Biconductor Exercises

Gopuraja Dharmalingam

04 February 2015

1. Print few gene names from org.Hs.eg.db

2. Print non-redundant list of chromosomes from org.Mm.eg.db

3. Retrieve gene name, chromosome and Ensembl gene identifiers for "HEBP2" and "PRND" from org.Hs.eg.db

4. Retrieve gene symbol, gene name and gene alias for genes in chromosome 2 for human using org.Hs.eg.db

5. Retrieve genomic coordinates for human protein coding genes from Ensembl biomart and build GRanges object. Include genes in only main chromosomes (1-22,X,Y).

Tips:

- You can select main chromosomes and "protein coding" genes by using appropriate filter and value.
- Search for "biotype" in available filters using `grep()`
- Run `filterOptions("biotype",selectedmart)` to see the accepted values for "biotype" filter
- When multiple filters specified, "values" argument should be a list of vectors; each vector corresponds to each specified filter.
- Annotation fields to retrieve: "chromosome_name", "start_position", "end_position", "ensembl_gene_id", "strand", "external_gene_name"
- Before creating GRanges object, add "chr" prefix to chromosome using `paste` function, Ex: change 1 to chr1 (required for next task)

6. Filter the above GRanges object for genes in chr1:1544000-2371000

7. Create a GRanges of Transcription start sites (1 bp range) for the GRanges object created in Q5.

- How to identify TSS for genes in forward/reverse strand?

8. Create a GRanges object of human promoters with TSS ± 2000bp (using the GRanges object created in Q5). Tip: Read the documentation for `promoters` function.

9. Import ELF1 binding sites in K562 cell from Encode (ELF1_K562.bed) and create GRanges object.

Tips:

- Import the ELF1 binding sites using `import.bed()` function from `rtracklayer` package and compare it with the above GRanges object
- Find ELF1 binding sites overlap with promoters (TSS ± 1kb) using `findOverlaps` and `subsetByOverlaps`
- Remember BED format uses 0-based coordinates

10. Retrieve the transcript coordinates for genes as GRangesList from `TxDb.Hsapiens.UCSC.hg19.knownGene` (install it from Bioconductor if required)

11. Retrieve 200 bp upstream promoter sequences for the given gene symbols `AQP1, ASNSP2, KPNA2, FRMD4A, NSUN5, VAC14` from Ensembl human biomart

Tips:

- Read documentation for `getSequence`
- Use `type="hgnc_symbol"` and `seqType="coding_gene_flank"`

Additional Exercise

- Download the following BAM and index files (*.bai) (ENCODE data - ChIP-Seq of CTCF in Ag04449 human fibroblast cells)
  - http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/wgEncodeUwTfbsAg04449CtcfS bam
  - http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/wgEncodeUwTfbsAg04449CtcfS bam.bai
- Use `readGAlignments` to read the bam. Construct an `ScanBamParam` object that accepts only aligned reads, passing quality control and not duplicates.
- Compute genome wide coverage using `coverage` function
- Compute number of reads overlapping with hg19 promoters (TSS ± 1kb) and export the results as text file.