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

**Liu Mo and Steve Rozen**

**October 27, 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. However it uses a different approach to combining “raw clusters” into “components” (the clusters of mutations due to particular mutational signatures.) The hdpx and mSigHdp packages can be installed from github:

remotes::install\_github(repo = "steverozen/hdpx", ref = "v0.3.0")

remotes::install\_github(repo = "steverozen/mSigHdp", ref = "v1.1.2")

Broadly speaking, the hdpx package is agnostic to the problem that HDP is being applied to, while mSigHdp is specific to the application of 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 HDP.

**How to extract signatures**

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

The main function for extracting mutational signatures is mSigHdp::RunHdpxParallel. There is a demo script at

<https://raw.githubusercontent.com/steverozen/mSigHdp/master/data-raw/RunHdpxParallel.example.R>

This can be run as follows:

R --vanilla < RunHdpxParallel.example.R > out.txt 2> err.txt &

This will create the folder ./output.from.RunHdpxParallel.example containing (1) information on extracted signatures and their assignments; (2) diagnostic plots; and (3) burn-in and Gibbs sampling checkpoints.

**How to evaluate the results**

The files extracted.signatures.csv and extracted.signature.pdf contain the extracted signatures. Signatures named e.g “hdp.999” were found in ≥ 90% of the posterior samples. Signatures named e.g. “potential hdp.999” were found in ≥ 50% but < 90% of posterior samples and should be considered in light of other evidence. They may have been generated by mutational processes with low activity or may be variants of the high confidence signatures.

inferred.exposure.count.pdf and inferred.exposure.proportion.pdf show the activity of each signature in each sample. This exposure information was retrieved from the comp\_dp\_counts matrix in each posterior sample. (the plots are made from inferred.exposures.csv). We do not think exposures (assignments) output from hdpx are very accurate, and we would recommend a separate step using other software to estimate assignments.

If the argument ground.truth.sig is non-NULL, there will be several plots and files generated for comparing extracted signatures with ground truth signatures.

The Diagnostic\_Plots folder will contain:

1.diagnostic.likelihood.pdf

2.diagnostic.signatures.pdf

3.diagnostics.data.assigned.pdf

4.diagnostic.numcluster.pdf

The four plots above were inherited from Nicola Roberts hdp package. Please refer to the vignettes for the original hdp package for details

5. diagnostic.hdp.signature.exposure.each.sample.pdf: For each hdp signature, this file plots the five tumors with highest exposure proportion of the signature. This is useful in understanding which tumors contribute to the signature extraction most and a preliminary decision on if a potential hdp signature should be included.

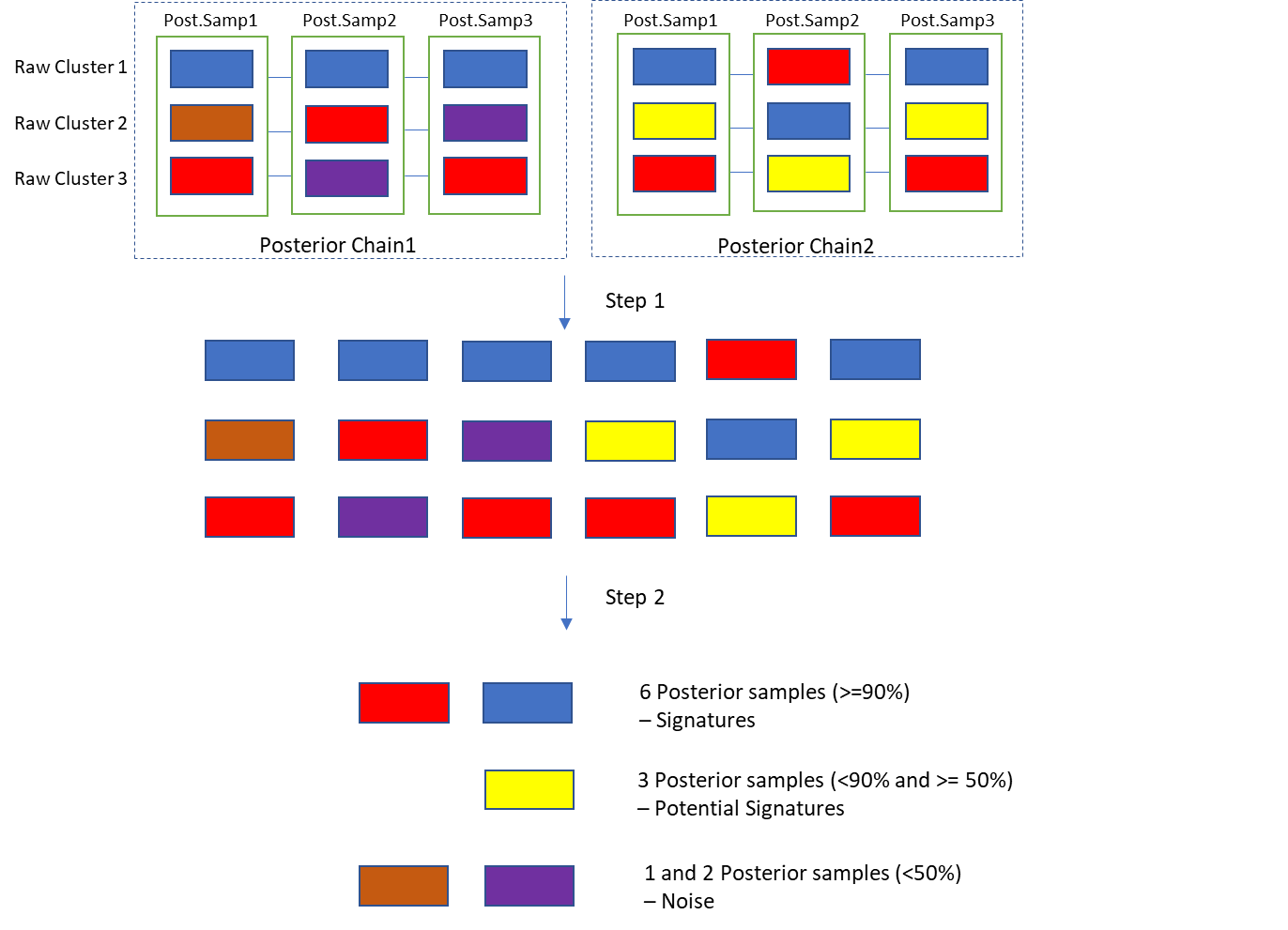
6. component.distribution.in.posterior.samples.pdf: this plotting is still under development. Its purpose is to show for each signature the presence of the raw clusters that contributed to it across the posterior chains. For example, some signatures have contributions from all posterior chains while signatures have contributions from raw clusters only in a single posterior chain.

**Background**

mSigHdp::RunHdpxParallel incorporates burnin and Gibbs sampling processes, followed up by extracting signatures from raw clusters collected by Gibbs sampling, and improved diagnostic plotting with visualization on signature and exposures.

We recommended the gamma distribution with alpha (shape) = 1 and beta (rate) = 20 for the concentration hyperparameters for single base substitution signature extraction, and alpha = 1 and beta = 50 for doublet base substitution and indel signature extraction. These seem to encourage fewer raw clusters, and therefore fewer extracted signatures than default values in the original hdp R package.

The diagram below shows the new approach to combining “raw clusters” into “components”.



This very simple example shows two independent Gibbs sampling chains with 3 posterior samples per chain. There are 3 raw clusters in each posterior sample. Each color represents a raw cluster of mutations generated by a mutational process. Step 1 collects all raw clusters from all posterior samples across all chains into a matrix. Step 2 uses hierarchical clustering to combine raw clusters with high similarity into “components”. We keep track of the number of posterior samples that contribute to each component. In the diagram, the red component was contributed to by red raw clusters in all 6 posterior samples. If we interpret the posterior chains as sampling the posterior distribution of clusters of mutations generated by particular mutational process, this would indicate strong evidence for the mutational signature that generated the red cluster of mutations. In contrast, the brown component comprises the brown raw cluster from only 1 posterior sample, which indicates little evidence for the mutational signature it represents. The yellow component indicates an intermediate situation: there is some but not overwhelming evidence to support a signature that generated this component.