

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

NGS analysis for gene regulation and epigenomics
Physalia 2021

SummarizedExperiment

- Free, open source repository

GRanges

- Open development software project

Biostrings

- Based primarily on the statistical R programming language

- Bioconducting packages are bi-annual

DESeq2

rtracklayer

- Provides tools for the analysis and comprehension of genomic data

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>

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

Bioconductor is not only for analysis packages

- [Software \(1974\)](#)

- [AssayDomain \(791\)](#)
- [BiologicalQuestion \(822\)](#)
- [Infrastructure \(456\)](#)
- [ResearchField \(902\)](#)
- [StatisticalMethod \(727\)](#)
- [Technology \(1251\)](#)
- [WorkflowStep \(1081\)](#)

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

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

▼ Software (1974)

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

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

Visualization tools

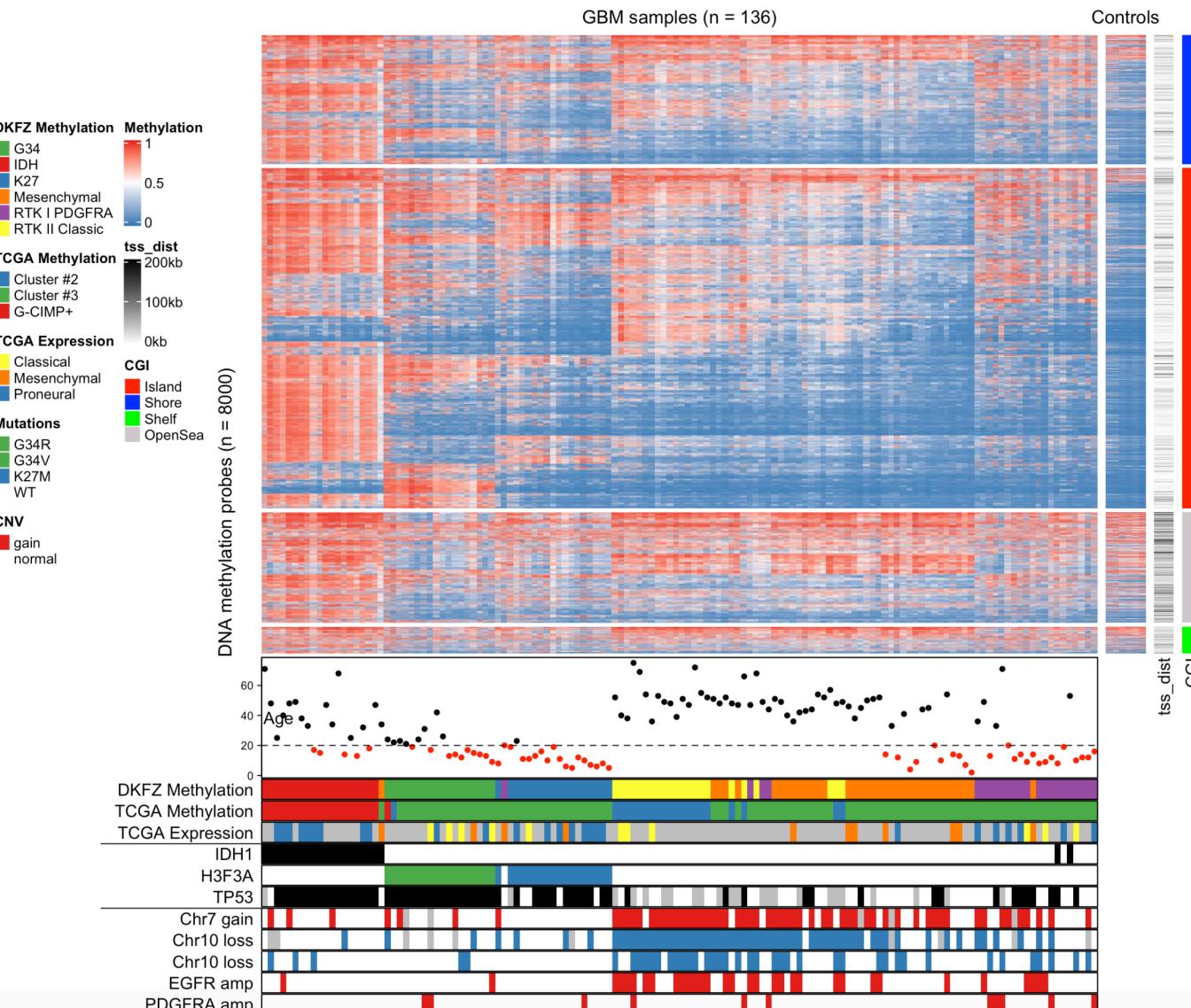
Complex heatmaps reveal patterns and correlations
in multidimensional genomic data 

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
# ... 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>
  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
```

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
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():
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 Annotation packages

	Package	Maintainer	Title
BS	BSgenome.Scerevisiae.UCSC.sacCer3	Bioconductor Package Maintainer	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer3)
BS	BSgenome.Scerevisiae.UCSC.sacCer2	Bioconductor Package Maintainer	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer2)
TxDb	TxDb.Scerevisiae.UCSC.sacCer3.sgdGene	Bioconductor Package Maintainer	Annotation package for TxDb object(s)
BS	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 from Inparanoid
	MeSH.Sce.S288c.eg.db	Koki Tsuyuzaki	Mapping table for Saccharomyces cerevisiae S288c Gene ID to MeSH
TxDb	TxDb.Scerevisiae.UCSC.sacCer2.sgdGene	Bioconductor Package Maintainer	Annotation package for TxDb object(s)
org	org.Sc.sgd.db	Bioconductor Package Maintainer	Genome wide annotation for Yeast

org 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
```

org 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
```

org 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
```

org 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
```

org 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
| CENTRALTD: 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
```

org 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
```

org 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
```

org 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
```

org 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
```

org 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
```

org 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

org packages

- Different sources → different nomenclatures

```
> keytypes(org.Sc.sgd.db)
[1] "ALIAS"          "COMMON"        "DESCRIPTION"    "ENSEMBL"      "ENSEMLPROT"
[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
```

org packages

- Different sources → different nomenclatures

```
> keytypes(org.Sc.sgd.db)
[1] "ALIAS"          "COMMON"         "DESCRIPTION"    "ENSEMBL"        "ENSEMLPROT"
[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
```

org 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

- 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, ...)
```

```
> 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
```

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) %>%
+   summarise(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      Q0010,00017 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 CCACACCACACCCACACACACACACACACAC...GTGTGGGTGTGGTGTGGTGTGGGTGTGGGTGTGGTGTGGG chrI
[2] 813184 AAATAGCCCTCATGTACGTCTCTCCAAGCCCTGTTGTCTCTTACCCGGATG...AGGGTGTGGGTGTGGTGTGGTGTGGGTGTGGG chrII
[3] 316620 CCCACACACCCACACCCACACACACACACACAC...GGTGTGGGTGTGGGTGTGGGTGTGGGTGTGGGTGTGGGTGTGTG chrIII
[4] 1531933 ACACCACACCCACACCCACACACACACACACAC...AGACAATCTATAAAAAGTAAACATAAAATAAGGTAGTAAGTAGCTTTGG chrIV
[5] 576874 CGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGATGTTCAACCAAAAGCTAC...TAGAAGTGCTGTAGGGCTAAAGAACAGGGTTTCATT... chrV
...
[13] 924431 CCACACACACACCCACACACACACACACACACAC...TGGGTGTGGTGTGGTGTGTGGGTGTGGGTGTGGGTGTGGG chrXIII
[14] 784333 CCGGCTTCTGACCGAAATTAAAAAAAAAAAAATGAAAATGAAACCCCTGTTCT...GGGTGTGGTGTGGGTGTGGGTGTGGGTGTGGG chrXIV
[15] 1091291 ACACCACACCCACACCCACACACACACACACAC...ATGACAGCGAGAGTAGAGGTAGATGTGAGAGAGTGTGTGGGTGTGGTGTGT chrXV
[16] 948066 AAATAGCCCTCATGTACGTCTCTCCAAGCCCTGTTGTCTCTTACCCGGATG...CTAAAGAACAGGGTTTCATTTCATTTTTTAATTTCGGTCAGAAA chrXVI
[17] 85779 TTCATAATTAATTTTATATATTATATTATAATTAATTTAT...ATAGTTATATTATACAGAAATATGCTTAATTATAATATAATATCCATA 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
[1]   464 ATATATTATATTATTTTATATAATATATTATAATTATTATTTAATT... TATAAACTAACCTTAATTAAATAATTAAATAATATAACTTTATTTTATAA YAL001C
[2]   291 ATATTAATAATATATATTATTATTATAATTGAAAACCTATCCTATATT... ATAAATATTCAATTCTTACTAATTAAATAAAAAAGTTTTATATTCAATTAA YAL002W
[3] 12884 ATGGTACAAAGATGATTATATTCAACAAATGAAAAGATATTGCAGTATTAT... ACTTCTCCACCAGCTGTACACTCATTTAATACACCAGCTGTACAATCTAA YAL003W
[4]  1100 ATATTAATATTATTAATAATAATTAACTAATAATAAAAGTTTTATAGAAA... AATAGAGTTAAAATTATTATAATAAACTTCATTATAACAATATCGAATAA YAL004W
[5]   147 ATGCCACAATTAGTTCATTATTATTTATGAATCAATTAAACATATGGTTCT... CCTATGATCTTAAGATTATGTATCTAGATTATTATTTCTAAATTATAA YAL005C
...
[6530]   393 ATGGTAAGTTTCATAACGTCTAGGCAACTCAAGGGCTAATTGAAAATCAGA... TGGGAGACCCATTGTAGAGAGAGTAACCTTAAATTGATTGTTAGTGGTTGA Q0160
[6531] 1215 ATGTCAGAAGATCAAAAAAGTGAAAATTGGTACCTTCTAAGGTTAATATGG... ATAGTCGCAGGGATCCTTAAACCATATTATATGGAACAAATAGAAATTAA Q0182
[6532] 1157 ATGGAAATTGAAAACGAACGTACGTACCGACTACTTTATTGGCAGGCCGG... AACAAAAAGGACCGATTGGGTGATGTTTGATGTGCTGCCAAGTTGA Q0255
[6533]  483 ATGATGCCCTGCTAAACTGCAGCTTGACGTACTGCAGGCCCTGCAGTCCAGCG... CAAGTATTCTTGTATTCCCTGTCATTTCGCAGCATTCTCCACAGCTAG Q0275
[6534] 3099 ATGGCAGACACACCCCTCTGGCAGTACAGGGCCCACCGGGCTATGGTAAGA... AGATTGATATATTATTATGTAGAGATTGAGCAGAGAAGTTGGAGAGTGA 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...TATTTATAA
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> ATGATGCC...CCACAGCTAG
YPR204W  chrXVI  944603-947701     + |  YPR204W    YPR204W    S000006408    <NA> ATGGCAGACA...TGGAGAGTGA
-----
seqinfo: 17 sequences (1 circular) from sacCer3 genome
```

TxDb + BSgenome

```
# Get all introns from TxDb
int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()
```

TxDb + BSgenome

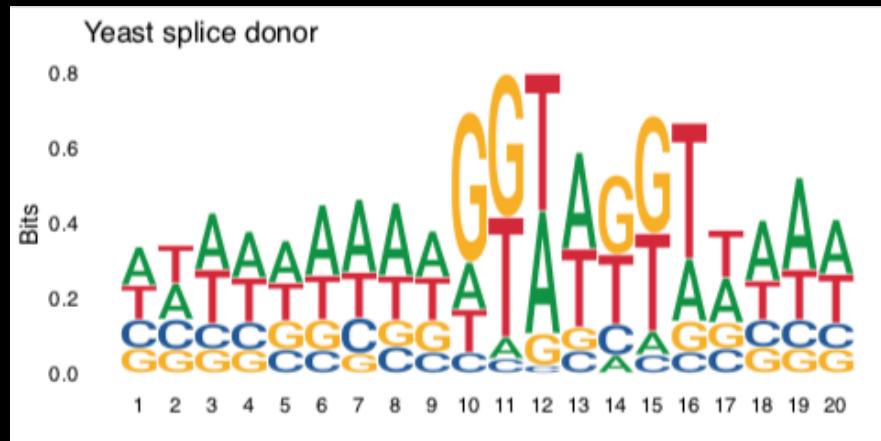
```
# Get all introns from TxDb
int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()

# Get all intron 5' ends
int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')
int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')
```

TxDb + BSgenome

```
# Get all introns from TxDb
int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()

# Get all intron 5' ends and plot their seqLogo
int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')
int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')
p_donor <- bs[int_donor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +
  ggtitle(paste0(org, ' splice donor'))
```



TxDb + BSgenome

```
# Get all introns from TxDb
int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()

# Get all intron 5' ends and plot their seqLogo
int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')
int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')
p_donor <- bs[int_donor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +
  ggtitle(paste0(org, ' splice donor'))

# Get all intron 3' ends and plot their seqLogo
int_acceptor <- resize(int, fix = 'end', 1) %>% resize(20, fix = 'center')
int_acceptor <- subsetByOverlaps(int_acceptor, as(seqinfo(bs), 'GRanges'), type = 'within')
p_acceptor <- bs[int_acceptor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +
  ggtitle(paste0(org, ' splice acceptor'))
```

TxDb + BSgenome

```
function(tx, bs, org) {  
  
  # Get all introns from TxDb  
  int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()  
  
  # Get all intron 5' ends and plot their seqLogo  
  int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')  
  int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')  
  p_donor <- bs[int_donor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +  
    ggtitle(paste0(org, ' splice donor'))  
  
  # Get all intron 3' ends and plot their seqLogo  
  int_acceptor <- resize(int, fix = 'end', 1) %>% resize(20, fix = 'center')  
  int_acceptor <- subsetByOverlaps(int_acceptor, as(seqinfo(bs), 'GRanges'), type = 'within')  
  p_acceptor <- bs[int_acceptor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +  
    ggtitle(paste0(org, ' splice acceptor'))  
  
  # Return both plots as a list  
  return(list(p_donor, p_acceptor))  
  
}
```

TxDb + BSgenome

```
TXs <- list(
  TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::TxDb.Scerevisiae.UCSC.sacCer3.sgdGene,
  TxDb.Celegans.UCSC.ce11.ensGene::TxDb.Celegans.UCSC.ce11.ensGene,
  TxDb.Dmelanogaster.UCSC.dm3.ensGene::TxDb.Dmelanogaster.UCSC.dm3.ensGene,
  TxDb.Drerio.UCSC.danRer11.refGene::TxDb.Drerio.UCSC.danRer11.refGene,
  TxDb.Rnorvegicus.UCSC.rn6.refGene::TxDb.Rnorvegicus.UCSC.rn6.refGene,
  TxDb.Mmusculus.UCSC.mm10.knownGene::TxDb.Mmusculus.UCSC.mm10.knownGene,
  TxDb.Hsapiens.UCSC.hg38.knownGene::TxDb.Hsapiens.UCSC.hg38.knownGene,
  TxDb.Athaliana.BioMart.plantsmart28::TxDb.Athaliana.BioMart.plantsmart28
)
BSs <- list(
  Biostrings::getSeq(BSgenome.Scerevisiae.UCSC.sacCer3)::BSgenome.Scerevisiae.UCSC.sacCer3),
  Biostrings::getSeq(BSgenome.Celegans.UCSC.ce11)::BSgenome.Celegans.UCSC.ce11),
  Biostrings::getSeq(BSgenome.Dmelanogaster.UCSC.dm3)::BSgenome.Dmelanogaster.UCSC.dm3),
  Biostrings::getSeq(BSgenome.Drerio.UCSC.danRer11)::BSgenome.Drerio.UCSC.danRer11),
  Biostrings::getSeq(BSgenome.Rnorvegicus.UCSC.rn6)::BSgenome.Rnorvegicus.UCSC.rn6),
  Biostrings::getSeq(BSgenome.Mmusculus.UCSC.mm10)::BSgenome.Mmusculus.UCSC.mm10),
  Biostrings::getSeq(BSgenome.Hsapiens.UCSC.hg38)::BSgenome.Hsapiens.UCSC.hg38),
  Biostrings::getSeq(BSgenome.Athaliana.TAIR.TAIR9)::BSgenome.Athaliana.TAIR.TAIR9)
)
orgs <- list(
  "Yeast",
  "Nematode",
  "Fly",
  "Fish",
  "Rat",
  "Mouse",
  "Human",
  "Plant"
)
```

TxDb + BSgenome

```
plotlist <- mapply(  
  function(tx, bs, org) {  
  
    # Get all introns from TxDb  
    int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()  
  
    # Get all intron 5' ends and plot their seqLogo  
    int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')  
    int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')  
    p_donor <- bs[int_donor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +  
      ggtitle(paste0(org, ' splice donor'))  
  
    # Get all intron 3' ends and plot their seqLogo  
    int_acceptor <- resize(int, fix = 'end', 1) %>% resize(20, fix = 'center')  
    int_acceptor <- subsetByOverlaps(int_acceptor, as(seqinfo(bs), 'GRanges'), type = 'within')  
    p_acceptor <- bs[int_acceptor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +  
      ggtitle(paste0(org, ' splice acceptor'))  
  
    # Return both plots as a list  
    return(list(p_donor, p_acceptor))  
  },  
  TXs, BSs, orgs,  
  SIMPLIFY = FALSE  
)
```

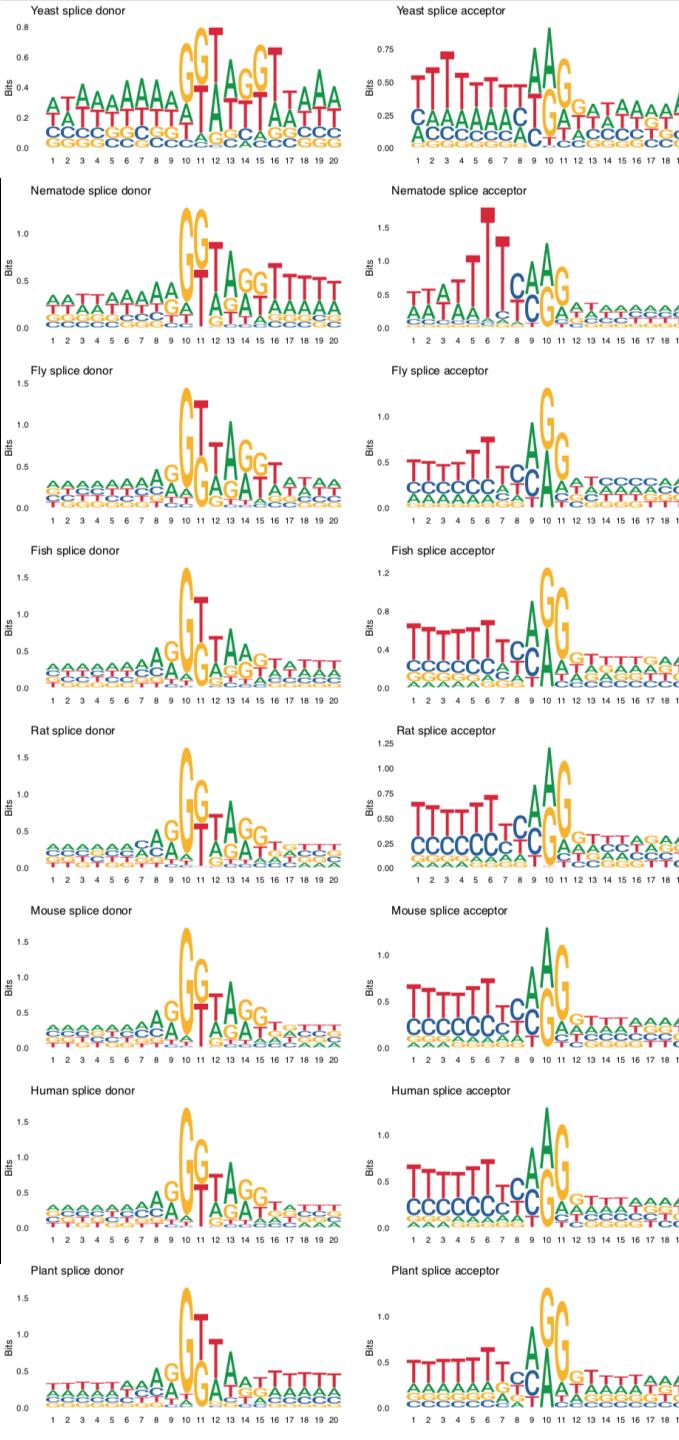
TxDb + BSgenome

```
plotlist <- mapply(
  function(tx, bs, org) {
    # Get all introns from TxDb
    int <- GenomicFeatures::intronsByTranscript(tx) %>% unlist()

    # Get all intron 5' ends and plot their seqLogo
    int_donor <- resize(int, fix = 'start', 1) %>% resize(20, fix = 'center')
    int_donor <- subsetByOverlaps(int_donor, as(seqinfo(bs), 'GRanges'), type = 'within')
    p_donor <- bs[int_donor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +
      ggtitle(paste0(org, ' splice donor'))

    # Get all intron 3' ends and plot their seqLogo
    int_acceptor <- resize(int, fix = 'end', 1) %>% resize(20, fix = 'center')
    int_acceptor <- subsetByOverlaps(int_acceptor, as(seqinfo(bs), 'GRanges'), type = 'within')
    p_acceptor <- bs[int_acceptor] %>% as.character() %>% ggseqlogo::ggseqlogo(namespace = 'ATCGN') +
      ggtitle(paste0(org, ' splice acceptor'))

    # Return both plots as a list
    return(list(p_donor, p_acceptor))
  },
  TXs, BSs, orgs,
  SIMPLIFY = FALSE
)
cowplot::plot_grid(plotlist = purrr::flatten(plotlist), ncol = 2)
```

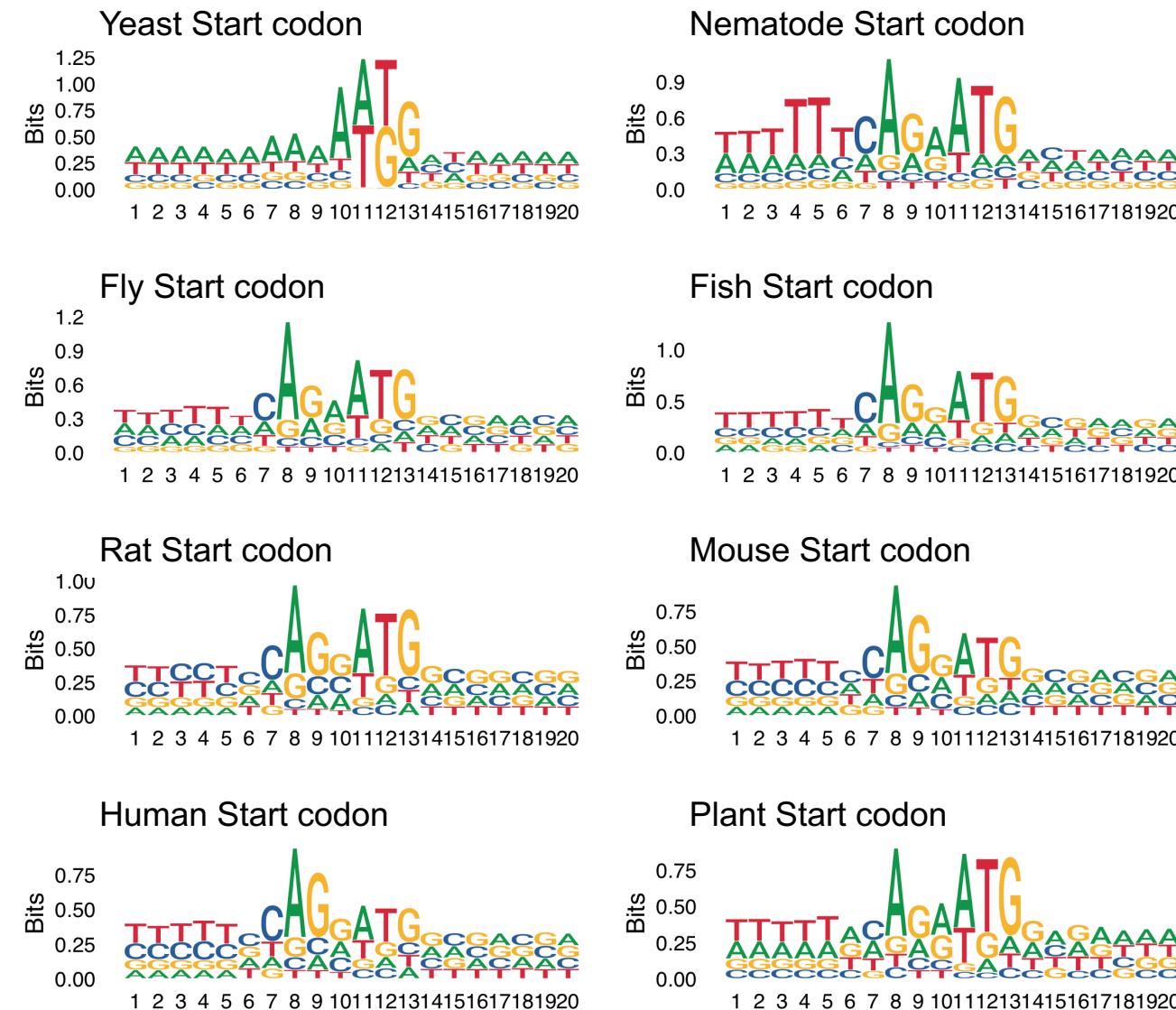


2020/01/13

BioC resources and access to public databases

Jacques Serizay

TxDb + BSgenome



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
```

AnnotationHub: 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
# $datatype: UCSC
# $species: Saccharomyces cerevisiae
# $rdataclass: TwoBitFile
# $rdatadateadded: 2014-12-15
# $title: sacCer3.2bit
# $description: UCSC 2 bit file for sacCer3
# $taxononyid: 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[["..."]]`

AnnotationHub: retrieving release-specific files

- Many **many** resources available on AnnotationHub

AnnotationHub: retrieving release-specific files

```
> 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  
# $dataprovier: 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 | wgEncodeCsh1LongRnaSeqNhemfm2CellTotalGeneGencV10.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