Supporting Information

Thompson et al. 10.1073/pnas.0912386106

SI Text

Additional Details on Data Filtering and Analysis Background was measured using the Edwards calculation (1); data were then normalized by intensity-dependent Loess estimation and scaled by the median absolute deviation for each array. The $\log_2$ Cy5/Cy3 ratios were median-centered across each array in a given set hybridized on the same day. Samples from each time point were evenly distributed across hybridization sets, and 10 replicate arrays were done on different days. The centering did not induce bias as 8/10 of the replicate pairs clustered next to each other in an unsupervised hierarchical cluster of all 183 arrays (data not shown) and the other two pairs were found within the same cluster arms.

Genes and arrays were organized using average linkage hierarchical clustering using Cluster 3.0 (2). To show changes in gene expression over time in comparison to healthy controls (HC), separate clusters of arrays from each time point were performed preserving the gene order within each set. Correlation coefficients were calculated between the gene expression of every gene and each measured clinical parameter as described in refs. 3 and 4. Gene expression data were randomly permuted 1,000 times; the correlation coefficients derived from each permutation of the data were compared with that from the actual data. The *P* value for each gene was calculated as the fraction of the permutations that resulted in a correlation coefficient as high as was observed in the actual data.

SAM (5) was used to identify transcripts that were significantly more or less abundant in a specified group of samples. The nonparametric Mann–Whitney statistical test was used to assess significant differences in CBCs. Gene Ontology analysis was performed with Genetrail (6) and a false discovery rate (FDR) of 5%. The Molecular Signatures Database MSigDB (7) and NextBio were used to identify relevant published gene expression datasets, using a *P* value of 0.05.

1. Edwards D (2003) Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics* 19:825–833.
2. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868.
3. Simmons CP, et al. (2007) Patterns of host genome-wide gene transcript abundance in the peripheral blood of patients with acute dengue hemorrhagic fever. *J Infect Dis* 195:1097–1107.
4. Popper SJ, et al. (2007) Gene-expression patterns reveal underlying biological processes in Kawasaki disease. *Genome Biol* 8:R261.
5. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
6. Backes C, et al. (2007) GeneTrail—Advanced gene set enrichment analysis. *Nucleic Acids Res* 35:W186–W192.
7. Rose FR, et al. (1998) Potential role of epithelial cell-derived histone H1 proteins in innate antimicrobial defense in the human gastrointestinal tract. *Infect Immun* 66:3255–3263.

Fig. S1. An extension of Fig. 2 showing all time points from the typhoid patients as well as the T1 malaria samples. Transcripts (1,082) determined by SAM analysis to vary significantly in abundance from the HC samples at T1, T28, or T9M in the typhoid samples were used to cluster samples from each time point, preserving the gene order across all arrays and clusters. Red indicates high expression, and green indicates low expression as shown in the legend; gray indicates missing data. Pearson correlation coefficients and *P* values were calculated for the expression of every gene and each clinical parameter (percent neutrophils, percent lymphocytes, body temperature, gender, percent HCT, and PLT) across a selected set of 85 samples that had full clinical data. The plots to the right of the clusters show the negative $\log_{10}$ of the *P* value signed according to the sign of the calculated correlation coefficient. The *P* values are plotted as moving averages of three genes (along the vertical axis). The red vertical lines on each plot indicate a *P* value of 0.05.

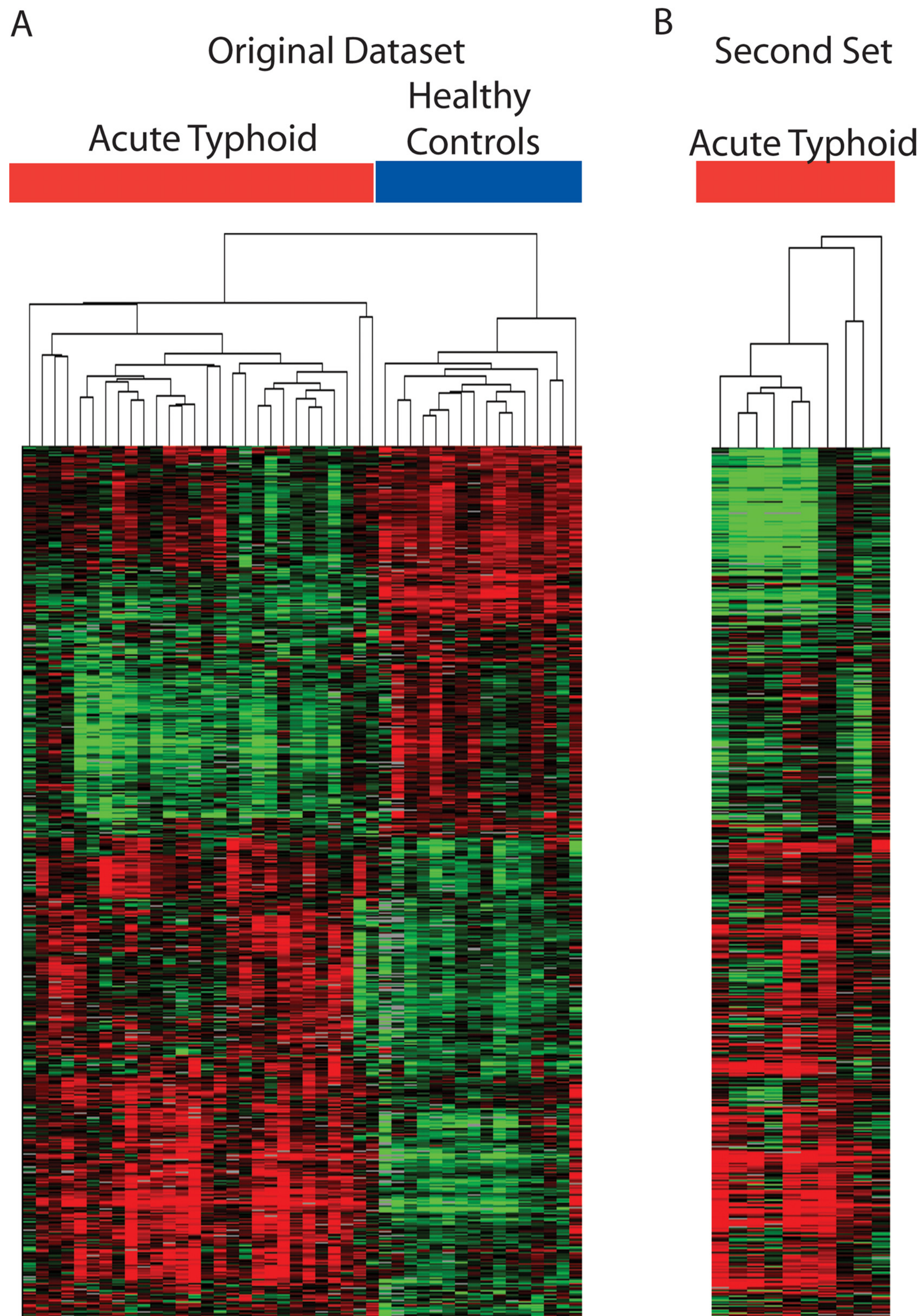

Fig. S2. Comparison of the transcriptional response during acute typhoid in the original set of patients (29 individuals) with that in the second set of typhoid patients (10 individuals) obtained a full year later. (*A*) Hierarchical clustering using the transcripts that differ significantly in abundance in the original set of acute typhoid samples as compared with healthy controls (list in Dataset S1). The dendrogram separates the acute typhoid from the control samples on separate branches as indicated at the top of the figure. (*B*) Clustering of the second set of acute typhoid samples maintaining the same gene order as in panel *A*.

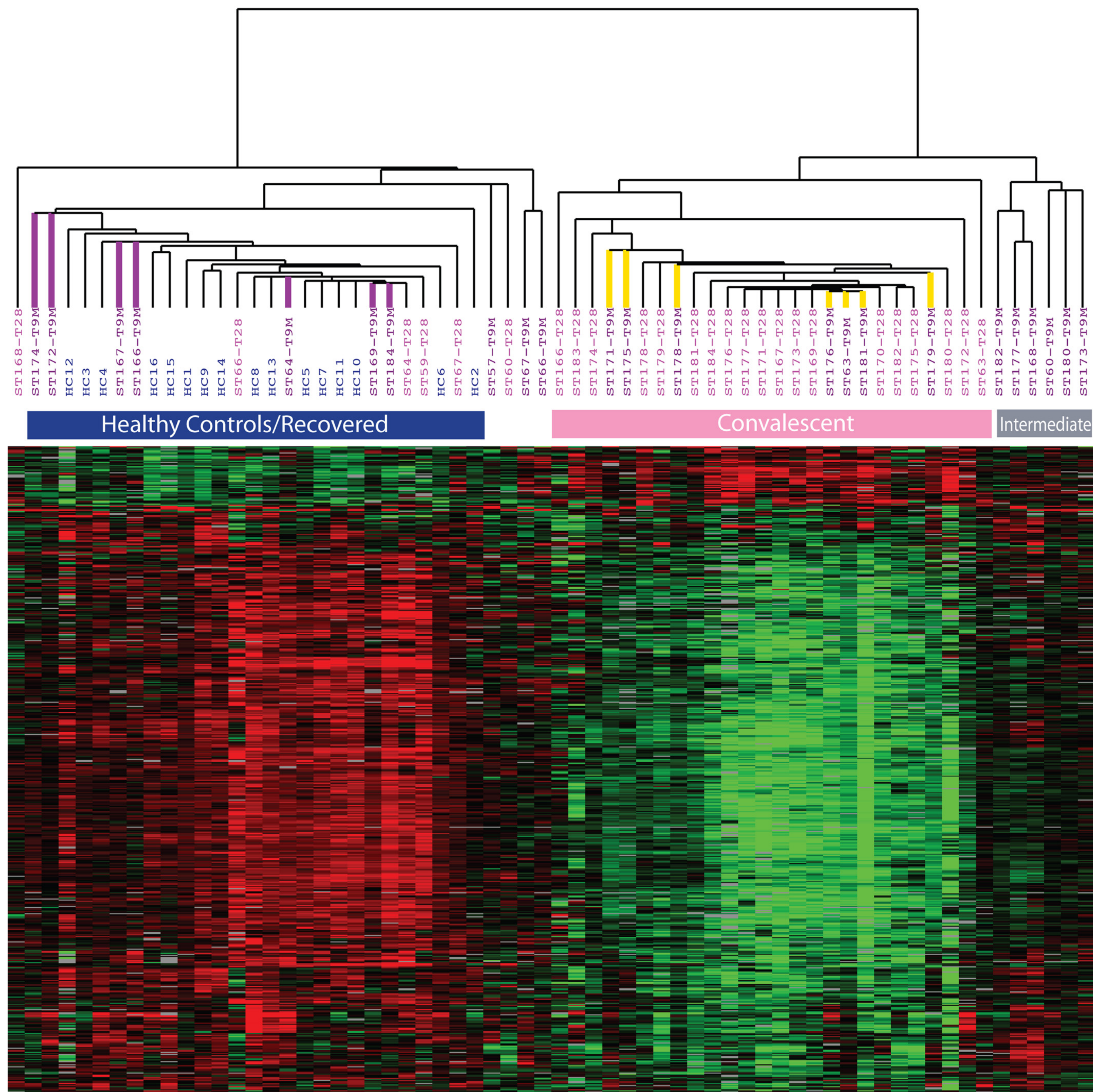

Fig. S3. Classification of T9M typhoid samples. Transcripts determined by SAM analysis to vary significantly between T28 typhoid samples and healthy controls (517 features) were used to hierarchically cluster T28 and T9M typhoid samples and HC samples. Red indicates high expression, and green indicates low expression; gray indicates missing data. Individual samples are color coded: HC are blue, T28 typhoid samples are pink, and T9M typhoid samples are purple. Colored horizontal bars indicate particular subsets of array samples: HC/Recovered (blue), Convalescent (pink), and Intermediate (gray). The blue cluster contains all of the HC samples as well as seven T9M samples that are indicated by purple vertical lines in the dendogram. These seven T9M samples were designated as having returned to ''normal.'' The pink cluster contains the majority of the T28 samples plus seven T9M samples indicated by yellow vertical lines in the dendogram. The latter were designated as ''unusual'' samples.

Other Supporting Information Files

Dataset S1
Dataset S2