APMIS JOURNAL OF PATHOLOGY, MICROBIOLOGY AND IMMUNOLOGY

APMIS 130: 524–534 *© 2022 The Authors. APMIS published by John Wiley & Sons Ltd on behalf of Scandinavian Societies for Medical Microbiology and Pathology.*

DOI 10.1111/apm.13234

Transcriptomic fingerprint of bacterial infection in lower extremity ulcers

BLAINE G. FRITZ,[1] JULIUS B. KIRKEGAARD,[2] CLAUS HENRIK NIELSEN,[3] KLAUS KIRKETERP-MØLLER,[4] MATTHEW MALONE[5,6] and THOMAS BJARNSHOLT[1,7]

[1]Department of Immunology and Microbiology, Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark; [2]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark; [3]Department of Rheumatology and Spine Diseases, Institute for Inflammation Research, Rigshospitalet, Copenhagen, Denmark; [4]Center for Wound Healing, Bispebjerg Hospital, Copenhagen, Denmark; [5]South West Sydney Limb Preservation and Wound Research, Liverpool Hospital, Sydney, Australia; [6]Infectious Diseases and Microbiology, School of Medicine, Western Sydney University, Sydney, Australia; and [7]Department of Clinical Microbiology, Rigshospitalet, Copenhagen, Denmark

Fritz BG, Kirkegaard JB, Nielsen CH, Kirketerp-Møller K, Malone M, Bjarnsholt T. Transcriptomic fingerprint of bacterial infection in lower extremity ulcers. APMIS. 2022; 130: 524–534.

Clinicians and researchers utilize subjective, clinical classification systems to stratify lower extremity ulcer infections for treatment and research. The purpose of this study was to examine whether these clinical classifications are reflected in the ulcer's transcriptome. RNA sequencing (RNA-seq) was performed on biopsies from clinically infected lower extremity ulcers (n = 44). Resulting sequences were aligned to the host reference genome to create a transcriptome profile. Differential gene expression analysis and gene ontology (GO) enrichment analysis were performed between ulcer severities as well as between sample groups identified by k-means clustering. Lastly, a support vector classifier was trained to estimate clinical infection score or k-means cluster based on a subset of genes. Clinical infection severity did not explain the major sources of variability among the samples and samples with the same clinical classification demonstrated high inter-sample variability. High proportions of bacterial RNA were identified in some samples, which resulted in a strong effect on transcription and increased expression of genes associated with immune response and inflammation. K-means clustering identified two clusters of samples, one of which contained all of the samples with high levels of bacterial RNA. A support vector classifier identified a fingerprint of 20 genes, including immune-associated genes such as *CXCL8*, *GADD45B*, and *HILPDA*, which accurately identified samples with signs of infection *via* cross-validation. This study identified a unique, host-transcriptome signature in the presence of infecting bacteria, often incongruent with clinical infection-severity classifications. This suggests that stratification of infection status based on a transcriptomic fingerprint may be useful as an objective classification method to classify infection severity, as well as a tool for studying host–pathogen interactions.

Key words: Diabetic foot; ulcer; chronic wounds; transcriptomics; RNA sequencing; biofilm; infection; machine learning.

Thomas Bjarnsholt, Department of Immunology and Microbiology, Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark. e-mail: tbjarnsholt@sund.ku.dk

Lower extremity ulcers present both humanistic and economic burdens to society. A UK study identified a prevalence of chronic lower extremity wounds in 6% of the population with management costs amounting to £328.8 million[1]. Diabetic-related foot ulcers (DFUs) are one type of lower extremity ulcer with a high burden. DFUs arise due to several factors, including compromised arterial circulation, peripheral neuropathy, and repeated injury. Many individuals also demonstrate dysregulation of the immune response and poor glycemic control. The combination of these complex comorbidities likely increases the potential of DFUs to develop infection and osteomyelitis [2,3]. More than 50% of DFUs will become infected [4], which increases the likelihood of poor clinical outcomes, risk of hospital admission, and lower extremity amputation [5,6]. DFUs may take weeks, months, or years to heal and 65% of patients develop a new ulcer within 5 years [3,7]. These cycles of re-

Received 20 January 2022. Accepted 11 May 2022

ulceration and/or infection contribute to reduced quality of life, increased morbidity, and mortality. Similarly, persons with leg ulcers can experience prolonged pain, social challenges, and decreased psychological well-being [6,8–10].

Several classification systems exist to stratify DFUs and grade infection severity. These include the Infectious Diseases Society of America/International Working Group on the Diabetic Foot (IDSA/IWGDF) guidelines [11], Wagner classification system [12], University of Texas system (UT) [13], site/ischemia/neuropathy/bacterial infection/depth (SINBAD) system [14], diabetic ulcer severity score (DUSS) [15], and perfusion/extent/depth/infection/sensation (PEDIS) systems [16]. These classifications combine clinical data and observations to stratify wounds, classify infections, guide treatment, and predict outcomes [17]. A well-known limitation of classification systems is that clinical observations are open to observer interpretation and may vary greatly, depending on the observer and/or the patient's physiological state [18]. Underlying comorbidities, such as diabetes mellitus or peripheral arterial disease, further compound the clinical picture and have been implicated as causes of immune dysfunction and/or reduced infection symptoms [19,20].

Recent advances in RNA sequencing technologies have allowed high-resolution examination of gene expression in ulcer tissue and infecting bacteria at the bulk and single-cell level. For example, stratification by infection severity demonstrated increased microbial diversity and a unique host response in severe DFIs [21,22]. Wound healing has also been studied as an important outcome. Dysregulation of major transcriptional networks, such as those associated with migration of neutrophils and macrophages, and inflammation have been identified in non-healing wounds [23,24]. Single-cell RNA sequencing has associated differential macrophage polarization and unique subpopulations of fibroblasts with improved wound healing [23,25,26]. Furthermore, transcriptomic studies have identified unique transcriptomes in bacteria during chronic infection and bacteria-specific responses [27,28]. Studies examining transcription as a factor of infection severity often utilize clinical scores to classify infected samples, though there is evidence that clinical infection scores do not correspond with bacterial load [18]. Studies of the human transcriptome also generally perform host mRNA enrichment, making the simultaneous observation of bacterial and host RNA transcripts impossible.

In this study, we analyzed RNA sequencing data from clinically infected DFU biopsies to examine the effect of bacterial load on host transcription and correlation to clinical parameters. We assessed both the clinical infection severity score (IDSA) as well the quantity of bacterial mRNA in the samples to explain changes in the host transcriptome. We then identified differentially expressed genes and pathways which are characteristic for DFUs with high proportions of bacterial RNA. From these genes, we develop a transcriptomic fingerprint of ~20 genes, which could be utilized to identify infected DFUs and guide treatment.

METHODS

Sample collection and external data sources

This study included tissue biopsies (n = 30) collected from the Liverpool Hospital (LHS) High-Risk Foot Service, Liverpool Hospital, Liverpool, Australia. External RNA-sequencing data (n = 16) were obtained from Heravi *et al.* [22]. This external data included clinically infected DFUs graded by infection severity score (PEDIS/IDSA) and processed with similar methodologies for RNA sequencing as the LHS data. These external data were also included to provide a source of inter-laboratory variation. LHS samples were collected from patients presenting with an infected foot ulcer, where a punch biopsy was taken post-debridement from the edge of the ulcer. Detailed tissue collection methods are described in Supplementary Information (SI). All tissue samples were immediately placed into RNAlater (ThermoFisher Scientific, Vinius, Lithuania), incubated at 4°C for 24 h, and then frozen at −80°C until RNA extraction. The clinical metadata categories of interest were ulcer duration (0 = <2 weeks, 1 = 2–4 weeks, and 2 = >4 weeks), PEDIS/IDSA infection score (2 = mild, 3 = moderate, and 4 = severe) [11].

RNA extraction, library preparation, and sequencing

Frozen samples in RNAlater were thawed on ice. The tissue was removed and placed inside an empty, sterile Petri plate. A scalpel was used to cut a piece of tissue with a volume of approximately 0.5 cm^3. RNA extraction was then performed chloroform/phenol phase separation, as previously described [27,29] with some modifications. The tissue was placed in a 2 mL microtube (Sarstedt, Nuembrecht, Germany) filled one-third with 2 mm and 0.1 mm diameter zirconia beads (Biospec, OK, USA). One milliliter of RNABee (Amsbio Europe, Alkmaar, the Netherlands) containing 10 μL/mL β-mercaptoethanol was added to the tubes. The tubes were placed in a MagNA Lyzer instrument (Roche, Zug, Switzerland) and homogenized at maximum power for 3 × 30 s. Tubes were placed on ice for 1 min after each homogenization step. 200 μL of chloroform was added, and the tubes were shaken vigorously for 30 s, incubated for 5 min on ice, and then centrifuged at 13 000 *g* for 30 min at 4 °C. The upper aqueous phase was collected and placed in a new, 1.5 μL DNA Lo-Bind® centrifuge tubes (Eppendorf, Hamburg, Germany). Ice-cold ethanol (0.5 mL) and 2 μL of 5 mg/mL linear acrylamide was added (ThermoFisher Scientific). The tubes were inverted several times and stored at −80 °C overnight. The samples were then thawed on ice and centrifuged (13 000 *g*, 30 min, +4 °C). The supernatant was discarded, and the pellet was washed twice with fresh, ice-cold 75% ethanol. The pellet was then resuspended in 20–65 μL nuclease-free water. RNA concentration was quantified with a NanoDrop™ spectrophotometer (ThermoFisher Scientific). Contaminating DNA was removed by combining ~2.5 μg RNA with 3 μL RQ-1 RNAse Free DNase (Promega, Madison, WI, USA), 3 μL DNAse buffer solution, 1 μL RiboGuard™ RNAse inhibitor (Lucigen, Middleton, WI, USA), and nuclease-free water to a final volume of 30 μL. The DNAse-treated RNA was then repurified with the RNABee protocol described above and then stored at −80°C. Ribosomal RNA depletion was performed with the 10:1 human:pan-prokaryote riboPOOLs (protocol version: 1.4.2; siTOOLS Biotech, Planegg, Germany). The NEBNext® Ultra II RNA library preparation kit (cat: E7775S; New England Biolabs, Ipswich, MA, USA) was used to prepare cDNA libraries. Concentrations and quality of the final libraries were assessed with a Qubit 4 fluorometer (ThermoFisher Scientific) and Agilent Bioanalyzer (Agilent, Santa Clara, CA, USA). Samples were sequenced on a Novaseq6000 sequencing instrument (Illumina, San Diego, CA, USA) with an S2 flow cell for 200 cycles to generate 100 bp, paired-end reads.

QC processing, alignment, quantification, and bias control of sequence data

All raw sequence data, including external data, were processed at the same time and with the same pipeline (DOI: 10.5281/zenodo.6586732). Adapter and quality trimming was performed with cutadapt 2.4 [30]. Reads less than 20 nucleotides after trimming were discarded. For analysis of host reads, in-silico rRNA depletion was performed with SortMeRNA v 2.1b against the SILVA rRNA databases to remove eukaryotic, bacterial, or archaeal ribosomal sequences. The reads were then aligned to the GRCh38 human genome assembly (GCA_000001405.15, RefSeq, full analysis set), including all alternative haplotypes and unlocalized scaffolds, with bwa-mem using default settings (v. 0.7.16a). Aligned reads mapping to exon features in the NCBI RefSeq annotation were quantified with featureCounts [31]. Any samples containing <1 million reads were discarded from the analysis. Any exonic features which represented other transcript types, such as ncRNA and tRNA, were also removed from the analysis. To normalize for batch differences between data sets, differential gene expression was performed between LHS data and the external data from Heravi *et al.* [22]. Any genes identified as differentially expressed between source were removed from the data set. These data were then used for further host analysis. For analysis and classification of bacterial reads, kraken2 [32] was applied to the raw sequence data. The number and percentage of bacterial and human reads were calculated as the number or percentage of reads covered by the clade rooted at the kingdom Bacteria or the species *Homo sapiens*, respectively. Percent relative activity was calculated as the number of reads classified to the clade rooted at a given bacterial species divided by the sum of all reads rooted at a species-level clade for bacteria.

Principal component analysis, k-means clustering, and differential gene expression

The count data were normalized using DESeq2's variance stabilizing transformation (vst, blind = TRUE, nsub = 1000) and analyzed with the prcomp function in R with no additional scaling. The first two principal components were plotted (Fig. 1). Component loading analysis and principal component correlations were performed with the R package, PCAtools. Spearman correlation with benjamini Hochberg correction for multiple comparisons was used for the correlation analysis of metadata with principal component positioning. The optimal number of clusters for k-means was selected by the fviz_nbclust function the factoextra package (v. 1.0.7) using the "silhouette" method. Differential gene expression analysis was performed with DEseq2 (v. 1.28.1) using default settings. First, the DESeq2 model was fitted to the categorical IDSA/PEDIS score, categorical ulcer duration and a binary variable for if the sample contained >10 percent bacterial reads (formula: ~ IDSA Score + Ulcer Duration + > 10% bacteria). To identify differentially expressed genes due to k-means cluster, the only variable in the formula was k-means cluster (C1/C2). Genes with an adjusted p-value of <0.05 (Wald test) and estimated ¦log2 fold-change¦ > 2 were considered significantly differentially expressed and used for further analysis.

To test whether genes differentially expressed between clusters represented an enrichment of known biological pathways, a statistical overrepresentation test was performed using PANTHER [33] and the "GO biological process complete" data set (GO Ontology database DOI: 10.5281/zenodo.4081749, Released 2020-10-09). Input to PANTHER was a list of significantly differentially expressed genes. The analysis was performed twice, using either genes with positive (*i.e.*, higher expression in Cluster 1) or negative (increased expression in Cluster 2) log2 fold changes, respectively. The reference list input to PANTHER was the list of all gene names inputted into the DESeq2 analysis. The analysis was performed with Fisher's exact test and FDR correction for multiple testing (FDR < 0.05).

Feature selection and testing

To identify gene features that could be used to identify specific levels of clinical metadata or clusters of samples, a support vector machine (SVM) with a linear kernel was fitted to the normalized data for PEDIS/IDSA score or k-means cluster. The most important features in the model were then selected based on the SVM coefficients to select features for classification. To evaluate the optimal number of features to include in the model, this process was repeated for up to 100 features. For each number of features (1–100), the given number of features was selected by the SVC and used for a stratified, sixfold cross-validation. A fingerprint based on twenty gene features was selected by the authors as the optimal number to avoid overfitting and account for additional variability when using external data. All analyses were performed with the python library scikit-learn (v. 1.0) [34].

RESULTS

RNA sequencing was performed on 30 DFU biopsies obtained from Liverpool Hospital (LHS), Liverpool Australia (samples annotated as P500–

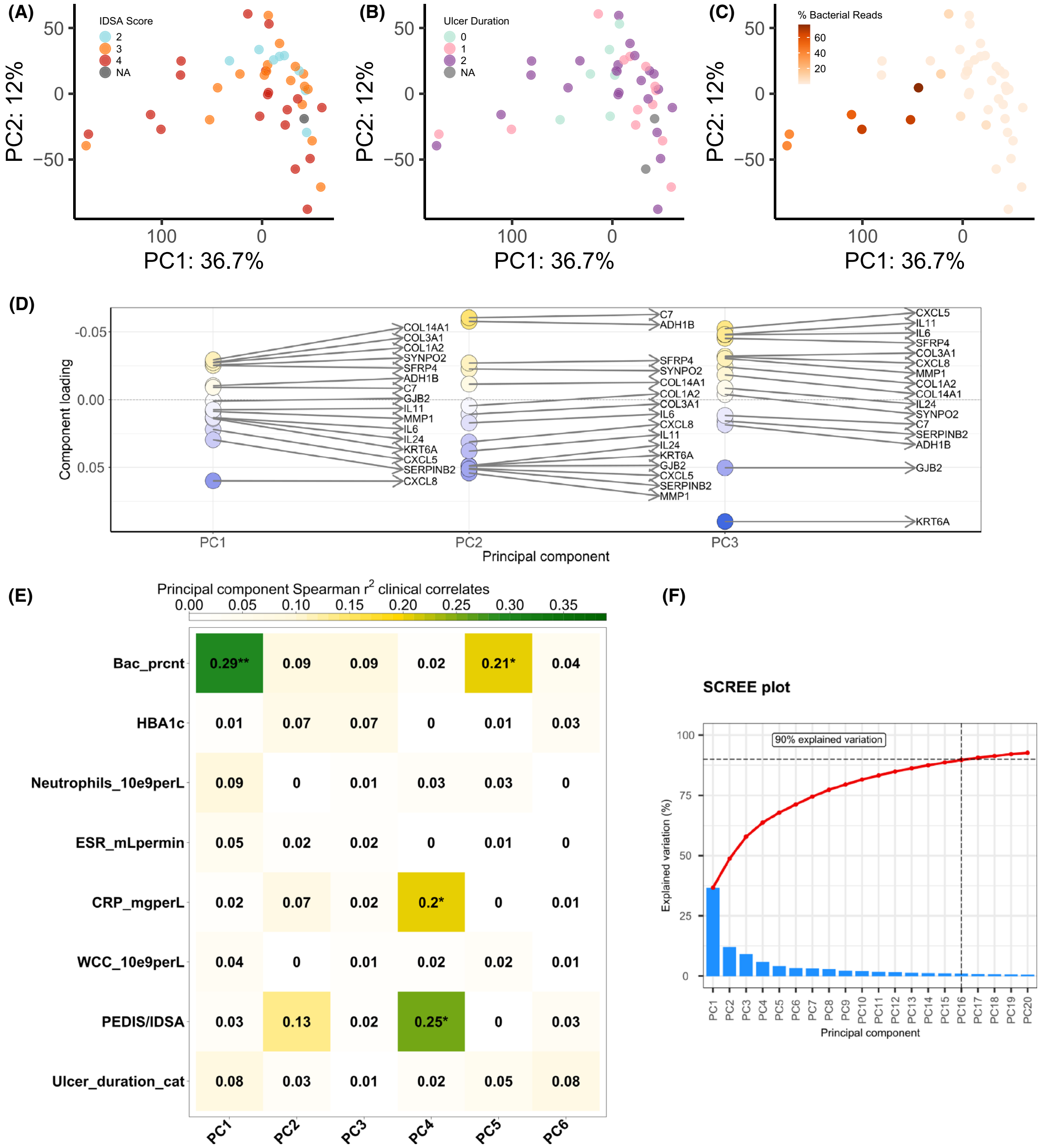


Fig. 1. (A–C) Characterization of host gene expression ulcer transcriptomes (n = 12,378 genes) by principal component analysis. Points are colored by: (A) IDSA/PEDIS infection severity score [2: Mild, 3: Moderate, and 4: Severe], (B) ulcer duration [0: Less than 2 weeks, 1: 2 to 6 weeks, and 2: Greater than 6 weeks], and (C) percentage of all RNA-seq reads classified to bacteria. (D) Component loadings for the top 5% of positive and negatively weighted genes for PC1 to PC3. Points are shaded by component loading value. (E) Spearman correlation coefficients of metadata variables with positioning of a samples along PC1 to PC6. Significance tests were performed with benjamini–hochberg p-value correction for multiple comparisons [**: $p < 0.01$, *: $p < 0.05$]. (F) Scree plot demonstrating the percent of explained variance for PC1 to PC20. The red line represents the cumulative percentage of explained variance across these PCs.

P529). Raw RNA sequence data (n = 16) from [22] were also included in the data set (samples prefixed with HH*). The combined data set yielded an average of 153 ± 23.9 M reads per sample passing quality filters. LHS samples showed significantly higher rRNA contamination than the HH data with mean percentages of rRNA contamination of 57.4 ± 17 and 5.0 ± 1, respectively. The non-rRNA reads were then aligned to the human reference genome, and all reads mapping to exonic gene features were counted. One sample with less than 1 M reads was excluded (P525). Additionally, sample HH5 was excluded, as its gene expression profile contributed a disproportionate amount of variability, relative to any of the other samples included in the study (Fig. S5). To control for variability due to source, differential gene expression analysis was performed to identify genes differentially expressed due to sample source. This identified 14 210 genes identified as differentially expressed due to source, which were then removed from the analysis. The remaining 32 841 genes were then used for further analysis (unless otherwise noted). The final data set contained a mean and median of 26.7 ± 21 M and 20.9 M reads per sample, respectively.

Presence of bacteria is associated with shift in host transcriptome

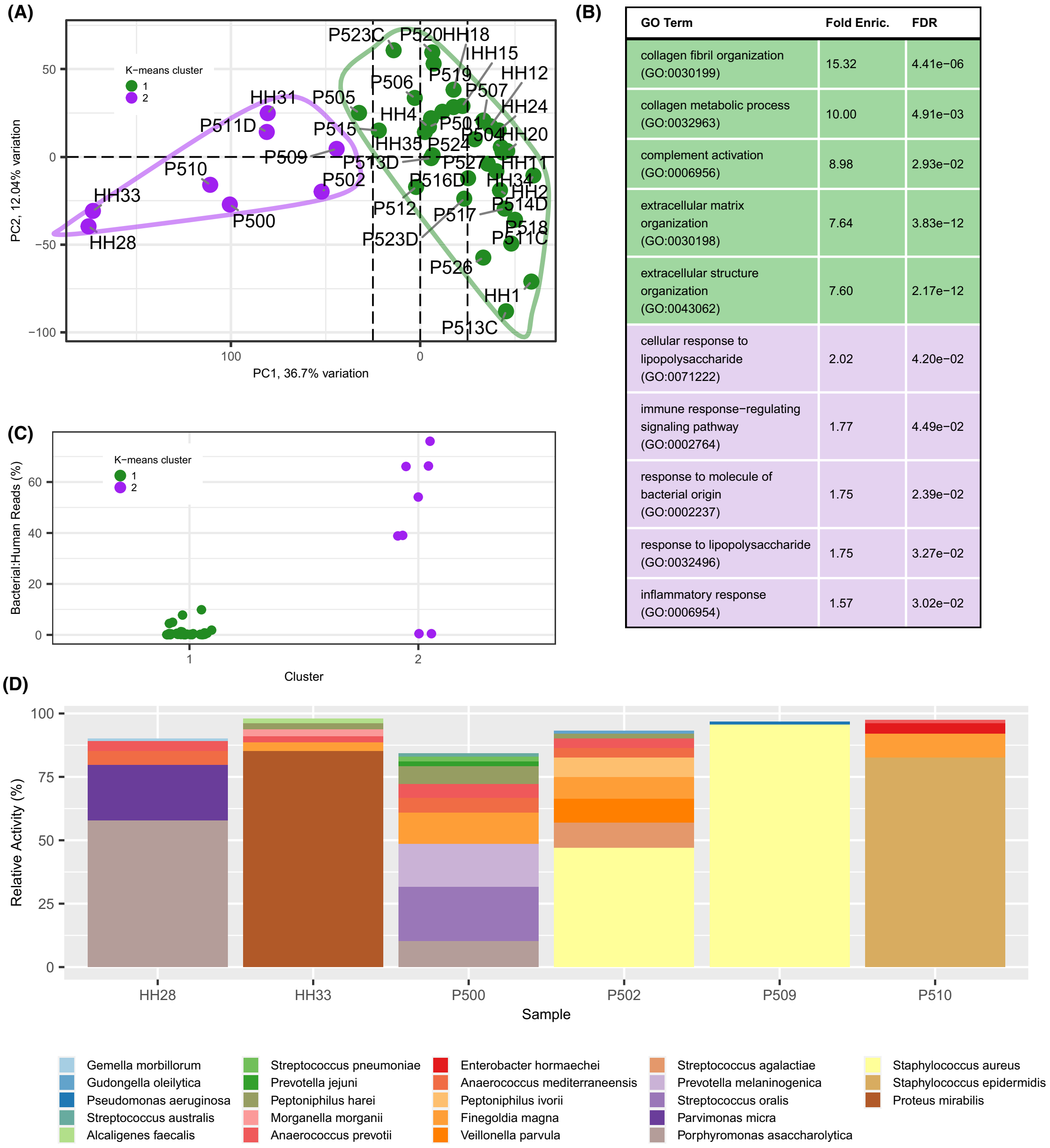
To evaluate the effect of bacteria on the host transcriptome, the vst-transformed host expression data (n = 44) were summarized by principal component analysis (PCA). The results of the PCA analysis are plotted in Fig. 1. The first two principal components summarized 49.1% of the overall variability. We observed that the majority of samples clustered positively (0–50) along the first principal component (PC1), while a subset of samples was spread across PC1 in the negative direction (−175 to 0). The samples were evenly distributed along PC2. To determine which factors affected this positioning, we tested whether clinical metadata variables or percentage of bacterial reads in the samples correlated with any of the first 10 principal components. The proportion of bacteria/human reads was the only factor showing significant correlation along either PC1 or PC2 ($r^2 = 0.29$, $p < 0.05$, spearman correlation with benjamini–hochberg correction). Levels of C-reactive protein as well as infection classification score showed slightly significant correlation to PC4, but this represented only a small proportion of the overall variability in the data. Additionally, the effect of increased proportions of bacterial reads was more prominent than batch variability between sources (Fig. 1C, Fig. S1).

To further demonstrate that the positioning represented biological effects of bacteria rather than confounding factors, such as decreased sequencing depth of the host due to the high presence of bacteria, we examined which genes drive the variation across PC1 (Fig. 1D). Of the top 5% most weighted genes, the genes with the strongest loading in the negative direction included inflammatory cytokines (*CXCL8*, *CXCL5*, *IL6*, *IL11*), keratinocyte factors associated with bacterial infection (*KRT6A*), and a matrix metalloprotease induced under inflammation (*MMP1*). The genes with the strongest positive loading included collagens associated with extracellular matrix deposition (*COL14A1*, *COL3A1*, *COL1A2*), actin-binding (*SYNPO2)*, apoptosis (*SFRP4)*, alcohol metabolism (*ADH1B)*, complement (*C7)*, and cellular gap junctions (*GJB2)*. This suggested that shift in positioning across PC1 in the negative direction is due to increased expression of immune-related genes, likely in response to bacterial infection.

Enrichment of immune processes and inflammation in samples with high bacterial activity

To confirm that samples with increased bacterial activity represented a specific transcriptomic response and to generate groups for comparative analysis, we performed k-means clustering on the normalized expression data. This analysis identified two cluster of samples, C1 (n = 8) and C2 (n = 36) (Fig. 2A). The mean proportion of bacterial:human reads per sample was significantly higher in C2 (t = 4.0367, p-value = 0.005, Welch t-test; Fig. 2C). We then tested for genes that showed differential expression between these two clusters using DESeq2. This analysis identified 2665 genes differentially expressed (¦log2FoldChange¦ > 2, adjusted p-value <0.05) between these two groups. Of these, 238 and 2427 genes showed significantly increased expression in C1 and C2, respectively. In comparison, only 113 genes were significantly differentially expressed between different levels of IDSA/PEDIS, likely due to high variability among samples with the same clinical classification.

C1 demonstrated increased expression of the cytokines (*CXCL12*, *CXCL13*) and cadherins *(CD34*, *CD36*). C2 demonstrated increased expression of *S100A8/9* and *S100A12*. *ADAM8* showed significantly increased expression in C2. In C2, we identified significantly increased expression of several leukocyte-associated cytokines (*CXCL8*, *CXCL2*, *CXCL16*, *IL6*, etc.) and cadherins (*CD53*, *CD69*). We also observed increased expression of NFKB2, but also relatively higher expression NFKB inhibitors (*NFKBIA*, *NFKBIZ*). Though



GO Term	Fold Enric.	FDR
collagen fibril organization (GO:0030199)	15.32	4.41e−06
collagen metabolic process (GO:0032963)	10.00	4.91e−03
complement activation (GO:0006956)	8.98	2.93e−02
extracellular matrix organization (GO:0030198)	7.64	3.83e−12
extracellular structure organization (GO:0043062)	7.60	2.17e−12
cellular response to lipopolysaccharide (GO:0071222)	2.02	4.20e−02
immune response−regulating signaling pathway (GO:0002764)	1.77	4.49e−02
response to molecule of bacterial origin (GO:0002237)	1.75	2.39e−02
response to lipopolysaccharide (GO:0032496)	1.75	3.27e−02
inflammatory response (GO:0006954)	1.57	3.02e−02

Fig. 2. (A) Two clusters (C1 and C2) were identified by k-means analysis. Results are displayed projected over the principal component analysis plot of all genes (n = 12,378) with points colored by k-means cluster. Samples are labeled by ID, where samples prefixed with "P" and "HH" are from this study and Heravi *et al.* [22], respectively. (B) Gene ontology (GO) terms for GO biological processes showing a significant overrepresentation (Fisher's exact test) of genes identified as differentially expressed between C1 and C2. The top 5 overrepresented pathways for cluster 1 (green) and cluster 2 (violet) are shown. (C) Percentage of RNA-seq reads classified to bacteria relative to the total number of reads classified as either bacterial or host for C1 (green) and C2 (purple). The mean percentage of bacterial:human reads was significantly higher in C2 (t = 4.04, p = 0.004, Welch t-test). (D) Relative activity (percentage of RNA reads for a specific species relative to all bacterial reads, %) for bacterial species with relative activity >5%. Only samples with greater than 10% bacterial reads are shown.

differentially expressed in C2, expression of TNF-alpha was low.

To examine whether the genes differentially expressed between the C1 (low bacterial activity) and C2 (high bacterial activity) represented enrichment of known biological processes, we performed a Gene Ontology enrichment analysis using PANTHER (Fig. 2B). In the cluster with high bacterial activity (C2), we identified enrichment of 28 pathways (FDR < 0.05), including the specific subclasses, "cellular response to lipopolysaccharide" (GO: 0071222), "immune response-regulating signaling pathway" (GO:0002764), "inflammatory response" (GO: 0006954), and "innate immune response" (GO: 0045087). In C1, there were 15 significantly enriched pathways (FDR < 0.05), including "collagen fibril organization" (GO:0030199), "collagen metabolic process" (GO:0032963), "complement activation" (GO:0006956), and "cell-matrix adhesion" (GO:0007160).

Differential host response to *Staphylococcus aureus*

To investigate which bacterial species were present and active in the samples, KRAKEN was used to identify reads originating from bacteria and assign taxonomy (Fig. 2D). In four of the six samples with high bacterial signals, at least 50% of the reads classified to bacteria were identified to a single species (*S. aureus*, *Staphylococcus epidermidis*, *Proteus mirabilis*, or *Porphyromonas asaccharolytica*). We hypothesized that samples with increased signals of infection (*i.e.*, C2) would display decreased alpha diversity of active bacteria (*i.e.*, species with >5% of reads in a sample). There was, however, no significant difference in the number of active species between the clusters (t = –0.071, p = 0.94, Welch t-test) for species. We further investigated whether *S. aureus*, which was highly active in 2 of 8 samples in C2 (mean relative activity: P502–47.1%, P509–96.3%), was also present in other samples with lower signals of infection (*i.e.*, C1). Interestingly, *S. aureus* was also found with at least a 10% relative activity in 13 of 36 samples in C1 (mean relative activity: 37.4% ± 25). *S aureus* was only associated with increased immune response and inflammation in samples with a high proportion of bacterial:human reads (P502, P509) and high relative activity, supporting its dual role as both a pathogen and a commensal organism.

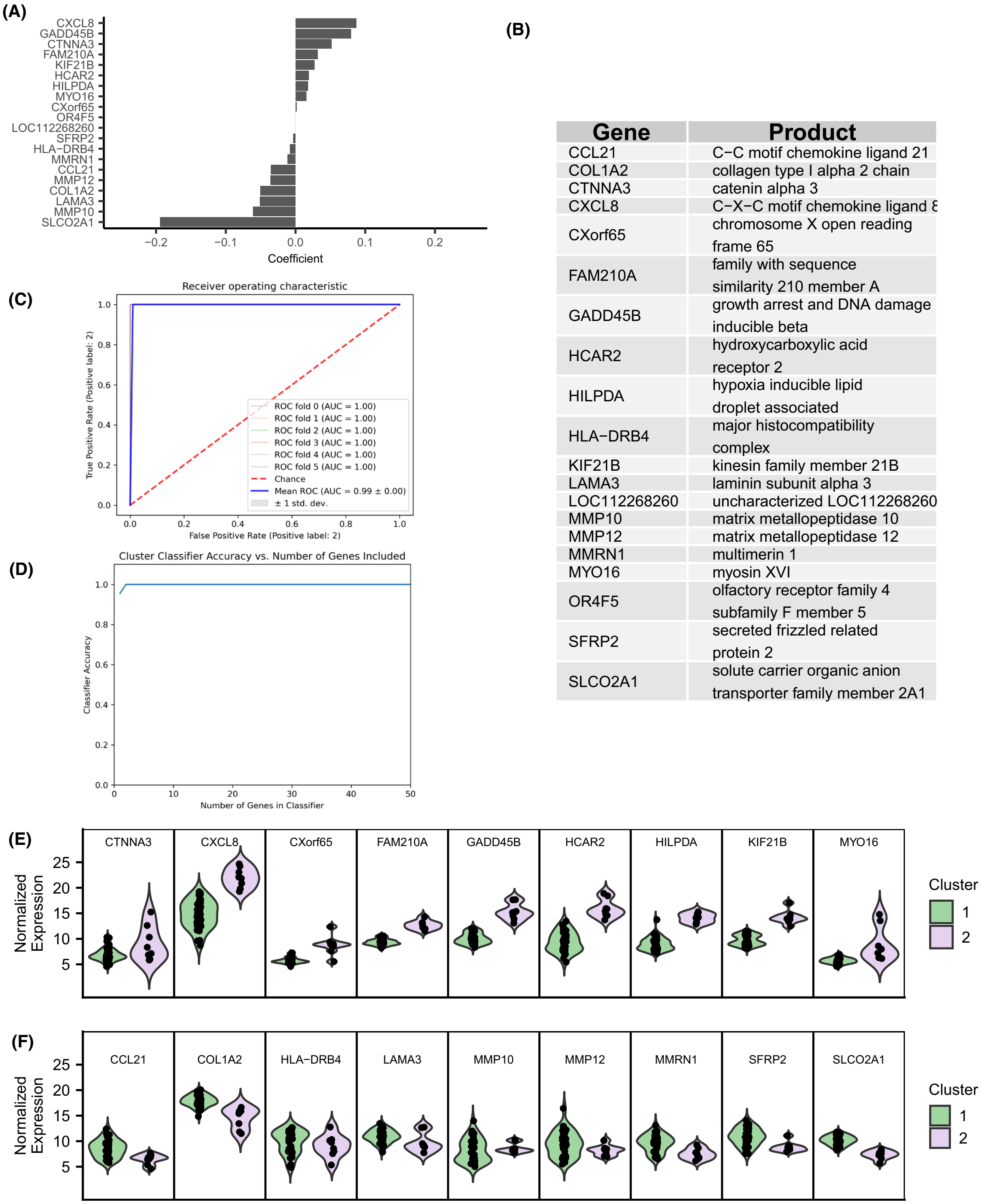
Definition and validation of transcriptomic fingerprint to classify ulcer status

To identify a small set of genes, which could be used to identify samples with a bacterial infection, a support vector classifier (SVC) was applied to the RNA-seq data to select a reduced set of gene features to define each cluster. Results of this analysis are displayed in Fig. 3. Twenty gene features were selected from the model as useful classifiers to differentiate between samples in C1 and C2 (Fig. 3A, B). We evaluated the accuracy of the classifier when trained with between 1 and 100 features and obtained high accuracy with less than 10 genes, but conservatively included 20 genes to increase the robustness of the model (Fig. 3D). Several of the identified genes were associated with immune cells and inflammation, including *CXCL8* (neutrophil recruitment [35]), *GADD45B* (stress-response[36]), *HILPDA* (macrophage infiltration [37]), and *KIF21B* (T-cell polarization [38]). The normalized expression of these genes was also clearly elevated in C2 relative to C1 (Fig. 3E). Genes that were negative classifiers for C2 (*i.e.*, demonstrated increased expression in C1) included metalloproteases *MMP10* and *MMP12*, the collagen matrix protein *COL1A2*, and the chemokine *CCL21*. The feature with the largest coefficient was *SLCO2A1*, which showed increased expression in C1.

We additionally performed the same analysis on the PEDIS/IDSA score to evaluate whether a similar model could predict clinical PEDIS/IDSA infection scores (Figs S2, S3). This model only achieved a maximum of 75–80% accuracy for predicting the PEDIS/IDSA score of the testing data with ~30 gene features (Fig. S4). This was also to be expected, as there did not appear to be a difference in gene expression for the features selected as good classifiers for PEDIS/IDSA score (Fig. S3).

DISCUSSION

This study examined clinically infected DFUs to investigate whether clinical infection classification or ulcer duration reflect ulcers' gene expression profiles, determined by RNA sequencing. We observed that most of the variability in the data was not described by infection severity classification or ulcer duration. Rather, that the proportion of bacterial reads was a driving force in transcriptomic variation. Unsupervised clustering identified two groups of samples in the data, where one group demonstrated significantly increased bacterial activity. These groups were inconsistent with an ulcer's infection classification and duration, despite the samples clustering in the same cluster. Samples with increased proportions of bacteria exhibited increased expression of genes associated with immune cells and inflammation, suggesting a direct response to the bacterial threat. We then identified



Gene	Product
CCL21	C−C motif chemokine ligand 21
COL1A2	collagen type I alpha 2 chain
CTNNA3	catenin alpha 3
CXCL8	C−X−C motif chemokine ligand 8
CXorf65	chromosome X open reading frame 65
FAM210A	family with sequence similarity 210 member A
GADD45B	growth arrest and DNA damage inducible beta
HCAR2	hydroxycarboxylic acid receptor 2
HILPDA	hypoxia inducible lipid droplet associated
HLA−DRB4	major histocompatibility complex
KIF21B	kinesin family member 21B
LAMA3	laminin subunit alpha 3
LOC112268260	uncharacterized LOC112268260
MMP10	matrix metallopeptidase 10
MMP12	matrix metallopeptidase 12
MMRN1	multimerin 1
MYO16	myosin XVI
OR4F5	olfactory receptor family 4 subfamily F member 5
SFRP2	secreted frizzled related protein 2
SLCO2A1	solute carrier organic anion transporter family member 2A1

Fig. 3. (A) Coefficient values for the top 20 genes extracted from the support vector classifier. (B) Gene symbols and products for identified features. (C) Receiver operating characteristic curve from cross-validation (sixfold, stratified) analysis. Given the 20 gene fingerprint, the classifier performed with 100% accuracy for classifying test samples in each fold. (D) Plot of classifier accuracy vs number of features included in the classifier and tested *via* stratified, sixfold crossvalidation for each number of features. (E) Normalized expression values for genes selected as positive predictive genes for cluster 2. (F) Normalized expression values for genes selected as positive predictive genes for cluster 1.

a fingerprint of 20 genes, including molecules of the immune system such as *CXCL8*, a chemoattractant for neutrophils [35], which accurately identified samples exhibiting a transcriptome consistent with high proportions of bacterial RNA.

Developing a robust system for the stratification of ulcers is essential for treatment and study. Several studies have performed comparisons between ulcers, based on stratification by clinical classification, but often do not account for the influence of bacterial activity on the local microenvironment [21,39–42]. For example, a previous study of infected ulcers identified increased expression of *GADD45B*, a DNA damage and stress response protein [36], (also identified in the present study) with IDSA/PEDIS scores of 4 [21]. Our findings suggest rather that this gene and others are expressed only in a subset of samples with an IDSA/PEDIS score of 4, specifically those with high bacterial activity. Similarly, a previous study found no differences between microbiological data including presence of gram-negative organisms or monomicrobial/polymicrobial infections among different grades of infection [42]. This is not to say that there is no difference in these parameters for severe infections, rather that high variability among samples with the same grade may lead to decreased statistical power to detect these differences.

Our results suggest an acute inflammatory response to bacteria in C2 relative to lower-level inflammation observed in C1. This is supported by the observation of CXCL8 and other molecules induced by an inflammatory environment in C2, such as BCL2 related protein A1 (*BCL2A1*), oncostatin M (*OSM*), prostaglandin G/H synthase (*PTSG2*), and S-100 calcium-binding protein A8/9 (*S100A8/9*). The increased expression of *CXCL8* and *ADAM8* also suggests active recruitment of neutrophils to the ulcer [43]. The presence of neutrophils in the samples is suggested by the hydroxycarboxylic acid receptor 2 (*HCAR2*) and oncostatin M (*OSM*) genes, which are expressed by neutrophils [44,45]. In addition, the increased expression of interleukin-6 (*IL6*) in C2 supports the presence of pro-inflammatory, type 1 macrophages. Contrarily, the presence of a reduced immune response and inflammatory environment was suggested in C1. For example, CXC13, a selective chemoattractant for B cells [46], showed increased expression C1 suggesting the recruitment of B cells. Previous research has suggested that the inhibition of immune response and recruitment of immune cells may lead to decreased wound healing, suggesting that C1 may represent a chronic-like state in comparison with C2, where we observed genes associated with acute inflammatory response and recruitment of neutrophils.

This study and the interpretation of the findings presents several limitations. First, distribution of bacteria is known to be heterogeneous within diabetic foot ulcers, but the heterogeneity of gene expression in an ulcer has not been studied. It is unclear whether the difference between clusters arises from global difference in wound gene expression or differences among sampling areas. Identification and quantification of bacteria in this study were performed with a k-mer based metagenomic classifier applied to RNA sequencing data. Thus, it is not clear whether increased proportions of bacterial reads are due to increased numbers of bacteria or increased bacterial activity. Bacteria with more genes and larger genomes may also be overrepresented. To classify samples, a support vector classifier was applied to k-means groups found in the host-transcriptome data. This classification may be affected by unseen confounding variables such as similar histocompatibility complexes between patients in the same group, underlying comorbidities, and uneven patient group distributions. Furthermore, though our model includes 20 genes, it is effective with very few genes. This may lead to overfitting and inaccuracy with respect to additional external data. Further research is required to overcome these limitations.

CONCLUSIONS

This study identified and characterized a unique, host-gene expression pattern resulting from host–pathogen interactions in lower extremity ulcer infection. High proportions of bacterial RNA in a sample resulted in a consistent shift in host gene expression toward increased immune response and inflammation. Patterns of host gene expression were often inconsistent with clinical infection severity classifications. Expression levels of a set of ~20 host genes could consistently identify samples with high proportions of bacterial RNA. Such a transcriptomic fingerprint may provide a useful tool for clinicians and researchers to classify infection state in lower extremity ulcers.

Work for this project was supported by grants from The Lundbeck Foundation to TB and BF. We thank the GeoGenetics Sequencing Core at University of Copenhagen for assistance in generating the RNA-seq data and Computerome (www.computerome.dk) for providing the computing infrastructure to analyze the data. We thank Elio Rossi for useful discussions and contributions to the code used in the analysis.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

1. Phillips CJ, Humphreys I, Fletcher J, Harding K, Chamberlain G, Macey S. Estimating the costs associated with the management of patients with chronic wounds using linked routine data. Int Wound J. 2016;13(6):1193–7.
2. Raffetto JD, Ligi D, Maniscalco R, Khalil RA, Mannello F. Why venous leg ulcers have difficulty healing: overview on pathophysiology, clinical consequences, and treatment. J Clin Med. 2020;10(1):29.
3. Armstrong DG, Boulton AJM, Bus SA. Diabetic foot ulcers and their recurrence. N Engl J Med. 2017;376 (24):2367–75.
4. Prompers L, Huijberts M, Apelqvist J, Jude E, Piaggesi A, Bakker K, et al. High prevalence of ischaemia, infection and serious comorbidity in patients with diabetic foot disease in Europe. Baseline results from the Eurodiale study. Diabetologia. 2007;50(1):18–25.
5. Skrepnek GH, Mills JL, Lavery LA, Armstrong DG. Health care service and outcomes among an estimated 6.7 million ambulatory care diabetic foot cases in the US. Diabetes Care. 2017;40(7):936–42.
6. Herber OR, Schnepp W, Rieger MA. A systematic review on the impact of leg ulceration on patients' quality of life. Health Qual Life Outcomes. 2007;5 (1):44.
7. Smith-Strøm H, Iversen MM, Igland J, Østbye T, Graue M, Skeie S, et al. Severity and duration of diabetic foot ulcer (DFU) before seeking care as predictors of healing time: a retrospective cohort study. PLoS One. 2017;12(5):e0177176.
8. Pedras S, Carvalho R, Pereira MG. Predictors of quality of life in patients with diabetic foot ulcer: the role of anxiety, depression, and functionality. J Health Psychol. 2018;23(11):1488–98.
9. Polikandrioti M, Vasilopoulos G, Koutelekos I, Panoutsopoulos G, Gerogianni G, Babatsikou F, et al. Quality of life in diabetic foot ulcer: associated factors and the impact of anxiety/depression and adherence to self-care. Int J Low Extrem Wounds. 2020;19(2):165–79.
10. Zhao H, McClure N, Johnson JA, Soprovich A, Al Sayah F, Eurich DT. A longitudinal study on the association between diabetic foot disease and health-related quality of life in adults with type 2 diabetes. Can J Diabetes. 2020;44(3):280–286 e1.
11. Lipsky BA, Berendt AR, Cornia PB, Pile JC, Peters EJ, Armstrong DG, et al. 2012 Infectious Diseases Society of America clinical practice guideline for the diagnosis and treatment of diabetic foot infections. Clin Infect Dis. 2012;54(12):e132–73.
12. Wagner FW Jr. The diabetic foot. Orthopedics. 1987;10(1):163–72.
13. Lavery LA, Armstrong DG, Harkless LB. Classification of diabetic foot wounds. J Foot Ankle Surg. 1996;35(6):528–31.
14. Treece KA, Macfarlane RM, Pound N, Game FL, Jeffcoate WJ. Validation of a system of foot ulcer classification in diabetes mellitus. Diabet Med. 2004;21(9):987–91.
15. Beckert S, Witte M, Wicke C, Königsrainer A, Coerper S. A new wound-based severity score for diabetic foot ulcers: a prospective analysis of 1000 patients. Diabetes Care. 2006;29(5):988–92.
16. Schaper NC. Diabetic foot ulcer classification system for research purposes: a progress report on criteria for including patients in research studies. Diabetes Metab Res Rev. 2004;20(S1):S90–5.
17. Monteiro-Soares M, Boyko EJ, Jeffcoate W, Mills JL, Russell D, Morbach S, et al. Diabetic foot ulcer classifications: a critical review. Diabetes-Metab Res Rev. 2020;36(1):e3272.
18. Gardner SE, Hillis SL, Frantz RA. Clinical signs of infection in diabetic foot ulcers with high microbial load. Biol Res Nurs. 2009;11(2):119–28.
19. Edmonds M. Infection in the Neuroischemic foot. Int J Low Extrem Wounds. 2005;4(3):145–53.
20. Lipsky BA, Senneville È, Abbas ZG, Aragón–Sánchez J, Diggle M, Embil JM, et al. Guidelines on the diagnosis and treatment of foot infection in persons with diabetes (IWGDF 2019 update). Diabetes Metab Res Rev. 2020;36(S1):e3280.
21. Radzieta M, Sadeghpour-Heravi F, Peters TJ, Hu H, Vickery K, Jeffries T, et al. A multiomics approach to identify host-microbe alterations associated with infection severity in diabetic foot infections: a pilot study. Npj Biofilms Microbiomes. 2021;7(1):29.
22. Heravi FS, Zakrzewski M, Vickery K, Malone M, Hu H. Metatranscriptomic analysis reveals active bacterial communities in diabetic foot infections. Front Microbiol. 2020;11(1):e1688.
23. Theocharidis G, Baltzis D, Roustit M, Tellechea A, Dangwal S, Khetani RS, et al. Integrated skin transcriptomics and serum multiplex assays reveal novel mechanisms of wound healing in diabetic foot ulcers. Diabetes. 2020;69(10):2157–69.
24. Sawaya AP, Stone RC, Brooks SR, Pastar I, Jozic I, Hasneen K, et al. Deregulated immune cell recruitment orchestrated by FOXM1 impairs human diabetic wound healing. Nat Commun. 2020;11(1):4678.
25. Theocharidis G, Thomas BE, Sarkar D, Mumme HL, Pilcher WJR, Dwivedi B, et al. Single cell transcriptomic landscape of diabetic foot ulcers. Nat Commun. 2022;13(1):181.
26. Januszyk M, Chen K, Henn D, Foster DS, Borrelli MR, Bonham CA, et al. Characterization of diabetic and non-diabetic foot ulcers using single-cell RNA-sequencing. Micromachines. 2020;11(9):815.
27. Cornforth DM, Dees JL, Ibberson CB, Huse HK, Mathiesen IH, Kirketerp–Møller K, et al. Pseudomonas aeruginosa transcriptome during human infection. Proc Natl Acad Sci. 2018;115(22):5125–34.
28. Ramirez HA, Pastar I, Jozic I, Stojadinovic O, Stone RC, Ojeh N, et al. Staphylococcus aureus triggers induction of miR-15B-5P to diminish DNA repair and deregulate inflammatory response in diabetic foot ulcers. J Invest Dermatol. 2018;138(5):1187–96.
29. Murray JL, Kwon T, Marcotte EM, Whiteley M. Intrinsic antimicrobial resistance determinants in the Superbug P. aeruginosa. MBio. 2015;6(6):e01603–15.
30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17 (1):3.

31. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
32. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20 (1):257.
33. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 2021;49(D1):D394–403.
34. Pedregosa F et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–30.
35. Murphy PM. Neutrophil receptors for interleukin-8 and related CXC chemokines. Semin Hematol. 1997;34(4):311–8.
36. Zumbrun SD, Hoffman B, Liebermann DA. Distinct mechanisms are utilized to induce stress sensor gadd45b by different stress stimuli. J Cell Biochem. 2009;108(5):1220–31.
37. Liu C et al. HILPDA is a prognostic biomarker and correlates with macrophage infiltration in pan-cancer. Front Oncol. 2021;11(1):597860.
38. Hooikaas PJ, Damstra HG, Gros OJ, van Riel WE, Martin M, Smits YT, et al. Kinesin-4 KIF21B limits microtubule growth to allow rapid centrosome polarization in T cells. elife. 2020;9(1):e62876.
39. Jnana A, Muthuraman V, Varghese VK, Chakrabarty S, Murali TS, Ramchandra L, et al. Microbial community distribution and Core microbiome in successive wound grades of individuals with diabetic foot ulcers. Appl Environ Microbiol. 2020;86(6):e02608–19.
40. Lavery LA, Crisologo PA, la Fontaine J, Bhavan K, Oz OK, Davis KE. Are we misdiagnosing diabetic foot osteomyelitis? Is the gold standard gold? J Foot Ankle Surg. 2019;58(4):713–6.
41. Noor S, Borse AG, Ozair M, Raghav A, Parwez I, Ahmad J, et al. Inflammatory markers as risk factors for infection with multidrug-resistant microbes in diabetic foot subjects. Foot. 2017;32(1):44–8.
42. Ismail AA, Meheissen MA, Elaaty TAA, Abd-Allatif NE, Kassab HS. Microbial profile, antimicrobial resistance, and molecular characterization of diabetic foot infections in a university hospital. Germs. 2021;11 (1):39–51.
43. Domínguez-Luis M, Lamana A, Vazquez J, García-Navas R, Mollinedo F, Sánchez-Madrid F, et al. The metalloprotease ADAM8 is associated with and regulates the function of the adhesion receptor PSGL-1 through ERM proteins. Eur J Immunol. 2011;41 (12):3436–42.
44. Grenier A, Dehoux M, Boutten A, Arce-Vicioso M, Durand G, Gougerot-Pocidalo MA, et al. Oncostatin M production and regulation by human polymorphonuclear neutrophils. Blood. 1999;93(4):1413–21.
45. Kostylina G, Simon D, Fey MF, Yousefi S, Simon HU. Neutrophil apoptosis mediated by nicotinic acid receptors (GPR109A). Cell Death Differ. 2008;15 (1):134–42.
46. Legler DF, Loetscher M, Roos RS, Clark-Lewis I, Baggiolini M, Moser B. B cell-attracting chemokine 1, a human CXC chemokine expressed in lymphoid tissues, selectively attracts B lymphocytes via BLR1/ CXCR5. J Exp Med. 1998;187(4):655–60.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1 PCA plot of vst-normalized count data prior to normalization for batch effects of (a) Source and (b) proportion of bacterial: human reads identified by RNA-seq.

Fig. S2 (a–c) Coefficients for the 20 genes selected as a “fingerprint” for IDSA/PEDIS infection severity score for classification of (a) IDSA 2- mild, (b) IDSA 3- moderate, and (c) IDSA 4 – severe infections. (d) Gene symbols and product names for the 20 genes identified as the IDSA/PEDIS fingerprint.

Fig. S3 VST-normalized gene expression values for the 20 genes selected to be effective classifiers of IDSA/PEDIS infection severity scores of 2(“mild,” red), 3(“moderate,” green), and 4 (“severe”, blue). Samples missing IDSA/PEDIS values are displayed in the far-right group for each gene. Despite being classified as effective classifiers based on their coefficient weights in the SVC model, the expression of the majority of these genes appears similar between groups.

Fig. S4 Classifier accuracy for identifying IDSA/ PEDIS scores of unknown samples with increasing number of gene features included in the model. Accuracy was taken as the mean accuracy between folds based on a six-fold stratified cross-validation for each number genes.

Fig. S5 PCA plot of vst-normalized count data for host gene expression prior to (left) and post (right) removal of the data point HH5. HH5 was identified as an extreme outlier, relative to any of the other samples in the study, by the authors. Due to the disproportionate amount of variance contributed by HH5 relative to similar samples from the same data set, it was excluded from the analysis.

Data S1 Supplementary data, including metadata, read count tables, gene ontology results, and adapter trimming/rRNA depletion statistics.