**SAINTexpress: Significance Analysis of INTeractome – Express Version**

Guoci Teo and Hyungwon Choi

June 24, 2013

**1. Installation**

The source code requires g++ version 4.4 or above for compilation. Makefile has been created in the source and a simple call of “make” at the source directory will compile the program. It is convenient to add the directory to the PATH variable for execution of software at any location on the file system. All the libraries have also been included in the distribution, and hence there is no further dependency.

*SAINTexpress* is devoid of time-consuming sampling steps and mostly runs very quickly. This new implementation no longer takes the optional arguments such as lowMode, minFold, and normalize.

**2. Input file preparation**

Similar to the original SAINT, *SAINTexpress* expects three mandatory input files and an additional input file (described in Section 4). These include bait, prey, and interaction files, and the formatting remains the same as the original version. Here we describe it again for reminder and also for new users.

*Bait file*

This file should have three columns: IP name, bait name, and the indicator for test and negative control purifications (T = test, C = control). See the example below with two bait proteins and seven negative control runs.

|  |  |  |
| --- | --- | --- |
| HDAC1-1 | HDAC1 | T |
| HDAC1-2 | HDAC1 | T |
| HDAC2-1 | HDAC2 | T |
| HDAC2-2 | HDAC2 | T |
| GFP\_1 | GFP\_1 | C |
| GFP\_2 | GFP\_2 | C |
| GFP\_3 | GFP\_3 | C |
| GFP\_4 | GFP\_4 | C |
| GFP\_5 | GFP\_5 | C |
| GFP\_6 | GFP\_6 | C |
| GFP\_7 | GFP\_7 | C |

In the example above, each of the two bait proteins HDAC1 and HDAC2 was analyzed in two replicates (preferably biological replicates).

*Prey file*

Prey file should also contain three columns: prey (protein) name, prey protein length, and prey gene name. In the example given below, we constructed the input files using gene IDs as the main prey identifier and thus the first and third columns are identical. Typically these two columns are different: the first column lists the protein identifiers such as GI accession number or Uniprot IDs.

|  |  |  |
| --- | --- | --- |
| GBE1 | 702 | GBE1 |
| HSPE1 | 102 | HSPE1 |
| EFTUD2 | 972 | EFTUD2 |
| AGPAT5 | 364 | AGPAT5 |
| PLCG1 | 1290 | PLCG1 |
| DECR1 | 335 | DECR1 |
| CNP | 421 | CNP |
| PDE12 | 609 | PDE12 |
| PSMC6 | 389 | PSMC6 |
| PSMC1 | 440 | PSMC1 |
| PSMC3 | 439 | PSMC3 |
| PSMC4 | 418 | PSMC4 |
| PSMC2 | 433 | PSMC2 |
| PSMC5 | 406 | PSMC5 |

*Interaction file*

The interaction file should contain four columns: IP name, bait name, prey name, and spectral counts or intensity values, depending on the mode of quantitation. The prey name should coincide with the first column of the prey file. Interactions with zero counts must be removed from the file.

|  |  |  |  |
| --- | --- | --- | --- |
| HDAC1-1 | HDAC1 | YWHAB | 13 |
| HDAC1-1 | HDAC1 | YWHAE | 14 |
| HDAC1-1 | HDAC1 | YWHAH | 5 |
| HDAC1-1 | HDAC1 | YWHAG | 15 |
| HDAC1-1 | HDAC1 | SFN | 3 |
| HDAC1-1 | HDAC1 | YWHAQ | 12 |
| HDAC1-1 | HDAC1 | YWHAZ | 19 |
| HDAC1-1 | HDAC1 | ACTL6A | 3 |
| HDAC1-1 | HDAC1 | ADNP | 9 |
| HDAC1-1 | HDAC1 | ADH5 | 7 |
| HDAC1-1 | HDAC1 | MKI67 | 1 |
| HDAC1-1 | HDAC1 | RERE | 27 |
| HDAC1-1 | HDAC1 | NARS | 3 |
| HDAC1-1 | HDAC1 | ARID4A | 29 |
| HDAC1-1 | HDAC1 | ARID4B | 64 |
| HDAC1-1 | HDAC1 | ARID5B | 45 |

**3. Running the analysis**

To run the analysis, the original SAINT (up to v2.3.4) requires a pre-processing step called “saint-reformat”. We automated this process into the main analysis module and therefore it is no longer necessary to run the reformat command. To analyze the data, use the following command line call:

> SAINTexpress-spc [OPTIONS] <interaction data> <prey data> <bait data>

At the moment, there are two options in SAINTexpress.

(1) –L option: this argument sets the number of virtual control purifications by compression. For instance, if the user wishes to take 4 largest spectral counts for controls, do

> SAINTexpress-spc –L4 inter.dat prey.dat bait.dat

(2) –R option: this argument sets the number of replicates (with largest spectral counts or intensities) to be used for probability calculation in each bait. This option is useful when some baits have more replicates than others. Default is 100, using all replicates in most realistic datasets.

**4. Incorporating known interaction data**

To incorporate external data sources for computing the topology-aware probability score (TopoAvgP), the user must also provide the interaction database file that contains two columns: interaction identifier column and interaction/grouping information column. The first column is just for formality and thus can be filled in with anything (no white space) and it will not be utilized in the scoring. The second column must be formatted as a string of prey identifiers (consistent with the first column of the prey file) separated by a white space. See an example below.

|  |  |
| --- | --- |
| GOID | EntrezGeneID |
| GO:0000002 | SLC25A4 TYMP MEF2A MPV17 LONP1 |
| GO:0000012 | LIG4 TNP1 XRCC1 APTX TDP1 TDP1 APLF LOC100133315 |
| GO:0000018 | IL7R KPNA1 KPNA2 SMARCAD1 |
| GO:0000019 | MRE11A RAD50 |
| GO:0000022 | PRC1 KIF23 |
| GO:0000028 | RPSA RPS6 RPS14 RPS14 RPS17 RPS25 ERAL1 |
| GO:0000038 | ACOX1 HSD17B4 CYP4F2 ACOT2 SLC27A5 SLC27A2 ACSBG1 SLC27A6 ACOT4 ACOT1 |
| GO:0000042 | RAB6A OPTN |

Given this additional input file named “GO.txt”, the user can run the analysis as:

> SAINTexpress-spc –L4 inter.dat prey.dat bait.dat GO.txt

**5. Field description in the output file**

The output file of SAINTexpress reports 16 columns for all observed interactions. Here is the description of the fields:

Bait: bait identifier

Prey: prey identifier

PreyGene: additional prey identifier

Spec: spectral counts for the bait-prey pair

SpecSum: sum of the spectral counts

AvgSpec: average spectral counts over replicates

NumReplicate: number of replicate purifications for the given bait

ctrlCounts: spectral counts in the negative controls

AvgP: main probability score

MaxP: maximal probability score of the interaction over replicates

TopoAvgP: topology-aware probability score incorporating known interaction data

TopoMaxP: topology-aware maximal probability score over replicates

SaintScore: larger of AvgP and TopoAvgP

FoldChange: average spectral count in test interaction divided by the average in controls

Boosted\_by: indicates which known interactors of the same bait contributed to TopoAvgP

FDR: Bayesian false discovery rate