PNAS
www.pnas.org

Supplementary Information for

An *in vivo* atlas of host-pathogen transcriptomes during *Streptococcus pneumoniae* colonization and disease

Adonis D'Mello[1], Ashleigh N. Riegler[2], Eriel Martínez[2], Sarah M. Beno[2], Tiffany D. Ricketts[2], Ellen F. Foxman[3], Carlos J. Orihuela* & Hervé Tettelin[1]*

*These authors contributed equally

[1]Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, United States of America
[2]Department of Microbiology, The University of Alabama at Birmingham, Birmingham, AL, United States of America.
[3]Department of Laboratory Medicine, Yale University School of Medicine, New Haven, CT 06520, USA.

Corresponding author: Hervé Tettelin

Email: tettelin@som.umaryland.edu

ORCID: 0000-0002-0615-3257

Classification

Biological Sciences: Microbiology

Keywords

Dual species RNA-seq; *in vivo* transcriptomics; *Streptococcus pneumoniae*; mouse models of colonization and invasive disease; host-pathogen interactions

This PDF file includes:

Supplemental text
Supplemental table legends
Figures S1 to S9
SI References

SI Appendix 1

Bacterial strains

Streptococcus pneumoniae (*Spn*) strains used in this study are listed in the Dataset S8. Pneumococci were grown in Todd-Hewitt broth supplemented with 0.5% yeast extract (THY) or on tryptic soy blood agar plates at 37°C in 5% CO_2. Isogenic mutants were created by allelic exchange in strain TIGR4 (1). Mutagenic PCR constructs were generated by amplifying the upstream and downstream DNA fragments flanking the gene(s) of interest followed by the 5' and 3' integration of these fragments with the Sweet Janus Cassette using a HiFi assembly master mix (NEB). Transformation of TIGR4 with the mutagenic construct (100ng/ml) was induced using competence-stimulating peptide variant 2 (CSP-2) as described (2). Blood agar plates supplemented with kanamycin (300 mg/L) were used for selection. To create the double mutant, *Δula*/SP_1675, the Sweet Janus Cassette from *Δula* mutant was replaced by a construct containing a clean deletion of the *ula* operon. In this instance streptomycin (200 mg/L) was used for selection.

Mice

Animal experiments were carried out using male and female mice 6 to 12 weeks of age. Wildtype C57BL/6 were supplied from The Jackson Labs (Ellsworth, Maine) and housed at the University of Alabama at Birmingham Animal Facilities for at least 1 week prior to their use. All animal experiments were performed using protocols approved by the University of Alabama at Birmingham IACUC (protocols #21152 and #21231).

RNA isolation

On the day of RNA extractions, samples were thawed and spun down to discard the supernatant. Pellets were then incubated in 100 µL of lysis buffer (10 µL of mutanolysin, 20 µL of proteinase K, 30 µL of lysozyme, 40 µL of TE buffer) for 10 minutes. Pellets were mechanically disrupted in 600 µL RLT buffer (RNeasy Mini Kit, Qiagen) containing 1% β-mercaptoethanol, using a motorized pestle for 30 seconds. Samples were then spun through a Qiashredder followed by addition of 590 µL of 80% ethanol, and RNA was captured on the RNeasy Mini Kit columns with DNase treatment on column (Qiagen protocol), followed by additional DNase treatment in solution. Extracted RNA was quantitated using a Bioanalyzer. Ribosomal RNA was depleted using the Ribo-Zero rRNA Removal Kits for Gram-positive bacteria and/or for human/mouse/rat (Illumina).

RNA library construction and sequencing

300 bp-insert RNA-seq Illumina libraries were constructed using 0.01 – 2.0 µg of enriched mRNA that was fragmented then used for synthesis of strand-specific cDNA using the NEBnext Ultra Directional RNA Library Prep Kit (NEB-E7420L). The cDNA was purified between enzymatic reactions and the size selection of the library performed with AMPure SpriSelect Beads (Beckman Coulter Genomics). The titer and size of the libraries was assessed on the LabChip GX (Perkin Elmer) and with the Library Quantification Kit (Kapa Biosciences). RNA-seq was conducted on 150 nt pair-end runs of the Illumina HiSeq 4000 platform using two or three biological replicates (organs from different mice) for each condition. The numbers of reads generated for each sample are provided in Dataset S1.

Transcriptomics data analyses and differential gene expression calculations

FASTQ files were mapped to their respective genomes using HISAT2 (3) for *Mus musculus* and Bowtie (4) for pneumococcal genomes of D39, TIGR4, and 6A-10 (see reference genome information in the Dataset S8). Reads mapped to each respective genome were tabulated in Dataset S1. Gene expression counts for all samples were then estimated using HTseq (5). All sample counts for the mouse data were tabulated in Dataset S2. For bacterial counts, core genes without paralogs were determined for D39, TIGR4, and 6A-10 using PanOCT (6). Paralogs were removed because they could lead to potential biases in differential gene expression from a core perspective, due to reads mapping to multiple paralogs in a respective genome. Counts were then tabularized for analysis based on core genes, using TIGR4 locus tags as identifiers in Dataset S2. Counts tables were then imported into Rstudio for analyses and estimation of differentially expressed (DE) genes. Rarefaction curves in Figure 1 were generated from HTseq counts data (See Data & Code Availability). Principal Component Analyses (PCAs) and dendrograms in Figures 2 & 3 were generated using various R packages (see Dataset S8) based on normalized Variance Stabilized Transformation (VST) counts acquired using the DESeq2 R package (7). For mouse DE gene estimation, infected/colonized tissues were compared to their respective uninfected tissue control samples as the baseline using DEseq2 and filtered using an FDR cutoff of 0.05. For bacterial DE gene estimation, each infected tissue was compared to the nasopharyngeal colonization state as the baseline, while accounting for strain differences within core genes using DEseq2 and filtered using an FDR cutoff of 0.05. Common and unique DE genes for both species were determined using Upset plots and individual heatmaps of DE genes were generated based on Z-scores of VST counts (Dataset S2). Pneumococcal and mouse DE genes are provided in Datasets S3, S4 & S5. See Dataset S8 for package source/documentation & Data and Code Availability for additional parameters used in DE gene estimation or heatmap generation.

Supplemental Dataset Legends

Dataset S1. Overview of the total number of 150nt HiSeq 4000 Illumina reads generated for each of the samples that were part of this study. The breakdown of reads properly mapping to the host (mouse) genome and to the appropriate Spn genome (TIGR4, D39, or 6A10) is also indicated. Measurements of TIGR4 in vivo Colony Forming Unit (CFU) counts at the 4 diseased anatomical sites are also provided.

Dataset S2. Exhaustive list of HTSeq's raw read counts and DESeq2's variance stabilized (VST) read counts for the 1,682 core Spn genes (shared across strains TIGR4, D39, and 6A10), as well as all bacterial genes for each individual strain, and all mouse genes, for each of the samples that were part of this study. VST counts are transformed counts normalized for sequencing depth and variance across biological replicates.

Dataset S3. Complete list of TIGR4 and 6A-10 whole genome, differentially expressed values for all genes, across disease anatomical sites when compared to the nasopharynx. D39 had no nasopharyngeal data. Orthologous gene information across the 3 pneumococcal strains, calculated using PanOCT (Fouts et al. 2012), is provided in columns A, B & C.

Dataset S4. List of Spn genes that were differentially expressed, passing an FDR cutoff of <0.05, across disease anatomical sites when compared to the nasopharynx. The last tab indicates the list of 69 Spn DE genes that were common across the 4 comparisons (Figure 4B).

Dataset S5. List of host genes that were differentially expressed, passing an FDR cutoff of <0.05, at each anatomical site when comparing the state colonized/infected with Spn, to the uninfected state. The last tab indicates the list of 190 host DE genes that were common across all comparisons (excluding nasopharyngeal comparisons) (Figure 4D).

Dataset S6. KEGG pathways significantly enriched in Spn DE genes from Dataset S4. The last tab lists the top 100 most highly expressed Spn genes at each anatomical site (Figure 5B), and the list of 52 Spn genes shared across all sites (Figure 5C).

Dataset S7. Ingenuity Pathway Analysis (IPA) of host differentially expressed genes from Dataset S5 (using an additional log fold change cutoff of +/- 2). Results of significantly enriched IPA canonical pathways are presented for each anatomical site, together with lists of host DE genes (labeled as molecules) that contributed to pathway enrichment.

Dataset S8. Resource table of relevant reagents, software and algorithms used in this study.

Legends to supplemental figures are provided at the bottom of each figure.

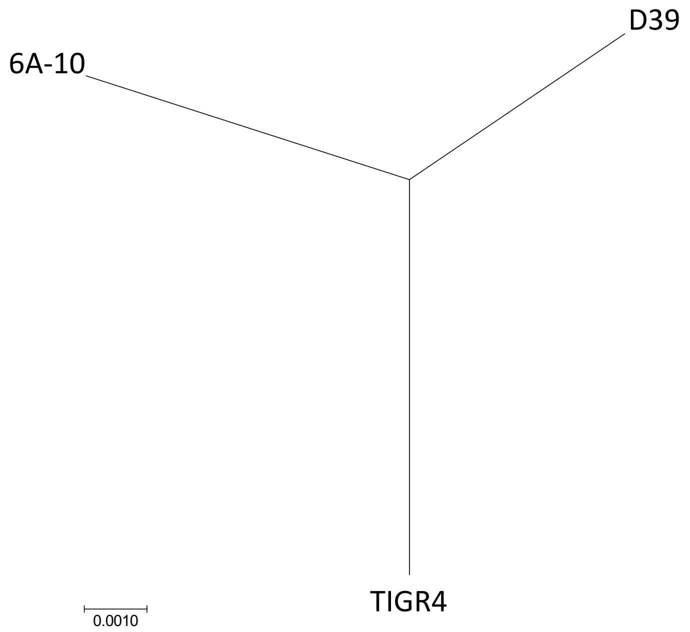

Fig. S1. Pneumococcal strains D39 and 6A10 are more genetically similar when compared to TIGR4. A phylogenetic tree based on complete genome sequences of all 3 pneumococcal strains shows a relatively smaller genetic distance between D39 and 6A10 strains. This potentially explains their more similar *in vivo* transcriptomic profiles relative to TIGR4.

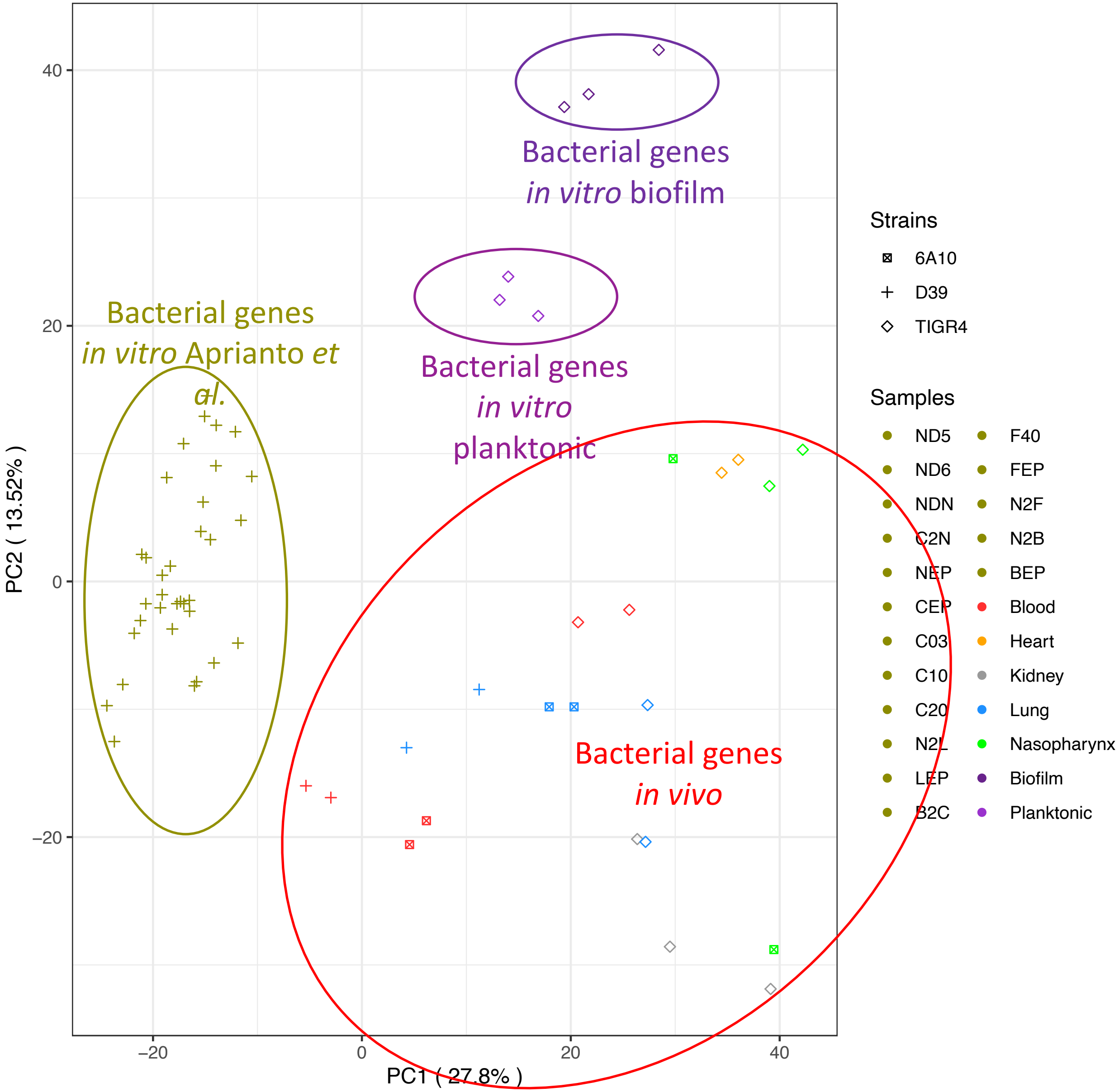

Fig. S2. Principal component analysis (PCA) of *in vivo* and *in vitro* representative datasets. A core *Spn* gene expression profile PCA of our dataset combined with in vitro data from Shenoy, Brissac et al. (2017) and a D39 study of *in vitro* conditions mimicking *in vivo* conditions (Aprianto, Slager et al. 2018). The PCA shows a clear separation between *in vitro* and *in vivo* datasets.

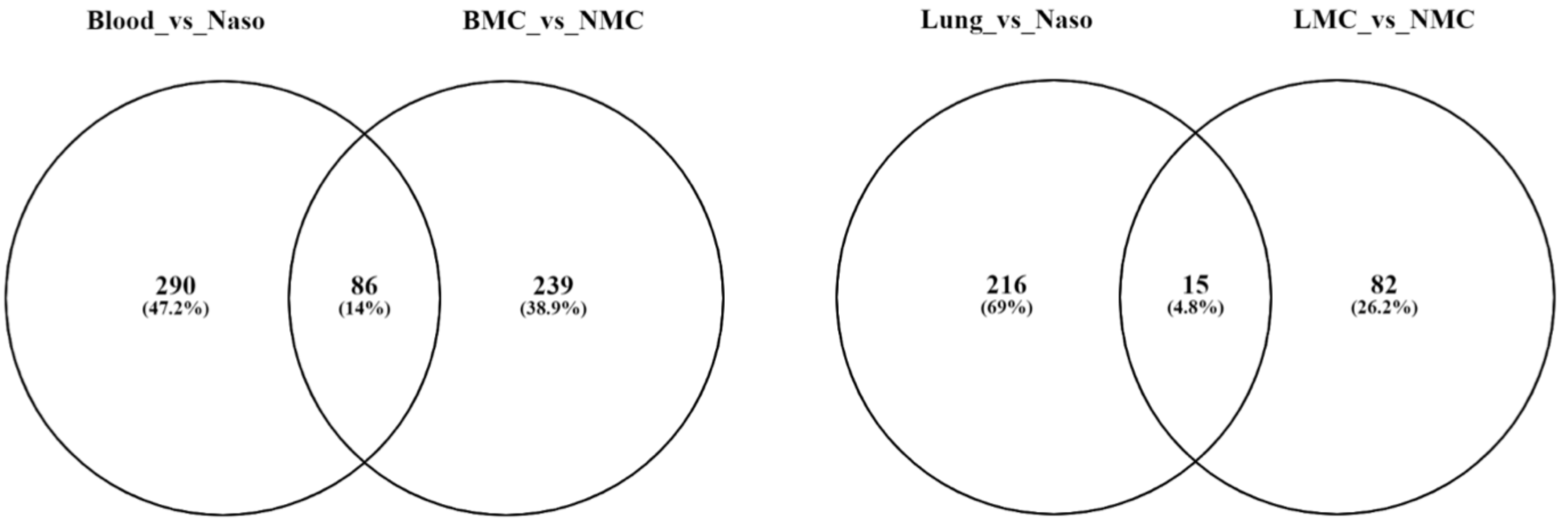

Fig. S3. Venn diagrams of pneumococcal differentially expressed (DE) genes detected *in vivo* and *in vitro* in mimicking conditions. Only 14% of *Spn* DE genes are shared between *in vivo* blood vs nasopharynx (our study) and *in vitro* blood mimicking conditions (BMC) vs nasopharynx mimicking conditions (NMC) (Aprianto, Slager et al. 2018). Similarly, only 4.8% of *Spn* DE genes are shared relative to lung mimicking conditions (LMC).

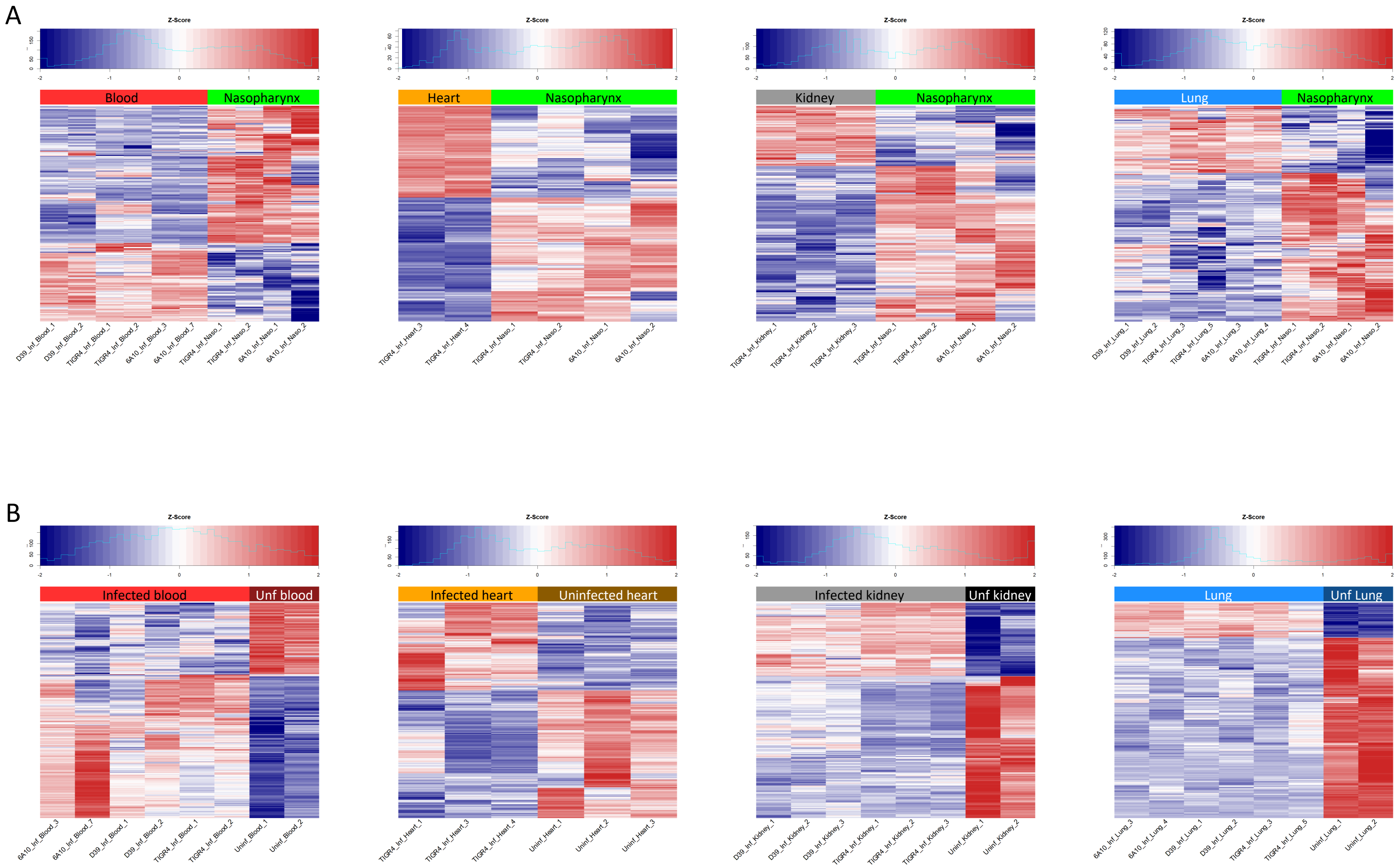

Fig. S4. Differentially expressed (DE) genes of the pneumococcus and the host. Z-scored heatmaps of expression levels of DE genes from host and pneumococcal Upset plots (Figure 4). **A.** *Spn* DE genes from Figure 4A (horizontal colored bars) in the blood, heart, kidneys, and lungs (left to right, respectively). **B.** Host DE genes unique to each organ site from Figure 4C (vertical colored bars) in the blood, heart, kidneys, and lungs (left to right, respectively).

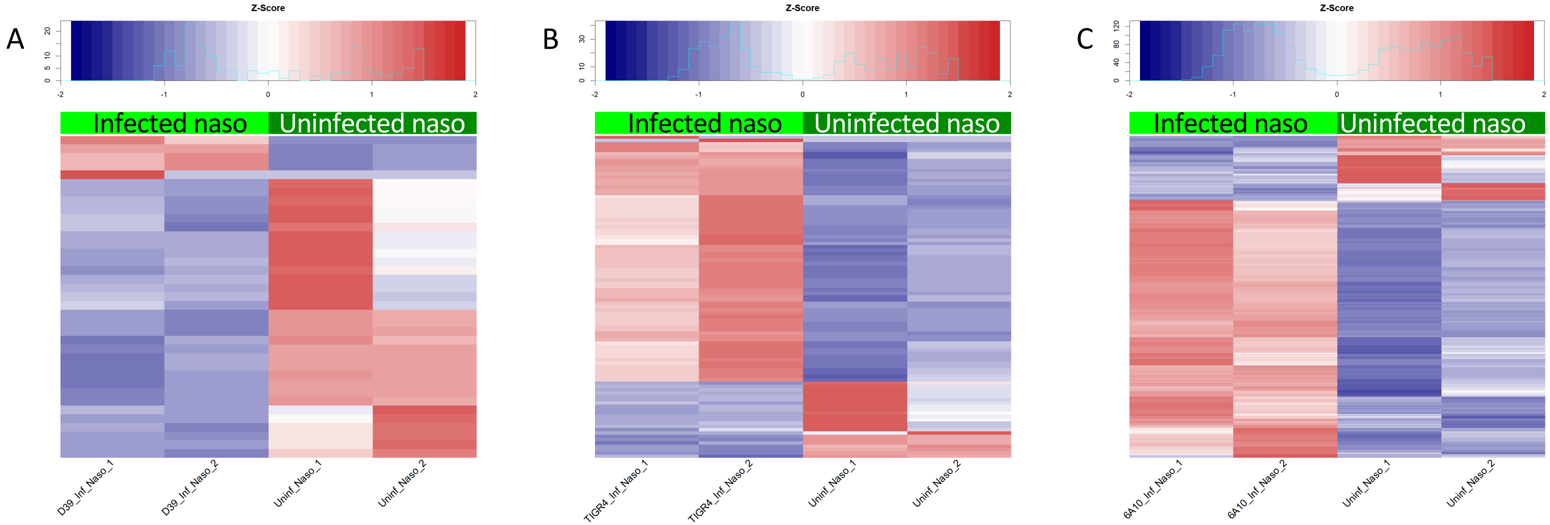

Fig. S5. Differentially expressed (DE) genes of the host in the nasopharynx. Z-scored heatmaps of expression levels of DE genes from host nasopharyngeal samples. Relatively fewer genes were shared across all host colonized samples versus diseased samples, therefore, unlike Figure S4 where organ-specific subsets of genes that were DE across 3 pneumococcal strains (TIGR4, D39, 6A-10), the heatmaps presented here encompass all host DE genes for each pneumococcal strain. **A.** Heatmap of D39 colonized nasopharynx. **B.** Heatmap of TIGR4 colonized nasopharynx. **C.** Heatmap of 6A-10 colonized nasopharynx.

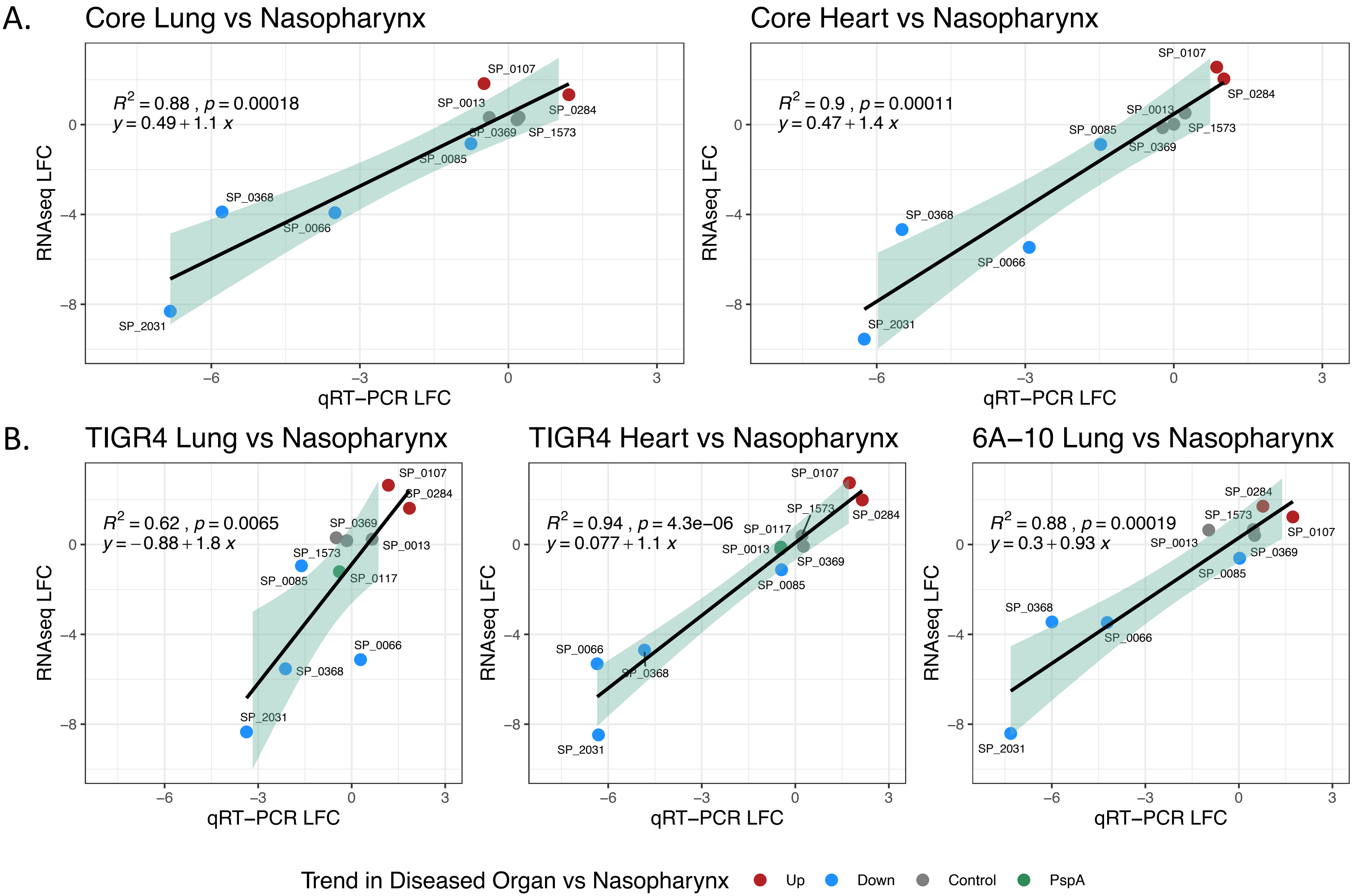

Fig. S6. Log$_2$ Fold Change (LFC) correlation plots of RNA-seq LFC vs qRT-PCR $\Delta\Delta$Ct (equivalent to LFC) in pneumococcal infected lung, heart and nasopharynx. Up- and downregulated genes were selected from the 69 genes differentially expressed at all sites relative to the nasopharynx (Figure 4B**).** qRT-PCR was performed on (i) the same RNA samples that were sequenced as validation of the RNA-seq measurements, (ii) additional TIGR4 and D39 infected hearts that were not subjected to RNA-seq (Note that these RNAs were depleted after completion of this experiment). Strong correlations with significant p-values are observed for the 10 genes tested. PspA primers were not specific in 6A-10 hence not shown in these plots. Green shadings indicate one standard error from the linear regression. **A.** RNA-seq LFC values calculated using core pneumococcal genes shared across 3 strains (TIGR4, D39, 6A-10). **B.** Strain-specific RNA-seq values.

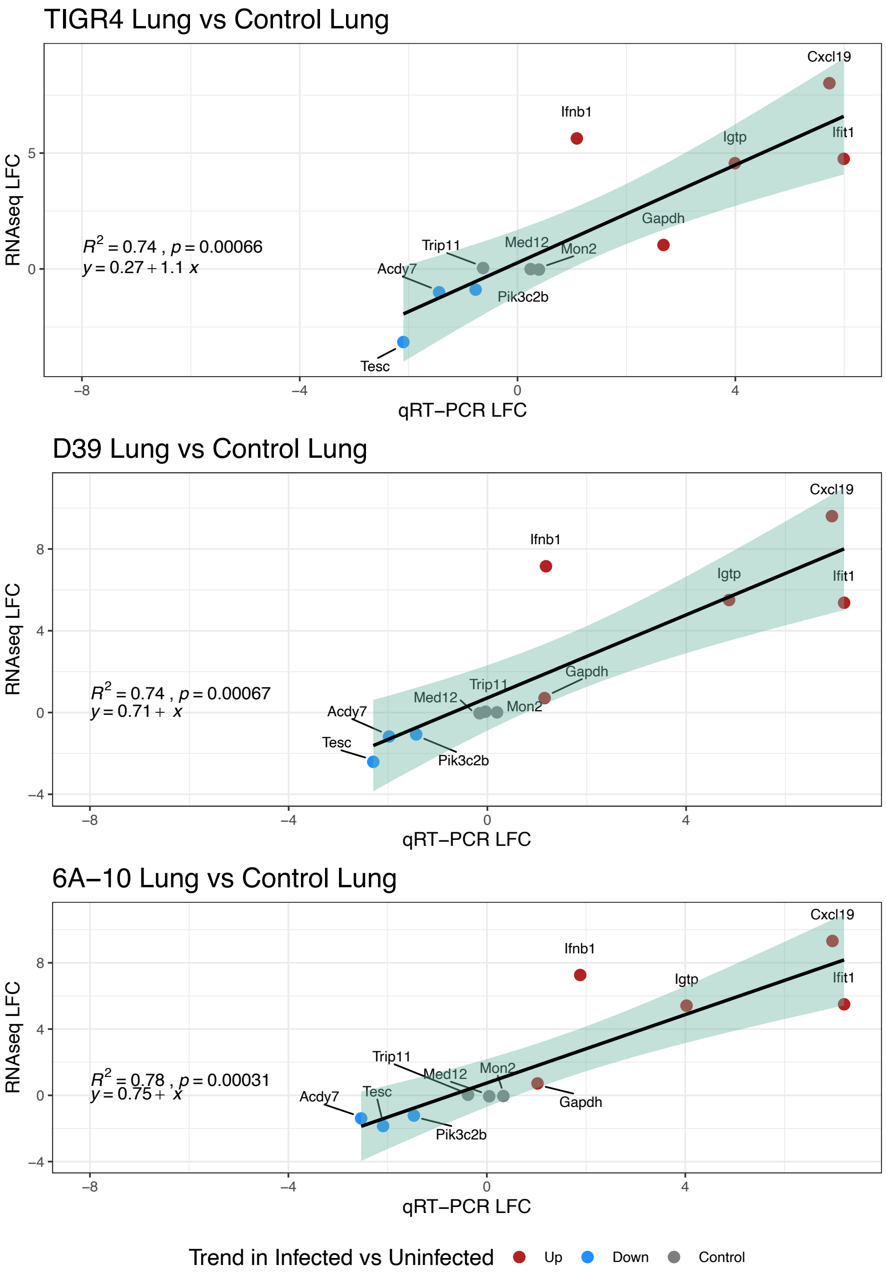

Fig. S7. Log$_2$ Fold Change (LFC) correlation plots of RNA-seq LFC vs qRT-PCR ΔΔCt (equivalent to LFC) in the mouse lung. Up- and downregulated genes were selected from the 190 genes differentially expressed at all infected sites relative to uninfected sites, excluding the nasopharynx (Figure 4D). qRT-PCR was performed on the same RNA samples that were sequenced as validation of the RNA-seq measurements. Strong correlations with significant p-values are observed for the 11 genes tested for each individual infecting pneumococcal strain. Green shadings indicate one standard error from the linear regression.

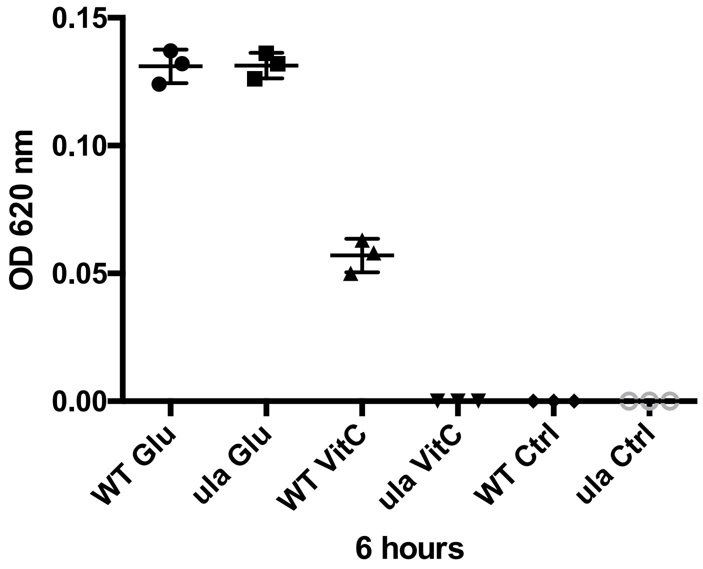

Fig. S8. The pneumococcus can rely on the *ula* operon for growth with ascorbate as a sole carbon source. TIGR4 and its isogenic mutant ∆*ula* were inoculated in Chemically Defined Medium (CDM) supplemented with glucose or ascorbic acid as the sole carbon source and allowed to grow for 6 hours. The optical density of bacterial cultures is shown. While both TIGR4 and ∆*ula* can grow on glucose at the same rate, ∆*ula* was unable to grow on ascorbic acid as a sole carbon source.

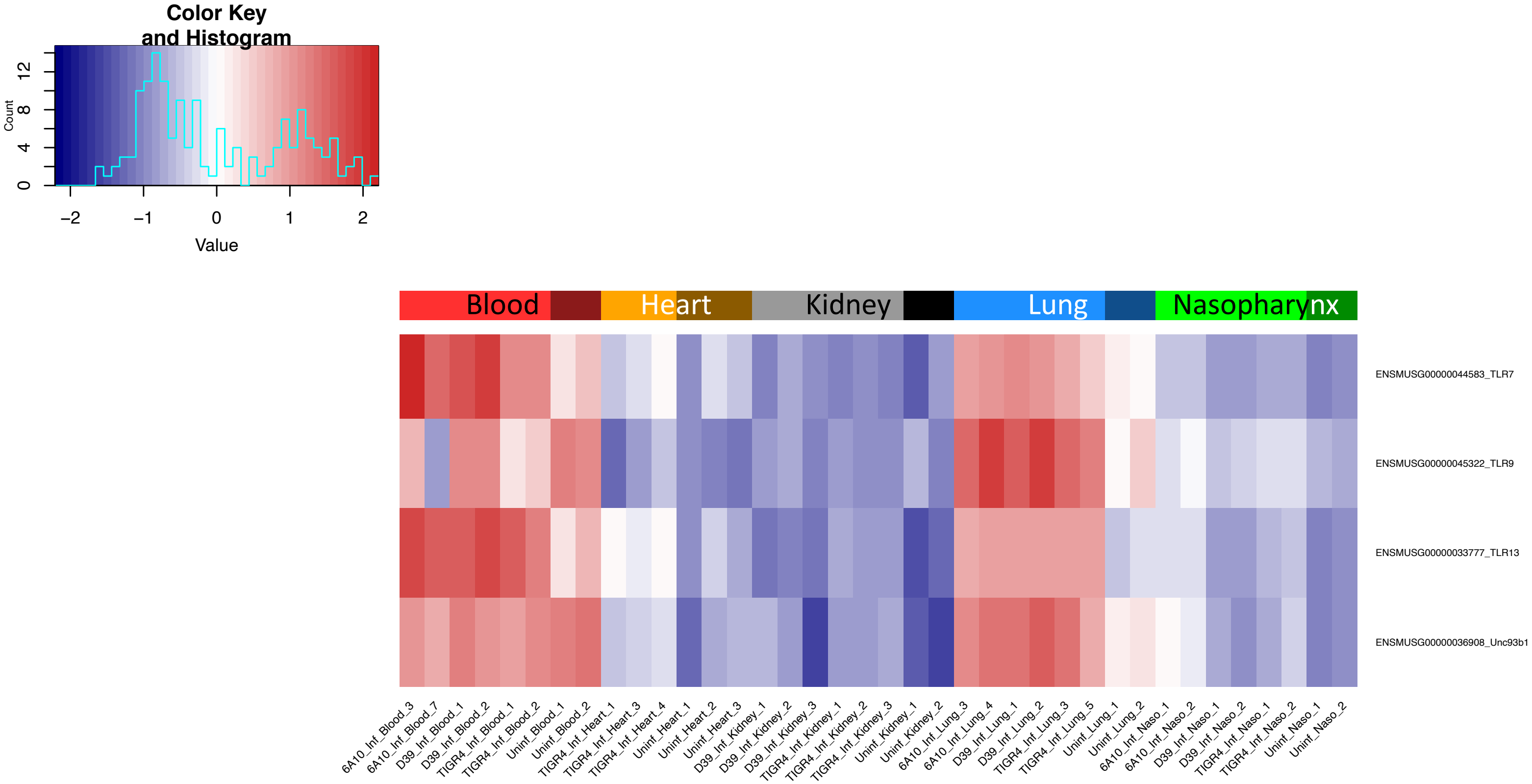

Fig. S9. Z-scored heatmap of host TLR 7, 9, 13, and chaperone protein Unc93b1. These host genes are upregulated under conditions of particular interest within the blood and lungs, as compared to the study by Famà et al. 2020.

SI References

1. Y. Li, C. M. Thompson, M. Lipsitch, A modified Janus cassette (Sweet Janus) to improve allelic replacement efficiency by high-stringency negative selection in Streptococcus pneumoniae. PLoS One 9, e100510 (2014).
2. A. L. Bricker, A. Camilli, Transformation of a type 4 encapsulated strain of Streptococcus pneumoniae. FEMS Microbiol Lett 172, 131-135 (1999).
3. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357-360 (2015).
4. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).
5. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169 (2015).
6. D. E. Fouts, L. Brinkac, E. Beck, J. Inman, G. Sutton, PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40, e172 (2012).
7. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).
8. A. T. Shenoy et al., Streptococcus pneumoniae in the heart subvert the host response through biofilm-mediated resident macrophage killing. PLoS Pathog 13, e1006582 (2017).
9. R. Aprianto, J. Slager, S. Holsappel, J. W. Veening, High-resolution analysis of the pneumococcal transcriptome under a wide range of infection-relevant conditions. Nucleic Acids Res 46, 9990-10006 (2018).
10. A. Famà et al., Nucleic Acid-Sensing Toll-Like Receptors Play a Dominant Role in Innate Immune Recognition of Pneumococci. mBio 11, e00415-00420 (2020).