbiome microRNA MiGSA mimager MIMOSA MineICA minet minfi MinimumDistance MiPP MIRA MiRaGe miRBaseConverter miRcompm miRIntegrator miRLAB miRmine miRNAmiC miRNAPath miRNAtap miRsponge Mirsynergy missMethyl mitoODE MLInterfaces MLP MLSeq MMDF2 MmPalateMiRNA MODA mogsa monodel MoonLightR MoPS mosaics motifbreakR motifcounter MotifDb motifmatchr motifRG motifStack MotIV MPFE mpra mQTl msa MSGGui MSGPlus msmsEDa msmsTests MSnbase MSnID msPurity MSstats Mulcom MultiAssayExperiment multiClust MultiDataSet MultiMed multiOmicsViz multiscan multitest muscle MutationalPatterns MVCClass mvGST MWASTools mygene myvariant mzID AFinder NanoStringDiff NanoStringQCPro NarrowPeks ncdfFlow NCgraph nidx nem netbenchmark netbioweb netbioweb netpathMiner netprioR netReg netresponse NetSAM networkBMA NGScopy nnNorm NOISeq nondetects normalize450K NormqPCR normr npGSEAN NT nucleoSim nucleR nudge NuPoP occgene OCplus odseq OGSA oligoClasses OLINOLgui omicade4 OmicCircos omiclpt RomicRepose OmicsMarkR omicsPrint Onassis oncomix OncoScore OncoSimulR oneSENSE ontoCAT ontoProc openCyto openPrimeR oboPrimeR UroRipa Mate oposSOM oppar OPWeight OrderedList Organism OrganismDbi OSAT Oscope OTUbase OutlierD PAA PADOG paircompvz panR panelC PApnBuilder panP PanVizGenerator PAPi pargms parody Path2PPI pathifier PathNet PathoStat pathprint pathRender pathVar pathview PathwaySplice paxtoolsR Pbbase pbcmc pcaExplorer pcaGoPromoter pcaMethods PCAN pco2 PCpheno pcnx pdInfoBuilder PECA pepStat pepXMLTab PGA pgca PGSEA PharmacoGx phenoDist phenopath phenoTest PhenStat phir phosphonormalizer phyleoPj piano pickgene PICS Pigengene PING pint pkdDepTools plateCore plethy plgem plier PLPE plrs plw pmrn podkatt posgos polyester Polyfit POST TPPInfer ppIStats ppSFinder prada prebs PREDA prediction preprocessCore Prize probBAMR PROcoil ProCoNa proFIA profileScoreDstd progeny pRoloC GUI PROMISE PROPER PROPS Prostar prot2D proteinProfiles ProteomicsAnnotationHubData proteoQC ProtGenerics PSEA psichomics PSICQUIC psenger2r purma PureCN pvcv Pviz PWEMrich pwicmetrics QDNAseq qcrNorm qgrpm qrcs qseaf QUALIFIER quanto quantsmooth QuartPAC QuasR QuaternaryProd QUBIC usagae qvalue RC3PCT R3Cseq R453Plus1Toolbox R4RNA RaggedExperiment rain rama RamIGO ramwas randPack RankProd RareVariantR RbcBook1 RBGL RBioinf RBioPaxParser RBM Rbowtie Rbowtie2 rbsrv Rcaude RCAS RCASPAR rclmliner rCGH Rchmcpp RchyOptimyx RcisTarget Rcpi RCy3 RCy5 RDAVIDWebService rDGdb Rdiosop RDRToolbox ReactomePA readat ReadqPCR reb recount reDeR REDseq RefNet RefPlus regioneR regionReport regsplice REMP Reptoids ReportingTools ReQON restfulSE rexosome rfPred rGADEM RGALaxy RGML RGraph2js Rgraphviz rGREAT RGSEA rgsepR rhdf5 client Rhdf5lib Rhitslb rHVDM RiboProfiling riboSeq rlmPort Ringo RIPSeeker Risa RITAN RIVER RJMCNCNucleosomes RLMM Rmagpie RMassBank rMAT RMr RNAinteract RNAither RNAPrR rnaseqcomp rnSeqMap RNASeqPower RnaSeqSampleSize RnBeads Rnits roar ROC Roleswitch rols romia ROntoTools roOTS RPA RProtoBufL RpsiXML rpX Rqc rqT rqbic RRHO Rsmarts tools rsbml RSFFreader Rsubread RSVsim RTCA RTCGA RTCGAToolbox RTN RTNduals RTNsurvival RTopper rtracklayer RtreeM rTRM rUnubic RUVcorr RUVnormalize RUVSeq R4Vectors safe sazenhaft SAGX samExploreR sampleClassifier SamSPECTRAL sangerseqR SANTA sapFinder savR SBMLR SC Scale4C SCAN scater scd scd scfFeatureFilter scfnd scfnd ScI scmp SCnorm scoreLhvMai scPcne scran scrs SegmentSeq SELEXEN semisup SEPA seq2pathway SeqArray seqbias seqCAT seqCNA seqcombo SeqGSEA seqLogo seqPattern seqplots SeqVarTools sevenbridges SGSeq shinyMethyl shinyTANDEM ShortRead SIctools sigaR SigCheck SigFuge siggenes sights signeR sign

Code Usage and Contributions

Software Download Rates

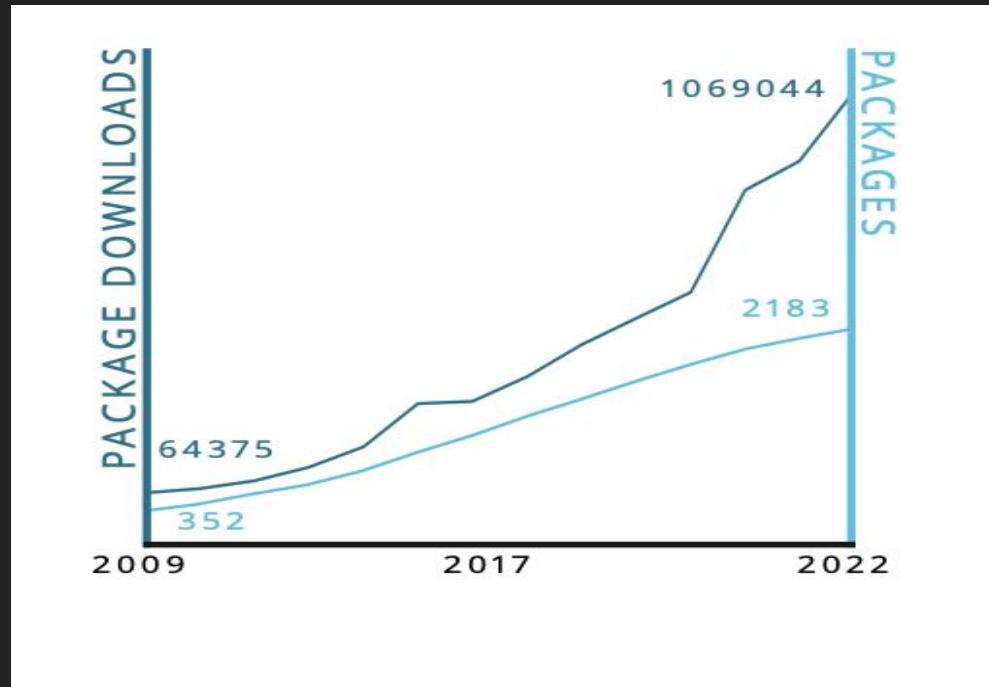
Indicator of usage and demand.

Community Code Contributions

Package contributions indicate collaborative development and innovation.

Challenges

Capturing the full picture. For example, downloads from mirrors



Community growth



100,000s users, clustered by location

Code Usage and Contributions

Software Download Rates

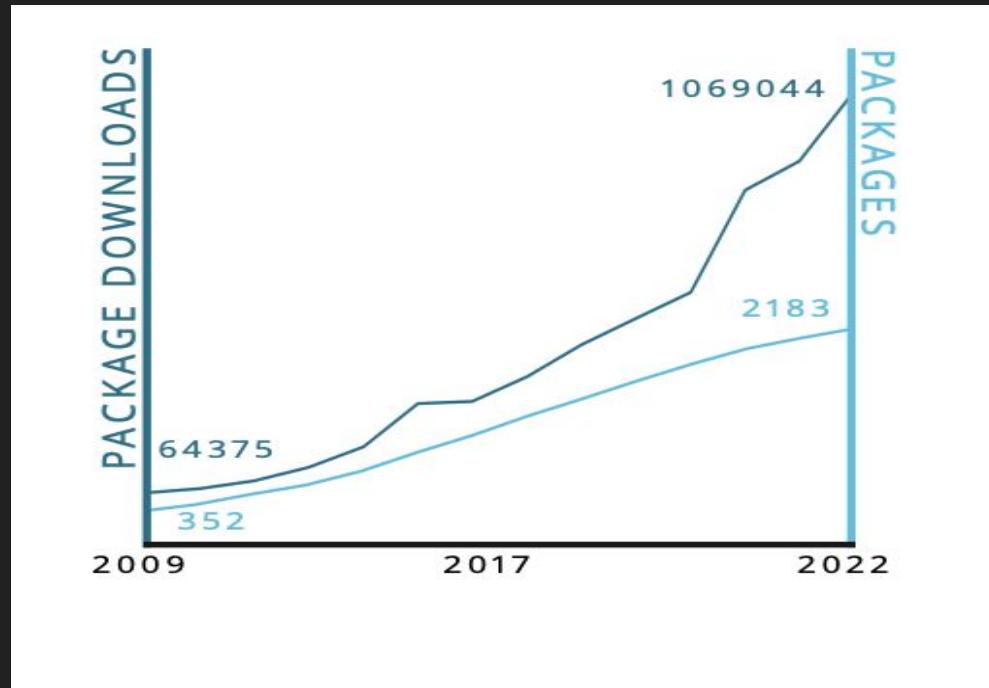
Indicator of usage and demand.

Community Code Contributions

Package contributions indicate collaborative development and innovation.

Challenges

Capturing the full picture. For example, downloads from mirrors



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*

The screenshot shows the Bioconductor website's "All Packages" page. At the top, there is a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation, the page title is "All Packages". A sidebar on the left lists categories under "Bioconductor version 3.6 (Release)": 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), and Phenomics (8). The main content area displays a table titled "Packages found under Software:". The table has columns for Package, Maintainer, and Title. The packages listed are:

Package	Maintainer	Title
a4	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Umbrella Package
a4Base	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Base Package
a4Classif	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Classification Package
a4Core	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Core Package
a4Preproc	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Preprocessing Package
a4Reporting	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Reporting Package
ABAEnrichment	Steffi Grote	Gene expression enrichment in human brain regions
ABarray	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (A81700) gene expression data.
ABSSeq	Wentao Yang	ABSSeq: 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 <ssdavis2@mail.nih.gov>

Maintainer: Sean Davis <ssdavis2@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	Use the check boxes to show only packages with the selected status types.
ERROR	INSTALL, BUILD or BUILD BIN of package failed, or CHECK produced errors	<input checked="" type="checkbox"/>
WARNINGS	CHECK of package produced warnings	<input checked="" type="checkbox"/>
OK	INSTALL, BUILD, CHECK or BUILD BIN of package was OK	<input checked="" type="checkbox"/>
NotNeeded	INSTALL of package was not needed (click on glyph to see why)	<input checked="" type="checkbox"/>
skipped	BUILD or BUILD BIN of package was skipped because the BUILD step failed	<input checked="" type="checkbox"/>
NA	BUILD, CHECK or BUILD BIN result is not available because of an anomaly in the Build System	<input checked="" type="checkbox"/>

Package propagation status is indicated by one of LEDs

YES:	Package was propagated because it didn't prevent bumping
NO:	Package was not propagated because of a problem (impossible dependencies, or version lower than what 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 no longer maintained.

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

Package 1/1554	Hostname OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
a4 1.27.0	malbec2 Linux (Ubuntu 16.04.1 LTS) / x86_64	OK	OK	OK	
Tobias Verbeke					
Last Commit: 52d6c2b					
Last Changed Date: 2017-10-30 12:52:11 -0400					
Package 2/1554	Hostname OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
a4Base 1.27.0	malbec2 Linux (Ubuntu 16.04.1 LTS) / x86_64	OK	OK	OK	
Tobias Verbeke					
Last Commit: 72d568e					
Last Changed Date: 2017-10-30 12:52:11 -0400					

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

My: messages • votes • posts • tags • following • bookmarks Sean Davis • 21k

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION LATEST 6 NEWS JOBS TUTORIALS TAGS USER

Limit Sort Search

0 votes	0 answers	2 views	Cannot install Rhtslib on Mac OS 10.13	installation compilation error rhstlib	written 2 minutes ago by Ryan C. Thompson • 6.5k
0 votes	1 answer	35 views	No CNAs or SNVs in results	purecn	written 21 hours ago by twtoal • 0
0 votes	0 answers	8 views	monocle estimateSizeFactors give Inf for all values	maseq	written 1 hour ago by jonesara770 • 10
0 votes	1 answer	14 views	Row clustering featureAlignedHeatmap function (ChipPeakAnno package)	chippeakanno heatmap	written 2 hours ago by gdeniz • 0 • updated 2 hours ago by Ou, Jianhong • 1.0k
0 votes	0 answers	7 views	Metabolite identification package	metabolomics xcms massspectrometry	written 2 hours ago by johnhamre3 • 0
1 vote	1 answer	10 views	LaTeX Error with BiocWorkflowTools	biocworkfowtools	written 3 hours ago by shbrief • 10 • updated 2 hours ago by Mike Smith • 2.6k
0 votes	0 answers	6 views	msa output formats for use down stream to create phylogenetic trees	R bioconductor phylogenetic	written 2 hours ago by cav3gh • 0
0 votes	1 answer	40 views	Results counts post DESeq same raw counts	deseq2 results	written 13 hours ago by A • 0 • updated 3 hours ago by Michael Love • 17k
0 votes	1 answer	14 views	Problem installing Minfi on Cluster	minfi centos local installation hpc	written 5 hours ago by Goku • 0 • updated 5 hours ago by Kasper Daniel Hansen • 6.3k
0 votes	1 answer	27 views	Compare groups of different RNAseq sets	limma batch effect	written 7 hours ago by b.notia • 290 • updated 5 hours ago by Aaron Lun • 19k

Recent... Replies
• A: No CNAs or SNVs in results
0
• C: Results from chippeakanno
Michael Li
• A: Row clustering featureAlignedHeatmap function (ChipPeakAnno package)
Jianhong Ou
• C: Results from monocle
• A: LaTeX Error with BiocWorkflowTools
Mike Smith • 2.6k

Votes
• LaTeX Errors
• DESeq2
• Bioconductor
• What's new
• EdgeR
• Scholarships
• Popular Questions
1.4k
• Scholarships

Awards • All
• Teacher Grants
• Scholar Grants
• Scholar Grants
• Scholar Grants
• Popular Questions
1.4k
• Scholar Grants

Locations • All
• Dana-Farber
USA, 3 months ago
• The Scripps
CA, 9 months ago
• Walter and
Research, 10 months ago

Support Site: <https://support.bioconductor.org/>

The screenshot shows a web browser window for the Bioconductor Forum at <https://support.bioconductor.org>. The page displays a list of posts from users like James W. MacDonald, Hervé Pages, and Gordon Singh. Posts are categorized by tags such as Bioconductor, Transcriptomics, DifferentialExpression, DESeq2, RNAseqData, Bio2023, and Awards. The interface includes a search bar, sorting options, and a sidebar with links for Ask a Question, Latest, News, Jobs, Tutorials, Tags, and Users.

Recent Posts:

- problem while retrieving the genes by using p value and log2fc as conditions (Bioconductor | Transcriptomics | DifferentialExpression | DESeq2) - updated 4 minutes ago by James W. MacDonald 62k + written 15 minutes ago by streetmara72 0
- Can the limma package be applied to Log2 RUV-normalized data? (RNAseqData | MicroarrayData | limma) - 1 hour ago pg45863 0
- News: Last Week to Nominate for Bio2023 Awards! (Bioconductor | Bio2023 | Recognition | Awards) - 3 hours ago shepherd 3.4k
- News: Junior Developer or New Package Developer Award (Bioconductor | Recognition | Bio2023 | Awards) - 11 days ago + updated 3 hours ago shepherd 3.4k
- News: Bioconductor Community Engagement, Outreach, and Diversity Award (Awards | Bioconductor | Bio2023 | Recognition) - 29 days ago + updated 3 hours ago shepherd 3.4k
- News: Bioconductor Long-term Contribution Award (Bio2023 | Bioconductor | Awards | Recognition) - 21 days ago + updated 3 hours ago shepherd 3.4k
- News: Nominations for Bioconductor 2023 Awards! (Awards | Bioconductor | Bio2023 | Recognition) - 5 weeks ago + updated 3 hours ago shepherd 3.4k
- DESeq2 about comparision of gene expression change on two-time points between two group samples. (DESeq2 | RNAseqData | iopsR | limma | DifferentialExpression) - 3 hours ago simon 0
- Could not find function "read_block_OLD" (DelayedArray | BioKNN) - updated 13 hours ago by Hervé Pages 16k + written 18 hours ago by xliuwen.zheng 0
- News: course - single-cell RNAseq data analysis with R and Bioconductor (Workshop | dCNAq) - 18 hours ago info 10
- Keep getting this error "Warning: unable to access index for repository?" (CATALYST) - updated 1 day ago by Hervé Pagès 18k + written 4 days ago by Hyosun 10
- Installing densvis on a Singularity container (densvis | docker) - updated 15 hours ago by jiazhou0116 0 + written 25 days ago by James W. MacDonald 62k

Traffic: 666 users visited in the last hour
SummarizedExperiment is now warning on attach about 'read_block' namespace conflict

The screenshot shows a web browser window for the Bioconductor Community at <https://support.bioconductor.org/user/list/order/reputation>. The page displays a grid of 20 user profiles, each with a small profile picture, name, reputation, and visit information. The users are from various countries and institutions, including the United States, Australia, and the European Molecular Biology Laboratory (EMBL). The interface includes a search bar, sorting options, and a sidebar with links for Ask a Question, Latest, News, Jobs, Tutorials, Tags, and Users.

User	Reputation	Last Visit	Location
James W. MacDonald	62k	visited 5 minutes ago	United States
Gordon Singh	47k	visited 2 hours ago	WEHI, Melbourne, Australia
Michael Livne	40k	visited 1 day ago	United States
Aaron Lun	29k	visited 8 hours ago	The city by the bay
Martin Morgan	25k	visited 7 days ago	United States
Seán Davis	21k	visited 9 months ago	United States
Dan Tenenbaum	8.2k	visited 2.3 years ago	United States
Hervé Pages	16k	visited 13 hours ago	Seattle, WA, United States
Wolfgang Huber	13k	visited 17 days ago	EMBL, European Molecular Biology Laboratory
Steve Langmead	13k	visited 12 weeks ago	United States
Guest User	12k	visited 8.7 years ago	
Michael Lawrence	11k	visited 17 months ago	United States
Ryan C. Thompson	7.8k	visited 9 weeks ago	Scopus Research, La Jolla, CA, United States
Seth Falcon	7.4k	visited 8.7 years ago	
Marc Carlson	7.2k	visited 6.8 years ago	United States
Valerie Obersteiner	6.7k	visited 16 months ago	United States
Vincent J. Carey, Jr.	6.6k	visited 15 hours ago	United States
Kasper Daniel Hansen	6.5k	Visited 20 days ago	United States
Naomi Altman	6.0k	visited 2.1 years ago	United States
Mike Smith	6.0k	visited 3 hours ago	EMBL, Heidelberg
Igor Gurevich	5.5k	visited 8.0 years ago	United States
Rory Stark	4.8k	visited 11 days ago	CRG, Cambridge, UK
Benton Carvalho	4.3k	visited 3.2 years ago	Brazil Comprida FNCAMP
Jule Zhu	4.3k	visited 6.7 months ago	United States

Ask questions about packages or data analysis
Thousands of Bioconductor users and maintainers are members

Community Slack: slack.bioconductor.org

The screenshot shows the Slack application window for the 'community-bioc' team. The left sidebar lists various channels, with '#general' being the active channel. The main pane displays messages from users like Maria Doyle, Ajda Prstavec, Michael Kesling, and PN. A message from Michael Kesling is highlighted, asking about querying genes by tissue-specific expression.

File Edit View Go History Window Help

general - community-bioc - Slack

Search community-bioc

general Link to join the slack team - <https://slack.bioconductor.org/>

4 Pinned +

installation or many development suites (cuda devtools, ktools, MSVC, compilers, tools) under windows and that the compilation time under windows is important (15 minutes or more). I am afraid that this discourages many people. I would like to be able to provide the Windows executables directly in the package... Under linux, the compilation seems to me unavoidable because of the risk of incompatibility with the libc library... Is it possible to indicate a compilation under linux and no compilation under windows in the DESCRIPTION file or should I make two different packages?

Thanks for your help!

Maria Doyle 10:16 AM 3 days until the annual (and 1st Bioconductor) Smörgåsbord training course starts! There is still time to register if you want to participate.

Join us for a week of free, online, self-paced #Bioconductor and #UseGalaxy #bioinformatics learning!

May 22-26

<https://gallantries.github.io/video-library/modules/bioconductor>

Ajda Prstavec 11:07 AM joined #general. Also, Michael Kesling and Yuka Takemon joined.

Michael Kesling 2:28 PM Hi everyone. I'm new here. I'm wondering if there's a package or function in bioconductor for querying genes that have a tissue-specific pattern. For most of what I've found on the web, I need to start with a single gene and then browse its attributes. At NCBI Gene, I can see tissue-specific expression for that particular gene. I'd rather perform a query for tissue-specific expression and then get a list of genes back. Is there a way to do that? Thanks very much!

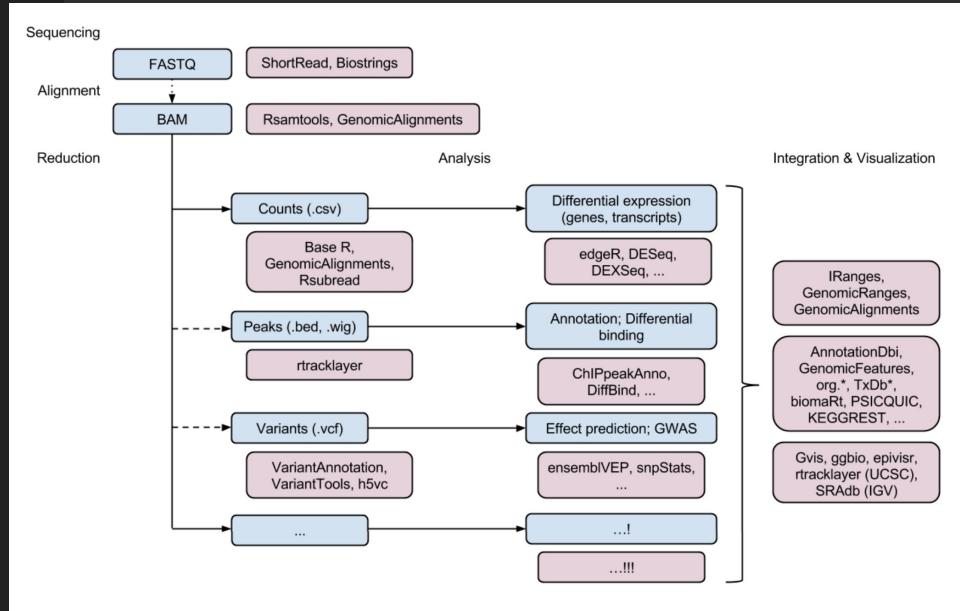
PN 6:58 PM joined #general. Also, Anna Powell joined.

Message #general

+ Aa 😊 @ | ⌂ 🔍

Qualities

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



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

[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](https://doi.org/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@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

GenomicDataCommons

[Home](#) > [Bioconductor 3.19](#) > [Software Packages](#) > [maftools](#)

maftools

Summarize, Analyze and Visualize MAF Files

platforms all rank 128 / 2300 support 1 / 1 in Bioc 7.5 years build warnings updated before release dependencies 16

DOI: [10.18129/B9.bioc.maftools](https://doi.org/10.18129/B9.bioc.maftools)

Bioconductor version: Release (3.19)

Analyze and visualize Mutation Annotation Format (MAF) files from large scale sequencing studies. This package provides various functions to perform most commonly used analyses in cancer genomics and to create feature rich customizable visualizations with minimal effort.

Author: Anand Mayakonda [aut, cre] 

Maintainer: Anand Mayakonda <anand_mt at hotmail.com>

Citation (from within R, enter citation("maftools")):

Mayakonda A, Lin D, Assenov Y, Plass C, Koeffler PH (2018). "Maftools: efficient and comprehensive analysis of somatic variants in cancer." *Genome Research*. doi:[10.1101/gr.239244.118](https://doi.org/10.1101/gr.239244.118).

maftools

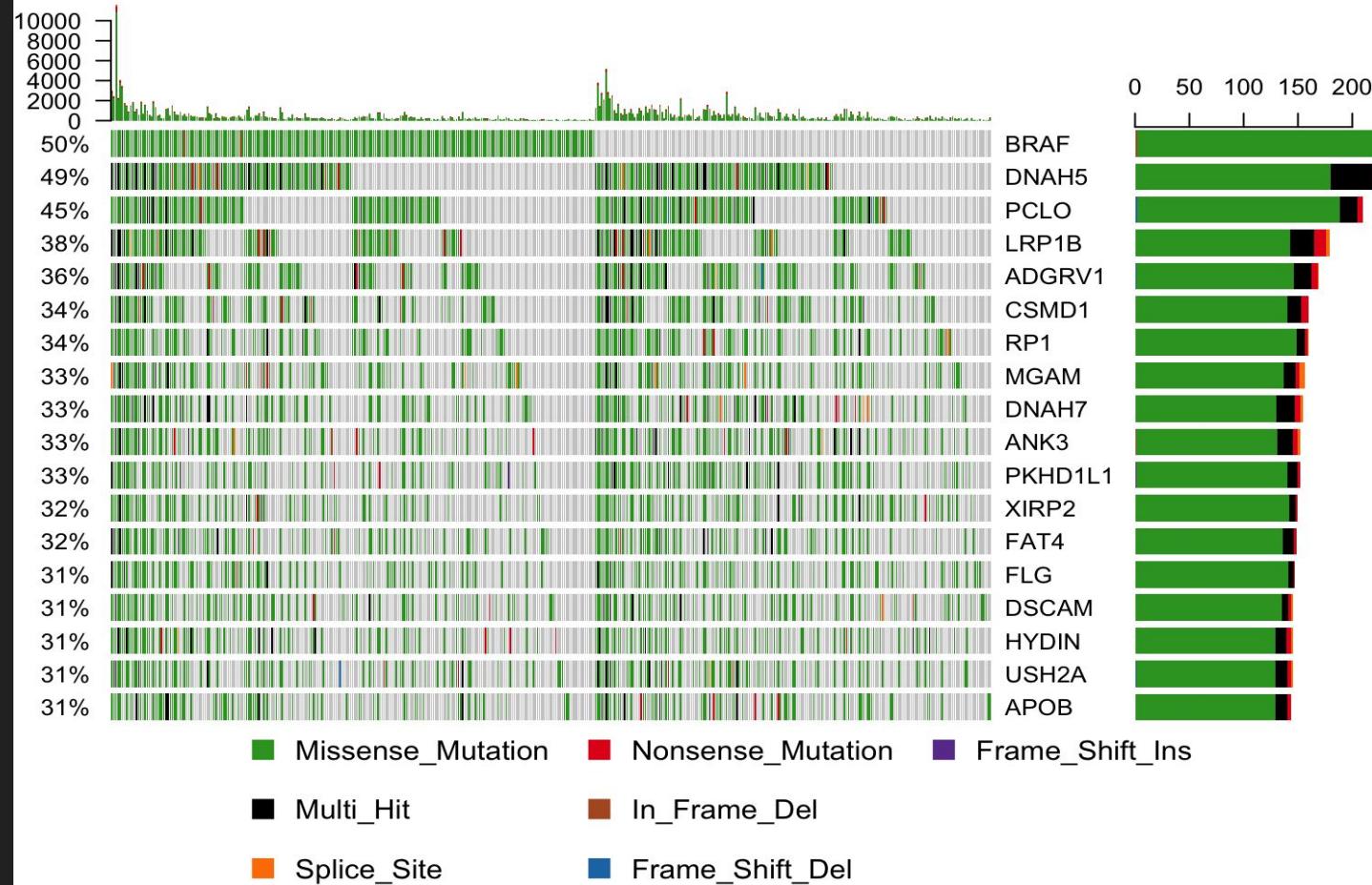
Use the `GenomicDataCommons` package to find and download variants from the TCGA cutaneous melanoma dataset.

```
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() %>%
  gcddata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

```
library(maf-tools)
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() %>%
  gcddata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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

Showing 1 - 20 of 11,265 files

Show More

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.vep.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_realm.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.vep.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
	controlled	1d0ed301-414c-4945-a65b-5bfb4360d65.FPKM.txt.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	535.3 KB	0
	controlled	28ace18-95c5-4690-8afa-814655399ca7.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_realm_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-bbfff51461313.vep.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-8540-2d37d8432f2c.vcf.gz	1	TCGA-SKCM	Simple Nucleotide Variation	VCF	78.87 KB	0
	controlled	C828.TCGA-EE-A17X-10A-01D-A199-08.2.gdc_realm.bam	1	TCGA-SKCM	Raw Sequencing Data	BAM	9.86 GB	0
	controlled	29bd9ed6-f0a8-4ff9-bb1c-79766f2e2dbe.vep.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() %>%
  gcddata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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

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
CommaCancer Variant Aggregation and Masking 1
VarScan2 Variant Aggregation and Masking 1

Data Format MAF 1

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

Add All Files to Cart Manifest View 470 Cases in Exploration Browse Annotations

Files (1) Cases (470) Primary Site Project Data Category Data Type Data Format 89.38 MB

Show More

Showing 1 - 1 of 1 files

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

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() %>%
  gcddata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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

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 Masking 1
VarScan2 Variant Aggregation and Masking 1

Data Format MAF 1

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

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

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

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() %>%
  gacdata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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

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

Platform

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

Add All Files to Cart Manifest View 470 Cases in Exploration [Browse Annotations](#)

Files (1) Cases (470) Primary Site Project Data Category Data Type Data Format 89.38 MB

Show More

Showing 1 - 1 of 1 files

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

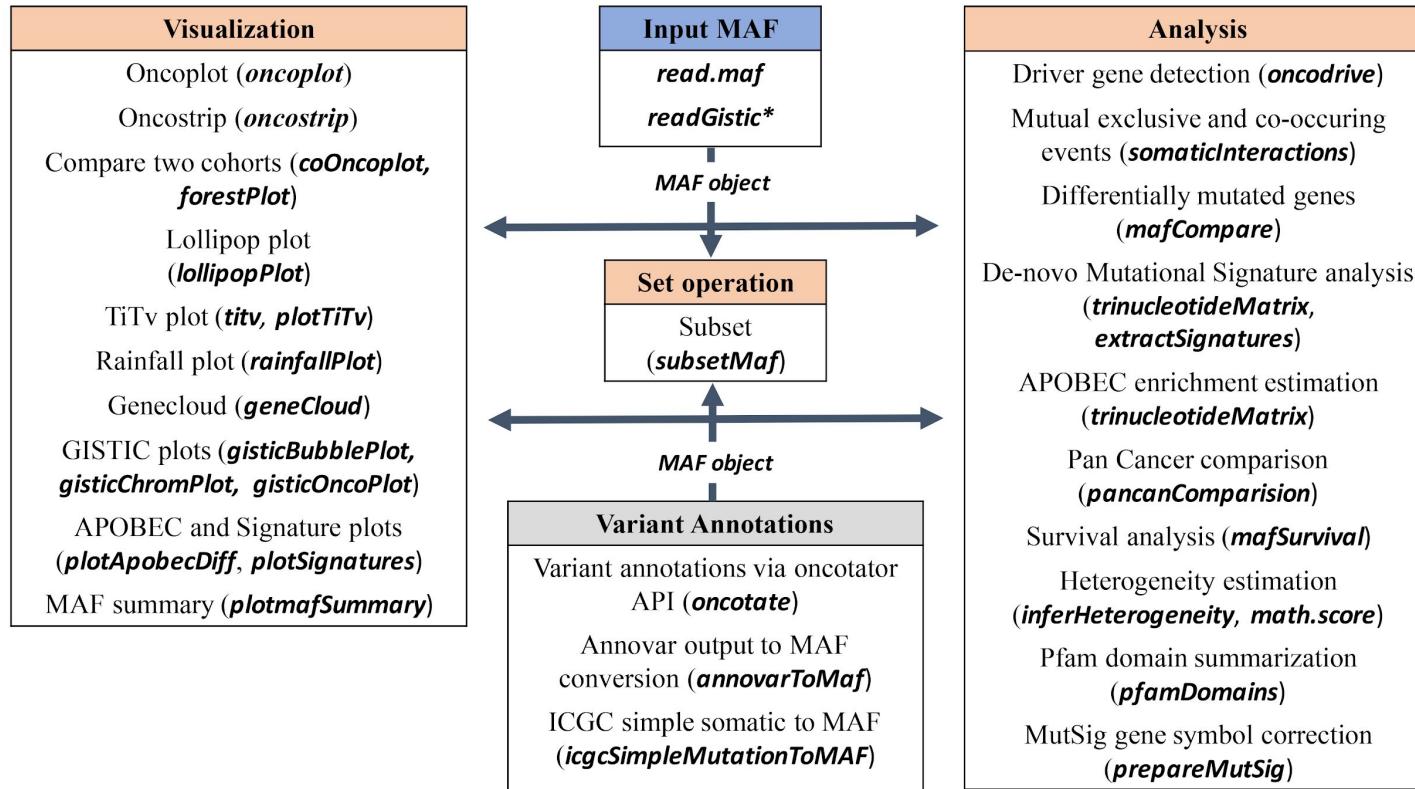
Show 20 entries [1](#)

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() %>%
  gdcdta()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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



GDC Programmatic access example:
Somatic profiles from TCGA melanoma samples (8 lines of code)

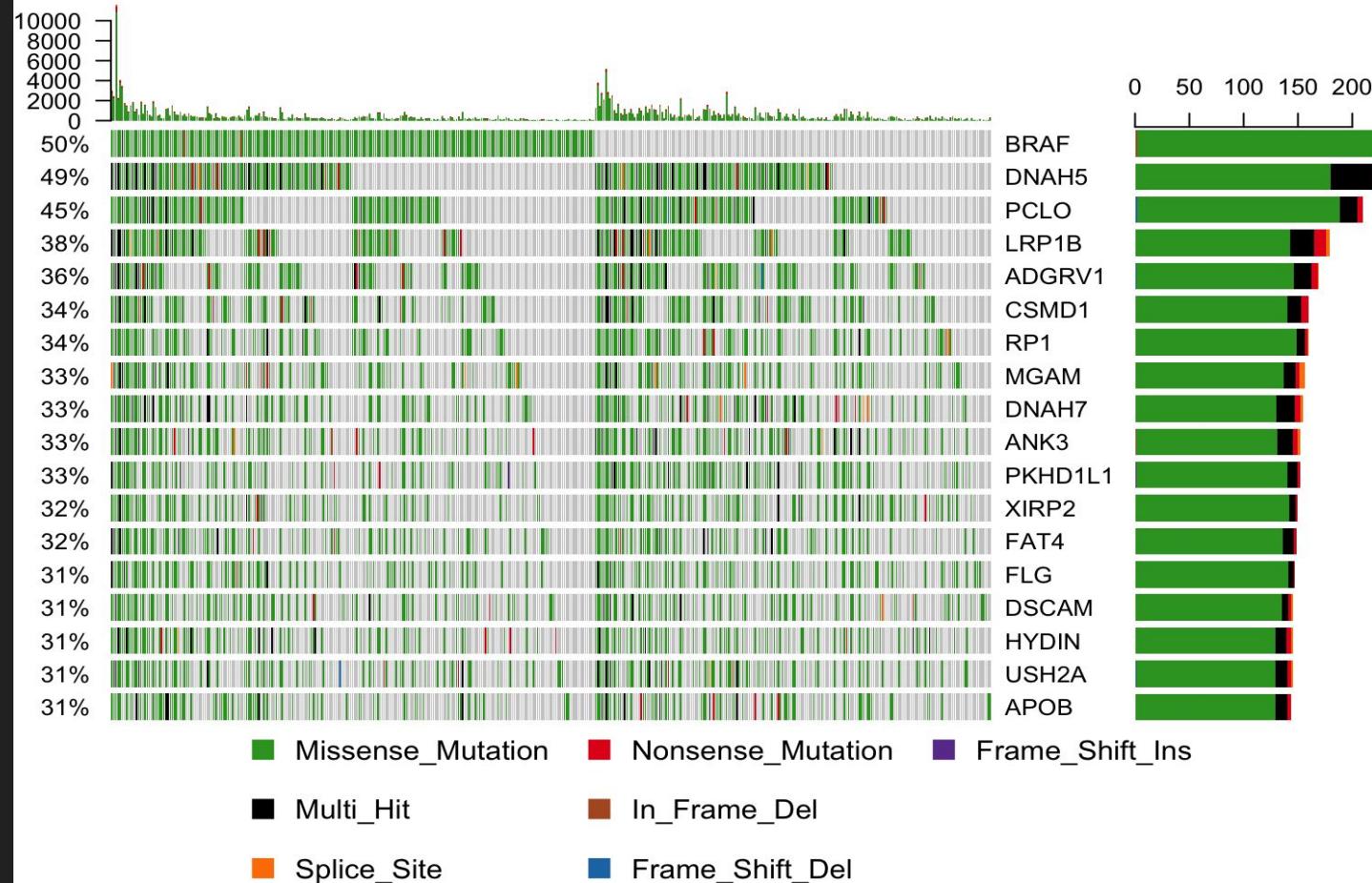
Use the `GenomicDataCommons` package to find and download variants from the TCGA cutaneous melanoma dataset.

```
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() %>%
  gcddata()
```

And now take those data directly to `maf-tools` for analysis and visualization.

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

Altered in 424 (90.79%) of 467 samples.



Variant Type

Variant Classification

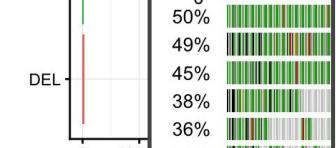
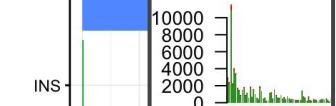
SNV Class

SNP

INS

DEL

Altered in 424 (90.79%) of 467 samples.



of Variants

-log₁₀(fdr)

0.0

0.5

1.0

1.5

Mutations

1

0.5

0.0

0

NRAS[2]

BRAF: [Somatic Mutation Rate: 49.89%]
NM_004333

198

198

198

198

198

198

198

198

198

198

Raf_RBD

C1

PKAlike

Each with 1-2

- In_Frame_Del
- Missense_Mutation
- Nonsense_Mutation

769

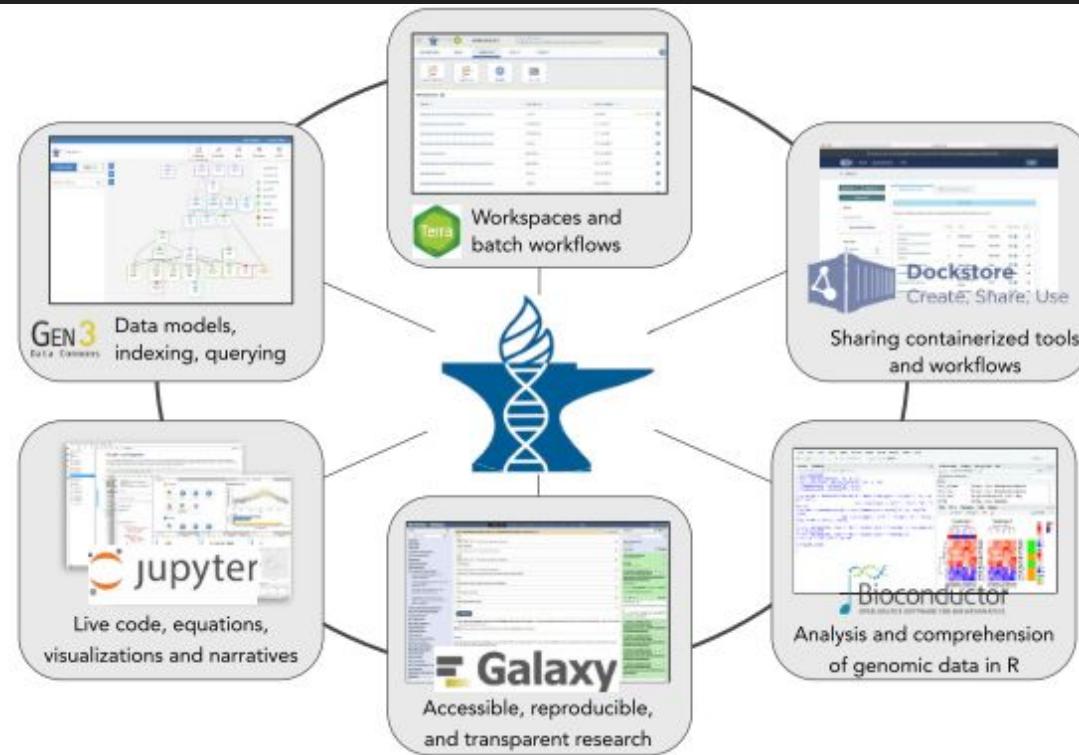
Qualities

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



NHGRI's AnVIL – inverting the data sharing model

cost controls,
security, data
volumes,
avoiding bitrot



Consortium	Datatypes	Cohorts	Samples	Participants	Size (TB)
1000 Genomes Project (1KGP)	WGS	1	3,202	3,202	72.98
Centers for Common Disease Genomics (CCDG)	WGS, WXS, Clinical Phenotypes	198	272,306	256,318	2,624.12
Centers for Mendelian Genomics (CMG)	WGS, Clinical Phenotypes	41	20,706	16,599	97.89
Convergent Neuroscience	WGS	2	304	300	5.32
Genotype-Tissue Expression (GTEx v8)	WGS, RNAseq	1	17,382	979	182.14
Human Pangenome Reference Consortium (HPRC)	Short & long-read WGS	1	57	47	223.47

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) [f](#) [t](#)

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>.

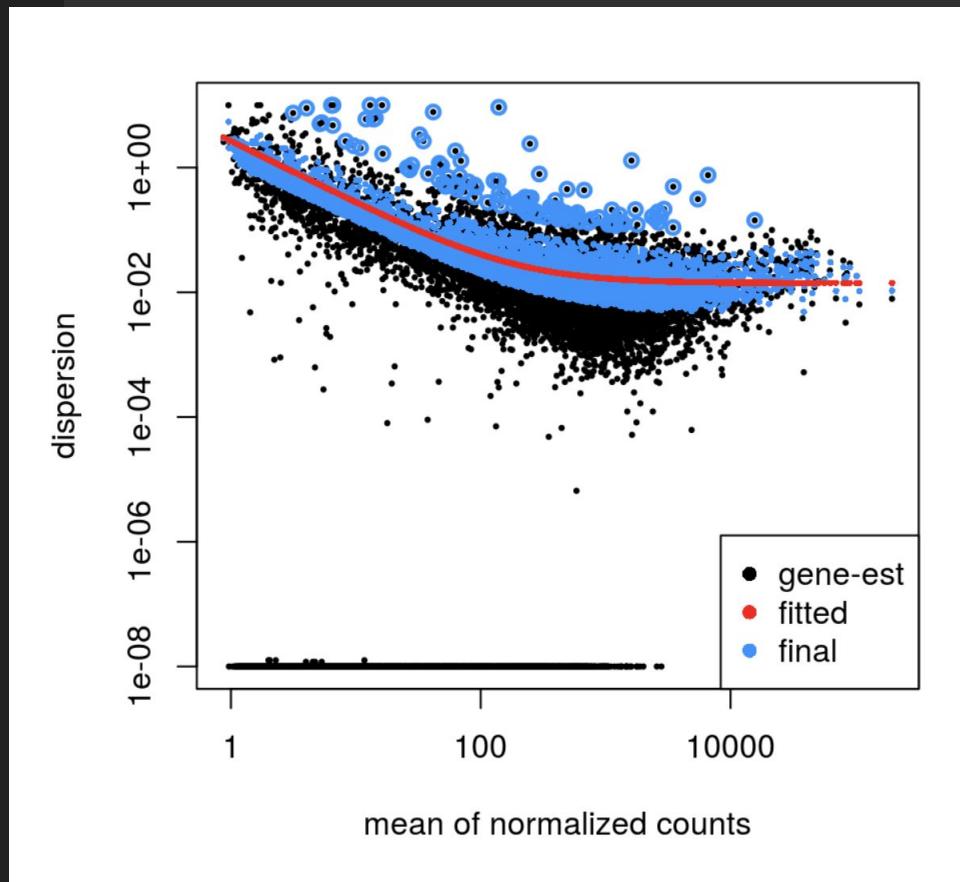
Installation

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 / Contributions'. At the top, there are tabs for 'Code' (selected), 'Issues 24', 'Pull requests 0', 'ZenHub', 'Projects 0', 'Wiki', and 'Insights'. A 'Watch' button with a count of 18 is also visible. Below the tabs, the repository name 'bioconductor' is shown along with statistics: 45 commits, 1 branch, and 0 releases. A dropdown menu shows the current branch is 'master'. There are buttons for 'New pull request', 'Create new file', and 'Upload files'. A list of files includes 'CONTRIBUTING.md', 'README.md', and 'issue_template.md', each with a brief description. A 'Table of Contents' section lists several items under the heading 'Contributing a Bioconductor Package':

- Contributing a *Bioconductor Package*
- Starting the Submission Process
- What to Expect
- Adding a Web Hook
- Submitting Related Packages
- Additional Actions
- Resources

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

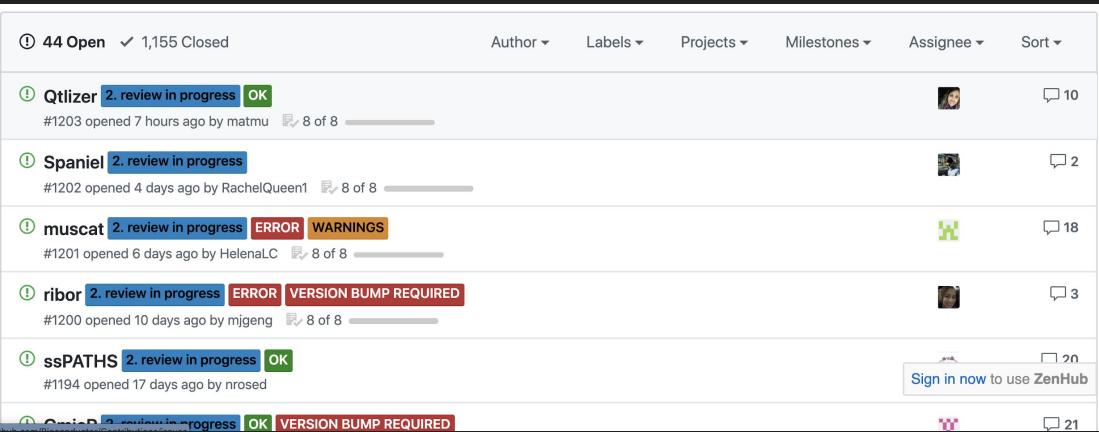
 Search or jump to... / Pull requests Issues Marketplace Explore

Bioconductor / Contributions

Code Issues 44 Pull requests 0 ZenHub

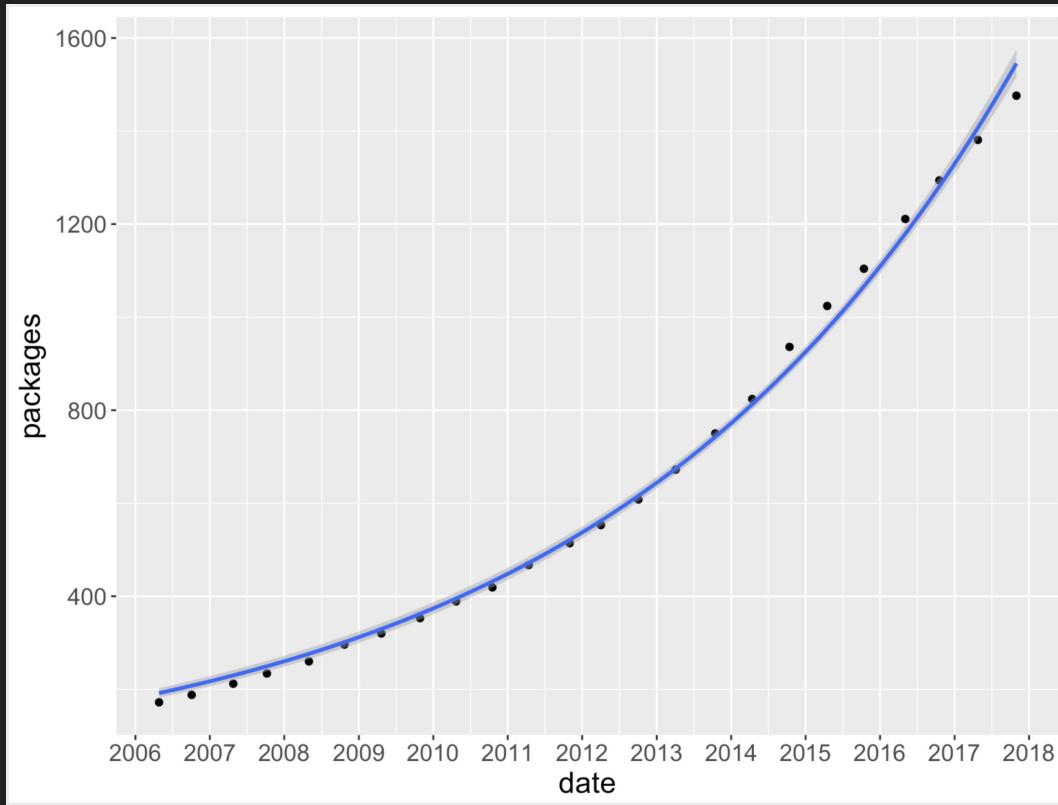
Contribute Packages to Bioconductor

bioconductor



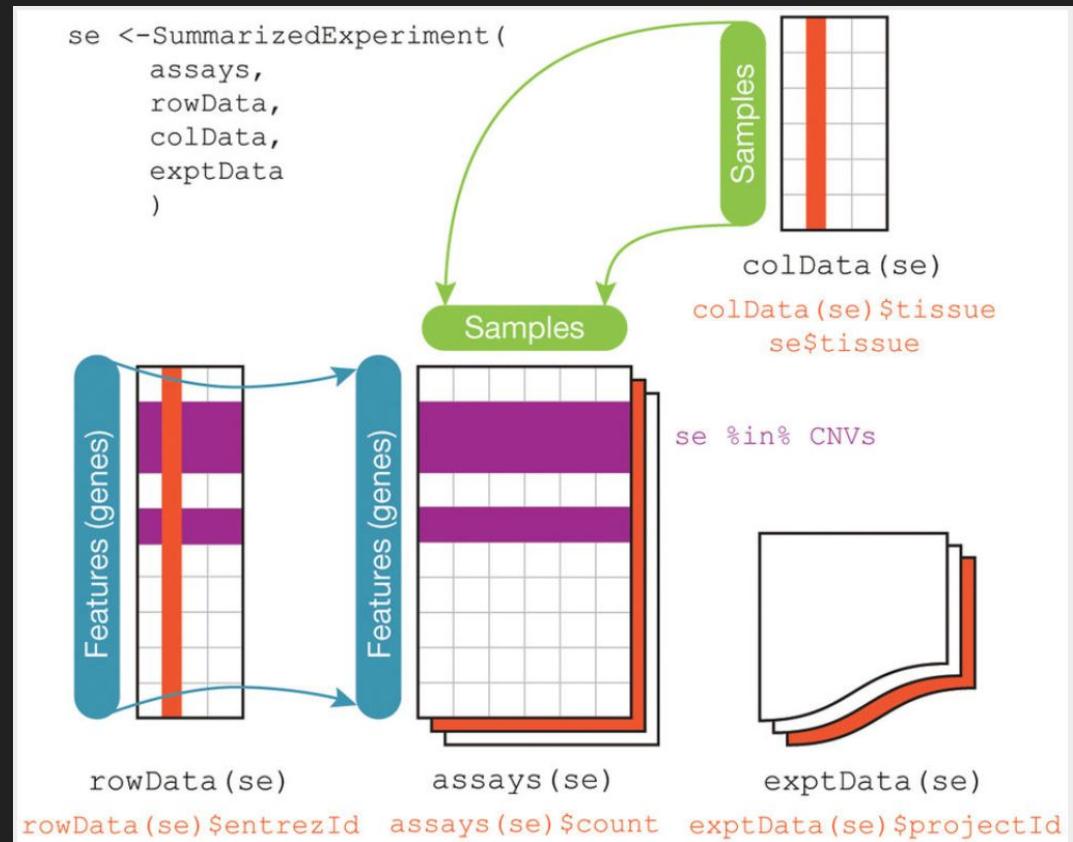
<https://github.com/Bioconductor/Contributions>

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

Genomic annotation, ingestible via R function calls

← → ⌂ https://shiny.sph.cuny.edu/BioHubsShiny/ 67% ⭐

Select A Bioconductor Hub
AnnotationHub ExperimentHub

Download *Hub Resources
Select the rows of interest and then run the code below the table within an R session.

*Hub Metadata
Select rows and click 'Download metadata'
Download metadata
Send metadata
Click 'Send metadata' to interactively add selected rows to the current R session. If viewing the app on a webpage, use the 'Download metadata' button instead to obtain an Rd file of the selections.
Tip: Use the search box at the top right of the table to filter records.

Stop BioHubsShiny

Search through 70762 AnnotationHub resources from 3008 distinct species in Bioconductor
Snapshot Date: 2023-10-20

Show 6 entries Search: ENCODE

HUBID	title	dataprovider	species	taxonomyid	genome	description	coordinate_1_based	maintainer	rdatadateadded	tags
AH5016	ENCODE Pilot	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'ENCODE Pilot'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	encodeRegi...
AH5073	Affy RNA Loc	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'Affy RNA Loc'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	wgEncodeAf...
AH5079	CSDL Small RNA-seq	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'CSDL Small RNA-seq'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	wgEncodeCs...
AH5080	GIS RNA PET	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'GIS RNA PET'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	wgEncodeGi...
AH5084	RIKEN CAGE Loc	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'RIKEN CAGE Loc'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	wgEncodeRi...
AH5088	SUNY SwitchGear	UCSC	Homo sapiens	9606	hg19	GRanges object from UCSC track 'SUNY SwitchGear'	1	Marc Carlson <mcarlson@fhcrc.org>	2013-03-26	wgEncodeSu...

Showing 1 to 6 of 6,964 entries (filtered from 70,762 total entries)

Previous 1 2 3 4 5 ... 1,161 Next

"CSHL small RNA-seq", a UCSC "track" to compute

```
> library(AnnotationHub)
4/32 packages newly attached/loaded, see sessionInfo() for details.
> ah = AnnotationHub()
snapshotDate(): 2023-10-20
> smr = ah[["AH5079"]]
downloading 1 resources
retrieving 1 resource
|=====| 100%
```

```
loading from cache
require("GenomicRanges")
```

```
> smr
GRanges object with 32537 ranges and 2 metadata columns:
```

	seqnames	ranges	strand	name	score
	<Rle>	<IRanges>	<Rle>	<character>	<numeric>
[1]	chr1	566845-566876	-	chr1:566844-566876(u..	20
[2]	chr1	567544-567588	-	chr1:567543-567588(u..	28
[3]	chr1	567549-567602	+	chr1:567548-567602(u..	45
[4]	chr1	568187-568230	-	chr1:568186-568230(u..	96

ExperimentHub ... 'coherent' representations

https://shiny.sph.cuny.edu/BiocHubsShiny/ 67%  Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Bioconductor *Hub Resources

The online shop for AnnotationHub and ExperimentHub Data

Bioconductor Hub About

Select A Bioconductor Hub

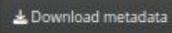
AnnotationHub ExperimentHub

Download *Hub Resources

Select the rows of interest and then run the code below the table within an R session.

*Hub Metadata

Select rows and click 'Download metadata'

 Download metadata

Click 'Send metadata' to interactively add selected rows to the current R session. If viewing the app on a webpage, use the 'Download metadata' button instead to obtain an Rds of the selections.

Search through 7282 ExperimentHub resources from 17 distinct species in Bioconductor

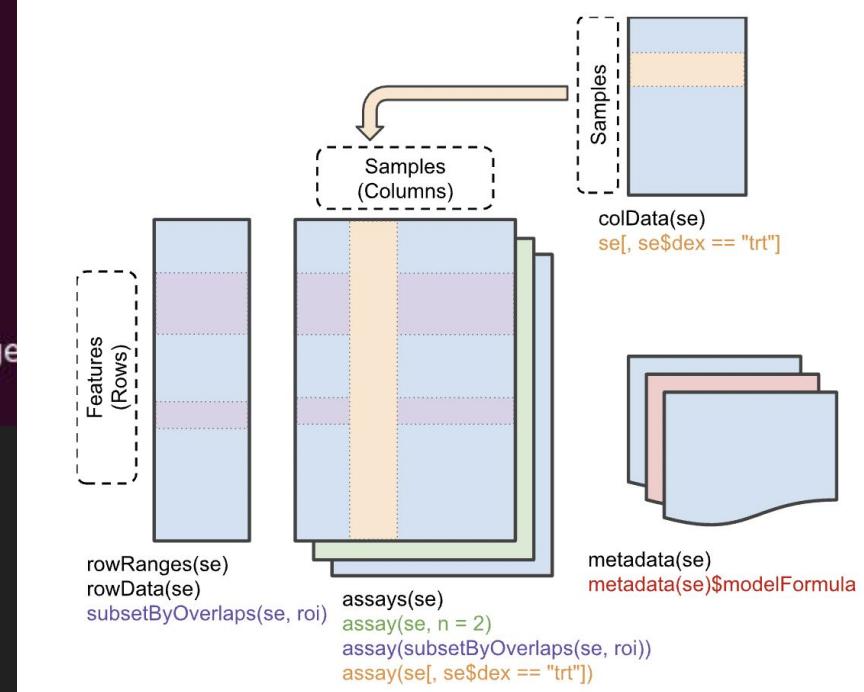
Snapshot Date: 2023-10-24

Show	6	entries	Search: cancer						
HUBID	title	dataprovider	species	taxonomyid	genome	description	coordinate_1_based	maintainer	rdatadateadded
EH8527	cao_esophageal_wgbs_hg19	University...	Homo sapiens	9606	hg19	A HDF5-backed RangedSummarizedExperiment for WGBS Data (CpG sites only) from 10 esophageal squamous carcinomas and 9 matching normal esophageal samples	1	Richard Heery <richardheery@gmail.com>	2023-10-06
EH8528	mcrpc_wgbs_hg38_chr11	University...	Homo sapiens	9606	hg38	A HDF5-backed RangedSummarizedExperiment for WGBS Data for chromosome 11 (CpG sites only) from 100 castration-resistant prostate cancer metastases	1	Richard Heery <richardheery@gmail.com>	2023-10-13

A package in development ... TumourMethData

```
> eso = TumourMethData::download_meth_dataset(dataset="cao_esophageal_wgbs_hg19")
[1] "A HDF5 SummarizedExperiment is already present in /home/vincent/TEMP/RtmpZeW
9 and is being returned"
> eso
class: RangedSummarizedExperiment
dim: 28217883 19
metadata(0):
assays(1): beta
rownames: NULL
rowData names(0):
colnames(19): N1 N2 ... T13 T15
colData names(6): patient_id disease_status ... tnm_stage
survival_months
```

A major element in pursuit of coherence



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](#)

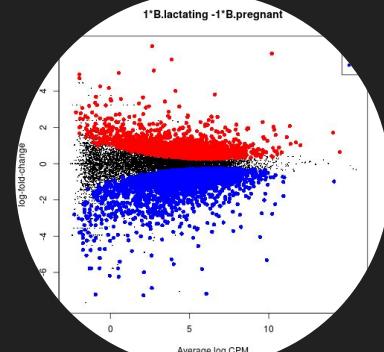
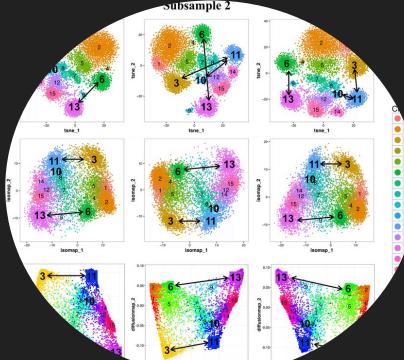
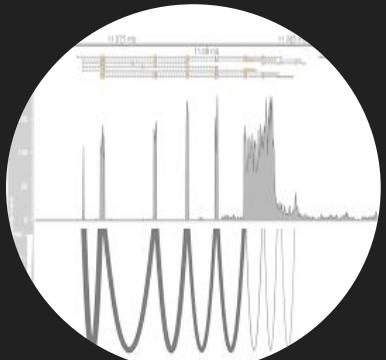
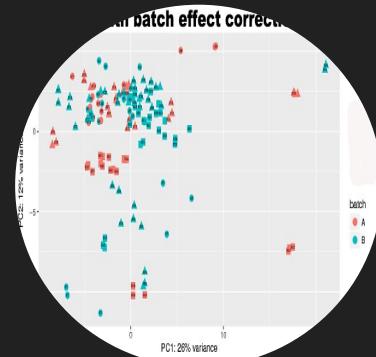
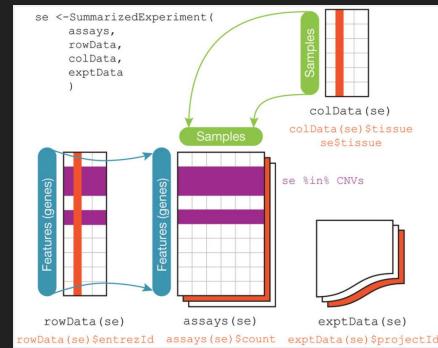
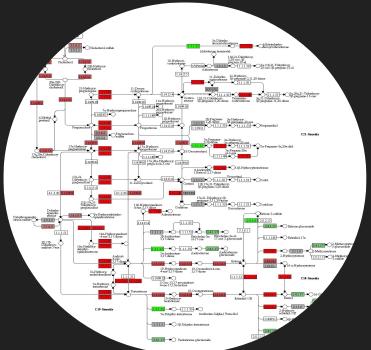
[7 Validating](#)

[8 Example metadata.csv file and more information](#)

Typically WGBS data can't "fit in memory"; HDF5 is

```
> str(assay(eso))
Formal class 'DelayedMatrix' [package "DelayedArray"] with 1 slot
..@ seed:Formal class 'DelayedSetDimnames' [package "DelayedArray"] with 2 slots
... . . . @ dimnames:List of 2
... . . . . $ : int -1
... . . . . $ : chr [1:19] "N1" "N2" "N6" "N7" ...
... . . . . @ seed :Formal class 'HDF5ArraySeed' [package "HDF5Array"] with 7 slots
... . . . . . @ filepath : chr "/home/vincent/.cache/R/ExperimentHub/611bdc2bc8_8590"
... . . . . . @ name : chr "beta"
... . . . . . @ as_sparse: logi FALSE
... . . . . . @ type : chr NA
... . . . . . @ dim : int [1:2] 28217883 19
... . . . . . @ chunkdim : int [1:2] 1000000 1
... . . . . . @ first_val: num NA
> head(rowRanges(eso))
GRanges object with 6 ranges and 0 metadata columns:
  seqnames      ranges strand
  <Rle> <IRanges>  <Rle>
[1]    chr1    10469      *
[2]    chr1    10471      *
[3]    chr1    10484      *
[4]    chr1    10486      *
```

Core value: reuse and interoperability



Bioconductor: Education, Training, and Community

The screenshot shows the Bioconductor website's 'About' page. The header features the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area has five sections: 'About Bioconductor' (describing the tools for analysis and comprehension of high-throughput genomic data), 'Install' (links to R installation and package exploration), 'Learn' (links to courses, support, and documentation), 'Use' (links to software annotation and experiment packages), and 'Develop' (links to developer resources and package submission guidelines). At the bottom, there are links for Support and Events, and a 'Tweets by @Bioconductor' feed.

About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1473 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

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

Install »

Get started with Bioconductor

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

Learn »

Master Bioconductor tools

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

Use »

Create bioinformatic solutions with Bioconductor

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

Develop »

Contribute to Bioconductor

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

Support

Events

Tweets by @Bioconductor

Bioconductor Retweeted

Core infrastructure

Community contribution

Bioconductor Blog

Bioconductor community blog

About Contributing       

Bioconductor community blog

Jul 12, 2024
Maria Doyle

Bioconductor projects funded by CZI EOSS Cycle 6

BIOCONDUCTOR

Announcing the Bioconductor projects funded in the Chan Zuckerberg Initiative EOSS Cycle 6



Chan
Zuckerberg
Initiative

Bioconductor

Categories

All (25)
ARM64 (2)
BioC (3)
Bioconductor (9)
CAB (1)
Diversity (2)
GitHub Actions (1)
Hackathon (1)
Japanese (1)

Goals, but challenging to measure

Turning users into (increasingly experienced) developers

- Developer mailing list & Slack #developers-forum
- Package review
- Conferences and Workshops

"Care" of the community

- Code of Conduct
- Outreach including to underrepresented areas
- Internships giving opportunities to learn

Events: 3 annual conferences

North American conference

BioC2024,
July 24-26,
Grand Rapids, USA

European conference

EuroBioC2024,
September 4-6,
Oxford, UK

Australasian conference

BioCAustralia2024
late 2024,
Sydney, Australia

BioC2023 by the numbers



Website Tour

Questions?

<https://bioconductor.org>

<https://seandavi.github.io>