Whole blood transcriptomic profiling identifies molecular pathways related to cardiovascular mortality in heart failure

Contents

Supplementary methods

Pre-processing stage: Checking of raw HTA array data

The HTA 2.0 gene-chip contains 6.9 million short 25mer 'probes' designed using several transcriptome databases. The probes are computationally assembled into groups (probe-sets) to provide a reliable quantification of a transcript or exon. First, the quality of the raw data from arrays were evaluated by standard protocols provided by Affymetrix. These included a hybridisation control (in which eukaryotic hybridisation controls were used to evaluate the sample hybridisation efficiency) and labelling control (Poly-A RNA controls used to monitor the entire target labelling process) and signal box plot (presenting the signal intensity for each sample). The raw data for each patient met the standard quality control features and all quality check data showed values within the acceptable range. The 944 CEL files from the present analysis are deposited at GEO (GSE181114) along with the custom CDF design utilized to process the CEL files.

Deriving the summarized HTA expression set data

To optimise hybridisation in a microarray experiment, probes are designed to constrain the GC content. However, large-format, high-density microarray such as HTA contain millions of probes spanning a wide range of GC content. Hence, GC Correction was applied (Affymetrix Power Tools, Release 1.20.6) to normalise the probe intensities to an intensity distribution that closely matches the distribution for probes with GC count of 12. This transformation, within the range of 20 to 80% GC content, allows for better comparison of expression values across probesets (e.g. ENSG defined genes) and with qPCR and RNA sequencing (http://tools.thermofisher.com/content/sfs/brochures/sst_gccn_whitepaper.pdf).

The 6.9 million 'probes' on the HTA array are combined into groups (probe-sets or transcripts) using a 'map' named the chip definition file (CDF). To optimise the signal 'summarisation' process, poor performing probes are removed as described previously[1]. Briefly, we checked the sequence specificity of the 6.9 million probes on the HTA 2.0 array to reference genome (GRCh38v82p3) using bowtie[2] We removed probes which did not map to a single location which resulted approximately 5M probes. We then assessed the signal characteristics for each

probe using the aroma.affymetrix package (www.aroma-project.org) in R[3], and probes with a very low and invariant signal (e.g. <10 signal units and with a coefficient of variation 25% or less) were removed. A small number of probes with extreme GC contents (~50,000, <20%, >80%) were also removed because the GC adjustment model is not linear at extreme probe compositions. Each remaining probe (2,311,328) was assigned to a gene level 'probe-set' (36,046) based on an ENSEMBL unique ENSG identifiers (all probe-sets that were represented by 3 or more probes were retained). For the current protein coding gene analysis, the HTA data for all patients were processed using this customised 'gene-level' CDF, which incorporates all probes into an artificial single gene 'signal'. Of the 36,046 probe-sets, 17,924 genes were annotated as protein-coding genes in the ensembl database (ENSG), then normalized and logarithm (base 2) transformed and these were used in the present analysis.

Quality checks of summarized HTA expression set data

We implemented several statistics and plots to assess the quality of the summarised RNA expression set data[4–6]. These included:

(1) Relative Log Expression (RLE) plot detects problems in arrays either have larger spread, or will not be centred at zero, or both.
(2) Normalized Unscaled Standard Error (NUSE) plot detects low quality arrays that are significantly elevated or more spread out, relative to the other arrays.
(3) Boxplots of summaries of the signal intensity distributions ($\log_2$ scale) of the genes in each array with outlier detection using the Kolmogorov-Smirnov statistic Ka between each array's distribution and the distribution of the pooled data.
(4) Density plots of signal intensity data (on the $\log_2$ scale)
(5) MA plot using the difference (M) and average of the intensity of a gene on the array and the median intensity of the gene for all arrays with outlier detection using computing Hoeffding's statistic on the joint distribution of A and M for each array.
(6) Heatmap to capture the distance between arrays and detecting outliers using the distance statistic.
(7) Dendrogram to capture distance between arrays and detecting obvious confounding of batch and group as well as strong clustering of batches.

(8) Principal component analysis and comparing principal component scores to detect clustering of array due to potential batch effects
(9) Evaluating the variance mean dependence to capture the density plot of the standard deviation of the intensities across arrays versus the rank of their mean

Post-processed data showed good array quality based on the above criteria. Overall, all plots and statistics were within reasonable and expected range. We also explored the between batch, between centre and between patient variability for the summarised RNA expression set data. Estimates of these variance components were reasonably low. No array consistently failed all the QC measures or identified as an outlier. Scenarios where possible outlying arrays (e.g. by PCA distribution) were identified by a given statistic, we conducted sensitivity analysis and compared the results by including and excluding the relevant arrays. We did not record any substantial changes in the overall outcomes. Therefore, we considered to include all arrays for subsequent analyses.

Additional QC analysis to evaluate batch effects

To evaluate any batch effect independently of the protocol described above, the normalised expression set data, obtained from the R package aroma.affymetrix implementation, were adjusted for potential batch effect using the ComBat function (mean and variance) from the R package sva. All the subsequent QC evaluations were conducted on the batch-adjusted expression set data as described above. The QC analyses showed very marginal changes as a result of batch adjustment and therefore, based on these statistical and bioinformatics outputs, we utilised the non-batch adjusted data for the main analysis.

Whole Blood cell type deconvolution

We utilized the deconvolution method of Monaco *et al* (10) to identify and then adjust the limma analysis for shifts in cell populations. Briefly, Monaco et al. build transcriptomic models of 29 cell types using microarrays and RNA sequencing and FACS cell sorting from two independent cohorts. In the process of developing the method they identified that RNA yield was distinct across cell types, with high yields from dendritic cells and monocytes, but less so from CD4 T cells and these require scaling before incorporating into a deconvolution model. Normalized abundance of marker genes were identified through iterative analysis of T-Cell and B-Cells versus monocyte subtypes, calibrated to deconvolute peripheral PBMC samples.

The method required identified marker genes with high absolute expression values in one cell type over the others, using brute force exhaustive searching. In the present study the absolute deconvolution values generated using the ABIS Shiny app (https://github.com/giannimonaco/ABIS) were scaled to 100% (all cells in the sample) and these scaled values were used as covariates in the limma analysis[8].

Supplementary References

1. Timmons JA, Atherton PJ, Larsson O, Sood S, Blokhin IO, Brogan RJ, Volmar CH, Josse AR, Slentz C, Wahlestedt C, Phillips SM, Phillips BE, Gallagher IJ, Kraus WE. A coding and non-coding transcriptomic perspective on the genomics of human metabolic disease. Nucleic Acids Res. 2018; 46:7772–7792.

2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359.

3. Bengtsson H, Simpson K, Bullard J, Hansen K. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Dep Stat Univ California, Berkeley. 2008; 745:1–9.

4. Kauffmann A, Huber W. Microarray data quality control improves the detection of differentially expressed genes. Genomics. 2010; 138–142.

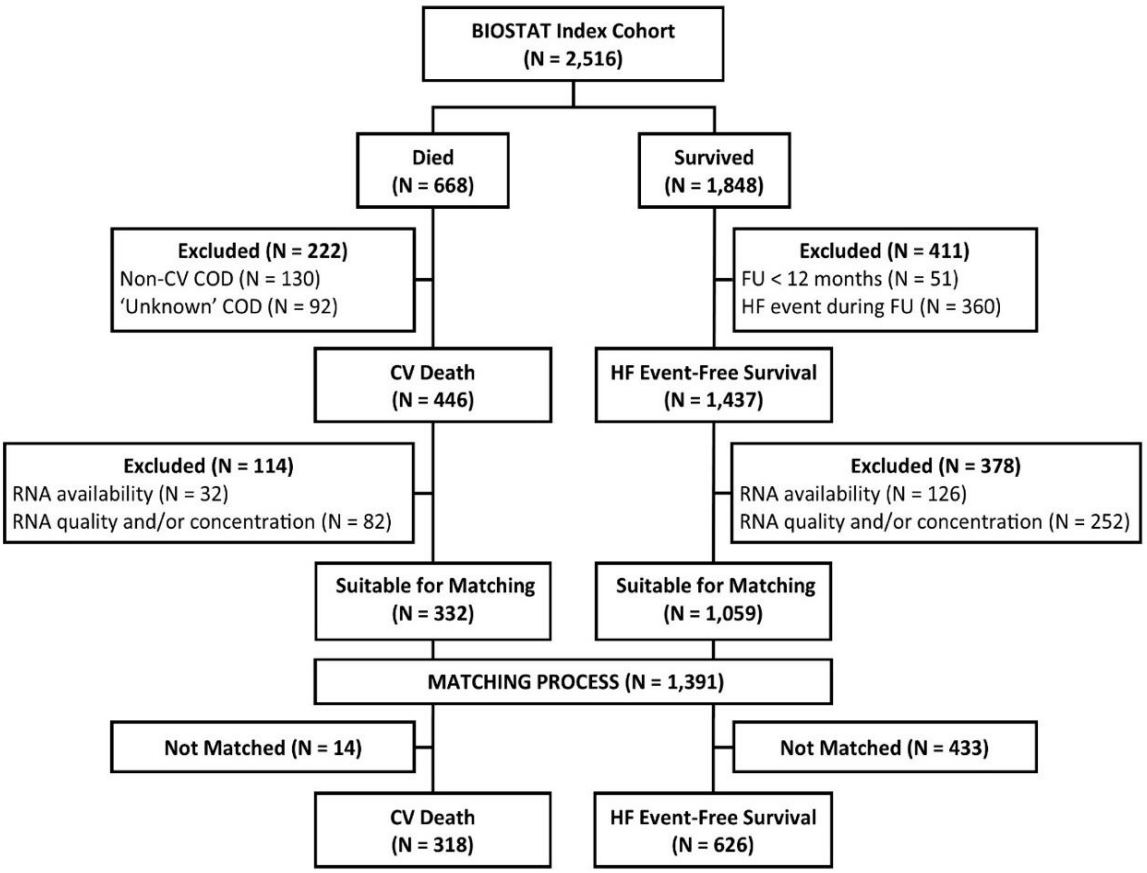
5. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–1735.

6. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleš AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M, Oleś AK, Pagès H, Reyes A, Shannon P, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015; 12:115–121.

7. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, Burdin N, Visan L, Ceccarelli M, Poidinger M, Zippelius A, Pedro de Magalhães J, Larbi A. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. 2019; 26:1627-1640.e7.

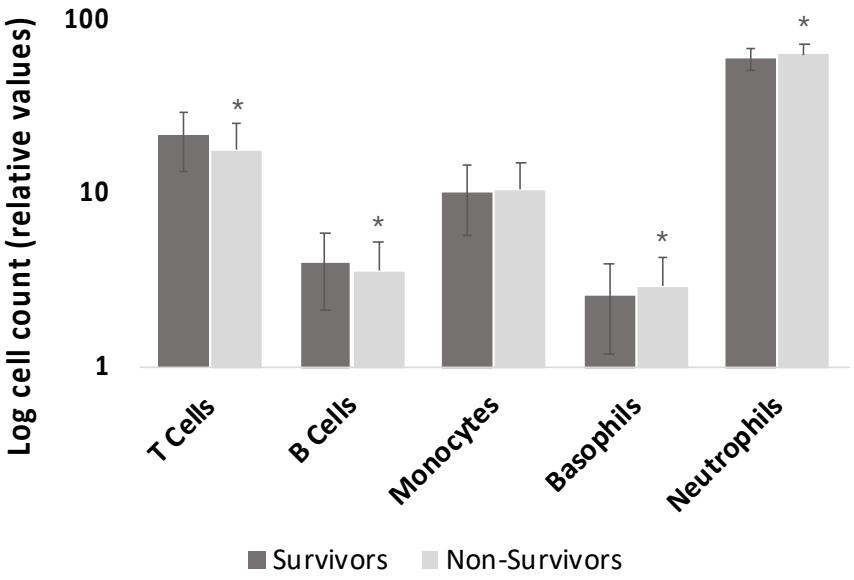
8. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, R. Irizarry WH, eds. Bioinformatics and Computational Biology Solutions using R and Bioconductor, New York Springer. 2005. p. 397—420.

Supplementary Figures 1 – 8

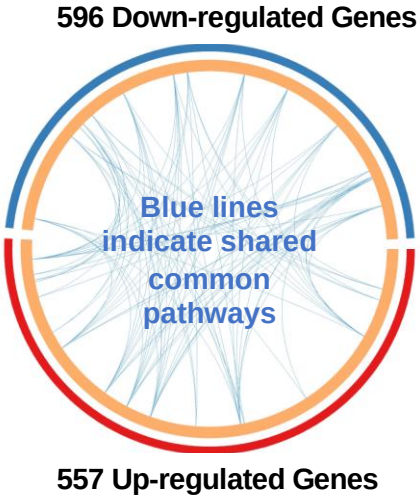


Supplementary Figure 1: CONSORT-style diagram showing the recruitment process and final selection of patients for the BIOSTAT-CHF cohort including the selection of samples for RNA profiling. A total of 318 that died of a cardiovascular cause, and 626 that survived at the point of follow-up, were profiled on the HTA 2.0 Affymetrix Gene Chip. COD = cause of death; CV = cardiovascular; FU = follow-up; HF = heart failure.

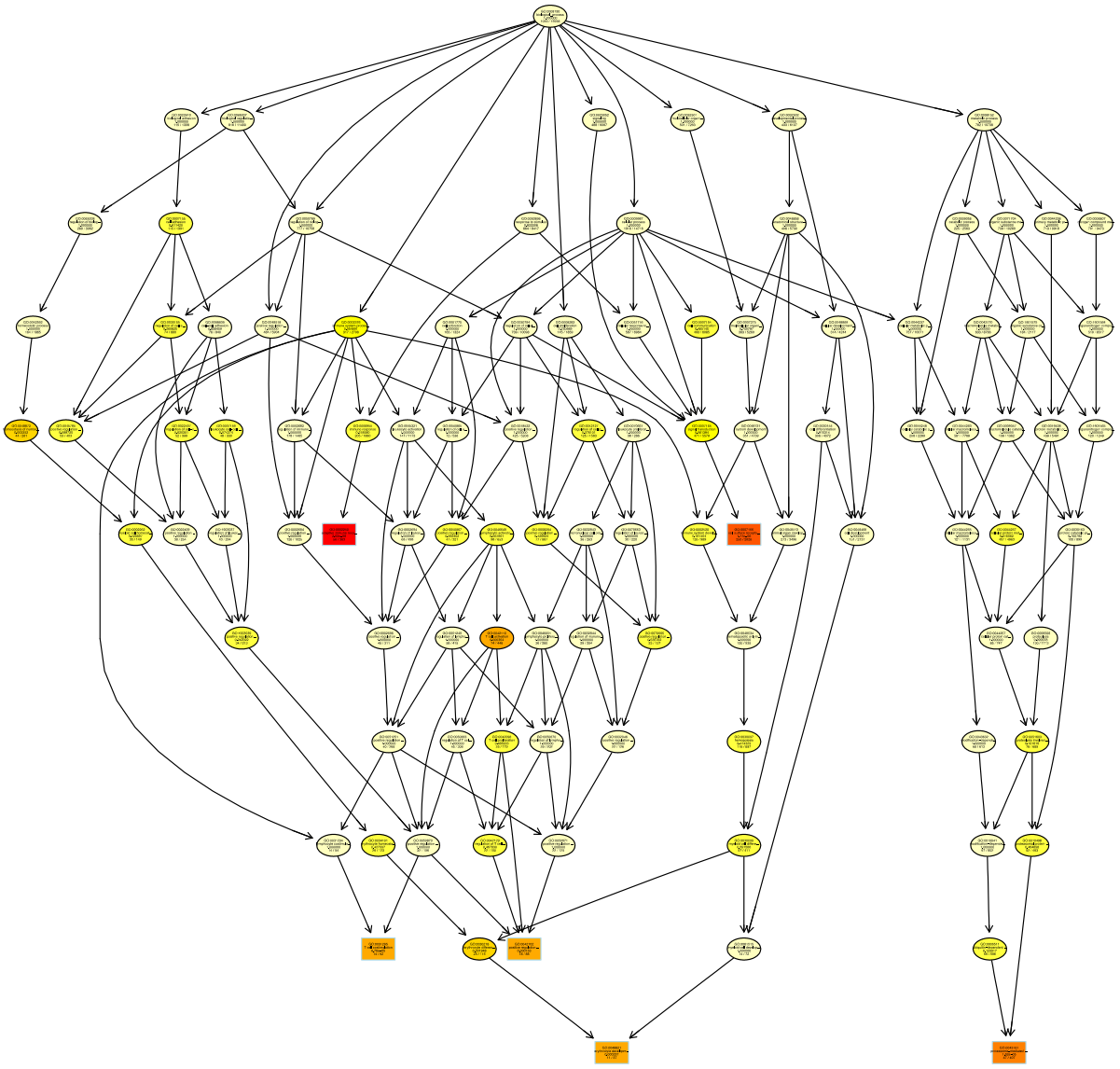
A



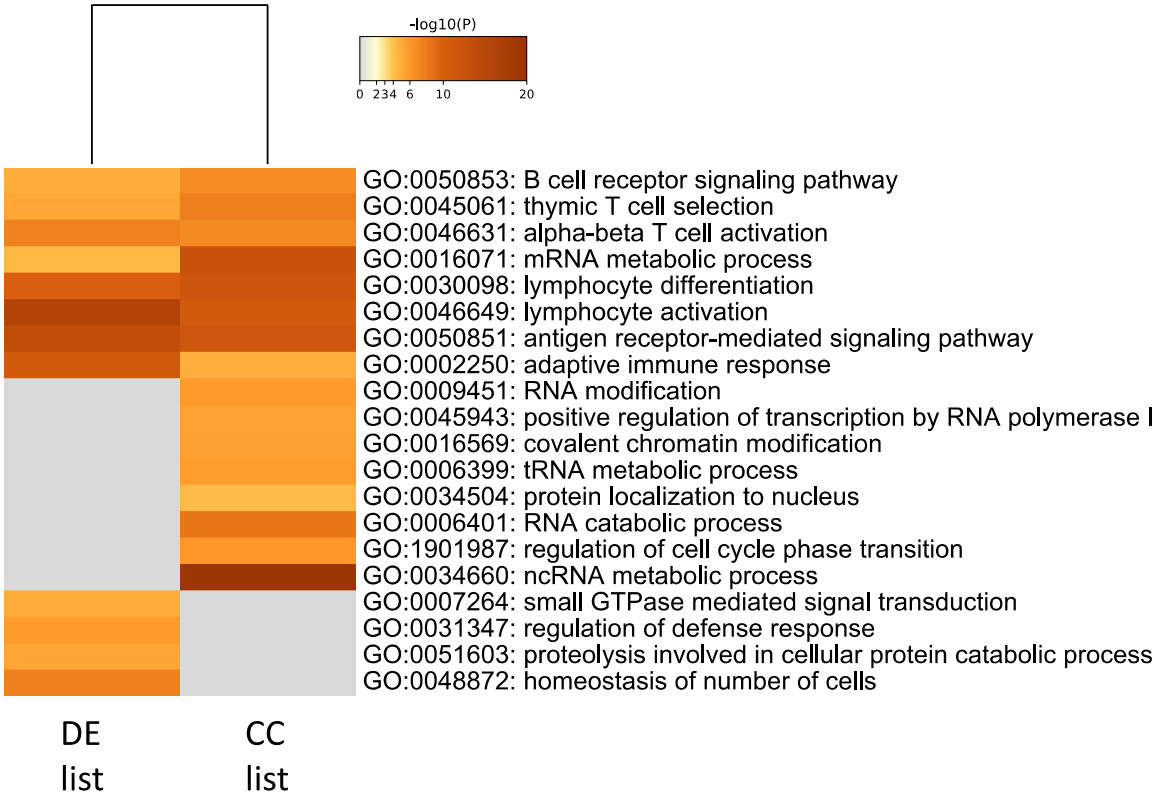
B



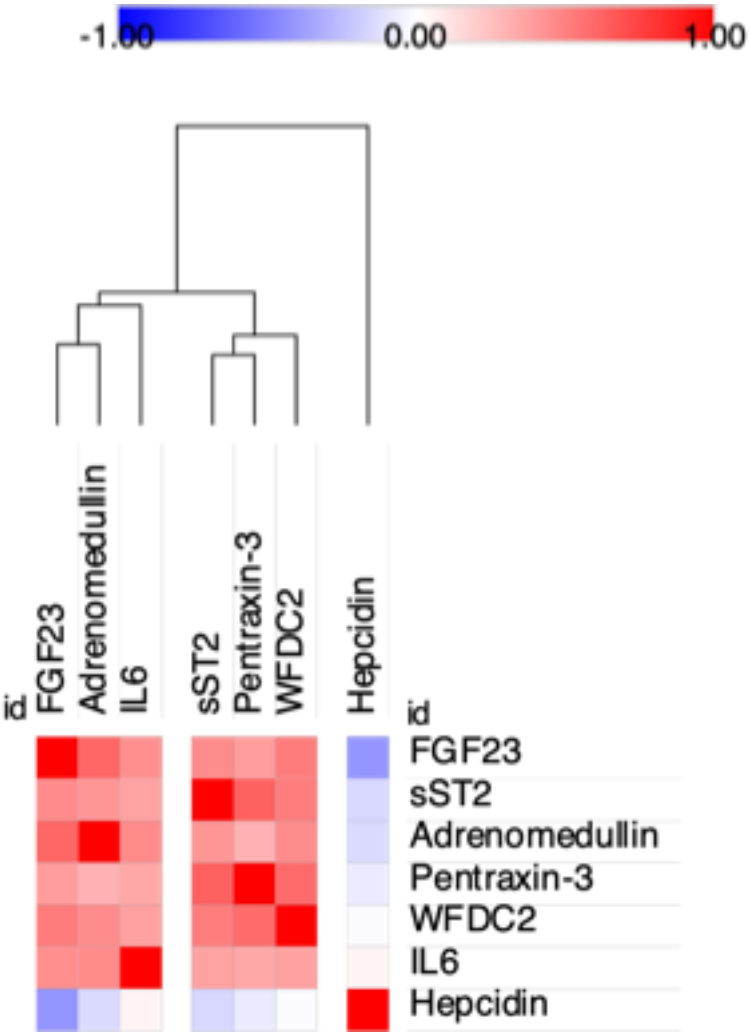
Supplementary Figure 2. Overview of the whole-blood transcriptomic cell count and gene expression differences between survivors versus non-survivors. A). Estimated white cell counts from whole blood RNA are plotted as the natural logarithm of the total count adjusted to 100. The most abundant cell types that demonstrated systematic differences between survivor and non-survivor groups were plotted. *$p<1x10^{-10}$- $p<1x10^{-3}$. Values are mean and standard deviation. **B)** The list of up and down regulated genes were non-overlapping, however the blue lines indicate that the two lists overlapped at the pathway level (biological ontologies).



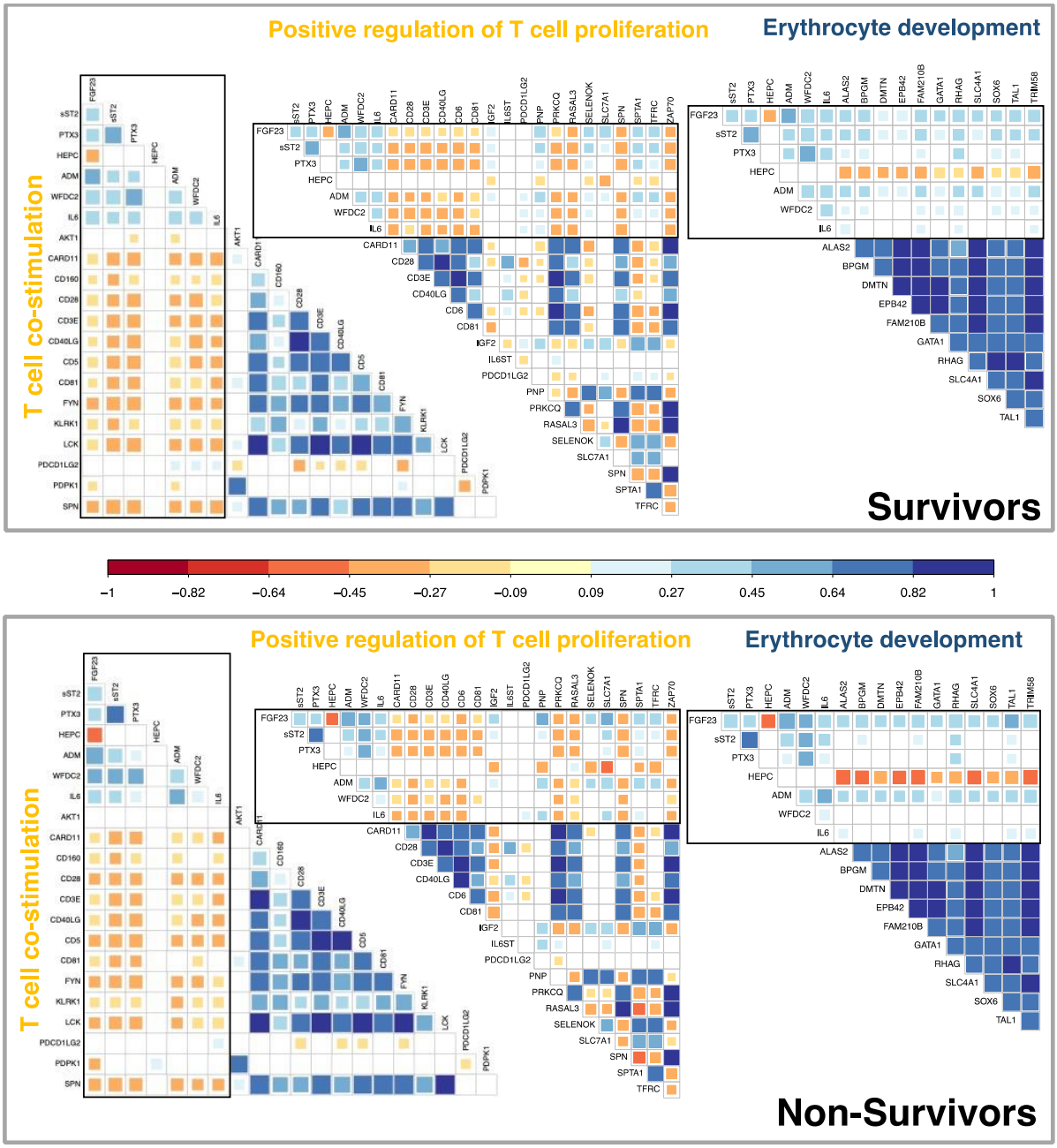
Supplementary Figure 3. GO hierarchical analysis. TopGO identifies 6 GO's (Square Boxes) driving the overall ontology results shown in Figure S1 with 5 demonstrating robust enrichment scores. The categories were: adaptive immune response (56 genes, GO:000225), cell surface receptor signaling pathway (250 genes, GO:0043161), proteasome-mediated ubiquitin-dependent protein catabolic process (47 genes, GO:0007166), T cell costimulation (14 genes, GO:0031295), positive regulation of T cell proliferation (18 genes, GO:0042102) and erythrocyte development (11 genes, GO:0048821). Of these top 6 GO categories 5 were also had robust enrichments ratios i.e. >1.5 more genes than proportionate by pathway content ((GO:000225 (p=2.0E-08), GO:0007166 (p=1.1E-06), GO:0031295 (p=6.8E-05), GO:0042102 (p=1.5E-04) and GO:0048821(p=2.4E-04)) while the sixth (GO:0043161, p=1.6E-05) was 1.3 fold enriched and a category with a large number of genes.



Supplementary Figure 4. Pathway enrichment for whole-blood transcriptomic expression relating to survival and the BIOSTAT Risk Score. Gene ontology (Biological Processes) analysis carried out using metascape.org pathway analysis of genes differentially regulated between survivors and non-survivors, or genes correlated with the BIOSTAT Risk Score (BRS) for all-cause mortality. Each gene list is compared with the 17,748 detected protein-coding genes which was used as the background for calculating enrichment scores and p-values. DE list is the differentially expressed list (FDR <5%, FC > 1.1) **CC list** are genes correlated with the BIOSTAT Risk Score (BRS) for mortality **(**R values <-0.3 or >0.3).

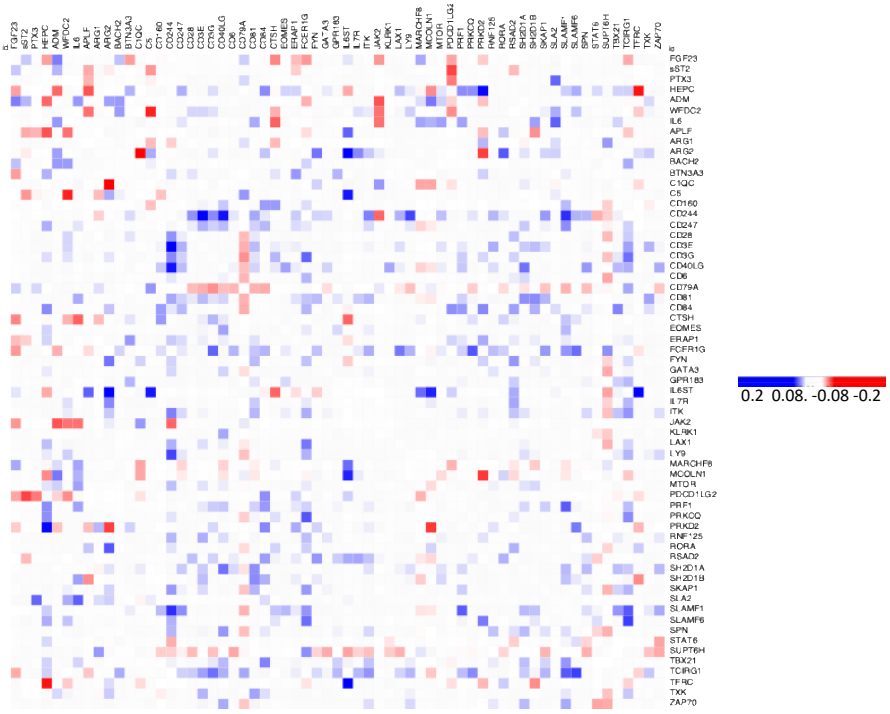


Supplementary Figure 5. Transcriptome identified protein biomarkers. The inter-relationship (Pearson correlation coefficients) using values from all patients, between the seven protein biomarkers that demonstrated an RV coefficient values of 0.2 or greater with the survivor/non-survivor associated differentially expressed pathways.

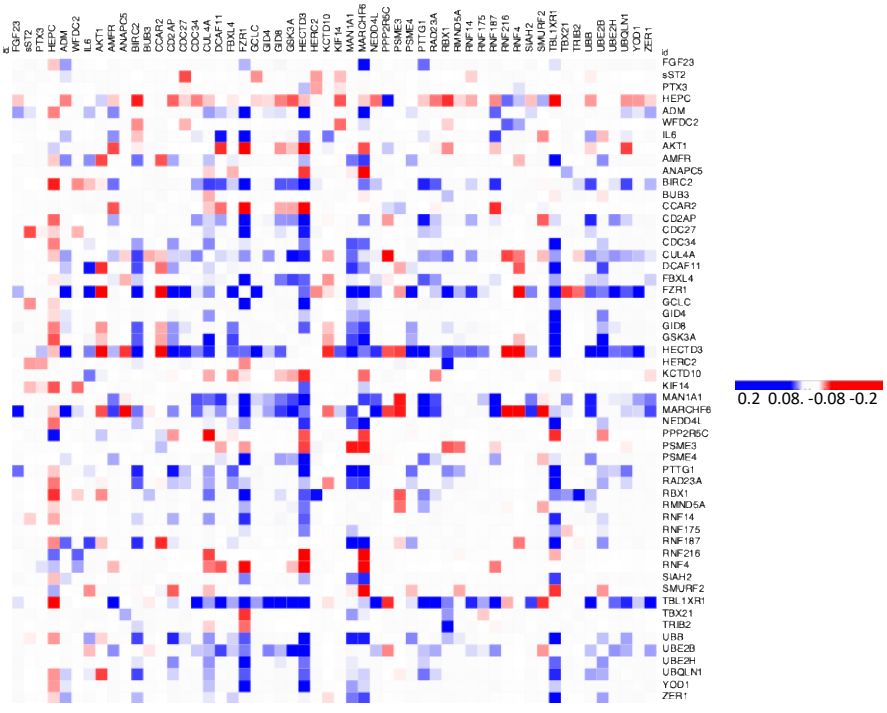


Supplementary Figure 6. Top GO pathway gene and protein biomarkers inter-relationships. Gene expression was correlated with the top protein biomarkers using Pearson correlation coefficients for three top pathways 'T-Cell co-stimulation', 'Positive regulation of T cell proliferation' and 'Erythrocyte development'. The protein values are enclosed by a black oblong box. Data is plotted separately (grey boxes) for survivors (n=626) and non-survivors (n=318). Correlation values are represented by color and are plotted for significant correlations (Bonferroni corrected threshold, $p<1.5x10^{-4}$).

A **Differential correlation for adaptive immune response GO**

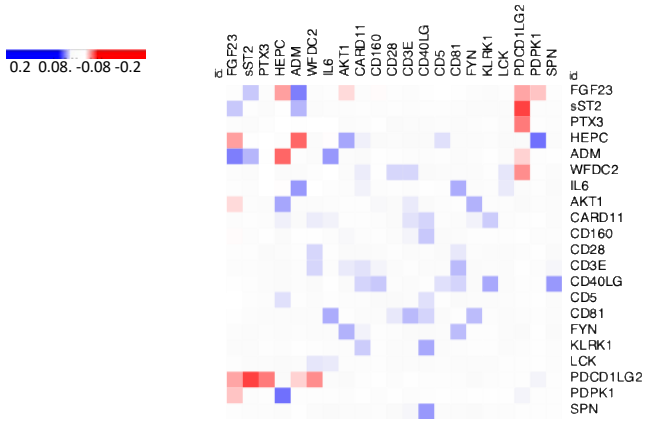


B **Differential correlation for proteasome-mediated ubiquitin-dependent protein catabolic process (UPP) GO**

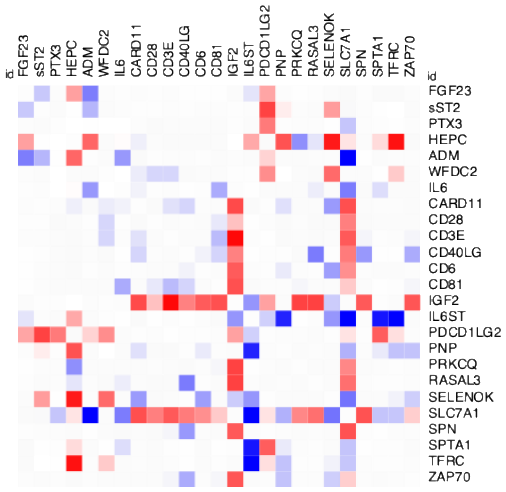


Supplementary Figure 7. Co-expression gene network heatmaps. To aid identification of shifts in the relationship between protein biomarkers and gene expression differential correlation coefficients were plotted from the data presented in Figure 3 (In Figure 3 correlations are only presented if they were significant after Bonferroni p-value correction).

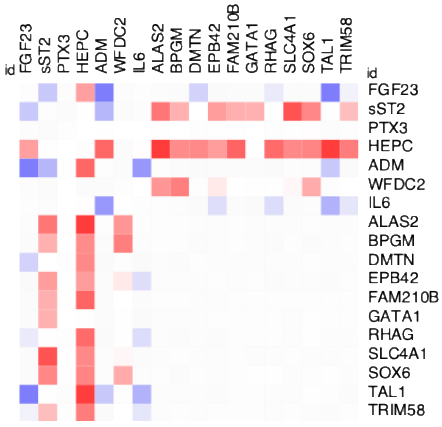
A Differential correlation for T cell co-stimulation GO



B Differential correlation for Positive regulation of T cell proliferation GO



C Differential correlation for Erythrocyte development GO



Supplementary Figure 8. Co-expression gene network heatmaps. To aid identification of shifts in the relationship between protein biomarkers and gene expression differential correlation coefficients were plotted from the data presented in Figure S6 (In Figure S5 correlations are only presented if they were significant after Bonferroni p-value correction)