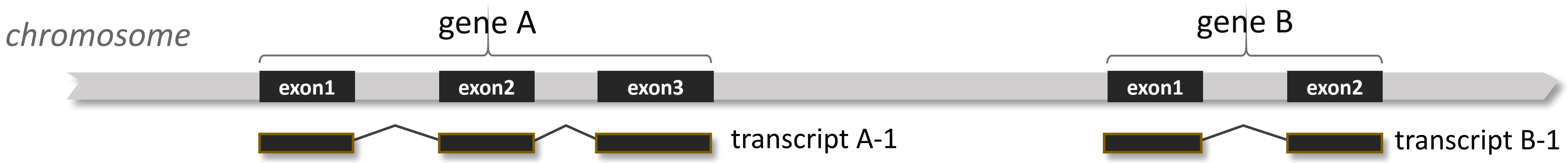
Power Ranges in R

Using GRanges for efficient handling of genomic annotation data

Geo Pertea

2023/03/31

https://github.com/gpertea/PowerRangesR



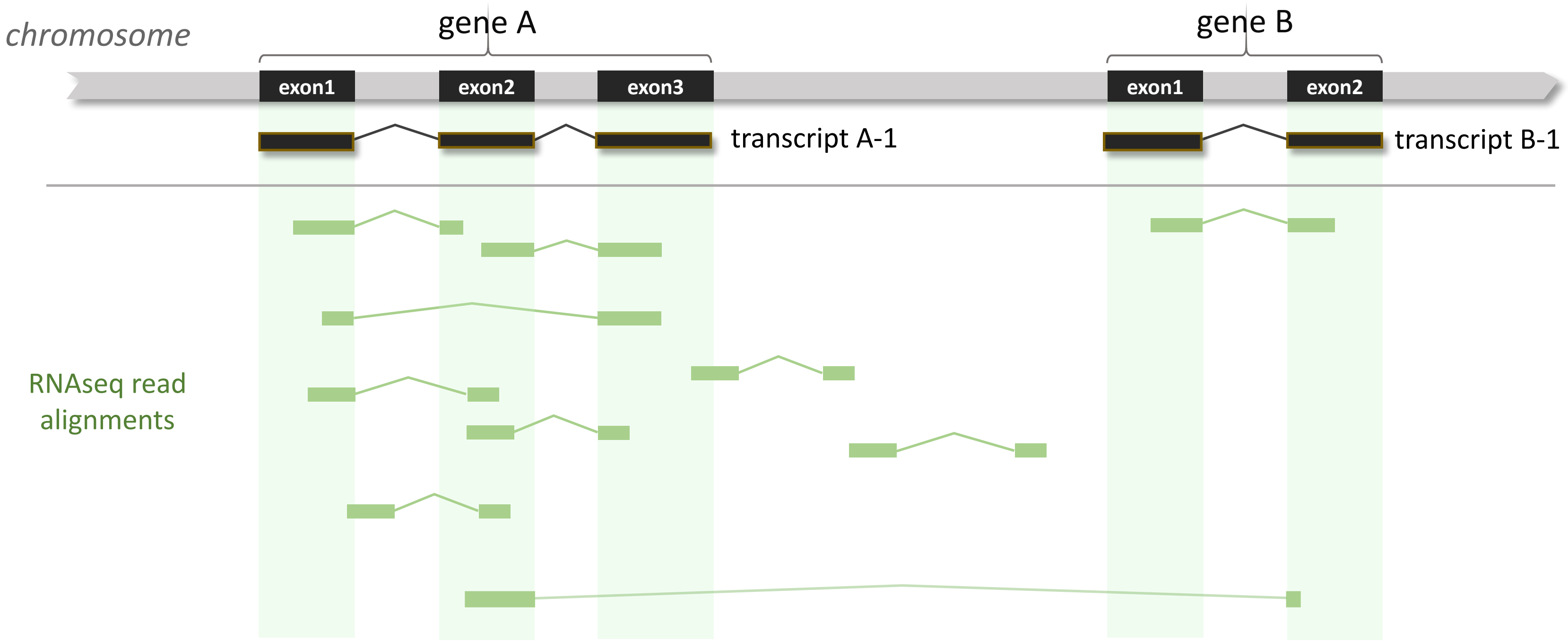
GTF (Gene Transfer Format)

track

feature type

9th column: attributes

```
chr1   HAVANA   transcript   9873504   9891995   .   +   .   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   exon         9873504   9874841   .   +   .   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   CDS          9873504   9874841   .   +   0   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   exon         9877488   9877679   .   +   .   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   CDS          9877488   9877679   .   +   0   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   exon         9888412   9888586   .   +   .   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   CDS          9888412   9888586   .   +   0   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   exon         9891475   9891998   .   +   .   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1   HAVANA   CDS          9891475   9891995   .   +   2   gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
```



We need efficient algorithms & data structures for overlap queries on large sets of genomic ranges

GTF

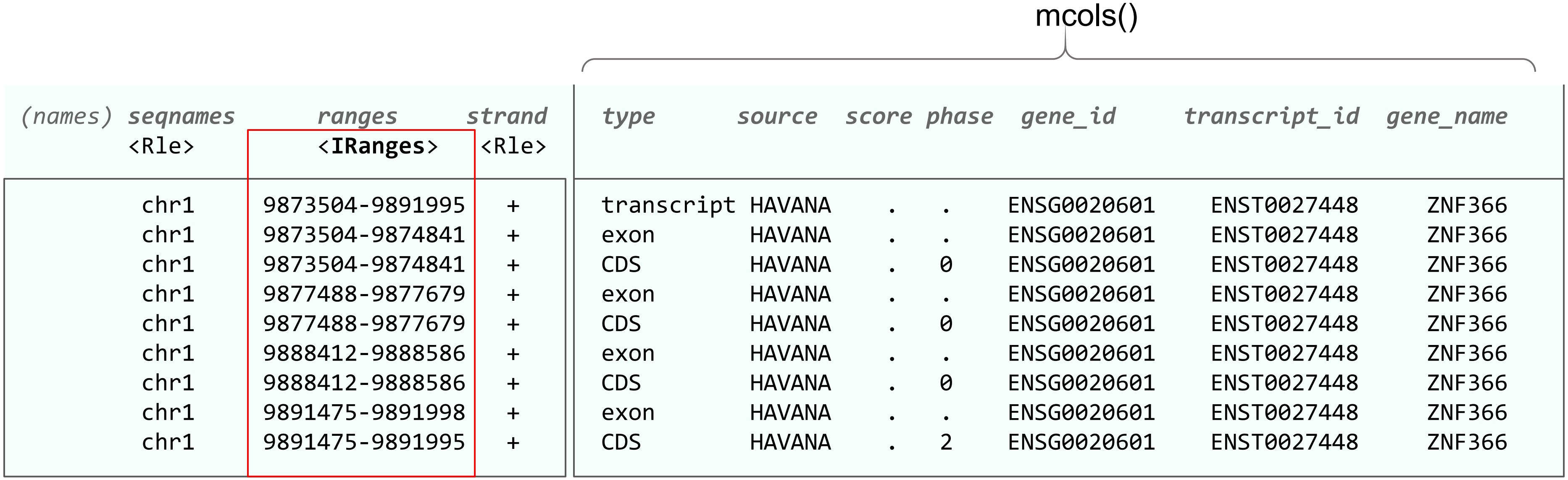
```
chr1  HAVANA  transcript  9873504  9891995  .  +  .  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  exon        9873504  9874841  .  +  .  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  CDS         9873504  9874841  .  +  0  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  exon        9877488  9877679  .  +  .  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  CDS         9877488  9877679  .  +  0  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  exon        9888412  9888586  .  +  .  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  CDS         9888412  9888586  .  +  0  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  exon        9891475  9891998  .  +  .  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
chr1  HAVANA  CDS         9891475  9891995  .  +  2  gene_id "ENSG0020601"; transcript_id "ENST0027448"; gene_name "ZNF366";
```

GRanges

metadata columns: **mcols**(GRanges)

```
(names) seqnames         ranges        strand | type       source  score phase gene_id      transcript_id  gene_name
        <Rle>            <IRanges>      <Rle> |
          chr1   9873504-9891995    +         | transcript HAVANA    .    .   ENSG0020601  ENST0027448    ZNF366
          chr1   9873504-9874841    +         | exon       HAVANA    .    .   ENSG0020601  ENST0027448    ZNF366
          chr1   9873504-9874841    +         | CDS        HAVANA    .    0   ENSG0020601  ENST0027448    ZNF366
          chr1   9877488-9877679    +         | exon       HAVANA    .    .   ENSG0020601  ENST0027448    ZNF366
          chr1   9877488-9877679    +         | CDS        HAVANA    .    0   ENSG0020601  ENST0027448    ZNF366
          chr1   9888412-9888586    +         | exon       HAVANA    .    .   ENSG0020601  ENST0027448    ZNF366
          chr1   9888412-9888586    +         | CDS        HAVANA    .    0   ENSG0020601  ENST0027448    ZNF366
          chr1   9891475-9891998    +         | exon       HAVANA    .    .   ENSG0020601  ENST0027448    ZNF366
          chr1   9891475-9891995    +         | CDS        HAVANA    .    2   ENSG0020601  ENST0027448    ZNF366
```

GRanges



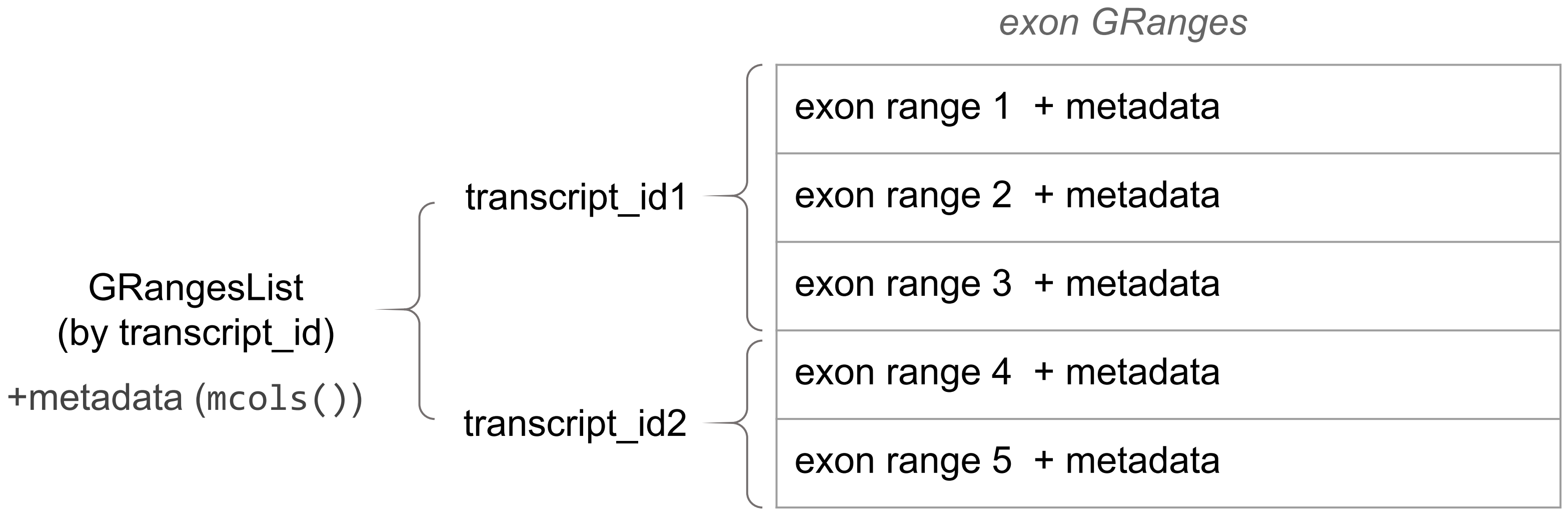
mcols()

```
(names) seqnames              ranges  strand | type       source   score phase   gene_id       transcript_id   gene_name
           <Rle>           <IRanges>   <Rle> |
            chr1   9873504-9891995      +   | transcript HAVANA      .     .    ENSG0020601   ENST0027448       ZNF366
            chr1   9873504-9874841      +   | exon       HAVANA      .     .    ENSG0020601   ENST0027448       ZNF366
            chr1   9873504-9874841      +   | CDS        HAVANA      .     0    ENSG0020601   ENST0027448       ZNF366
            chr1   9877488-9877679      +   | exon       HAVANA      .     .    ENSG0020601   ENST0027448       ZNF366
            chr1   9877488-9877679      +   | CDS        HAVANA      .     0    ENSG0020601   ENST0027448       ZNF366
            chr1   9888412-9888586      +   | exon       HAVANA      .     .    ENSG0020601   ENST0027448       ZNF366
            chr1   9888412-9888586      +   | CDS        HAVANA      .     0    ENSG0020601   ENST0027448       ZNF366
            chr1   9891475-9891998      +   | exon       HAVANA      .     .    ENSG0020601   ENST0027448       ZNF366
            chr1   9891475-9891995      +   | CDS        HAVANA      .     2    ENSG0020601   ENST0027448       ZNF366
```

IRanges base class - implemented as a Nested Containment List (NCList)

$O(n + \log N)$ query time

GRangesList



exon GRanges

GRangesList
(by transcript_id)

+metadata (`mcols()`)

transcript_id1

exon range 1 + metadata

exon range 2 + metadata

exon range 3 + metadata

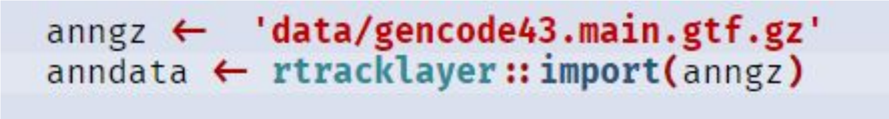
transcript_id2

exon range 4 + metadata

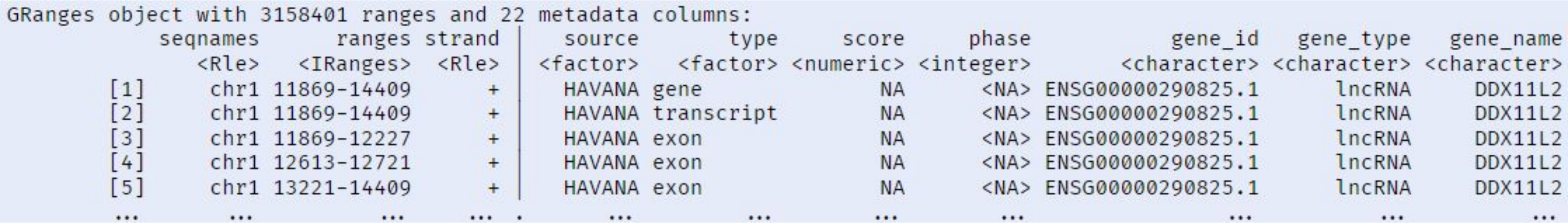
exon range 5 + metadata

```
annexgr <- anndata[anndata$type=='exon'] # exon GRanges only
## build a GRangesList of transcripts, with exons grouped by transcript_id
txgrl <- split(annexgr, annexgr$transcript_id)
```

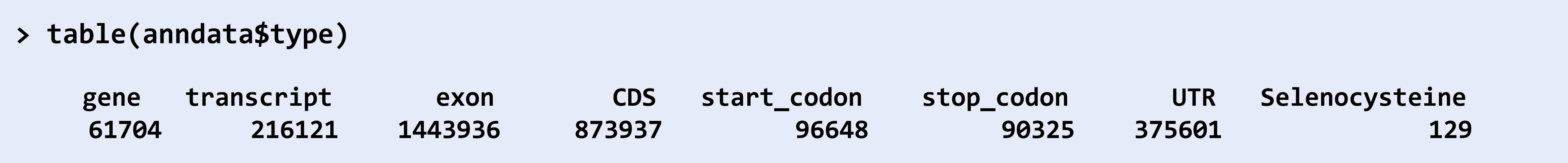
Loading a reference annotation dataset



```
anngz ←   'data/gencode43.main.gtf.gz'
anndata ← rtracklayer::import(anngz)
```

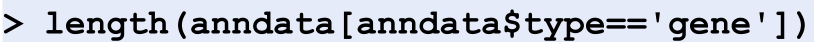


```
GRanges object with 3158401 ranges and 22 metadata columns:
            seqnames        ranges strand |   source       type     score     phase            gene_id   gene_type   gene_name
               <Rle>     <IRanges>  <Rle> | <factor>   <factor> <numeric> <integer>        <character> <character> <character>
        [1]     chr1 11869-14409      + |   HAVANA gene              NA      <NA> ENSG00000290825.1      lncRNA     DDX11L2
        [2]     chr1 11869-14409      + |   HAVANA transcript        NA      <NA> ENSG00000290825.1      lncRNA     DDX11L2
        [3]     chr1 11869-12227      + |   HAVANA exon              NA      <NA> ENSG00000290825.1      lncRNA     DDX11L2
        [4]     chr1 12613-12721      + |   HAVANA exon              NA      <NA> ENSG00000290825.1      lncRNA     DDX11L2
        [5]     chr1 13221-14409      + |   HAVANA exon              NA      <NA> ENSG00000290825.1      lncRNA     DDX11L2
        ...      ...           ...    ... .      ...        ...       ...       ...                ...         ...         ...
```

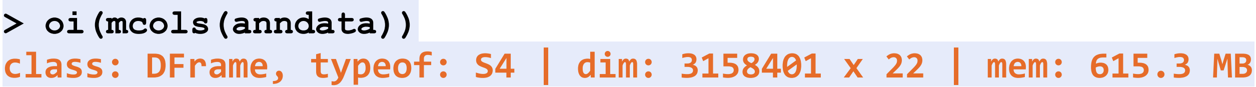


```
> table(anndata$type)

    gene   transcript        exon         CDS   start_codon    stop_codon       UTR  Selenocysteine
   61704       216121     1443936      873937         96648         90325    375601             129
```



```
> length(anndata[anndata$type=='gene'])
```



```
> oi(mcols(anndata))
class: DFrame, typeof: S4 | dim: 3158401 x 22 | mem: 615.3 MB
```

findOverlaps()

ovls <- **findOverlaps**(qryGr, subjGr, type = c("any", "start", "end", "within", "equal"), ...)

```
Hits object with 8 hits and 0 metadata columns:
      queryHits subjectHits
      <integer>   <integer>
  [1]         1        4295
  [2]         1      103503
  [3]         3      106387
  [4]         7      107678
  [5]        12      109652
  [6]        12      109728
  [7]        25      110692
  [8]        36      104995
```

`qryGr[queryHits(ovls)]` is paired with: **`subjGr[subjectHits(ovls)]`**

`qryGr[queryHits(ovls)]$gene_name <- subjGr[subjectHits(ovls)]$gene_name`

Example use cases

- getting the exonic length of all transcripts and genes in Gencode
- getting the set of known introns in Gencode
- I have a set of (novel) genomic features (introns, transcripts), what genes it overlaps in Gencode v43 and how?

- writing a FASTA file with the intergenic sequences (e.g. for Salmon decoy)
- writing a transcriptome FASTA file (transcript sequences)