Supplemental Material

Data S1.

Supplemental Methods

Processing of transcriptomic data sets

For microarray studies CEL files were read using R's *oligo* package and normalized using Robust Multi-array Average (RMA) [18]. Probes were annotated to their corresponding HUGO Gene Nomenclature Committee (HGNC) gene symbols using platform specific annotations. For duplicated measurements the mean intensity was calculated. For RNA-Seq studies, reads were aligned using BioJupies [10]. BioJupies works with the ARCHS4 pipeline utilizing Kallisto to map reads onto the human GRCh38 cdna reference. All studies have been processed by Illumina platforms except for Tarazon14, which utilized AB 5500xl Genetic Analyzer. Here the nucleotide sequence is coded in color space that could not be handled by the BioJupies pipeline and the alignment of Tarazon14 was therefore performed with R's Rsubread package [66], TMM normalization factors were calculated with R's edgeR package[67]. All RNAseq datasets were transformed using voom from R's limma package to obtain continuous measurements [17].

Hannenhalli06 only provided processed data, but followed identical normalization methods. In the case of Kittleson05, processed data was used since raw available data was incomplete. Identical normalization procedures were followed. Read alignment wasn't performed for vanHeesch19 since they only provided raw transcript counts, but identical normalization procedures were followed. One sample from Liu15_R was excluded due to technical reasons.

For each experiment, sample quality was assessed by visually comparing the distribution of gene expression values. Multidimensional scaling was performed to visualize the separation of HF and control samples. No samples were excluded based on these metrics and no additional quality control was performed.

Sample variability

To evaluate the study specific batch effects, we used principal component analysis (PCA) on the union of all pre-processed datasets and the genes that were shared among all the studies (Figure S4A). Each principal component was then tested for association with the study labels using Analyses of Variance (ANOVAs) (p-value < 0.05). To obtain a simple data integration we performed a z-transformation of all genes independently for each study including only HF-samples. Principal component analysis was performed on this transformation and each principal component was tested for associations with study or technology labels using ANOVAs (Figure S4B). t-Distributed Stochastic Neighbor Embedding was used for alternative visualization (Figure S4C).

In an additional analysis, we first standardized (mean = 0, sd = 1) all genes independently for each study including all samples and then merged them into a single matrix. Principal component analysis was performed on this transformation and each principal component was tested for associations with study or technology labels using ANOVAs (Figure S5).

To quantify how much of the variability of the samples within a study can be explained by the covariates used in their differential expression analysis, we fitted linear models to a reduced data representation (Figure S6). For each study, first, we standardized its gene expression and performed dimensionality reduction using PCA. Then we tested each principal component for association with each covariate using linear models. If a covariate was associated with a Principal Component (p-value < 0.05), then we assigned the proportion of explained variance to it. We also applied the same methodology for HF patients only. Underestimations of proportion of explained variance are expected in small studies, since the number of evaluated principal components equals the number of samples. However this is a fair approximation for most of the studies.

Gene-specific expression variability

We merged studies after processing and gene standardization. Independent two-way ANOVAs were fitted to each gene using disease status as a first factor, and for samples with available information, sample's

study, transcriptional profiling technology, sex, age or occasion of sample acquisition, as a second factor. The proportion of explained variance of each independent variable was measured with eta-squared values (Figure S12). Additionally, to evaluate the bias of the HF consensus signature towards dilated cardiomyopathy, we performed independent two-way analysis of variance (ANOVAs) to quantify the amount of explained variance in gene expression that could be accounted to differences in heart failure etiology (Figure S13). First, we selected 8 studies in our curation that profiled sufficient ICM and DCM patients (at least 3 patients of each etiology). Then, for each selected study we fitted to each gene an ANOVA with HF and etiology as covariates. Eta-squared values of each covariate were used as a proxy of the proportion explained variance.

Differential expression analysis

Samples with incomplete clinical information from vanHeesch19 were excluded from the analysis to be able to account for the clinical information of the remaining samples in the DEA. We excluded the age information in the DEA of the samples from Kim16. Here, excluding samples with unknown age information would have reduced the sample size drastically.

Between study consistency and replicability

The disease score is an expression footprint based transfer learning approach that compares the observed expression patterns in the samples of one experiment (B) with the expected disease patterns observed in an independent sample from another experiment (A). First, for an experiment *A*, *k* differentially expressed genes between the healthy and disease condition are defined using linear models. The t-values of these *k* genes are used as the expected disease pattern to be used for transfer learning. Then, for each sample *i* in experiment B we calculate its disease score by making a linear combination of the t-values from these k genes with their expression values in sample i, for genes present in both the reference signature and the expression values (Figure S8). All disease scores were standardized after calculation. The robustness of the disease score classification and the enrichment analysis was tested using 50, 100, 200, 500, and 1000 differentially expressed genes (Figure S10).

Meta-analysis

We evaluated the importance of the top genes of the meta-ranking in the description of HF patients by repeating the classifications made with the disease score described before. Samples of each study were classified using a disease score defined by the first *n* or *total-n* genes in the meta-ranking and study-specific t-values. AUROCs were averaged for each predicted study and n ranged from 50 to the total number of genes in the meta-ranking (Figure S11).

To evaluate the added value of the meta-analysis, we tested if the selection of the top 500 genes from the consensus signature defined a better transcriptional signature of HF compared to signatures obtained from individual experiments. We tested if the AUROCs obtained were greater than the ones coming from classifications made by the top 500 genes coming from individual studies using a Wilcoxon paired test. To show that the top genes of the consensus signature shared a more consistent direction of differential regulation than signatures coming from individual studies, we separated the 500 top genes from the consensus signature into up and downregulated independently for each dataset , and enriched them into the sorted gene-level statistics of each of the other studies using Gene Set Enrichment Analysis (GSEA) as in Figure 2 C. We compared the enrichment scores of these pairwise comparisons to the ones obtained using the top 500 differentially expressed genes of individual experiments using a Wilcoxon paired test.

Functional analysis

Gene sets with less than 15 or more than 300 genes were excluded from the GSEA analysis. A, B, C and D regulons from DoRothEA with less than 20 genes were excluded from the viper analysis. Pathway activities were estimated using 200 footprint genes from PROGENy. Empirical p-values for PROGENy scores were calculated from pathways' null distributions calculated after permuting 1000 times the labels of the directed-meta-ranking. BH-corrected p-values were calculated for each test and are available in table S3.

Extrapolation of the HF consensus signature to other etiologies, HF-related processes or technologies.

Studies from the query results that did not match inclusion criteria due to differences in HF etiology, biopsy location or profiling platform were used for further exploration of the disease score classifier (GSE10161, GSE4172, GSE76701, GSE84796, GSE9800, GSE52601) (Figure S15, table S1). We calculated the mean disease score of each sample of these excluded studies using the top 500 genes of the meta-ranking and the gene level statistics of the studies included in the meta-analysis. AUROCs were used to evaluate the ability of the disease score to differentiate between healthy and HF patients in each data set.

Additionally, we proposed a framework to use the HF consensus signature as a resource to build and confirm hypotheses. First, dysregulated features are identified in an independent study. Next, a test for enrichment of these features is performed in the HF consensus signature using GSEA. Finally, highly consistent features can be filtered by dysregulation direction and significance levels. We used a combination of the leading edge of GSEA and the ranking of the HF consensus signature.

For the analysis of plasma biomarkers, we used the result tables from Egerstedt, *et al* [57] that contained protein-level statistics of the comparison of plasma proteomics of healthy and HF patients (manifest HF), and the results of the prospective analysis of proteins during HF development (early HF). Proteins that mapped to a gene symbol in the HF consensus signature and had a BH corrected p-value<0.01 were tested for enrichment as described above. For the analysis of fetal transcriptional responses (Figure S17), we used the expression matrices of two studies (Spurrell19, GSE52601) that compared healthy human hearts with fetal hearts. Differential expression analysis and estimation of TF activities of these two studies were performed as described before. Genes with a BH corrected p-value < 0.05 were tested for enrichment and TF activities with a p-value < 0.05 were compared to the ones estimated from the HF consensus signature.

Statistical analysis

All correlations and Wilcoxon paired tests were performed using *stats* package. *sjstats* package was used to calculate ANOVAs and eta-squared values, *ROCR* package was used to calculate receiver operating characteristic curves [68].

Supplemental Results

Study Description

Gene expression of all studies was measured with RNA-seq and microarray (eight datasets each) on eight different platforms (table S1). The age of HF patients is noticeably younger than what would be expected, since HF prevalence increases with age (Figure S3). This might be connected to age restrictions in transplantation guidelines and LVAD treatment recommendations.

Study Comparability

Despite identical normalization and analysis procedures for all datasets, we visualized variation due to study and technology as we expected it might impact our study. In a PCA of all unified gene expression values after processing, 85% of the variance of the samples was explained by the first two components representing study of origin and applied technology (Figure S4A). These differences among cohorts reflect the expected inherent interaction that technical and sample heterogeneity have with gene expression and reinforce the importance of adjusting for technology when combining samples. Due to the study and technology bias of untransformed gene expression values, HF samples were z-transformed and again analyzed via PCA (Figure S4B). 74% of the variance captured by the principal components explained differences of HF samples by study (ANOVA p-value <0.05). The difference of samples by study was better visualized when a t-SNE was performed to this data (Figure S4C). We did not use this approach of data integration for any downstream analysis, due to the strong technical variation.

Next, we compared studies on the level of differential gene expression (HF vs. control) to explore how technical and sample variability affected gene level statistics. A strong difference in the distributions of t-values and p-values of the genes compared is visible in the largest study in our analysis (Liu15_M) (Figure S7). This difference in distributions persists after adjustment for all available clinical covariates, though it is consistent with expectations based on study sample size. These results together establish expected bias among datasets, likely dependent on technical differences rather than biology.

Gradient of information in the meta-analysis

We tested the performance of sample classifiers using different numbers of top genes from the consensus signature with our previously defined disease score. We observed a constant decrease in the mean AUROCs

of classifiers that excluded genes at the top of the consensus signature or included genes at the bottom (Figure S11), confirming that a gradient of meaningful information is present in this ranking.

Gene-level variability

A series of independent two-way ANOVAs were fitted to a complete data set that combined each study individually after gene-standardization to quantify the proportion of variability in gene expression that can be explained by HF and other clinical or technical covariates (Figure S12). Gene standardization cancels the effect that the study of origin and technology have on gene expression (Figure S5) and can be confirmed by the low eta-squared values in all genes (Figure S12 upper panels). For the top 500 genes in the meta-ranking we observed a higher eta-squared value for HF than any other additional clinical covariate (Figure S12 lower panels), suggesting that the expression of top-ranked genes in our consensus transcriptional signature is mostly influenced by HF than any other covariate measured in the analysis. Similar trends were observed when analyzing the effects of etiology differences in individual studies (Figure S13).

Supplemental Tables – see Excel files

Table S1. Complete description of the studies included in the meta-analysis.

Table S2. Summary statistics and rankings from the meta-analysis.

Table S3. Functional characterization of the consensus signature. GSEA gene set level statistics for MSigDB's canonical pathways and gene ontology terms, DoRothEA's transcription factor level statistics**,** PROGENy's signalling pathway level statistics, and micro-RNA level statistics.

Table S4. Full results from validation analysis.

Figure S1. Schematic representation on how the disease score was defined. AUROC, area under the receiver operating characteristic. HF, heart failure.

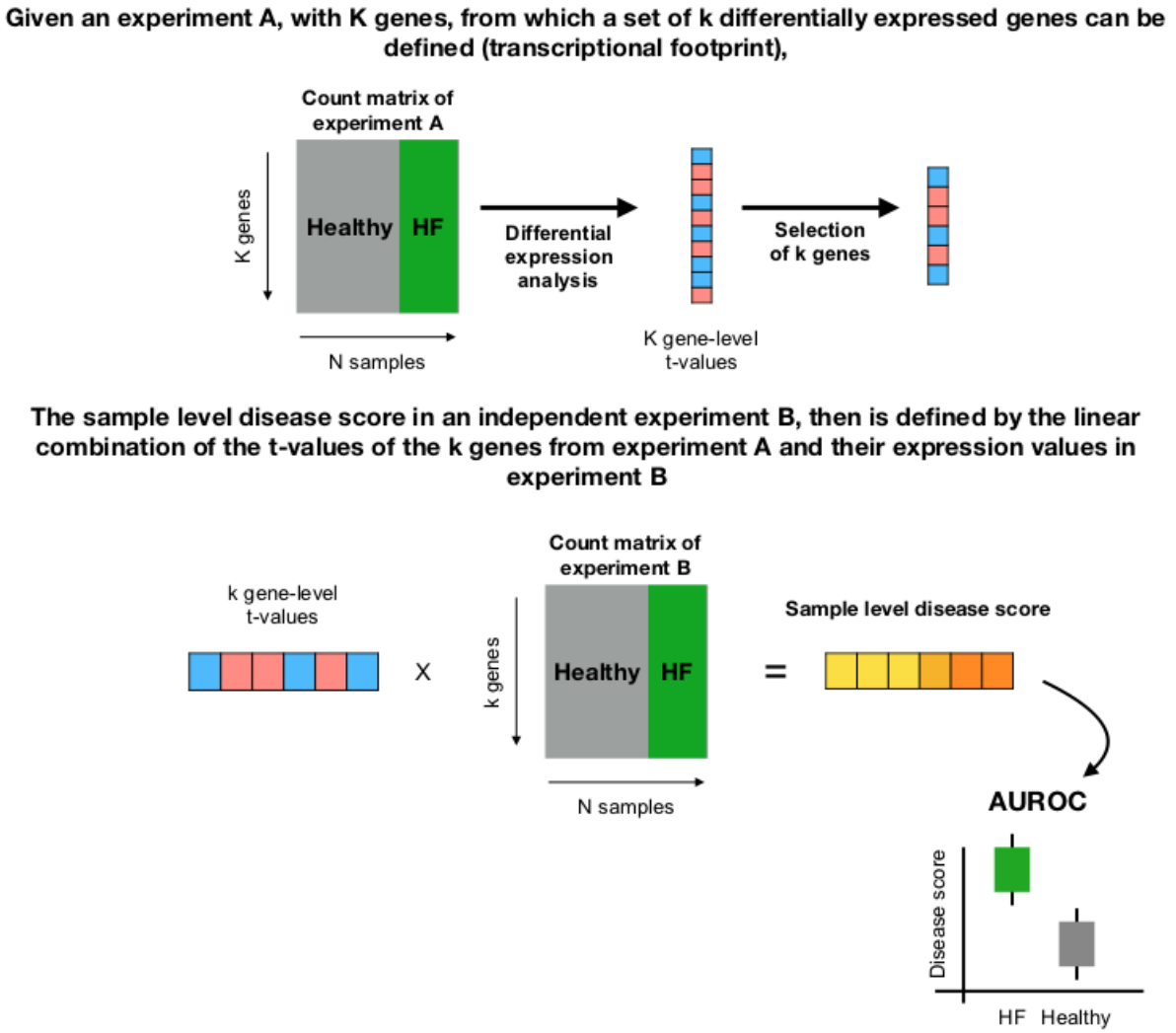
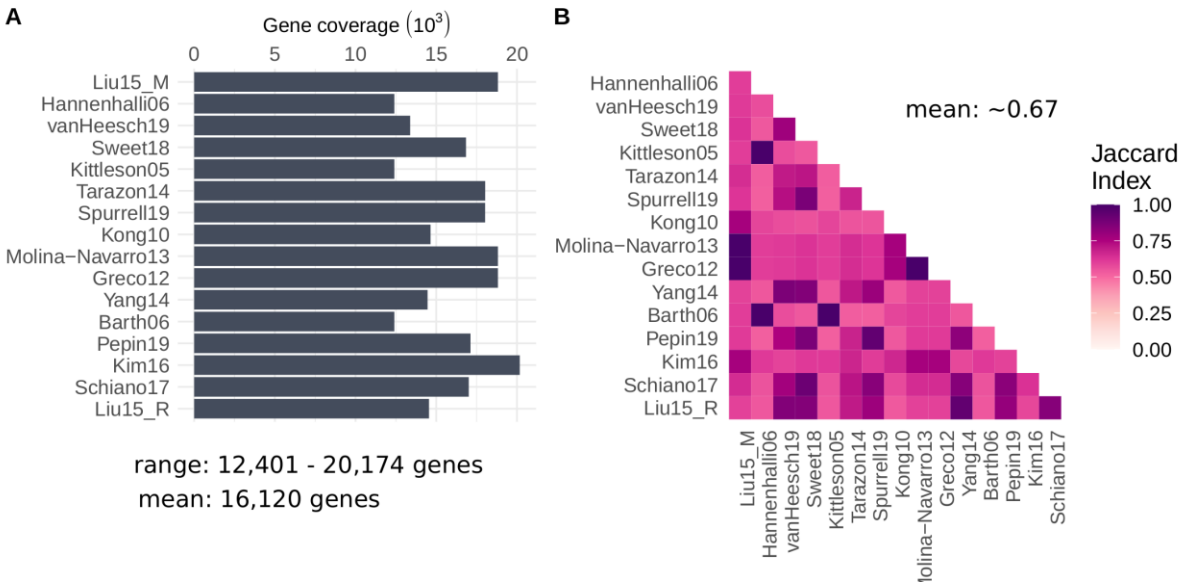
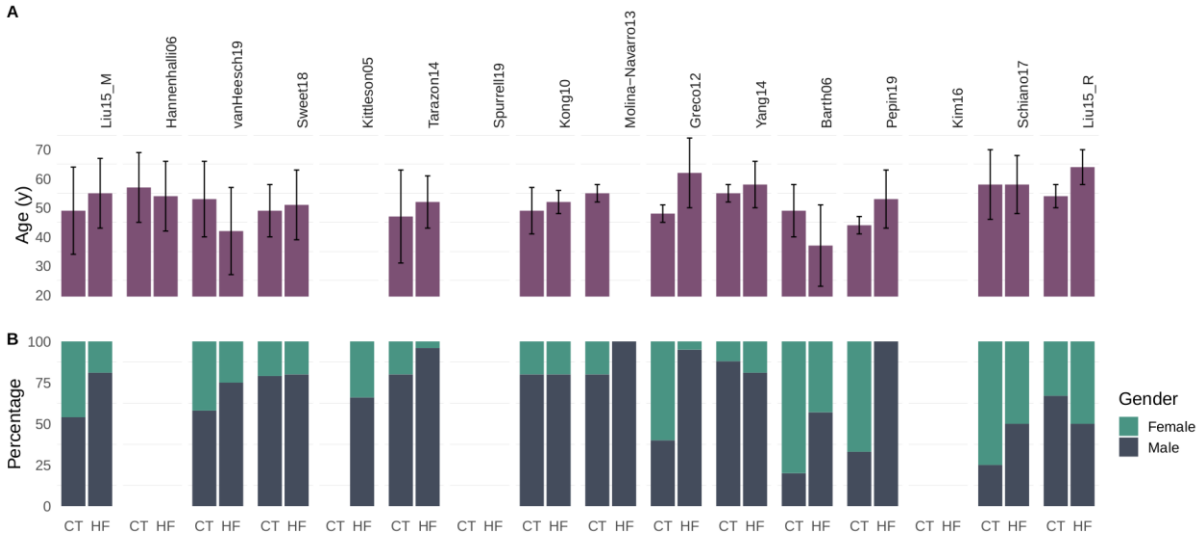


Figure S2. Overview of gene coverage of studies included in meta-analysis.



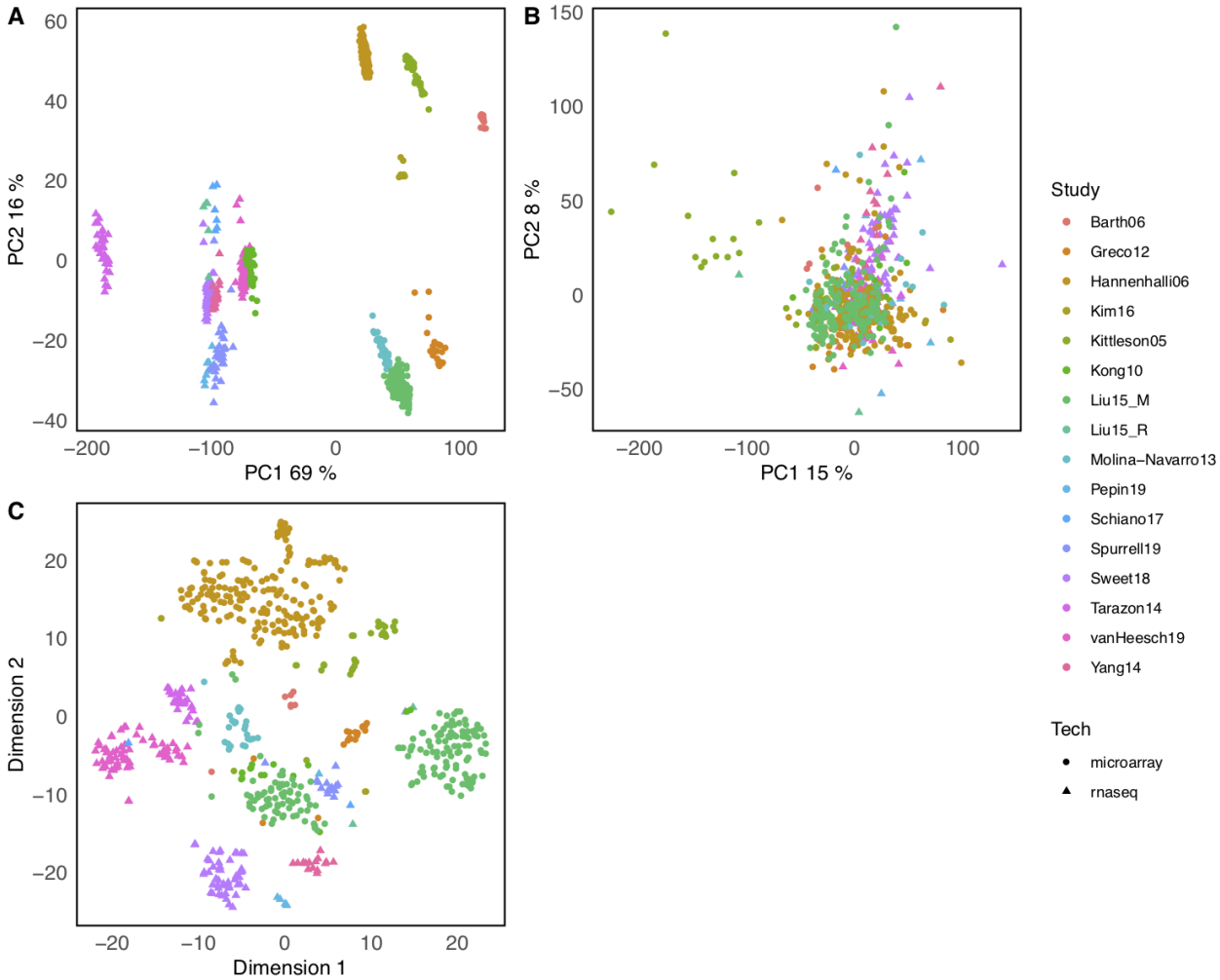
A) Absolute gene coverage per study after processing. B) Pairwise comparison of covered genes measured with Jaccard Index.

Figure S3. Age and sex distribution per study.



A) Age distribution in years of control (CT) and heart failure samples (HF) per study. Displayed is mean and standard deviation. B) Sex of patients in % per study.

Figure S4. Differences in samples included in the study.

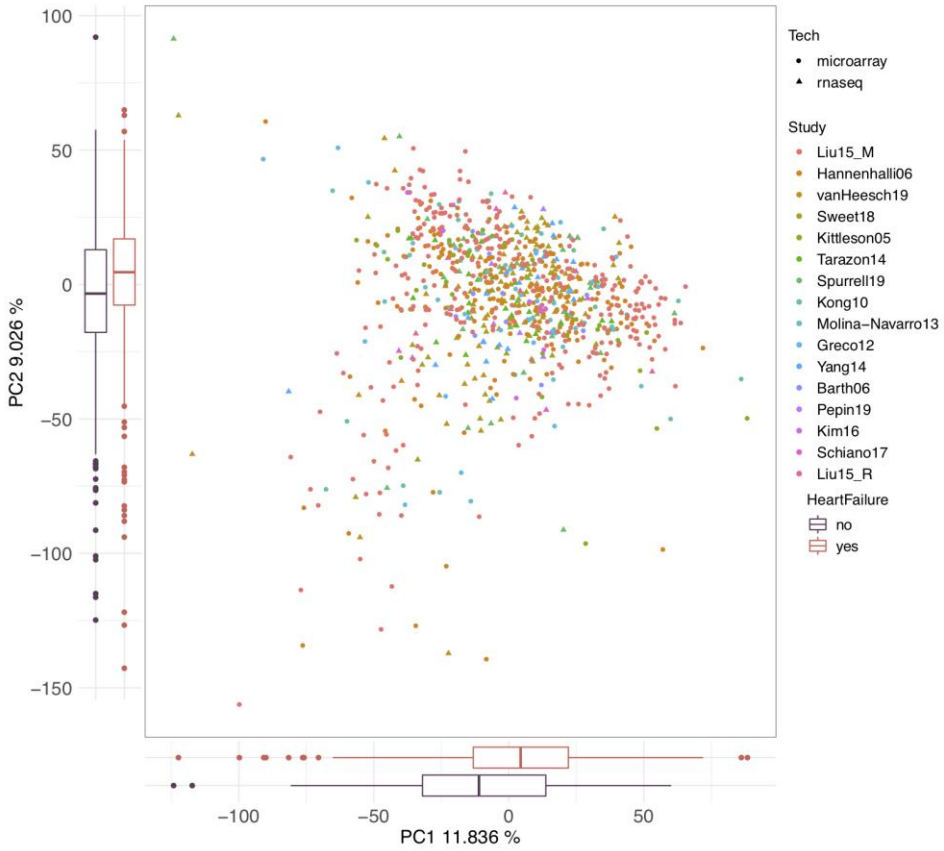


A) First two components from a Principal Component Analysis (PCA) done to all samples

B) First two components from a PCA done to all z-transformed heart failure samples

C) t-distributed stochastic neighbor embedding of all z-transformed heart failure simples

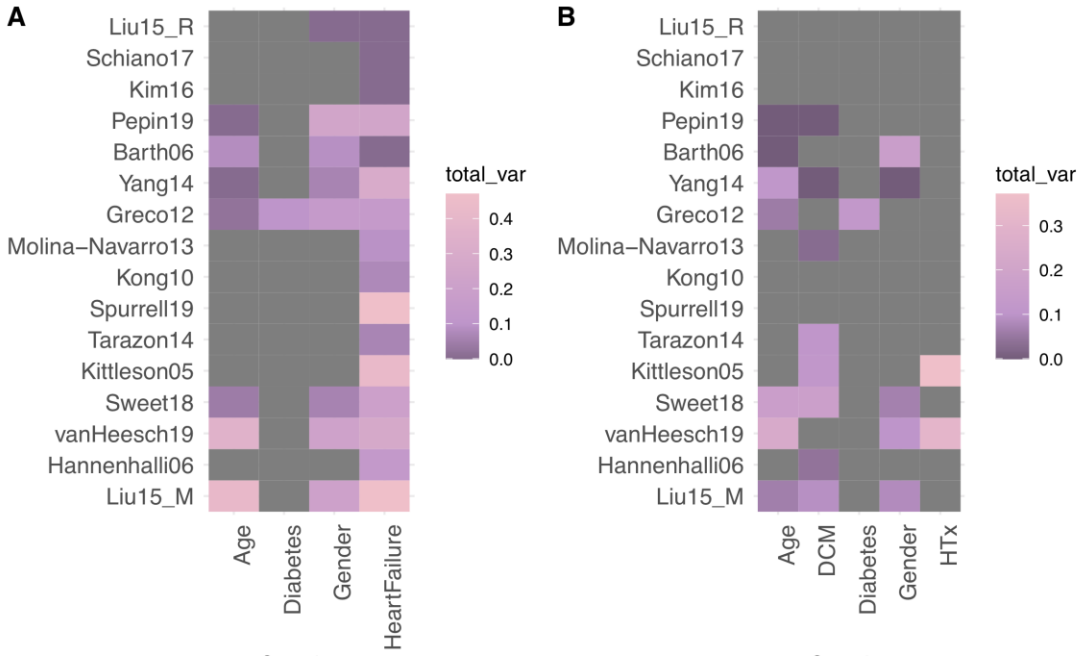
Figure S5. Principal Component Analysis of all samples analyzed after gene standardization.



Factor	Total PCs	Cumulative Proportion of Variance
Heart Failure	26	0.43534
study	16	0

The scatter plot shows the first two principal components and the percentage of variance explained by them. In the table is showed the cumulative proportion of variance that is explained by components associated to Heart Failure and study (Analysis of variance, p-value<0.05)

Figure S6. Contribution of the covariates to the variability of individual studies.



Estimated proportion of explained variance associated with the different covariates used in the differential expression analysis (See Supplemental Methods) in A) all patients and B) only **heart failure.** patients. Grey tiles represent missing reported data. HTx, heart transplantation

Figure S7. Distributions of -log10(p-values), t-values and log2(fold-changes) [LFC] from the differential expression analysis of all genes measured in each study.

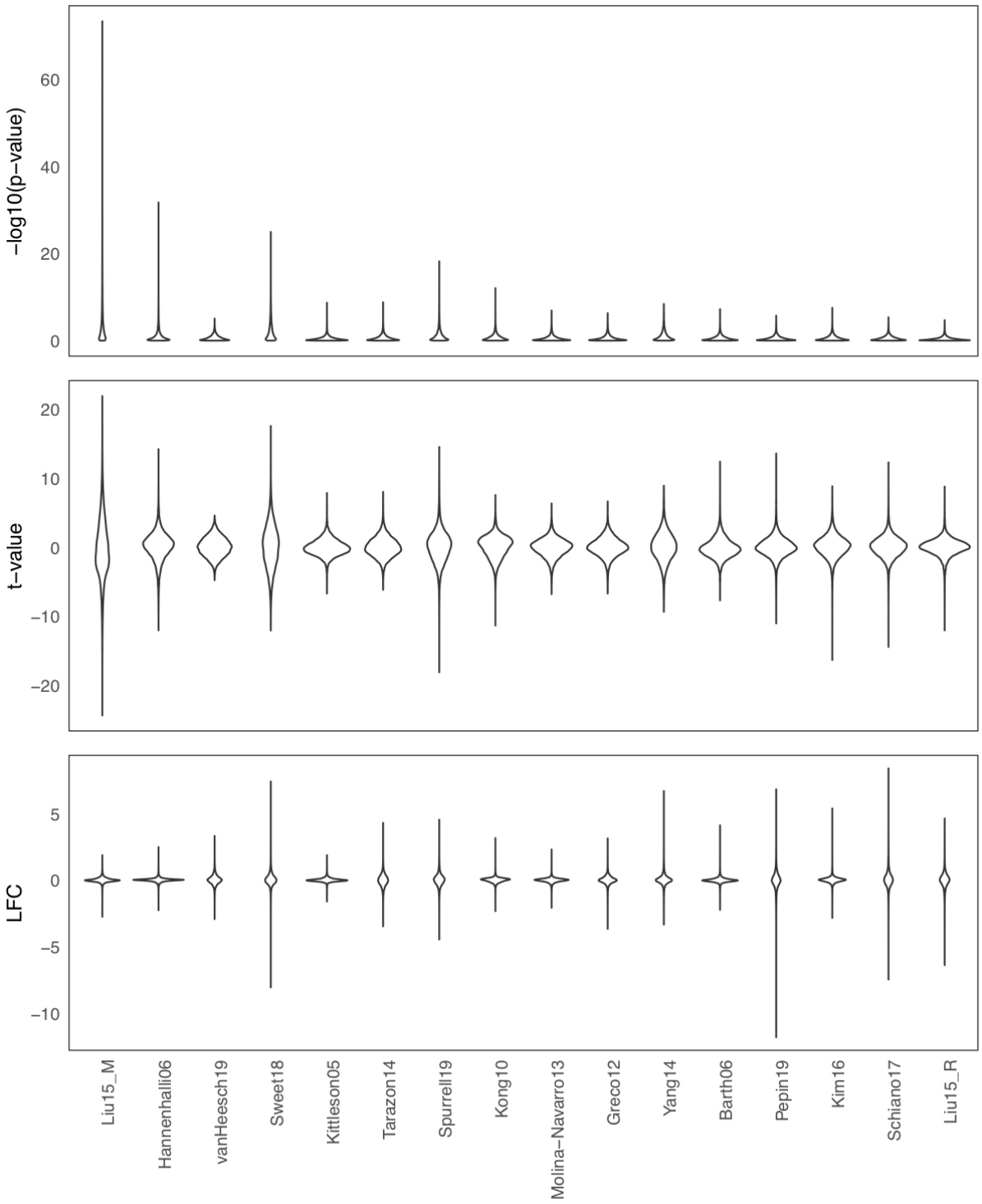
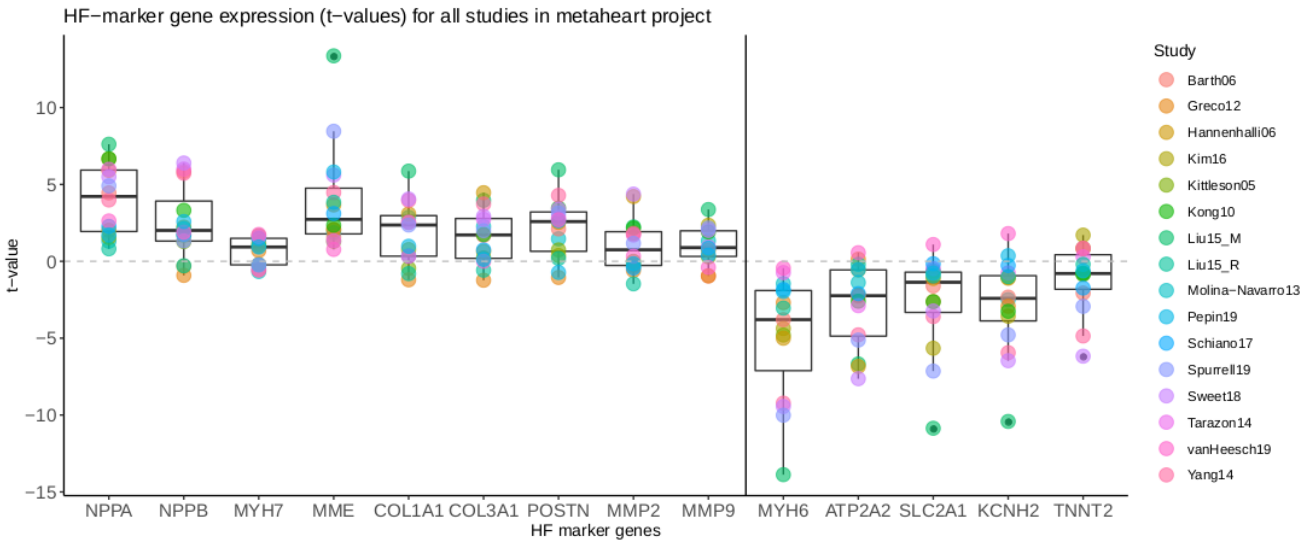
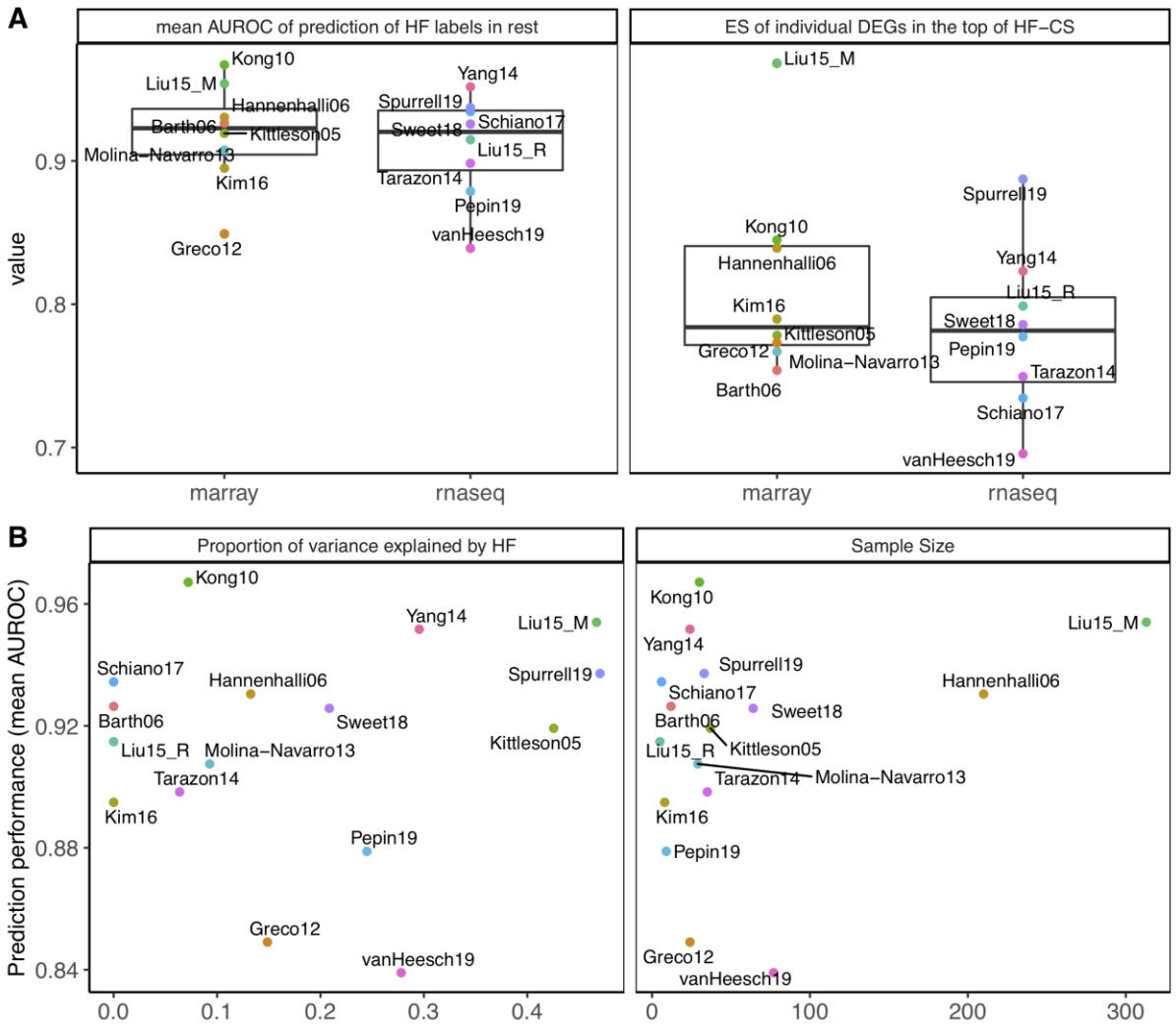


Figure S8. t-values from the differential expression analysis of genes that are established as dysregulated in heart failure (HF).



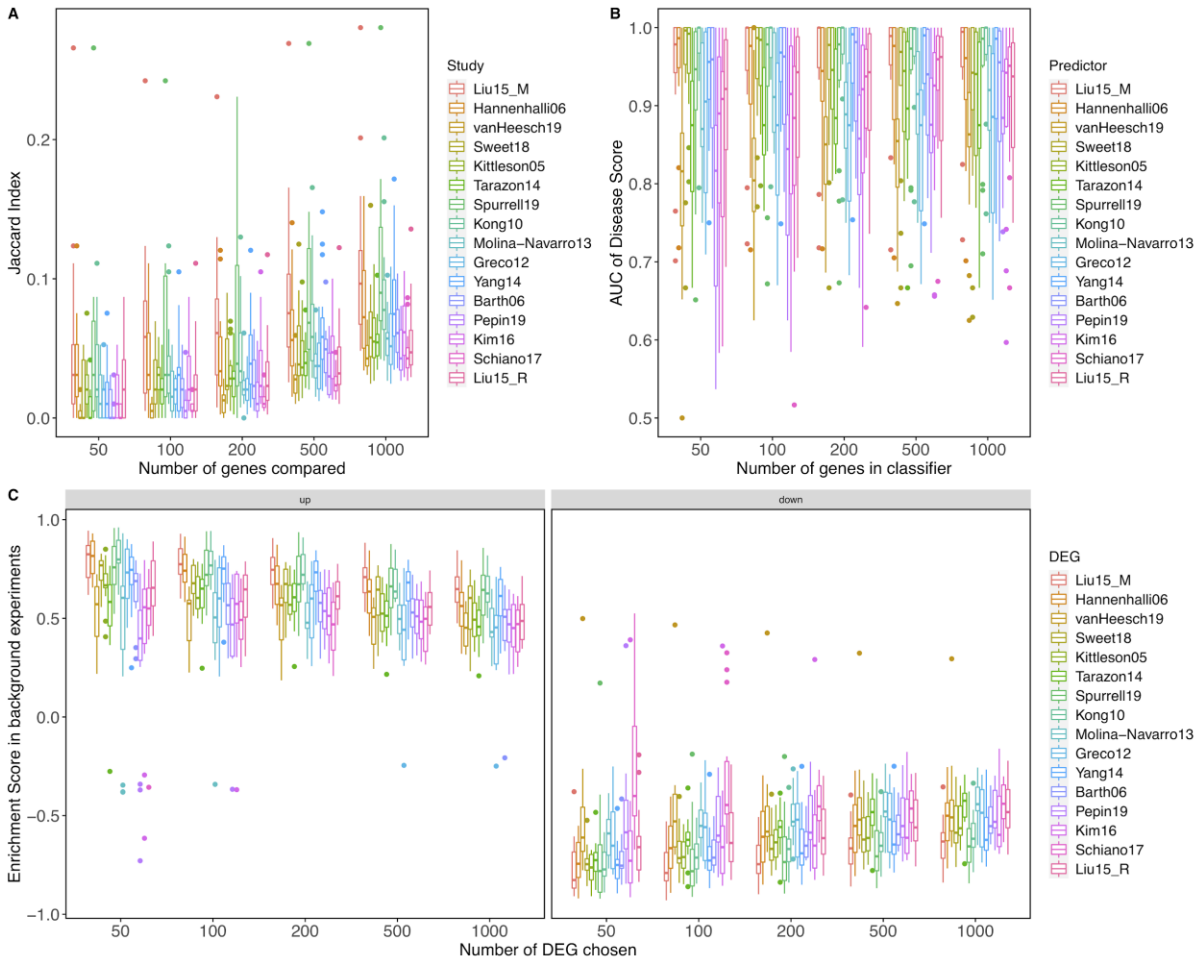
Expected up and downregulated genes are in the left and right panel, respectively.

Figure S9. Comparison of the studies included in the meta-analysis.



A) Distribution of predictor performances and enrichment of differentially expressed genes in the heart failure consensus signature (HF-CS) grouped by technology. In the left panel each dot represents the mean area under the receiver operating characteristic curve (AUROC) of the disease score classifier trained in a study and tested in the rest (See Methods). In the right panel each dot represents the enrichment scores of the top 500 differentially expressed genes of the study in the HF-CS. B) Relationship between the predictive performance of each study and its proportion of explained variance associated with HF (See Supplemental Methods, Figure S6) and sample size.

Figure S10. Test of robustness of the replicability measures used to compare the studies included in the meta-analysis.



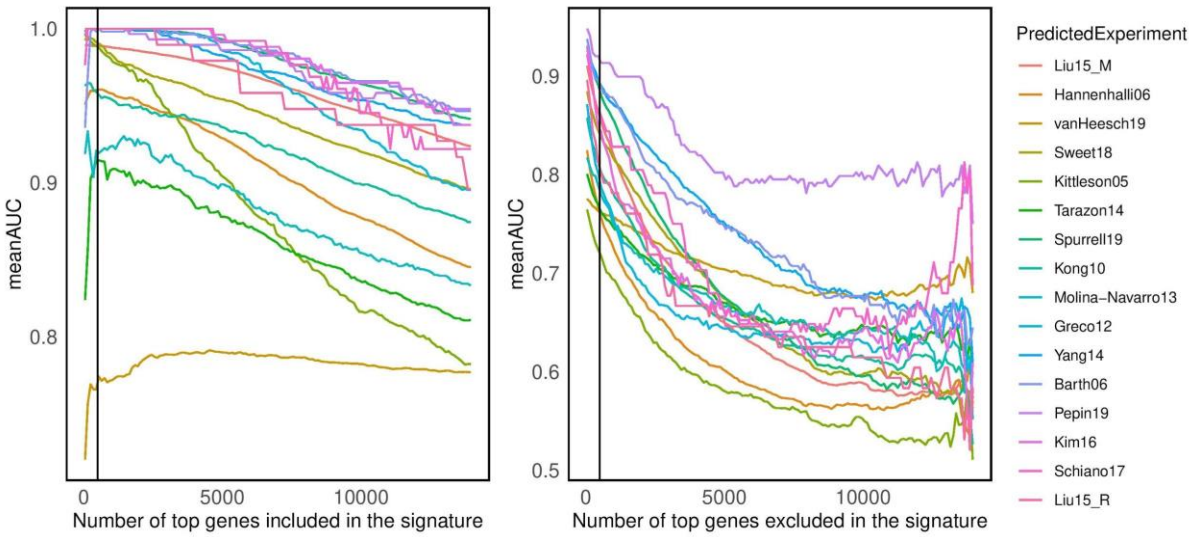
Each dot represents a pairwise comparison using:

A) Jaccard Index

B) Disease Score

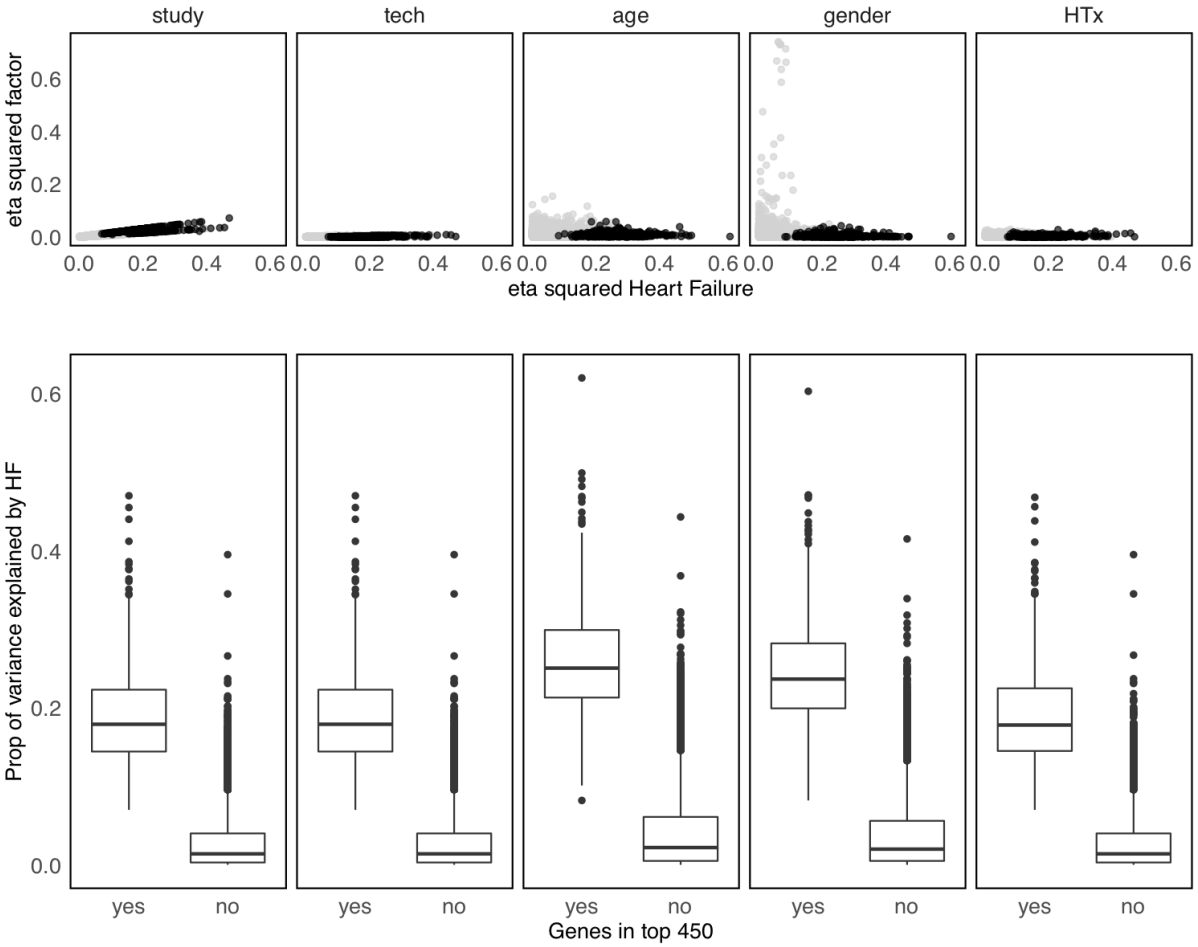
C) Enrichment Score

Figure S11. Mean area under the receiver operating characteristic curve (meanAUC) of predictions using the disease score with n (left panel) or total-n (right panel) genes of the consensus signature from the meta-analysis and gene-level statistics of all studies except the one being predicted to avoid overfitting.



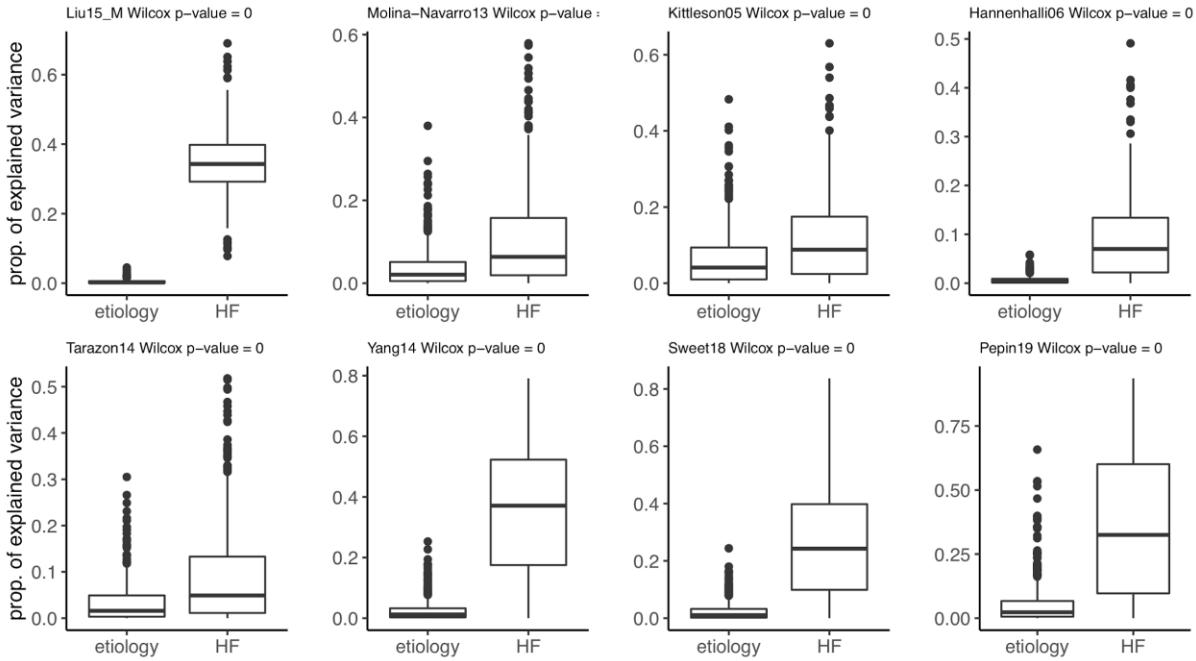
The line shows where we defined the cut-off for the rest of the tests (500). A general decrease of the meanAUC is observed as top genes of the meta-analysis are excluded from the calculation of the disease score.

Figure S12. Proportion of gene expression variance explained by heart failure (HF) and additional clinical and confounding factors.



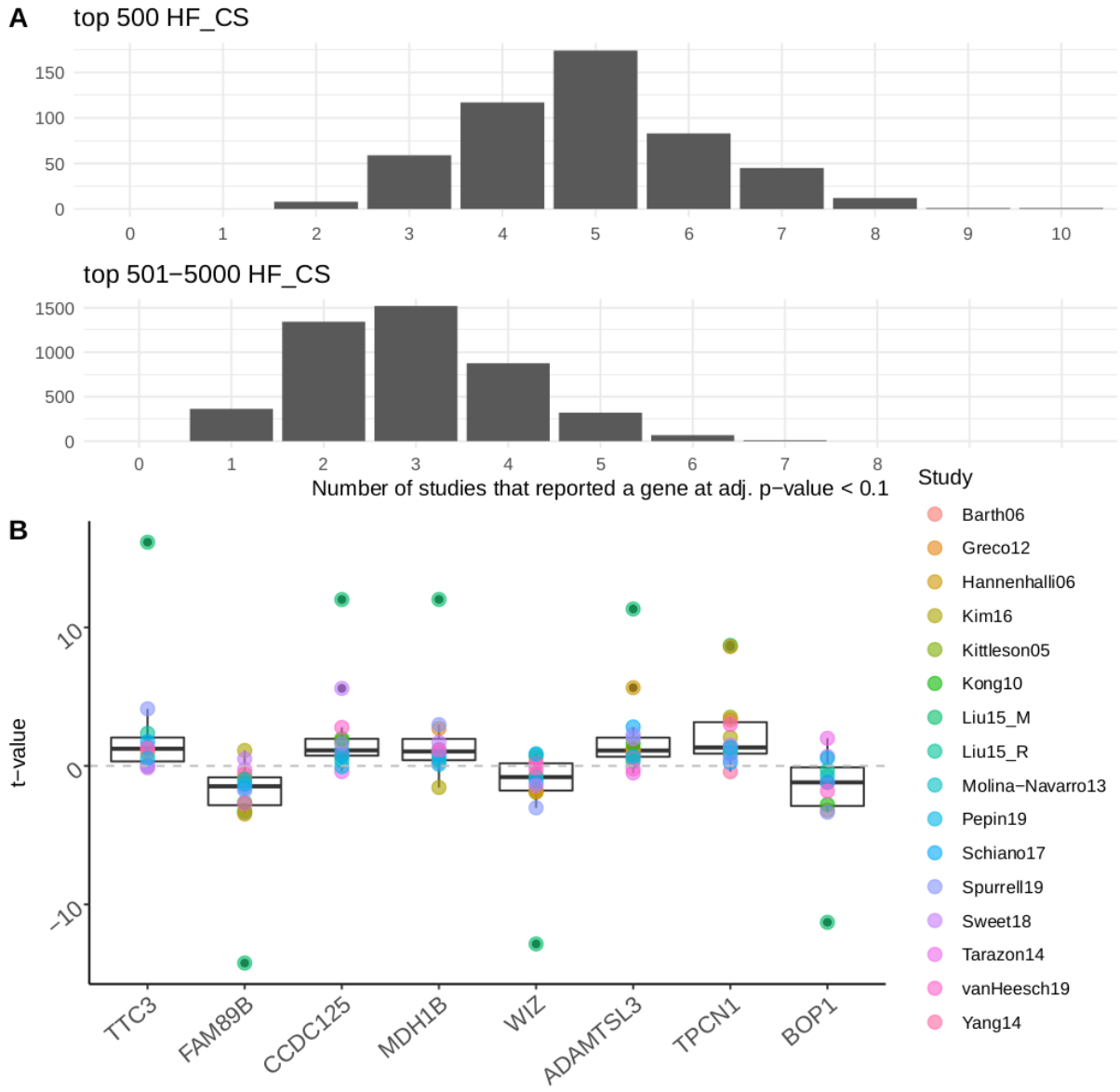
Each vertical panel shows the results of an independent 2-way analysis of variance with HF and another clinical or technical covariate, from an integrated gene standardized data set that only included samples with available information. Upper panels show the proportion of explained variance from each factor as shown by their eta-squared values. Lower panels show the difference in the proportion of variance explained by HF between the top 500 genes of our consensus signatures and the rest.

Figure S13. Proportion of gene expression variance explained by heart failure HF and etiology (DCM [dilated cardiomyopathy] or ICM [ischemic cardiomyopathy]).



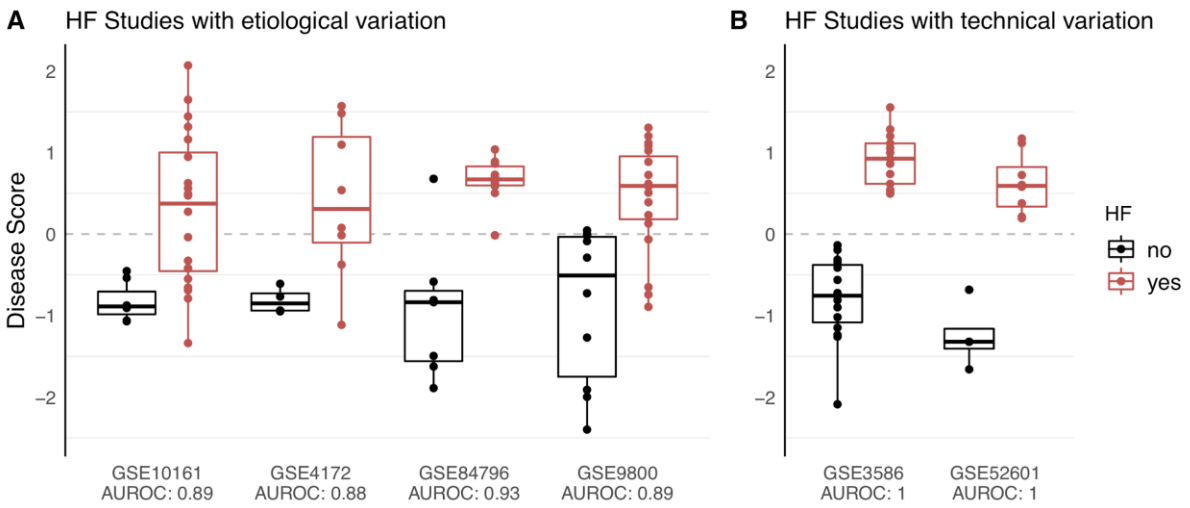
Each panel shows the results of independent 2-way ANOVAs fitted to the top 500 genes from the heart failure consensus signature with HF and DCM as covariates. Each dot represents a different gene and the y-axis is the eta-squared value of each covariate in the ANOVA model.

Figure S14. Added value of the heart failure consensus signature (HF-CS) on single gene level.



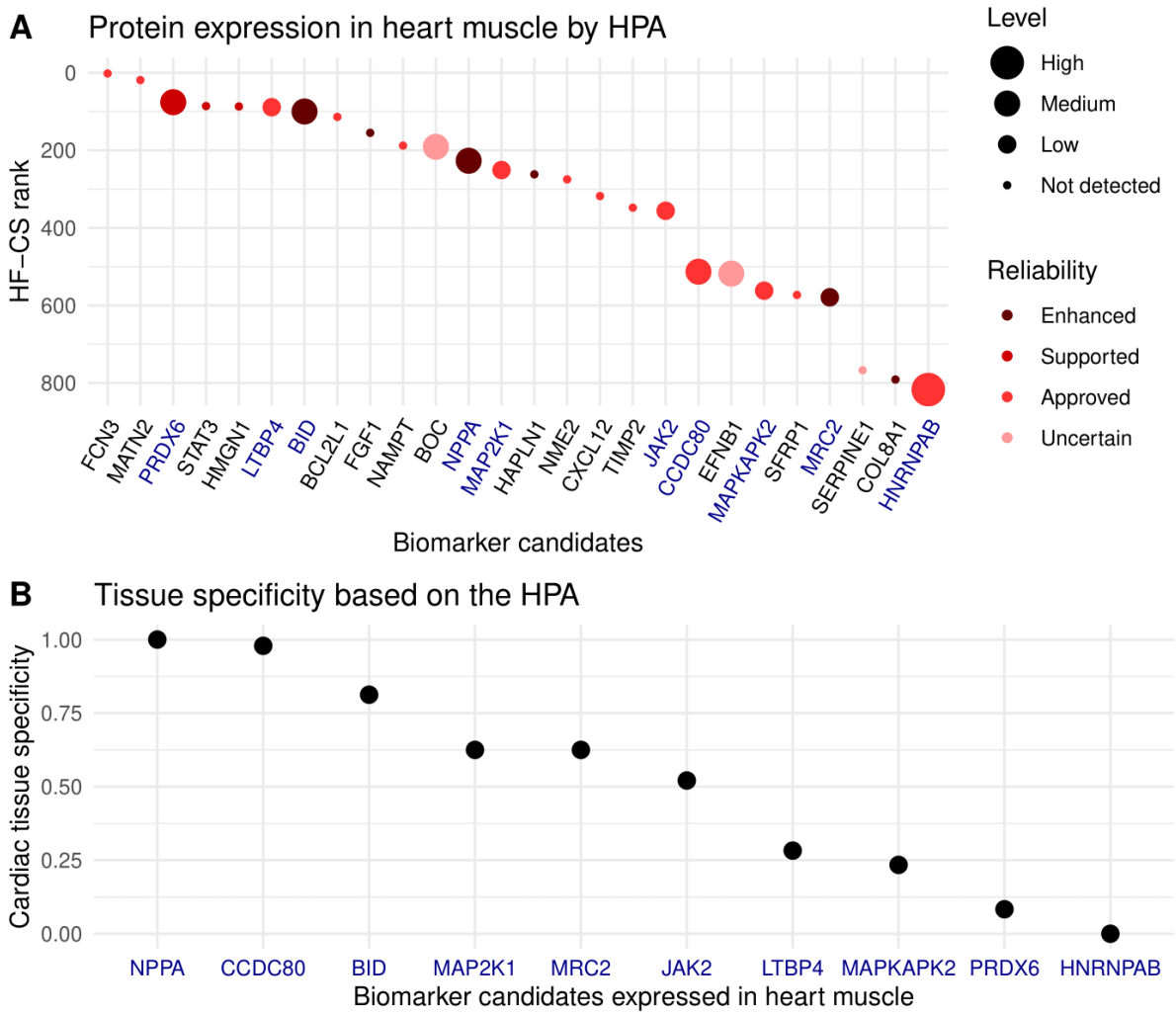
A) Histogram of genes that were reported by single studies (with adj. p-value <0.1), grouped by HF-CS rank < 501 (upper panel) and rank between 501-5000 (lower panel). Distribution of both groups varies significantly (p-value <0.0001, Wilcoxon test). **B)** Genes that were reported by only 2 individual studies (adj. p-value <0.1) and with a HF-CS rank < 500. Single study t-values are displayed for each gene to visualize consistency in expression.

Figure S15. Disease score calculation based on the top 500 genes from the consensus signature for diverse heart failure (HF) studies.



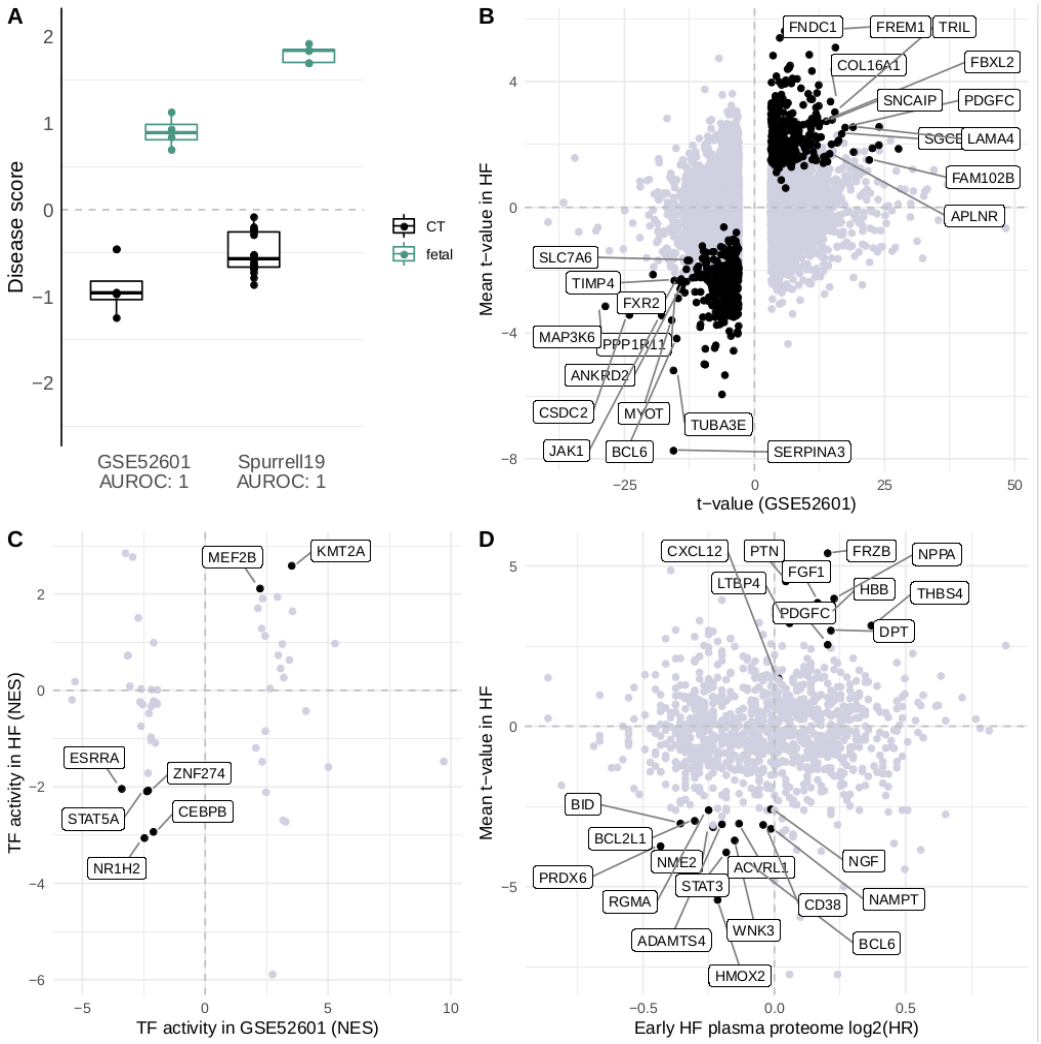
A) HF with diverse etiologies: aortic stenosis (GSE10161); PVB19 infection (GSE4172); chagas disease (GSE84796); eosinophilic myocarditis, alcoholic cardiomyopathy, hypertrophic cardiomyopathy, sarcoidosis, peripartum cardiomyopathy, ischemic cardiomyopathy (ICM), dilatative cardiomyopathy (DCM) (GSE84796). B) HF studies with ICM and DCM samples but processed with different bioinformatic pipelines (GSE3586, GSE52601).

Figure S16. Biomarker candidates and their expression in the Human Protein Atlas (HPA).



A) Relevant biomarker candidates taken from figure 5 and analyzed for their reported protein expression in heart muscle tissue in the HPA. Protein expression was reported for genes labeled in red including PRDX6, LTBP4, BID, BOC, NPPA, MAP2K1, JAK2 with a rank in the heart failure consensus signature (HF-CS) < 500 and CCDC80, MAPKAPK2, MRC2, HNRNPAB with rank between 500-1000. Expression of FRZB, TIMP3, F3 and DPT were not assessed by the HPA. **B)** Assessment of tissue specificity of protein expression using the HPA. The total number of measured non-cardiac tissues in the HPA per candidate ranged between 46 and 48. Tissue specificity was calculated as the ratio of tissues not expressing the protein (Low or Not detected) to the total number of measured tissues. NPPA is not expressed in any non-cardiac tissue. CCDC80 and BID are showing high to moderate specificity while HNRNPAB is suggested to be unsuitable for a cardiac biomarker as it is reported in all non-cardiac tissues.

Figure S17. Heart failure consensus signature (HF-CS) as a reference that complements independent studies.



A) Disease score calculation for fetal experiments Spurrell19 and GSE52601. CT, control (adult non failing heart samples); fetal, fetal heart samples. See supplemental methods for details. B) Significant genes in GSE52601 mapped to the HF-CS. Black dots indicate correlated genes in the enrichment leading edge. Labels indicate genes with a rank < 500 in HF-CS and adjusted p-value < 10e-4.3. C) Significant transcription factors (TFs) in GSE52601 mapped to TFs derived from the HF-CS. Black dots and labels indicate significant and correlated TFs in GSE52601 and HF-CS. D) Plasma proteome of early heart failure patients mapped to the HF-CS. All plasma proteins are displayed. Black dots and labels indicate correlated proteins with a rank < 500 in the HF-CS.