PH525.5x Section 4: Genomic annotation with Bioconductor

Lauren Kemperman

4/13/2021

Representing Reference Sequence

- Annotation concept hierarchy
- Base - reference genomic sequence for an organism
- Above this, organize the chromosomal sequence into regions of interest - i.e. genes, transcripts
- SNPs and CpG sites are also regions of interest
- SNPS are single nucleotide
- Other varients – indels, structural variants, fusions can constitute regions of interest but are more complicated to express + represent
- Within ROI, identify platform oriented annotation provided by assay manufacturer
- Once manufacturing happens, genomic annotation proceeds and annotations must be updated to account for ambiguities or updates for assay probe elements
- Above genomic sequence ROIs, annotations concerning groups with shared structural or functional properties
- Pathways with nodes being genes and paths being relationships between gene products, i.e. protein protein interaction, promotion, enhancement, repression (3rd level of hierarchy)
- Begin with reference genomes
- Biostrings package - **available.genomes** - packages that represent reference genomic sequences for many different organisms
- Homo sapiens reference - some have repeat masking and there are versions which include the masked regions
  - different numbers of sequences in the two builds due to contigs that haven't been placed on chromosomes yet
- Operations defined for BSGenome objects - substring, extract chromosomal information
- Bases in full sequence aren't completely resolved
- Application of iteration - count the number of bases in a number of chromosomes
- If you have enough RAM, it is possible to operate on chromosomes in parallel and performing operations using multicore programming

```
library(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.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

## Warning: package 'GenomeInfoDb' was built under R version 4.0.5

## 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
```

```
library(Biostrings)
ag = available.genomes()
grep("Scerev", ag, value=TRUE)
```

```
## [1] "BSgenome.Scerevisiae.UCSC.sacCer1" "BSgenome.Scerevisiae.UCSC.sacCer2"
## [3] "BSgenome.Scerevisiae.UCSC.sacCer3"
```

```
grep("Hsap", ag, value=TRUE)
```

```
##  [1] "BSgenome.Hsapiens.1000genomes.hs37d5"
##  [2] "BSgenome.Hsapiens.NCBI.GRCh38"
##  [3] "BSgenome.Hsapiens.UCSC.hg17"
##  [4] "BSgenome.Hsapiens.UCSC.hg17.masked"
##  [5] "BSgenome.Hsapiens.UCSC.hg18"
##  [6] "BSgenome.Hsapiens.UCSC.hg18.masked"
##  [7] "BSgenome.Hsapiens.UCSC.hg19"
##  [8] "BSgenome.Hsapiens.UCSC.hg19.masked"
##  [9] "BSgenome.Hsapiens.UCSC.hg38"
## [10] "BSgenome.Hsapiens.UCSC.hg38.masked"
```

```
# inspect the human genome
library(BSgenome.Hsapiens.UCSC.hg19)
Hsapiens
```

```
## Human genome:
## # organism: Homo sapiens (Human)
## # genome: hg19
## # provider: UCSC
## # release date: June 2013
## # 298 sequences:
## #   chr1                 chr2                 chr3
## #   chr4                 chr5                 chr6
## #   chr7                 chr8                 chr9
## #   chr10                chr11                chr12
## #   chr13                chr14                chr15
## #   ...                  ...                  ...
## #   chr19_gl949749_alt   chr19_gl949750_alt   chr19_gl949751_alt
## #   chr19_gl949752_alt   chr19_gl949753_alt   chr20_gl383577_alt
## #   chr21_gl383578_alt   chr21_gl383579_alt   chr21_gl383580_alt
## #   chr21_gl383581_alt   chr22_gl383582_alt   chr22_gl383583_alt
## #   chr22_kb663609_alt
## # (use 'seqnames()' to see all the sequence names, use the '$' or '[[' operator
## # to access a given sequence)
```

```
length(Hsapiens)
```

```
## [1] 298
```

```
class(Hsapiens)
```

```
## [1] "BSgenome"
## attr(,"package")
## [1] "BSgenome"
```

```
methods(class="BSgenome")
```

```
##  [1] [[              $               as.list         bsgenomeName
##  [5] coerce          commonName      countPWM        export
##  [9] extractAt       getSeq          injectSNPs      length
## [13] masknames       matchPWM        metadata        metadata<-
## [17] mseqnames       names           organism        provider
## [21] providerVersion releaseDate     releaseName     seqinfo
## [25] seqinfo<-       seqnames        seqnames<-      show
## [29] snpcount        SNPlocs_pkgname snplocs         sourceUrl
## [33] vcountPattern   vcountPDict     Views           vmatchPattern
## [37] vmatchPDict
```

```
## see '?methods' for accessing help and source code
```

```
# inspect human genome
Hsapiens$chrX
```

```
## 155270560-letter DNAString object
## seq: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

```
substr(Hsapiens$chrX, 5e6, 5.1e6)
```

```
## 100001-letter DNAString object
## seq: GCCTCAATGTCAGAATTATGCTGTTGCCCAAAATTG...TACTAAAAATACAAAAATTAGCTGGGCATGGTGGTG
```

```
nchar(Hsapiens$chrY)
```

```
## [1] 59373566
```

```
nchar(Hsapiens[[24]])
```

```
## [1] 59373566
```

```
library(parallel)
options(mc.cores=detectCores())

system.time(sum(unlist(mclapply(18:24, function(x) nchar(Hsapiens[[x]])))))
```

```
##    user  system elapsed
##   4.885   3.661   8.209
```

Assessment: Reference Genomes

```
library(BSgenome)
library(Biostrings)
ag = available.genomes()
library(BSgenome)
grep("mask", grep("Drerio", available.genomes(), value=TRUE), invert=TRUE, value=TRUE) # exclude masked
```

```
## [1] "BSgenome.Drerio.UCSC.danRer10" "BSgenome.Drerio.UCSC.danRer11"
## [3] "BSgenome.Drerio.UCSC.danRer5"  "BSgenome.Drerio.UCSC.danRer6"
## [5] "BSgenome.Drerio.UCSC.danRer7"
```

```
library(BSgenome.Hsapiens.UCSC.hg19.masked)
c17m = BSgenome.Hsapiens.UCSC.hg19.masked$chr17

c22m = BSgenome.Hsapiens.UCSC.hg19.masked$chr22
round(100*sum(width(masks(c22m)$AGAPS))/length(c22m),0)
```

```
## [1] 32
```

Gene, Transcript and Exon Databases

- Can find information about reference genome regions such as genes, transcripts and exons on annotation packages
- UCSC Genome Browser - major source of reference genome structure annotation
- **TxDb.Hsapiens.UCSC.hg19** - collection of well documented protein coding genes, transcripts and exons on the hg19 build of the human genome. Additional TxDb packages exist for other organisms and genome builds

- Introduction to TxDb package architecture

```
# Import TxDb transcript database
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
```

```
## Loading required package: GenomicFeatures
```

```
## Warning: package 'GenomicFeatures' was built under R version 4.0.4
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
txdb = TxDb.Hsapiens.UCSC.hg19.knownGene
class(txdb)
```

```
## [1] "TxDb"
## attr(,"package")
## [1] "GenomicFeatures"
```

```
methods(class="TxDb")
```

```
##  [1] $                     $<-                   annotatedDataFrameFrom
##  [4] as.list               asBED                 asGFF
##  [7] assayData             assayData<-           cds
## [10] cdsBy                 cdsByOverlaps         coerce
## [13] columns               combine               contents
## [16] dbconn                dbfile                dbInfo
## [19] dbmeta                dbschema              disjointExons
## [22] distance              exons                 exonsBy
## [25] exonsByOverlaps       ExpressionSet         extractUpstreamSeqs
## [28] featureNames          featureNames<-        fiveUTRsByTranscript
## [31] genes                 initialize            intronsByTranscript
## [34] isActiveSeq           isActiveSeq<-         isNA
## [37] keys                  keytypes              mapIds
## [40] mapIdsToRanges        mappedkeys            mapRangesToIds
## [43] mapToTranscripts      metadata              microRNAs
## [46] nhit                  organism              promoters
## [49] revmap                sample                sampleNames
## [52] sampleNames<-         saveDb                select
## [55] seqinfo               seqinfo<-             seqlevels<-
## [58] seqlevels0            show                  species
## [61] storageMode           storageMode<-         taxonomyId
## [64] threeUTRsByTranscript transcripts           transcriptsBy
## [67] transcriptsByOverlaps tRNAs                 updateObject
## see '?methods' for accessing help and source code
```

```
# extract and inspect genes from TxDb
genes(txdb)
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
```

```
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
## GRanges object with 23056 ranges and 1 metadata column:
##         seqnames              ranges strand |     gene_id
##            <Rle>           <IRanges>  <Rle> | <character>
##       1    chr19   58858172-58874214      - |           1
##      10     chr8   18248755-18258723      + |          10
##     100    chr20   43248163-43280376      - |         100
##    1000    chr18   25530930-25757445      - |        1000
##   10000     chr1 243651535-244006886      - |       10000
##     ...      ...                 ...    ... .         ...
##    9991     chr9 114979995-115095944      - |        9991
##    9992    chr21   35736323-35743440      + |        9992
##    9993    chr22   19023795-19109967      - |        9993
##    9994     chr6   90539619-90584155      + |        9994
##    9997    chr22   50961997-50964905      - |        9997
##   -------
##   seqinfo: 93 sequences (1 circular) from hg19 genome
```

```
table(strand(genes(txdb)))
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
##
##     +     -     *
## 11737 11319     0
```

```
summary(width(genes(txdb)))
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##       20     5666    20116    60660    58175 24187703
```

```
# inspect larges gene in genome
id = which.max(width(genes(txdb)))
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
genes(txdb)[id]
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
```

```
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
## GRanges object with 1 range and 1 metadata column:
##          seqnames            ranges strand |     gene_id
##             <Rle>         <IRanges>  <Rle> | <character>
##   286297     chr9 42844370-67032072      - |      286297
##   -------
##   seqinfo: 93 sequences (1 circular) from hg19 genome
```

```
library(org.Hs.eg.db)
```

```
##
```

```
select(org.Hs.eg.db, keys="286297", keytype="ENTREZID", columns=c("SYMBOL", "GENENAME"))
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
##   ENTREZID    SYMBOL
## 1   286297 LOC286297
##                                                                  GENENAME
## 1 methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1 like pseudogene
```

```
# compare total size of exons to total size of genes
ex = exons(txdb)
rex = reduce(ex)
ex_width = sum(width(rex)) # bases in exons
gene_width = sum(width(genes(txdb))) # bases in genes
```

```
##   403 genes were dropped because they have exons located on both strands
##   of the same reference sequence or on more than one reference sequence,
##   so cannot be represented by a single genomic range.
##   Use 'single.strand.genes.only=FALSE' to get all the genes in a
##   GRangesList object, or use suppressMessages() to suppress this message.
```

```
ex_width/gene_width
```

```
## [1] 0.06380062
```