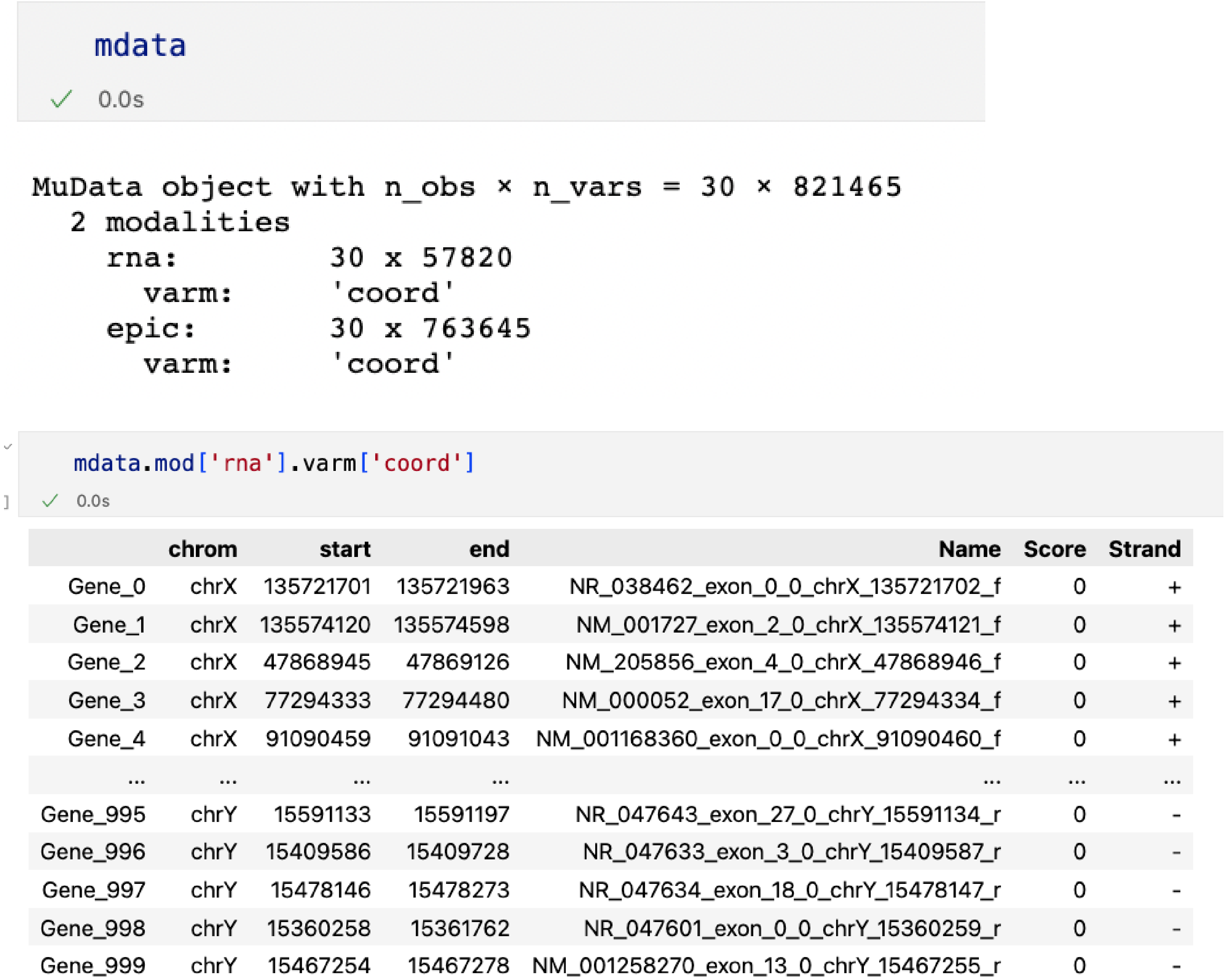
Genomic ranges support in `AnnData` and `MuData`

Qi An

2023.04.18

Set coordinates as `varm`



```
mdata
```

✓ 0.0s

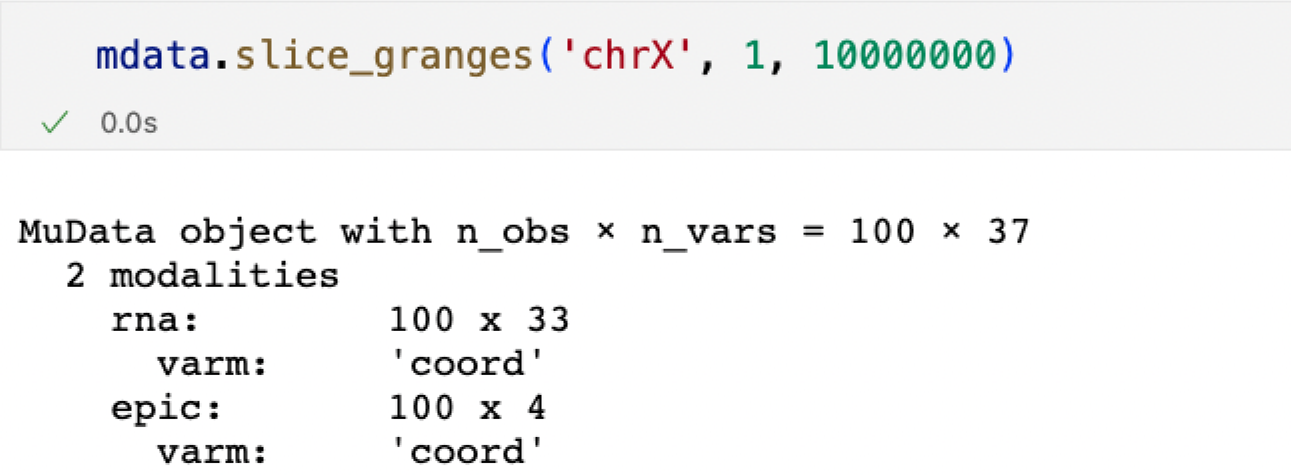
```
MuData object with n_obs × n_vars = 30 × 821465
  2 modalities
    rna:          30 x 57820
      varm:       'coord'
    epic:         30 x 763645
      varm:       'coord'
```

```
mdata.mod['rna'].varm['coord']
```

✓ 0.0s

	chrom	start	end	Name	Score	Strand
Gene_0	chrX	135721701	135721963	NR_038462_exon_0_0_chrX_135721702_f	0	+
Gene_1	chrX	135574120	135574598	NM_001727_exon_2_0_chrX_135574121_f	0	+
Gene_2	chrX	47868945	47869126	NM_205856_exon_4_0_chrX_47868946_f	0	+
Gene_3	chrX	77294333	77294480	NM_000052_exon_17_0_chrX_77294334_f	0	+
Gene_4	chrX	91090459	91091043	NM_001168360_exon_0_0_chrX_91090460_f	0	+
...	...	...	...	...	...	...
Gene_995	chrY	15591133	15591197	NR_047643_exon_27_0_chrY_15591134_r	0	-
Gene_996	chrY	15409586	15409728	NR_047633_exon_3_0_chrY_15409587_r	0	-
Gene_997	chrY	15478146	15478273	NR_047634_exon_18_0_chrY_15478147_r	0	-
Gene_998	chrY	15360258	15361762	NR_047601_exon_0_0_chrY_15360259_r	0	-
Gene_999	chrY	15467254	15467278	NM_001258270_exon_13_0_chrY_15467255_r	0	-

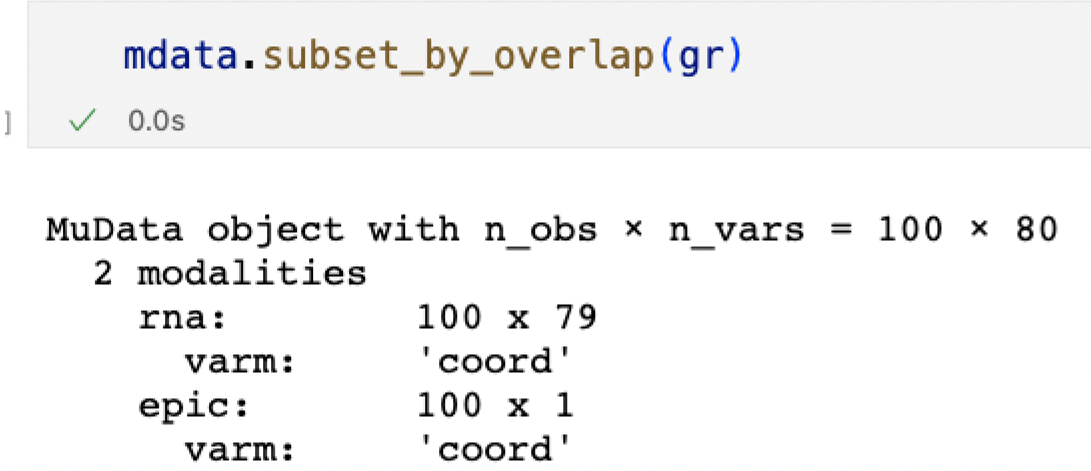
Slice on Coordinates/Subset by overlapping with granges object



```
mdata.slice_granges('chrX', 1, 10000000)
```

✓ 0.0s

```
MuData object with n_obs × n_vars = 100 × 37
  2 modalities
    rna:          100 x 33
      varm:       'coord'
    epic:         100 x 4
      varm:       'coord'
```

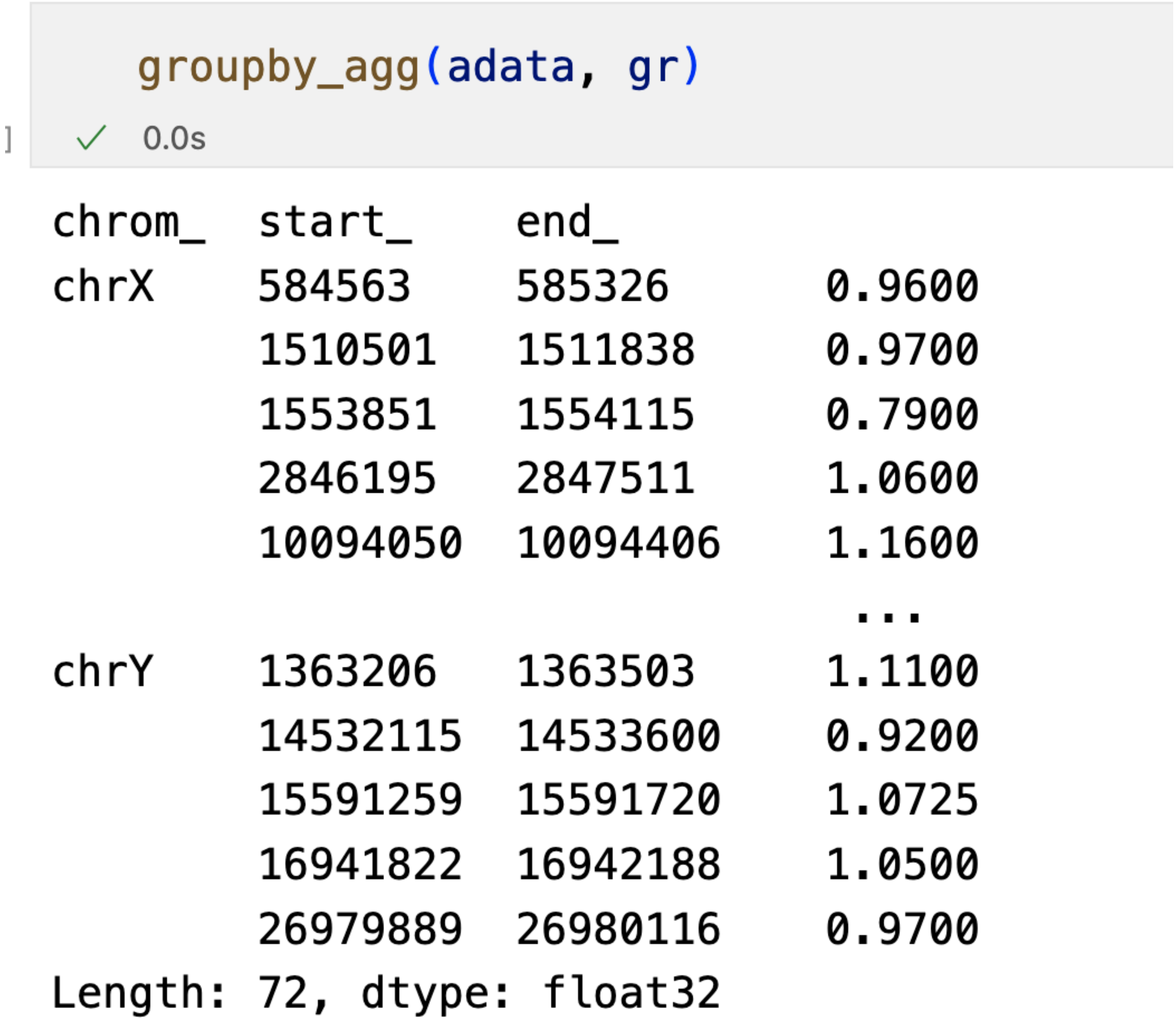


```
mdata.subset_by_overlap(gr)
```

✓ 0.0s

```
MuData object with n_obs × n_vars = 100 × 80
  2 modalities
    rna:          100 x 79
      varm:       'coord'
    epic:         100 x 1
      varm:       'coord'
```

Groupby aggregation of each grange in given list



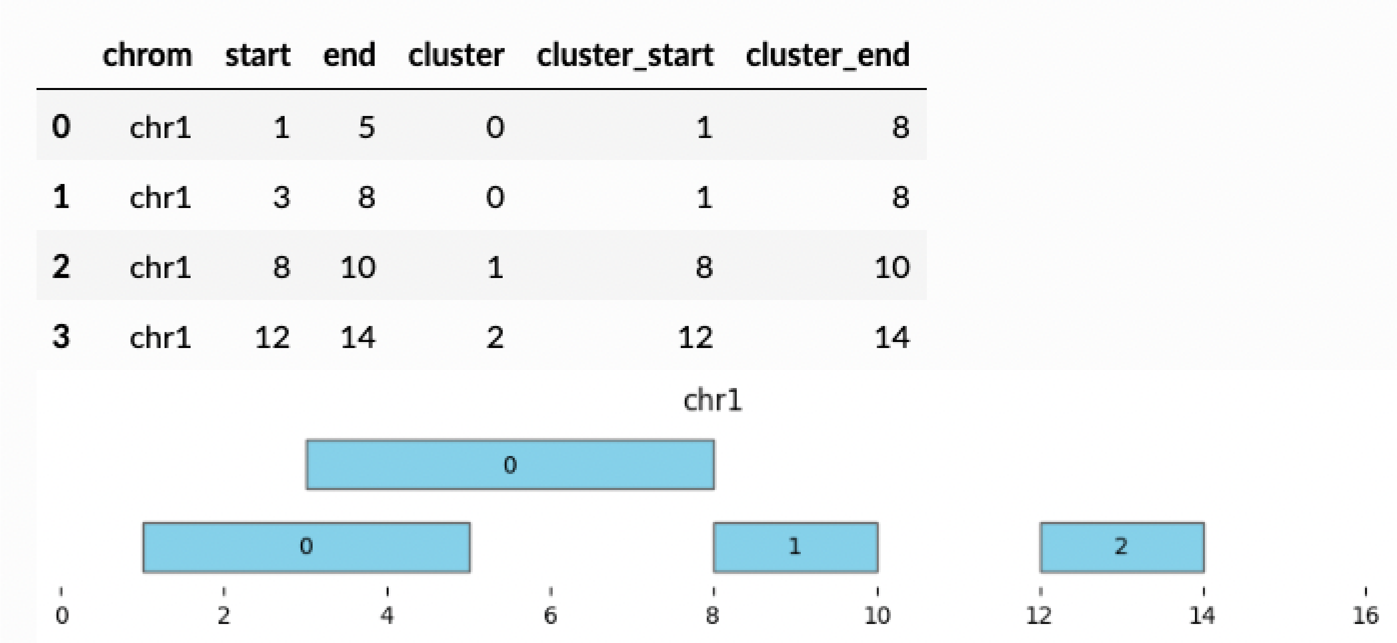
```
groupby_agg(adata, gr)
```

✓ 0.0s

```
chrom_   start_     end_
chrX     584563     585326       0.9600
         1510501    1511838      0.9700
         1553851    1554115      0.7900
         2846195    2847511      1.0600
         10094050   10094406     1.1600
                                  ...
chrY     1363206    1363503      1.1100
         14532115   14533600     0.9200
         15591259   15591720     1.0725
         16941822   16942188     1.0500
         26979889   26980116     0.9700
Length: 72, dtype: float32
```

Problems and outlook

- Exportation: `h5ad` doesn't support serialization of categorical variables; chromosome data cannot be exported
- Cross layer analysis
- Peak analysis: For DMR and ATAC seq, peaks called from each sample not overlapping.



	chrom	start	end	cluster	cluster_start	cluster_end
0	chr1	1	5	0	1	8
1	chr1	3	8	0	1	8
2	chr1	8	10	1	8	10
3	chr1	12	14	2	12	14