Haematologica

Supplemental Data

Gender-specific transcriptional profiles identified in β-thalassemia patients.

A. Nanou, C. Toumpeki, P. Fanis, N. Bianchi, L.C. Cosenza, C. Zuccato, G. Sentis, G. Giagkas, C. Stephanou, M. Phylactides, S. Christou, M. Hadjigavriel, M. Sitarou, C. W. Lederer, R. Gambari, M. Kleanthous and E. Katsantoni.

Supplemental Methods

Participants

All participants were enrolled to the study after submission of written informed consent according to the Declaration of Helsinki and approval by the Ethical Committee of Ferrara Hospital, the Ethical Committee of Rovigo Hospital or the Cyprus National Bioethics Committee, as appropriate. 24 samples were collected from the Ferrara and Rovigo Hospitals in Italy (8 healthy subjects, 8 TI patients and 8 TM patients) and 30 samples were collected from the Thalassemia Clinics in Nicosia and Larnaca, the Limassol General Hospital and the Cyprus Institute of Neurology and Genetics in Cyprus (10 healthy subjects, 10 TI patients and 10 TM patients). All samples were organized in 18 groups, each one consisting of one TI patient, one TM patient and one healthy subject (Table S1). All samples per group originated from the same research center or hospital were cultured at the same time, and were gender- and age-matched. The average SD for age per group of matched samples was 1.83 years (range 0.58-3.46 years).

Erythroid Precursor Cell (ErPC) cultures

Separation of PBMCs from peripheral blood was performed using Lympholyte-H Cell Separation Media (Cedarlane Labs), and $CD34^+$ cells were isolated using anti-human $CD34^+$ beads and two rounds of enrichment on pre-chilled and equilibrated MACS LS columns (Miltenyi Biotec). The cells were resuspended in 5 ml expansion medium StemSpan SFEM II (STEMCELL Technologies) supplemented with 1% CC-100 Cytokine Cocktail (STEMCELL Technologies), 2 U/mL erythropoietin, 10^{-6} M dexamethasone and 1x Penicillin/Streptomycin solution. Cell confluency was maintained below $0.5x10^6$ cells/ml during expansion, and erythroid differentiation was initiated around the 14^{th} day of expansion. Expansion time was variable across all samples and unrelated to the analysis group (healthy, TI, TM).

For erythroid differentiation, cells were resuspended in differentiation medium containing 70% MEMα (Corning Cellgro), 30% defined FBS (HyClone Defined FBS), 10^{-5} M 2-mercapto-ethanol, 10 U/ml erythropoietin, 10 ng/ml Stem Cell Factor and 1x Penicillin/Streptomycin solution. ErPC cultures were characterized using flow cytometry staining for surface markers prior to and after differentiation (at the 11^{th} and 13^{th} day of expansion, and at the 4^{th} day of differentiation). The antibodies used for characterization of differentiation stage included APC-conjugated mouse anti-human CD235a monoclonal antibody (BD Pharmingen, 551336), PE-conjugated mouse anti-human CD117 monoclonal antibody (eBioscience, 12-1178-42), PE-conjugated mouse anti-human CD29 monoclonal antibody (eBioscience, 12-0299-42) and PE/Cy7-conjugated rat anti-human CD44 monoclonal antibody (Biolegend, 103030). Cells for RNA-seq analysis were collected on the 4^{th} day of differentiation, where levels of differentiation and cell death were similar between all three analysis groups, as assessed by flow cytometry and visual inspection after cytocentrifugation.

Library Preparation and Sequencing

Total RNA extraction was performed using Tri Reagent (Sigma) and RNA quality was verified prior to library construction using spectrophotometry, electrophoresis and measurement of RNA Integrity Number (RIN) values using an Agilent Bioanalyzer 2100 (RNA 6000 Nano Kit, Agilent, 5067-1511). The RNA-seq libraries were constructed using the TruSeq RNA Sample Preparation kit v2 (Illumina RS-122-2001) using 1.5-2.0 μg of total RNA according to manufacturer's instructions. All sequenced libraries contained single-end reads with no strand specificity and the read

lengths were 50-51 bp depending on the library (Table S2). An Agilent Bioanalyzer 2100 was used to perform quality control of the RNA-seq libraries (DNA chips 1000, Agilent, 5067-1504) and all libraries were sequenced on the Illumina HiSeq2000 high-throughput sequencer.

NGS Data Analysis

After sequencing, quality control of all libraries was performed using the FastQC algorithm.[1] Trimming of library reads due to low base-calling quality (1-3 bp from the start of the read) and removal of primer/adapter sequences was performed using Trimmomatic (v0.30),[2] if necessary. The reads were then aligned to the human transcriptome (hg38), allowing for split reads, using TopHat2.[3] HTSeq (v0.5.4)[4] was used for expression quantification and DESeq2 (v1.8.1)[5] for differential expression analysis with normalization steps for eliminating batch effects using the 'groups' as blocking factor. After analysis, five samples (1 healthy subject, 3 TI patients and 1 TM patient) were excluded from further studies due to the low quality of sequencing or alignment efficiency. In addition, one TM patient was re-classified by medical personnel to TI after the patient selection and library construction. After DE analysis and multiple testing correction, differentially expressed genes were defined as significant when padj < 0.1 in all the performed analyses. More information regarding the patients and libraries quality can be found in the Supplementary Data (Tables S1-S2).

For visualization of expression levels, data matrices were created and represented as heatmaps using the Java TreeView software.[6] Gene set enrichment analysis (GSEA) was performed by ranking the genes according to their $\log_2$FoldChange values (pre-ranked analysis option) and testing them against datasets from the Molecular Signatures Database (MSigDB v6.1).[7] Gene Ontology (GO) analysis was also performed using Metascape[8] either for single gene lists or for comparison of GO terms between multiple lists. Moreover, the data were further explored using the Ingenuity Pathway Analysis (IPA, Qiagen Inc.) for additional interpretation.[9] Data have been deposited in NCBI[10] and are accessible through accession number GSE117221.

Validation by Reverse Transcriptase quantitative PCR (RT-qPCR)

19 participants from Cyprus and Italy were recruited and $CD34^+$ cells isolation/cultures were performed as described above (ErPC cultures). Total RNA was extracted using Trizol (Sigma), treated with RQ1 RNase-Free DNaseI (Promega) and reverse transcribed with MMLV Reverse Transcriptase (Invitrogen), as previously described.[11, 12] Real-time PCR was performed with SYBR Green on an ABI PRISM 7000 Sequence Detection System (Applied Biosystems). The amount of template was normalized using primers for *HPRT*[11] and specific gene primer sequences (5'-3') are: *ELF3*: CAGATGTCATTGGAGGGTACAG (F), CTTCTCCACTTGGTAGCTGATC (R); *SAA1*: CCATTCTGAAGGTGTCTTATCTCC (F), GCCAAGGAACGAA AAGAAGC (R); *TACC2*: AAAAGGAAGCAGCAGGACA (F), CAGAACTC TCAGAAGCGGTG (R). The relative quantitation was performed using the ΔΔCt method.[13] The log2FC values were calculated from the ratios of the ΔΔCt values for TMM vs HM, TMF vs HF, TIM vs HM, TIF vs HF, and compared to RNA-seq data (DESeq2 log2FC).

Supplemental Results

To validate the RNA-seq data, RT-qPCR was performed in randomly selected differentially expressed genes. The genes tested show a corresponding change in expression for RT-qPCR and RNA-seq analyses, with different levels for each method, owing to differences in sample numbers and methodology. However, similar trend was seen for both methods validating gender-related differences that will be further explored with greater patient numbers in future studies (Figure S5).

References

1. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data.[cited; from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]
2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20.
3. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36.
4. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166-9.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.
6. Saldanha AJ. Java Treeview--extensible visualization of microarray data. Bioinformatics. 2004 Nov 22;20(17):3246-8.
7. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.
8. Tripathi S, Pohl MO, Zhou Y, et al. Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. Cell Host Microbe. 2015 Dec 9;18(6):723-35.
9. Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014 Feb 15;30(4):523-30.
10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002 Jan 1;30(1):207-10.
11. Theodorou M, Speletas M, Mamara A, et al. Identification of a STAT5 target gene, Dpf3, provides novel insights in chronic lymphocytic leukemia. PLoS One. 2013;8(10):e76155.
12. Nanou A, Toumpeki C, Lavigne MD, et al. The dual role of LSD1 and HDAC3 in STAT5-dependent transcription is determined by protein interactions, binding affinities, motifs and genomic positions. Nucleic Acids Res. 2017 Jan 9;45(1):142-54.
13. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001 Dec;25(4):402-8.

Supplemental Figures

Figure S1. Protein networks in β-thalassemia. Key proteins in β-thalassemia are depicted and colored according to their gene expression levels (green depicts down-regulation and red depicts up-regulation), as produced by the Ingenuity Pathway Analysis (IPA) program. Only significantly differentially expressed genes are colored according to $\log_2$(Fold Change) values (padj < 0.1). Solid lines show direct interactions, whereas dotted lines show indirect interactions. (A) Gene expression levels were determined after differential expression analysis of TI patients (N = 16) against healthy (H) participants (N = 17). (B) Gene expression levels were determined after differential expression analysis of TM patients (N = 16) against healthy participants (N = 17).

Figure S2. Molecular pathways affected by β-thalassemia. Gene Set Enrichment Analysis (GSEA) was performed by ranking all genes according to their $\log_2$FoldChange values (pre-ranked analysis option) and testing them against datasets from the Molecular Signatures Database. Positive Normalized Enrichment Score (NES) shows enrichment of term in the up-regulated genes, whereas negative NES shows enrichment of term in the down-regulated genes. The molecular pathways identified were yielded after comparison of all TI patients against healthy subjects (A), all TM patients against healthy subjects (B) or after all TM patients against all TI patients (C). Only the top statistically significant terms are shown (FDR q-value < 0.01).

Figure S3. Comparing differentially expressed genes in male and female β-thalassemia patients. (A-B) Venn diagrams depicting common and unique differentially expressed genes showing down- or up-regulation when comparing analyses of all TI patients against healthy subjects (16 TI vs. 17 H), males only (7 TI vs. 8 H) or females only (9 TI vs. 9 H). (C-D) Venn diagrams depicting common and unique differentially expressed genes showing down- or up-regulation when comparing analyses of all TM patients against healthy subjects (16 TM vs. 17 H), males only (8 TM vs. 8 H) or females only (8 TM vs. 9 H).

Figure S4. Gene ontology analysis and gender bias in β-thalassemia. Mosaic graphs produced by IPA depicting enriched terms regarding diseases and body functions per gender. For better visualization, category labels are not shown in full, but detailed enrichment terms can be found in Table S7. (A) On the left, female TI patients (N = 9) were compared to healthy female participants (N = 9) and on the right male TI patients (N = 7) were compared to healthy male participants (N = 8). (B) On the left, female TM patients (N = 8) were compared to healthy female participants (N = 9) and on the right male TM patients (N = 8) were compared to healthy male participants (N = 8). The z-score depicts predicted inhibition or activation of disease/function, whereas the size of the box signifies the possibility of a non-random association ($-\log_{10}$pValue).

Figure S5. Validation of RNA-seq data by RT-qPCR. Relative mRNA expression levels for selected genes were measured by RT-qPCR. The samples used (N=19) included TM patients [N=7, males (N=4), females (N=3)], TI patients [N=6, males (N=3), females (N=3)] and healthy participants [N=6, males (N=3), females (N=3)]. For RT-qPCR (left) the histograms represent the $\log_2$FoldChange values calculated from the ratios of the ΔΔCt values for TMM vs HM, TMF vs HF, TIM vs HM, TIF vs

HF and for RNA-seq (right) the histograms represent the DESeq2 log_2FoldChange values for the same comparisons; M: male, F: female.

Supplemental Tables

Table S1. List of participants used in transcriptomics analysis. The table presents all the participants recruited, as organized in age- and gender-matched groups.

Table S2. List of RNA-seq libraries generated from healthy participants, TI and TM patients. The library size is shown, as well as alignment information to reference genome of each library.

Table S3. Differential expression analysis of TI patients (N=16) versus healthy participants (N=17). The analysis is produced by DESeq2. All significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S4. Differential expression analysis of TM patients (N=16) versus healthy participants (N=17). The analysis is produced by DESeq2. All significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S5. Common and unique genes that show significant differential gene expression between different analyses. The following comparisons are included in the Table: a. DE genes of (TI vs. H) and (TM vs. H) analyses, b. down-regulated DE genes of (TI vs. H) analysis present in males, females and all participants, c. up-regulated DE genes of (TI vs. H) analysis present in males, females and all participants, d. down-regulated DE genes of (TM vs. H) analysis present in males, females and all participants and up-regulated DE genes of (TM vs. H) analysis present in males, females and all participants.

Table S6. Differential expression analysis of TM patients (N=16) versus TI patients (N=16). The analysis is produced by DESeq2.

Table S7. List of enriched disease terms presented in mosaic plots (Figure 1C-D, Figure S4) according to differential expression levels, as produced by IPA software. The lists include the terms, the p-value generated, the z-score, as well as the molecules represented by each term; IPA: Ingenuity Pathway Analysis.

Table S8. Differential expression analysis of female TI patients (N=9) versus female healthy participants (N=9). The analysis is produced by DESeq2. All significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S9. Differential expression analysis of male TI patients (N=7) versus male healthy participants (N=8). The analysis is produced by DESeq2. All significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S10. Differential expression analysis of female TM patients (N=8) versus female healthy participants (N=9). The analysis is produced by DESeq2. All significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S11. Differential expression analysis of male TM patients (N=8) versus male healthy participants (N=8). The analysis is produced by DESeq2. All

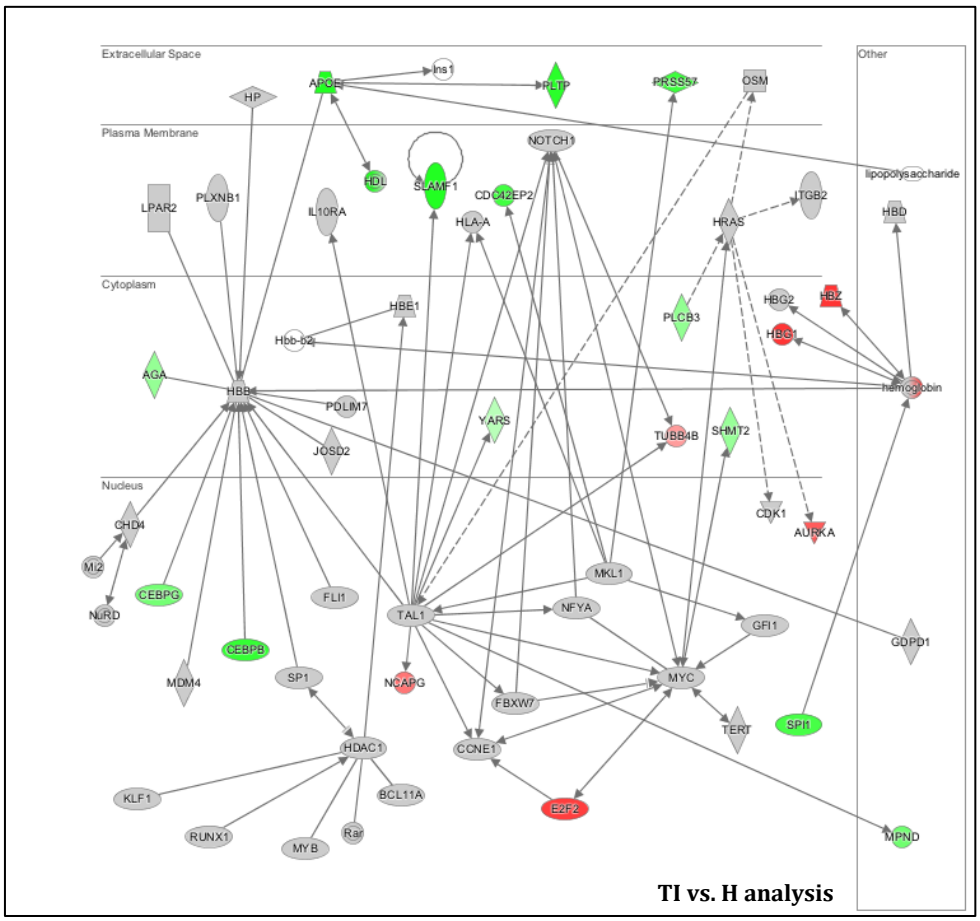
significantly differentially expressed genes with p-adjusted value below 0.1 are highlighted in red.

Table S12. Differential expression analysis of female TM patients (N=8) versus female TI patients (N=9). The analysis is produced by DESeq2. The significantly differentially expressed gene with p-adjusted value below 0.1 is highlighted in red.

Table S13. Differential expression analysis of male TM patients (N=8) versus male TI patients (N=7). The analysis is produced by DESeq2.

Figure S1

A


Extracellular Space
Other
Plasma Membrane
Cytoplasm
Nucleus
APOE
Ins1
PLTP
PRSS57
OSM
HP
NOTCH1
HDL
SLAMF1
lipopolysaccharide
LPAR2
PLXNB1
IL10RA
CDC42EP2
HLA-A
HRAS
ITGB2
HBD
HBE1
HBG2
HBZ
PLCB3
Hbb-b2
HBG1
AGA
HBB
hemoglobin
PDLIM7
YARS
TUBB4B
SHMT2
JOSD2
CDK1
AURKA
CHD4
MI2
MKL1
CEBPG
FLI1
TAL1
NFYA
GFI1
NuRD
GDPD1
CEBPB
MDM4
SP1
NCAPG
MYC
FBXW7
TERT
SPI1
HDAC1
CCNE1
KLF1
BCL11A
E2F2
RUNX1
MYB
Rar
MPND
TI vs. H analysis

B

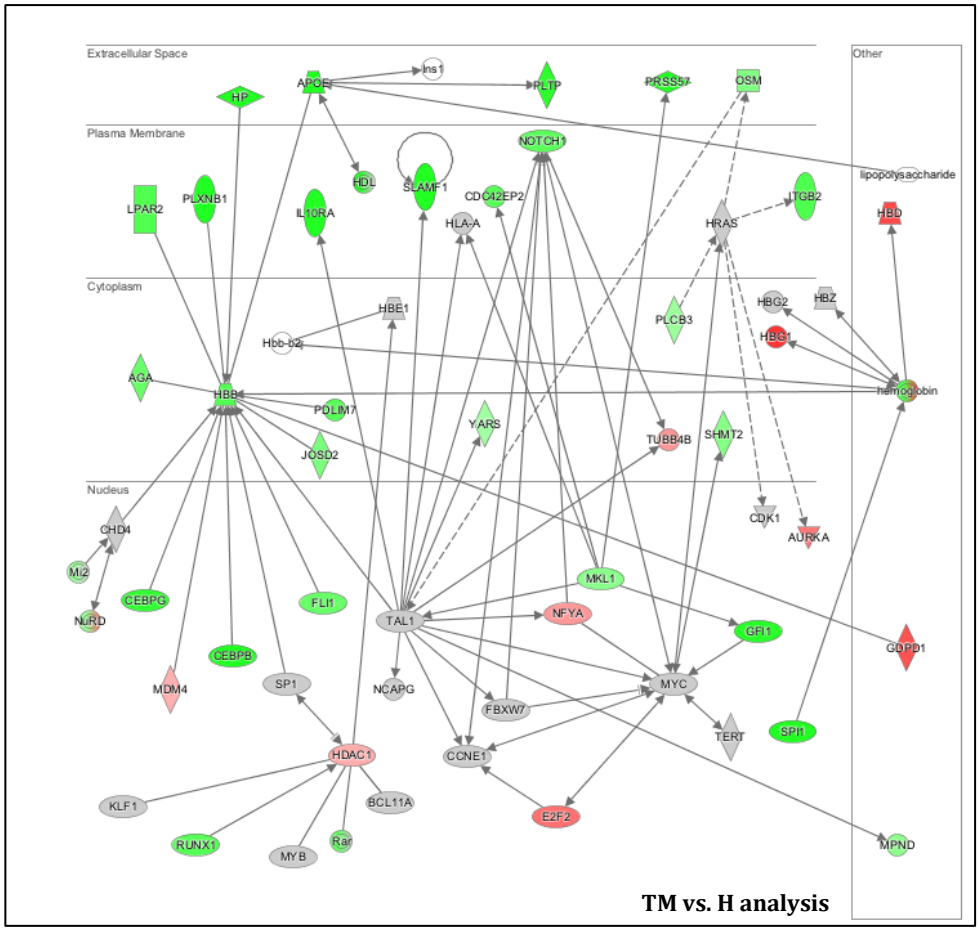

Extracellular Space
Other
Plasma Membrane
Cytoplasm
Nucleus
APOE
Ins1
PLTP
PRSS57
OSM
HP
NOTCH1
HDL
SLAMF1
lipopolysaccharide
LPAR2
PLXNB1
IL10RA
CDC42EP2
HLA-A
HRAS
ITGB2
HBD
HBE1
HBG2
HBZ
PLCB3
Hbb-b2
HBG1
AGA
HBB
hemoglobin
PDLIM7
YARS
TUBB4B
SHMT2
JOSD2
CDK1
AURKA
CHD4
MI2
MKL1
CEBPG
FLI1
TAL1
NFYA
GFI1
NuRD
GDPD1
CEBPB
MDM4
SP1
NCAPG
MYC
FBXW7
TERT
SPI1
HDAC1
CCNE1
KLF1
BCL11A
E2F2
RUNX1
MYB
Rar
MPND
TM vs. H analysis

Figure S2

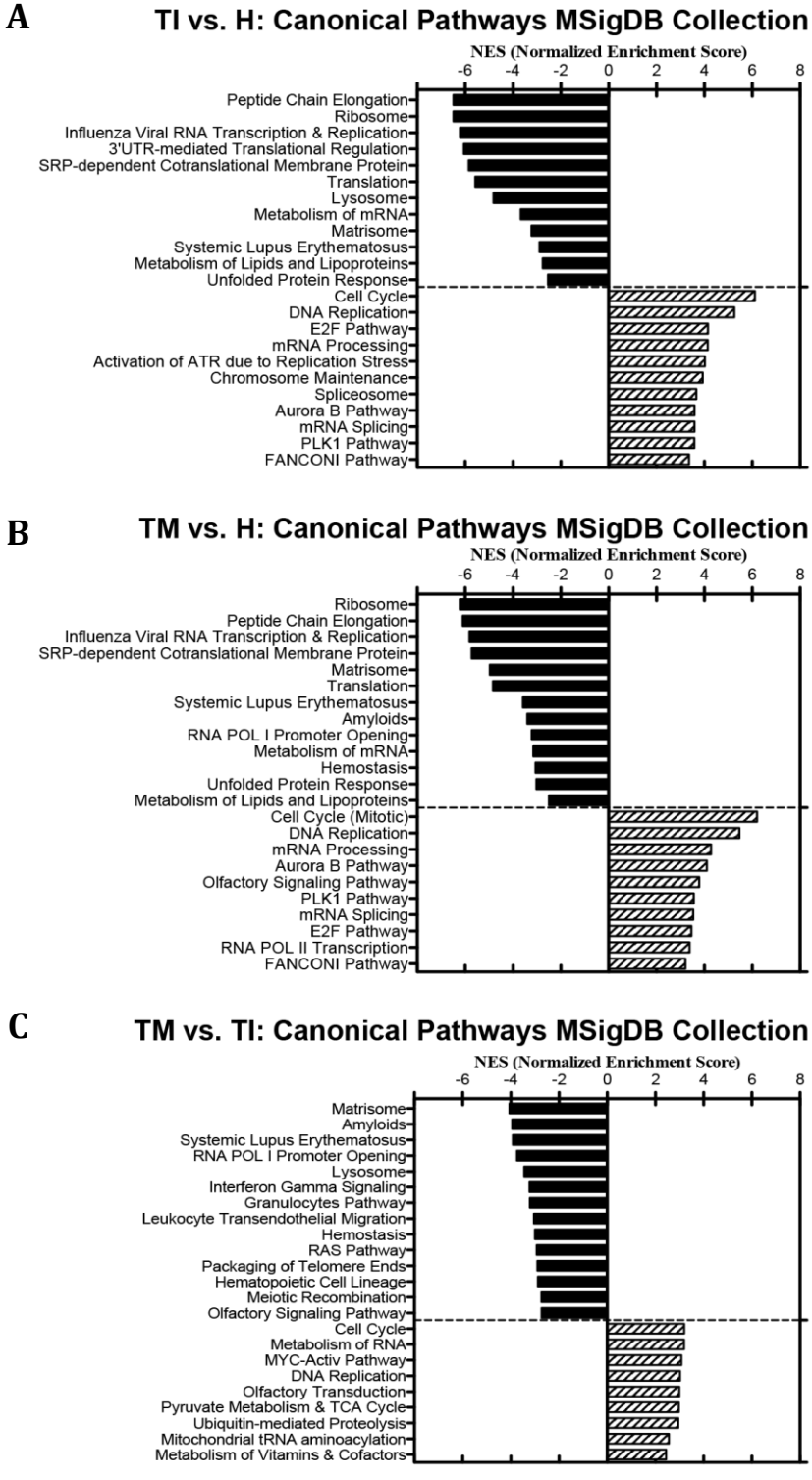

A
TI vs. H: Canonical Pathways MSigDB Collection
NES (Normalized Enrichment Score)
-6 -4 -2 0 2 4 6 8
Peptide Chain Elongation
Ribosome
Influenza Viral RNA Transcription & Replication
3'UTR-mediated Translational Regulation
SRP-dependent Cotranslational Membrane Protein
Translation
Lysosome
Metabolism of mRNA
Matrisome
Systemic Lupus Erythematosus
Metabolism of Lipids and Lipoproteins
Unfolded Protein Response
Cell Cycle
DNA Replication
E2F Pathway
mRNA Processing
Activation of ATR due to Replication Stress
Chromosome Maintenance
Spliceosome
Aurora B Pathway
mRNA Splicing
PLK1 Pathway
FANCONI Pathway
B
TM vs. H: Canonical Pathways MSigDB Collection
NES (Normalized Enrichment Score)
-6 -4 -2 0 2 4 6 8
Ribosome
Peptide Chain Elongation
Influenza Viral RNA Transcription & Replication
SRP-dependent Cotranslational Membrane Protein
Matrisome
Translation
Systemic Lupus Erythematosus
Amyloids
RNA POL I Promoter Opening
Metabolism of mRNA
Hemostasis
Unfolded Protein Response
Metabolism of Lipids and Lipoproteins
Cell Cycle (Mitotic)
DNA Replication
mRNA Processing
Aurora B Pathway
Olfactory Signaling Pathway
PLK1 Pathway
mRNA Splicing
E2F Pathway
RNA POL II Transcription
FANCONI Pathway
C
TM vs. TI: Canonical Pathways MSigDB Collection
NES (Normalized Enrichment Score)
-6 -4 -2 0 2 4 6 8
Matrisome
Amyloids
Systemic Lupus Erythematosus
RNA POL I Promoter Opening
Lysosome
Interferon Gamma Signaling
Granulocytes Pathway
Leukocyte Transendothelial Migration
Hemostasis
RAS Pathway
Packaging of Telomere Ends
Hematopoietic Cell Lineage
Meiotic Recombination
Olfactory Signaling Pathway
Cell Cycle
Metabolism of RNA
MYC-Activ Pathway
DNA Replication
Olfactory Transduction
Pyruvate Metabolism & TCA Cycle
Ubiquitin-mediated Proteolysis
Mitochondrial tRNA aminoacylation
Metabolism of Vitamins & Cofactors

Figure S3

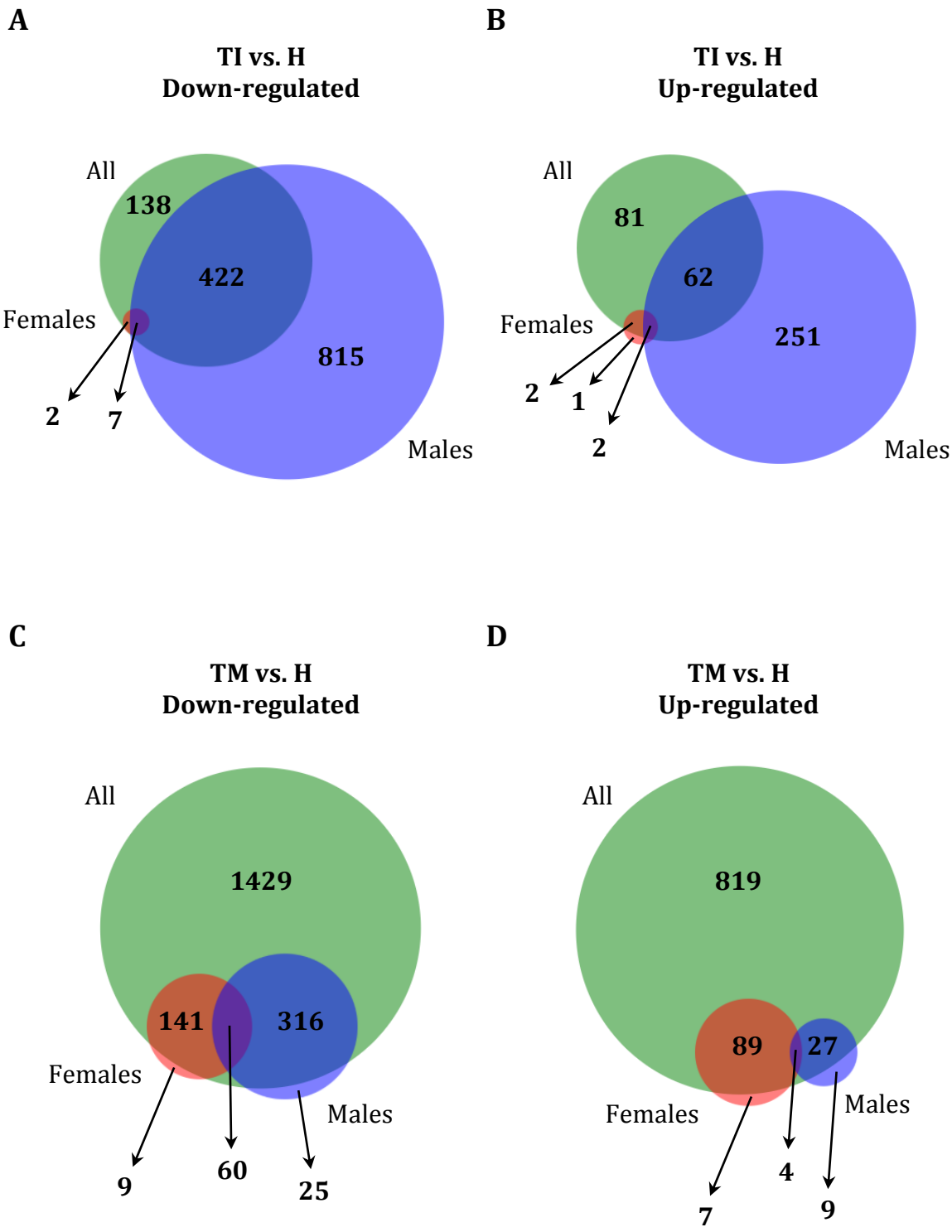

A
TI vs. H
Down-regulated
All
138
422
Females
815
2
7
Males
B
TI vs. H
Up-regulated
All
81
62
Females
251
2
1
2
Males
C
TM vs. H
Down-regulated
All
1429
141
316
Females
Males
9
60
25
D
TM vs. H
Up-regulated
All
819
89
27
Females
Males
7
4
9

Figure S4

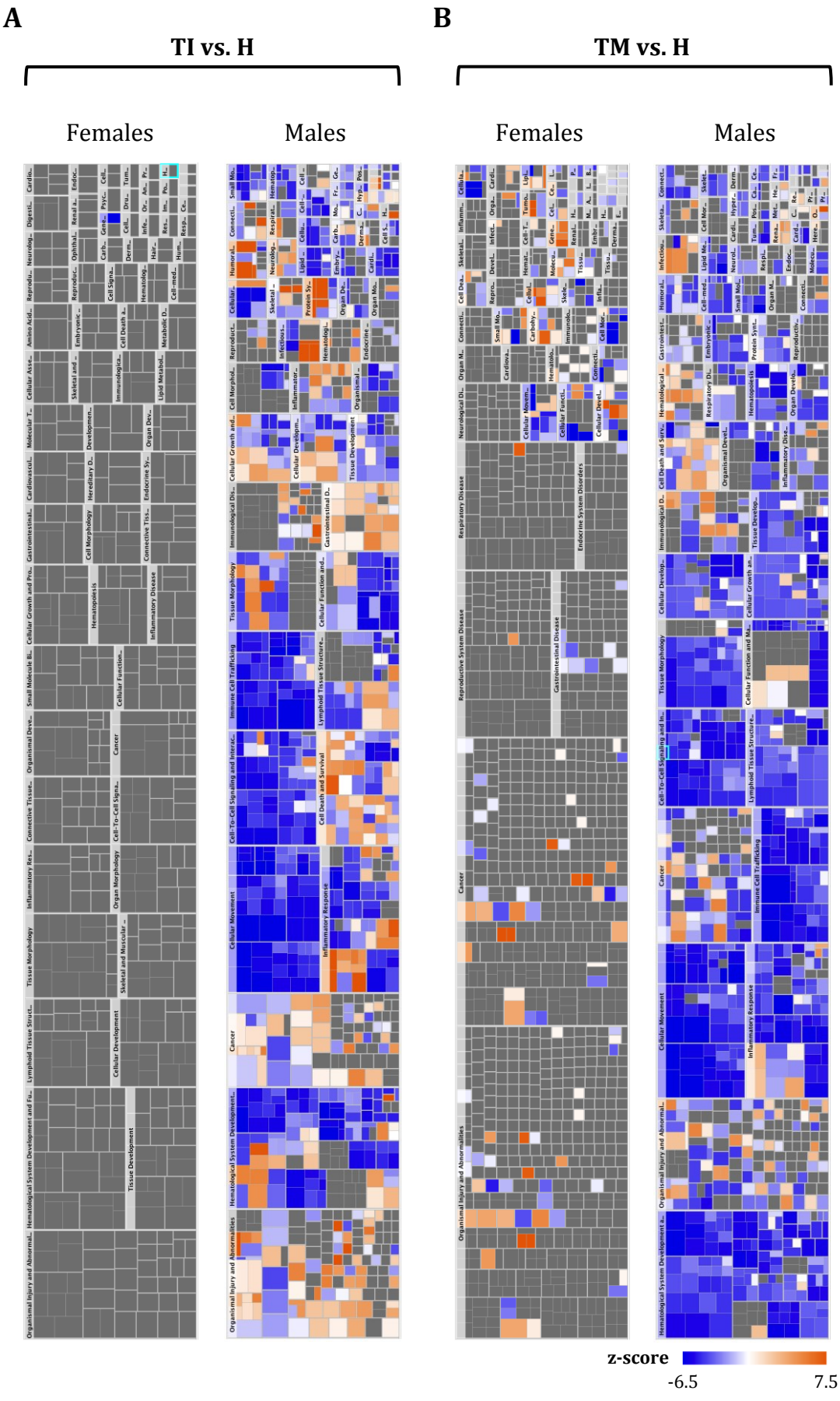
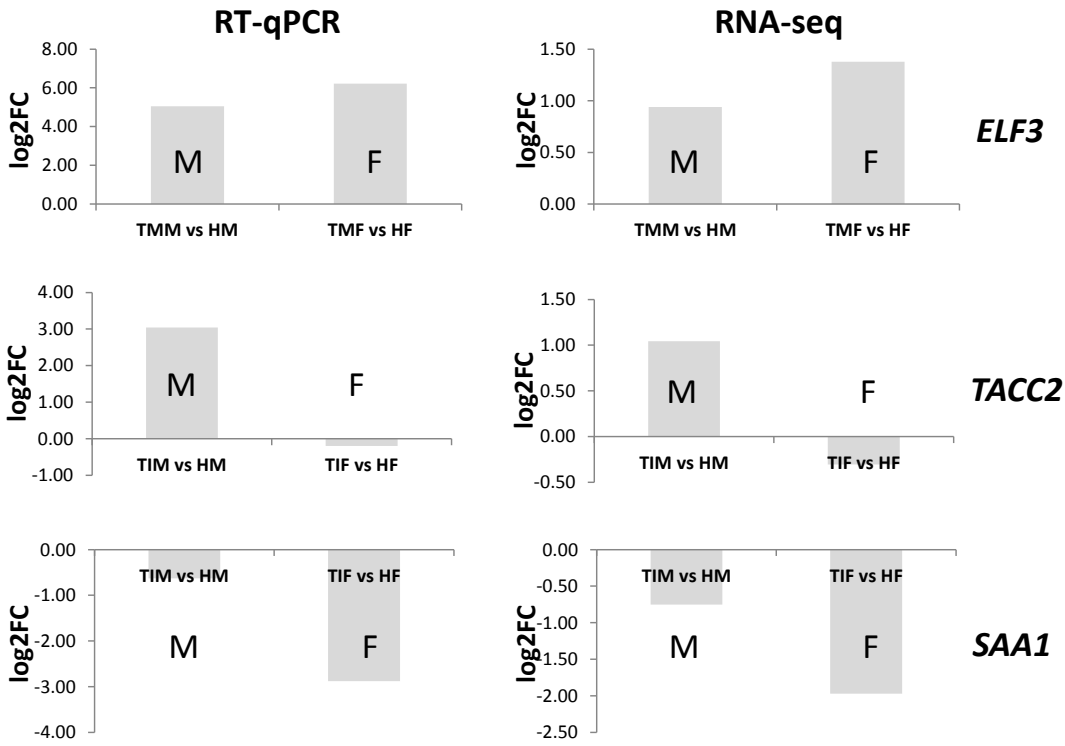


Figure S5


RT-qPCR
RNA-seq
log2FC
8.00
6.00
4.00
2.00
0.00
M
F
TMM vs HM
TMF vs HF
1.50
1.00
0.50
0.00
ELF3
4.00
3.00
2.00
1.00
0.00
-1.00
TIM vs HM
TIF vs HF
-0.50
TACC2
-2.00
-3.00
-4.00
SAA1