Table 1. Enrichment of Gene Ontology (GO) annotation terms in the bacterial diversity data set

GO#	Annotation Term	Example Induced Genes	p value
6960	Antiapoptosis	c-IAP2, Bfl-1, Bcl-XL	8.3×10^{-3}
7276	Antimicrobial humoral response	SLAM, CD40, Adenosine deaminase	9.0×10^{-6}
2125	Cell-cell signaling	TNFα, IL6, MIP1α, MIP1β	2.9×10^{-6}
5615	Chemokine	PARC, LARC, LD78β	8.3×10^{-3}
7165	Cytokine	IL1α, IL1β, IL10	2.6×10^{-4}
6916	Extracellular space	TGFα, G-CSF, HDGF	1.0×10^{-3}
6960	Signal transduction	STAT3, CD69, Thioredoxin, CREB	5.2×10^{-4}
GO#	**Annotation Term**	**Example Repressed Genes**	p **value**
6954	Inflammatory response	5-LO, CD14, CYBB, CCR1, FPR	1.0×10^{-5}
9596	Perception of pest/pathogen/	MHC II transactivator, HLA-DMα, HLA-DMβ	1.0×10^{-2}
5886	Plasma membrane	TNFR1, CD13, CD31, RGS14, CD105	2.9×10^{-4}

Using the bacterial diversity data set, 1,208 unique genes were selected based on having a Locus Link assignment (http://www.ncbi.nlm.nih.gov/LocusLink). p values for the frequency of 184 GO annotation terms (those that had at least five occurrences in the data set) were calculated based on the hypergeometric distribution. Those annotation terms found to have p values less than 1% within the common induction (red) and common repression (green) clusters are displayed with their GO identification numbers and example genes from the 1,208-gene data set. The GO annotations for genes in the entire bacterial diversity data set were extracted in batch from SOURCE (http://genome-www.stanford.edu/source), using the Locus Link identification number for each gene. The resulting 1,208 unique annotated genes were used for the analysis, among which 184 annotation terms occurred for at least five genes. We compared the frequency of these 184 annotations in the data set as a whole to that in the common induction and common repression clusters, using the hypergeometric distribution to calculate p values [Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285] (see equation below). Let $N = 1{,}208$ denote the total number of genes under consideration and A, the number of these genes with a particular annotation. The chance of observing at least x genes with that annotation in a random subset of n genes is given by

$$p(x; N, A, n) = 1 - \sum_{i=0}^{x-1} \frac{\binom{A}{i}\binom{N-A}{n-i}}{\binom{N}{n}}, \text{ where } \binom{N}{n} = N!/n!(N-n)!.$$