**Working Document on Hierarchical Dirichlet Process for Mutational Signature Discovery**

**Liu Mo and Steve Rozen**

**October 25, 2020**

This document describes the use of the R packages mSigHdp and hdpx to apply the hierarchical Dirichlet process (HDP) approach to the problem of discovering (“extracting”) mutational signatures from mutational spectra. We assume the reader is familiar with mutational signatures and the issues around extracting mutational signatures from “catalogs” of mutational spectra. The hdpx package was forked from Nicola Roberts’s hdp package, but uses a different approach to combining “raw clusters” into “components” (the clusters of mutations due to particular mutational signatures.) The packages can be installed from github repos steverozen/hdpx and steverozen/mSigHdp. <Need to create stable branches by Monday and refer to them here>

Broadly speaking, the hdpx package is agnostic to the application of HDP, while mSigHdp adapts HDP to mutational signature extraction. There is one important exception to this general design decision, however: parallelization is implemented in mSigHdp, even though this could be used for any application of the HDP.

**How extract signatures**

The main function for extracting mutational signatures is mSigHdp::RunHdpParallel <Mo, change NewRunHdpParallel to this – if there is an old version rename it to DepcreatedRunHdpParallel. Define NewRunHdpParalle <- RunHdpParallel.>

**Important**: hdpx has only been tested on Linux. Not recommended to run it on MS Windows.

Mo, can you provide a working toy example?

**How to interpret and evaluate the results**