Gender study gene expression data from Vawter et al. [2004]

Laurent Jacob

November 18, 2026

1 Introduction

Vawter et al. [2004] studied differences in gene expression between male and female patients.

This gender study is an interesting benchmark for methods aiming at removing unwanted variation as it expected to be affected by several technical and biological factors: two microarray platforms, three different labs, three tissue localizations in the brain. Most of the 10 patients involved in the study had samples taken from the anterior cingulate cortex (a), the dorsolateral prefontal cortex (d) and the cerebellar hemisphere (c). Most of these samples were sent to three independent labs: UC Irvine (I), UC Davis (D) and University of Michigan, Ann Arbor (M).

Gene expression was measured using either HGU-95A or HGU-95Av2 Affymetrix arrays with $12,600$ genes shared between the two platforms ($12,626$ on the HG-U95A and $12,625$ on the HGu-95Av2). Six of the $10 \times 3 \times 3$ combinations were missing, leading to 84 samples.

Gagnon-Bartsch and Speed [2012] used the resulting dataset to study the performances of RUV-2: the number of genes from the X and Y chromosomes which were among the most differentially expressed genes between male and female patients was used to assess how much each correction method helped. Following this paper, we pre-processed each array using RMA, and log transformed the probe intensities.

This data package also provides negative control probeset indices. These indices correspond to the 799 housekeeping probesets which were provided in Eisenberg and Levanon [2003] and used in Gagnon-Bartsch and Speed [2012].

The data in this package is used in the vignette and examples of the *RUVnormalize* package. *RUVnormalize* implements normalization methods from Jacob et al. [2012], intended for the case where neither the unwanted variation sources nor the factors of interest are observed. This situation arises when performing unsupervised estimation tasks such as clustering or PCA, in the presence of unwanted variation. It can also be the case that one needs to normalize a dataset without knowing which factors of interest will be studied. The objective is then to correct the gene expression by estimating and removing the unwanted variation, without removing the — unobserved — variation of interest.

2 Object

The package contains a single *ExpressionSet* object `gender` which describes the data from Vawter et al. [2004].

The assayData field contains the 12600×84 gene expression matrix.

The phenoData field contains an *AnnotatedDataFrame* object describing the samples. The first column indicates the gender ('F' for female, 'M' for male). The next three columns indicate the lab: a one in the second, third or fourth column indicates that the sample was hybridized and scanned at UC Davis, UC Irvine or University of Michigan, Ann Arbor respectively. The last three columns contain brain regions. A one in the fifth, sixth or seventh column indicates that the sample was extracted from the anterior cingulate cortex, cerebellum or dorsolateral prefrontal cortex respectively.

The featureData field contains an *AnnotatedDataFrame* object with a single logical vectors indicating which probesets where used as negative controls in Gagnon-Bartsch and Speed [2012].

The annotation field indicates the chip type, among HGU-95A and HGU-95Av2 Affymetrix arrays.

3 Session Information

```
R version 4.6.2 (2026-10-29)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 26.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.32.so;  LAP

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: Etc/UTC
tzcode source: system (glibc)

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

```
loaded via a namespace (and not attached):
 [1] compiler_4.6.2   cli_3.6.6        tools_4.6.2      maketools_1.3.2
 [5] buildtools_1.0.0 knitr_1.52       xfun_0.62        rlang_1.3.0
 [9] sys_3.4.3        evaluate_1.0.5
```

References

Eli Eisenberg and Erez Y Levanon. Human housekeeping genes are compact. *Trends Genet*, 19 (7):362–365, Jul 2003. doi: 10.1016/S0168-9525(03)00140-9. URL http://dx.doi.org/10.1016/S0168-9525(03)00140-9.

Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, Jul 2012. ISSN 1468-4357. doi: 10.1093/biostatistics/kxr034. URL http://dx.doi.org/10.1093/biostatistics/kxr034.

L. Jacob, J. Gagnon-Bartsch, and T. P. Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Technical report, arXiv, 2012. URL http://arxiv.org/abs/1211.4259.

Marquis P. Vawter, Simon Evans, Prabhakara Choudary, Hiroaki Tomita, Jim Meador-Woodruff, Margherita Molnar, Jun Li, Juan F. Lopez, Rick Myers, David Cox, Stanley J. Watson, Huda Akil, Edward G. Jones, and William E. Bunney. Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology*, 29(2):373–384, Feb 2004. ISSN 0893-133X. doi: 10.1038/sj.npp.1300337. URL http://dx.doi.org/10.1038/sj.npp.1300337.