When is hub gene selection better than standard meta-analysis? Supporting Text S1: supplementary figures

Peter Langfelder* and Steve Horvath
*Corresponding author: peter (.) langfelder (a) gmail (.) com

This document collects the supplementary figures (and their captions) for the main article *When is hub gene selection better than standard meta-analysis?*

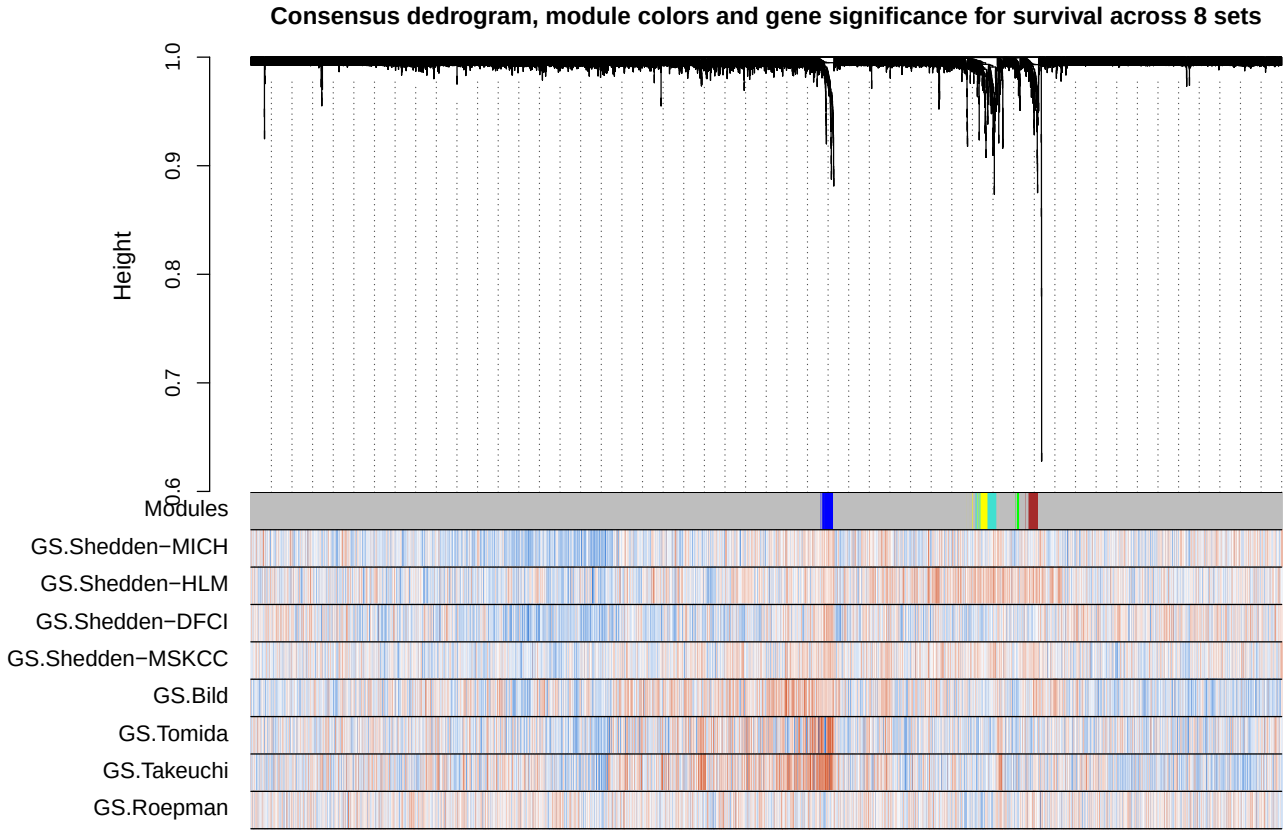

Figure S1. Gene clustering tree based on the consensus Topological Overlap similarity across **8 adenocarcinoma (lung cancer) data sets**. Each short vertical line ("leaf" of the tree) corresponds to a single gene. Branches of the clustering tree group together genes with high consensus similarity and hence define modules. The modules are indicated by colors below the clustering tree (line Modules). Grey color corresponds to genes not assigned to any of the modules. In this analysis, likely because of the poor reproducibility among the 8 data sets, we only find 5 small modules and most genes remain unassigned. The 8 color rows below the module indicator indicate the gene significance, defined as the correlation of survival time deviance with gene expression. Green color denotes negative correlation with survival deviance (i.e., the gene over-expression is associated with lower risk of death) while red color denotes positive correlation with survival deviance (i.e., the gene over-expression is associated with higher risk of death). This representation makes apparent the poor reproducibility of gene significance for survival time between the 8 data sets.

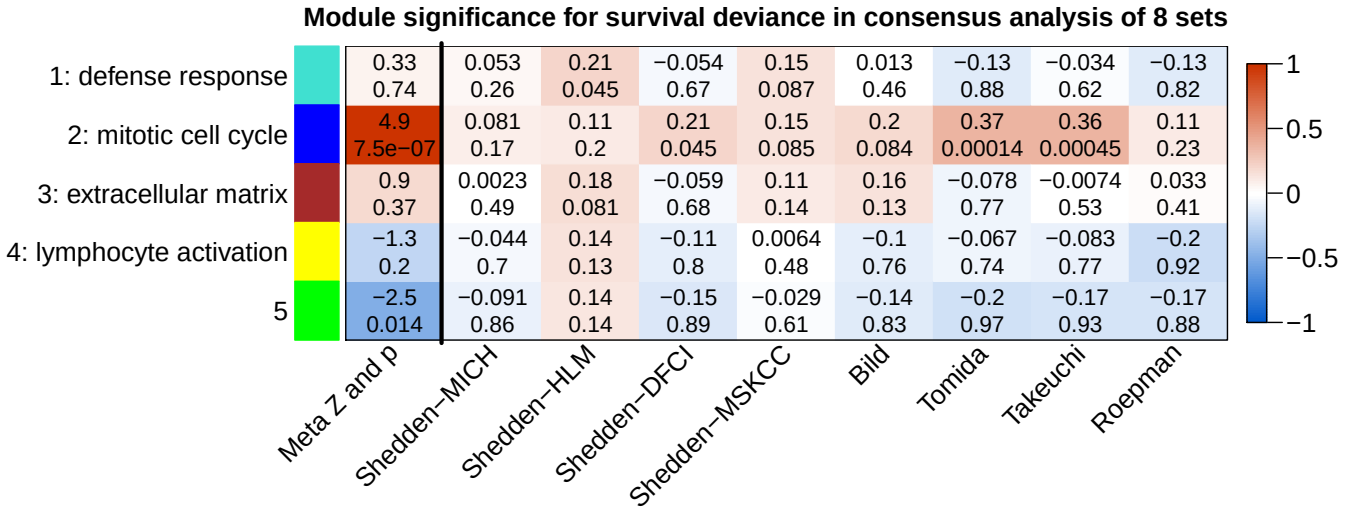

Figure S2. Module significance for survival deviance in consensus analysis of 8 adenocarcinoma expression data sets. Each row of this table corresponds to one module identified in the consensus module analysis. Modules are labeled by the module number and (where appropriate) a GO-derived functional label. Columns 2–9 give correlations between the survival time deviance and the module eigengenes and the associated p-values. The first column shows the meta-analysis Z statistic and the corresponding p-value. We observe that module 2 exhibits a relatively weak but consistent association with survival time deviance across the 8 data sets. The meta-analysis Z score and p-value for module 2 are significant. The significant association and cell cycle-related functional annotation We note that while module 5 also attains nominal significance at the 0.05 level, a Bonferroni correction for the number of tested modules (5) would render the p-value non-significant. Functional enrichment analysis (Supplementary Table ??) did not reveal strong enrichment in biologically plausible categories, and we do not consider this module in the following.

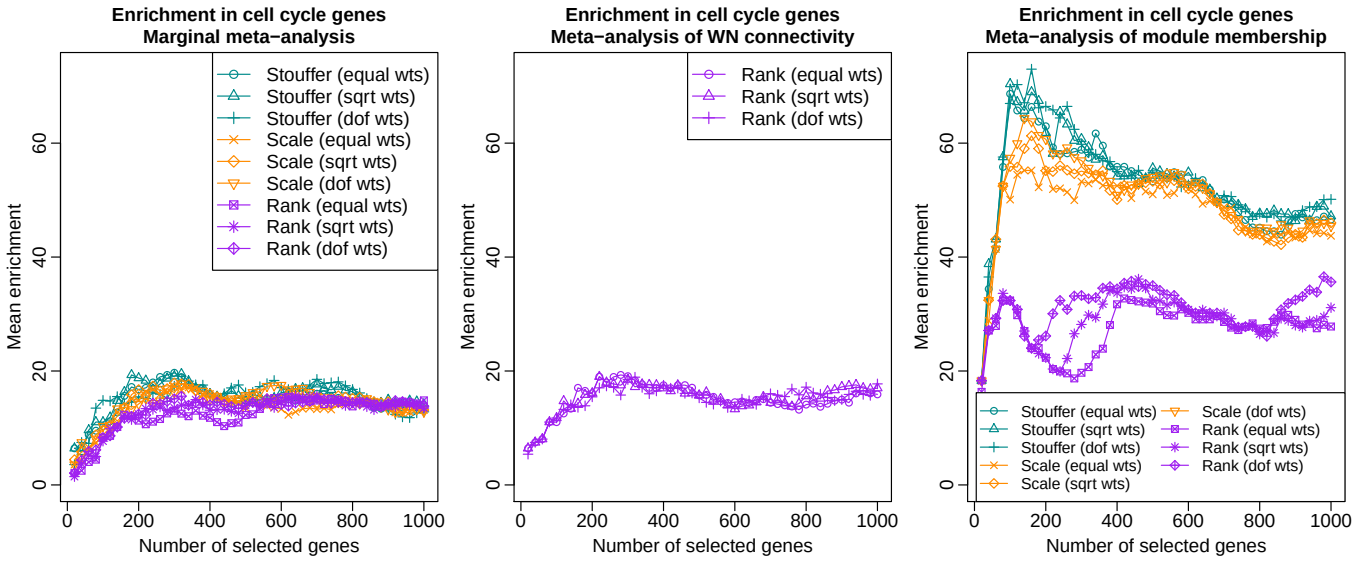

Figure S3. Enrichment score (defined as $-log_{10}(p)$ with p being the enrichment p-value) in GO term "cell cycle" as a function of the number of top selected genes by a variety of meta-analysis (MA) methods applied to **8 adenocarcinoma (lung cancer) data sets**. The left panel shows enrichment of genes selected by 6 marginal MA methods, while the right panel shows the analogous enrichment of genes selected by MA of module membership in consensus modules. Clearly, MA of module membership in consensus modules selects genes with much higher enrichment in GO term "cell cycle".

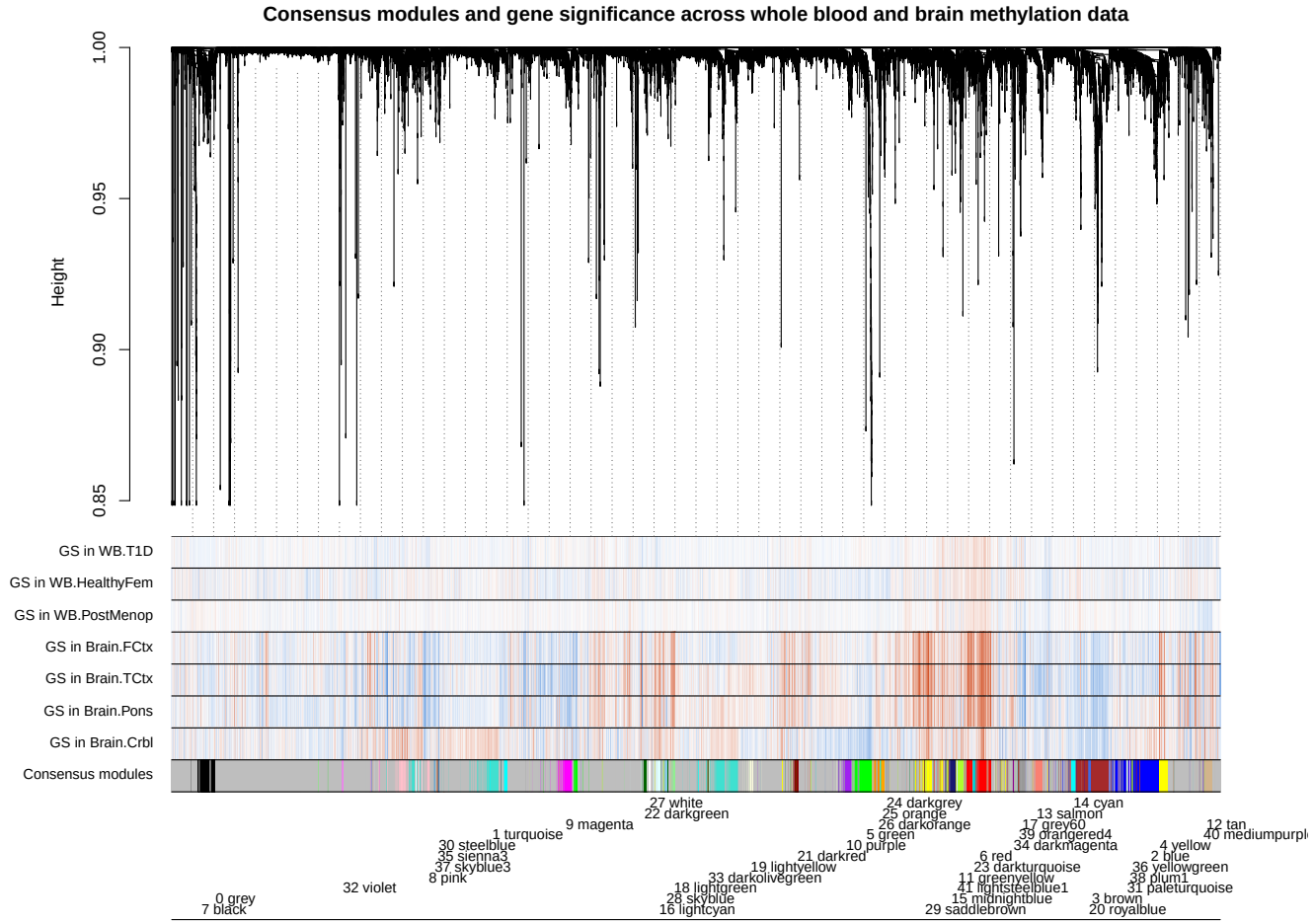

Figure S4. Gene clustering tree based on the consensus Topological Overlap similarity across **7 methylation data sets**. Each short vertical line ("leaf" of the tree) corresponds to a single gene. Branches of the clustering tree group together genes with high consensus similarity and hence define modules. The modules are indicated by colors below the clustering tree (line Modules). Grey color corresponds to genes not assigned to any of the modules. In this analysis we find a relatively large number (41) of modules. The 7 color rows above the module indicator indicate the gene significance, defined as the correlation of age with probe methylation. Blue color denotes negative correlation with age while red color denotes positive correlation with age.

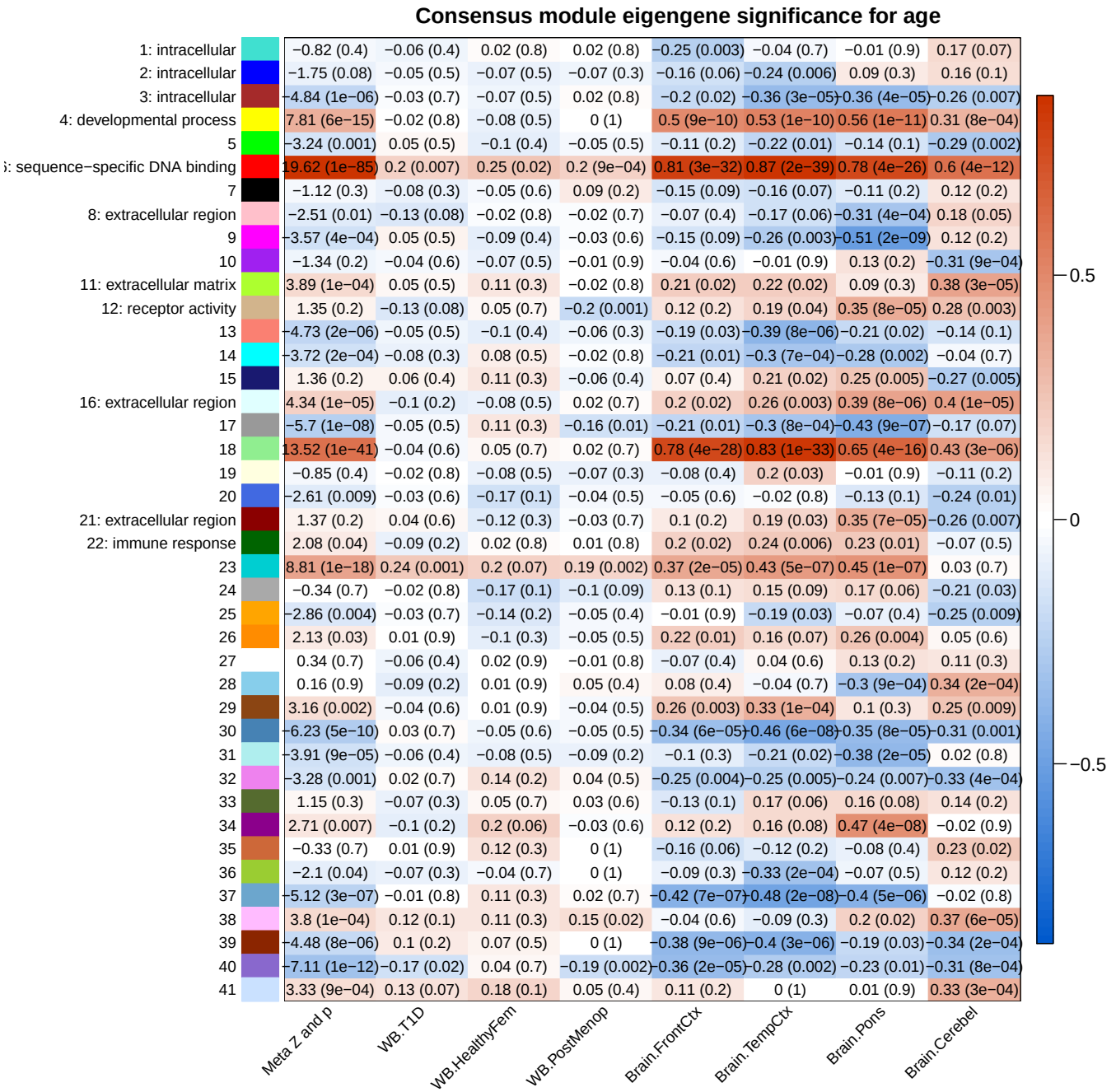

Figure S5. Consensus eigengene significance for age across 7 genome-wide methylation data sets. Each row of this table corresponds to one module identified in the consensus module analysis. Modules are labeled by the module number and (where appropriate) a GO-derived functional label. Columns 2–8 give robust correlations between age and the module eigengenes and the associated p-values. The first column shows the meta-analysis Z statistic and the corresponding p-value. We observe that several modules attain strong statistical significance in the meta-analysis of the 7 data sets. The most strongly associated module is module 6. The association between age and the module eigengene is consistent in all 7 data sets, lending further credence to the meta-analysis result. Hence, we focus our analysis on this module.

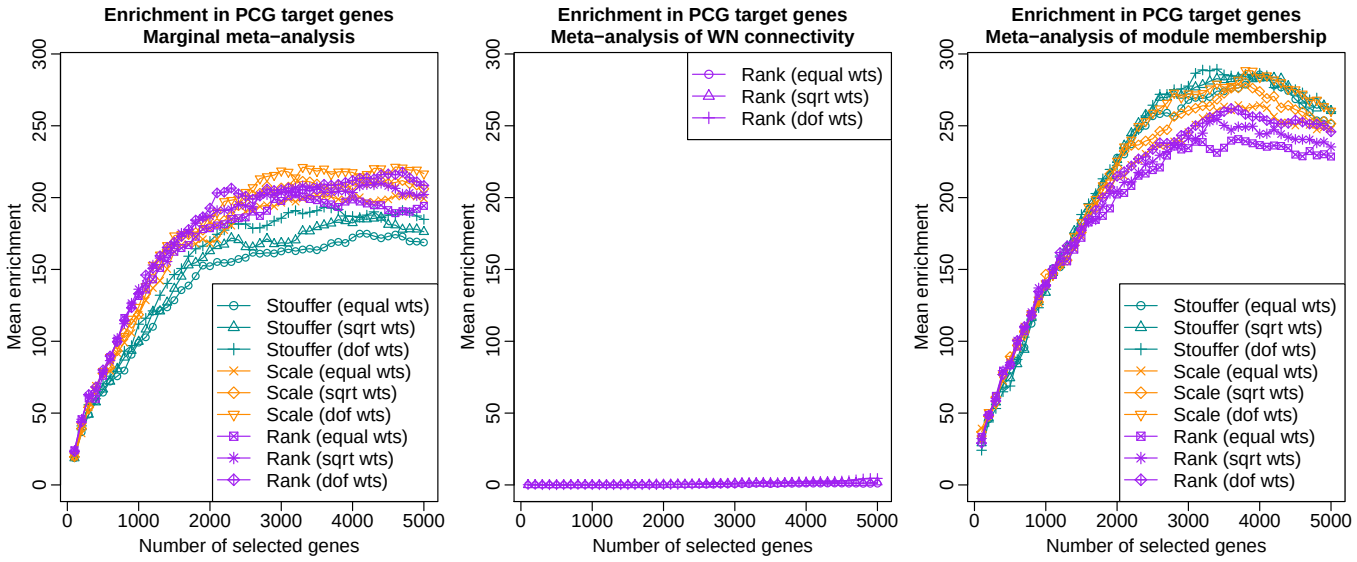

Figure S6. Enrichment score (defined as $-log_{10}(p)$ with p being the Bonferroni-corrected enrichment p-value) in Polycomb group (PCG) target genes as a function of the number of top selected genes by a variety of meta-analysis (MA) methods applied to **7 methylation data sets**. The left panel shows enrichment of genes selected by 6 marginal MA methods, while the right panel shows the analogous enrichment of genes selected by MA of module membership in consensus modules. Although marginal MA methods results in gene sets with very strong enrichment, meta-analysis of module membership in modules selects genes with somewhat higher enrichment.

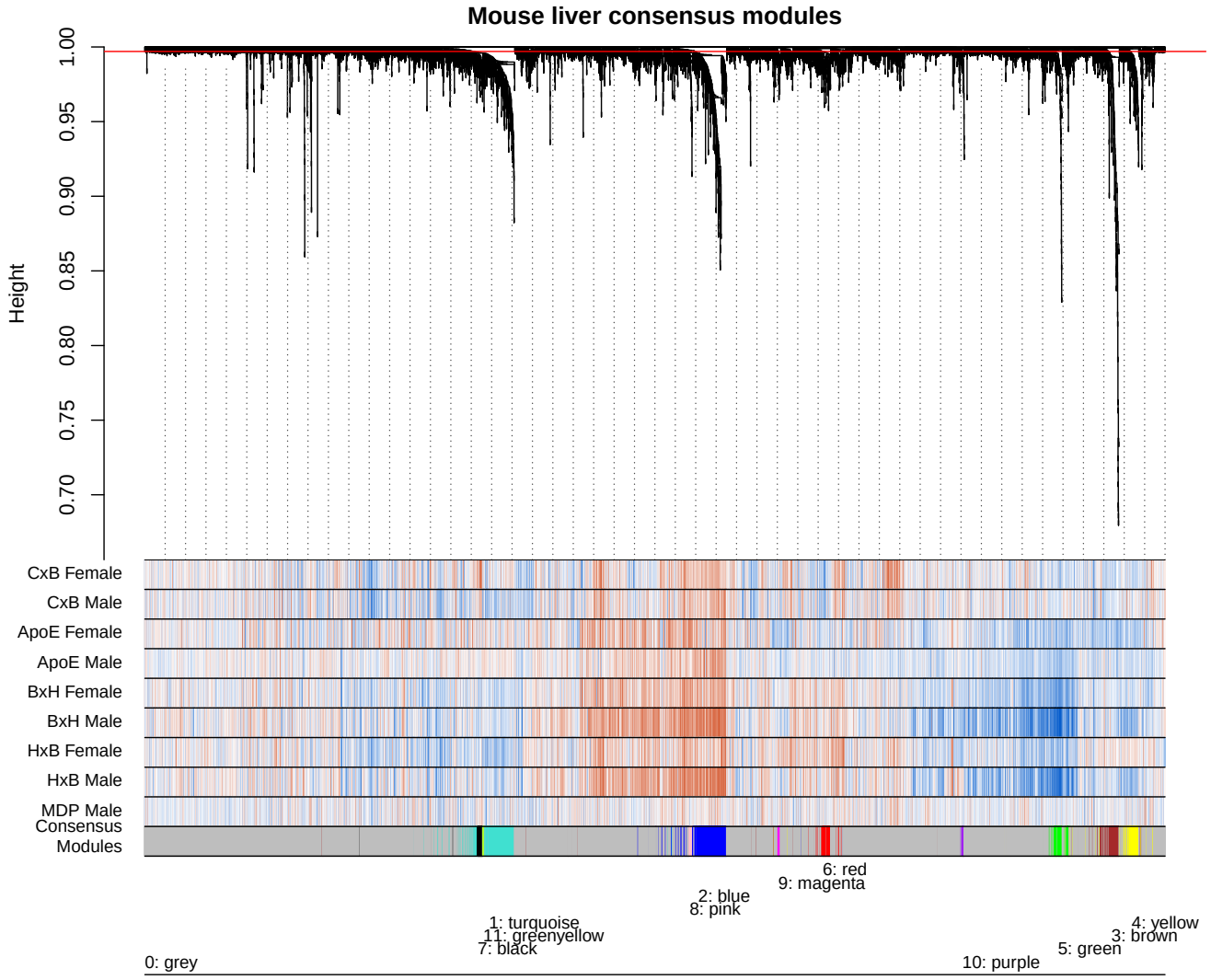

Figure S7. Gene clustering tree based on the consensus Topological Overlap similarity across **9 mouse liver expression data sets**. Each short vertical line ("leaf" of the tree) corresponds to a single gene. Branches of the clustering tree group together genes with high consensus similarity and hence define modules. The modules are indicated by colors below the clustering tree (line Modules). Grey color corresponds to genes not assigned to any of the modules. In this analysis we find a moderate number (11) of modules. The 9 color rows above the module indicator indicate the gene significance, defined as the correlation of total cholesterol with gene expression. Blue color denotes negative correlation with total cholesterol while red color denotes positive correlation with total cholesterol.

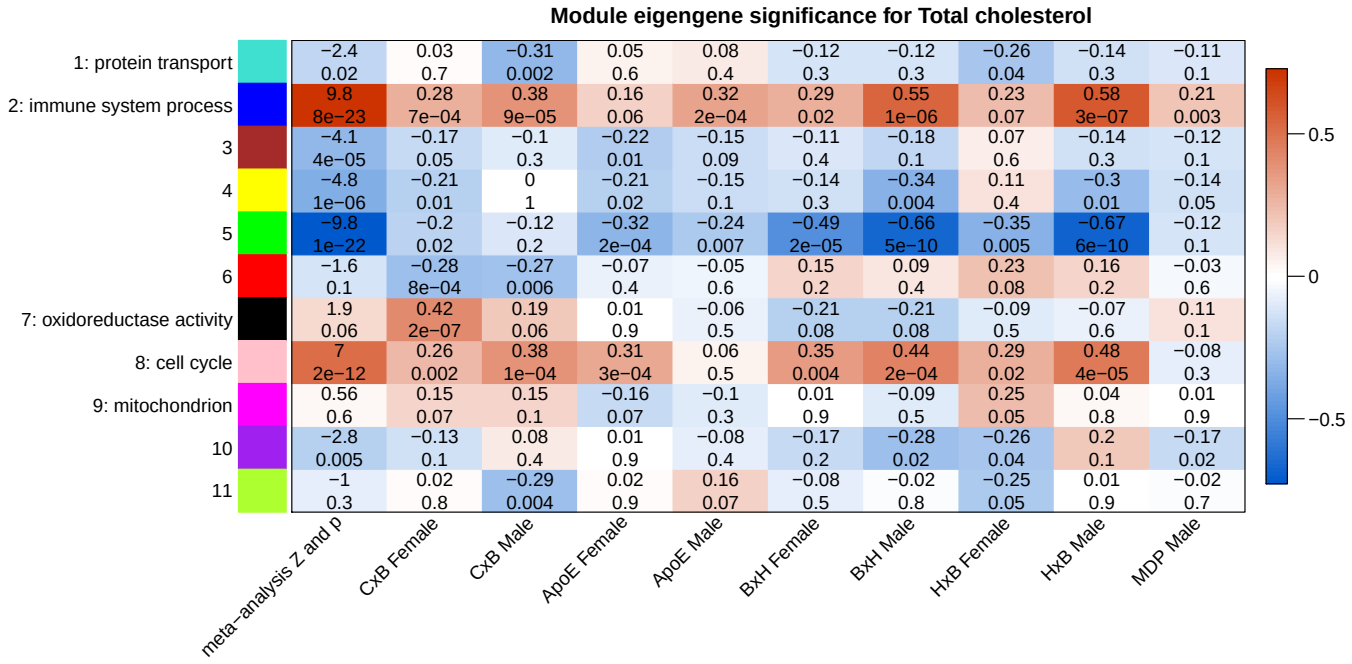

Figure S8. Consensus eigengene significance for age across 9 mouse liver expression data sets. Each row of this table corresponds to one module identified in the consensus module analysis. Modules are labeled by the module number and (where appropriate) a GO-derived functional label. Columns 2–10 give robust correlations between total cholesterol and the module eigengenes as well as the associated p-values. The first column shows the meta-analysis Z statistic and the corresponding p-value. We observe that several modules attain strong statistical significance in the meta-analysis of the 9 data sets. The most strongly associated module is module 2. The association between age and the module eigengene is consistent in all 9 data sets, lending further credence to the meta-analysis result. Hence, we focus our analysis on this module.

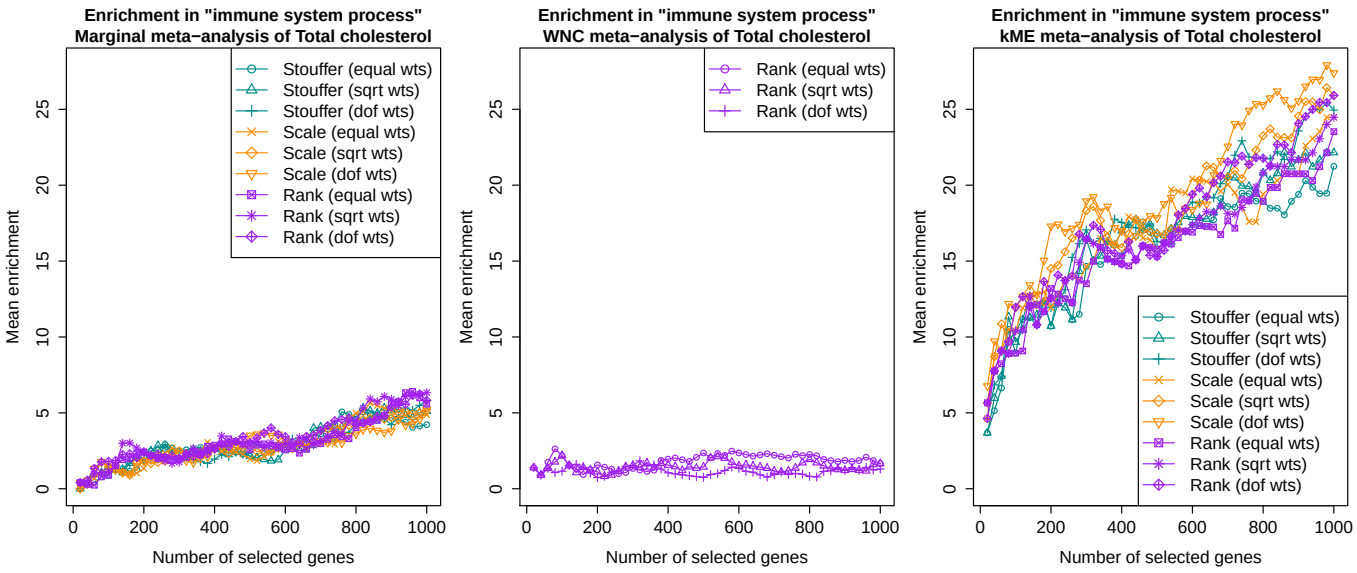

Figure S9. Enrichment score (defined as $-log_{10}(p)$ with p being the Bonferroni-corrected enrichment p-value) in in GO term "immune system process" as a function of the number of top selected genes by a variety of meta-analysis (MA) methods applied to **9 mouse liver expression data sets**. The left panel shows enrichment of genes selected by 6 marginal MA methods, while the right panel shows the analogous enrichment of genes selected by MA of module membership in consensus modules. Meta-analysis of module membership in consensus modules selects genes with much higher enrichment than marginal meta-analysis methods.

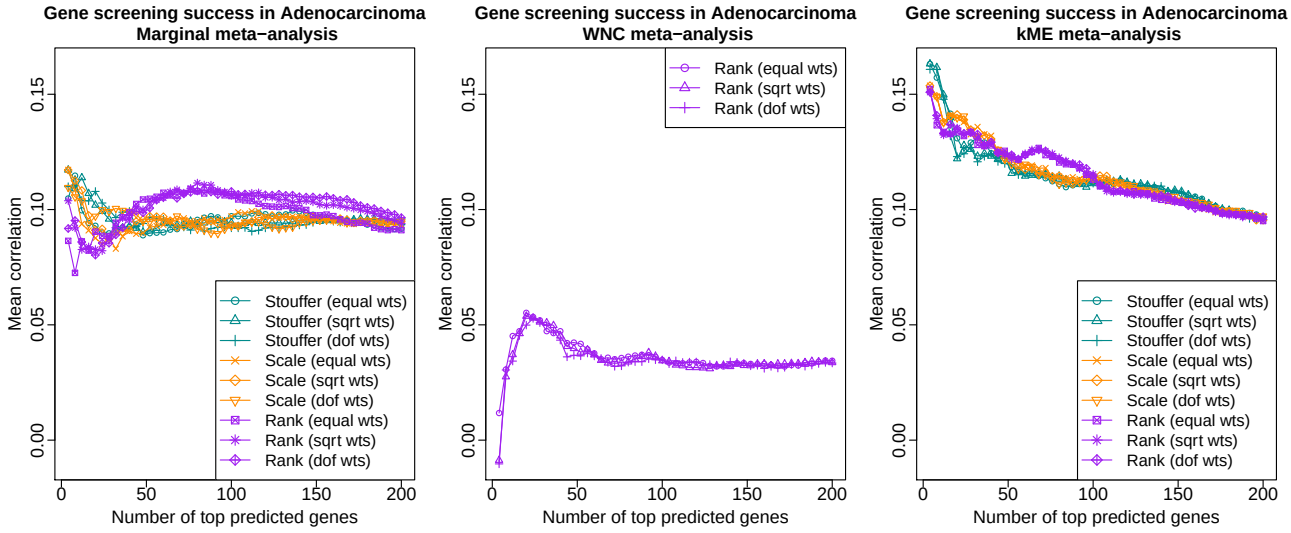

Figure S10. Validation success (y-axis), measured as the mean correlation of selected genes with survival time deviance in an independent validation data set, as a function of the number of top selected genes (x-axis) in the adenocarcinoma expression data (Application 1). In this application, meta-analysis of module membership (right) outperforms marginal meta-analysis (left).

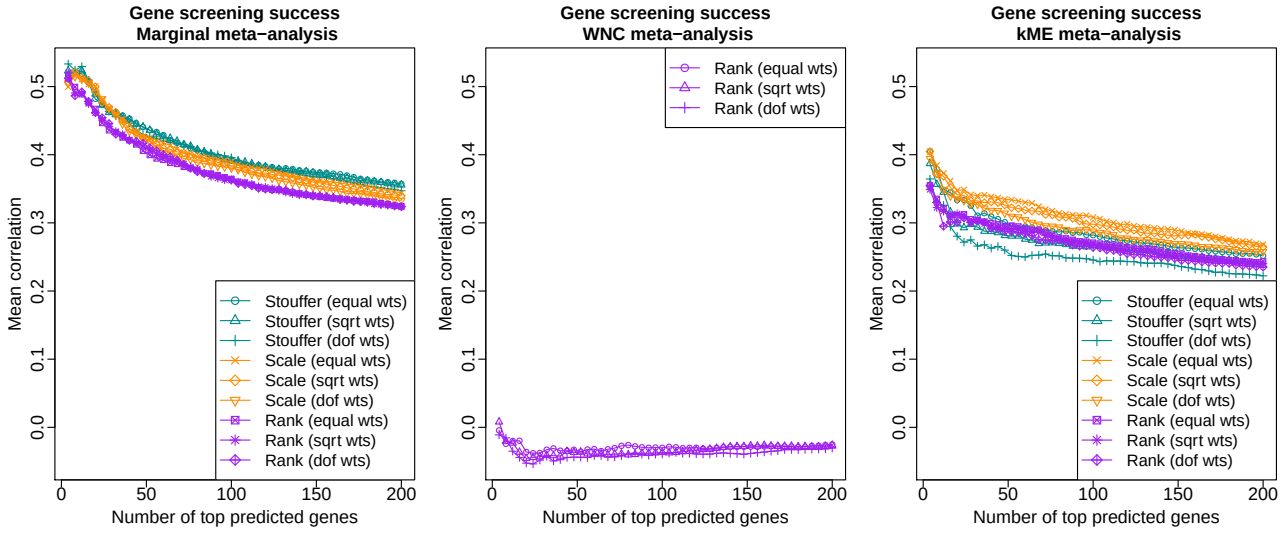

Figure S11. Validation success (y-axis), measured as the mean correlation of selected genes with survival time deviance in an independent validation data set, as a function of the number of top selected genes (x-axis) in the methylation data (Application 2). In this application, marginal meta-analysis (left) outperforms meta-analysis of module membership (right).

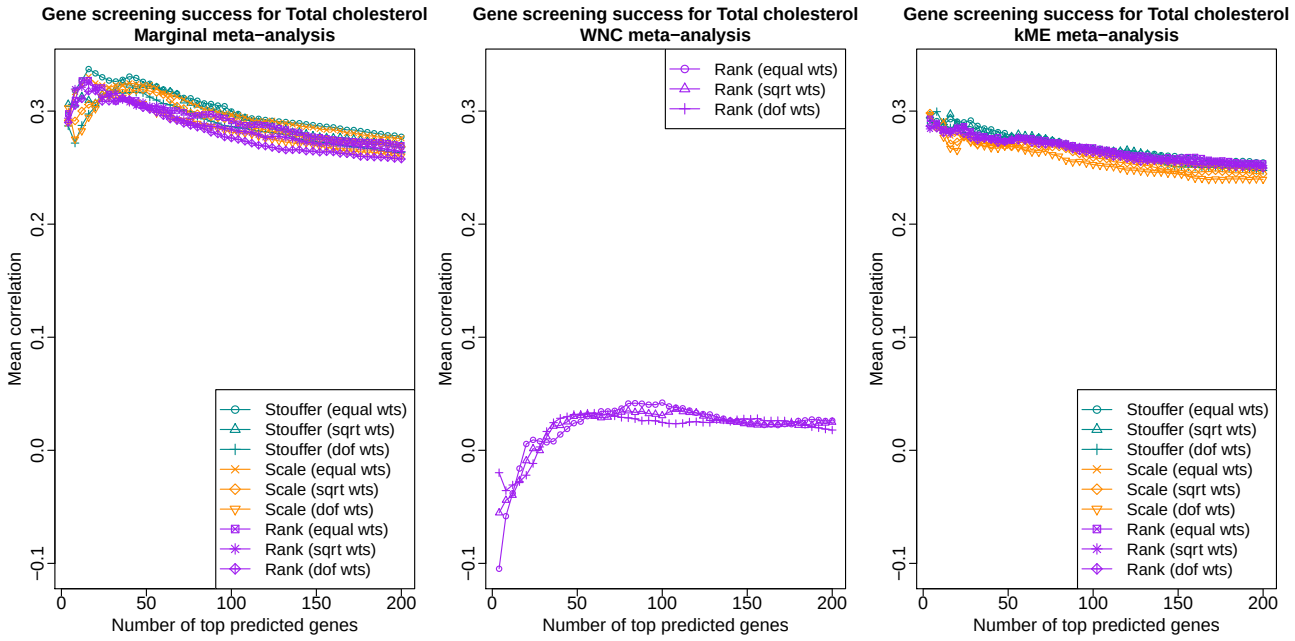

Figure S12. Validation success (y-axis), measured as the mean correlation of selected genes with survival time deviance in an independent validation data set, as a function of the number of top selected genes (x-axis) in the mouse liver expression data (Application 3). In this application, marginal meta-analysis (left) outperforms meta-analysis of module membership (right).

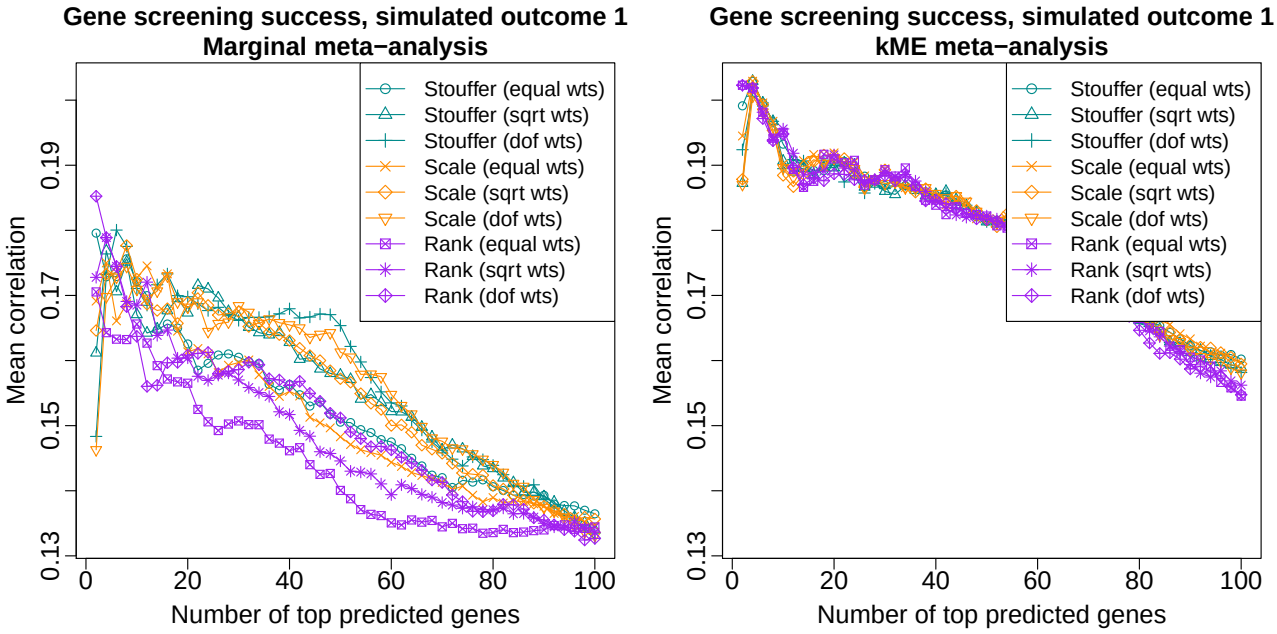

Figure S13. Validation success (y-axis), measured as the mean correlation of selected genes with clinical trait in an independent validation data set, as a function of the number of top selected genes (x-axis), for simulated clinical trait 1. In this simulation, meta-analysis of module membership (right) outperforms marginal meta-analysis (left).

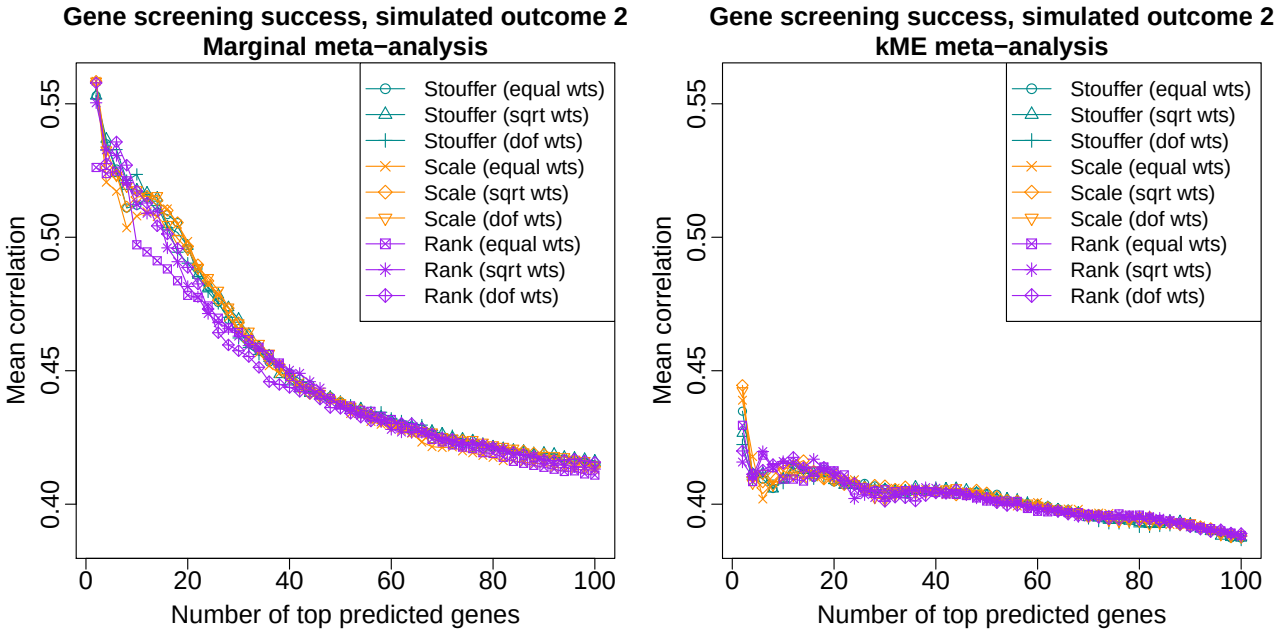

Figure S14. Validation success (y-axis), measured as the mean correlation of selected genes with clinical trait in an independent validation data set, as a function of the number of top selected genes (x-axis), for simulated clinical trait 2. In this simulation, marginal meta-analysis (left) outperforms meta-analysis of module membership (right).