GRanges - Overview

Kasper D Hansen

@KasperDHansen

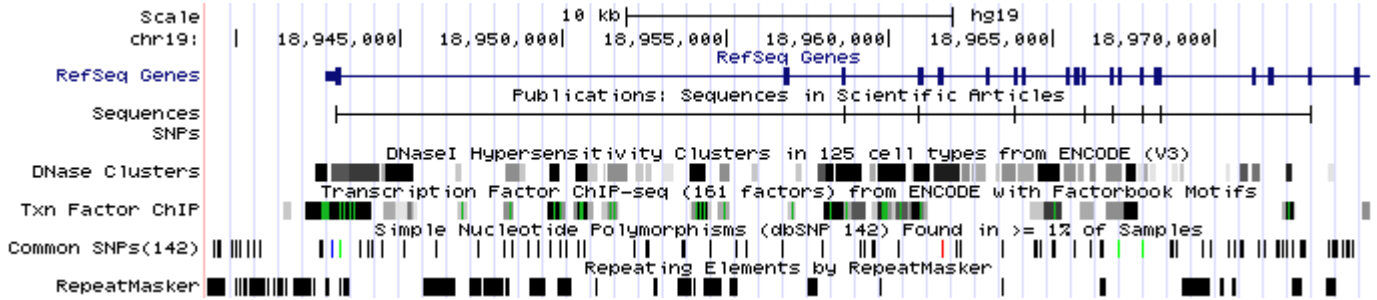
www.hansenlab.org

A GRanges is a data structure for storing genomic intervals.

They are fast and efficient and have transformed my own work.

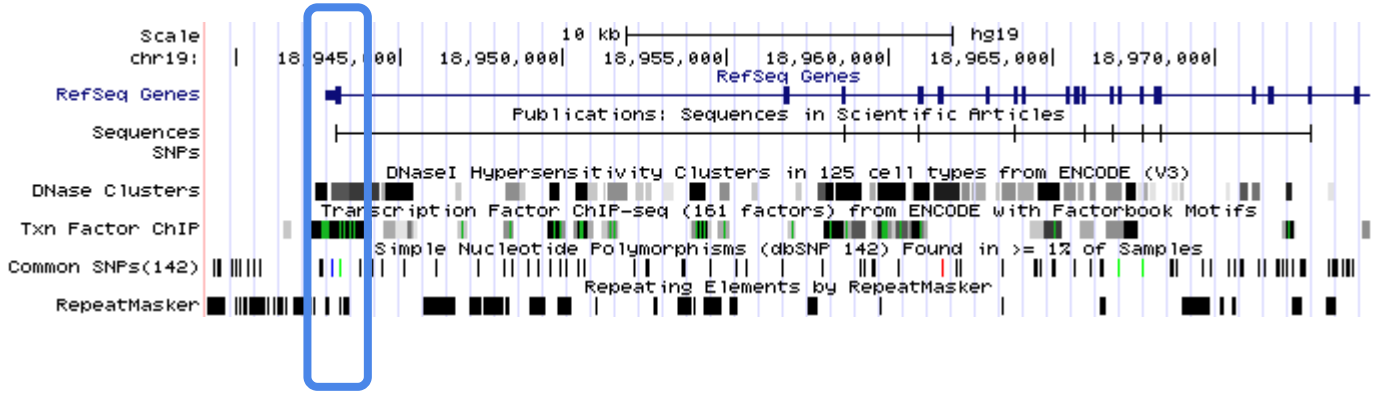
Every R user dealing with genomic data needs to master this material.

Many entities in genomics are intervals or sets of intervals (of integers):



Promoters, Genes, SNPs, CpG Islands, ….
Sequencing reads; mapped and processed.

Many tasks involves relating sets of intervals to each other:



Which promoters contains SNPs?
Which TF binding sites overlap a promoter?
Which genes are covered by sequencing reads?

GRanges

```
GRanges with 3 ranges and 0 metadata columns:
     seqnames    ranges strand
       <Rle> <IRanges>  <Rle>
A1     chr1          [1, 3]        +
A2     chr1          [3, 5]        -
A3     chr1          [5, 7]        +
---
seqlengths:
  chr1
    NA
```

Functionality in the GenomicRanges and IRanges packages.

These packages are fast and efficient, but can appear complicated.

Software described in

Lawrence et al. (2013) PLoS Comp Bio.

Alternative: much functionality overlaps with bedtools; a popular command line tool.