Advanced R/Bioconductor: BioC resources and access to public databases

NGS analysis for gene regulation and epigenomics

Physalia 2021

Jacques Serizay

Bioconductor

- Free, open-source repository
- Open development software project
- Based primarily on the statistical R programming language
- Bioconductor’s releases are bi-annual
- **Provides tools for the analysis and comprehension of genomic data**

Bioconductor

SummarizedExperiment

- Free, open source repository
- Open development software project
- Based primarily on the statistical R programming language
- Bioconductor packages are bi-annual
- **Provides tools for the analysis and comprehension of genomic data**

GRanges

Biostrings

DESeq2

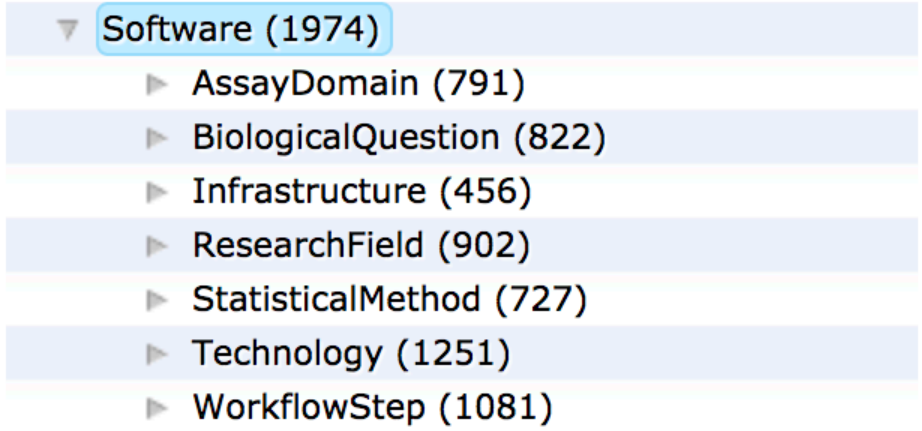
rtracklayer

Bioconductor

- Free, open-source repository
- Open development software project
- Based primarily on the statistical R programming language
- Bioconductor's releases are bi-annual
- **Provides tools for the analysis and comprehension of genomic data**

Bioconductor is not only for analysis packages

https://www.bioconductor.org/packages/release/BiocViews.html



Bioconductor is not only for analysis packages

https://www.bioconductor.org/packages/release/BiocViews.html

- Software (1974)
  - AssayDomain (791)
  - BiologicalQuestion (822)
  - Infrastructure (456)
  - ResearchField (902)
  - StatisticalMethod (727)
  - Technology (1251)
  - WorkflowStep (1081)
- AnnotationData (971)
  - ChipManufacturer (388)
  - ChipName (196)
  - CustomArray (2)
  - CustomDBSchema (6)
  - FunctionalAnnotation (31)
  - Organism (634)
  - PackageType (682)
  - SequenceAnnotation (1)
- Workflow (28)
  - AnnotationWorkflow (3)
  - BasicWorkflow (5)
  - EpigeneticsWorkflow (4)
  - GeneExpressionWorkflow (11)
  - GenomicVariantsWorkflow (2)
  - ImmunoOncologyWorkflow (14)
  - ProteomicsWorkflow (2)
  - ResourceQueryingWorkflow (2)
  - SingleCellWorkflow (2)
- ExperimentData (398)
  - AssayDomainData (81)
  - DiseaseModel (90)
  - OrganismData (139)
  - PackageTypeData (41)
  - RepositoryData (94)
  - ReproducibleResearch (22)
  - SpecimenSource (103)
  - TechnologyData (266)

NGS workflow management tools

- Software (1974)
  - AssayDomain (791)
  - BiologicalQuestion (822)
  - Infrastructure (456)
  - ResearchField (902)
  - StatisticalMethod (727)
  - Technology (1251)
  - WorkflowStep (1081)

systemPipeR

platforms all | rank 148 / 1974 | posts 0 | in Bioc 6 years

build ok | updated < 3 months | dependencies 154

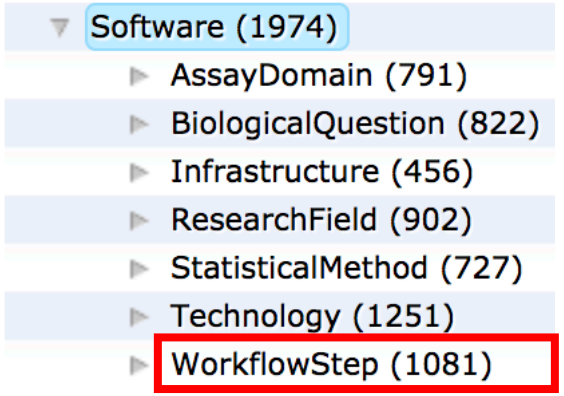
DOI: 10.18129/B9.bioc.systemPipeR

systemPipeR: NGS workflow and report generation environment

Bioconductor version: Release (3.12)

R package for building and running automated end-to-end analysis workflows for a wide range of next generation sequence (NGS) applications such as RNA-Seq, ChIP-Seq, VAR-Seq and Ribo-Seq. Important features include a uniform workflow interface across different NGS applications, automated report generation, and support for running both R and command-line software, such as NGS aligners or peak/variant callers, on local computers or compute clusters. Efficient handling of complex sample sets and experimental designs is facilitated by a consistently implemented sample annotation infrastructure. Instructions for using systemPipeR are given in the Overview Vignette (HTML). The remaining Vignettes, linked below, are workflow templates for common NGS use cases.

Reporting tools



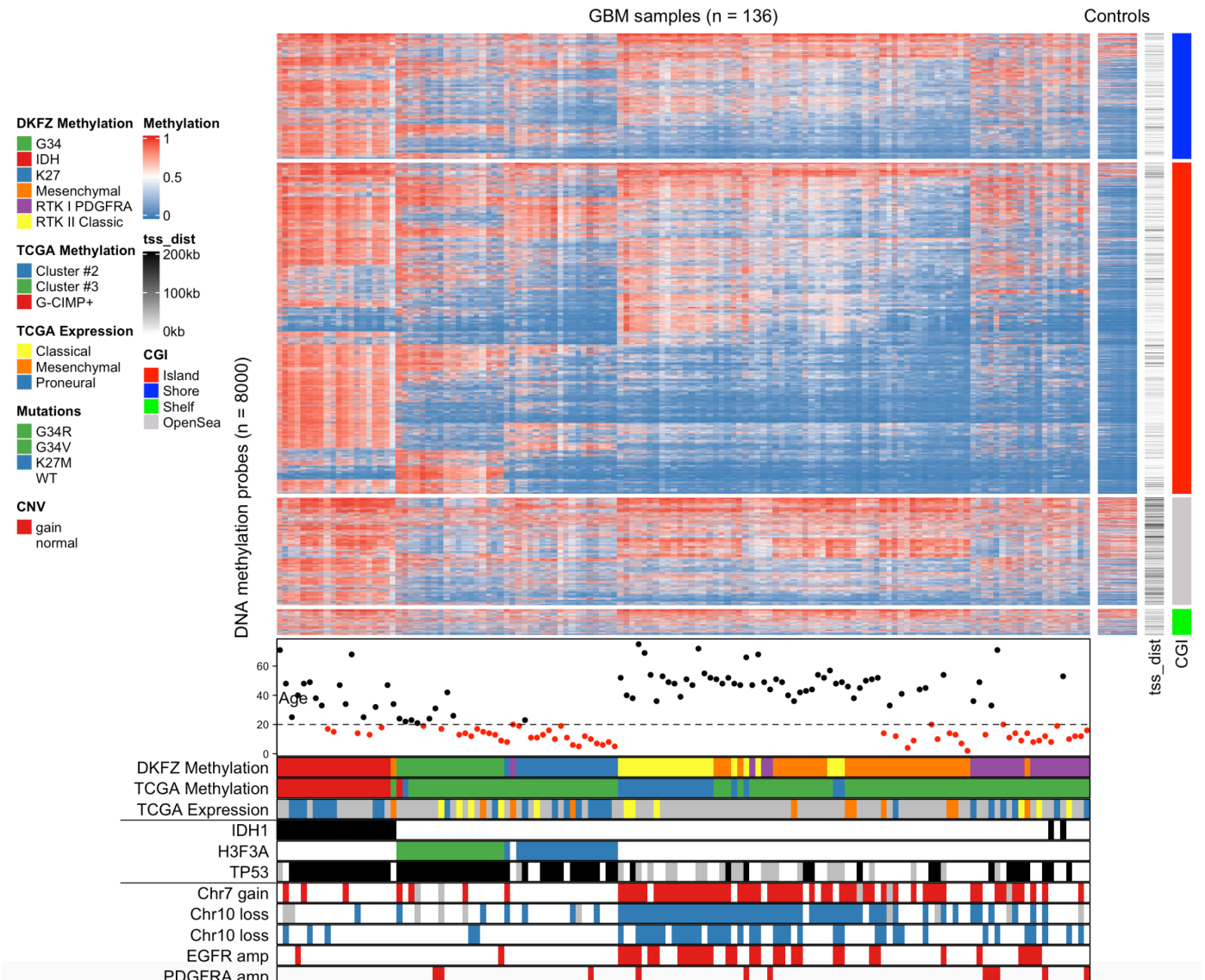

```
DESeq2Report::DESeq2Report(
        dds = dds,
        project = "OsmoResponse",
        intgroup = "timepoint"
)
```

Visualization tools

Complex heatmaps reveal patterns and correlations in multidimensional genomic data FREE

Zuguang Gu, Roland Eils, Matthias Schlesner ✉ Author Notes

Bioinformatics, Volume 32, Issue 18, 15 September 2016, Pages 2847–2849,
https://doi.org/10.1093/bioinformatics/btw313
Published: 20 May 2016 **Article history** ▾



Bioconductor is not only for analysis packages

https://www.bioconductor.org/packages/release/BiocViews.html

- Software (1974)
- AnnotationData (971)
- ExperimentData (398)
  - AssayDomainData (81)
  - DiseaseModel (90)
  - OrganismData (139)
  - PackageTypeData (41)
  - RepositoryData (94)
  - ReproducibleResearch (22)
  - SpecimenSource (103)
  - TechnologyData (266)
- Workflow (28)

Retrieving specific experiments

RangedSummarizedExperiment for time course RNA-Seq of fission yeast in response to stress, by Leong et al., Nat Commun 2014.

Bioconductor version: Release (3.12)

This package provides a RangedSummarizedExperiment object of read counts in genes for a time course RNA-Seq experiment of fission yeast (Schizosaccharomyces pombe) in response to oxidative stress (1M sorbitol treatment) at 0, 15, 30, 60, 120 and 180 mins. The samples are further divided between a wild-type group and a group with deletion of atf21. The read count matrix was prepared and provided by the author of the study: Leong HS, Dawson K, Wirth C, Li Y, Connolly Y, Smith DL, Wilkinson CR, Miller CJ. "A global non-coding RNA system modulates fission yeast protein levels in response to stress". Nat Commun 2014 May 23;5:3947. PMID: 24853205. GEO: GSE56761.

Author: Michael Love

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

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

Leong, S. H, Dawson, K., Wirth, C., Li, Y., Connolly, Y., Smith, L. D, Wilkinson, R. C, Miller, J. C (2014). "A global non-coding RNA system modulates fission yeast protein levels in response to stress." *Nat Commun*, **5**, 3947. http://www.ncbi.nlm.nih.gov/pubmed/24853205.

```
 > library(fission)
 > fission
class: RangedSummarizedExperiment
dim: 7039 36
metadata(1): ''
assays(1): counts
rownames(7039): SPAC212.11 SPAC212.09c ... SPMITTRNAGLU.01 SPMIT.11
rowData names(2): symbol biotype
colnames(36): GSM1368273 GSM1368274 ... GSM1368307 GSM1368308
colData names(4): strain minute replicate id
 > rowRanges(fission)
GRanges object with 7039 ranges and 2 metadata columns:
                 seqnames      ranges strand |          symbol        biotype
                    <Rle>   <IRanges>  <Rle> |     <character>       <factor>
      SPAC212.11        I      1-5662      - |            tlh1 protein_coding
     SPAC212.09c        I   7619-9274      + |     SPAC212.09c pseudogene
      SPNCRNA.70        I 11027-11556      - |      SPNCRNA.70 ncRNA
      SPAC212.12        I 15855-16226      + |      SPAC212.12 protein_coding
     SPAC212.04c        I 21381-23050      + |     SPAC212.04c protein_coding
             ...      ...         ...    ... .             ...             ...
  SPMITTRNATYR.01       MT 17257-17342      + | SPMITTRNATYR.01 tRNA
  SPMITTRNAILE.02       MT 17542-17613      + | SPMITTRNAILE.02 tRNA
       SPMIT.10         MT 17806-18030      + |            atp9 protein_coding
  SPMITTRNAGLU.01       MT 18404-18475      + | SPMITTRNAGLU.01 tRNA
       SPMIT.11         MT 18561-19307      + |            cox2 protein_coding
  -------
  seqinfo: 4 sequences from an unspecified genome; no seqlengths
```

Retrieving specific experiments

```
library(VariantAnnotation)
vcf <- readVcf(
         system.file("extdata", "SonVariantsChr21.vcf.gz", package = "AshkenazimSonChr21"),
         genome = "hg19"
)
info(vcf)
```

```
# A tibble: 94,527 x 35
   AC    AF       AN    DP    QD BLOCKAVG_min30p… BaseQRankSum DS     Dels   END    FS  HRun HaplotypeScore
   <I<l> <chr> <int> <int> <dbl> <lgl>                   <dbl> <lgl> <dbl> <int> <dbl> <int>          <dbl>
 1 <int… 0.50      2    38  8.25 FALSE                  -0.923 FALSE     0    NA  0         0          1.98
 2 <int… 0.50      2    37 19.5  FALSE                  -0.334 FALSE     0    NA  1.44      1          1.00
 3 <int… 0.50      2    49 23.0  FALSE                  -0.683 FALSE     0    NA 11.8       1          0.867
 4 <int… 0.50      2    62 20.0  FALSE                   1.40  FALSE     0    NA  1.00      0          0
 5 <int… 0.50      2    57 10.8  FALSE                  -1.44  FALSE     0    NA  0         0          0
 6 <int… 0.50      2    56 10.8  FALSE                  -1.46  FALSE     0    NA  0         1         12.0
 7 <int… 0.50      2    55  7.13 FALSE                  -0.141 FALSE     0    NA  0         0         14.0
 8 <int… 0.50      2    50 16.8  FALSE                   0.842 FALSE     0    NA  0         0          0
 9 <int… 0.50      2    73 18.0  FALSE                   0.456 FALSE     0    NA  9.32      2          0.789
10 <int… 0.50      2    86  8.44 FALSE                  -0.005 FALSE     0    NA  0         2          5.86
# … with 94,517 more rows, and 22 more variables: InbreedingCoeff <dbl>, MQ <dbl>, MQ0 <int>, MQRankSum <dbl>,
#   ReadPosRankSum <dbl>, SB <dbl>, VQSLOD <dbl>, culprit <chr>, set <chr>, CSQT <I<list>>, CSQR <I<list>>, AA <chr>,
#   GMAF <I<list>>, EVS <I<list>>, cosmic <I<list>>, clinvar <I<list>>, phastCons <lgl>, Variant.type <I<list>>,
#   Gene.name <I<list>>, Gene.component <I<list>>, phyloP <dbl>, SNP.Frequency <dbl>
```

Retrieving specific experiments

```
 > scRNAseq::listDatasets()
DataFrame with 46 rows and 5 columns
                Reference  Taxonomy                Part    Number                 Call
              <character> <integer>         <character> <integer>          <character>
1   @aztekin2019identifi..      8355                tail     13199      AztekinTailData()
2   @bach2017differentia..     10090        mammary gland     25806      BachMammaryData()
3     @baron2016singlecell      9606             pancreas      8569 BaronPancreasData('h..
4     @baron2016singlecell     10090             pancreas      1886 BaronPancreasData('m..
5   @buettner2015computa..     10090 embryonic stem cells       288      BuettnerESCData()
...                   ...       ...                  ...       ...                    ...
42      @wu2019advantages     10090               kidney     17542        WuKidneyData()
43            @xin2016rna      9606             pancreas      1600      XinPancreasData()
44       @zeisel2015brain     10090                brain      3005      ZeiselBrainData()
45  @zilionis2019singlec..      9606                 lung    173954     ZilionisLungData()
46  @zilionis2019singlec..     10090                 lung     17549 ZilionisLungData('mo..
```

```
 > ZeiselBrainData()
snapshotDate(): 2020-10-02
see ?scRNAseq and browseVignettes('scRNAseq') for documentation
loading from cache
see ?scRNAseq and browseVignettes('scRNAseq') for documentation
loading from cache
see ?scRNAseq and browseVignettes('scRNAseq') for documentation
loading from cache
snapshotDate(): 2020-10-02
see ?scRNAseq and browseVignettes('scRNAseq') for documentation
loading from cache
class: SingleCellExperiment
dim: 20006 3005
metadata(0):
assays(1): counts
rownames(20006): Tspan12 Tshz1 ... mt-Rnr1 mt-Nd4l
rowData names(1): featureType
colnames(3005): 1772071015_C02 1772071017_G12 ... 1772066098_A12 1772058148_F03
colData names(10): tissue group # ... level1class level2class
reducedDimNames(0):
altExpNames(2): ERCC repeat
```

Bioconductor is not only for analysis packages

https://www.bioconductor.org/packages/release/BiocViews.html

- Software (1974)
- AnnotationData (971)
- ExperimentData (398)
- Workflow (28)
  - AnnotationWorkflow (3)
  - BasicWorkflow (5)
  - EpigeneticsWorkflow (4)
  - GeneExpressionWorkflow (11)
  - GenomicVariantsWorkflow (2)
  - ImmunoOncologyWorkflow (14)
  - ProteomicsWorkflow (2)
  - ResourceQueryingWorkflow (2)
  - SingleCellWorkflow (2)

Standard pipeline boilerplates

rnaseqGene

platforms all | rank 1 / 28 | posts 1 / 1 / 2 / 0 | build ok

updated before release | dependencies 225

DOI: 10.18129/B9.bioc.rnaseqGene

RNA-seq workflow: gene-level exploratory analysis and differential expression

Bioconductor version: Release (3.12)

Here we walk through an end-to-end gene-level RNA-seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were aligned to the reference genome, and prepare a count matrix which tallies the number of RNA-seq reads/fragments within each gene for each sample. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

Author: Michael Love [aut, cre]

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

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

Love MI, Anders S, Kim V, Huber W (2015). "RNA-Seq workflow: gene-level exploratory analysis and differential expression." *F1000Research*. doi: 10.12688/f1000research.7035.

RNA-seq workflow: gene-level exploratory analysis and differential expression

***Michael I. Love*[1,2], *Simon Anders*[3], *Vladislav Kim*[4] *and Wolfgang Huber*[4]**

[1]Department of Biostatistics, UNC-Chapel Hill, Chapel Hill, NC, US
[2]Department of Genetics, UNC-Chapel Hill, Chapel Hill, NC, US
[3]Zentrum für Molekulare Biologie der Universität Heidelberg, Heidelberg, Germany
[4]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

16 October, 2019

Abstract

Here we walk through an end-to-end gene-level RNA-seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were quantified to the reference transcripts, and prepare gene-level count datasets for downstream analysis. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

Contents

Standard pipeline boilerplates

simpleSingleCell

platforms all | rank 9 / 28 | posts 0 | build ok
updated before release | dependencies 38

DOI: 10.18129/B9.bioc.simpleSingleCell

A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor

Bioconductor version: Release (3.12)

Once a proud workflow package, this is now a shell of its former self. Almost all of its content has been cannibalized for use in the "Orchestrating Single-Cell Analyses with Bioconductor" book at https://osca.bioconductor.org. Most vignettes here are retained as reminders of the glory that once was, also providing redirection for existing external links to the relevant OSCA book chapters.

Author: Aaron Lun [aut, cre], Davis McCarthy [aut], John Marioni [aut]

Maintainer: Aaron Lun <infinite.monkeys.with.keyboards at gmail.com>

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

Lun ATL, McCarthy DJ, Marioni JC (2016). "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor." *F1000Res.*, **5**, 2122. doi: 10.12688/f1000research.9501.2.

Standard pipeline boilerplates

proteomics

platforms all | rank 4 / 28 | posts 5 / 0.4 / 0.8 / 0 | build ok
updated before release | dependencies 230

DOI: 10.18129/B9.bioc.proteomics

Mass spectrometry and proteomics data analysis

Bioconductor version: Release (3.12)

This workflow illustrates R / Bioconductor infrastructure for proteomics. Topics covered focus on support for open community-driven formats for raw data and identification results, packages for peptide-spectrum matching, data processing and analysis.

Author: Laurent Gatto [aut, cre]

Maintainer: Laurent Gatto <laurent.gatto at uclouvain.be>

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

Laurent Gatto (2021). *proteomics: Mass spectrometry and proteomics data analysis*. R package version 1.14.0, https://www.bioconductor.org/help/workflows/proteomics/.

Mass spectrometry and proteomics data analysis

Laurent Gatto[1*]

[1]de Duve Institute, UCLouvain, Belgium

[*]laurent.gatto@uclouvain.be

Fri Jan 1 09:22:50 2021

Contents

Bioconductor Annotation packages

Packages found under AnnotationData:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries Search table: cerevi

Package	Maintainer	Title
BSgenome.Scerevisiae.UCSC.sacCer3	Bioconductor Package Maintainer	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer3)
BSgenome.Scerevisiae.UCSC.sacCer2	Bioconductor Package Maintainer	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer2)
TxDb.Scerevisiae.UCSC.sacCer3.sgdGene	Bioconductor Package Maintainer	Annotation package for TxDb object(s)
BSgenome.Scerevisiae.UCSC.sacCer1	Bioconductor Package Maintainer	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer1)
hom.Sc.inp.db	Bioconductor Package Maintainer	Homology information for Saccharomyces cerevisiae fror Inparanoid
MeSH.Sce.S288c.eg.db	Koki Tsuyuzaki	Mapping table for Saccharomyces cerevisiae S288c Gene ID to MeSH
TxDb.Scerevisiae.UCSC.sacCer2.sgdGene	Bioconductor Package Maintainer	Annotation package for TxDb object(s)

Bioconductor **AnnotationDbi** package

Gene centric AnnotationDbi packages include:

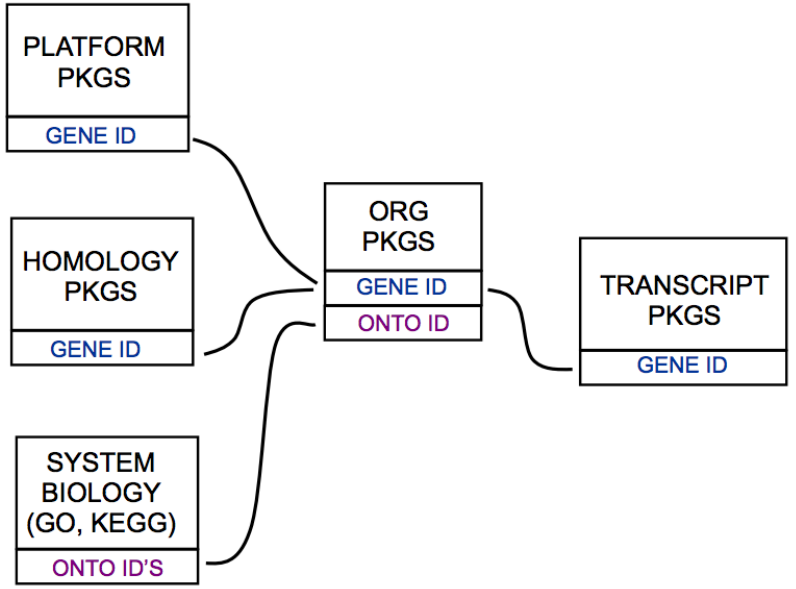
- Organism level: e.g. org.Mm.eg.db
- Platform level: e.g. hgu133plus2.db, hgu133plus2.probes, hgu133plus2.cdf
- Homology level: e.g. hom.Dm.inp.db
- System-biology level: GO.db

Genome centric GenomicFeatures packages include

- Transcriptome level: e.g. TxDb.Hsapiens.UCSC.hg19.knownGene, EnsDb.Hsapiens.v75
- Generic genome features: Can generate via GenomicFeatures

biomaRt:

- Query web-based 'biomart' resource for genes, sequence, SNPs, etc



Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> library(org.Sc.sgd.db)
> org.Sc.sgd.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: YEAST_DB
| ORGANISM: Saccharomyces cerevisiae
| SPECIES: Yeast
| YGSOURCENAME: Yeast Genome
| YGSOURCEURL: http://sgd-archive.yeastgenome.org
| YGSOURCEDATE: 2019-Oct25
| CENTRALID: ORF
| TAXID: 559292
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: http://current.geneontology.org/ontology/go-basic.obo
| GOSOURCEDATE: 2020-09-10
| EGSOURCEDATE: 2020-Sep23
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| ENSOURCEDATE: 2020-Aug18
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| UPSOURCENAME: Uniprot
| UPSOURCEURL: http://www.UniProt.org/
| UPSOURCEDATE: Mon Oct  5 00:25:28 2020
```

Bioconductor AnnotationDb packages

```
> keytypes(org.Sc.sgd.db)
 [1] "ALIAS"        "COMMON"       "DESCRIPTION"  "ENSEMBL"      "ENSEMBLPROT"
 [6] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"
[11] "GENENAME"     "GO"           "GOALL"        "INTERPRO"     "ONTOLOGY"
[16] "ONTOLOGYALL"  "ORF"          "PATH"         "PFAM"         "PMID"
[21] "REFSEQ"       "SGD"          "SMART"        "UNIPROT"
```

- Different sources →

  different nomenclatures

Bioconductor AnnotationDb packages

- Different sources → different nomenclatures

```
> keytypes(org.Sc.sgd.db)
 [1] "ALIAS"        "COMMON"       "DESCRIPTION"  "ENSEMBL"      "ENSEMBLPROT"
 [6] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"
[11] "GENENAME"     "GO"           "GOALL"        "INTERPRO"     "ONTOLOGY"
[16] "ONTOLOGYALL"  "ORF"          "PATH"         "PFAM"         "PMID"
[21] "REFSEQ"       "SGD"          "SMART"        "UNIPROT"
```

```
> AnnotationDbi::select(
+     org.Sc.sgd.db::org.Sc.sgd.db,
+     columns = c("ENSEMBL", "ENTREZID", "UNIPROT"),
+     keys = c('HOG1', 'MSN4'),
+     keytype = "GENENAME"
+ )
'select()' returned 1:1 mapping between keys and columns
  GENENAME ENSEMBL ENTREZID UNIPROT
1     HOG1 YLR113W   850803  P32485
2     MSN4 YKL062W   853803  P33749
```

Bioconductor AnnotationDb packages

- Different sources → different nomenclatures

```
> keytypes(org.Sc.sgd.db)
 [1] "ALIAS"        "COMMON"       "DESCRIPTION"  "ENSEMBL"      "ENSEMBLPROT"
 [6] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"
[11] "GENENAME"     "GO"           "GOALL"        "INTERPRO"     "ONTOLOGY"
[16] "ONTOLOGYALL"  "ORF"          "PATH"         "PFAM"         "PMID"
[21] "REFSEQ"       "SGD"          "SMART"        "UNIPROT"
```

```
> AnnotationDbi::select(
+     org.Sc.sgd.db::org.Sc.sgd.db,
+     columns = c("ENSEMBL", "ENTREZID", "UNIPROT"),
+     keys = c('HOG1', 'MSN4'),
+     keytype = "GENENAME"
+ )
'select()' returned 1:1 mapping between keys and columns
  GENENAME ENSEMBL ENTREZID UNIPROT
1     HOG1 YLR113W   850803  P32485
2     MSN4 YKL062W   853803  P33749
```

```
> AnnotationDbi::select(
+     org.Sc.sgd.db::org.Sc.sgd.db,
+     columns = c("ENSEMBL", "COMMON", "ENTREZID", "UNIPROT"),
+     keys = c('HOG1', 'MSN4'),
+     keytype = "GENENAME"
+ )
'select()' returned 1:many mapping between keys and columns
  GENENAME ENSEMBL                                      COMMON        SGD ENTREZID UNIPROT
1     HOG1 YLR113W                                        HOG1 S000004103   850803  P32485
2     HOG1 YLR113W           mitogen-activated protein kinase HOG1 S000004103   850803  P32485
3     HOG1 YLR113W                                        SSK3 S000004103   850803  P32485
4     MSN4 YKL062W                                        MSN4 S000001545   853803  P33749
5     MSN4 YKL062W stress-responsive transcriptional activator MSN4 S000001545   853803  P33749
```

Bioconductor AnnotationDb packages

- Different sources → different nomenclatures

```
convertIDs <- function(orgDb, ID, fromType, toType) {
        df <- AnnotationDbi::select(
                orgDb,
                columns = toType,
                keys = ID,
                keytype = fromType
        )
}
```

```
> convertIDs(ID = ids, orgDb = db, fromType = "ORF", toType = "REFSEQ")
'select()' returned 1:many mapping between keys and columns
     ORF        SGD       REFSEQ
1   HRA1 S000119380    NR_132149
2 YLR113W S000004103 NM_001182000
3 YLR113W S000004103    NP_013214
4  RUF22 S000130131    NR_132174
```

TxDb packages

- A TxDb package connects a set of genomic coordinates to various transcript oriented features.
- In other words, TxDb packages provide **gene annotation models**

TxDb packages

```
> TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: sacCer3
# Organism: Saccharomyces cerevisiae
# Taxonomy ID: 4932
# UCSC Table: sgdGene
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Name of canonical transcript in cluster
# Full dataset: yes
# miRBase build ID: NA
# transcript_nrow: 6692
# exon_nrow: 7034
# cds_nrow: 7034
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-10-07 18:20:42 +0000 (Wed, 07 Oct 2015)
# GenomicFeatures version at creation time: 1.21.30
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
```

TxDb packages

```
> TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: sacCer3
# Organism: Saccharomyces cerevisiae
# Taxonomy ID: 4932
# UCSC Table: sgdGene
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Name of canonical transcript in cluster
# Full dataset: yes
# miRBase build ID: NA
# transcript_nrow: 6692
# exon_nrow: 7034
# cds_nrow: 7034
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-10-07 18:20:42 +0000 (Wed, 07 Oct 2015)
# GenomicFeatures version at creation time: 1.21.30
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
```

TxDb packages

```
> TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: sacCer3
# Organism: Saccharomyces cerevisiae
# Taxonomy ID: 4932
# UCSC Table: sgdGene
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Name of canonical transcript in cluster
# Full dataset: yes
# miRBase build ID: NA
# transcript_nrow: 6692
# exon_nrow: 7034
# cds_nrow: 7034
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-10-07 18:20:42 +0000 (Wed, 07 Oct 2015)
# GenomicFeatures version at creation time: 1.21.30
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
```

TxDb packages

```
> TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: sacCer3
# Organism: Saccharomyces cerevisiae
# Taxonomy ID: 4932
# UCSC Table: sgdGene
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Name of canonical transcript in cluster
# Full dataset: yes
# miRBase build ID: NA
# transcript_nrow: 6692
# exon_nrow: 7034
# cds_nrow: 7034
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-10-07 18:20:42 +0000 (Wed, 07 Oct 2015)
# GenomicFeatures version at creation time: 1.21.30
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
```

TxDb packages

- TxDb databases can be explored with AnnotationDbi functions

```
> AnnotationDbi::keys(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene, keytype = "GENEID") %>% glimpse
 chr [1:6534] "Q0010" "Q0032" "Q0055" "Q0075" "Q0080" "Q0085" "Q0092" "Q0120" "Q0130" "Q0140" "Q0142" "Q0143"
> AnnotationDbi::keys(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene, keytype = "TXNAME") %>% glimpse
 chr [1:6692] "YAL069W" "YAL068W-A" "YAL067W-A" "YAL066W" "YAL064W-B" "YAL064W" "YAL062W" "YAL061W" "YAL060W"
```

TxDb packages

- TxDb databases can be explored with AnnotationDbi functions

```
> dat <- AnnotationDbi::select(
+     tx,
+     columns = c("GENEID", "TXNAME"),
+     keys = keys(tx),
+     keytype = "GENEID"
+ ) %>%
+     group_by(GENEID) %>%
+     summarize(TXNAME = list(TXNAME))
'select()' returned 1:many mapping between keys and columns
`summarise()` ungrouping output (override with `.groups` argument)
> dat
# A tibble: 6,534 x 2
   GENEID TXNAME
   <chr>  <list>
 1 Q0010  <chr [2]>
 2 Q0032  <chr [1]>
 3 Q0055  <chr [6]>
 4 Q0075  <chr [1]>
 5 Q0080  <chr [1]>
 6 Q0085  <chr [1]>
 7 Q0092  <chr [1]>
 8 Q0120  <chr [4]>
 9 Q0130  <chr [1]>
10 Q0140  <chr [1]>
# … with 6,524 more rows
```

TxDb packages

```
> genes <- GenomicFeatures::genes(tx)
> genes
GRanges object with 6534 ranges and 1 metadata column:
           seqnames        ranges strand |     gene_id
              <Rle>     <IRanges>  <Rle> | <character>
     Q0010     chrM     3952-4415      + |       Q0010
     Q0032     chrM   11667-11957      + |       Q0032
     Q0055     chrM   13818-26701      + |       Q0055
     Q0075     chrM   24156-25255      + |       Q0075
     Q0080     chrM   27666-27812      + |       Q0080
       ...      ...           ...    ... .         ...
   YPR200C   chrXVI 939279-939671      - |     YPR200C
   YPR201W   chrXVI 939922-941136      + |     YPR201W
   YPR202W   chrXVI 943032-944188      + |     YPR202W
 YPR204C-A   chrXVI 946856-947338      - |   YPR204C-A
   YPR204W   chrXVI 944603-947701      + |     YPR204W
  -------
  seqinfo: 17 sequences (1 circular) from sacCer3 genome
```

- **GenomicFeatures** package is used to **interact with** TxDb databases

transcripts(x, columns=c("tx_id", "tx_name"), filter=NULL, use.names=FALSE)

exons(x, columns="exon_id", filter=NULL, use.names=FALSE)

cds(x, columns="cds_id", filter=NULL, use.names=FALSE)

genes(x, columns="gene_id", filter=NULL, single.strand.genes.only=TRUE)

promoters(x, upstream=2000, downstream=200, use.names=TRUE, ...)

TxDb packages

- **TxDb** and **org** packages can be intersected
- Tidyverse makes that easy

```
> genes <- GenomicFeatures::genes(tx)
> genes
GRanges object with 6534 ranges and 1 metadata column:
           seqnames        ranges strand |     gene_id
              <Rle>     <IRanges>  <Rle> | <character>
     Q0010     chrM     3952-4415      + |       Q0010
     Q0032     chrM   11667-11957      + |       Q0032
     Q0055     chrM   13818-26701      + |       Q0055
     Q0075     chrM   24156-25255      + |       Q0075
     Q0080     chrM   27666-27812      + |       Q0080
       ...      ...           ...    ... .         ...
   YPR200C   chrXVI 939279-939671      - |     YPR200C
   YPR201W   chrXVI 939922-941136      + |     YPR201W
   YPR202W   chrXVI 943032-944188      + |     YPR202W
 YPR204C-A   chrXVI 946856-947338      - |   YPR204C-A
   YPR204W   chrXVI 944603-947701      + |     YPR204W
  -------
  seqinfo: 17 sequences (1 circular) from sacCer3 genome
> mcols(genes) <- AnnotationDbi::select(
+     tx,
+     columns = c("GENEID", "TXNAME"),
+     keys = keys(tx),
+     keytype = "GENEID"
+ ) %>%
+     group_by(GENEID) %>%
+     summarize(TXNAME = list(TXNAME)) %>%
+     left_join(
+         .,
+         AnnotationDbi::select(org, columns = "GENENAME", keys = .$GENEID, keytype = "ORF"),
+         by = c("GENEID" = "ORF")
+     )
'select()' returned 1:many mapping between keys and columns
`summarise()` ungrouping output (override with `.groups` argument)
'select()' returned 1:1 mapping between keys and columns
> genes
GRanges object with 6534 ranges and 4 metadata columns:
           seqnames        ranges strand |      GENEID               TXNAME         SGD    GENENAME
              <Rle>     <IRanges>  <Rle> | <character>              <list> <character> <character>
     Q0010     chrM     3952-4415      + |       Q0010          Q0010,Q0017  S000007257        <NA>
     Q0032     chrM   11667-11957      + |       Q0032                Q0032  S000007259        <NA>
     Q0055     chrM   13818-26701      + |       Q0055 Q0050,Q0055,Q0060,...  S000007262         AI2
     Q0075     chrM   24156-25255      + |       Q0075                Q0075  S000007266    AI5_BETA
     Q0080     chrM   27666-27812      + |       Q0080                Q0080  S000007267        ATP8
       ...      ...           ...    ... .         ...                 ...         ...         ...
   YPR200C   chrXVI 939279-939671      - |     YPR200C              YPR200C  S000006404        ARR2
   YPR201W   chrXVI 939922-941136      + |     YPR201W              YPR201W  S000006405        ARR3
   YPR202W   chrXVI 943032-944188      + |     YPR202W       YPR202W,YPR203W  S000006406        <NA>
 YPR204C-A   chrXVI 946856-947338      - |   YPR204C-A            YPR204C-A  S000028727        <NA>
   YPR204W   chrXVI 944603-947701      + |     YPR204W              YPR204W  S000006408        <NA>
  -------
  seqinfo: 17 sequences (1 circular) from sacCer3 genome
```

BSgenome packages

- Data packages that contain the full genome sequences of a given organism

```
> genome <- BSgenome.Scerevisiae.UCSC.sacCer3::BSgenome.Scerevisiae.UCSC.sacCer3
> genome
Yeast genome:
# organism: Saccharomyces cerevisiae (Yeast)
# genome: sacCer3
# provider: UCSC
# release date: April 2011
# 17 sequences:
#   chrI    chrII   chrIII  chrIV   chrV    chrVI   chrVII  chrVIII chrIX   chrX    chrXI   chrXII  chrXIII chrXIV  chrXV   chrXVI  chrM
# (use 'seqnames()' to see all the sequence names, use the '$' or '[[' operator to access a given sequence)
```

BSgenome packages

- **Biostrings** package is used to **interact with** BSgenome databases

```
> genome <- BSgenome.Scerevisiae.UCSC.sacCer3::BSgenome.Scerevisiae.UCSC.sacCer3
> genome
Yeast genome:
# organism: Saccharomyces cerevisiae (Yeast)
# genome: sacCer3
# provider: UCSC
# release date: April 2011
# 17 sequences:
#   chrI    chrII   chrIII  chrIV   chrV    chrVI   chrVII  chrVIII chrIX   chrX    chrXI   chrXII  chrXIII chrXIV  chrXV   chrXVI  chrM
# (use 'seqnames()' to see all the sequence names, use the '$' or '[[' operator to access a given sequence)
```

BSgenome packages

```
> genome <- BSgenome.Scerevisiae.UCSC.sacCer3::BSgenome.Scerevisiae.UCSC.sacCer3
> genome
Yeast genome:
# organism: Saccharomyces cerevisiae (Yeast)
# genome: sacCer3
# provider: UCSC
# release date: April 2011
# 17 sequences:
#   chrI    chrII   chrIII  chrIV   chrV    chrVI   chrVII  chrVIII chrIX   chrX    chrXI   chrXII  chrXIII chrXIV  chrXV   chrXVI  chrM
# (use 'seqnames()' to see all the sequence names, use the '$' or '[[' operator to access a given sequence)
> Biostrings::getSeq(genome)
DNAStringSet object of length 17:
       width seq                                                                                                                  names
 [1]  230218 CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACCCA...GTGTGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGGG chrI
 [2]  813184 AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATG...AGGGTGTGTGGGTGTGGGTGTGGTGTGTGTGGGTGTGGTGTGTGGGTGTGT chrII
 [3]  316620 CCCACACACCACACCCACACCACACCCACACACCACACACACCACACCCACA...GGTGTGGTGTGTGGGTGTGGGTGTGTGGGTGTGGTGGGTGTGGTGTGTGTG chrIII
 [4] 1531933 ACACCACACCCACACCACACCCACACACACCACACCCACACACCACACCCAC...AGACAATCTATAAAAAGTAAACATAAAATAAAGGTAGTAAGTAGCTTTTGG chrIV
 [5]  576874 CGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATGTTCAACCAAAAGCTAC...TAGAAGTGCTGTAGGGCTAAAGAACAGGGTTTCATTTTCATTTTTTTTTTT chrV
  ...     ... ...
[13]  924431 CCACACACACCACACCCACACCACACCCACACCCACACCCACACCCACAC...TGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGTGGGG chrXIII
[14]  784333 CCGGCTTTCTGACCGAAATTAAAAAAAAAAAATGAAAATGAAACCCTGTTCT...GGGTGTGGTGTGTGGGTGTGGTGTGTGGGTGTGTGTGGGTGTGGTGTGGGT chrXIV
[15] 1091291 ACACCACACCCACACCACACCCACACCCACACACCACACCCACACACACACA...ATGACAGCGAGAGTAGAGGTAGATGTGAGAGAGTGTGTGGGTGTGGTGTGT chrXV
[16]  948066 AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATG...CTAAAGAACAGGGTTTCATTTTCATTTTTTTTTTTTAATTTCGGTCAGAAA chrXVI
[17]   85779 TTCATAATTAATTTTTTATATATATATTATATTATAATATTAATTTATATTA...ATAGTTATATTATTATACAGAAATATGCTTAATTATAATATAATATCCATA chrM
```

BSgenome packages

- You can extract sequences for a given Granges with …[…]

```
> seqs <- Biostrings::getSeq(BSgenome.Scerevisiae.UCSC.sacCer3::BSgenome.Scerevisiae.UCSC.sacCer3)
> seqs[genes]
DNAStringSet object of length 6534:
       width seq                                                                                 names
   [1]   464 ATATATTATATTATATTTTTATATAATATATTATTAATTATTATTTTAATTT...TATAAATCTAACTTAATTAATAATTTAAATAATATACTTTATATTTTATAA YAL001C
   [2]   291 ATATTAATAATATATATATTATTTATTATAATGAAAACCTATCCTATATTAT...ATAAATATTCATTTCTTACTAATTAATAAAAAGTTTTTATATTCATTATAA YAL002W
   [3] 12884 ATGGTACAAAGATGATTATATTCAACAAATGCAAAAGATATTGCAGTATTAT...ACTTCTCCACCAGCTGTACACTCATTTAATACACCAGCTGTACAATCTTAA YAL003W
   [4]  1100 ATATTAATATTATTAATAATAATTTTAACTAATAATAAAGTTTTTATAGAAA...AATAGAGTTAAAATTATTATTAATAAACTTCATTATAACAATATCGAATAA YAL004W
   [5]   147 ATGCCACAATTAGTTCCATTTTATTTTATGAATCAATTAACATATGGTTTCT...CCTATGATCTTAAGATTATATGTATCTAGATTATTTATTTCTAAATTATAA YAL005C
   ...   ... ...
[6530]   393 ATGGTAAGTTTCATAACGTCTAGGCAACTCAAGGGCCTAATTGAAAATCAGA...TGGGAGACCCATTGTAGAGAGAGTAACCTTAAATTGATTGTTAGTGGTTGA Q0160
[6531]  1215 ATGTCAGAAGATCAAAAAAGTGAAAATTCGGTACCTTCTAAGGTTAATATGG...ATAGTCGCGAGGATCCTTAAACCATATTATATATGGAACAATAGAAATTAA Q0182
[6532]  1157 ATGGAAATTGAAAACGAACGTACGTACCGACTACTTTATTTTTGCAGGCCGG...AACAAAAAGGACCGATTGGGTGATGTTTTGCATGTGTGCTGCCCAAGTTGA Q0255
[6533]   483 ATGATGCCTGCTAAACTGCAGCTTGACGTACTGCGGACCCTGCAGTCCAGCG...CAAGTATTCTTTGTATTCCTCGTCATTTCGCAGCATTCTCTCCACAGCTAG Q0275
[6534]  3099 ATGGCAGACACACCCTCTGTGGCAGTACAGGCCCCACCGGGCTATGGTAAGA...AGATTTGATATATTATTATGTAGAGATTCGAGCAGAGAAGTTGGAGAGTGA Q0297
```

BSgenome packages

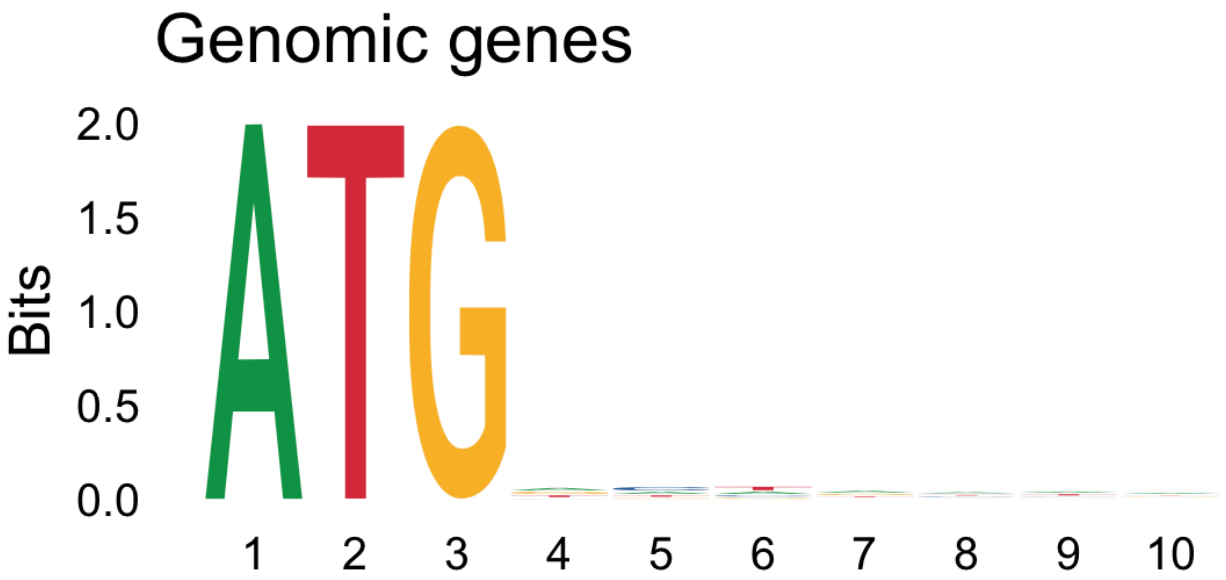
- You can extract sequences for a given Granges with …[…]

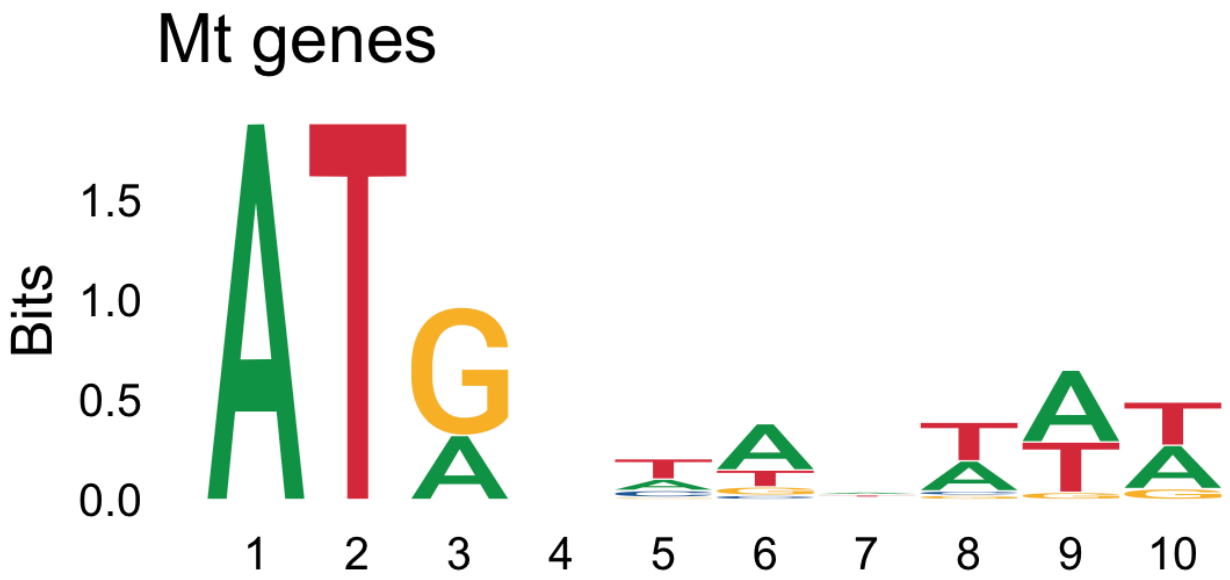
```
> genes$seq <- seqs[genes]
> genes
GRanges object with 6534 ranges and 5 metadata columns:
            seqnames        ranges strand |      GENEID                TXNAME         SGD    GENENAME                     seq
               <Rle>     <IRanges>  <Rle> | <character>                <list> <character> <character>          <DNAStringSet>
      Q0010     chrM     3952-4415      + |       Q0010          Q0010,Q0017  S000007257        <NA> ATATATTATA...TATTTTATAA
      Q0032     chrM   11667-11957      + |       Q0032                Q0032  S000007259        <NA> ATATTAATAA...TTCATTATAA
      Q0055     chrM   13818-26701      + |       Q0055 Q0050,Q0055,Q0060,...  S000007262         AI2 ATGGTACAAA...ACAATCTTAA
      Q0075     chrM   24156-25255      + |       Q0075                Q0075  S000007266    AI5_BETA ATATTAATAT...TATCGAATAA
      Q0080     chrM   27666-27812      + |       Q0080                Q0080  S000007267        ATP8 ATGCCACAAT...TAAATTATAA
        ...      ...            ...    ... .        ...                 ...         ...         ...                      ...
    YPR200C   chrXVI 939279-939671      - |     YPR200C              YPR200C  S000006404        ARR2 ATGGTAAGTT...TAGTGGTTGA
    YPR201W   chrXVI 939922-941136      + |     YPR201W              YPR201W  S000006405        ARR3 ATGTCAGAAG...TAGAAATTAA
    YPR202W   chrXVI 943032-944188      + |     YPR202W      YPR202W,YPR203W  S000006406        <NA> ATGGAAATTG...CCCAAGTTGA
  YPR204C-A   chrXVI 946856-947338      - |   YPR204C-A            YPR204C-A  S000028727        <NA> ATGATGCCTG...CCACAGCTAG
    YPR204W   chrXVI 944603-947701      + |     YPR204W              YPR204W  S000006408        <NA> ATGGCAGACA...TGGAGAGTGA
  -------
  seqinfo: 17 sequences (1 circular) from sacCer3 genome
```

BSgenome packages

- CRAN et BioC packages can be used together to investigate biological questions such as start codon usage over mitochondrial/genomic genes in Yeast

```
cowplot::plot_grid(
    genes$seq[seqnames(genes) != 'chrM'] %>% Biostrings::subseq(1, 10) %>% as.character() %>% ggseqlogo::ggseqlogo(.) + ggtitle('Genomic genes'),
    genes$seq[seqnames(genes) == 'chrM'] %>% Biostrings::subseq(1, 10) %>% as.character() %>% ggseqlogo::ggseqlogo(.) + ggtitle('Mt genes')
)
```





AnnotationHub: retrieving release-specific files

- The AnnotationHub package provides a client interface to resources stored at the AnnotationHub web service

- It is different from **AnnotationDbi-supported packages** (e.g. orgDb or TxDb packages), since it allows access to **files on top of databases**

AnnotationHub: retrieving release-specific files

```
> ah <- AnnotationHub::AnnotationHub()
snapshotDate(): 2020-10-27
> ah
AnnotationHub with 54989 records
# snapshotDate(): 2020-10-27
# $dataprovider: Ensembl, BroadInstitute, UCSC, ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, Haemcode, FungiDB, Inparanoid8, TriTrypDB, Plasmo...
# $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus, Pan troglodytes, Rattus norvegicus, Danio rerio, Gallus gal...
# $rdataclass: GRanges, TwoBitFile, BigWigFile, EnsDb, Rle, OrgDb, ChainFile, TxDb, Inparanoid8Db, data.frame
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH5012"]]'

            title
  AH5012  | Chromosome Band
  AH5013  | STS Markers
  AH5014  | FISH Clones
  AH5015  | Recomb Rate
  AH5016  | ENCODE Pilot
  ...       ...
  AH89321 | Ensembl 102 EnsDb for Xiphophorus couchianus
  AH89322 | Ensembl 102 EnsDb for Xiphophorus maculatus
  AH89323 | Ensembl 102 EnsDb for Xenopus tropicalis
  AH89324 | Ensembl 102 EnsDb for Zonotrichia albicollis
  AH89325 | Ensembl 102 EnsDb for Zalophus californianus
```

Annotation**Hub**: retrieving release-specific **files**

- Queries are done using query(ah, “keyword”)

AnnotationHub: retrieving release-specific files

- Queries are done using query(ah, “keyword”)

```
> query(ah, c('sacCer3', 'TwoBitFile'))
AnnotationHub with 1 record
# snapshotDate(): 2020-10-27
# names(): AH14104
# $dataprovider: UCSC
# $species: Saccharomyces cerevisiae
# $rdataclass: TwoBitFile
# $rdatadateadded: 2014-12-15
# $title: sacCer3.2bit
# $description: UCSC 2 bit file for sacCer3
# $taxonomyid: 4932
# $genome: sacCer3
# $sourcetype: TwoBit
# $sourceurl: http://hgdownload.cse.ucsc.edu/goldenpath/sacCer3/bigZips/sacCer3.2bit
# $sourcesize: NA
# $tags: c("2bit", "UCSC", "genome")
# retrieve record with 'object[["AH14104"]]'
```

AnnotationHub: retrieving release-specific files

- Queries are done using query(ah, “keyword”)
- Objects are retrieved using ah[[“…”]]

```
> query(ah, c('sacCer3', 'TwoBitFile'))
AnnotationHub with 1 record
# snapshotDate(): 2020-10-27
# names(): AH14104
# $dataprovider: UCSC
# $species: Saccharomyces cerevisiae
# $rdataclass: TwoBitFile
# $rdatadateadded: 2014-12-15
# $title: sacCer3.2bit
# $description: UCSC 2 bit file for sacCer3
# $taxonomyid: 4932
# $genome: sacCer3
# $sourcetype: TwoBit
# $sourceurl: http://hgdownload.cse.ucsc.edu/goldenpath/sacCer3/bigZips/sacCer3.2bit
# $sourcesize: NA
# $tags: c("2bit", "UCSC", "genome")
# retrieve record with 'object[["AH14104"]]'
> twobit <- ah[['AH14104']]
loading from cache
> twobit
TwoBitFile object
resource: /Users/jacquesserizay/Library/Caches/AnnotationHub/a6c5475f3d0f_18199
> seqs <- getSeq(twobit)
> seqs
DNAStringSet object of length 17:
       width seq                                                                                names
 [1]  230218 CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACCCA...GTGTGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGGG chrI
 [2]  813184 AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATG...AGGGTGTGTGGGTGTGGGTGTGGTGTGTGTGGGTGTGGTGTGTGGGTGTGT chrII
 [3]  316620 CCCACACACCACACCCACACCACACCCACACACCACACACACCACACCCACA...GGTGTGGTGTGTGGGTGTGGGTGTGTGGGTGTGGTGGGTGTGGTGTGTGTG chrIII
 [4] 1531933 ACACCACACCCACACCACACCCACACACCACACCCACACACCACACCCAC...AGACAATCTATAAAAAGTAAACATAAAATAAAGGTAGTAAGTAGCTTTTGG chrIV
 [5]  439888 CACACACACCACACCCACACCACACCACACCACACCCACACCCACACACC...GGATGTGGTGTGGATGTGGTGTGGGTGTGGATGTGGGTGTGGTGTGTGTGT chrIX
 ...     ... ...
[13]  924431 CCACACACACACCACACCCACACCACACCCACACCCACACCCACACCAC...TGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGTGGGG chrXIII
[14]  784333 CCGGCTTTCTGACCGAAATTAAAAAAAAAAAAATGAAAATGAAACCCTGTTCT...GGGTGTGGTGTGTGGGTGTGGTGTGTGGGTGTGTGTGGGTGTGGTGTGGGT chrXIV
[15] 1091291 ACACCACACCCACACCACACCCACACCCACACCACACCCACACACACACA...ATGACAGCGAGAGTAGAGGTAGATGTGAGAGAGTGTGTGGGTGTGGTGTGT chrXV
[16]  948066 AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATG...CTAAAGAACAGGGTTTCATTTTCATTTTTTTTTTTTAATTTCGGTCAGAAA chrXVI
[17]   85779 TTCATAATTAATTTTTTATATATATTATATTATAATATTAATTTATATTA...ATAGTTATATTATTATACAGAAATATGCTTAATTATAATATAATATCCATA chrM
```

Annotation**Hub**: retrieving release-specific **files**

- Many **many** resources available on AnnotationHub

AnnotationHub: retrieving release-specific files

```
> query(ah, 'VcfFile'
+
> query(ah, 'VcfFile')
AnnotationHub with 8 records
# snapshotDate(): 2020-10-27
# $dataprovider: dbSNP
# $species: Homo sapiens
# $rdataclass: VcfFile
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH57956"]]'

            title
  AH57956 | clinvar_20160203.vcf.gz
  AH57957 | clinvar_20160203_papu.vcf.gz
  AH57958 | common_and_clinical_20160203.vcf.gz
  AH57959 | common_no_known_medical_impact_20160203.vcf.gz
  AH57960 | clinvar_20160203.vcf.gz
  AH57961 | clinvar_20160203_papu.vcf.gz
  AH57962 | common_and_clinical_20160203.vcf.gz
  AH57963 | common_no_known_medical_impact_20160203.vcf.gz
```

AnnotationHub: retrieving release-specific files

```
> query(ah, c('bigwig', 'UCSC') )
AnnotationHub with 2198 records
# snapshotDate(): 2020-10-27
# $dataprovider: UCSC
# $species: Homo sapiens, Drosophila melanogaster, Mus musculus
# $rdataclass: Rle, GRanges
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH23256"]]'

            title
  AH23256 | wgEncodeBroadHistoneGm12878H3k4me3StdPk.broadPeak.gz
  AH23257 | wgEncodeBroadHistoneGm12878H3k9acStdPk.broadPeak.gz
  AH23262 | wgEncodeBroadHistoneGm12878H3k36me3StdPk.broadPeak.gz
  AH23367 | wgEncodeBroadHistoneHuvecH3k27me3StdPk.broadPeak.gz
  AH24345 | wgEncodeCshlLongRnaSeqNhemfm2CellTotalGeneGencV10.gtf.gz
  ...       ...
  AH78698 | phastCons30way.UCSC.hg38.chrX.rds
  AH78699 | phastCons30way.UCSC.hg38.chrX_KI270880v1_alt.rds
  AH78700 | phastCons30way.UCSC.hg38.chrX_KI270881v1_alt.rds
  AH78701 | phastCons30way.UCSC.hg38.chrX_KI270913v1_alt.rds
  AH78702 | phastCons30way.UCSC.hg38.chrY.rds
```

AnnotationHub: retrieving release-specific files

```
> query(ah, c('TxDb', 'GENCODE'))
AnnotationHub with 20 records
# snapshotDate(): 2020-10-27
# $dataprovider: GENCODE
# $species: Homo sapiens
# $rdataclass: TxDb
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH75134"]]'

            title
  AH75134 | TxDb for Gencode v23 on hg19 coordinates
  AH75137 | TxDb for Gencode v23 on hg38 coordinates
  AH75140 | TxDb for Gencode v24 on hg19 coordinates
  AH75143 | TxDb for Gencode v24 on hg38 coordinates
  AH75146 | TxDb for Gencode v25 on hg19 coordinates
  ...       ...
  AH75179 | TxDb for Gencode v30 on hg38 coordinates
  AH75182 | TxDb for Gencode v31 on hg19 coordinates
  AH75185 | TxDb for Gencode v31 on hg38 coordinates
  AH75188 | TxDb for Gencode v32 on hg19 coordinates
  AH75191 | TxDb for Gencode v32 on hg38 coordinates
```