**matter:** A flexible interface to complex data-on-disk

Kylie A. Bemis

A common challenge in bioinformatics is the proliferation of large, heterogeneous datasets stored in disjoint files and in specialized file formats. Frequently, such datasets exceed computer memory, either individually or in aggregate. Such data poses a major challenge to statisticians attempting to develop domain-specific statistical methods in R. A traditional R solution for this problem has been to convert the data into a file format compatible with R packages built for this purpose (e.g., *bigmemory* and *rhdf5*). Indeed, the Bioconductor package *HDF5Array* simplifies the process of accessing array-like data stored in HDF5 files.

However, experimental data is frequently shared in domain-specific file formats incompatible with solutions like *bigmemory* or *rhdf5*. Here, we argue it is important to enable direct interaction with such data in their native file formats wherever possible. Direct interactions avoid the time and storage cost of creating converted files. This approach minimizes the loss of information that can occur during the conversion, and therefore improves the accuracy and reproducibility of the analytical results.

We present *matter*, a novel paradigm and package for direct interactions with complex, larger-than-memory data on disk in R. The *matter* package provides a flexible interface to data on disk without requiring conversion, and allows aggregation of arbitrarily many files into a single R data structure addressable as an on-disk matrix, array, or data frame. This is achieved by a flexible data representation that abstracts the structure of data-on-disk away from the in-memory structure represented and accessible in R.

We will present a case study showing the integration of *matter* with the *Cardinal* package for statistical analysis of larger-than-memory mass spectrometry imaging (MSI) data. We will also discuss how *matter* fits into the overall ecosystem of Bioconductor with existing packages for on-disk data like *DelayedArray* and *HDF5Array*.