Introduction to Bioconductor

Maruf Ahmed Bhuiyan

7/24/2020

Table of Contents

# Install and Load BiocManager and BSgenome

# Install BiocManager & Load using library  
library(BiocManager)

## Bioconductor version 3.10 (BiocManager 1.30.10), ?BiocManager::install for help

## Bioconductor version '3.10' is out-of-date; the current release version '3.11'  
## is available with R version '4.0'; see https://bioconductor.org/install

# Install BSgenome  
#BiocManager::install("BSgenome")  
  
# Install Scerevisiae genome from BSgenome  
#BiocManager::install("BSgenome.Scerevisiae.UCSC.sacCer3")  
  
packageVersion("BSgenome")

## [1] '1.54.0'

# Load the genome  
library(BSgenome.Scerevisiae.UCSC.sacCer3)

## Loading required package: BSgenome

## Loading required package: BiocGenerics

## Loading required package: parallel

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, basename, cbind, colnames,  
## dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,  
## grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,  
## order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,  
## rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,  
## union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':  
##   
## expand.grid

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: GenomicRanges

## Loading required package: Biostrings

## Loading required package: XVector

##   
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':  
##   
## strsplit

## Loading required package: rtracklayer

# S4 class definition

We will use the class BSgenome. Let’s check the formal definition of this class by using the function showClass(“className”). Check the BSgenome class results and find its parent classes (Extends) and the classes that inherit from it (Subclasses).

showClass("BSgenome")

## Class "BSgenome" [package "BSgenome"]  
##   
## Slots:  
##   
## Name: pkgname single\_sequences multiple\_sequences  
## Class: character OnDiskNamedSequences RdaCollection  
##   
## Name: source\_url user\_seqnames injectSNPs\_handler  
## Class: character character InjectSNPsHandler  
##   
## Name: .seqs\_cache .link\_counts nmask\_per\_seq  
## Class: environment environment integer  
##   
## Name: masks organism common\_name  
## Class: RdaCollection character character  
##   
## Name: provider provider\_version release\_date  
## Class: character character character  
##   
## Name: release\_name seqinfo  
## Class: character Seqinfo  
##   
## Extends: "GenomeDescription"  
##   
## Known Subclasses: "MaskedBSgenome"

#Extends: "GenomeDescription"  
  
#Known Subclasses: "MaskedBSgenome"  
  
isS4(BSgenome)

## [1] FALSE

# BSgenome is S3

# Discovering the Yeast genome

Let’s continue to explore the yeast genome using the package BSgenome.Scerevisiae.UCSC.sacCer3 which is already installed for you.

As with other data in R, we can use head() and tail() to explore the yeastGenome object. We can also subset the genome by chromosome by using $ as follows: object\_name$chromosome\_name. If you need the names of the chromosomes use the names() function.

Another nifty function is nchar(), used to count the number of characters in a sequence.

yeast <- BSgenome.Scerevisiae.UCSC.sacCer3  
  
# How many chromosomes  
length(yeast)

## [1] 17

# What are the name of the chromosomes  
names(yeast)

## [1] "chrI" "chrII" "chrIII" "chrIV" "chrV" "chrVI" "chrVII"   
## [8] "chrVIII" "chrIX" "chrX" "chrXI" "chrXII" "chrXIII" "chrXIV"   
## [15] "chrXV" "chrXVI" "chrM"

# Length of each chromosome  
seqlengths(yeast)

## chrI chrII chrIII chrIV chrV chrVI chrVII chrVIII chrIX chrX   
## 230218 813184 316620 1531933 576874 270161 1090940 562643 439888 745751   
## chrXI chrXII chrXIII chrXIV chrXV chrXVI chrM   
## 666816 1078177 924431 784333 1091291 948066 85779

head(yeast$chrM)

## 6-letter "DNAString" instance  
## seq: TTCATA

tail(yeast$chrM)

## 6-letter "DNAString" instance  
## seq: TCCATA

# sequence length of chr M  
nchar(yeast$chrM)

## [1] 85779

print(yeast$chrM)

## 85779-letter "DNAString" instance  
## seq: TTCATAATTAATTTTTTATATATATATTATATTATA...TACAGAAATATGCTTAATTATAATATAATATCCATA

# Partitioning the Yeast genome

Genomes are often big, but interest usually lies in specific regions of them. Therefore, we need to subset a genome by extracting parts of it. To pick a sequence interval use getSeq() and specify the name of the chromosome, and the start and end of the sequence interval:

getSeq(yeastGenome, names = “chrI”, start = 100, end = 150) Notice that names is optional; if not specified, it will return all chromosomes. The parameters start and end are also optional and, if not specified, will take the default values 1 and the length of the sequence respectively.

# getSeq(yeastGenome, names = "chrI", start = 100, end = 150)  
  
getSeq(yeast, end = 30)

## A DNAStringSet instance of length 17  
## width seq names   
## [1] 30 CCACACCACACCCACACACCCACACACCAC chrI  
## [2] 30 AAATAGCCCTCATGTACGTCTCCTCCAAGC chrII  
## [3] 30 CCCACACACCACACCCACACCACACCCACA chrIII  
## [4] 30 ACACCACACCCACACCACACCCACACACAC chrIV  
## [5] 30 CGTCTCCTCCAAGCCCTGTTGTCTCTTACC chrV  
## ... ... ...  
## [13] 30 CCACACACACACCACACCCACACCACACCC chrXIII  
## [14] 30 CCGGCTTTCTGACCGAAATTAAAAAAAAAA chrXIV  
## [15] 30 ACACCACACCCACACCACACCCACACCCAC chrXV  
## [16] 30 AAATAGCCCTCATGTACGTCTCCTCCAAGC chrXVI  
## [17] 30 TTCATAATTAATTTTTTATATATATATTAT chrM

# Available Genomes

As a recap, the BSgenome package makes available various public genomes. If you want to explore the available genomes from this package, you can use:

available.genomes() The list of names will appear in the following format: BSgenome.speciesName.provider.version.

After running this function, can you tell which is the major provider of available genomes?

available.genomes()

## [1] "BSgenome.Alyrata.JGI.v1"   
## [2] "BSgenome.Amellifera.BeeBase.assembly4"   
## [3] "BSgenome.Amellifera.UCSC.apiMel2"   
## [4] "BSgenome.Amellifera.UCSC.apiMel2.masked"   
## [5] "BSgenome.Aofficinalis.NCBI.V1"   
## [6] "BSgenome.Athaliana.TAIR.04232008"   
## [7] "BSgenome.Athaliana.TAIR.TAIR9"   
## [8] "BSgenome.Btaurus.UCSC.bosTau3"   
## [9] "BSgenome.Btaurus.UCSC.bosTau3.masked"   
## [10] "BSgenome.Btaurus.UCSC.bosTau4"   
## [11] "BSgenome.Btaurus.UCSC.bosTau4.masked"   
## [12] "BSgenome.Btaurus.UCSC.bosTau6"   
## [13] "BSgenome.Btaurus.UCSC.bosTau6.masked"   
## [14] "BSgenome.Btaurus.UCSC.bosTau8"   
## [15] "BSgenome.Btaurus.UCSC.bosTau9"   
## [16] "BSgenome.Carietinum.NCBI.v1"   
## [17] "BSgenome.Celegans.UCSC.ce10"   
## [18] "BSgenome.Celegans.UCSC.ce11"   
## [19] "BSgenome.Celegans.UCSC.ce2"   
## [20] "BSgenome.Celegans.UCSC.ce6"   
## [21] "BSgenome.Cfamiliaris.UCSC.canFam2"   
## [22] "BSgenome.Cfamiliaris.UCSC.canFam2.masked"   
## [23] "BSgenome.Cfamiliaris.UCSC.canFam3"   
## [24] "BSgenome.Cfamiliaris.UCSC.canFam3.masked"   
## [25] "BSgenome.Cjacchus.UCSC.calJac3"   
## [26] "BSgenome.Dmelanogaster.UCSC.dm2"   
## [27] "BSgenome.Dmelanogaster.UCSC.dm2.masked"   
## [28] "BSgenome.Dmelanogaster.UCSC.dm3"   
## [29] "BSgenome.Dmelanogaster.UCSC.dm3.masked"   
## [30] "BSgenome.Dmelanogaster.UCSC.dm6"   
## [31] "BSgenome.Drerio.UCSC.danRer10"   
## [32] "BSgenome.Drerio.UCSC.danRer11"   
## [33] "BSgenome.Drerio.UCSC.danRer5"   
## [34] "BSgenome.Drerio.UCSC.danRer5.masked"   
## [35] "BSgenome.Drerio.UCSC.danRer6"   
## [36] "BSgenome.Drerio.UCSC.danRer6.masked"   
## [37] "BSgenome.Drerio.UCSC.danRer7"   
## [38] "BSgenome.Drerio.UCSC.danRer7.masked"   
## [39] "BSgenome.Ecoli.NCBI.20080805"   
## [40] "BSgenome.Gaculeatus.UCSC.gasAcu1"   
## [41] "BSgenome.Gaculeatus.UCSC.gasAcu1.masked"   
## [42] "BSgenome.Ggallus.UCSC.galGal3"   
## [43] "BSgenome.Ggallus.UCSC.galGal3.masked"   
## [44] "BSgenome.Ggallus.UCSC.galGal4"   
## [45] "BSgenome.Ggallus.UCSC.galGal4.masked"   
## [46] "BSgenome.Ggallus.UCSC.galGal5"   
## [47] "BSgenome.Ggallus.UCSC.galGal6"   
## [48] "BSgenome.Hsapiens.1000genomes.hs37d5"   
## [49] "BSgenome.Hsapiens.NCBI.GRCh38"   
## [50] "BSgenome.Hsapiens.UCSC.hg17"   
## [51] "BSgenome.Hsapiens.UCSC.hg17.masked"   
## [52] "BSgenome.Hsapiens.UCSC.hg18"   
## [53] "BSgenome.Hsapiens.UCSC.hg18.masked"   
## [54] "BSgenome.Hsapiens.UCSC.hg19"   
## [55] "BSgenome.Hsapiens.UCSC.hg19.masked"   
## [56] "BSgenome.Hsapiens.UCSC.hg38"   
## [57] "BSgenome.Hsapiens.UCSC.hg38.masked"   
## [58] "BSgenome.Mdomestica.UCSC.monDom5"   
## [59] "BSgenome.Mfascicularis.NCBI.5.0"   
## [60] "BSgenome.Mfuro.UCSC.musFur1"   
## [61] "BSgenome.Mmulatta.UCSC.rheMac10"   
## [62] "BSgenome.Mmulatta.UCSC.rheMac2"   
## [63] "BSgenome.Mmulatta.UCSC.rheMac2.masked"   
## [64] "BSgenome.Mmulatta.UCSC.rheMac3"   
## [65] "BSgenome.Mmulatta.UCSC.rheMac3.masked"   
## [66] "BSgenome.Mmulatta.UCSC.rheMac8"   
## [67] "BSgenome.Mmusculus.UCSC.mm10"   
## [68] "BSgenome.Mmusculus.UCSC.mm10.masked"   
## [69] "BSgenome.Mmusculus.UCSC.mm8"   
## [70] "BSgenome.Mmusculus.UCSC.mm8.masked"   
## [71] "BSgenome.Mmusculus.UCSC.mm9"   
## [72] "BSgenome.Mmusculus.UCSC.mm9.masked"   
## [73] "BSgenome.Osativa.MSU.MSU7"   
## [74] "BSgenome.Ptroglodytes.UCSC.panTro2"   
## [75] "BSgenome.Ptroglodytes.UCSC.panTro2.masked"  
## [76] "BSgenome.Ptroglodytes.UCSC.panTro3"   
## [77] "BSgenome.Ptroglodytes.UCSC.panTro3.masked"  
## [78] "BSgenome.Ptroglodytes.UCSC.panTro5"   
## [79] "BSgenome.Ptroglodytes.UCSC.panTro6"   
## [80] "BSgenome.Rnorvegicus.UCSC.rn4"   
## [81] "BSgenome.Rnorvegicus.UCSC.rn4.masked"   
## [82] "BSgenome.Rnorvegicus.UCSC.rn5"   
## [83] "BSgenome.Rnorvegicus.UCSC.rn5.masked"   
## [84] "BSgenome.Rnorvegicus.UCSC.rn6"   
## [85] "BSgenome.Scerevisiae.UCSC.sacCer1"   
## [86] "BSgenome.Scerevisiae.UCSC.sacCer2"   
## [87] "BSgenome.Scerevisiae.UCSC.sacCer3"   
## [88] "BSgenome.Sscrofa.UCSC.susScr11"   
## [89] "BSgenome.Sscrofa.UCSC.susScr3"   
## [90] "BSgenome.Sscrofa.UCSC.susScr3.masked"   
## [91] "BSgenome.Tgondii.ToxoDB.7.0"   
## [92] "BSgenome.Tguttata.UCSC.taeGut1"   
## [93] "BSgenome.Tguttata.UCSC.taeGut1.masked"   
## [94] "BSgenome.Tguttata.UCSC.taeGut2"   
## [95] "BSgenome.Vvinifera.URGI.IGGP12Xv0"   
## [96] "BSgenome.Vvinifera.URGI.IGGP12Xv2"   
## [97] "BSgenome.Vvinifera.URGI.IGGP8X"

# UCSC- 74 genomes