Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data

Supplemental Material

Markus List[*1,2,3], Anne-Christin Hauschild[4], Torben A. Kruse[1,3], Qihua Tan[3,5], Jan Mollenhauer[1,2], Jan Baumbach[†6], and Richa Batra[†6]

[1]Lundbeckfonden Center of Excellence in Nanomedicine (NanoCAN), University of Southern Denmark, Odense, Denmark
[2]Institute of Molecular Medicin (IMM), University of Southern Denmark, Odense, Denmark
[3]Clinical Institute (CI), University of Southern Denmark, Odense, Denmark
[4]Computational Systems Biology Group, Max Planck Institute for Informatics, Saarbrücken, Germany
[5]Epidemiology, Biostatistics and Biodemography, Institute of Public Health, University of Southern Denmark, Odense, Denmark
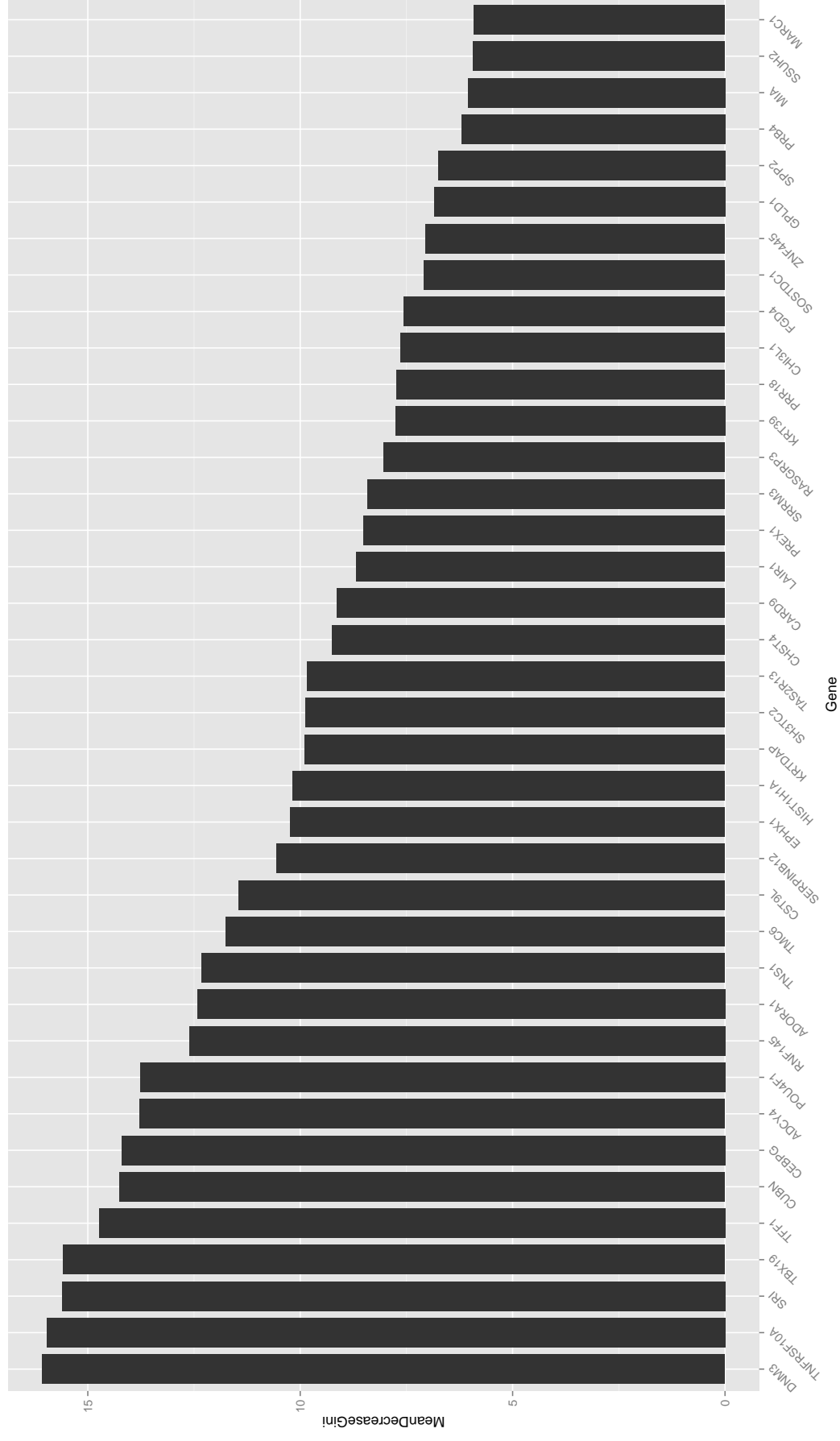[6]Department of Mathematics and Computer Science (IMADA), University of Southern Denmark, Odense, Denmark
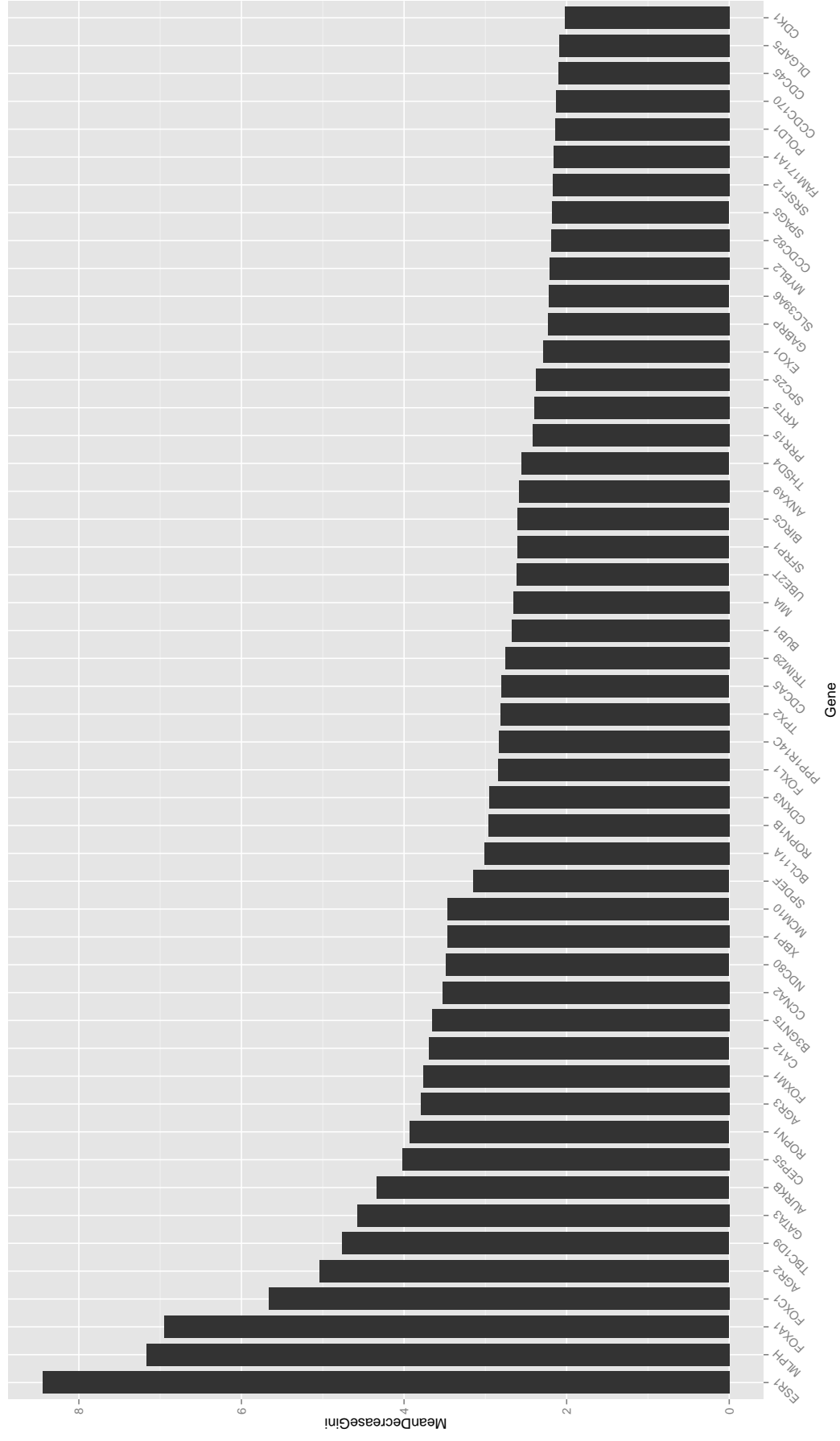
[*]mlist@health.sdu.dk
[†]joint first authors
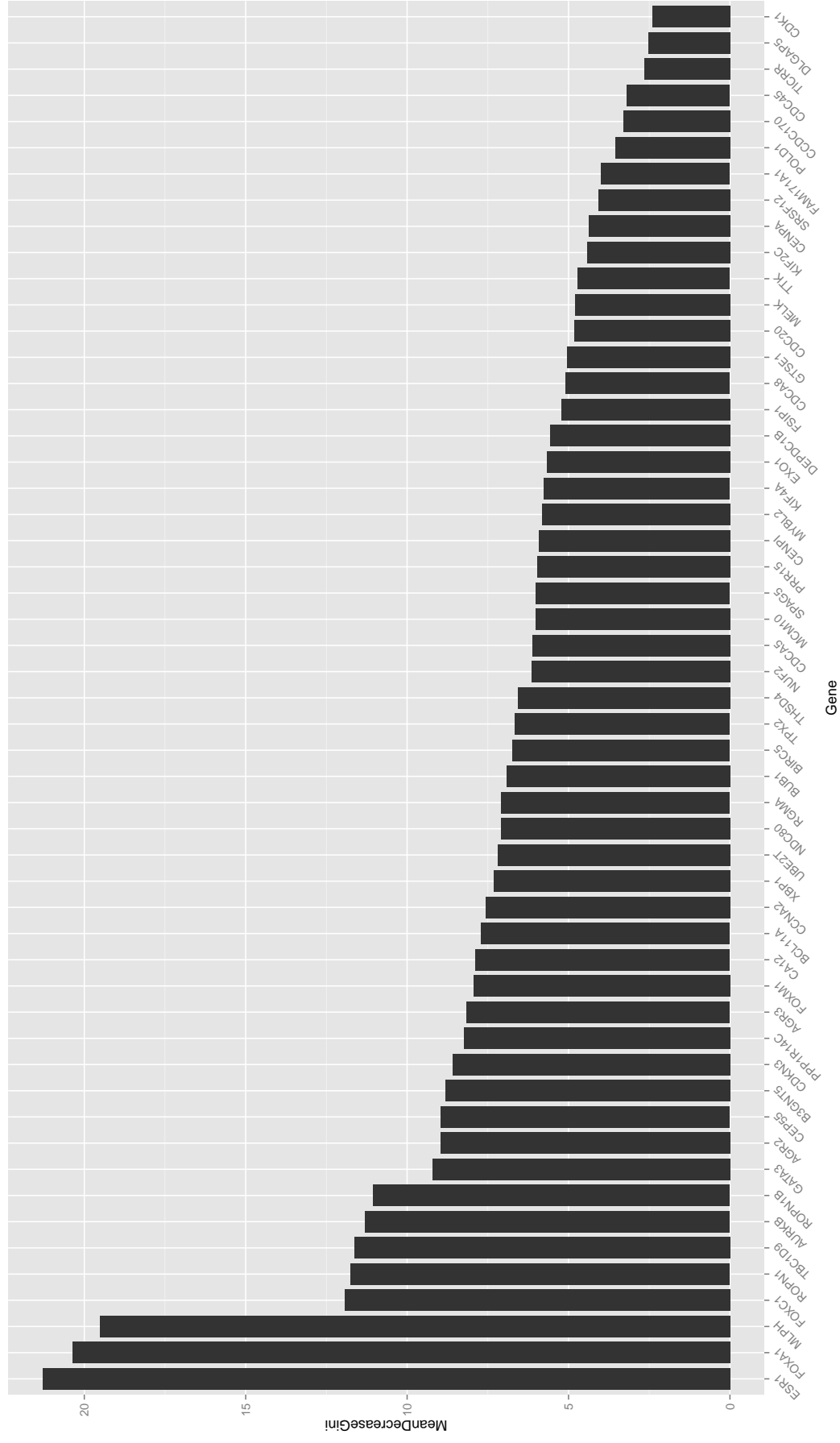
Supplemental Figure 1: All 53 selected features of the *gene expression model.*
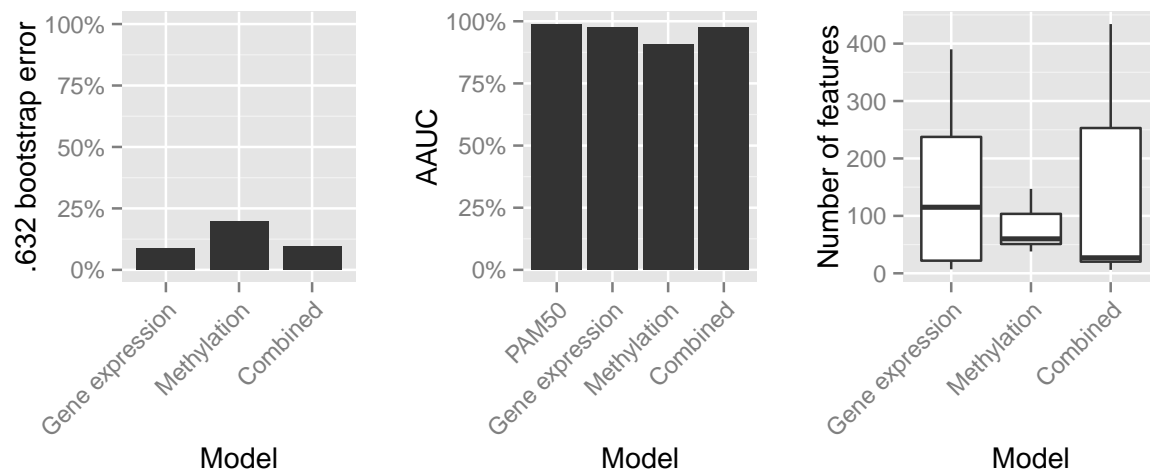
Supplemental Figure 2: All 38 selected features of the *methylation model.*

Supplemental Figure 3: The top 50 features of the *combined model* that have a mean decrease in gini index > 2.

Results when omitting all samples labelled as normal

Supplemental Figure 4: The .632 bootstrap error over 10 iterations (left), the AAUC (middle), and the distribution of the number of variables across bootstrap iterations (right) after applying random forests to the different models. Here, samples labelled as normal have been excluded from model creation. For comparison, the AAUC of the *control model* is also shown.
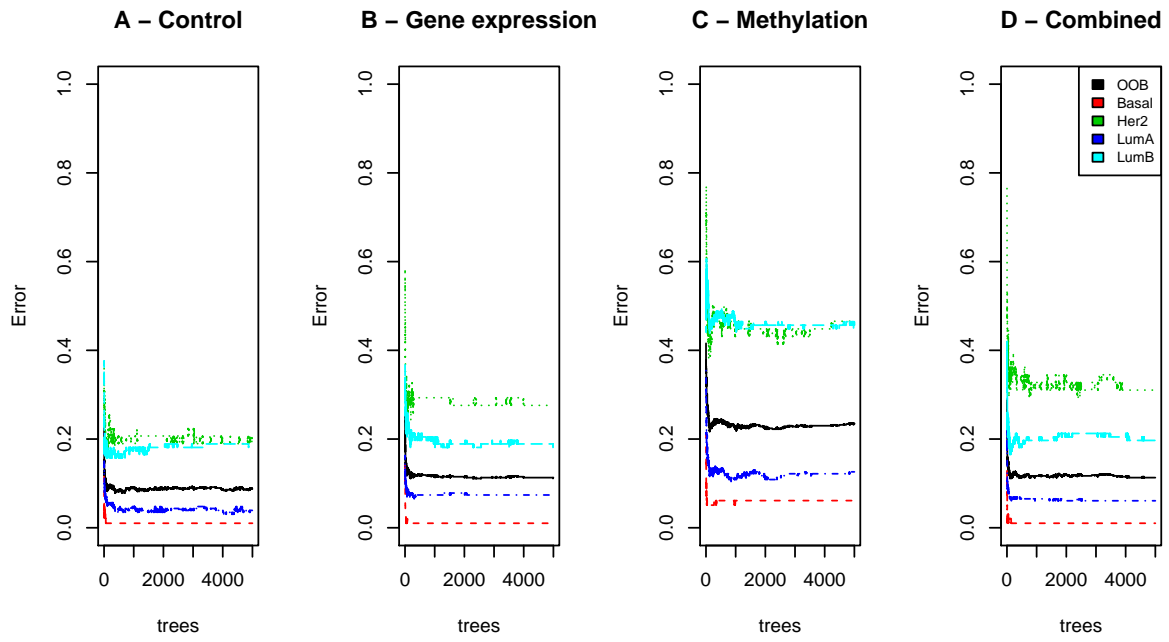
Supplemental Figure 5: The classification and out of bag error (OOB) of each subtype for the *control model* (A), the *gene expression model* (B), the *methylation model* (C) and the *combined model* (D) excluding samples labelled by PAM50 as normal and depending on the number of trees already created in the feature selection process.