Description of scripts used to evaluate SNACS

Script “functions.R” has functions that are used by the scripts that run SNACS and estimate its performance for the core data illustrated in the paper

Script “truthCall.R” is used to make “truth calls” for the multi-sample experiments. The main function to make the calls is named “makeTruthCall.R”.

Run SNACS on the core multi-sample experiment data illustrated in the paper:

Script "runSNACS.R" runs SNACS on the data used to demonstrate SNACS. It runs SNACS on the core multi-sample experiments (Experiments 5-7) (Ref: section 3.1). It creates heatmaps based on SNPs with cells annotated with the SNACS calls which can be used as a visual tool to see how well SNACS has worked. The script first creates raw SNACS objects from hdf5 files. Then it runs SNACS on those objects to make calls. It then generates heatmaps of the SNP data with the SNPs used to make the calls with cells annotated with the hash data and SNACS calls. Next it makes SNACS + doubletD calls. Then it regenerates the heatmaps annotated with those calls. It creates a legend of the colors used to depict the samples using the sampleColorLegend function in heatmap4 package.

Estimate the performance of SNACS run on the above data:

Script “accuracy\_script.R” estimates the accuracy with which SNACS makes calls (Ref: section 3.2). It uses functions “getAnnoForAccuracy”, “getTruthCall” and “createAnnotatedHeatmap” from script “functions.R” and “makeTruthCall” from script “truthCall.R”. This script compares the multi-sample experiments (Experiments 5-7) to the single-sample experiments (Experiments 1-4). It first creates raw SNACS objects of the single-sample experiments from hdf5 files. It then filters and imputes the missing values in the SNP data as done by SNACS with for the multi-sample experiments (Ref: section 4.5) using the “filterData” and “imputeMissingMutations” functions from SNACS package. It then creates annotation tables containing the SNACS calls and genotypes used to make the calls for each cell in a multi-sample experiment. It uses “getAnnoForAccuracy” function from script “functions.R” to create the tables. Next, it creates tables of truth calls from the single-sample experiments using “makeTruthCall” function in “truthCall.R” script. It then adds the truth calls to the above cell annotation tables using “getAnnoForAccuracy” function from script “functions.R”. To visually evaluate the performance of SNACS, it creates heatmaps, based on the SNP data, of the multi-sample experiment annotated with the truth calls and of the single-sample experiments. Lastly, the script computes the accuracy, sensitivity and specificity for the multi-sample experiments.

Run SNACS on multi-sample experiments 5-7 while varying the number of SNPs:

Script “runSNACS\_varySNPs.R” runs SNACS in a similar manner to script “runSNACS.R” but it varies the number of top SNPs (from 1 to 5) that best distinguishes the hash-antibody-defined groups (Ref: section 2.2).

Estimate the performance of SNACS run on the above data where the number of SNPs is varied:

Script “accuracy\_script\_varySNPs.R” estimates the performance of SNACS when varying the number of SNPs in a manner similar to script “accuracy\_script.R”.

Make “truth calls” for all cells:

Script “truthCallForAllCells.R” makes truth calls for all the cells in the raw data of the multi-sample experiments 5-7. It creates annotation tables with truth calls for all cells in a multi-sample experiment. It is used to estimate the performance of the [XXX] method.

Generate simulated data:

Scripts “run\_simulation.R ““, “functions\_sim.R”, “setup\_simulation.R”, “create\_sim.sh “, “create\_sim.R“, “run\_snacs\_sim.sh “ and “run\_snacs\_sim.R” are used to generate simulated data (Ref: section 3.4)

Script “run\_simulation.R” generates simulates data for a single multi-sample experiment. It is used by script “create\_sim.R” to simulate multiple multi-sample experiments.

Script “functions\_sim.R” has functions that are used by the simulation scripts.

Script “setup\_simulation.R “ creates initial data from single-sample experiments 1-4 that are used to simulate multi-sample experiments.

“create\_sim.sh” runs script “create\_sim.R”. The latter script simulates the multi-sample experiments. In order to that, it runs function “sim" from script “run\_simulation.R” for each simulation. It uses function “truthName2snacsCallName” from script “functions\_sim.R “ to properly format the names of the truth calls.

“run\_snacs\_sim.sh” runs script “run\_snacs\_sim.R”. The latter script runs SNACS on the simulated data. It uses functions “modifyCallName”, “createHeatmapForSim” and “getColVarInfo” from script “functions\_sim.R” to properly format the names of the truth calls, to generate heatmaps based on the SNPs and to get the appropriate variables and their colors for cell level annotation in the heatmaps respectively.