Supplementary materials and methods

Participants

Sepsis patients were prospectively recruited between January 2015 and August 2017 from Persistent Inflammation and Immunosuppression in Sepsis (PICS; NCT02276066) study that examines the immunological mechanisms of chronic critical illness in a prospective longitudinal cohort of surgical patients with sepsis at University of Florida Health (UFH). For the control group, we used preoperative urine samples from patients prospectively recruited between July 2015 and February 2018 to Network Analysis of Urinary Molecular Signature (NavigateAKI; NCT02114138) study characterizing the urinary molecular response to surgical stress among patients undergoing high-risk vascular surgery at UFH (Supplementary Figure S1). The study protocols were finalized and ethics approvals were obtained from UF Institutional Review Board (IRB201400611 and IRB201400127) prior to the recruitment of patients, and all methods were performed in accordance to relevant guidelines and regulations. All study participants provided written informed consent. There was no overlap of patients between the two cohorts.

The inclusion criteria for the sepsis cohort were admission to the surgical intensive care unit, age ≥ 18 years, and a clinical diagnosis of sepsis by attending intensivist with subsequent initiation of the computerized sepsis protocol. Patients who were taking specified immunosuppressive drugs (e.g., prednisone with a dose greater than 40mg) and those with inherited diseases (e.g, severe Crohn's disease, lupus) that affect the immune system were not enrolled. Patients with advanced liver or heart disease were excluded. For this analysis we excluded patients with end stage renal disease or urinary tract infection as the primary source of sepsis. The final sepsis diagnosis was clinically adjudicated by investigators during weekly adjudication meetings according to the American College of Chest Physicians consensus criteria. The NavigateAKI study recruited patients ≥ 18 years prior to high-risk vascular surgery at UF

Health and followed them for one year after surgery. All patients were adjudicated as having no evidence of infection prior to surgery by attending surgeons and investigators. Supplementary Table S3 shows the spectrum of surgeries these patients were scheduled for. This table contains Current Procedural Terminology (CPT) codes for the respective surgery for ease of understanding.

All relevant clinical data was prospectively collected. Severity of illness was defined within the first 24 hours using the Sequential Organ Failure Assessment Score. Patient outcomes, including hospital and 12-months mortality, were prospectively recorded. The first blood and urine samples for experimental analyses were collected within twelve hours of sepsis onset for sepsis patients and within four hours of scheduled surgery for control patients.

Discovery and Validation Cohorts

The discovery cohort consisted of RNA isolated from 238 patients recruited between January 2015 and March 2016 (Supplementary Figure S1). The prospective validation cohort consisted of RNA isolated from 110 patients recruited between February 2017 and February 2018. Complete data was available for 146 sepsis and 32 control patients in the discovery cohort and 41 sepsis and 32 control patients in the validation cohort. This sample size enabled us to ensure that for at least 85% of probes we have power greater than 80% to detect a twofold-change between the mean expressions for sepsis and control patients using a two-sided independent T-test with Bonferroni adjustment at a family-wise type 1 error of 0.05.

Processing of urine samples and ribonucleic acid purification

Using standardized protocols to separate cell pellets from urine supernatant (Figure 1A), approximately 50 mL of urine were collected in sterile manner at the bedside and processed within two hours of collection. We used previously described protocols to isolate total cellular RNA from the urinary cell pellet containing all cellular elements. In brief, the 50mL Urine was spun down at 1500g for 30 min at 4°C. The pellet was collected, lysed using 1mL RLT lysis buffer with 10uL β-mercaptoethanol from the kit and processed according the manufacturer's protocol. Total RNA

was extracted using Qiagen RNeasy mini kit (250) Catalog Number - 74106 according to manufacturer's protocol. To determine quality of isolated cellular RNA, we measured the quantity (absorbance at 260 nm) and purity (ratio of the absorbance at 260 and 280 nm). An RNA sample was classified as having passed quality control if the optical density 260 to 280 ratio was between 1.5 and 2.2 and final concentration was at least 8.7 µg/ml (25) (Supplementary Table S1).

Microarrays

Biotin-labeled sense strand complementary deoxyribonucleic acid (cDNA) was prepared from 300 ng of total RNA per sample using an Affymetrix GeneChip™ Whole Transcript Sense Target Labeling Assay per standard protocol. In brief, cDNA was amplified using GeneChip™ WT PLUS Reagent Kit, from ThermoFisher Scientific following manufacturer's instructions. Amplified cDNA was fragmented with UDG and APE1 enzymes according to ThermoFisher Scientific instructions. Labeling was performed with TdT with biotin reagent followed by streptavidin/phycoerythrin. Hybridization to GeneChip™ Human Transcriptome Array (HTA 2.0) was carried out at 45°C for 16 hours and the arrays were scanned on an Affymetrix GeneChip Scanner 3000 7G using AGCC software, which produced a set of DAT, CEL, JPG, and XML files for each array. Image analysis and probe quantification were performed using the Affymetrix software that produced raw probe intensity data in the Affymetrix CEL files. Transcriptome Analysis Console version 4.0.1 (Thermo Fisher Scientific, Santa Clara, CA) was used for microarray signal summarization and normalization using (Figure 1B) robust multi-array average (RMA) [28]. RMA equalizes test distributions by aligning their quantiles after ranking genes by their expression values. It does not make use of endogenous control or house-keeping genes. The final microarray dataset consisted of log2 transformed expression values for 67,528 probes of which 33,494 were mapped to one or more known gene(s) (available as GSE112098, GSE112099 and GSE112100 GEO series accessions).

Identification of cell-specific transcripts

The 33,494 probes mapped to known genes were used to estimate the immune and kidney cell composition of the samples (Figure 1B). The immune response in silico (IRIS) repository of 1,622 genes, classified by their specific expression in multiple immune cell lineages and previously described transcript sets of 637 genes for kidney-specific cell lineages were utilized to estimate the immune and renal cell composition in urine, respectively. Transcript sets which are unique to either one of the different immune cell types characterized by the IRIS dataset namely – Neutrophil, Dendritic cell, B cell, T cell, NK (Natural Killer) cell and Monocytes were used for analysis. Transcript sets which belonged to Myeloid, Lymphoid or Multiple immune cell groups were not used to prevent ambiguity in interpreting the results. Some transcripts did not map to known gene symbols and these were identified using *alias2symbol* function in the *Linear Models for Microarray Analysis (LIMMA)* package in R/Bioconductor. These transcripts could not be used for further analysis. Averaged expression values of cell lineage-specific transcripts were used to estimate changes in composition of their respective cell types in urine. The number of upregulated and downregulated probes from each cell type were shown to demonstrate the overall direction of regulation of that cell type. ComplexHeatmap package was used to generate heatmaps using scaled expression values. As a proof of concept that there are immune cells in urine of septic individuals, urine samples from 10 random septic patients were analyzed via flow cytometry using an LSR II flow cytometer (Becton Dickinson, Franklin, NJ). Approximately 50-200 mL of urine was collected and processed within 30 minutes of sample collection. The samples were stained with CD3-AF488 (#557694, BD) for T

cells, CD4-AF700 (#566318, BD) for T-helper cells, CD8-BV650 (#565289, BD) for T cells, CD14-PE (#561707, BD) for monocytes and macrophages, CD19-APC (#561742, BD) for B cells, and Sytox Blue (#S34857, Invitrogen) for dead cells.

Identification and characterization of discriminating set of genes in sepsis

We applied empirical Bayes method in the Linear Models for Microarray Analysis (LIMMA) to identify differentially expressed probes between sepsis and vascular patients. The significance threshold was adjusted for multiple testing using the Benjamini-Hochberg false discovery rate (FDR). Probes with FDR of ≤ 0.01 and an absolute fold change ≥ 2 were considered differentially expressed. Gene expression patterns were elucidated using Euclidean distance heatmaps with ComplexHeatmap. Ingenuity pathway analysis (IPA) software was used to identify significantly enriched biological functions, pathways (right-tailed Fisher's exact test, P value < .05), molecular networks and regulatory molecules concerning the differentially expressed genes (Figure 1B). The differentially expressed probes which map to known genes were subjected to feature selection using multiple machine learning techniques to find subsets that are highly unique for sepsis patients. Two different types of machine learning feature selection methods were used to sieve out important features – i) minimal-optimal methods – these methods select features that minimize the post-selection error rate of a given classifier, and ii) all-relevant method which finds all strongly and weakly related features to a given condition. The features selected in minimal-optimal method is always dependent on the classifier used, hence multiple classifiers namely Random Forest, Logistic Regression with L1-norm (lasso) penalty and Support Vector Machines were used. Boruta is used as an all-relevant method.

Four different feature selection techniques namely random forest, recursive feature elimination using support vector classifier, logistic regression with lasso and Boruta (Figure 1B) were used in parallel to generate four different lists of selected features. Random Forest provides an importance score for every feature. Therefore, features were ranked according to this importance and this list was truncated where the cumulative importance did not increase by 0.1% upon inclusion of the next feature. This gave 200 features from Random Forest. Boruta is a wrapper around ensemble methods (we used random forest), where in every iteration, it drops all features which are less "important" (according to the feature ranking provided by the internal ensemble) than self-generated randomly shuffled features called 'phantom features'**.** This algorithm gave us 49 features at the end of 100 iterations of Boruta. Recursive feature elimination reduces the feature set by recursively dropping low-importance features that do not contribute to the model performance. This algorithm applied to coefficients obtained from support vector machine gave 200 features. Lasso employed an L1-regularization to the coefficients obtained from logistic regression to select 266 features. Due to non-usage of a performance metric other than relative feature importances, n-fold cross validation is not suitable for training Boruta. Each of the other methods were parametrized inside a 5-fold cross validation design. The AUCs for every fold of these cross validation experiments are tabulated in Supplementary Table S2. Finally, we employed a simple voting strategy and retained features that appear in at least two of the four feature lists. PubMed was searched using text mining to identify articles that match this final subset of genes to the keyword "sepsis" using the *rentrez* package in R. The resulting articles were reviewed by authors (AB, SB, KF, HH) to provide an

overview of biologic functions of identified genes in sepsis. Ingenuity pathway analysis on these 239 genes show that many of them are involved in sepsis-specific canonical pathways (Supplementary Figure S9). The total number of pathways that each gene participated in was manually tabulated. This revealed the widespread involvement of transcription factor p65 (RELA) (16 of 17 pathways), interleukin 1 beta (IL1B) (8 of 17), protein kinase C delta (PRKCD) (8 of 17), prostaglandin-endoperoxide synthase 2 (*PTGS2)* (8 of 17), transforming growth factor beta 1 (TGFB1) (8 of 17), interleukin 8 (CXCL8) (6 of 17), toll-like receptor 2 (TLR2) (6 of 17), glycogen synthase kinase 3 beta (GSK3B) (5 of 17) and G-protein subunit alpha 13 (GNA13) (5 of 17). Upon searching co-expression networks for these genes, we discovered the presence of two tightly-knit regulatory co-expression networks containing these genes (Supplementary Figure S10). Descriptive analysis was performed in SAS (v.9.4, Cary, NC). An automated analytical framework for the entire process in Figure 1B was implemented using Bioconductor (version 3.7) in R (version 3.4.2) and scikit-learn (version 0.19.2) in Python (version 2.7) and is available on Github.