Notes on Data Structures SIG at Bioc Europe 2017

1. Background on MultiAssayExperiment

curatedTCGAData paradigm: ExperimentHub -> cache -> MultiAssayExperiment

Issues of out-of-memory representation: use DelayedArray for experiments and “save”, and saveHDF5SummarizedExperiment as different approaches relevant to different persistence conditions

Remote HDF5 in HDF5 server

2. DNA Variants

Slides

OnDiskLongTable is a custom representation of many rows, few columns data, not yet exported, but related to SNPlocs applications

Do we need to consider other representations of variants?

Filters are important use cases: activation/soft filter concept

GDS format (vs. HDF5?)

MonetDb and other approaches to handling annotation and location information out of memory

GoogleGenomics, GA4GH APIs hail,..., GATK

3. Pharmacogenomics -- use cases!

- drug x cell summarized sensitivity data can be a SummarizedExperiment
- raw drug x cell x dose x response can be a SummarizedExperiment if there are no multi-drug combinations.
  - But for arbitrarily complicated raw data, need a long form: cellid|drug1|drug2|conc|viability, could extend to more drugs with a separate table and key
- simple example of PharmacoSet -> SummarizedExperiment: gist
- Merck public dataset with multiple drugs per experiment
  - http://mct.aacrjournals.org/content/15/6/1155.long
- RESTful API for pharmacoDb

4. Other back ends for DelayedArray fstArray, matterArray, …, RleArray … new doc in DelayedArray