KB & Seqspec for share-seq

scRNA-SEQ

scRNA-SEQ (TBD – with the newer version?)

!kb ref - differences

- Adding multiple references files Vs. cDNA only

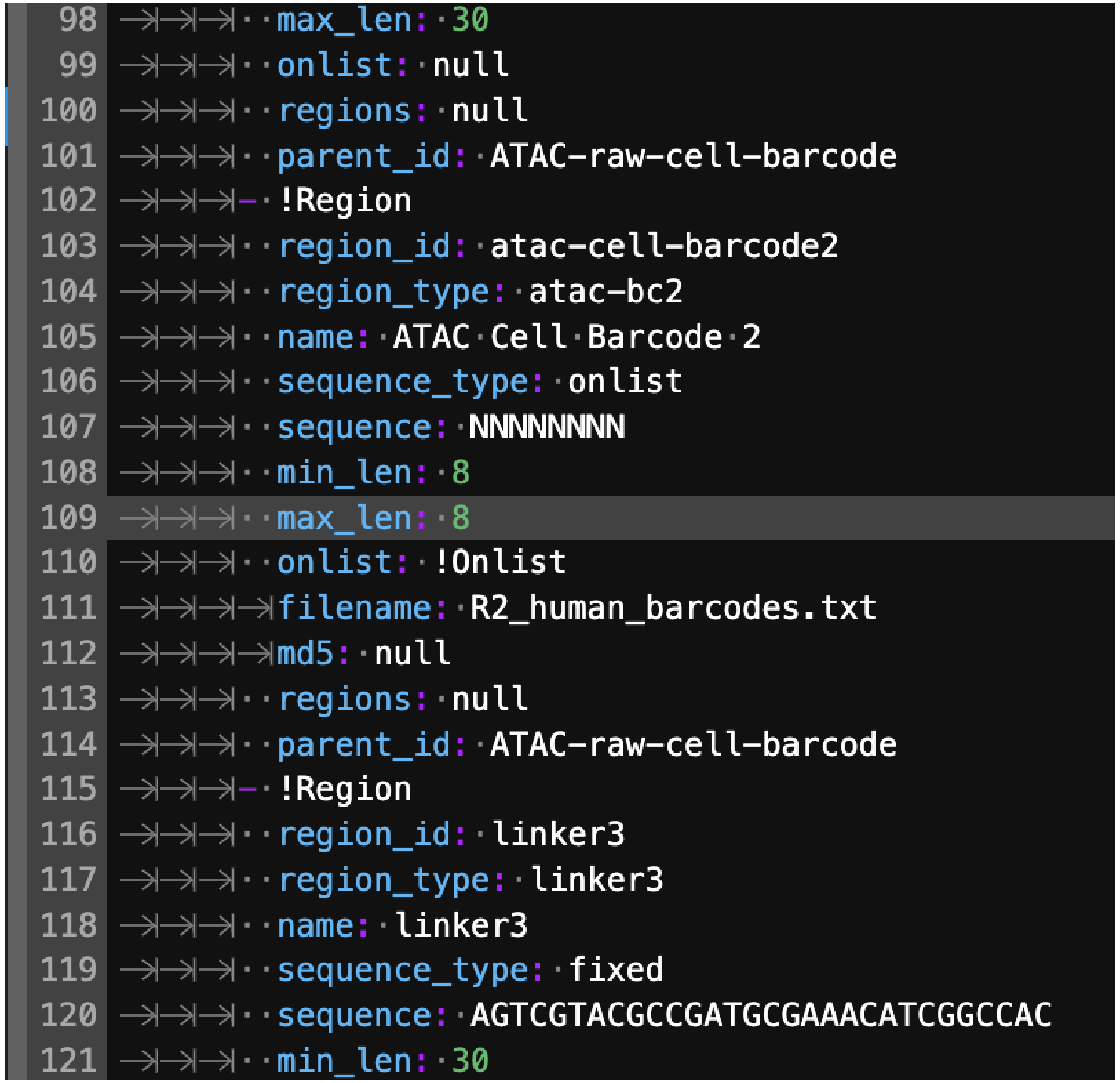
  -
    ```
    !kb ref -i ref_cDNA/transcriptome.idx -g ref_cDNA/transcripts_to_genes.txt \ -
    f1 ref_cDNA/Homo_sapiens.GRCh38.cdna.all.fa.gz \
    reference/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz \
    reference/Homo_sapiens.GRCh38.109.gtf.gz
    ```
- CDNA Reference file is generated from DNA and GTF Vs. Bulk - cDNA only (Downloaded from ENSEMBL)
  - One time effort

Fastq files

- ATAC:
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC.barcodes_annotated.end1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC.barcodes_annotated.end2.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC_L001_R1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC_L001_R2.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC_L002_R1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_ATAC_L002_R2.fastq.gz
- RNA:
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA.barcodes_annotated.UMI.end1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA.barcodes_annotated.UMI.end2.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA_L001_R1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA_L001_R2.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA_L002_R1.fastq.gz
- /oak/stanford/groups/akundaje/marinovg/ENCODE4/single-cell/2023-06-02-scJamboree/BMMC_single_donor_RNA_L002_R2.fastq.gz

Issues with YAML

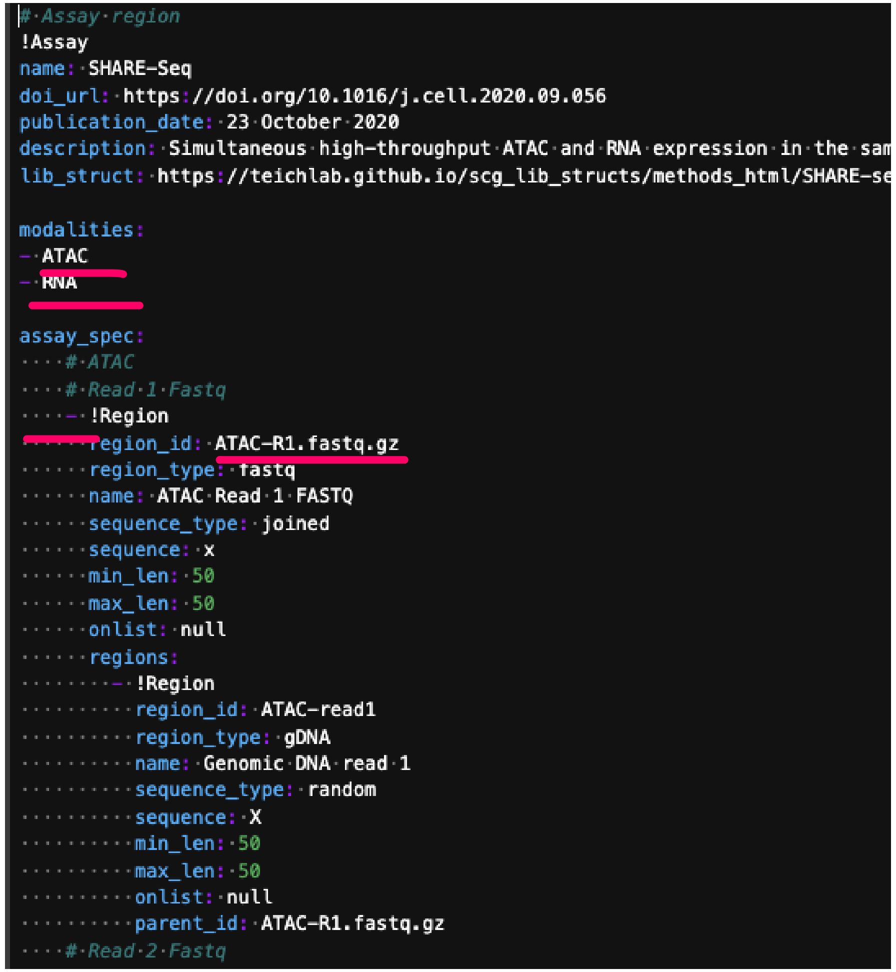
- Can not have tabs (->)
- File structure is important. Assays in the root (RNA, ATAC)
- YAML debugging: https://www.yamllint.com/
- !seqspec print shows the YAML structure



```
⇥⇥⇥··max_len:·30
⇥⇥⇥··onlist:·null
⇥⇥⇥··regions:·null
⇥⇥⇥··parent_id:·ATAC-raw-cell-barcode
⇥⇥⇥-·!Region
⇥⇥⇥··region_id:·atac-cell-barcode2
⇥⇥⇥··region_type:·atac-bc2
⇥⇥⇥··name:·ATAC·Cell·Barcode·2
⇥⇥⇥··sequence_type:·onlist
⇥⇥⇥··sequence:·NNNNNNNN
⇥⇥⇥··min_len:·8
⇥⇥⇥··max_len:·8
⇥⇥⇥··onlist:·!Onlist
⇥⇥⇥⇥filename:·R2_human_barcodes.txt
⇥⇥⇥⇥md5:·null
⇥⇥⇥··regions:·null
⇥⇥⇥··parent_id:·ATAC-raw-cell-barcode
⇥⇥⇥-·!Region
⇥⇥⇥··region_id:·linker3
⇥⇥⇥··region_type:·linker3
⇥⇥⇥··name:·linker3
⇥⇥⇥··sequence_type:·fixed
⇥⇥⇥··sequence:·AGTCGTACGCCGATGCGAAACATCGGCCAC
⇥⇥⇥··min_len:·30
```

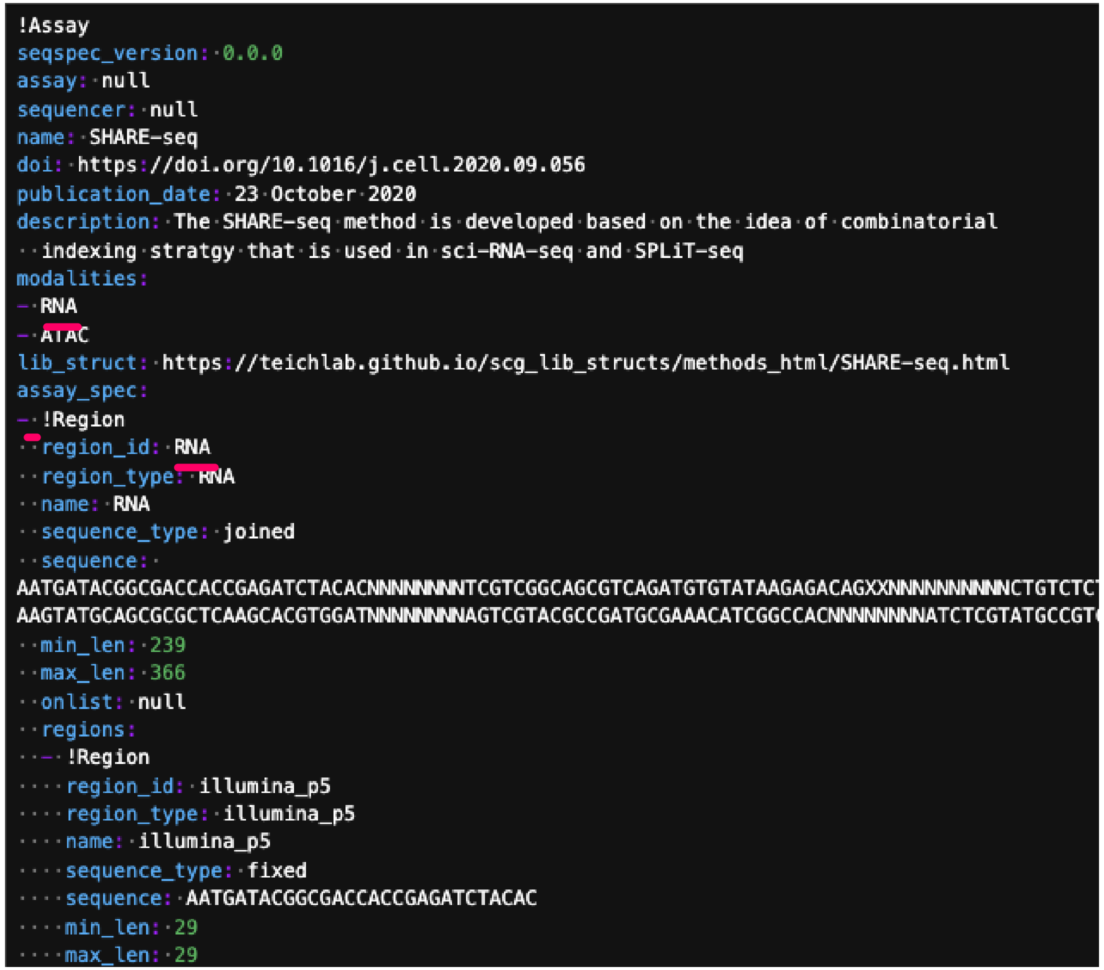
High level comparison

Broad



```
# Assay region
!Assay
name: SHARE-Seq
doi_url: https://doi.org/10.1016/j.cell.2020.09.056
publication_date: 23 October 2020
description: Simultaneous high-throughput ATAC and RNA expression in the sam
lib_struct: https://teichlab.github.io/scg_lib_structs/methods_html/SHARE-se

modalities:
- ATAC
- RNA

assay_spec:
    # ATAC
    # Read 1 Fastq
    - !Region
      region_id: ATAC-R1.fastq.gz
      region_type: fastq
      name: ATAC Read 1 FASTQ
      sequence_type: joined
      sequence: x
      min_len: 50
      max_len: 50
      onlist: null
      regions:
        - !Region
          region_id: ATAC-read1
          region_type: gDNA
          name: Genomic DNA read 1
          sequence_type: random
          sequence: X
          min_len: 50
          max_len: 50
          onlist: null
          parent_id: ATAC-R1.fastq.gz
    # Read 2 Fastq
```

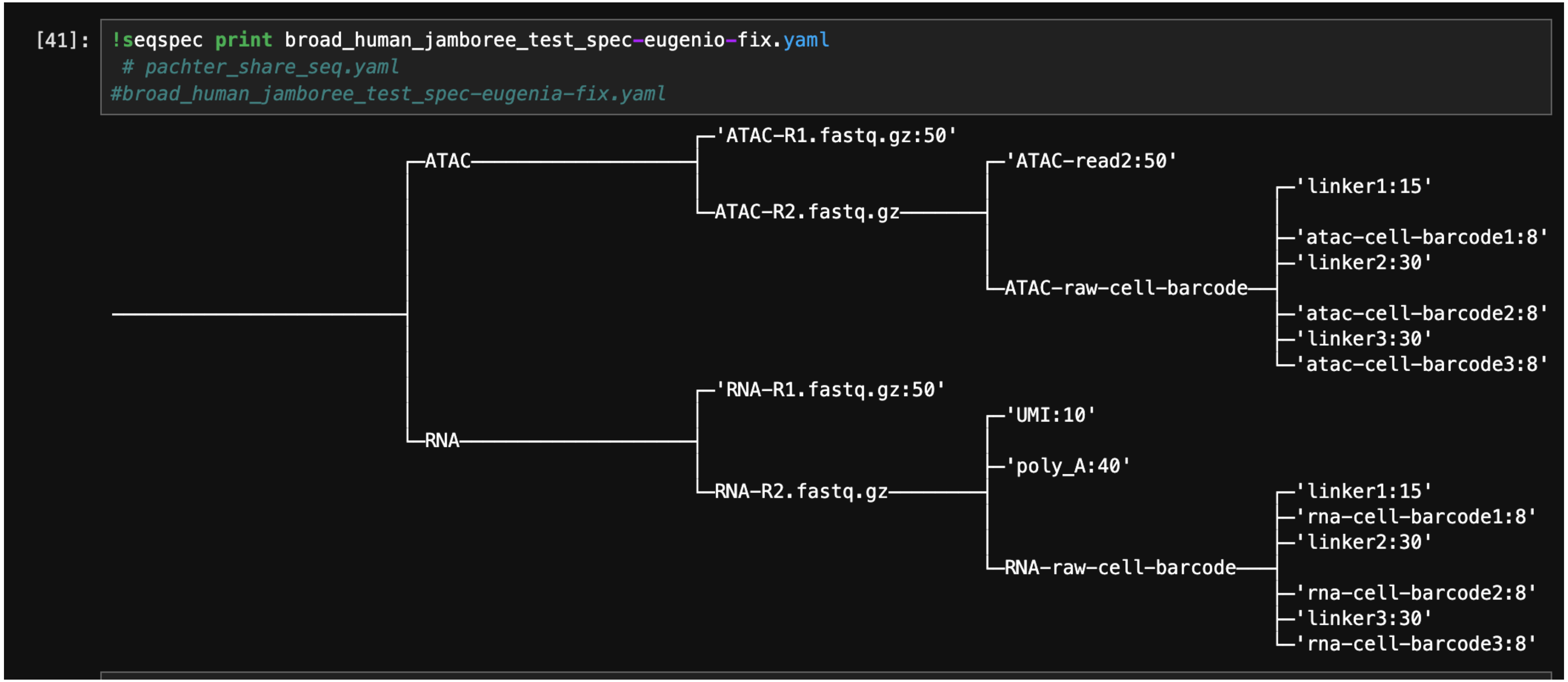
Pachter



```
!Assay
seqspec_version: 0.0.0
assay: null
sequencer: null
name: SHARE-seq
doi: https://doi.org/10.1016/j.cell.2020.09.056
publication_date: 23 October 2020
description: The SHARE-seq method is developed based on the idea of combinatorial
  indexing stratgy that is used in sci-RNA-seq and SPLiT-seq
modalities:
- RNA
- ATAC
lib_struct: https://teichlab.github.io/scg_lib_structs/methods_html/SHARE-seq.html
assay_spec:
- !Region
  region_id: RNA
  region_type: RNA
  name: RNA
  sequence_type: joined
  sequence: 
AATGATACGGCGACCACCGAGATCTACACNNNNNNNNTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGXXNNNNNNNNNNCTGTCTCT
AAGTATGCAGCGCGCTCAAGCACGTGGATNNNNNNNNAGTCGTACGCCGATGCGAAACATCGGCCACNNNNNNNNATCTCGTATGCCGT
  min_len: 239
  max_len: 366
  onlist: null
  regions:
  - !Region
    region_id: illumina_p5
    region_type: illumina_p5
    name: illumina_p5
    sequence_type: fixed
    sequence: AATGATACGGCGACCACCGAGATCTACAC
    min_len: 29
    max_len: 29
```

After fix (collaborative effort with the Broad)

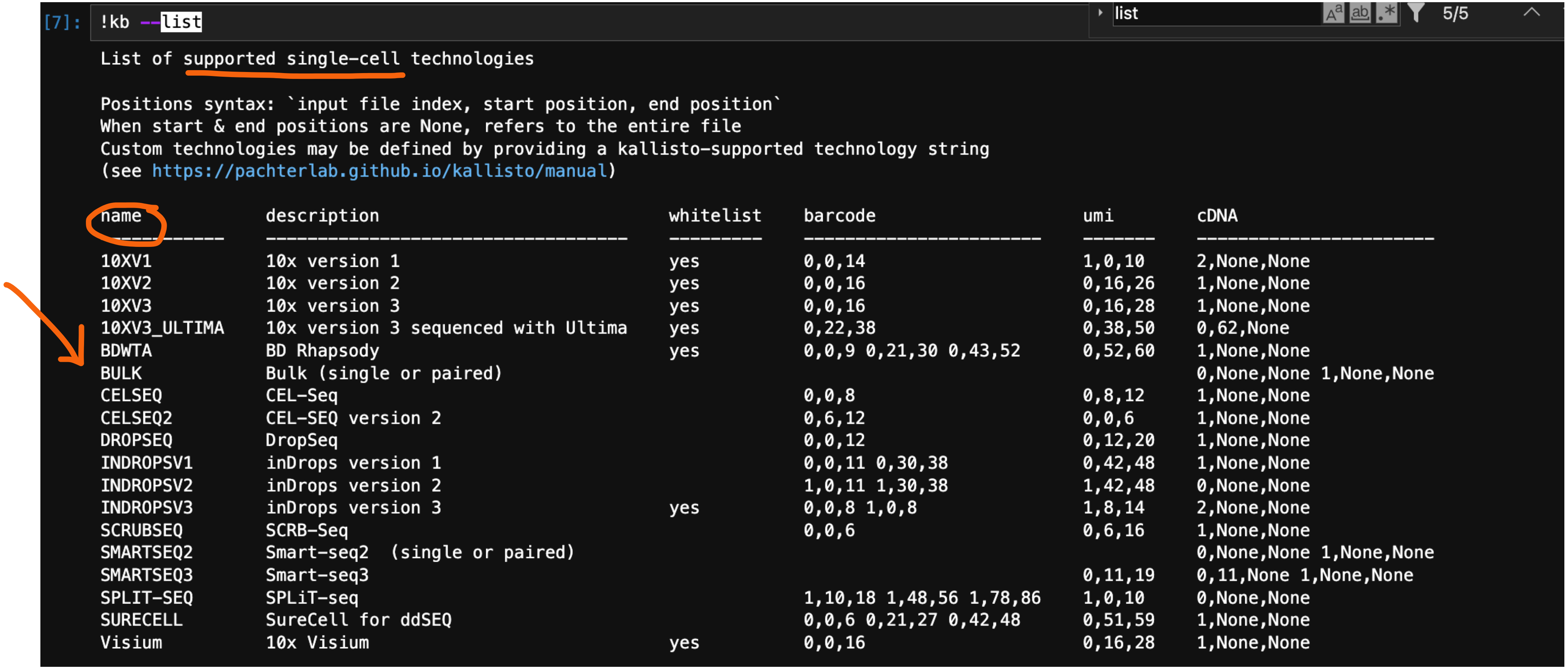
>Seqspec print



```
[41]: !seqspec print broad_human_jamboree_test_spec-eugenio-fix.yaml
       # pachter_share_seq.yaml
      #broad_human_jamboree_test_spec-eugenia-fix.yaml
```

```
                            ┌─'ATAC-R1.fastq.gz:50'
            ┌─ATAC──────────┤                     ┌─'ATAC-read2:50'
            │               │                     │                       ┌─'linker1:15'
            │               └─ATAC-R2.fastq.gz────┤                       │
            │                                     │                       ├─'atac-cell-barcode1:8'
            │                                     │                       ├─'linker2:30'
            │                                     └─ATAC-raw-cell-barcode─┤
────────────┤                                                             ├─'atac-cell-barcode2:8'
            │                                                             ├─'linker3:30'
            │                                                             └─'atac-cell-barcode3:8'
            │               ┌─'RNA-R1.fastq.gz:50'
            │               │                     ┌─'UMI:10'
            └─RNA───────────┤                     │
                            │                     ├─'poly_A:40'
                            └─RNA-R2.fastq.gz─────┤                       ┌─'linker1:15'
                                                  │                       ├─'rna-cell-barcode1:8'
                                                  │                       ├─'linker2:30'
                                                  └─RNA-raw-cell-barcode──┤
                                                                          ├─'rna-cell-barcode2:8'
                                                                          ├─'linker3:30'
                                                                          └─'rna-cell-barcode3:8'
```

>Kb --list – supported by name



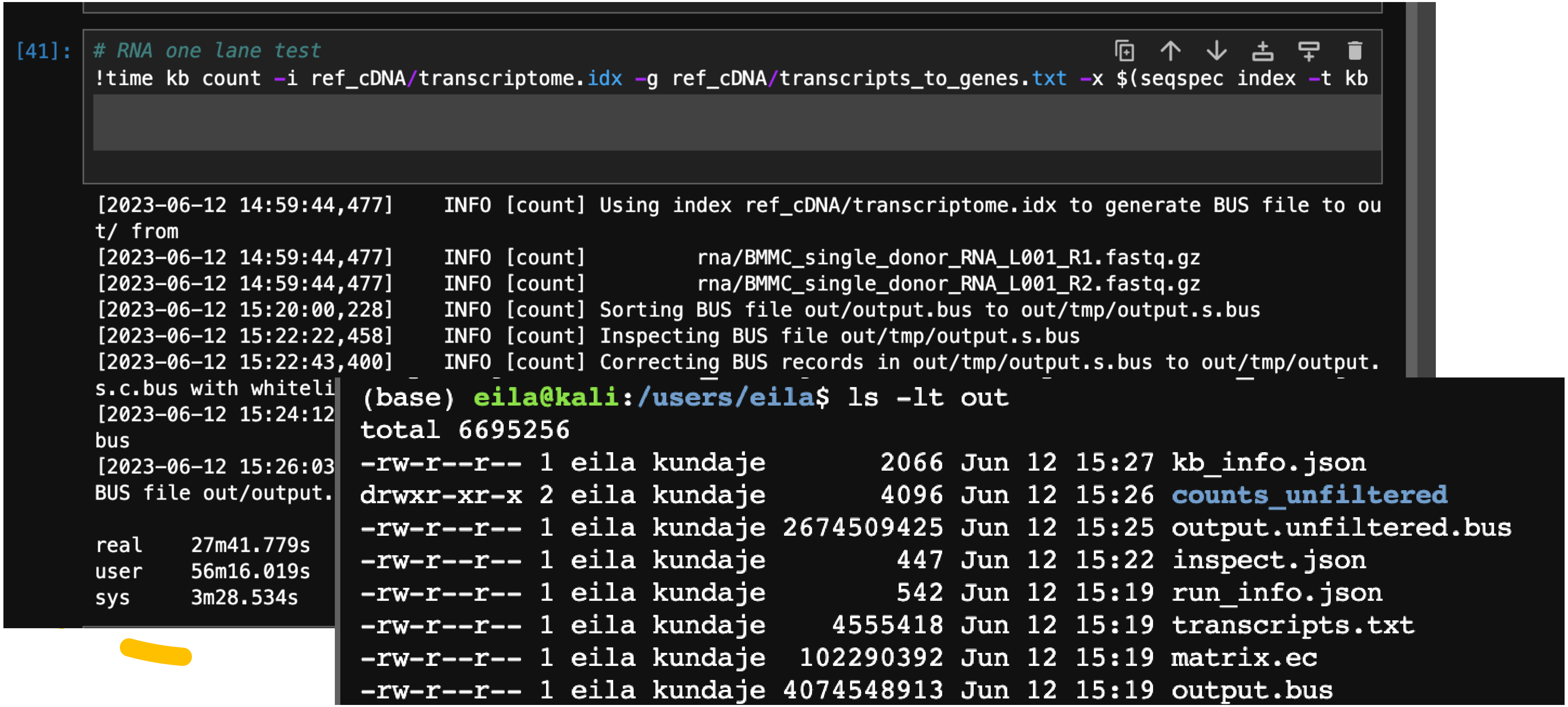
```
[7]: !kb --list

     List of supported single-cell technologies

     Positions syntax: `input file index, start position, end position`
     When start & end positions are None, refers to the entire file
     Custom technologies may be defined by providing a kallisto-supported technology string
     (see https://pachterlab.github.io/kallisto/manual)

     name            description                            whitelist    barcode                   umi        cDNA
     ------------    -----------------------------------    ---------    -----------------------   -------    -----------------------
     10XV1           10x version 1                          yes          0,0,14                    1,0,10     2,None,None
     10XV2           10x version 2                          yes          0,0,16                    0,16,26    1,None,None
     10XV3           10x version 3                          yes          0,0,16                    0,16,28    1,None,None
     10XV3_ULTIMA    10x version 3 sequenced with Ultima    yes          0,22,38                   0,38,50    0,62,None
     BDWTA           BD Rhapsody                            yes          0,0,9 0,21,30 0,43,52     0,52,60    1,None,None
     BULK            Bulk (single or paired)                                                                  0,None,None 1,None,None
     CELSEQ          CEL-Seq                                             0,0,8                     0,8,12     1,None,None
     CELSEQ2         CEL-SEQ version 2                                   0,6,12                    0,0,6      1,None,None
     DROPSEQ         DropSeq                                             0,0,12                    0,12,20    1,None,None
     INDROPSV1       inDrops version 1                                   0,0,11 0,30,38            0,42,48    1,None,None
     INDROPSV2       inDrops version 2                                   1,0,11 1,30,38            1,42,48    0,None,None
     INDROPSV3       inDrops version 3                      yes          0,0,8 1,0,8               1,8,14     2,None,None
     SCRUBSEQ        SCRB-Seq                                            0,0,6                     0,6,16     1,None,None
     SMARTSEQ2       Smart-seq2  (single or paired)                                                           0,None,None 1,None,None
     SMARTSEQ3       Smart-seq3                                                                    0,11,19    0,11,None 1,None,None
     SPLIT-SEQ       SPLiT-seq                                           1,10,18 1,48,56 1,78,86   1,0,10     0,None,None
     SURECELL        SureCell for ddSEQ                                  0,0,6 0,21,27 0,42,48     0,51,59    1,None,None
     Visium          10x Visium                             yes          0,0,16                    0,16,28    1,None,None
```

Does not include Share-seq. But it is supported. You will get the relevant string for share-seq

!kb count

- Require whitelist that matches the barcode lengths on the YAML file
- Few minutes execution (without limiting threads on server machine)
  - 3+ min for one lane (Kali, I didn’t set up the threads)
  - 6+ min on 2 lanes (Kali, I didn’t set up the threads)

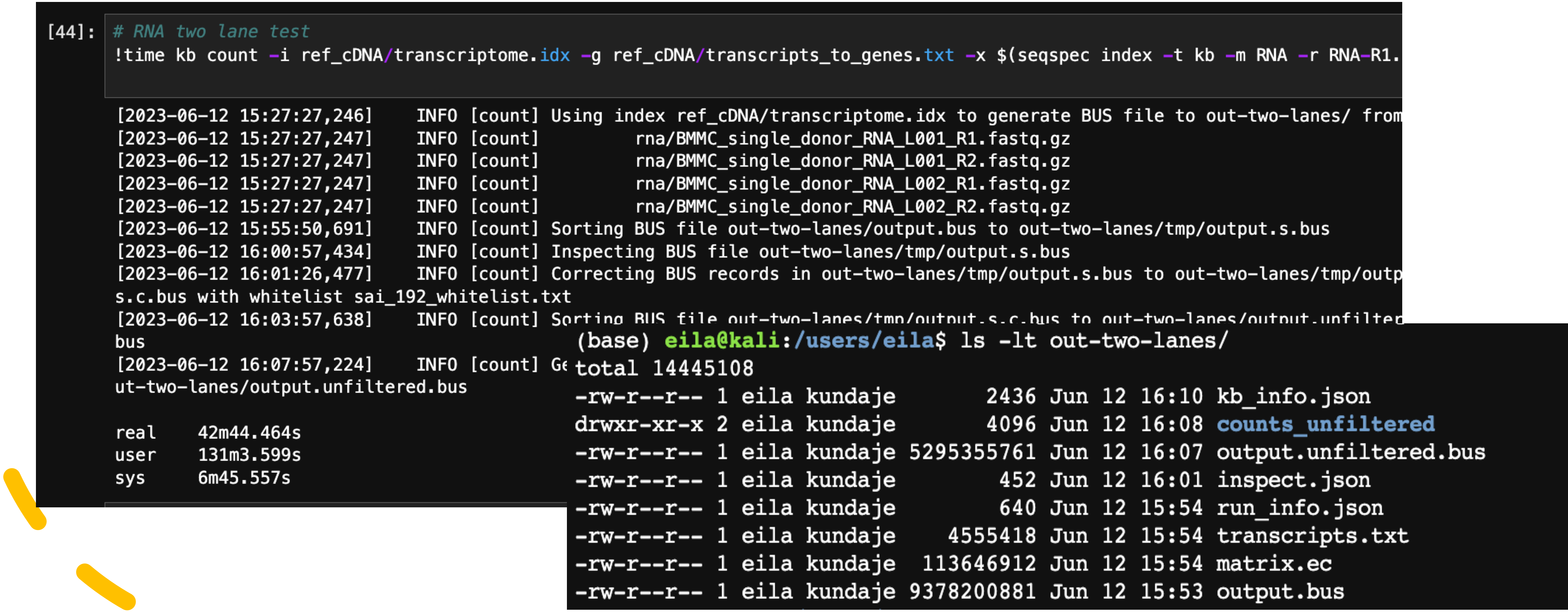
One lane execution and output



```
[41]: # RNA one lane test
      !time kb count -i ref_cDNA/transcriptome.idx -g ref_cDNA/transcripts_to_genes.txt -x $(seqspec index -t kb

[2023-06-12 14:59:44,477]    INFO [count] Using index ref_cDNA/transcriptome.idx to generate BUS file to ou
t/ from
[2023-06-12 14:59:44,477]    INFO [count]         rna/BMMC_single_donor_RNA_L001_R1.fastq.gz
[2023-06-12 14:59:44,477]    INFO [count]         rna/BMMC_single_donor_RNA_L001_R2.fastq.gz
[2023-06-12 15:20:00,228]    INFO [count] Sorting BUS file out/output.bus to out/tmp/output.s.bus
[2023-06-12 15:22:22,458]    INFO [count] Inspecting BUS file out/tmp/output.s.bus
[2023-06-12 15:22:43,400]    INFO [count] Correcting BUS records in out/tmp/output.s.bus to out/tmp/output.
s.c.bus with whiteli
[2023-06-12 15:24:12
bus
[2023-06-12 15:26:03
BUS file out/output.

real    27m41.779s
user    56m16.019s
sys     3m28.534s
```

```
(base) eila@kali:/users/eila$ ls -lt out
total 6695256
-rw-r--r-- 1 eila kundaje       2066 Jun 12 15:27 kb_info.json
drwxr-xr-x 2 eila kundaje       4096 Jun 12 15:26 counts_unfiltered
-rw-r--r-- 1 eila kundaje 2674509425 Jun 12 15:25 output.unfiltered.bus
-rw-r--r-- 1 eila kundaje        447 Jun 12 15:22 inspect.json
-rw-r--r-- 1 eila kundaje        542 Jun 12 15:19 run_info.json
-rw-r--r-- 1 eila kundaje    4555418 Jun 12 15:19 transcripts.txt
-rw-r--r-- 1 eila kundaje  102290392 Jun 12 15:19 matrix.ec
-rw-r--r-- 1 eila kundaje 4074548913 Jun 12 15:19 output.bus
```

Two lanes execution and output



```
[44]: # RNA two lane test
      !time kb count -i ref_cDNA/transcriptome.idx -g ref_cDNA/transcripts_to_genes.txt -x $(seqspec index -t kb -m RNA -r RNA-R1.

[2023-06-12 15:27:27,246]    INFO [count] Using index ref_cDNA/transcriptome.idx to generate BUS file to out-two-lanes/ from
[2023-06-12 15:27:27,247]    INFO [count]         rna/BMMC_single_donor_RNA_L001_R1.fastq.gz
[2023-06-12 15:27:27,247]    INFO [count]         rna/BMMC_single_donor_RNA_L001_R2.fastq.gz
[2023-06-12 15:27:27,247]    INFO [count]         rna/BMMC_single_donor_RNA_L002_R1.fastq.gz
[2023-06-12 15:27:27,247]    INFO [count]         rna/BMMC_single_donor_RNA_L002_R2.fastq.gz
[2023-06-12 15:55:50,691]    INFO [count] Sorting BUS file out-two-lanes/output.bus to out-two-lanes/tmp/output.s.bus
[2023-06-12 16:00:57,434]    INFO [count] Inspecting BUS file out-two-lanes/tmp/output.s.bus
[2023-06-12 16:01:26,477]    INFO [count] Correcting BUS records in out-two-lanes/tmp/output.s.bus to out-two-lanes/tmp/outp
s.c.bus with whitelist sai_192_whitelist.txt
[2023-06-12 16:03:57,638]    INFO [count] Sorting BUS file out-two-lanes/tmp/output.s.c.bus to out-two-lanes/output.unfilter
bus
[2023-06-12 16:07:57,224]    INFO [count] Ge
ut-two-lanes/output.unfiltered.bus

real    42m44.464s
user    131m3.599s
sys     6m45.557s
```

```
(base) eila@kali:/users/eila$ ls -lt out-two-lanes/
total 14445108
-rw-r--r-- 1 eila kundaje       2436 Jun 12 16:10 kb_info.json
drwxr-xr-x 2 eila kundaje       4096 Jun 12 16:08 counts_unfiltered
-rw-r--r-- 1 eila kundaje 5295355761 Jun 12 16:07 output.unfiltered.bus
-rw-r--r-- 1 eila kundaje        452 Jun 12 16:01 inspect.json
-rw-r--r-- 1 eila kundaje        640 Jun 12 15:54 run_info.json
-rw-r--r-- 1 eila kundaje    4555418 Jun 12 15:54 transcripts.txt
-rw-r--r-- 1 eila kundaje  113646912 Jun 12 15:54 matrix.ec
-rw-r--r-- 1 eila kundaje 9378200881 Jun 12 15:53 output.bus
```