An Introduction to the SNAGEE Package

David Venet

November, 2012

1 Introduction

The SNAGEEpackage is designed to estimate studies and samples signal-to-noise ratios for gene expression data. Those SNRs are related to the statistical strength of the biological conclusions the data support, and so can be used as a proxy for study and sample quality.

As microarray studies can be used to answer many biological questions, possibly unrelated to the ones treated in the original study, the strength of the biological signal is not determined based on sample annotations, but directly using the values of the microarray experiments. Whether SNR is a reliable estimate of data quality, in the sense of well-made experiments, depends on the study type. The signal-to-noise ratio depends not only on the amount of noise, which could be seen as a direct measure of quality, but also on the amount of signal. For instance, a high-quality study comparing a cell line in two different conditions could have very little variability, and so a low SNR. However, we have shown [1] that SNR is a good proxy of quality for studies that comprise a large number of diverse samples, like for instance large studies on cancer tissues, and can reliably be used to rate comparable studies. It can also be used to flag problematic samples inside a study.

SNAGEEestimates SNR using correlations of gene-gene correlations. It has been shown [2] that gene-gene correlations are not random, but that sets of genes are often found to be similarly correlated across different studies and biological conditions. This can be expressed by saying that the gene-gene correlation matrix has a certain distribution, with some genes likely to be correlated while others are not. SNAGEEuses the distribution of the gene correlations as the basis of an SNR measure for all studies and platforms. The distribution of the gene correlations is estimated by using a large number of studies and platforms. The SNR of a study is obtained by comparing its gene correlations to the expected gene correlations. The SNR of individual samples is assessed by observing the difference in the SNR of the study they are part of when they are removed. The sample SNR is a measure of the relative contribution by a sample to the signal and noise of its study, so it is not a ratio, but we still use the term signal-to-noise ratio as it conveys the idea behind the measure.

SNAGEEestimates the study SNR direcly using the gene measurments, which has many advantages compared to existing quality measures techniques: it is

based on a biologically meaningful concept, it works across studies, protocols and platforms, it can be applied to both studies and samples, it is sensitive to probe misannotation, it does not require access to the raw files, and it is fully automated.

Gene-gene correlation matrices calculated on many studies are available in SNAGEEdata, which should be installed alongside SNAGEE. SNAGEEcan be used for any platform and does not use platform-specific information, nor usual quality metrics. The only requirement is that the probes are mapped to gene IDs.

This vignette is intended to give a quick glance of the most useful tools. A more detailed help can be obtained with `help(SNAGEE)` after loading the library.

2 Study SNR

This section details the calculation of the SNR of the study in the data package ALL. This is a study of 128 samples on the Affymetrix U95A platform. The SNR of a study is based on the correlation between its gene-gene correlation matrix and the expected matrix, and so is a number between -1 and 1. Practically, numbers near or below 0 are symptomatic of seriously problematic studies (e.g. gene annotation problems, serious normalization issues). Numbers around 20-30% are average, depending on the platform. For instance, primary Affymetrix platforms (e.g. U133A) have usually higher SNRs, while secondary Affymterix platforms (e.g. U133B) are usually poorer.

```
R> library(SNAGEE)
R> library(ALL)
R> data(ALL)
```

And now the quality of the study can be calculated:

```
R> q = qualStudy(ALL)
```

```
Some genes appear more than once - collapsing using the mean.
```

```
R> print(q)
```

```
[1] 0.4291029
```

43% is about average for a U95A study.

3 Sample SNRs

SNAGEEcan also be used to determine the relative SNRs of samples inside a study. Contrary to the study SNR, the data cannot contain NAs. It is possible to use the imputepackage for instance to remove them. The calculation is longer, but can be parallelized using the parallel package.

Sample SNRs are obtained by comparing the SNR of their study with and without that sample. Those SNR differences are renormalized by dividing them

by their median absolute deviation (mad in R), so that SNRs below -3 could be considered as slightly suspicious, and below -5 as seriously suspicious. Sample SNRs are relative to their study, so a very problematic study could have no particularly suspicious samples—they could all be equally bad. Similarly, a suspicious sample in a high-quality study could still be better than an average sample in a low-quality study.

Calculation of the SNRs of samples is straightforward:

```
R> qs = qualSample(ALL, multicore=FALSE)

Some genes appear more than once - collapsing using the mean.

R> sum(qs < -5)

[1] 6
```

6 samples may have quality problems. This is quite large for a study with 128 samples.

4 Details

This document was written using:

```
R> sessionInfo()

R version 4.4.0 beta (2024-04-15 r86425)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 22.04.4 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats4    stats     graphics  grDevices utils
[6] datasets  methods   base
```

```
other attached packages:
 [1] hgu95av2.db_3.13.0   org.Hs.eg.db_3.19.1
 [3] AnnotationDbi_1.66.0 IRanges_2.38.0
 [5] S4Vectors_0.42.0     ALL_1.45.0
 [7] Biobase_2.64.0       BiocGenerics_0.50.0
 [9] SNAGEE_1.44.0        SNAGEEdata_1.39.0

loaded via a namespace (and not attached):
 [1] crayon_1.5.2            vctrs_0.6.5
 [3] httr_1.4.7              cli_3.6.2
 [5] rlang_1.1.3             DBI_1.2.2
 [7] UCSC.utils_1.0.0        png_0.1-8
 [9] jsonlite_1.8.8          bit_4.0.5
[11] Biostrings_2.72.0       KEGGREST_1.44.0
[13] fastmap_1.1.1           GenomeInfoDb_1.40.0
[15] memoise_2.0.1           compiler_4.4.0
[17] RSQLite_2.3.6           blob_1.2.4
[19] pkgconfig_2.0.3         XVector_0.44.0
[21] R6_2.5.1                GenomeInfoDbData_1.2.12
[23] tools_4.4.0             bit64_4.0.5
[25] zlibbioc_1.50.0         cachem_1.0.8
```

References

[1] David Venet, Vincent Detours, Hugues Bersini *A measure of the signal-to-noise of microarray samples and studies using gene correlations.* PLoS One, 2012.

[2] Lee, Homin K, Hsu, Amy K, Sajdak, Jon, Qin, Jie, Pavlidis, Paul Genome Res, 6, 1085–1094, 2004.