adding genomic ranges to muon or anndata

Stephen Kraemer

August 17, 2022

Contents

1 Public API / features **1**
- 1.1 Slicing . . . 1
- 1.2 Subset by overlap . . . 1
- 1.3 Groupby-Agg by overlap . . . 1
- 1.4 Muon only: muon.subset$_{\text{tointersect}}$ . . . 1
- 1.5 Muon only: muon.subset$_{\text{byfeatures}}$ . . . 1

2 Less visible features **2**
- 2.1 Support serialization . . . 2
- 2.2 Compatibility with standard indexing . . . 2
- 2.3 Memory performance . . . 2
- 2.4 Include in conversion to other container formats, eg from Bioconductor . . . 2
- 2.5 Support ragged genomic ranges annotation for "indirect features" . . . 2

3 Implementation notes **2**
- 3.1 How to talk to . . . 2
- 3.2 Where to store the pyranges objects? . . . 3
- 3.3 Use existing AnnData indexing capabilities . . . 3
- 3.4 Groupby-Agg operations require synchronization of the variable metadata objects . . . 3

1 Public API / features

1.1 Slicing

anndata.slice('chr1', 1_{000000}:2_{000000})

1.2 Subset by overlap

anndata.subset$_{\text{by overlap}}$(gr)

1.3 Groupby-Agg by overlap

anndata.groupby$_{\text{overlap}}$(gr).agg(['mean', 'var', 'std'])
(maybe also groupby-iteration?)

1.4 Muon only: muon.subset$_{\textbf{tointersect}}$

harmonize all modalities to overlapping intervals ie only keep DMRs, ATAC peaks etc. which overlap

1.5 Muon only: muon.subset$_{\textbf{byfeatures}}$

= subset by genes or similar, where each gene is associated with multiple transcripts via indexable data structure

2 Less visible features

2.1 Support serialization

- AnnData objects must be serializable to HDF5-like format
- One could store the dict of dataframes from which the nested containment list is created
- Or one could simply recreate the pyranges object on demand from the genomic interval metadata, after loading from HDF5 or only on demand when required during data analysis

2.2 Compatibility with standard indexing

- nested containment list are immutable
  - each variable indexing operation invalides the data structure
    - could check whether anndata.var has changed when function is called
    - could modify core indexing functions to remove invalid NCLs

2.3 Memory performance

- see serialization and compatibility with standard indexing

2.4 Include in conversion to other container formats, eg from Bioconductor

2.5 Support ragged genomic ranges annotation for "indirect features"

3 Implementation notes

3.1 How to talk to

- Max Frank (Oli Stegle)
- Issac Virshup
  - `https://scverse.zulipchat.com/`

3.2 Where to store the pyranges objects?

- AnnData.uns
- see performance considerations above

3.3 Use existing AnnData indexing capabilities

- use pyranges to get integer index for indexing
- then use AnnData integer indexing

3.4 Groupby-Agg operations require synchronization of the variable metadata objects

- discard all metadata which are no longer compatible with the aggregated features