Supplementary Information

Imputing Gene Expression to Maximize Platform Compatibility

Weizhuang Zhou[1,‡], Lichy Han[2,‡] and Russ B. Altman[1,3,*]

[1]Department of Bioengineering, [2]Biomedical Informatics Training Program, [3]Department of Genetics, Stanford University, Stanford, CA 94305, USA

[‡]Both authors contributed equally to this paper.

*To whom correspondence should be addressed.

Table of Contents

Figure S1: Venn diagram of the probe sets in the HG-U133A, HG-U133B, and HG-U133 Plus 2.0 platforms

The data was obtained from Affymetrix's technical note for the platform. (media.affymetrix.com/support/technical/technotes/hgu133_p2_technote.pdf)

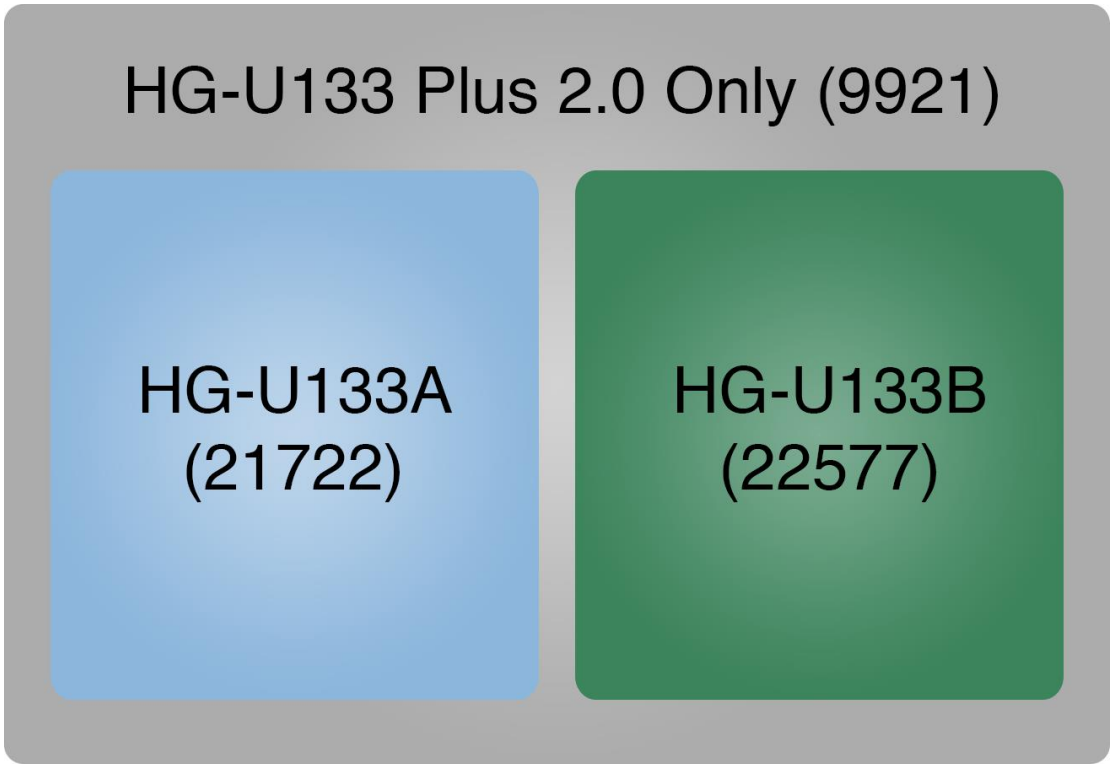
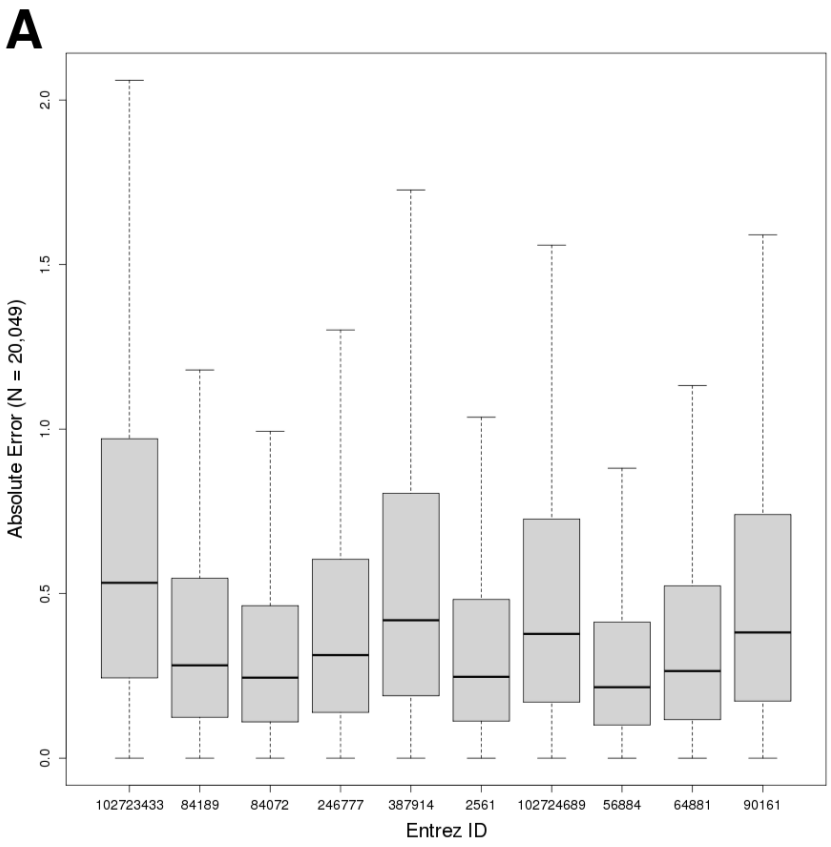


Figure S2: (A) Boxplots of squared errors for the ten genes with highest test CV(RMSE); (B) Boxplots of the normalized squared errors

Figure S2A shows the distribution of absolute errors for the test set (N = 20,049) for the top ten genes with the highest test CV(RMSE). Figure S2B shows the same with normalization, where the absolute errors are divided by the mean expression level for the gene. The boxplots show that the distributions of the errors are heavy-tailed. Apart from the first gene (Entrez ID: 102723433), the absolute errors of the remaining nine genes are less than 10% of the respective mean gene expression levels.



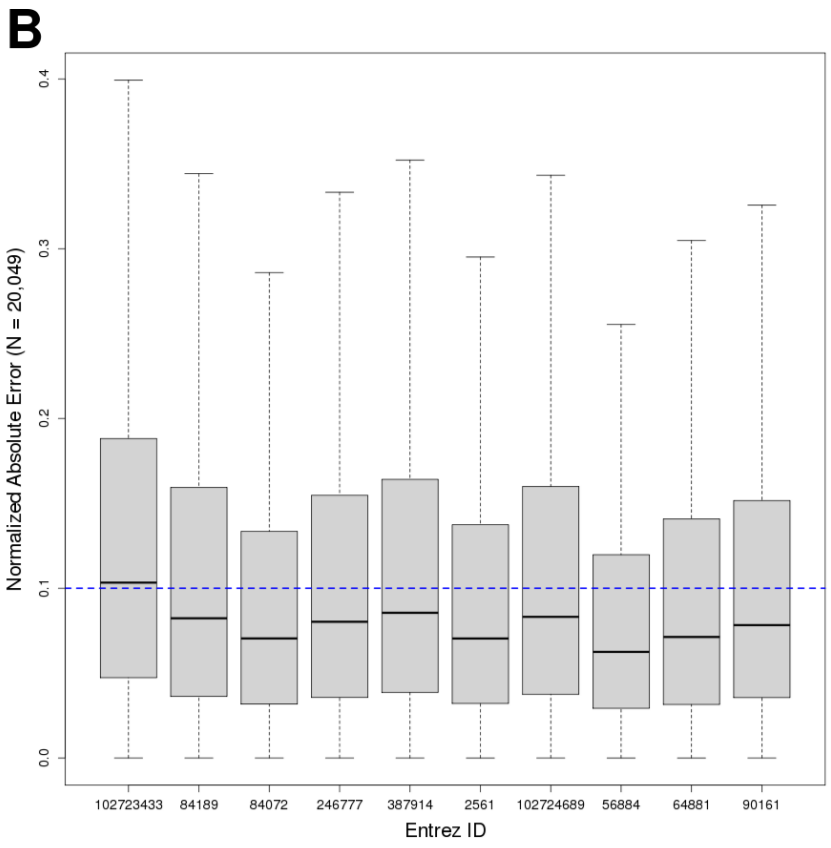
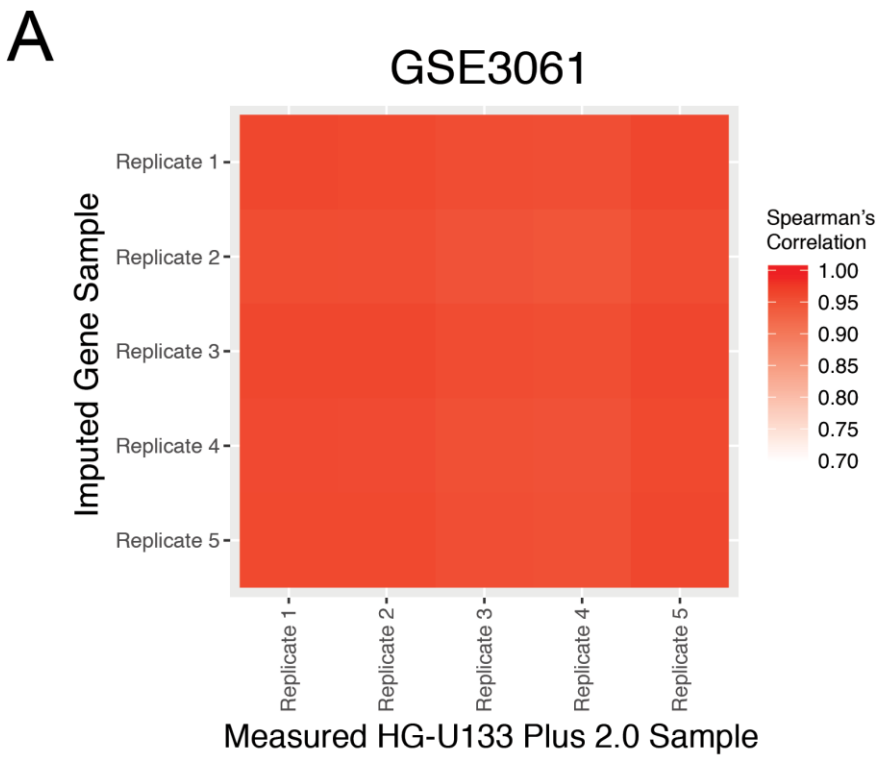


Figure S3: (A) Heatmap of Spearman's correlation coefficients between measured and imputed samples in GSE3061. (B) Gene-gene scatterplot for replicate 1.

GSE3061 contains five replicates of Stratagene's Universal Human Reference RNA measured on the HG-U133A and HG-U133 Plus 2.0 platforms. All replicates are highly inter-correlated (Spearman's $\rho = 0.96 \pm 0.005$, Figure S3).

Figure S3B plots the imputed array values against the actual measured values for replicate 1, with orange dots corresponding to genes that are measured on both platforms and blue dots corresponding to genes to the genes measured only on the HG-U133 Plus 2.0 platform.



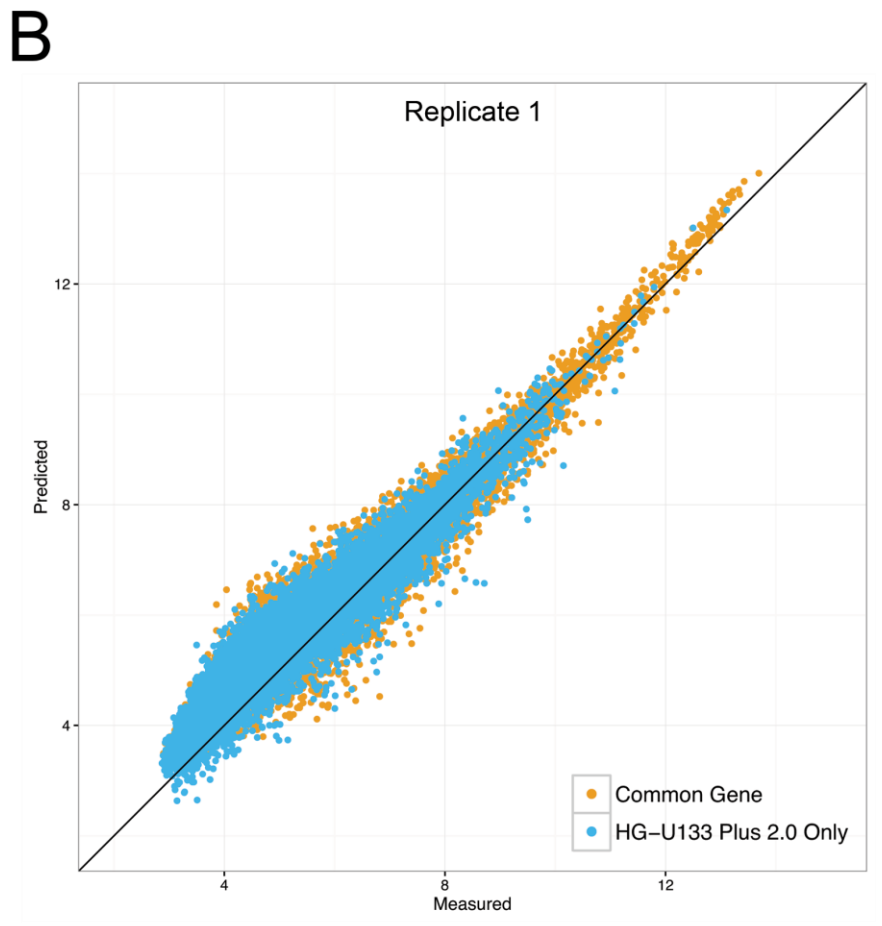


Figure S4: Heatmap of Spearman's correlation coefficients between measured and imputed samples in GSE17700

GSE17700 contains sixteen breast cancer samples measured using both the HG-U133A and the HG-U133 Plus 2.0 array. Additionally, all sixteen samples were measured at the Jules Bordet Institute and at the MD Anderson Cancer Center, for a total of 64 GSMs. As all samples are breast cancer in origin, modest correlation coefficients (Spearman's ρ = 0.86 ± 0.03) were observed when comparing different sample numbers, as shown by the off-diagonal elements in Figure S2. When comparing the same sample number, each imputed sample was consistently highly correlated with its corresponding measured sample (Spearman's ρ = 0.90 ± 0.012). Notably, the high correlation was preserved even when the sample was measured at a different institute.

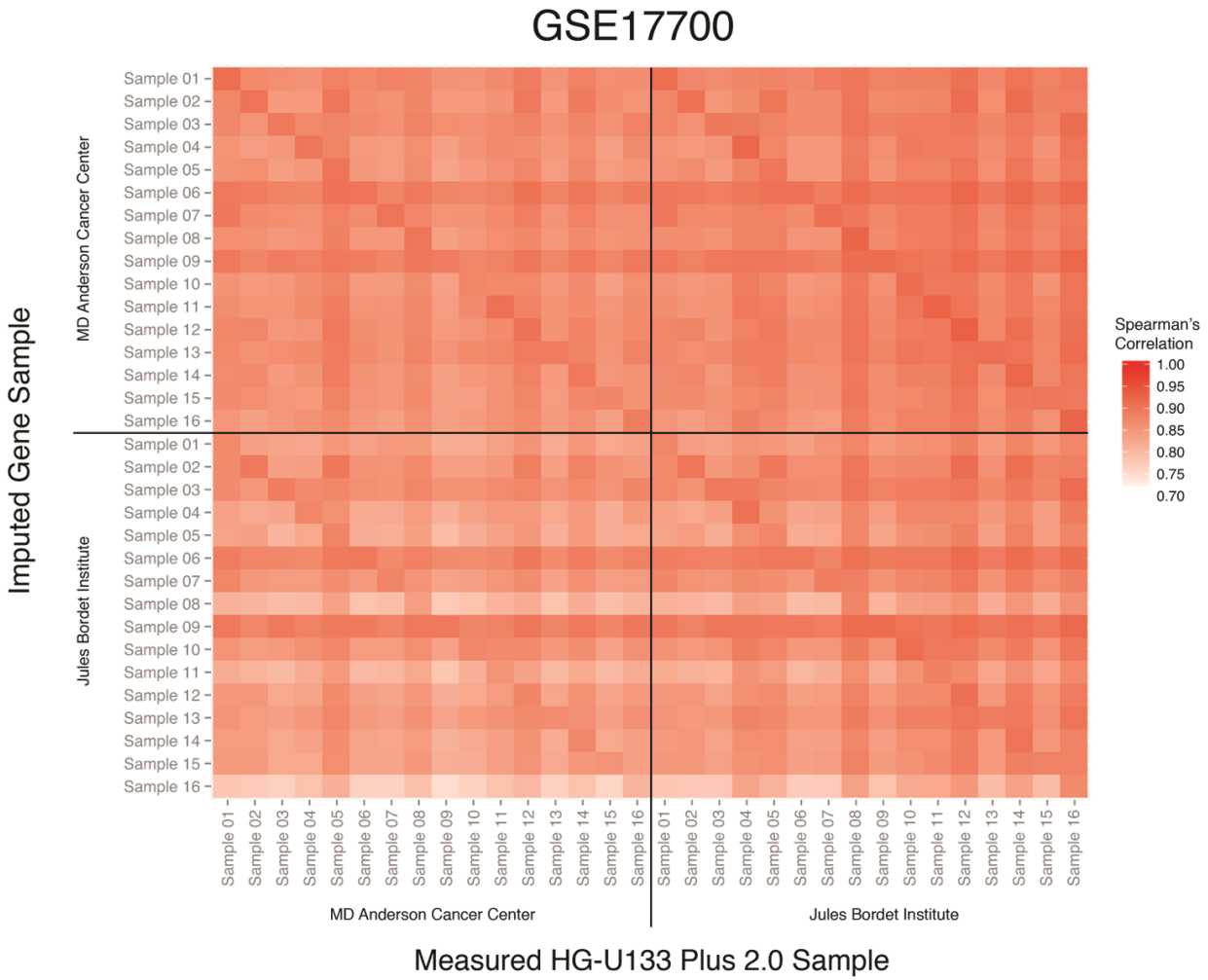
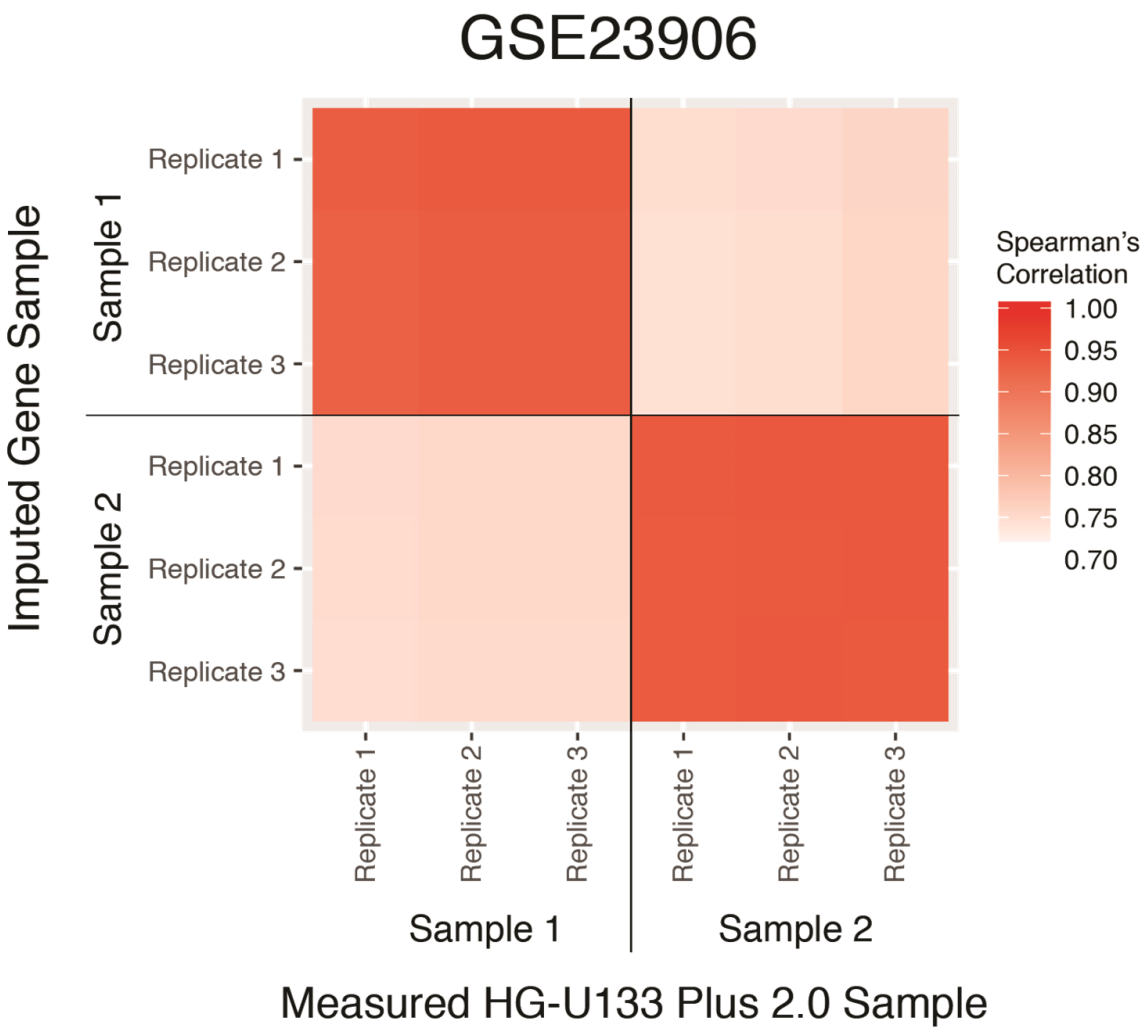


Figure S5: Heatmap of Spearman's correlation coefficients between measured and imputed samples in GSE23906.

GSE23906 studied the effects of using expired microarrays by comparing the results of expired HG-U133A arrays with previously published results and newly measured HG-U133 Plus 2.0 samples. The study used Stratagene's Universal Human Reference RNA (Sample 1) and Ambion's Human Brain Reference RNA (Sample 2), performing three replicates of each. We observed that imputed samples are highly correlated to their corresponding measured samples within the same sample type across all replicates (Spearman's ρ = 0.94 ± 0.004, Figure S4).



Additional Details on MSigDB Signatures' Coverage (Fig S6-S13)

For each gene signature in a particular MSigDB collection, a yellow/blue dot represents the percentage of genes found on the HG-U133 Plus 2.0/A platform. The two vertical red lines are the default limits for a valid signature, and the black curve is the minimum coverage percentage required in order for a signature to be retained for GSEA. Anything below the black curve is rejected from GSEA by default.

Figure S6: MSigDB C1 collection

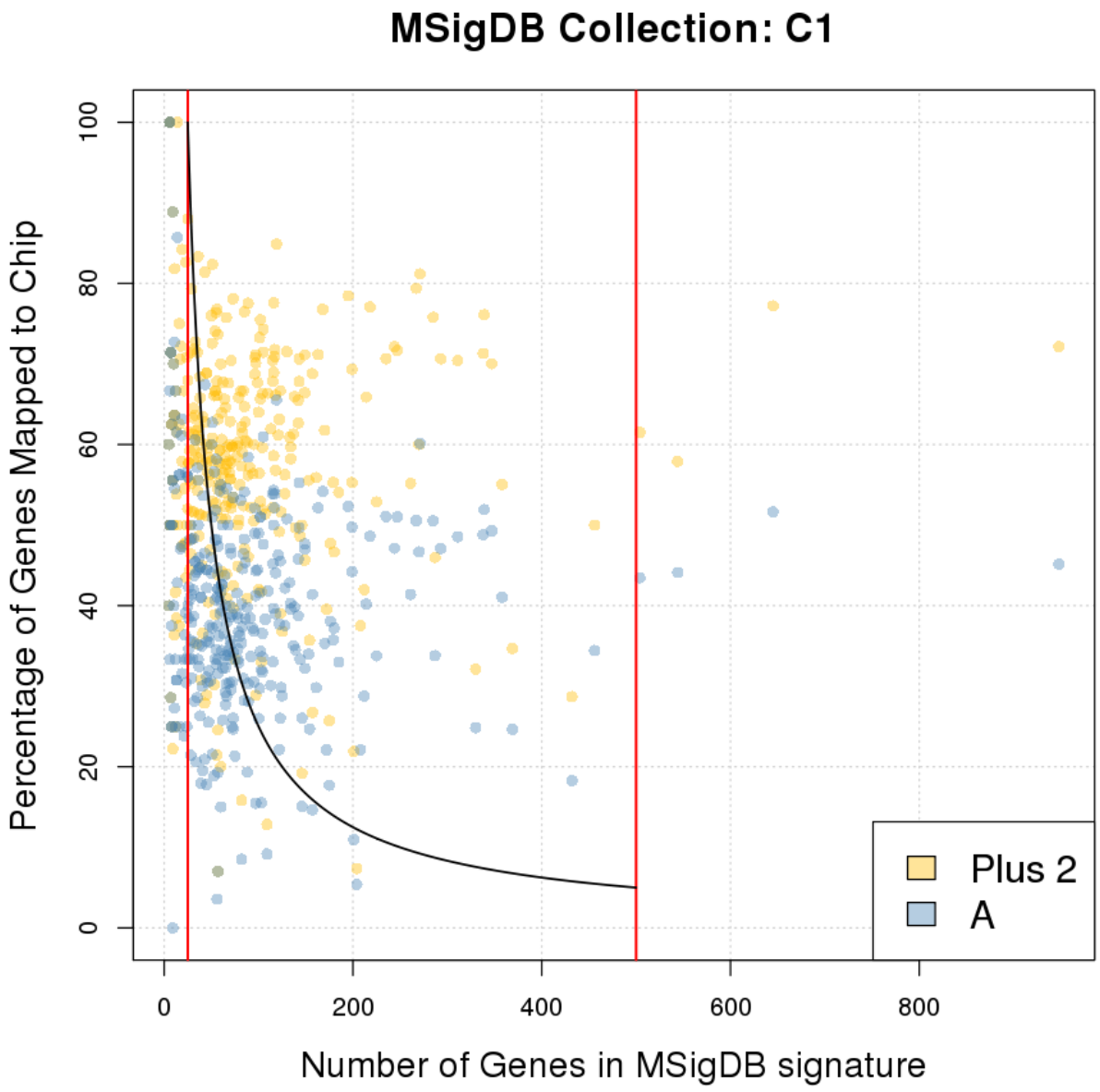


Figure S7: MSigDB C2 collection

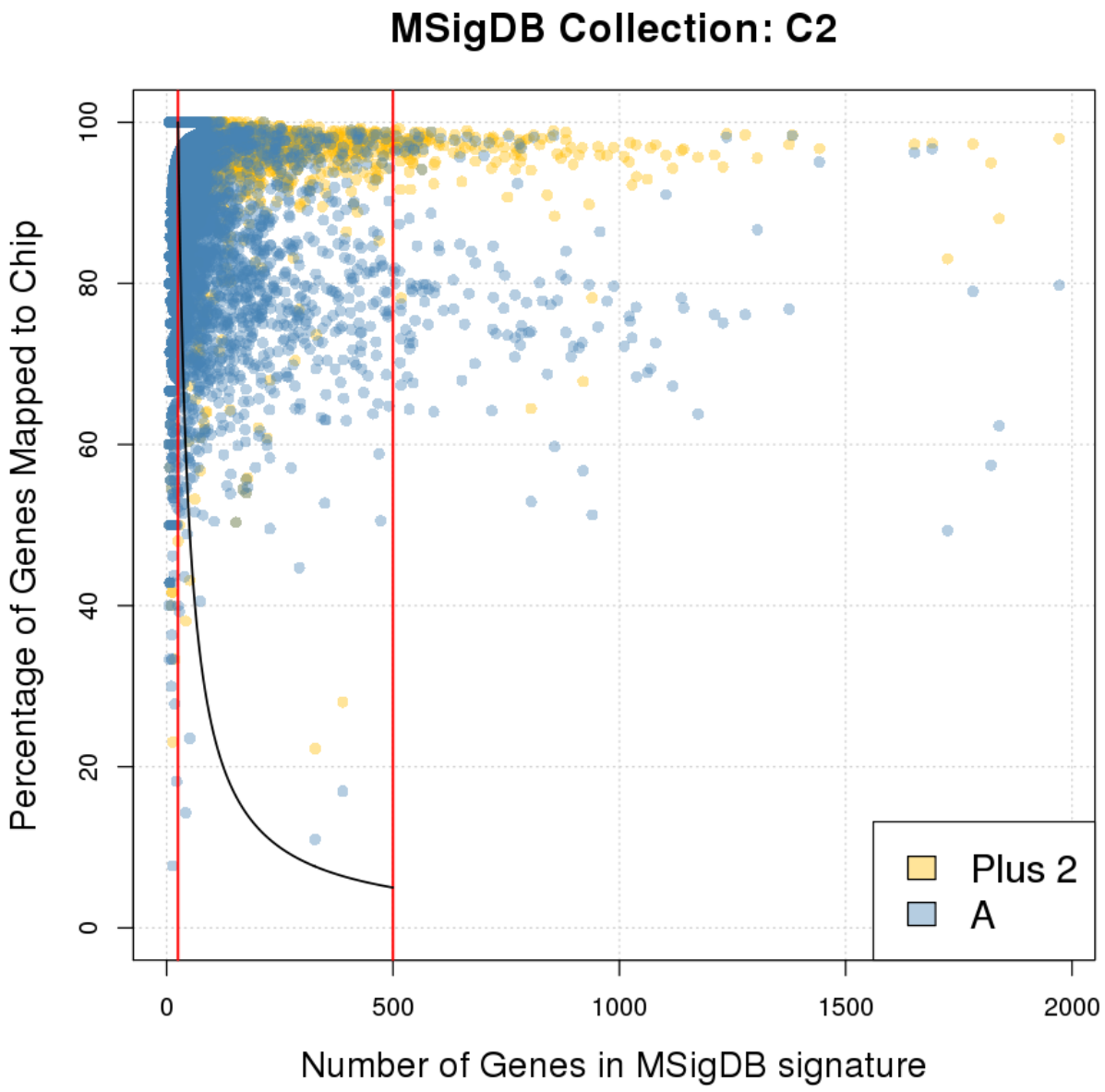


Figure S8: MSigDB C3 collection

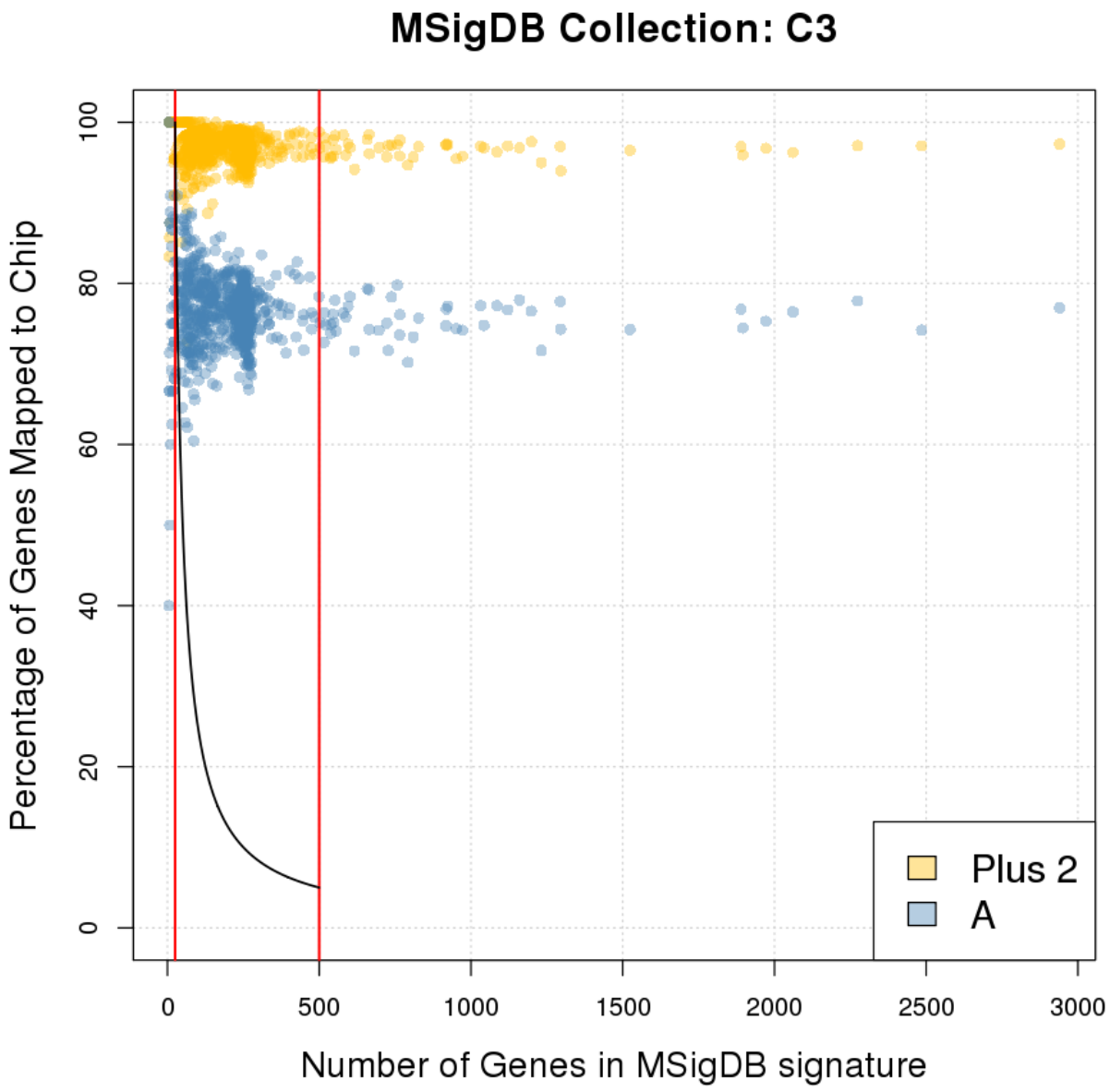


Figure S9: MSigDB C4 collection

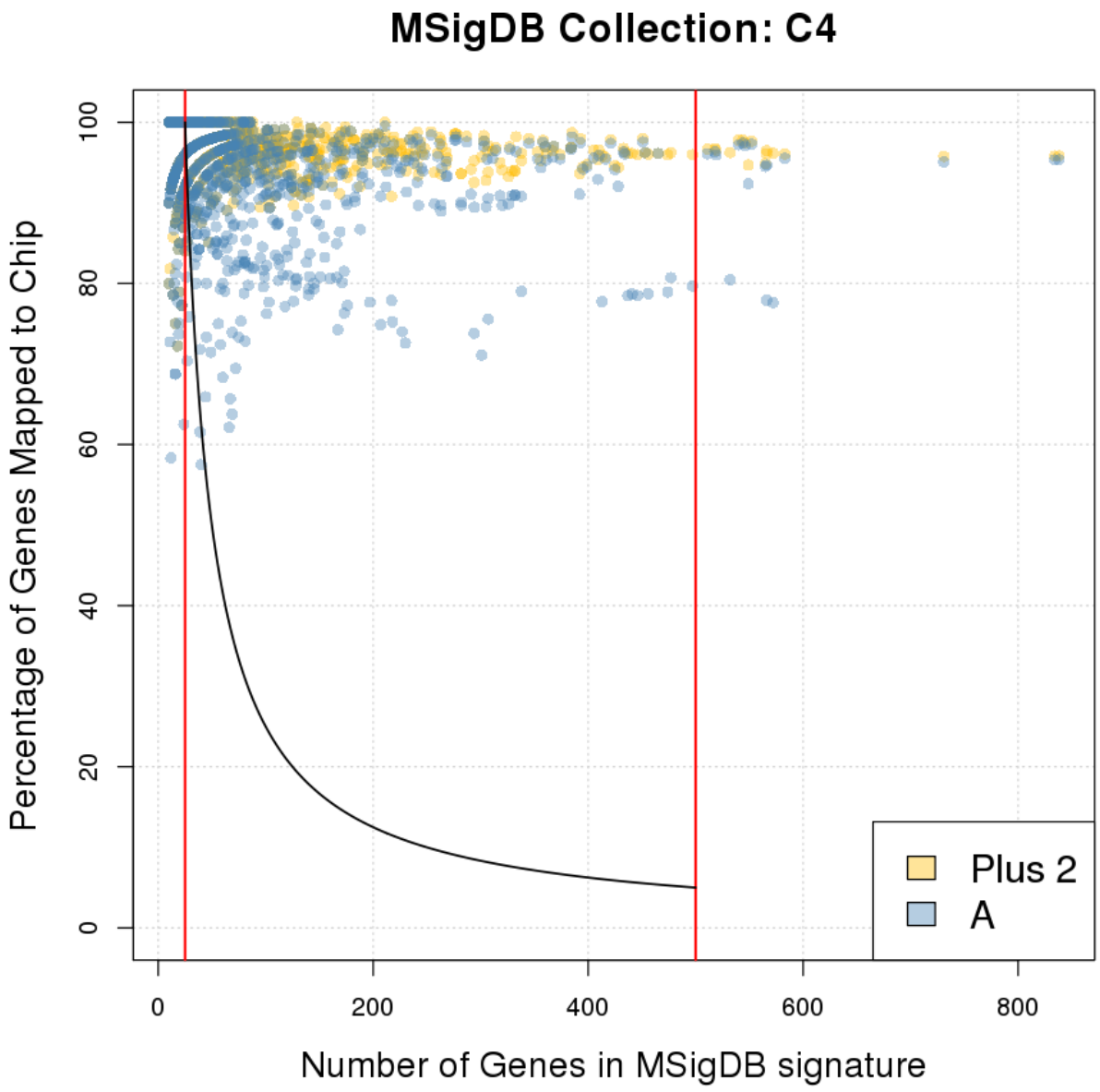


Figure S10: MSigDB C5 collection

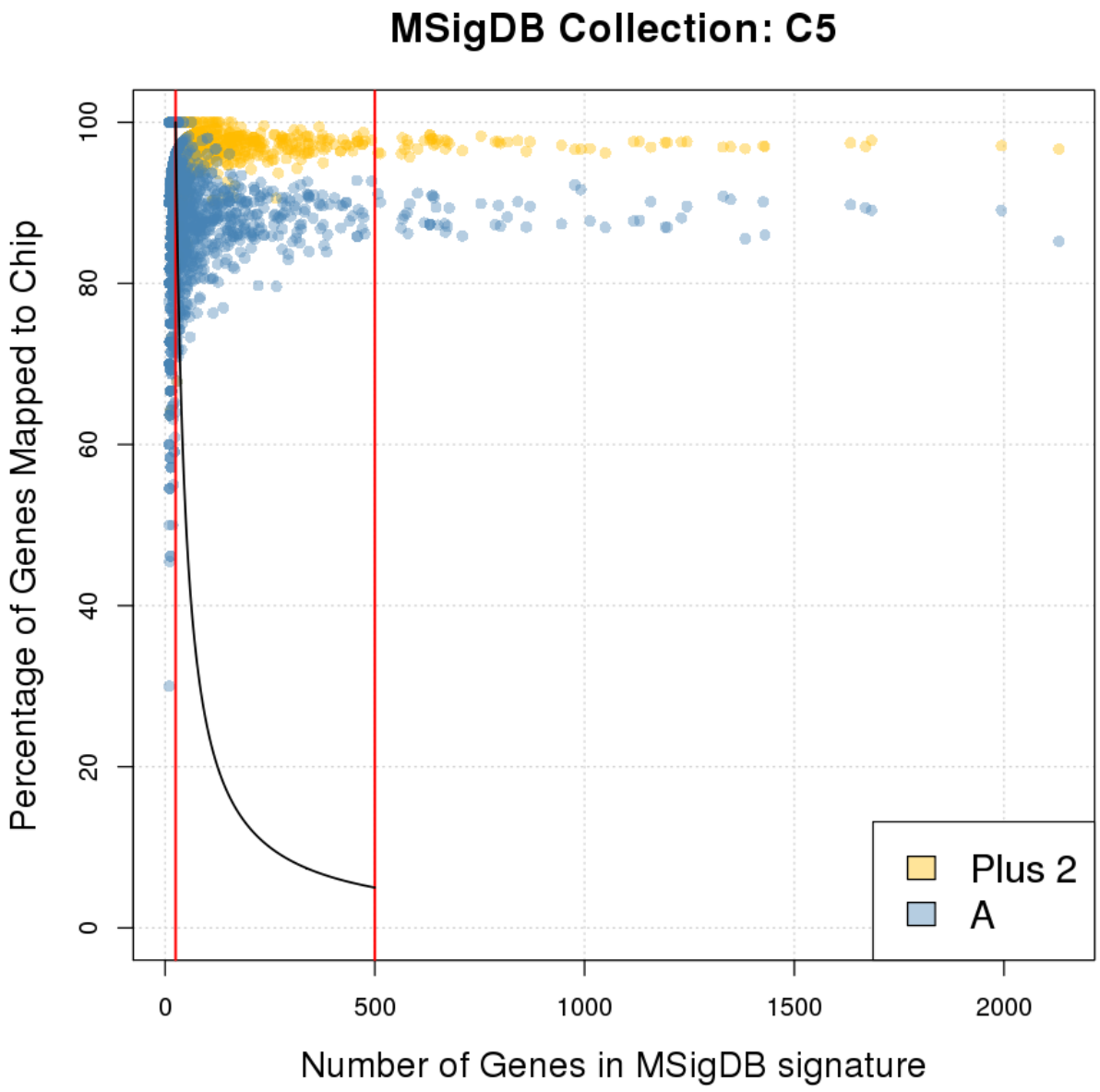


Figure S11: MSigDB C6 collection

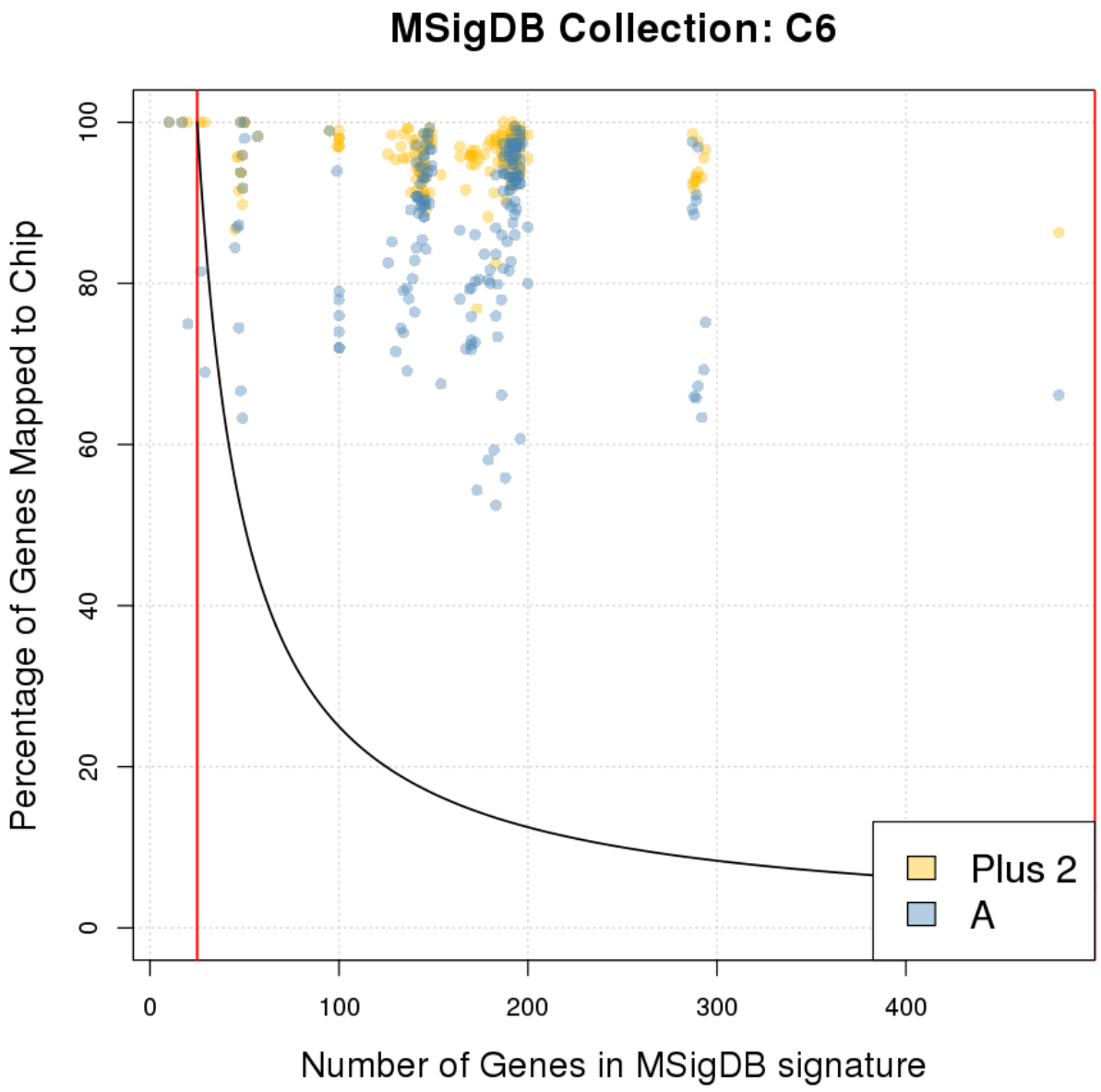


Figure S12: MSigDB C7 collection

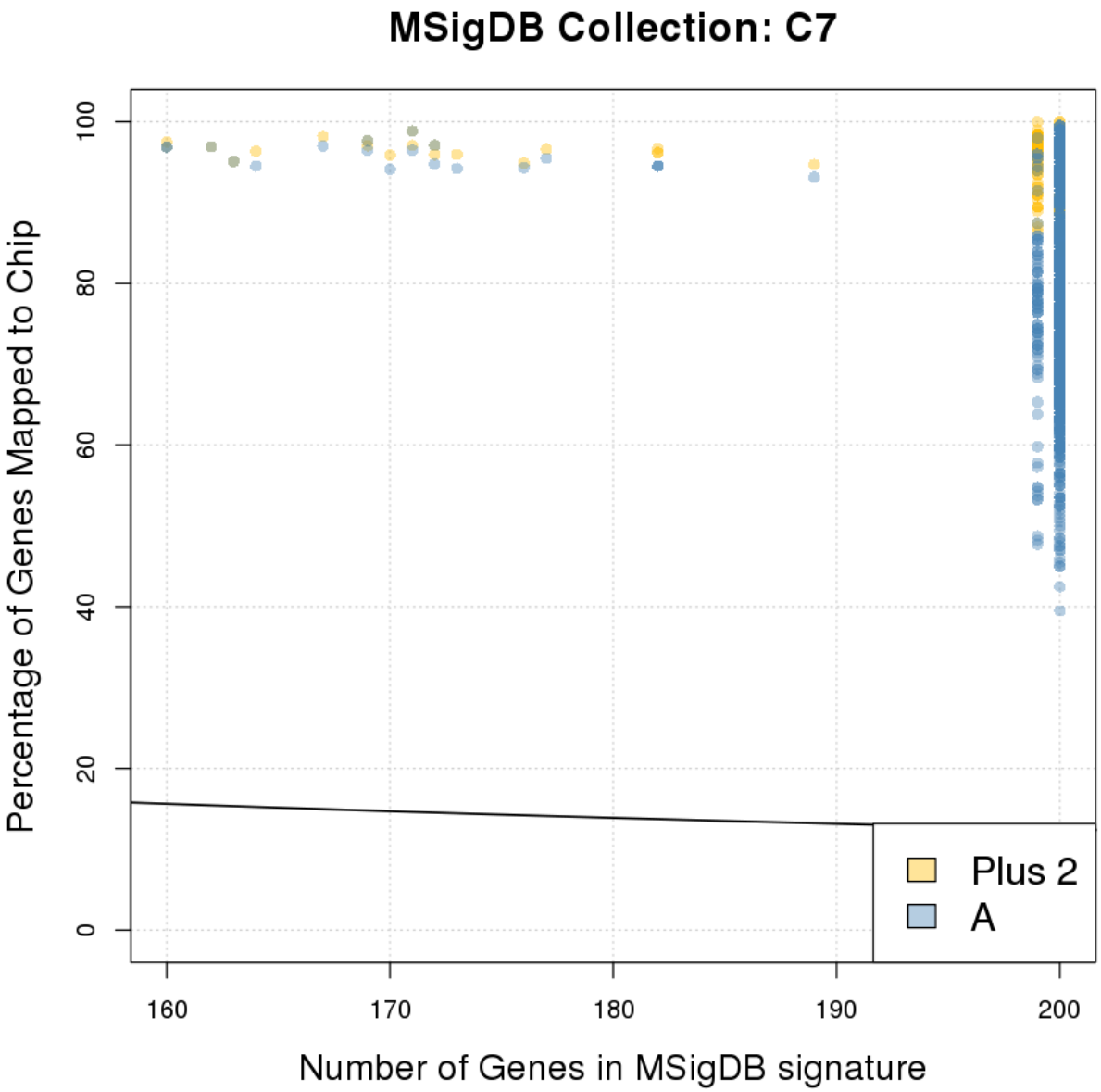


Figure S13: MSigDB H collection

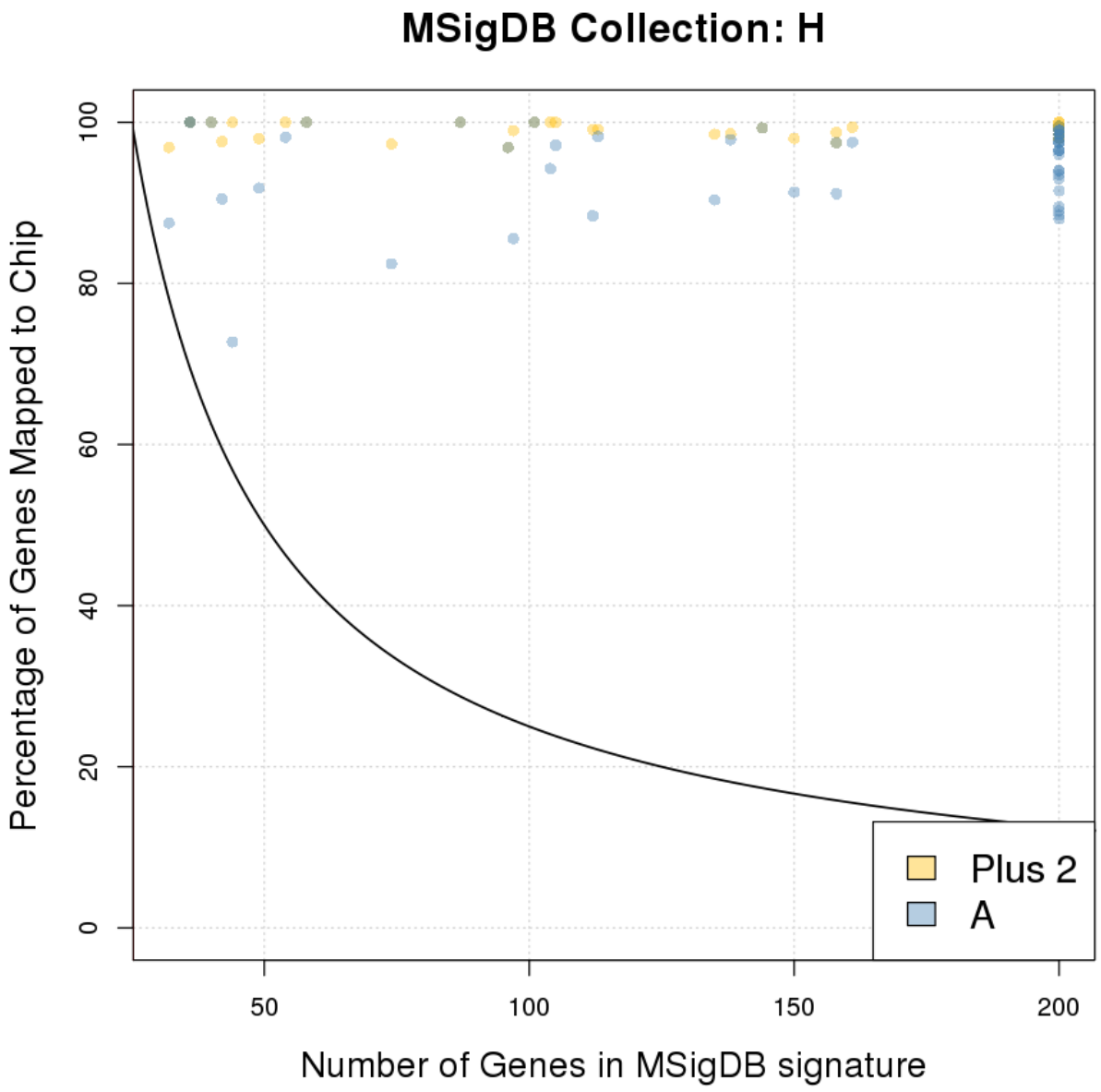


Figure S14: Test and training CV(RMSE) based on λ_{min} coefficients

Each colored circle represents a gene model. The marginal histograms show the distribution of errors across the 9986 gene models. The 365 gene models from the Human Disease Network are depicted in orange. The errors are comparable, but slightly higher, than the ones obtained when the λ_1se coefficients are used (Figure 3 in main paper).

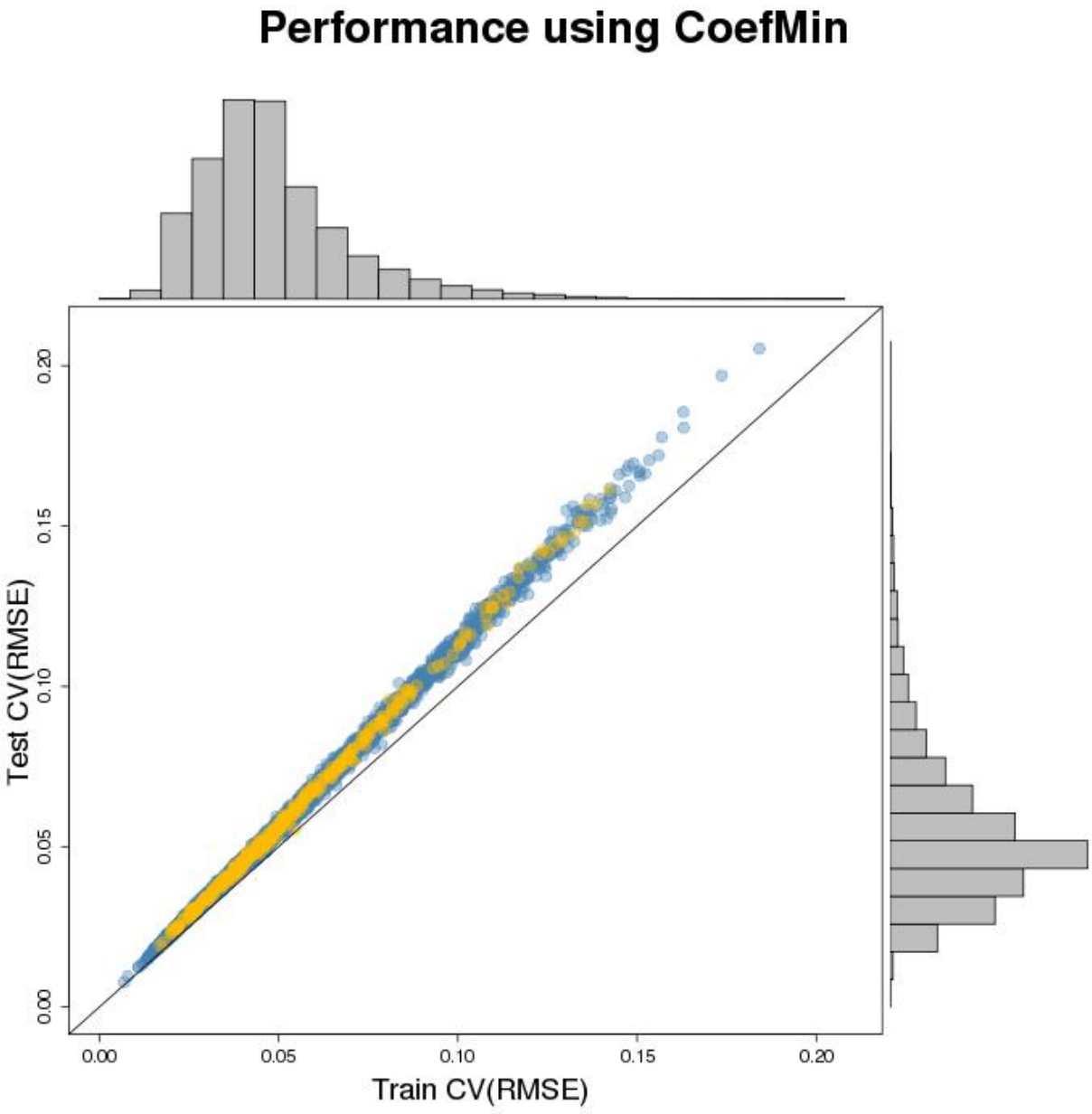


Figure S15: Sparsity of coefficient matrix and the effect on test CV(RMSE)

We considered two additional coefficient matrices that were selected to have 80% and 60% sparsity respectively. (In contrast, the Coef 1se matrix used in the paper is approximately 40% sparse.) The means of the test set CV(RMSE) for each matrix is plotted below: