Data Science Using R – Class 3 / 3

Arvind R. Subramaniam

Assistant Member

Basic Sciences Division and Computational Biology Program

Fred Hutchinson Cancer Research Center

What we learned in the last two classes

`Tidyverse` functions for working with tabular data

Import	Visualize	Transform
`read_tsv`	`geom_point`	`select`
	`geom_line`	`filter`
	`facet_grid`	`arrange`
		`mutate`
		`left_join`, `inner_join`
		`summarize`
		`group_by`

- Concept of `Tidy` data

What we will learn today

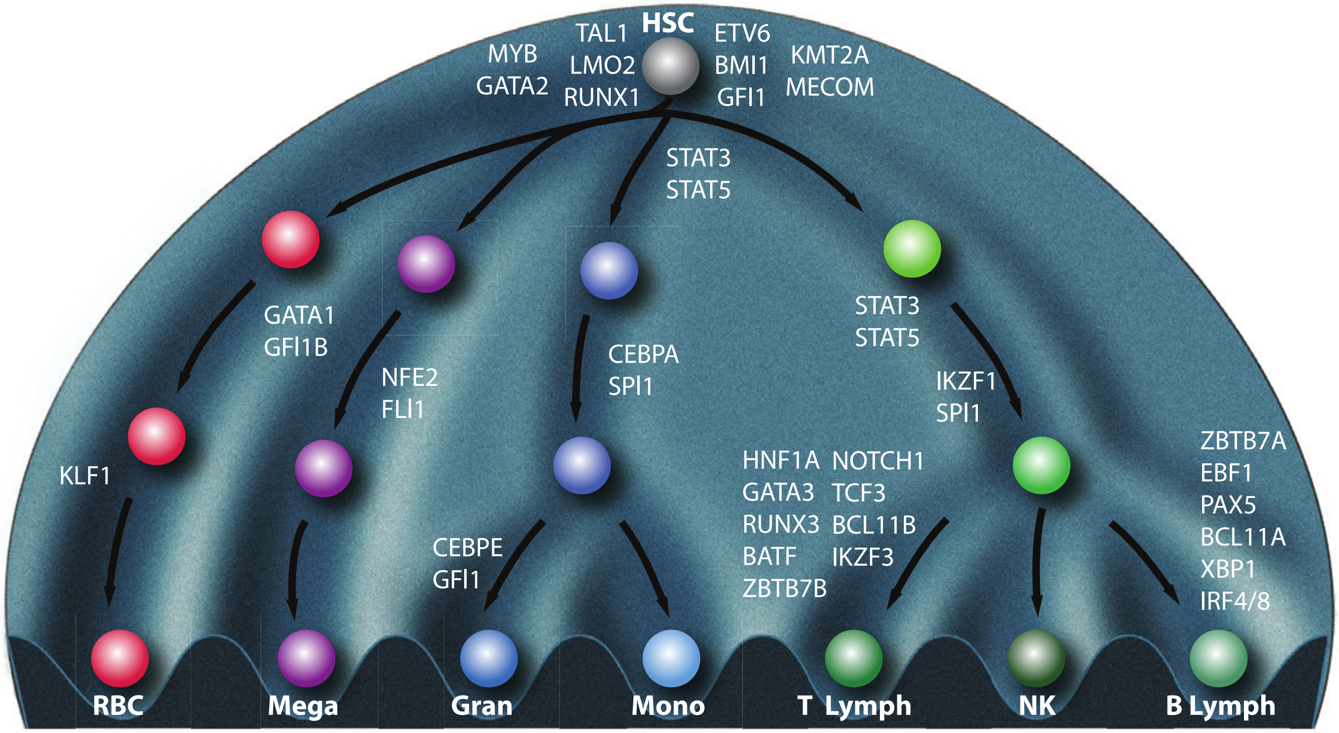
`Bioconductor` functions for working with genomic data.

Package	**Use**
`GenomicRanges`	Manipulating data along genome
`rtracklayer`	Reading and writing annotations along genome
`GenomicAlignments`	Reading and writing short sequences aligned to genome
`Biostrings`	Manipulating DNA sequences

How to install Bioconductor packages?

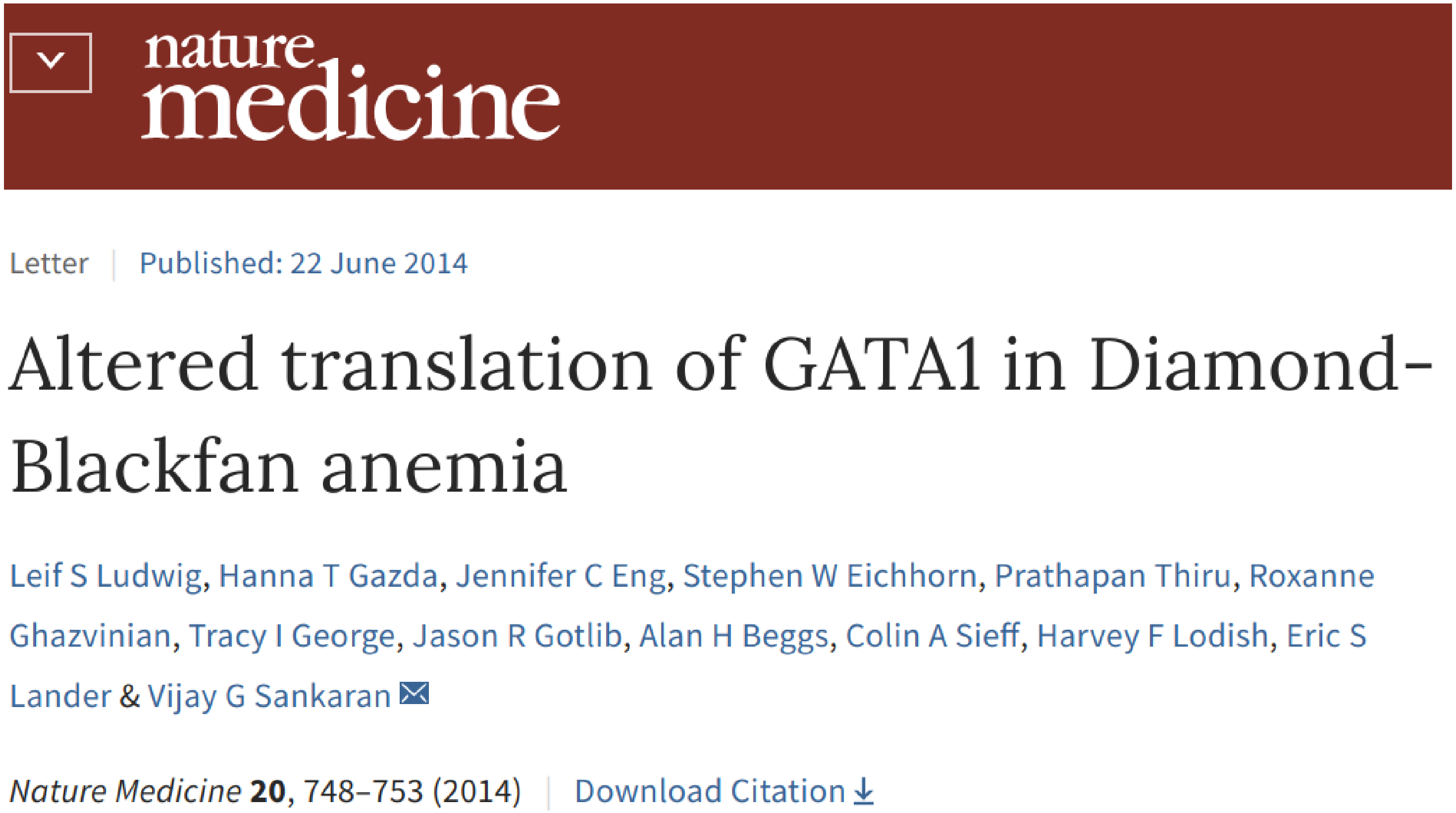
```
source("https://bioconductor.org/biocLite.R")

biocLite("GenomicRanges")

biocLite("plyranges")

biocLite("rtracklayer")

biocLite("GenomicAlignments")

biocLite("GenomicFeatures")

biocLite("BSgenome.Hsapiens.UCSC.hg19")
```

Goal: Identify 5′UTR sequence of *GATA1*



Khajuria *et al.* Cell 2018

Translation of *GATA1* mRNA is dysregulated in Diamond-Blackfan Anemia



nature medicine

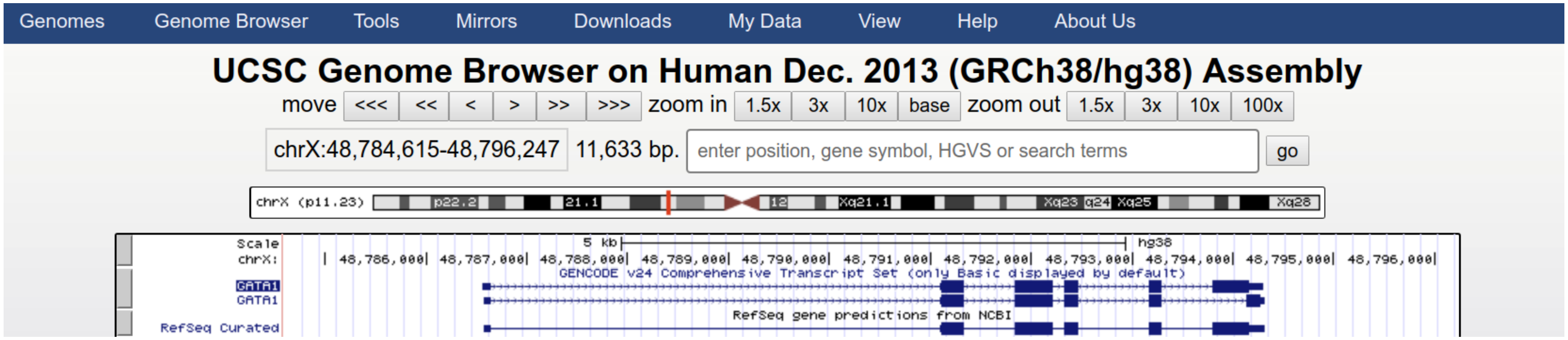
Letter | Published: 22 June 2014

Altered translation of GATA1 in Diamond-Blackfan anemia

Leif S Ludwig, Hanna T Gazda, Jennifer C Eng, Stephen W Eichhorn, Prathapan Thiru, Roxanne Ghazvinian, Tracy I George, Jason R Gotlib, Alan H Beggs, Colin A Sieff, Harvey F Lodish, Eric S Lander & Vijay G Sankaran ✉

Nature Medicine **20**, 748–753 (2014) | Download Citation ⤓

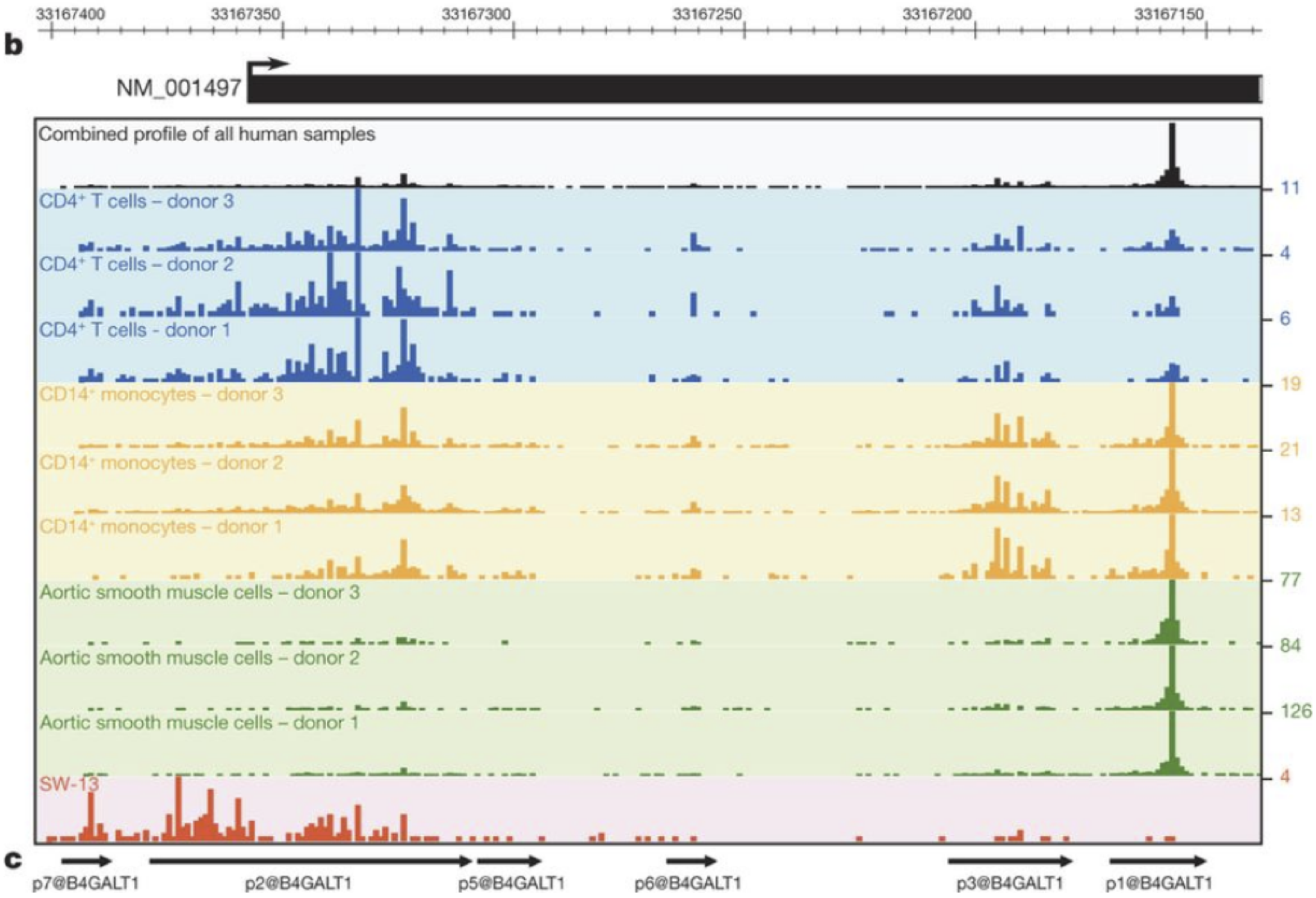
Consensus 5′UTR of *GATA1* mRNA



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

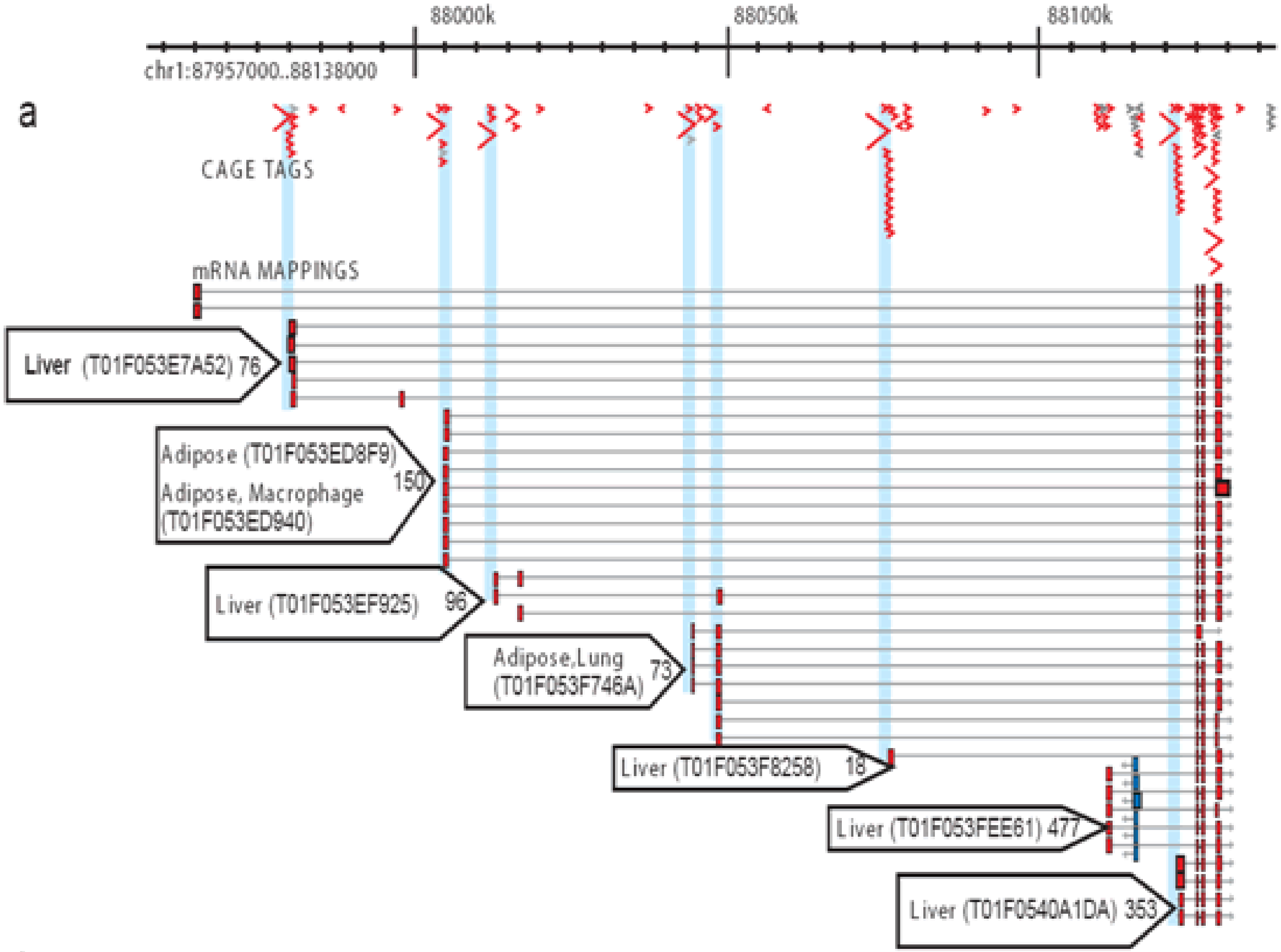
chrX:48,784,615-48,796,247 11,633 bp.

5′ UTRs tend to be cell-type specific



FANTOM5 Project Nature 2014

DeepCAGE – Cap Analysis of Gene Expression



Genomic data is often in standardized and processed formats

We want to perform experiments in K562 cells.

FANTOM5 processed data is available at http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.cell_line.hCAGE/.

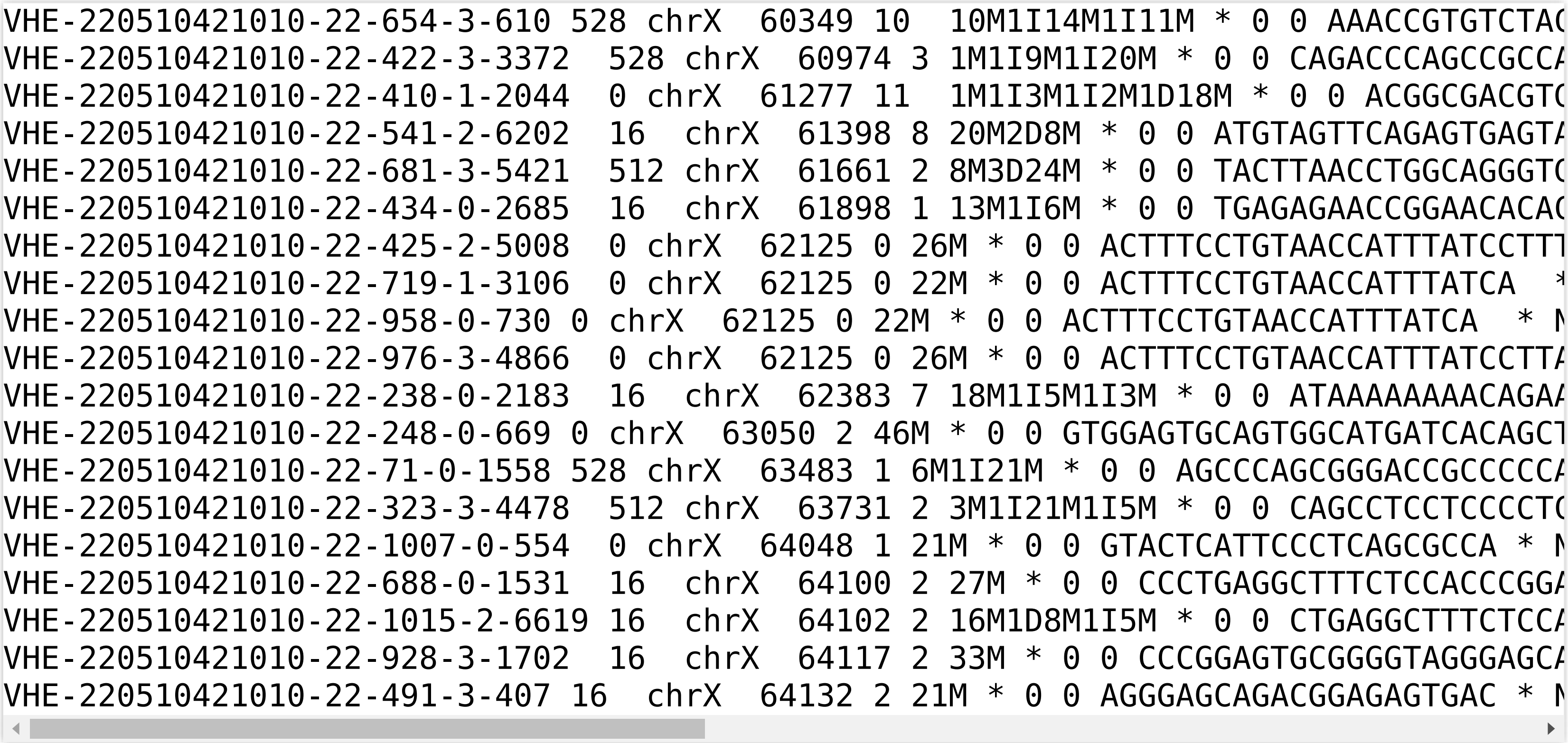
Annotation tables are available at http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/.

Several repositories to find well known datasets, eg. biomart.

Raw sequencing results are often in FASTQ format

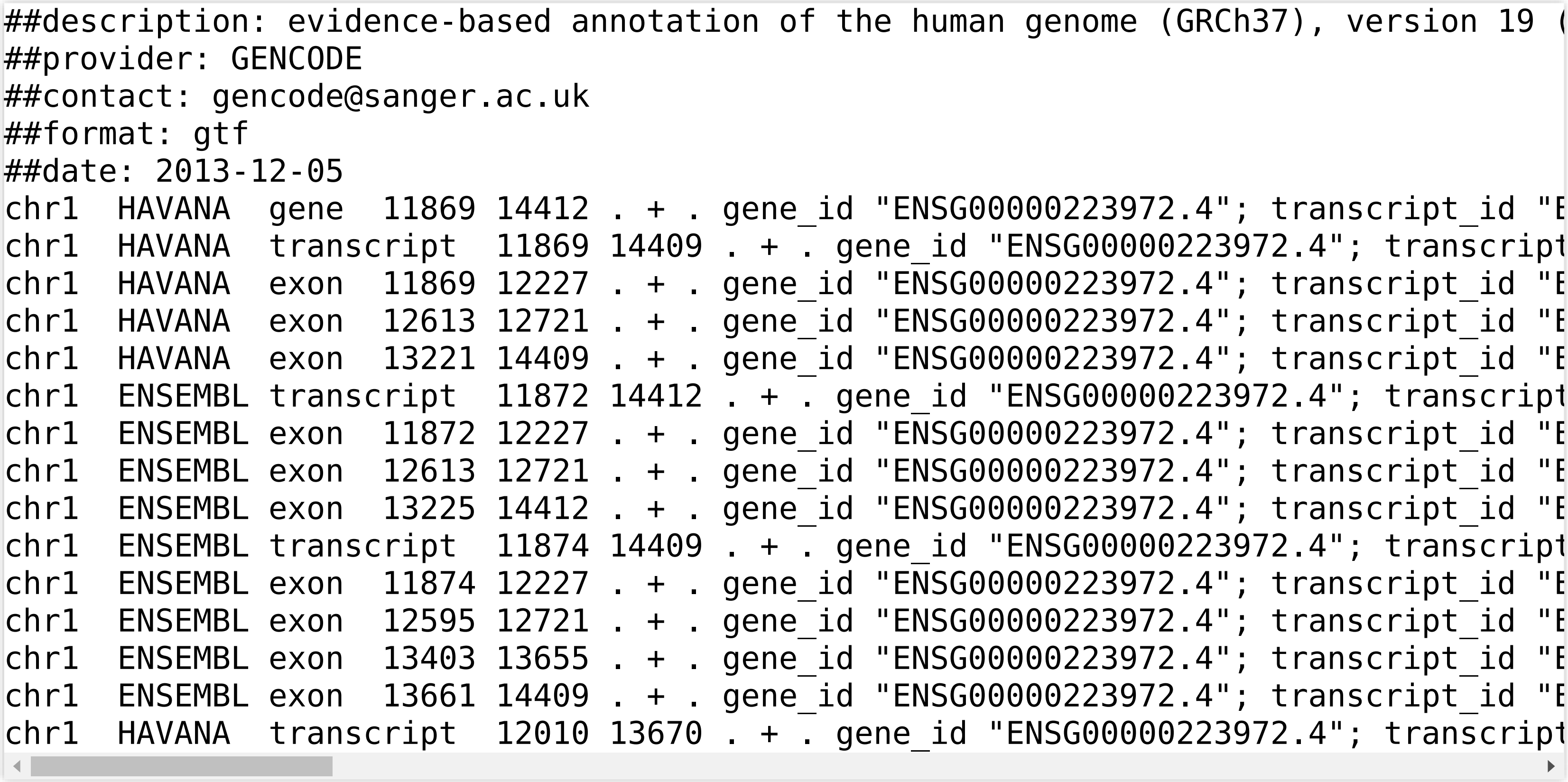
```
@K00153:11:H3FLGBBXX:1:1101:5801:1352 1:N:0:TGACCA
NATTCCAGGGGGCATGCCTGTTTGAGCGTCATTTCTGTAGGCACCATCAA
+
#<<,FFAKKFKFKKFFFKAAKKKKKAFKKKFF<KKKKKKKKKKKKKKKKK
@K00153:11:H3FLGBBXX:1:1101:5922:1352 1:N:0:TGACCA
NTGTGGCGTCGCTGAACCATAGCAGGCTCGCTGTAGGCACCATCAATATC
+
#<A<FAF<KFFF7AFAKKKFKKKKKAAK,(,AKKKKKKKKKFKK7FFA,A
@K00153:11:H3FLGBBXX:1:1101:6146:1352 1:N:0:TGACCA
NTTTCCACGTTCTAGCATTCAAGGTCCCACTGTAGGCACCATCAATATCT
+
#AAFAFKKKKKKKFKKKKKKKKKKFKKKKKKKKKKFKKKF7FKKKKKAAK
@K00153:11:H3FLGBBXX:1:1101:4836:1369 1:N:0:TGACCA
AAGAGGTGCACAATCGACCGATCCTGACTGTAGGCACCATCAATATCTCG
+
AAAFFAAFKAFFF<KA<FKKK<KKFAKA<7,FF<FFKAKKK,<K777,AF
@K00153:11:H3FLGBBXX:1:1101:4918:1369 1:N:0:TGACCA
ACATACATACACAATGGTCGCTCAAGTTCAACTGTAGGCACCATCAATAT
+
A,AFFKKKKKKFFKFAAKAA(,AKF,AF<FFKKFFKKKKKF<AKAAKKKK
```

Alignments are often in `SAM / BAM` format



```
VHE-220510421010-22-654-3-610 528 chrX  60349 10  10M1I14M1I11M * 0 0 AAACCGTGTCTAC
VHE-220510421010-22-422-3-3372  528 chrX  60974 3 1M1I9M1I20M * 0 0 CAGACCCAGCCGCCA
VHE-220510421010-22-410-1-2044  0 chrX  61277 11  1M1I3M1I2M1D18M * 0 0 ACGGCGACGTC
VHE-220510421010-22-541-2-6202  16  chrX  61398 8 20M2D8M * 0 0 ATGTAGTTCAGAGTGAGTA
VHE-220510421010-22-681-3-5421  512 chrX  61661 2 8M3D24M * 0 0 TACTTAACCTGGCAGGGTC
VHE-220510421010-22-434-0-2685  16  chrX  61898 1 13M1I6M * 0 0 TGAGAGAACCGGAACACAC
VHE-220510421010-22-425-2-5008  0 chrX  62125 0 26M * 0 0 ACTTTCCTGTAACCATTTATCCTTT
VHE-220510421010-22-719-1-3106  0 chrX  62125 0 22M * 0 0 ACTTTCCTGTAACCATTTATCA  *
VHE-220510421010-22-958-0-730 0 chrX  62125 0 22M * 0 0 ACTTTCCTGTAACCATTTATCA  * N
VHE-220510421010-22-976-3-4866  0 chrX  62125 0 26M * 0 0 ACTTTCCTGTAACCATTTATCCTTA
VHE-220510421010-22-238-0-2183  16  chrX  62383 7 18M1I5M1I3M * 0 0 ATAAAAAAAACAGAA
VHE-220510421010-22-248-0-669 0 chrX  63050 2 46M * 0 0 GTGGAGTGCAGTGGCATGATCACAGCT
VHE-220510421010-22-71-0-1558 528 chrX  63483 1 6M1I21M * 0 0 AGCCCAGCGGGACCGCCCCCA
VHE-220510421010-22-323-3-4478  512 chrX  63731 2 3M1I21M1I5M * 0 0 CAGCCTCCTCCCCTC
VHE-220510421010-22-1007-0-554  0 chrX  64048 1 21M * 0 0 GTACTCATTCCCTCAGCGCCA * N
VHE-220510421010-22-688-0-1531  16  chrX  64100 2 27M * 0 0 CCCTGAGGCTTTCTCCACCCGGA
VHE-220510421010-22-1015-2-6619 16  chrX  64102 2 16M1D8M1I5M * 0 0 CTGAGGCTTTCTCCA
VHE-220510421010-22-928-3-1702  16  chrX  64117 2 33M * 0 0 CCCGGAGTGCGGGGTAGGGAGCA
VHE-220510421010-22-491-3-407 16  chrX  64132 2 21M * 0 0 AGGGAGCAGACGGAGAGTGAC * N
```

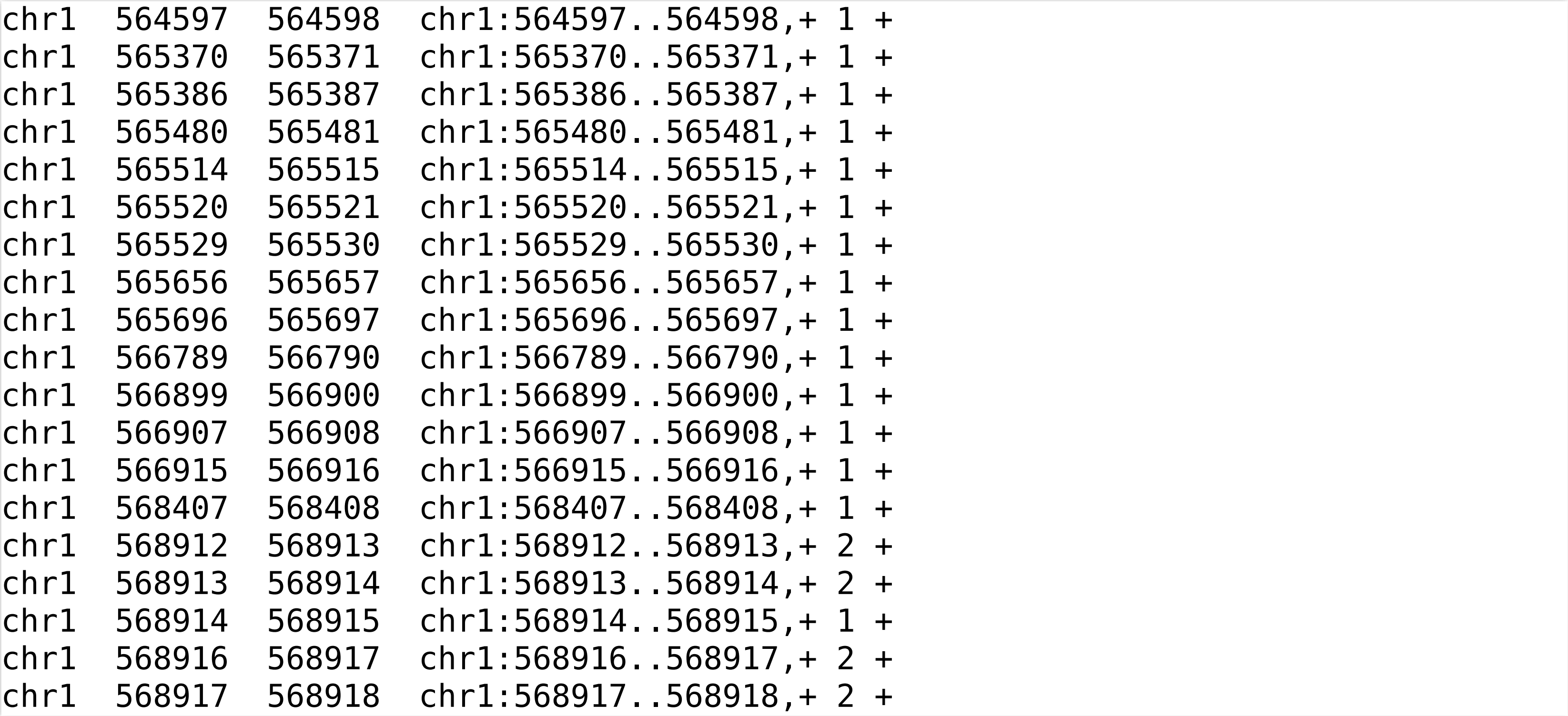
Annotations are often stored in GTF or BED format



```
##description: evidence-based annotation of the human genome (GRCh37), version 19 (
##provider: GENCODE
##contact: gencode@sanger.ac.uk
##format: gtf
##date: 2013-12-05
chr1  HAVANA  gene  11869 14412 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  HAVANA  transcript  11869 14409 . + . gene_id "ENSG00000223972.4"; transcript
chr1  HAVANA  exon  11869 12227 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  HAVANA  exon  12613 12721 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  HAVANA  exon  13221 14409 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL transcript  11872 14412 . + . gene_id "ENSG00000223972.4"; transcript
chr1  ENSEMBL exon  11872 12227 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL exon  12613 12721 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL exon  13225 14412 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL transcript  11874 14409 . + . gene_id "ENSG00000223972.4"; transcript
chr1  ENSEMBL exon  11874 12227 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL exon  12595 12721 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL exon  13403 13655 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  ENSEMBL exon  13661 14409 . + . gene_id "ENSG00000223972.4"; transcript_id "E
chr1  HAVANA  transcript  12010 13670 . + . gene_id "ENSG00000223972.4"; transcript
```

Gencode provides up-to-date gene annotations for human genes.

Annotations are often stored in GTF or BED format



```
chr1  564597  564598  chr1:564597..564598,+ 1 +
chr1  565370  565371  chr1:565370..565371,+ 1 +
chr1  565386  565387  chr1:565386..565387,+ 1 +
chr1  565480  565481  chr1:565480..565481,+ 1 +
chr1  565514  565515  chr1:565514..565515,+ 1 +
chr1  565520  565521  chr1:565520..565521,+ 1 +
chr1  565529  565530  chr1:565529..565530,+ 1 +
chr1  565656  565657  chr1:565656..565657,+ 1 +
chr1  565696  565697  chr1:565696..565697,+ 1 +
chr1  566789  566790  chr1:566789..566790,+ 1 +
chr1  566899  566900  chr1:566899..566900,+ 1 +
chr1  566907  566908  chr1:566907..566908,+ 1 +
chr1  566915  566916  chr1:566915..566916,+ 1 +
chr1  568407  568408  chr1:568407..568408,+ 1 +
chr1  568912  568913  chr1:568912..568913,+ 2 +
chr1  568913  568914  chr1:568913..568914,+ 2 +
chr1  568914  568915  chr1:568914..568915,+ 1 +
chr1  568916  568917  chr1:568916..568917,+ 2 +
chr1  568917  568918  chr1:568917..568918,+ 2 +
```

See UCSC File Format descriptions.

How to install Bioconductor packages?

```r
source("https://bioconductor.org/biocLite.R")

biocLite("GenomicRanges")

biocLite("plyranges")

biocLite("rtracklayer")

biocLite("GenomicAlignments")

biocLite("GenomicFeatures")

biocLite("BSgenome.Hsapiens.UCSC.hg19")
```

☞ Use rtracklayer to read in annotations

```r
# rtraclayer:: instead of using library(rtracklayer)
annotations <- rtracklayer::import.gff(
                                 "gencode.v19.chrX.gtf.gz") %>%
  print()
```

```
GRanges object with 80100 ranges and 21 metadata columns:
          seqnames              ranges strand |   source       type     score
             <Rle>           <IRanges>  <Rle> | <factor>   <factor> <numeric>
      [1]     chrX       170410-171758      + |   HAVANA       gene      <NA>
      [2]     chrX       170410-171758      + |   HAVANA transcript      <NA>
      [3]     chrX       170410-170513      + |   HAVANA       exon      <NA>
      [4]     chrX       171604-171758      + |   HAVANA       exon      <NA>
      [5]     chrX       192989-220023      + |   HAVANA       gene      <NA>
      ...      ...                 ...    ... .      ...        ...       ...
  [80096]     chrX 155257495-155257542      - |   HAVANA       exon      <NA>
  [80097]     chrX 155257025-155257109      - |   HAVANA       exon      <NA>
  [80098]     chrX 155256671-155256747      - |   HAVANA       exon      <NA>
  [80099]     chrX 155256349-155256502      - |   HAVANA       exon      <NA>
  [80100]     chrX 155255329-155256270      - |   HAVANA       exon      <NA>
              phase           gene_id     transcript_id      gene_type
          <integer>       <character>       <character>    <character>
      [1]      <NA> ENSG00000228572.2 ENSG00000228572.2     pseudogene
      [2]      <NA> ENSG00000228572.2 ENST00000431238.2     pseudogene
      [3]      <NA> ENSG00000228572.2 ENST00000431238.2     pseudogene
      [4]      <NA> ENSG00000228572.2 ENST00000431238.2     pseudogene
      [5]      <NA> ENSG00000182378.8 ENSG00000182378.8 protein_coding
      ...       ...               ...               ...            ...
  [80096]      <NA> ENSG00000227159.3 ENST00000507418.1     pseudogene
  [80097]      <NA> ENSG00000227159.3 ENST00000507418.1     pseudogene
```

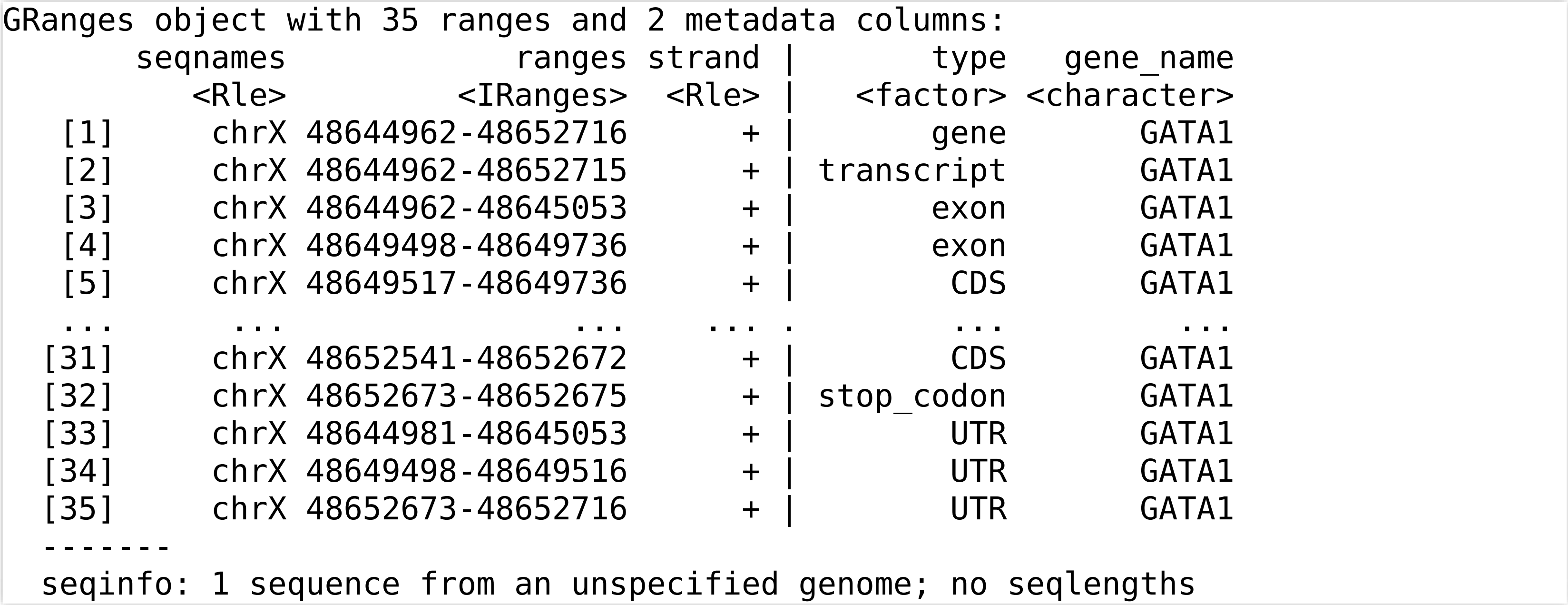
☞ plyranges : tidyverse + Bioconductor

```r
library(tidyverse)
library(plyranges)

annotations <- rtracklayer::import.gff(
                             "gencode.v19.chrX.gtf.gz") %>%
  select(type, gene_name) %>%
  print()
```

```
GRanges object with 80100 ranges and 2 metadata columns:
          seqnames              ranges strand |       type       gene_name
             <Rle>           <IRanges>  <Rle> |   <factor>     <character>
      [1]     chrX       170410-171758      + |       gene LL0YNC03-29C1.1
      [2]     chrX       170410-171758      + | transcript LL0YNC03-29C1.1
      [3]     chrX       170410-170513      + |       exon LL0YNC03-29C1.1
      [4]     chrX       171604-171758      + |       exon LL0YNC03-29C1.1
      [5]     chrX       192989-220023      + |       gene          PLCXD1
      ...      ...                 ...    ... .        ...             ...
  [80096]     chrX 155257495-155257542      - |       exon        DDX11L16
  [80097]     chrX 155257025-155257109      - |       exon        DDX11L16
  [80098]     chrX 155256671-155256747      - |       exon        DDX11L16
  [80099]     chrX 155256349-155256502      - |       exon        DDX11L16
  [80100]     chrX 155255329-155256270      - |       exon        DDX11L16
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Filtering annotations for a specific gene

```
gata1 <- rtracklayer::import.gff("gencode.v19.chrX.gtf.gz") %>%
  select(type, gene_name) %>%
  filter(gene_name == "GATA1") %>%
  print()
```



```
GRanges object with 35 ranges and 2 metadata columns:
       seqnames            ranges strand |       type   gene_name
          <Rle>         <IRanges>  <Rle> |   <factor> <character>
   [1]     chrX 48644962-48652716      + |       gene       GATA1
   [2]     chrX 48644962-48652715      + | transcript       GATA1
   [3]     chrX 48644962-48645053      + |       exon       GATA1
   [4]     chrX 48649498-48649736      + |       exon       GATA1
   [5]     chrX 48649517-48649736      + |        CDS       GATA1
   ...      ...               ...    ... .        ...         ...
  [31]     chrX 48652541-48652672      + |        CDS       GATA1
  [32]     chrX 48652673-48652675      + | stop_codon       GATA1
  [33]     chrX 48644981-48645053      + |        UTR       GATA1
  [34]     chrX 48649498-48649516      + |        UTR       GATA1
  [35]     chrX 48652673-48652716      + |        UTR       GATA1
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

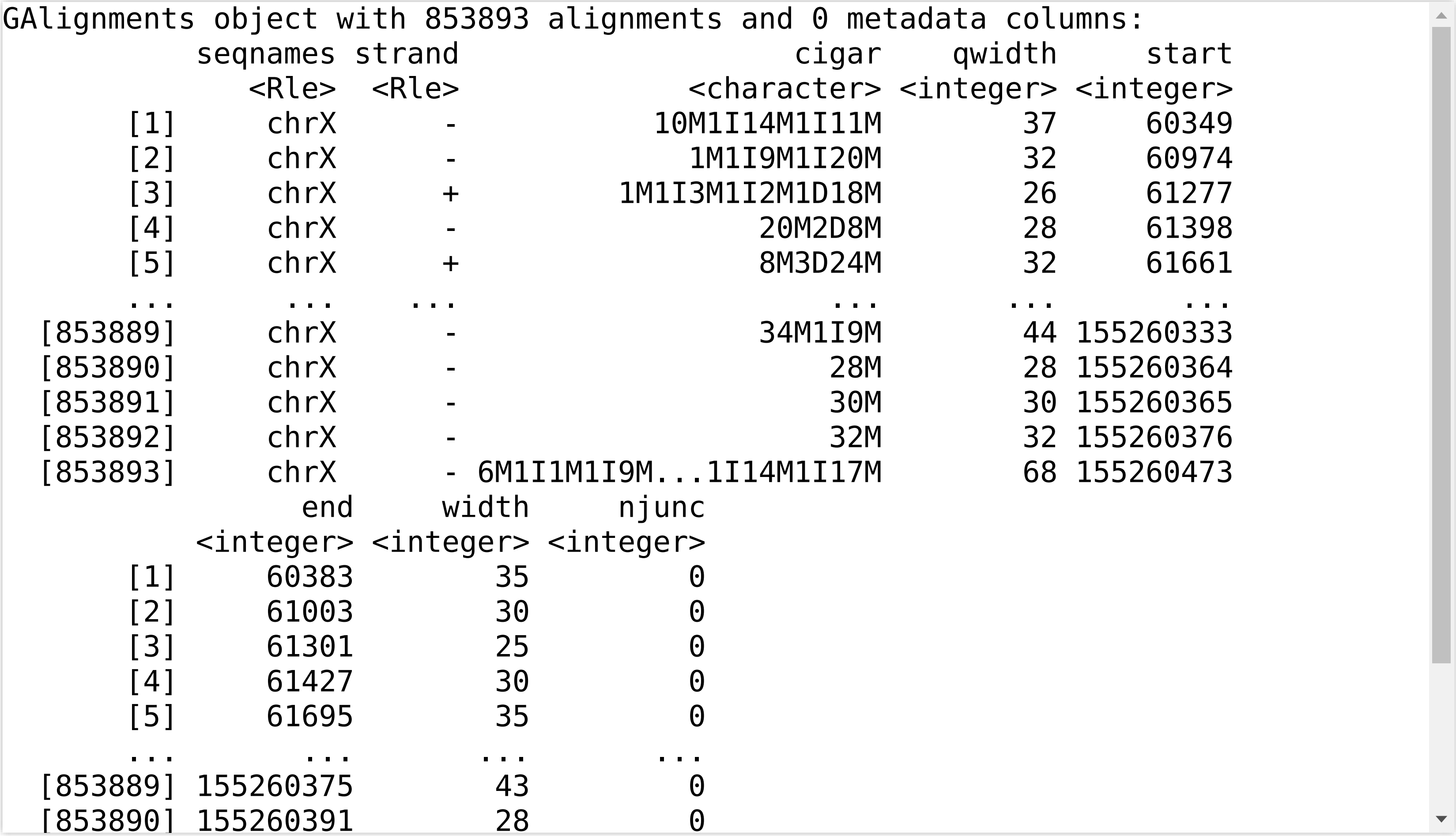
Filtering annotations for a specific gene

```r
gata1 <- rtracklayer::import.gff("gencode.v19.chrX.gtf.gz") %>%
  select(type, gene_name) %>%
  filter(gene_name == "GATA1" & type == "gene") %>%
  print()
```

```
GRanges object with 1 range and 2 metadata columns:
      seqnames            ranges strand |     type   gene_name
         <Rle>         <IRanges>  <Rle> | <factor> <character>
  [1]     chrX 48644962-48652716      + |     gene       GATA1
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

☞ Use GenomicAlignments for BAM files

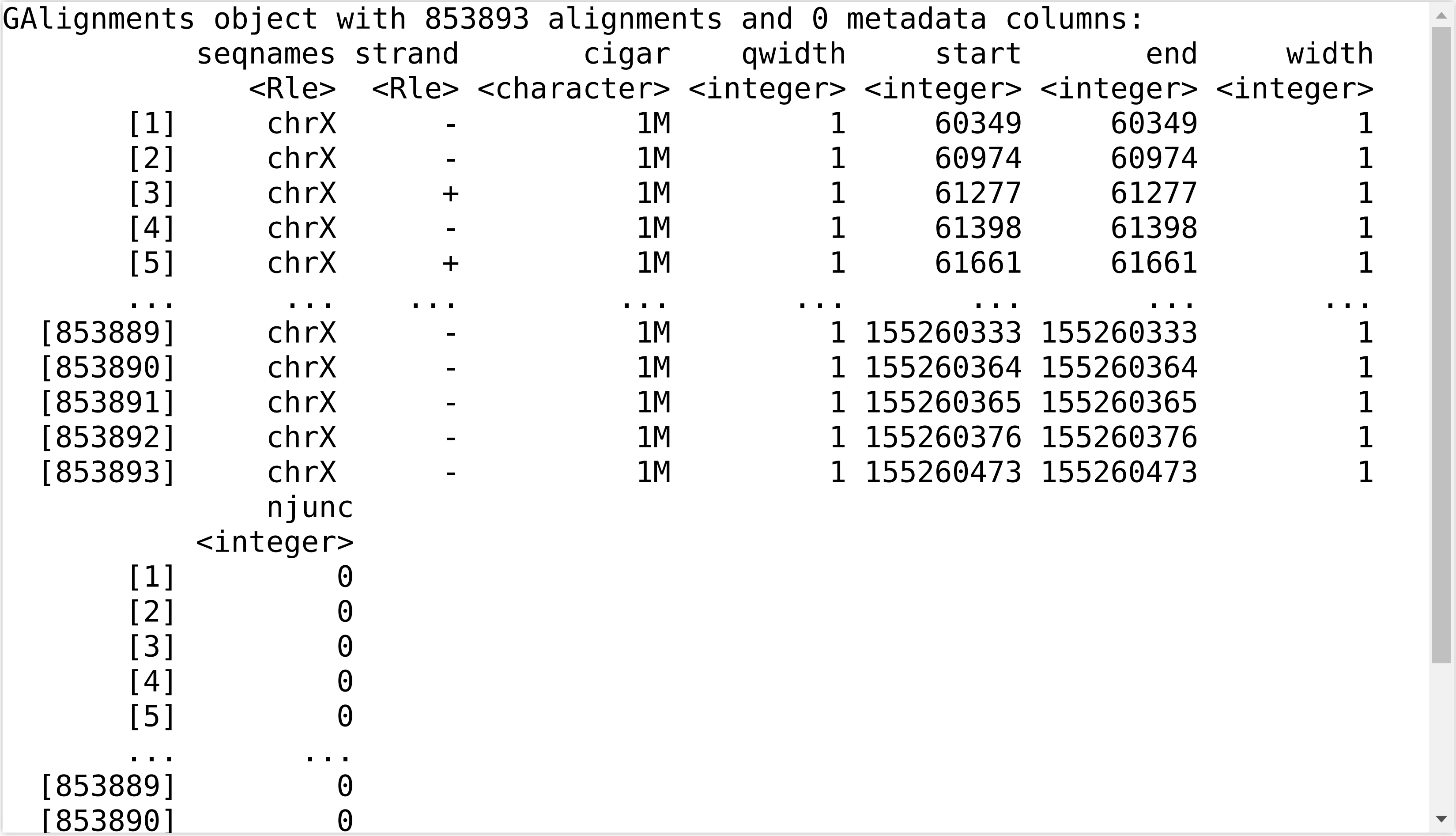
```r
library(GenomicAlignments)
aln <- readGAlignments("chrX.bam") %>%
  print()
```



```
GAlignments object with 853893 alignments and 0 metadata columns:
           seqnames strand                   cigar    qwidth     start
              <Rle>  <Rle>             <character> <integer> <integer>
       [1]     chrX      -           10M1I14M1I11M        37     60349
       [2]     chrX      -             1M1I9M1I20M        32     60974
       [3]     chrX      +         1M1I3M1I2M1D18M        26     61277
       [4]     chrX      -                 20M2D8M        28     61398
       [5]     chrX      +                 8M3D24M        32     61661
       ...      ...    ...                     ...       ...       ...
  [853889]     chrX      -                 34M1I9M        44 155260333
  [853890]     chrX      -                     28M        28 155260364
  [853891]     chrX      -                     30M        30 155260365
  [853892]     chrX      -                     32M        32 155260376
  [853893]     chrX      - 6M1I1M1I9M...1I14M1I17M        68 155260473
                 end     width     njunc
           <integer> <integer> <integer>
       [1]     60383        35         0
       [2]     61003        30         0
       [3]     61301        25         0
       [4]     61427        30         0
       [5]     61695        35         0
       ...       ...       ...       ...
  [853889] 155260375        43         0
  [853890] 155260391        28         0
```

Trim reads using `qnarrow`

```r
aln %>%
  qnarrow(start = 1, width = 1) %>%
  print()
```



```
GAlignments object with 853893 alignments and 0 metadata columns:
           seqnames strand       cigar    qwidth     start       end     width
              <Rle>  <Rle> <character> <integer> <integer> <integer> <integer>
       [1]     chrX      -          1M         1     60349     60349         1
       [2]     chrX      -          1M         1     60974     60974         1
       [3]     chrX      +          1M         1     61277     61277         1
       [4]     chrX      -          1M         1     61398     61398         1
       [5]     chrX      +          1M         1     61661     61661         1
       ...      ...    ...         ...       ...       ...       ...       ...
  [853889]     chrX      -          1M         1 155260333 155260333         1
  [853890]     chrX      -          1M         1 155260364 155260364         1
  [853891]     chrX      -          1M         1 155260365 155260365         1
  [853892]     chrX      -          1M         1 155260376 155260376         1
  [853893]     chrX      -          1M         1 155260473 155260473         1
               njunc
           <integer>
       [1]         0
       [2]         0
       [3]         0
       [4]         0
       [5]         0
       ...       ...
  [853889]         0
  [853890]         0
```

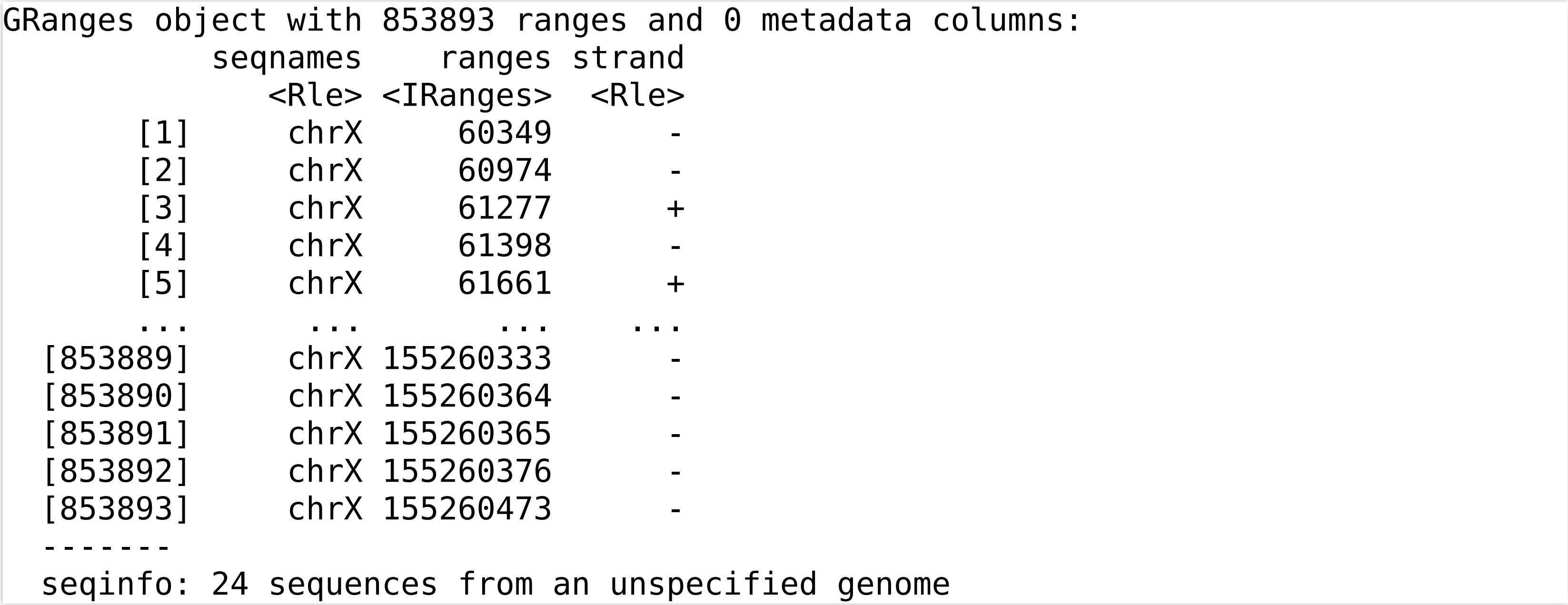
`GRanges` is a versatile structure similar to `tibble`

```r
aln %>%
  qnarrow(start = 1, width = 1) %>%
  GRanges() %>%
  print()
```

```
GRanges object with 853893 ranges and 0 metadata columns:
           seqnames    ranges strand
              <Rle> <IRanges>  <Rle>
       [1]     chrX     60349      -
       [2]     chrX     60974      -
       [3]     chrX     61277      +
       [4]     chrX     61398      -
       [5]     chrX     61661      +
       ...      ...       ...    ...
  [853889]     chrX 155260333      -
  [853890]     chrX 155260364      -
  [853891]     chrX 155260365      -
  [853892]     chrX 155260376      -
  [853893]     chrX 155260473      -
  -------
  seqinfo: 24 sequences from an unspecified genome
```

Use filter to filter for single nt width

```
aln %>%
  qnarrow(start = 1, width = 1) %>%
  GRanges() %>%
  filter(width == 1) %>%
  print()
```



```
GRanges object with 853893 ranges and 0 metadata columns:
           seqnames    ranges strand
              <Rle> <IRanges>  <Rle>
       [1]     chrX     60349      -
       [2]     chrX     60974      -
       [3]     chrX     61277      +
       [4]     chrX     61398      -
       [5]     chrX     61661      +
       ...      ...       ...    ...
  [853889]     chrX 155260333      -
  [853890]     chrX 155260364      -
  [853891]     chrX 155260365      -
  [853892]     chrX 155260376      -
  [853893]     chrX 155260473      -
  -------
  seqinfo: 24 sequences from an unspecified genome
```

overlap in Bioconductor ≡ join in tidyverse

```
gata1_aln <- aln %>%
  qnarrow(start = 1, width = 1) %>%
  GRanges() %>%
  filter(width == 1) %>%
  filter_by_overlaps(gata1) %>%
  print()
```

```
GRanges object with 1987 ranges and 0 metadata columns:
         seqnames    ranges strand
            <Rle> <IRanges>  <Rle>
     [1]     chrX  48644962      +
     [2]     chrX  48644962      +
     [3]     chrX  48644962      +
     [4]     chrX  48644962      +
     [5]     chrX  48644962      +
     ...      ...       ...    ...
  [1983]     chrX  48652480      +
  [1984]     chrX  48652486      +
  [1985]     chrX  48652635      -
  [1986]     chrX  48652636      +
  [1987]     chrX  48652661      +
  -------
  seqinfo: 24 sequences from an unspecified genome
```

`coverage` tallies up reads at each genomic position

```
gata1_aln %>%
  coverage() %>%
  print()
```

```
RleList of length 24
$chr1
integer-Rle of length 249250621 with 1 run
  Lengths: 249250621
  Values :         0

$chr2
integer-Rle of length 243199373 with 1 run
  Lengths: 243199373
  Values :         0

$chr3
integer-Rle of length 198022430 with 1 run
  Lengths: 198022430
  Values :         0

$chr4
integer-Rle of length 191154276 with 1 run
  Lengths: 191154276
  Values :         0

$chr5
integer-Rle of length 180915260 with 1 run
  Lengths: 180915260
```

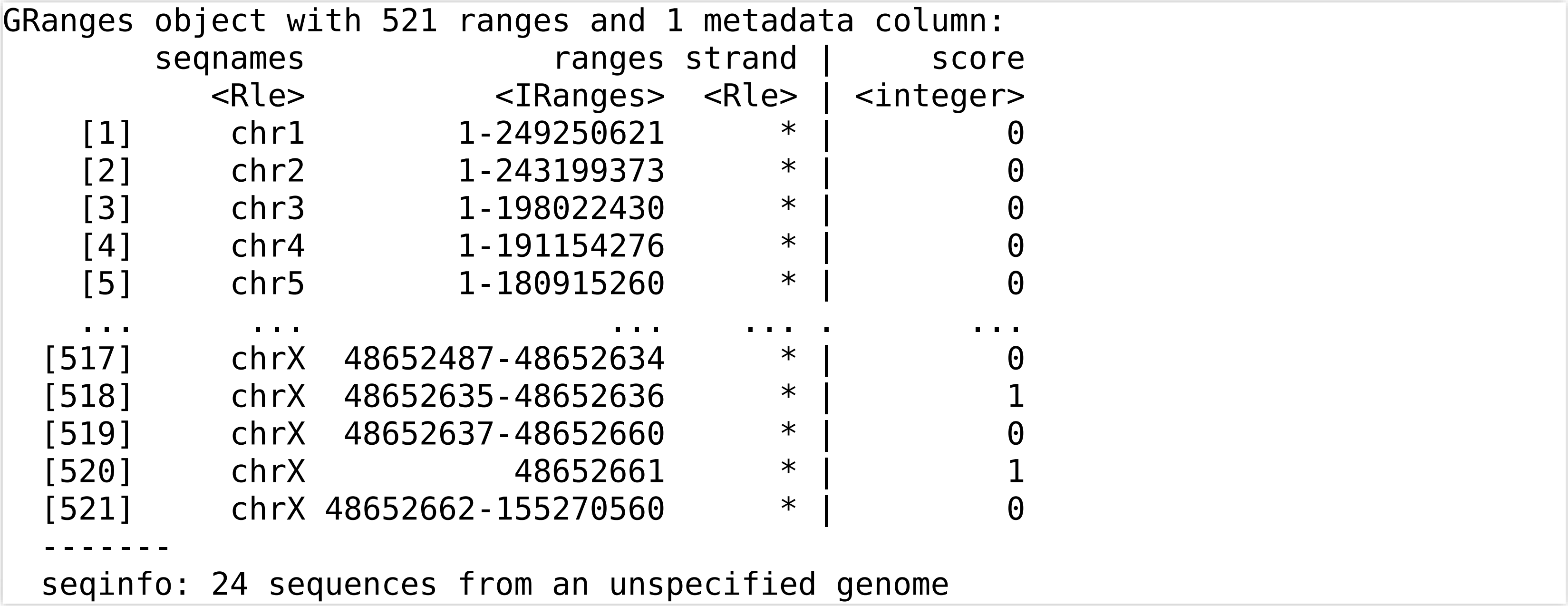
`coverage` tallies up reads at each genomic position

```r
gata1_aln %>%
  coverage() %>%
  magrittr::extract("chrX") %>%
  print()
```

```
RleList of length 1
$chrX
integer-Rle of length 155270560 with 498 runs
  Lengths:  48644961         1         1 ...        24         1 106617899
  Values :         0         5         9 ...         0         1         0
```

Most structures can be converted to GRanges

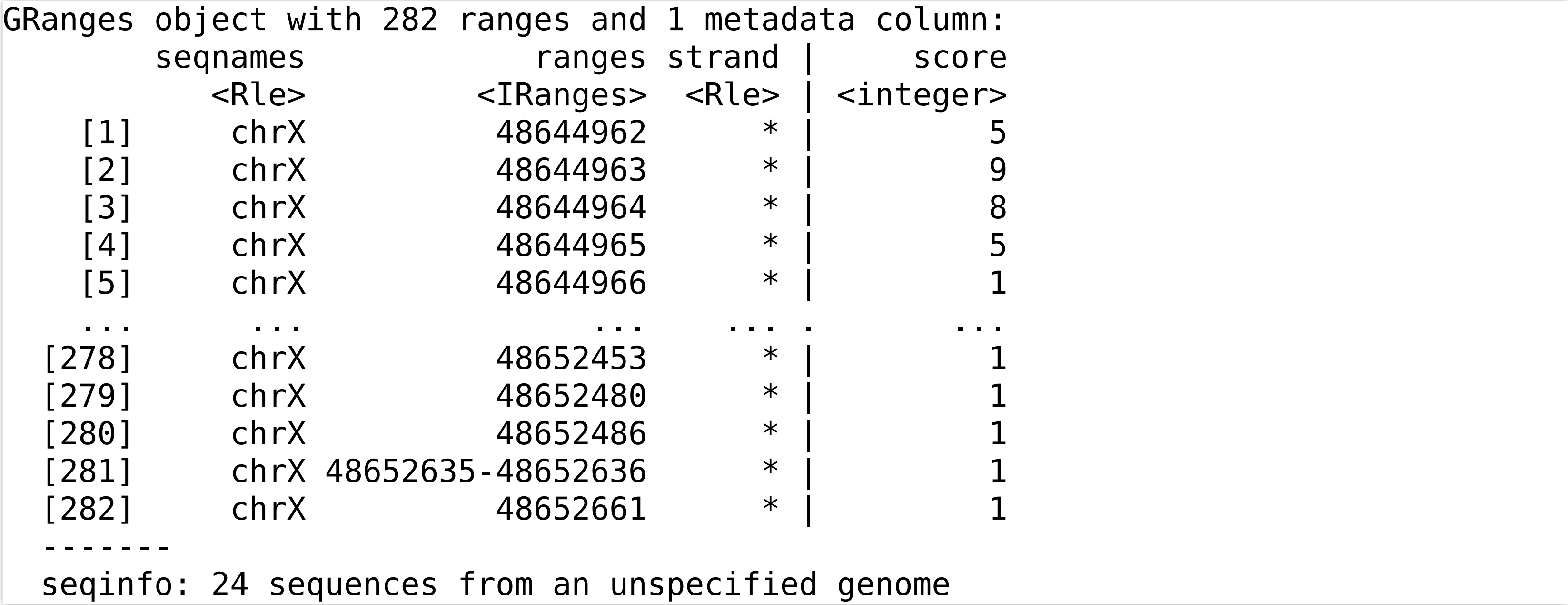
```
gata1_aln %>%
  coverage() %>%
  GRanges() %>%
  print()
```



```
GRanges object with 521 ranges and 1 metadata column:
        seqnames              ranges strand |     score
           <Rle>           <IRanges>  <Rle> | <integer>
    [1]     chr1         1-249250621      * |         0
    [2]     chr2         1-243199373      * |         0
    [3]     chr3         1-198022430      * |         0
    [4]     chr4         1-191154276      * |         0
    [5]     chr5         1-180915260      * |         0
    ...      ...                 ...    ... .       ...
  [517]     chrX   48652487-48652634      * |         0
  [518]     chrX   48652635-48652636      * |         1
  [519]     chrX   48652637-48652660      * |         0
  [520]     chrX            48652661      * |         1
  [521]     chrX  48652662-155270560      * |         0
  -------
  seqinfo: 24 sequences from an unspecified genome
```

Filter for non-zero read counts

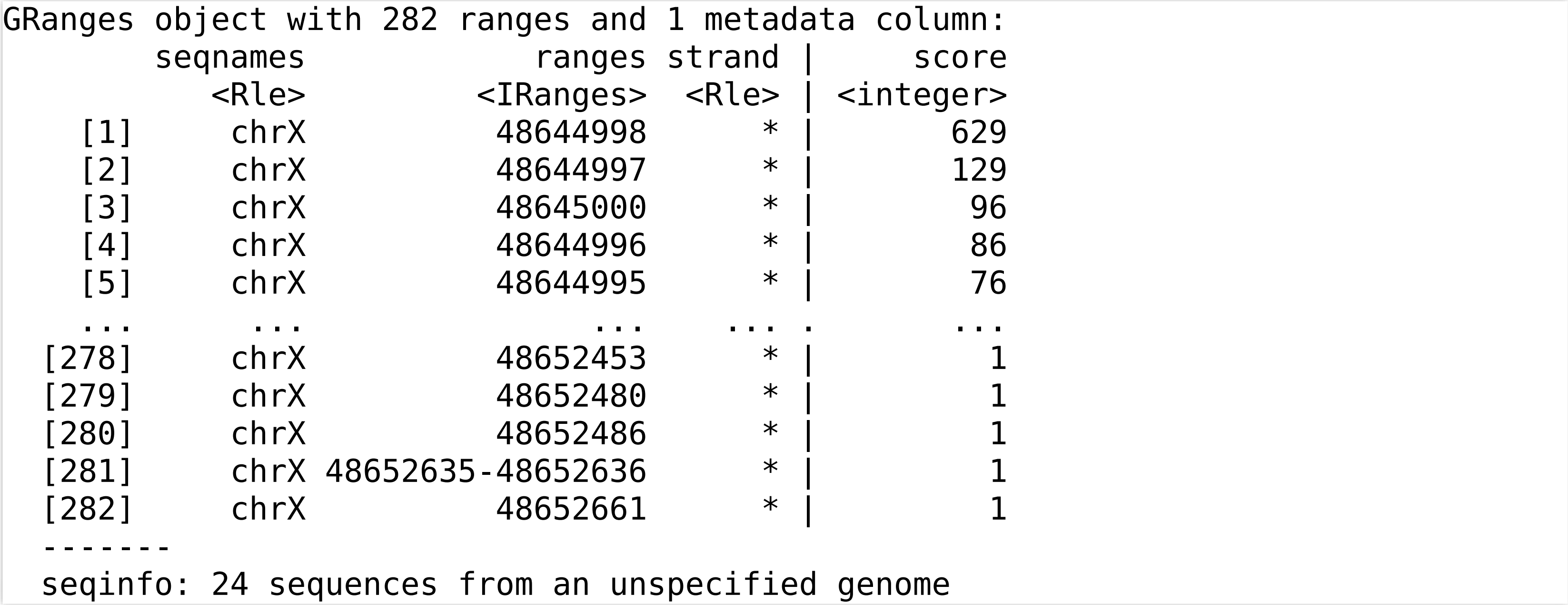
```
gata1_aln %>%
  coverage() %>%
  GRanges() %>%
  filter(score > 0) %>%
  print()
```



```
GRanges object with 282 ranges and 1 metadata column:
        seqnames              ranges strand |     score
           <Rle>           <IRanges>  <Rle> | <integer>
    [1]     chrX            48644962      * |         5
    [2]     chrX            48644963      * |         9
    [3]     chrX            48644964      * |         8
    [4]     chrX            48644965      * |         5
    [5]     chrX            48644966      * |         1
    ...      ...                 ...    ... .       ...
  [278]     chrX            48652453      * |         1
  [279]     chrX            48652480      * |         1
  [280]     chrX            48652486      * |         1
  [281]     chrX 48652635-48652636      * |         1
  [282]     chrX            48652661      * |         1
  -------
  seqinfo: 24 sequences from an unspecified genome
```

Find location with maximum counts

```
gata1counts <- gata1_aln %>%
  coverage() %>%
  GRanges() %>%
  filter(score > 0) %>%
  arrange(-score) %>%
  print()
```



```
GRanges object with 282 ranges and 1 metadata column:
        seqnames            ranges strand |     score
           <Rle>         <IRanges>  <Rle> | <integer>
    [1]     chrX          48644998      * |       629
    [2]     chrX          48644997      * |       129
    [3]     chrX          48645000      * |        96
    [4]     chrX          48644996      * |        86
    [5]     chrX          48644995      * |        76
    ...      ...               ...    ... .       ...
  [278]     chrX          48652453      * |         1
  [279]     chrX          48652480      * |         1
  [280]     chrX          48652486      * |         1
  [281]     chrX 48652635-48652636      * |         1
  [282]     chrX          48652661      * |         1
  -------
  seqinfo: 24 sequences from an unspecified genome
```

Cross-check against FANTOM5 processed counts

```r
fantom5counts <- rtracklayer::import.bed("ctss.bed.gz") %>%
  print()
```

```
GRanges object with 892902 ranges and 2 metadata columns:
           seqnames    ranges strand |                        name     score
              <Rle> <IRanges>  <Rle> |                 <character> <numeric>
       [1]     chr1    564598      + |       chr1:564597..564598,+         1
       [2]     chr1    565371      + |       chr1:565370..565371,+         1
       [3]     chr1    565387      + |       chr1:565386..565387,+         1
       [4]     chr1    565481      + |       chr1:565480..565481,+         1
       [5]     chr1    565515      + |       chr1:565514..565515,+         1
       ...      ...       ...    ... .                           ...       ...
  [892898]     chrX 155110863      - | chrX:155110862..155110863,-         1
  [892899]     chrX 155119239      - | chrX:155119238..155119239,-         1
  [892900]     chrX 155172471      - | chrX:155172470..155172471,-         1
  [892901]     chrX 155188784      - | chrX:155188783..155188784,-         1
  [892902]     chrX 155191106      - | chrX:155191105..155191106,-         1
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

GRanges enables standardized workflow

```
fantom5counts <- rtracklayer::import.bed("ctss.bed.gz") %>%
  filter_by_overlaps(gata1) %>%
  arrange(-score) %>%
  print()
```

```
GRanges object with 189 ranges and 2 metadata columns:
        seqnames    ranges strand |                   name     score
           <Rle> <IRanges>  <Rle> |            <character> <numeric>
    [1]     chrX  48644998      + | chrX:48644997..48644998,+       492
    [2]     chrX  48645000      + | chrX:48644999..48645000,+        77
    [3]     chrX  48644999      + | chrX:48644998..48644999,+        72
    [4]     chrX  48644997      + | chrX:48644996..48644997,+        71
    [5]     chrX  48644992      + | chrX:48644991..48644992,+        61
    ...      ...       ...    ... .                       ...       ...
  [185]     chrX  48651603      - | chrX:48651602..48651603,-         1
  [186]     chrX  48651807      - | chrX:48651806..48651807,-         1
  [187]     chrX  48651814      - | chrX:48651813..48651814,-         1
  [188]     chrX  48651815      - | chrX:48651814..48651815,-         1
  [189]     chrX  48652279      - | chrX:48652278..48652279,-         1
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

Compare our processing vs FANTOM5 processing

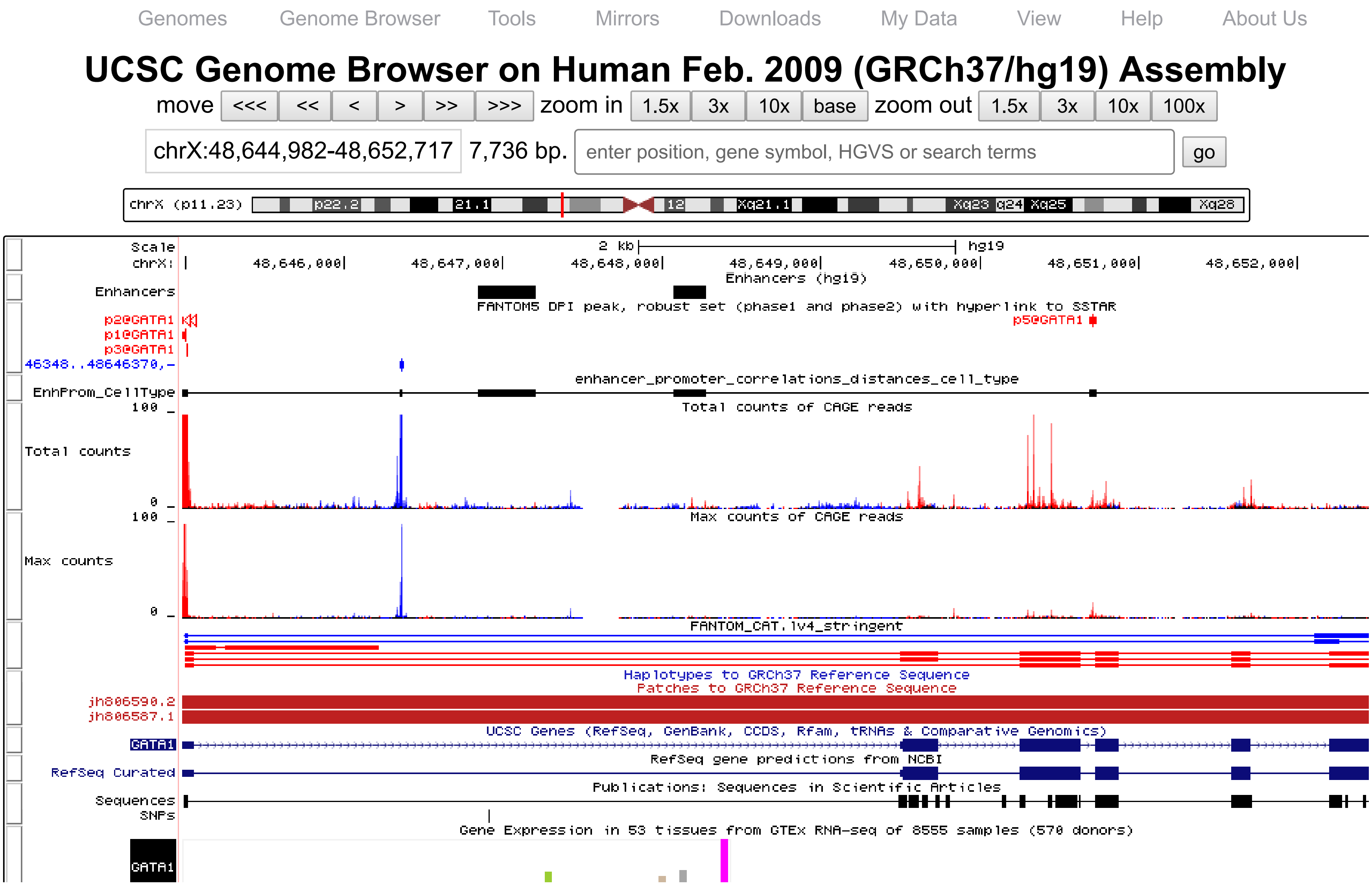
```
print(gata1counts)
```

```
GRanges object with 282 ranges and 1 metadata column:
        seqnames            ranges strand |     score
           <Rle>         <IRanges>  <Rle> | <integer>
    [1]     chrX          48644998      * |       629
    [2]     chrX          48644997      * |       129
    [3]     chrX          48645000      * |        96
    [4]     chrX          48644996      * |        86
    [5]     chrX          48644995      * |        76
    ...      ...               ...    ... .       ...
  [278]     chrX          48652453      * |         1
  [279]     chrX          48652480      * |         1
  [280]     chrX          48652486      * |         1
  [281]     chrX 48652635-48652636      * |         1
  [282]     chrX          48652661      * |         1
  -------
  seqinfo: 24 sequences from an unspecified genome
```

```
print(fantom5counts)
```

```
GRanges object with 189 ranges and 2 metadata columns:
        seqnames    ranges strand |                     name     score
           <Rle> <IRanges>  <Rle> |              <character> <numeric>
    [1]     chrX  48644998      + | chrX:48644997..48644998,+       492
    [2]     chrX  48645000      + | chrX:48644999..48645000,+        77
    [3]     chrX  48644999      + | chrX:48644998..48644999,+        72
    [4]     chrX  48644997      + | chrX:48644996..48644997,+        71
    [5]     chrX  48644992      + | chrX:48644991..48644992,+        61
    ...      ...       ...    ... .                       ...       ...
  [185]     chrX  48651603      - | chrX:48651602..48651603,-         1
```

Check against FANTOM5 identification of TSS



Genomes Genome Browser Tools Mirrors Downloads My Data View Help About Us

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chrX:48,644,982-48,652,717 7,736 bp. enter position, gene symbol, HGVS or search terms go

chrX (p11.23)
p22.2
21.1
12
Xq21.1
Xq23
q24
Xq25
Xq28
Scale
2 kb
hg19
chrX:
48,646,000|
48,647,000|
48,648,000|
48,649,000|
48,650,000|
48,651,000|
48,652,000|
Enhancers (hg19)
Enhancers
FANTOM5 DPI peak, robust set (phase1 and phase2) with hyperlink to SSTAR
p2@GATA1
p1@GATA1
p3@GATA1
p5@GATA1
46348..48646370,-
enhancer_promoter_correlations_distances_cell_type
EnhProm_CellType
Total counts of CAGE reads
100 _
Total counts
0 _
100 _
Max counts of CAGE reads
Max counts
0 _
FANTOM_CAT.lv4_stringent
Haplotypes to GRCh37 Reference Sequence
Patches to GRCh37 Reference Sequence
jh806590.2
jh806587.1
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
GATA1
RefSeq gene predictions from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
GATA1

Do it yourself using `tidyverse`

1. Go to http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/.

2. Download TSS coordinates `hg19.cage_peak_phase1and2combined_coord.bed.gz`

3. Download TSS annotations `hg19.cage_peak_phase1and2combined_ann.txt.gz`

4. Join the two tables and identify the `p1` peak for *GATA1*

How do we extract 5′UTR sequence?

1. Get the new beginning of the 5′UTR.

2. Get the end of the 5′UTR.

3. Account for transcript splicing.

4. Get genome → transcript → 5′UTR sequence.

Get the new beginning of the 5′UTR

```
gata1_5utr_start <- rtracklayer::import.bed("ctss.bed.gz") %>%
  filter_by_overlaps(gata1) %>%
  arrange(-score) %>%
  filter(score == max(score)) %>%
  print()
```

```
GRanges object with 1 range and 2 metadata columns:
      seqnames    ranges strand |                   name     score
         <Rle> <IRanges>  <Rle> |            <character> <numeric>
  [1]     chrX  48644998      + | chrX:48644997..48644998,+       492
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

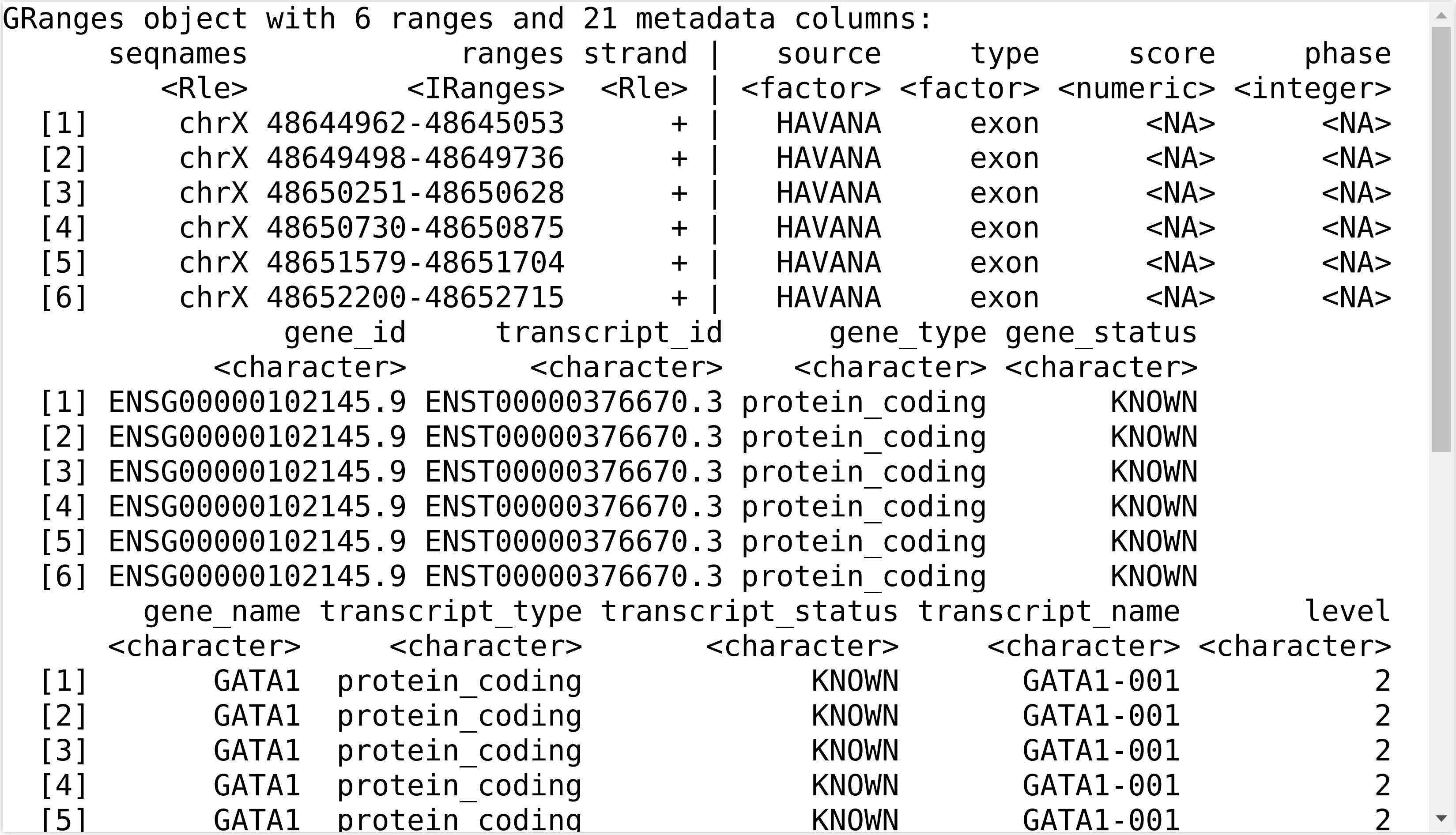
Get the end of the 5’UTR

```
gata1_5utr_end <- rtracklayer::import.gff(
                                "gencode.v19.chrX.gtf.gz") %>%
  filter(gene_name == "GATA1" &
         type == "start_codon" & transcript_status == "KNOWN") %>%
  select(-everything()) %>%
  print()
```

```
GRanges object with 1 range and 0 metadata columns:
      seqnames            ranges strand
         <Rle>         <IRanges>  <Rle>
  [1]     chrX 48649517-48649519      +
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Account for splicing by using only exons

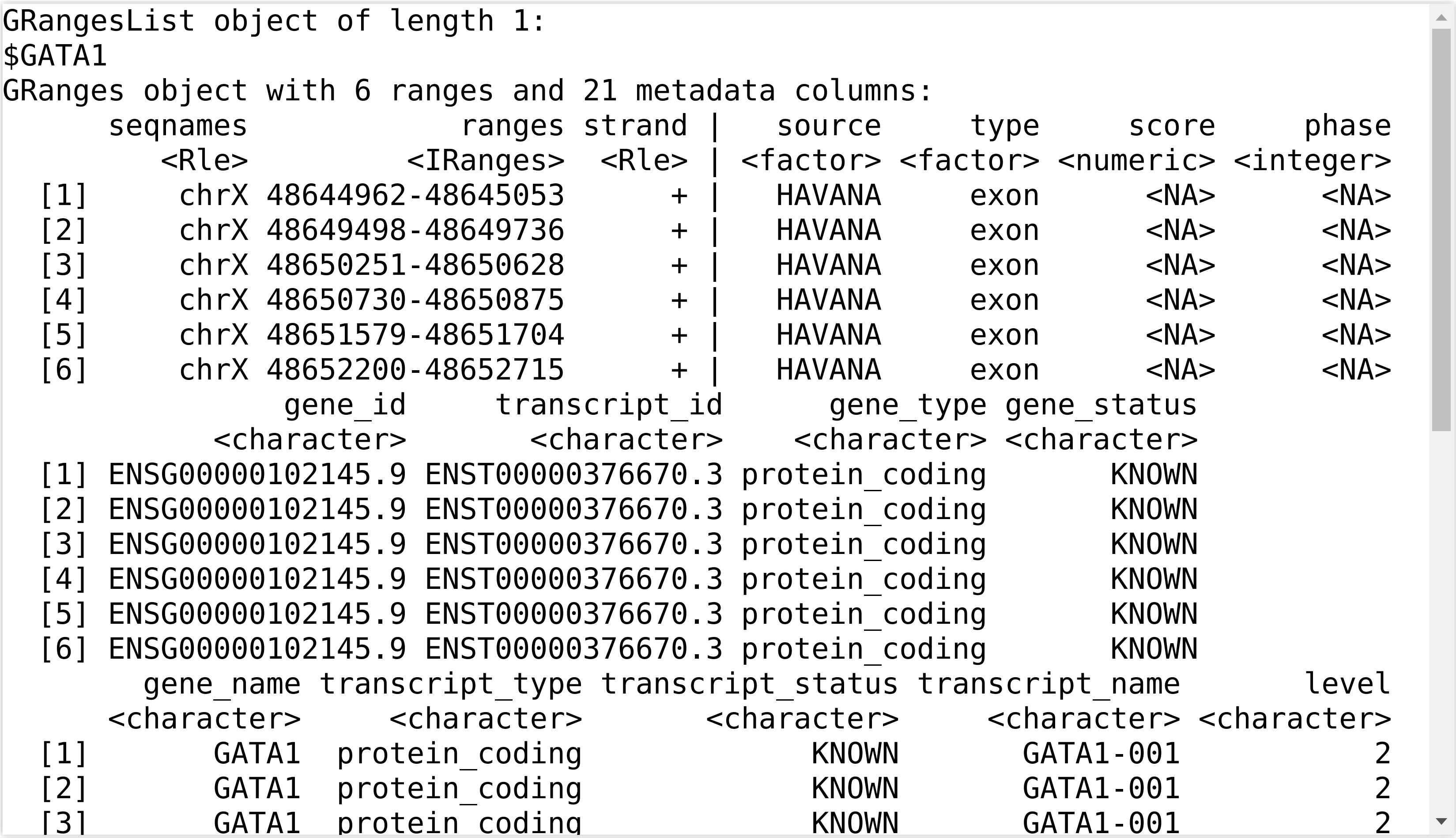
```r
gata1_tx <- rtracklayer::import.gff(
                        "gencode.v19.chrX.gtf.gz") %>%
  filter(gene_name == "GATA1" &
         type == "exon" & transcript_status == "KNOWN") %>%
  print()
```



```
GRanges object with 6 ranges and 21 metadata columns:
      seqnames            ranges strand |   source     type     score     phase
         <Rle>         <IRanges>  <Rle> | <factor> <factor> <numeric> <integer>
  [1]     chrX 48644962-48645053      + |   HAVANA     exon      <NA>      <NA>
  [2]     chrX 48649498-48649736      + |   HAVANA     exon      <NA>      <NA>
  [3]     chrX 48650251-48650628      + |   HAVANA     exon      <NA>      <NA>
  [4]     chrX 48650730-48650875      + |   HAVANA     exon      <NA>      <NA>
  [5]     chrX 48651579-48651704      + |   HAVANA     exon      <NA>      <NA>
  [6]     chrX 48652200-48652715      + |   HAVANA     exon      <NA>      <NA>
                gene_id     transcript_id      gene_type gene_status
            <character>       <character>    <character> <character>
  [1] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [2] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [3] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [4] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [5] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [6] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
        gene_name transcript_type transcript_status transcript_name       level
      <character>     <character>       <character>     <character> <character>
  [1]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [2]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [3]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [4]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [5]       GATA1  protein_coding             KNOWN       GATA1-001           2
```

Group related coordinates by GRangesList

```
gata1_tx <- rtracklayer::import.gff(
                          "gencode.v19.chrX.gtf.gz") %>%
  filter(gene_name == "GATA1" &
         type == "exon" & transcript_status == "KNOWN") %>%
  split(.$gene_name) %>%
  print()
```



```
GRangesList object of length 1:
$GATA1
GRanges object with 6 ranges and 21 metadata columns:
      seqnames            ranges strand |   source     type     score     phase
         <Rle>         <IRanges>  <Rle> | <factor> <factor> <numeric> <integer>
  [1]     chrX 48644962-48645053      + |   HAVANA     exon      <NA>      <NA>
  [2]     chrX 48649498-48649736      + |   HAVANA     exon      <NA>      <NA>
  [3]     chrX 48650251-48650628      + |   HAVANA     exon      <NA>      <NA>
  [4]     chrX 48650730-48650875      + |   HAVANA     exon      <NA>      <NA>
  [5]     chrX 48651579-48651704      + |   HAVANA     exon      <NA>      <NA>
  [6]     chrX 48652200-48652715      + |   HAVANA     exon      <NA>      <NA>
                gene_id     transcript_id      gene_type gene_status
            <character>       <character>    <character> <character>
  [1] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [2] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [3] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [4] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [5] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
  [6] ENSG00000102145.9 ENST00000376670.3 protein_coding       KNOWN
        gene_name transcript_type transcript_status transcript_name       level
      <character>     <character>       <character>     <character> <character>
  [1]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [2]       GATA1  protein_coding             KNOWN       GATA1-001           2
  [3]       GATA1  protein_coding             KNOWN       GATA1-001           2
```

Join beginning and end of 5′UTR

```
gata1_5utr <- union_ranges(gata1_5utr_start, gata1_5utr_end) %>%
  print()
```

```
GRanges object with 2 ranges and 0 metadata columns:
      seqnames            ranges strand
         <Rle>         <IRanges>  <Rle>
  [1]     chrX          48644998      *
  [2]     chrX 48649517-48649519      *
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

Convert 5′UTR coords to transcript coordinates

```
gata1_5utr <- union_ranges(gata1_5utr_start, gata1_5utr_end) %>%
  GenomicFeatures::mapToTranscripts(gata1_tx) %>%
  print()
```

```
GRanges object with 2 ranges and 2 metadata columns:
      seqnames    ranges strand |     xHits transcriptsHits
         <Rle> <IRanges>  <Rle> | <integer>       <integer>
  [1]    GATA1        37      + |         1               1
  [2]    GATA1   112-114      + |         2               1
  -------
  seqinfo: 1 sequence from an unspecified genome
```

Convert 5 ’UTR to a single GRanges

```
gata1_5utr <- union_ranges(gata1_5utr_start, gata1_5utr_end) %>%
  GenomicFeatures::mapToTranscripts(gata1_tx) %>%
  mutate(start = min(start), end = max(end)) %>%
  print()
```

```
GRanges object with 2 ranges and 2 metadata columns:
      seqnames    ranges strand |     xHits transcriptsHits
         <Rle> <IRanges>  <Rle> | <integer>       <integer>
  [1]    GATA1    37-114      + |         1               1
  [2]    GATA1    37-114      + |         2               1
  -------
  seqinfo: 1 sequence from an unspecified genome
```

Convert 5 ’UTR to a single GRanges

```
gata1_5utr <- union_ranges(gata1_5utr_start, gata1_5utr_end) %>%
  GenomicFeatures::mapToTranscripts(gata1_tx) %>%
  mutate(start = min(start), end = max(end)) %>%
  magrittr::extract(1) %>%
  print()
```

```
GRanges object with 1 range and 2 metadata columns:
      seqnames    ranges strand |     xHits transcriptsHits
         <Rle> <IRanges>  <Rle> | <integer>       <integer>
  [1]    GATA1    37-114      + |         1               1
  -------
  seqinfo: 1 sequence from an unspecified genome
```

Get genome sequence

```
BSgenome.Hsapiens.UCSC.hg19::BSgenome.Hsapiens.UCSC.hg19 %>%
  print()
```

```
Human genome:
# organism: Homo sapiens (Human)
# provider: UCSC
# provider version: hg19
# release date: Feb. 2009
# release name: Genome Reference Consortium GRCh37
# 93 sequences:
#   chr1                  chr2                  chr3
#   chr4                  chr5                  chr6
#   chr7                  chr8                  chr9
#   chr10                 chr11                 chr12
#   chr13                 chr14                 chr15
#   ...                   ...                   ...
#   chrUn_gl000235        chrUn_gl000236        chrUn_gl000237
#   chrUn_gl000238        chrUn_gl000239        chrUn_gl000240
#   chrUn_gl000241        chrUn_gl000242        chrUn_gl000243
#   chrUn_gl000244        chrUn_gl000245        chrUn_gl000246
#   chrUn_gl000247        chrUn_gl000248        chrUn_gl000249
# (use 'seqnames()' to see all the sequence names, use the '$' or '[[' operator
# to access a given sequence)
```

Get transcript sequence

```
genome <- BSgenome.Hsapiens.UCSC.hg19::BSgenome.Hsapiens.UCSC.hg19

tx_seq <- GenomicFeatures::extractTranscriptSeqs(
                                    genome, gata1_tx) %>%
  print()
```

```
  A DNAStringSet instance of length 1
    width seq                                                   names
[1]  1497 CAAAGGCCAAGGCCAGCCAGGAC...AAAATAAAACCACCAAAGTCCTG GATA1
```

Get 5′ UTR sequence

```
library(Biostrings)

genome <- BSgenome.Hsapiens.UCSC.hg19::BSgenome.Hsapiens.UCSC.hg19


utr_seq <- GenomicFeatures::extractTranscriptSeqs(
                                genome, gata1_tx) %>%
  subseq(start = start(gata1_5utr), end = end(gata1_5utr)) %>%
  print()
```

```
  A DNAStringSet instance of length 1
    width seq                                               names
[1]    78 ACACTGAGCTTGCCACATCCCCA...GGTTAATCCCCAGAGGCTCCATG GATA1
```

```
Biostrings::writeXStringSet(utr_seq, "gata1_5utr.fasta")
```