Novel Algorithms Reveal Streptococcal Transcriptomes and Clues about Undefined Genes

Patricia A. Ryan*, Brian W. Kirk, Chad W. Euler, Raymond Schuch, Vincent A. Fischetti

Department of Bacterial Pathogenesis and Immunology, Rockefeller University, New York, New York, United States of America

Bacteria–host interactions are dynamic processes, and understanding transcriptional responses that directly or indirectly regulate the expression of genes involved in initial infection stages would illuminate the molecular events that result in host colonization. We used oligonucleotide microarrays to monitor (in vitro) differential gene expression in group A streptococci during pharyngeal cell adherence, the first overt infection stage. We present neighbor clustering, a new computational method for further analyzing bacterial microarray data that combines two informative characteristics of bacterial genes that share common function or regulation: (1) similar gene expression profiles (i.e., co-expression); and (2) physical proximity of genes on the chromosome. This method identifies statistically significant clusters of co-expressed gene neighbors that potentially share common function or regulation by coupling statistically analyzed gene expression profiles with the chromosomal position of genes. We applied this method to our own data and to those of others, and we show that it identified a greater number of differentially expressed genes, facilitating the reconstruction of more multimeric proteins and complete metabolic pathways than would have been possible without its application. We assessed the biological significance of two identified genes by assaying deletion mutants for adherence in vitro and show that neighbor clustering indeed provides biologically relevant data. Neighbor clustering provides a more comprehensive view of the molecular responses of streptococci during pharyngeal cell adherence.

Introduction

Microarray technology is now commonly used to reveal genome-wide transcriptional changes in bacterial pathogens during interactions with the host. Several factors, however, limit the power of such analyses, including inadequate statistical analysis and insufficient sample replication, both of which do not account for experimental variability, and often result in arbitrary thresholds for significance [1,2]. In addition, unknown bacterial genes can confound the interpretation of expression profiles, restricting many microarray studies to the differential expression of well-characterized genes.

Several methods are available to organize gene expression profiles and to assist in extracting functional or regulatory gene information from microarray datasets. Clustering algorithms group genes by similarities in expression patterns, based on the assumption that co-expressed genes share common function or regulation [3,4]; however, clustering solely by co-expression patterns may not reveal a considerable amount of information contained in array data. These methods often: (1) produce unreliable data by missing known gene members of biological pathways; (2) fail to distinguish truly related gene clusters from coincidental groupings; and (3) identify clusters containing only unknown genes that may lack either common function or regulation, a considerable limitation for genomes containing a large percentage of undefined genes [1,2]. Because no tools exist to interpret unknown gene clusters or to assess their significance and completeness, a significant portion of bacterial expression profiles are not interpretable using current clustering methods.

We introduce neighbor clustering as a new tool for analyzing bacterial microarray data that addresses some of these limitations by incorporating the physical position of genes on the bacterial chromosome into the analysis of expression data. Information about gene function and regulation is stored intrinsically in the bacterial genome structure, as genes with common function or regulation tend to be physically proximate on the chromosome and often linked as operons [5,6]. We incorporated these positional data into a series of neighbor clustering algorithms, named GenomeCrawler, that identifies groupings of potentially related genes from array data by combining two informative characteristics of bacterial genes that share common function or regulation [3–6]: (1) similar gene expression profiles (i.e., co-expression); and (2) physical proximity of genes on the chromosome. The algorithms also recalculate the statistical significance of each gene as a member of a particular cluster,
Microarray technology is commonly used to reveal genome-wide transcriptional changes in bacterial pathogens during interactions with the host. Clustering algorithms, which group genes with similar expression patterns, facilitate microarray data organization and are based on assumptions that co-expressed genes share common function or regulation; however, clustering solely by co-expression may not reveal all of the information contained in bacterial array data. We introduce neighbor clustering, a new tool for analyzing bacterial gene expression profiles, which distinguishes itself from other programs by incorporating details unique to the architecture of bacterial chromosomes into the analysis. Neighbor clustering combines two informative characteristics of bacterial genes that share common function or regulation—(1) similar expression profiles and (2) physical proximity on the chromosome—and extracts statistically significant clusters of gene neighbors that are potentially related by function or regulation. We present the analysis of microarray data from group A streptococci during adherence to human pharyngeal cells, the first overt infection step. We show that neighbor clustering identifies more differentially expressed genes than rigorous statistical analyses alone, and can provide functional clues about unknown genes. We extended the analysis to include a previously published streptococcal array study to demonstrate the applicability of the method.

**Results/Discussion**

**Adherence-Mediated Differential Expression**

We developed spotted oligonucleotide arrays of the *S. pyogenes* SF370 (an M1 serotype) genome [14] and compared the transcriptionomes of streptococci that adhere to Detroit 562 human pharyngeal cells to non-adherent (“associated”) streptococci within the same experiment. Adherence assays were performed as described [15] with modifications to minimize eukaryotic cell disruption. We replicated experiments independently and used dye-swaps to incorporate biological and technical variation [16,17]. Following filtering and normalization [18,19], we analyzed data from four biological replicates [16] with robust summary statistics [20], Bayesian statistics [21,22], and permutation algorithms [19] to identify genes differentially expressed with significance during pharyngeal cell adherence.

This analysis identified 70 genes (4% of the genome) exhibiting statistically significant fold changes in expression ($P_F < 0.05$) during adherence from 1,769 open reading frames presented on the array (Table 1). We refer to such genes as “differentially expressed.” We present the entire dataset from all experiments as Table S1. Genes demonstrating upregulation ($n = 45$) and downregulation ($n = 34$) included virulence factors, prophage-encoded transcripts, metabolic genes, and transcriptional regulators (Table 2). Undefined or hypothetical genes comprised 27% of differentially expressed genes ($n = 21$; 11 chromosomally encoded genes, ten phase-encoded genes).

**Verification by Quantitative Real-Time PCR**

We conducted TaqMan (qRT-PCR) analysis [23] of 11 differentially expressed genes to validate selected microarray hybridization results (see Table S2 for genes and primer-probe sequences). Five genes chosen for validation demonstrated statistically significant fold changes in expression by microarray analysis ($P_F < 0.05$; two upregulated, three downregulated). The remaining six genes (four upregulated, two downregulated) did not have significant $P_F$ values, but were statistically significant as members of particular neighbor clusters in subsequent analyses ($P_E < 0.05$ as detailed in later sections). We averaged the data to generate a value for each gene, creating a set of 11 paired values from quantitative real-time (qRT)-PCR and microarray analyses (Table S3). Results of standard linear regression analysis demonstrated a strong positive correlation ($r = 0.9$) between data obtained using the different techniques (see Figure S1).

**Virulence Factors**

Streptococci elaborate several factors implicated in infection, including surface-exposed adhesins and secreted toxigenic proteins (reviewed in [7,14,24]). The initial statistical analysis identified four differentially expressed virulence genes (Tables 1 and 2). Genes encoding streptolysin O (*slo* or *spy0167*) and the SpeB protease (*spy2039*) were downregulated, while genes encoding pyrogenic exotoxin H (*speH* or *spy1008*) and a putative fibronectin-binding protein (*spy0130*) were upregulated. We verified the differential expression of *spy2039* and *spy0130* by qRT-PCR.
Table 1. Summary of Streptococcal Genes Exhibiting Significant Changes in Expression during Adherence to Pharyngeal Cells Compared with Associated Streptococcal Control

| Functiona | Gene productb | Gene Numberc | Log₂ Fold Changesd | P-valuee |
|-----------|---------------|--------------|--------------------|----------|
| Carbohydrate transport and metabolism | N-acetylmannosamine-6-P epimerase | spy0025 | 1.1726 | 0.023 |
| | PTS system, enzyme IB component | spy0105 | −1.5801 | 0.009 |
| | PTS system, enzyme IIC component | spy0109 | −1.716 | 0.002 |
| | Conserved hypothetical protein | spy1340 | −2.3914 | 0.006 |
| | lacD,1, tag-1,6-P aldolase | spy1704 | 5.3533 | <0.001 |
| | lacB, gal-6-P isomerase | spy1707 | 3.5548 | <0.001 |
| | lacA, gal-6-P isomerase | spy1708 | 3.5162 | <0.001 |
| | PTS system, enzyme IIC component | spy1709 | 3.7587 | 0.001 |
| | PTS system, enzyme IIB component | spy1710 | 3.6824 | 0.001 |
| | PTS system, enzyme IIA component | spy1711 | 3.6258 | 0.006 |
| | lacG, phospho-beta-D-galactosidase | spy1916 | 1.9633 | 0.037 |
| | lacA2, galactosidase acetyltransferase | spy1923 | 1.7327 | 0.039 |
| | mipB, transaldolase-like protein | spy2048 | −1.5419 | 0.001 |
| | Cell division | ftsH, cell division protein | spy0015 | 1.6818 | 0.001 |
| | | dgk, diacylglycerol kinase | spy0047 | 1.0861 | 0.021 |
| | | pbp2A, penicillin-binding protein | spy2059 | 1.4125 | 0.006 |
| | Coenzyme metabolism | folE, GTP cyclohydrolyase | spy1097 | 1.7146 | 0.001 |
| | | folP, dihydroteratoate synthase | spy1098 | 1.4278 | 0.009 |
| | | folQ, dihydropterin aldolase | spy1099 | 1.5245 | 0.009 |
| | | lplA, lipoate-protein ligase | spy1214 | 2.1793 | 0.006 |
| | | Hypothetical protein | spy1215 | 1.9938 | 0.021 |
| | DNA replication and repair | dnaE, DNA polymerase III (alpha subunit) | spy1284 | 1.4147 | 0.005 |
| | Energy production and conversion | gpsiA, glycerol-3-P dehydrogenase | spy0026 | 1.396 | 0.005 |
| | | atpH, proton-translocating ATPase | spy0757 | 1.8028 | 0.006 |
| | | atpG, proton-translocating ATPase | spy0759 | 1.6569 | 0.001 |
| | | atpD, proton-translocating ATPase | spy0760 | 1.6172 | 0.006 |
| | | atpc, proton-translocating ATPase | spy0761 | 1.8812 | <0.001 |
| | | Mg”/citrate complex transporter | spy1180 | −1.5147 | 0.018 |
| | | glpP, glycogen phosphorylase | spy1291 | −1.6642 | 0.017 |
| | | gidA, glyceral dehydrogenase | spy2047 | −1.6212 | 0.012 |
| | Function unknown | Hypothetical protein | spy0128 | 2.0769 | 0.006 |
| | | Hypothetical protein | spy0129 | 2.604 | <0.001 |
| | | Hypothetical protein | spy0241 | 1.6629 | 0.03 |
| | | Hypothetical protein | spy0646 | 1.3213 | 0.018 |
| | | Hypothetical protein | spy1216 | 1.7947 | 0.005 |
| | | Hypothetical protein | spy1701 | 1.8102 | 0.006 |
| | | Hypothetical protein | spy1936 | −1.9742 | 0.039 |
| | | Hypothetical protein | spy2115 | 2.2803 | 0.006 |
| | | Hypothetical protein | spy2215 | 1.838 | 0.031 |
| General function prediction | Signal peptide I | spy0127 | 1.9277 | 0.001 |
| | | Oxidoreductase | spy2107 | −2.237 | 0.004 |
| Inorganic ion transport and metabolism | phiA, alkaliphosphatase uptake | spy1277 | −2.4511 | <0.001 |
| Lipid metabolism | mvaS.1, HMG-CoA | spy0880 | 1.2361 | 0.006 |
| | | mvaS.2, HMG-CoA synthase | spy0881 | 1.3164 | 0.013 |
| | | atoA, acetyl-CoA:acetocetyl-CoA transferase | spy1639 | 1.1453 | 0.044 |
| | | accA, acetyl-CoA carboxylase subunit | spy1743 | −2.4735 | 0.041 |
| | | accC, acetyl-CoA carboxylase subunit | spy1745 | −2.3576 | 0.018 |
| | | fabH, beta ketoacyl-ACP synthase III | spy1754 | −2.4017 | 0.006 |
| | | phaB, enoyl CoA hydratase | spy1758 | −3.2466 | 0.001 |
| Nucleotide transport and metabolism | Phosphoribosylformylglycinamidine synthase | spy0025 | −1.3616 | 0.028 |
| | | purM, phosphoribosylformylglycinamidine cyclase | spy0027 | −1.3144 | 0.015 |
| | | purA, adenylsuccinate synthetase | spy0160 | −1.3497 | 0.044 |
| | | pyrE, orotate phosphoribosyltransferase | spy0901 | −1.9544 | 0.009 |
| Phage | Phage hypothetical protein | spy0940 | −2.8802 | <0.001 |
| | | Phage hypothetical protein | spy0947 | −2.9017 | 0.001 |
| | | Phage hypothetical protein | spy0952 | −3.3558 | <0.001 |
| | | Phage hypothetical protein | spy0956 | −3.3963 | <0.001 |
| | | Phage conserved hypothetical protein | spy0958 | −3.1246 | <0.001 |
| | | Phage hypothetical protein | spy0961 | −2.7622 | 0.006 |
| | | Phage hypothetical protein | spy0962 | −2.5092 | 0.001 |
| | | Phage hypothetical protein | spy0963 | −3.5341 | <0.001 |
| | | Phage hypothetical protein | spy0965 | −3.5579 | <0.001 |
| | | Phage conserved hypothetical protein | spy0967 | −2.324 | 0.009 |
| Posttranslational modification | Hypothetical protein | spy2037 | 2.6772 | 0.011 |
| | | nrdG, anaerobic ribonucleotide reductase activator | spy2105 | −1.648 | 0.018 |
| Signal transduction mechanism | Histidine kinase | spy1236 | 2.4838 | <0.001 |
| | | Response regulator (clAl homolog) | spy1237 | 2.0617 | 0.004 |
The downregulation of virulence loci during presumably inappropriate stages of infection was not surprising. Streptolysin O is a cytokin that damages human tissue and increases host cell cytotoxicity [7,25]. The resulting cellular damage, particularly to polymorphonuclear leukocytes [26], decreases internalization and subsequent intracellular killing of streptococci [27]. Based on its downregulation during adherence, we infer that slo was transcribed during pre-adherence associations, perhaps, as previously reported, to protect streptococci from phagocytic killing in vivo [27]. However, once adhered, our data suggest that streptococci downregulate production of this cytokin, presumably to prevent further host tissue destruction that could interfere with adherence.

SpeB (encoded by spy2039) is a multifunctional cysteine protease implicated in numerous infection strategies [28,29]. Although few studies have examined gene expression patterns during adherence, SpeB production (as detected by Western blot analysis) decreases during co-culture with human peripheral blood mononuclear cells [30] and in a mouse infection model [31]. When SpeB expression is limited, several streptococcal proteins necessary for adherence remain intact [24,32,33]; thus, decreased SpeB production (as indicated here) may promote pharyngeal cell attachment. Furthermore, SpeB abolishes internalization (following adherence) of certain streptococcal strains by epithelial cells (including Detroit 562 cells), a process mediated in part by the fibronectin-binding protein F [34,35]. We observed significant upregulation of the gene spy0130, encoding a protein recently found to be associated with the production of surface-exposed pili on strain SF370 [36]. The protein shares 60% sequence similarity to protein F, suggesting that it may coordinate a similar internalization mechanism or may be involved directly in adherence (discussed later in detail). SpeB downregulation also coincides with increased expression of pyrogenic exotoxins [33,37] that reportedly increase streptococcal survival in vivo. We observed that the exotoxin-encoding speH gene [38] was upregulated. Taken together, our results agree with previous reports on SpeB production during host cell interactions, suggesting that decreased expression may promote streptococcal adherence (by preventing proteolytic degradation of key virulence factors or adhesins), enhance internalization (perhaps through a fibronectin-mediated pathway), and increase survival (through increased pyrogenic exotoxin production, discussed below).

Phage-Encoded Genes

SF370 contains one inducible prophage (370.1) and three defective prophages (370.2, 370.3, and 370.4) that produce no infectious phage [39]. We identified 11 differentially expressed phage 370.2 genes, suggesting that this defective phage is not transcriptionally silent (Table 1). The speH gene (spy1008) was induced, and the remaining genes, hypothetically involved in replication and regulation [39], were downregulated. The speH gene encodes a mitogenic exotoxin [38] reportedly induced during polymorphonuclear leukocyte phagocytosis [8] but not implicated previously in adherence.

Allelic Replacement of speH

Increased expression of speH during pharyngeal cell adherence suggests that the SpeH exotoxin is either necessary for adherence, or is a component of a downstream infection process. Adherence-mediated upregulation of speH is likely not the result of phage induction, as the remaining phage 370.2 genes identified in our analysis were downregulated. To determine if SpeH plays a direct role in the adherence process, we created a deletion mutant in strain SF370 (SF370ΔspeH), which was confirmed by PCR (unpublished data) and RT-PCR (Figure 1A) and tested in vitro for adherence to human pharyngeal cells. We observed no significant difference in adherence between the wild-type (SF370) and mutant strains (Figure 1B), indicating that SpeH is not involved directly in attachment to the pharyngeal cell. The significant upregulation of the speH gene during adherence suggests that the gene product may function instead during a subsequent stage of infection.

### Table 1. Continued.

| Functiona | Gene productb | Gene Numberc | Log2-Fold Changedd | P valuee |
|-----------|---------------|--------------|-------------------|----------|
| Transcription | Histidine kinase | spy1622 | 1.1531 | 0.032 |
| Translation | Hypothetical protein | spy0228 | 1.1361 | 0.049 |
| Virulence | Hypothetical protein (ropB regulator) | spy0583 | 2.2766 | 0.001 |
| | Hypothetical protein (ropB regulator) | spy2401 | 1.1416 | 0.03 |
| | Hypothetical protein (ropB regulator) | spy2042 | 1.8458 | 0.032 |
| | infC, translation initiation factor 3 | spy0804 | −2.0177 | 0.047 |
| | rpsP, 30s ribosomal protein s16 | spy0840 | −1.1657 | 0.018 |
| | def, polypeptide deformylase | spy1958 | 1.8806 | 0.025 |
| | Hypothetical protein (protein F homolog) | spy0130 | 2.4112 | <0.001 |
| | slo, streptolysin O | spy0167 | −1.9713 | 0.037 |
| | speH, phage encoded pyrogenic exotoxin H | spy1008 | 1.4486 | 0.01 |
| | speB, pyrogenic exotoxin B | spy2039 | −2.3306 | |

---

*Function, gene product, and gene number designations from the annotated SF370 genome.

Fold change in expression ratio (adherent streptococci relative to associated control) calculated for each of the four biological replicates analyzed.

*P* values calculated as detailed in the Methods section. Genes with *P* values < 0.05 were considered to be undergoing a statistical fold change in expression during adherence compared with associated control.

doi:10.1371/journal.pcbi.0030132.s001
Differential Expression of Genes from Diverse Functional Categories

We identified a number of genes encoding proteins involved in housekeeping processes (such as carbohydrate and coenzyme metabolism) that were differentially expressed, indicating a shift in metabolic processes due to host cell adherence (Tables 1 and 2). For example, genes encoding proteins involved in folate biosynthesis [40] were upregulated, suggesting that certain cofactors that may be necessary during adherence were unavailable. Also upregulated were genes encoding subunits of the F0F1 ATPase [41] (discussed in more detail later), which may indicate an acid stress response to maintain cytoplasmic pH or a need to generate ATP in response to increased energy requirements.

We also identified the adherence-mediated upregulation of four transcriptional regulators (Table 1), suggestive of an adaptive response to host cell contact that is dynamic and complex. For example, RopB (encoded by spy2042), a member of the Rgg family of response regulators, interacts with a number of regulatory networks throughout the streptococcal genome (e.g., mga, csrRS, sagA, and fasBCA), affecting the transcription of numerous proteins, virulence factors, and two-component regulatory systems [42,43]. Although the delineation of genes influenced by RopB (or any identified transcriptional regulator) is beyond the scope of this study, our initial analysis did identify the upregulation of a two-component regulatory system, encoded by spy1236–1237. The functions of these particular loci are not yet known, and their adherence-mediated upregulation represents new targets in the study of regulators that function during host cell contact.

Table 2. Functional Categories of Streptococcal Genes Exhibiting Significant Changes in Expression during Adherence to Pharyngeal Cells Compared with Associated Streptococcal Control

| Functional Groupa | Total | Increased Expression | Decreased Expression |
|-------------------|-------|----------------------|----------------------|
| Metabolism: carbohydrate | 13    | 9                    | 4                    |
| Nucleotide        | 4     | 0                    | 4                    |
| Lipid             | 7     | 3                    | 4                    |
| Energy production/conversion | 8    | 5                    | 3                    |
| Inorganic ion transport | 1    | 0                    | 1                    |
| Coenzyme          | 5     | 5                    | 0                    |
| DNA replication and repair | 1    | 1                    | 0                    |
| Cell division     | 1     | 1                    | 0                    |
| Cell envelope/peptidoglycan biosynthesis | 2    | 2                    | 0                    |
| Signal transduction | 3   | 3                    | 0                    |
| Transcription     | 4     | 4                    | 0                    |
| Translation       | 3     | 1                    | 2                    |
| Posttranslational modification | 2   | 1                    | 1                    |
| Virulence (total) | 4     | 2                    | 2                    |
| Chromosomal       | 3     | 1                    | 2                    |
| Phage             | 1     | 1                    | 0                    |
| Chromosomal unknown function | 9   | 7                    | 2                    |
| Phage unknown function | 10  | 0                    | 10                   |
| Predicted function only | 2   | 1                    | 1                    |
| Total             | 79    | 45                   | 34                   |

Statistical significance calculated as described in Methods. Genes whose expression ratios had Bayesian P<0.05 were considered to be undergoing a statistically significant fold change in expression during adherence. See Table S1 for log2-fold change and P values for each gene calculated from four biological replicates.

Differential Expression of Genes from Diverse Functional Categories

We identified a number of genes encoding proteins involved in housekeeping processes (such as carbohydrate and coenzyme metabolism) that were differentially expressed, indicating a shift in metabolic processes due to host cell adherence (Tables 1 and 2). For example, genes encoding proteins involved in folate biosynthesis [40] were upregulated, suggesting that certain cofactors that may be necessary during adherence were unavailable. Also upregulated were genes encoding subunits of the F0F1 ATPase [41] (discussed in more detail later), which may indicate an acid stress response to maintain cytoplasmic pH or a need to generate ATP in response to increased energy requirements.

We also identified the adherence-mediated upregulation of four transcriptional regulators (Table 1), suggestive of an adaptive response to host cell contact that is dynamic and complex. For example, RopB (encoded by spy2042), a member of the Rgg family of response regulators, interacts with a number of regulatory networks throughout the streptococcal genome (e.g., mga, csrRS, sagA, and fasBCA), affecting the transcription of numerous proteins, virulence factors, and two-component regulatory systems [42,43]. Although the delineation of genes influenced by RopB (or any identified transcriptional regulator) is beyond the scope of this study, our initial analysis did identify the upregulation of a two-component regulatory system, encoded by spy1236–1237. The functions of these particular loci are not yet known, and their adherence-mediated upregulation represents new targets in the study of regulators that function during host cell contact.

Figure 1. Confirmation of speH Deletion Mutant and Pharyngeal Cell Adherence Assay

(A) Results of RT-PCR and PCR analyses of total RNA preparations isolated from mid-log (OD = 0.4) and stationary phase (OD = 1) cultures of the ΔspeH deletion mutant (ΔL and ΔS, respectively), and stationary phase cultures of the SF370 parental strain (P). RNA was reverse-transcribed as described in Methods. To assess genomic DNA contamination, control reactions containing Taq DNA polymerase instead of reverse transcriptase were included. cDNA products were separated on a 1% agarose gel and visualized by ethidium bromide staining. Lanes containing products from either the RT-PCR or PCR analysis are designated at the bottom of the panel. Lanes labeled MW contain 1 kb Plus DNA ladder (1 μg; Invitrogen).

(B) Results of the pharyngeal cell adherence assay (detailed in Methods), comparing parental strain SF370 with the deletion mutant SF370ΔspeH (abbreviated ΔspeH). Adherent streptococci are reported as the percentage of total number of streptococci added as inoculum to pharyngeal cell monolayers. Statistical significance (reported as p value) was determined by Student’s t-test.

doi:10.1371/journal.pcbi.0030132.g001

Neighbor Clustering

Our initial analysis revealed the differential expression of a wide range of functionally diverse genes and provided insight into the adaptive response of streptococci to host cell contact. However, despite a rigorous statistical approach, this analysis, like many previous microarray studies, identified the differential expression of a large number of unknown genes (n = 21) and a number of incomplete biological pathways (e.g., F0F1 ATPase [41] and folate biosynthesis...
by failing to detect the differential expression of a number of known gene pathway members (Table 1). To overcome these limitations and to extract more functional information from the array dataset (including more complete biological pathways), we developed the neighbor clustering algorithms to combine the physical position of genes on the streptococcal chromosome with gene expression data. Neighbor clustering was designed to identify expanded groupings of potentially related genes from our array data by incorporating two reliable predictors of genes that share common function or regulation, namely physical proximity and similar expression profiles [5,6].

We implemented this approach by developing an algorithm with dynamic windowing (GenomeCrawler) that sequentially stepped through the microarray data and identified clusters of adjacent genes exhibiting similar fold changes in expression. Because the genome contains many possible clusters, we restricted the algorithm's search space to identify only spatially related clusters. GenomeCrawler applied a separate permutation algorithm, using the sum of each gene's t-statistics to calculate adjusted P values ($P_k$) for each cluster, which corresponded to the probability of assembling a cluster by chance. Significance was assigned to clusters with $P_k < 0.05$, and the resulting groupings are listed in Table 3. Because individual genes could be members of many different significant clusters, GenomeCrawler then applied a distinct permutation algorithm to calculate the probability ($P_C$) that a gene was clustered coincidentally. Calculation of $P_C$ values relies on Bayes' Theorem, in which the probability of a gene’s $\log_2$-fold change ($P_F$ value) is combined with the cluster probability itself ($P_k$ value). We stress that $P_C$ reflects the significance of a gene based on its cluster context rather than a recapitulation of $P_F$. This ensures a strong dependency between $P_F$ and $P_C$, preventing a gene with a relatively low log$_2$-fold change from being scored as significant simply because it is clustered with a gene with a highly significant $P_F$ value. Finally, GenomeCrawler calculated the overall significance of differentially expressed genes ($P_E$ values) by integrating differential expression probabilities ($P_k$) and cluster context probabilities ($P_C$). We developed a plotting application (GenomeSpyer) that represents the chromosome as a linear molecule to visualize GenomeCrawler output, with genes displayed on the x-axis and their log$_2$-fold change magnitudes on the y-axis. Applications and all datasets are available for download at http://www.rockefeller.edu/vaf/streparray.php.

We visually inspected the resulting clusters and disqualified those that violated our neighbor cluster definition (see Methods for details). All output prior to cluster disqualification is included for comparison (see Table S4). Of the 309 qualifying clusters (Table S5), 197 (63.8%) were composed entirely of known, functionally defined genes; however, 26 (13%) of these were incorrectly assembled, as they contained known genes that are functionally unrelated. Because we did not incorporate functional annotations of genes into the algorithms (i.e., to keep the analysis "blind"), we anticipated the possibility that some groupings could be assembled incorrectly without the statistical framework for assigning clusters. Of the remaining 283 (91.6%) groupings, a number of differently sized clusters contained the same gene (Table S5). We report such clusters first by highest significance (lowest $P_k$ value), then by largest number of genes. Thus, if clusters containing a particular gene were of equal significance, we report the cluster with the most gene members. This method identified 47 significant clusters containing 173 differentially expressed genes (listed in Table 3 and visualized in Figures 2 and S2–S4), a considerably larger group than could have been compiled using only the initial 79 significant genes. A total of 56 of the original 79 significant genes became components of significant clusters, whereas 23 remained unclustered.

We subdivided all clusters into three qualitative types based on the functional annotation of gene members. We present examples of Type I and II clusters: Type I clusters (n = 25) contained only functionally defined and functionally related genes (as reported in published studies), such as biological pathways components (Figures 2B and S2); Type II clusters (n = 20) included both known and unknown genes (Figures 2C and S3). Type III clusters (n = 2) were composed entirely of unknown genes (Figures 2D and S4), and are not discussed in detail.

**Type I Clusters: Intact Metabolic Pathways and Multimeric Proteins**

We measured the performance of our algorithm by examining whether it identified gene groupings known to be functionally related (Type I clusters). Only four (16%) of 25 Type I clusters (spy0080–0081, spy1236–1237, spy1707–1711, spy2041–2042) could have been identified in entirety by significance analysis because all clustered genes exhibited significant differential expression ($P_F < 0.05$). A total of 11 (52.4%) of the remaining 21 clusters would not have been identified in their entirety without GenomeCrawler because we initially identified significant fold-changes in only a subset of genes necessary to encode particular pathways or loci; this is intuitively unreasonable if all genes are essential for functionality. GenomeCrawler expanded these clusters to contain more genes that encode intact loci (Table 3).

For example, we initially identified (Table 1) the significant upregulation of three of the five known gene members of the folate biosynthetic pathway [40] (spy1096–1100), but GenomeCrawler identified a significant cluster containing all five genes (Table 3 and Figure 2B). We obtained a similar result for the eight-gene operon encoding the F0F1-type proton translocating ATPase [41] (spy0754–0761). The initial significance analysis identified only four $atp$ genes (Table 1), but neighbor clustering identified a significant cluster containing all eight genes necessary to encode a functional ATPase (Table 3).

Each of the 11 neighbor clusters that could have been only partially identified by our initial analysis alone gained gene members after application of the algorithms and became more complete sets of functionally related genes than initially identified (Table 3). These clusters encompass various metabolic processes, including purine biosynthesis (spy0025–0028), lactose metabolism (spy1916–1923), fatty acid biosynthesis (spy1743–1747), lipoteichoic acid synthesis (spy1308–1312), and sugar phosphotransferase transport (spy1058–1060) [14], suggesting that specific changes occur in the streptococcal metabolic program as the bacteria adhere to human pharyngeal cells in vitro.

Notably, the remaining ten Type I clusters were composed entirely of genes that individually were not significant; however, after applying our algorithms, the combined
Table 3. Qualifying Neighbor Clusters Identified in the *S. pyogenes* SF370 Genome by GenomeCrawler Analysis of Adherent Versus Associated Microarray Data

| Cluster Type | Start–Stop* | Function* | $P_F$ Value$^b$ | Cluster Members* | $P_F$ Value$^c$ | $P_E$ Value$^d$ | Expression* |
|--------------|-------------|-----------|----------------|-----------------|----------------|----------------|-------------|
| Type I       | 0025–0028   | Purine biosynthesis | 0.0005 | 0025 | 0.028 | 0.0011 | Decreased |
|              |             |           |                | 0026 (purF) | 0.093 | 0.0037 |            |
|              |             |           |                | 0027 (purM) | 0.015 | 0.0005 |            |
|              |             |           |                | 0028 (purN) | 0.124 | 0.0077 |            |
|              | 0032–0034   | Purine biosynthesis | 0.0045 | 0032 (purD) | 0.246 | 0.0215 | Decreased |
|              |             |           |                | 0033 (purE) | 0.134 | 0.0057 |            |
|              |             |           |                | 0034 (purX) | 0.07* | 0.0038 |            |
|              | 0514–0516   | Carbohydrate metabolism | 0.0135 | 0514 (tcpA) | 0.405 | 0.0676 | Increased |
|              |             |           |                | 0515 | 0.324 | 0.0055 |            |
|              | 0738–0746   | Virulence | 0.028 | 0738 (sagA) | 0.534 | 0.0895 | Decreased |
|              |             |           |                | 0739 (sagB) | 1* | 1 |            |
|              |             |           |                | 0740 (sagC) | 1* | 1 |            |
|              |             |           |                | 0741 | 0.638 | 0.0766 |            |
|              |             |           |                | 0742 | 0.999 | 0.1272 |            |
|              |             |           |                | 0743 | 0.361 | 0.027 |            |
|              |             |           |                | 0744 | 0.066 | 0.0035 |            |
|              |             |           |                | 0745 (sagH) | 0.403 | 0.0435 |            |
|              |             |           |                | 0746 | 0.859 | 0.0614 |            |
|              | 0754–0761   | Energy production | 0.0005 | 0754 (atpE) | 0.769 | 0.0293 | Increased |
|              |             |           |                | 0755 (atpD) | 0.415 | 0.023 |            |
|              |             |           |                | 0756 (atpF) | 0.057 | 0.0018 |            |
|              |             |           |                | 0757 (atpH) | 0.006 | <0.0005 |            |
|              |             |           |                | 0758 (atpA) | 0.155 | 0.0072 |            |
|              |             |           |                | 0759 (atpG) | 0.001 | <0.0005 |            |
|              |             |           |                | 0760 (atpD) | 0.006 | 0.0001 |            |
|              |             |           |                | 0761 (atpC) | 0.001 | <0.0005 |            |
|              | 0776–0777   | DNA replication/repair | 0.048 | 0776 (rexB) | 0.196 | 0.0287 | Increased |
|              |             |           |                | 0777 (rexA) | 0.36* | 0.0521 |            |
|              | 0804–0806   | Translation | 0.0285 | 0804 (unC) | 0.047 | 0.007 | Decreased |
|              |             |           |                | 0805 (tpmH) | 0.608 | 0.0942 |            |
|              |             |           |                | 0806 (tpIF) | 0.656 | 0.1194 |            |
|              | 0880–0881   | Lipid metabolism | 0.0055 | 0880 (mvaS1) | 0.006 | 0.0001 | Increased |
|              |             |           |                | 0881 (mvaS2) | 0.013 | 0.0002 |            |
|              | 1007–1008   | Virulence | 0.0195 | 1007 (sagA) | 0.488 | 0.0739 | Increased |
|              |             |           |                | 1008 (sagB) | 0.237 | 0.0375 |            |
|              |             |           |                | 1009 (sagC) | 0.059 | 0.0011 |            |
|              |             |           |                | 1010 (sagD) | 0.286 | 0.0374 |            |
|              | 1058–1060   | Carbohydrate metabolism | 0.0035 | 1058 | 0.009 | 0.0008 | Decreased |
|              |             |           |                | 1059 | 0.002 | 0.0001 |            |
|              |             |           |                | 1060 | 0.393 | 0.0177 |            |
|              | 1096–1100   | Folate biosynthesis | 0.0005 | 1096 (folC1) | 0.051 | 0.0025 | Increased |
|              |             |           |                | 1097 (folD) | 0.001 | <0.0005 |            |
|              |             |           |                | 1098 (folP) | 0.009 | 0.0003 |            |
|              |             |           |                | 1099 (folQ) | 0.009 | 0.0002 |            |
|              |             |           |                | 1100 (folR) | 0.347 | 0.0311 |            |
|              | 1236–1237   | Signal transduction | 0.0005 | 1236 | 0.001 | <0.0005 | Increased |
|              |             |           |                | 1237 | 0.004 | 0.0001 |            |
|              | 1250–1251   | Nucleotide transfer | 0.043 | 1250 (nreA) | 0.09* | 0.0156 | Increased |
|              |             |           |                | 1251 (truB) | 0.684 | 0.1385 |            |
|              | 1294–1296   | Carbohydrate metabolism | 0.02 | 1294 | 0.231 | 0.0359 | Decreased |
|              |             |           |                | 1295 (malF) | 0.527 | 0.0967 |            |
|              |             |           |                | 1296 (malG) | 0.149 | 0.0191 |            |
|              | 1308–1312   | Lipoteichoic acid synthesis | 0.0015 | 1308 | 0.143 | 0.0044 | Increased |
|              |             |           |                | 1309 (dltD) | 0.199 | 0.022 |            |
|              |             |           |                | 1310 (dltC) | 0.051 | 0.0049 |            |
|              |             |           |                | 1311 (dltB) | 0.149 | 0.0085 |            |
|              |             |           |                | 1312 (dltA) | 0.35* | 0.0456 |            |
|              | 1599–1600   | Glucosidase/hyaluronidase | 0.049 | 1599 | 0.673 | 0.673 | Decreased |
|              |             |           |                | 1600 | 0.259 | 0.0416 |            |
|              | 1621–1622   | Signal transduction | 0.0105 | 1621 | 0.073 | 0.0115 | Increased |
|              |             |           |                | 1622 | 0.032 | 0.0044 |            |
|              | 1707–1711   | Carbohydrate metabolism | 0.0005 | 1707 (lucB1) | 0.001 | <0.0005 | Increased |
|              |             |           |                | 1708 (lucA1) | 0.001 | <0.0005 |            |
|              |             |           |                | 1709 | 0.001 | <0.0005 |            |
|              |             |           |                | 1710 | 0.001 | <0.0005 |            |
|              |             |           |                | 1711 | 0.006 | 0.0001 |            |
Table 3. Continued.

| Cluster Type | Start–Stopa | Functiona | PK Valueb | Cluster Membersa | PK Valuec | PK Valued | Expression* |
|--------------|------------|-----------|-----------|-----------------|-----------|----------|-------------|
| 1743–1747    | Lipid metabolism | 0.003     | 1743 (accA) | 0.041          | 0.0021    | Decreased |
|              |            |           | 1744 (accC) | 0.758*         | 0.0529    |           |
|              |            |           | 1745 (accC) | 0.018          | 0.0007    |           |
|              |            |           | 1746 (fabZ) | 0.558*         | 0.0385    |           |
|              |            |           | 1747 (acdB) | 0.201*         | 0.01      |           |
| 1753–1758    | Lipid metabolism | 0.006     | 1753 (accP) | 0.632*         | 0.1018    | Decreased |
|              |            |           | 1754 (fabH) | 0.006          | 0.0005    |           |
|              |            |           | 1755        | 1*             | 0.0385    |           |
|              |            |           | 1758 (phaB) | 0.001          | 0.0005    |           |
| 1916–1923    | Carbohydrate metabolism | 0.0005   | 1916 (lacG) | 0.037          | 0.0001    | Increased |
|              |            |           | 1917 (lacE) | 0.438*         | 0.0165    |           |
|              |            |           | 1918 (lacF) | 0.564*         | 0.0392    |           |
|              |            |           | 1919 (lacD2) | 0.37*         | 0.0126    |           |
|              |            |           | 1921 (lacC2) | 0.08*         | 0.0023    |           |
|              |            |           | 1922 (lacE2) | 0.555*         | 0.0254    |           |
| 2041–2042    | Transcription | 0.0095   | 2041        | 0.03           | 0.0006    | Increased |
|              |            |           | 2042 (ropB) | 0.032          | 0.0007    |           |
| 2047–2052    | Carbohydrate metabolism | 0.0015  | 2047 (gldA) | 0.012         | 0.0002    | Decreased |
|              |            |           | 2048 (mipB) | 0.001          | <0.005    |           |
|              |            |           | 2049 (pflD) | 0.571*         | 0.0405    |           |
|              |            |           | 2050        | 0.859*         | 0.037*    |           |
|              |            |           | 2051        | 0.839*         | 0.0353    |           |
|              |            |           | 2052        | 0.649*         | 0.026     |           |
| Type II      |            |           | 0039–0040   | Signal transduction | 0.041 | 0039     | 0.158* 0.0112 | Increased |
|              |            |           |              | 0040           | 0.456* 0.0537 |           |
|              |            |           | 0127–0130   | Virulence/surface proteins | 0.0005   | 0127     | 0.001 <0.0005 | Increased |
|              |            |           |              | 0128           | 0.006     | 0.0001    |
|              |            |           |              | 0129           | 0.001     | <0.0005   |
|              |            |           |              | 0130           | 0.001     | <0.0005   |
|              |            |           | 0238–0239   | Translation | 0.046   | 0238     | 0.137* 0.0166 | Increased |
|              |            |           |              | 0357–0362      | Membrane proteins | 0.038 | 0357     | 0.906* 0.965 | Increased |
|              |            |           |              | 0358           | 0.965     | 0.965     |
|              |            |           |              | 0359           | 0.424*    | 0.0617    |
|              |            |           |              | 0361 (ggr)     | 0.891*    | 0.891     |
|              |            |           |              | 0362           | 0.25*     | 0.0338    |
|              |            |           | 0421–0422   | Translation | 0.048   | 0421     | 0.03 0.0029 | Decreased |
|              |            |           |              | 0422           | 0.935* 0.935 |           |
|              |            |           | 0440–0441   | Fatty acid synthesis | 0.0495   | 0440     | 0.174* 0.0323 | Increased |
|              |            |           |              | 0441           | 0.681*    | 0.681     |
|              |            |           | 0472–0477   | Phospholipid synthesis | 0.0025   | 0472     | 0.051* 0.0068 | Decreased |
|              |            |           |              | 0473           | 0.748*    | 0.784     |
|              |            |           |              | 0475 (dgk)     | 0.021     | 0.0022    |
|              |            |           |              | 0476 (ara)     | 0.913*    | 0.913     |
|              |            |           |              | 0477           | 0.444*    | 0.0588    |
|              |            |           | 0504–0505   | Translation/energy production | 0.048 | 0504     | 0.611* 0.1165 | Decreased |
|              |            |           |              | 0505           | 0.086*    | 0.0119    |
|              |            |           | 0642–0645   | Cell division | 0.0225   | 0642     | 0.654* 0.1114 | Increased |
|              |            |           |              | 0643           | 0.245*    | 0.0246    |
|              |            |           |              | 0644           | 0.515*    | 0.0659    |
|              |            |           |              | 0645           | 0.707*    | 0.128     |
|              |            |           | 0792–0797   | Cell envelope biogenesis | 0.0275   | 0792     | 0.763* 0.763 | Increased |
|              |            |           |              | 0793           | 0.584*    | 0.1133    |
|              |            |           |              | 0794           | 0.769*    | 0.769     |
|              |            |           |              | 0796           | 0.725*    | 0.725     |
|              |            |           |              | 0797           | 0.54*     | 0.1014    |
|              |            |           | 0818–0822   | Translation | 0.0435   | 0818     | 0.874* 0.874 | Decreased |
|              |            |           |              | 0819           | 0.761*    | 0.1464    |
|              |            |           |              | 0821           | 0.605*    | 0.1229    |
|              |            |           | 0840–0841   | Translation | 0.0145   | 0840     | 0.018 0.0029 | Decreased |
|              |            |           |              | 0841           | 0.142*    | 0.0192    |
|              |            |           | 0873–0875   | Signal transduction | 0.044   | 0873     | 0.445* 0.0254 | Increased |
|              |            |           |              | 0874           | 0.733*    | 0.0442    |
|              |            |           |              | 0875           | 0.7*      | 0.0588    |
|              |            |           | 0882–0886   | Folate/DNA synthesis | 0.0115   | 0882     | 0.235* 0.01 | Increased |
|              |            |           |              | 0883           | 0.89*     | 0.0342    |
|              |            |           |              | 0884           | 0.146*    | 0.002     |
|              |            |           |              | 0885           | 0.245*    | 0.0053    |
Table 3. Continued.

| Cluster Type | Start–Stopa | Functiona | \(P_E\) Valueb | Cluster Membersa | \(P_E\) Valuec | \(P_F\) Valued | Expressione |
|--------------|-------------|------------|----------------|----------------|----------------|--------------|-------------|
| 0900–0901    | Pyrimidine biosynthesis | 0.015 | 0900 (pyrF) | 0.015* | 0.017 | Decreased |
|              |             |           | 0901 (pyrE) | 0.009 | 0.001 |
| 1212–1220    | Folate biosynthesis | 0.0005 | 1212 | 0.109* | 0.0067 | Increased |
|              |             |           | 1213 (ftsI) | 0.24* | 0.0041 |
|              |             |           | 1214 (lpIA) | 0.006 | 0.0004 |
|              |             |           | 1215 | 0.021 | 0.0012 |
|              |             |           | 1216 | 0.05 | 0.0033 |
|              |             |           | 1217 | 1* | 1 |
|              |             |           | 1218 | 0.576* | 0.0493 |
|              |             |           | 1219 | 0.702* | 0.1021 |
|              |             |           | 1220 | 0.99* | 0.07 |
| 1277–1281    | Amino acid metabolism | 0.0015 | 1277 | 0.001 | <0.0005 | Decreased |
|              |             |           | 1280 (glmS) | 0.094* | 0.019 |
|              |             |           | 1281 (piC) | 0.844* | 0.237 |
| 1701–1704    | Carbohydrate metabolism | 0.0005 | 1701 | 0.006 | 0.0001 | Increased |
|              |             |           | 1704 (lacD,1) | 0.001 | <0.0005 |
| 1719–1725    | Protein biosynthesis | 0.001 | 1719 (rbfA) | 0.079* | 0.0045 | Decreased |
|              |             |           | 1721 (nfb) | 0.13* | 0.0083 |
|              |             |           | 1722 | 0.367* | 0.0187 |
|              |             |           | 1723 | 0.607* | 0.0912 |
|              |             |           | 1724 | 0.59* | 0.0368 |
|              |             |           | 1725 | 0.18* | 0.0134 |
| 2059–2060    | Membrane protein/translation | 0.0355 | 2059 (pbp2A) | 0.006 | 0.0081 | Increased |
|              |             |           | 2060 | 0.908* | 0.1316 |
| 2105–2107    | Nucleotide metabolism/ DNA replication | 0.0015 | 2105 (nrpG) | 0.018 | 0.0013 | Decreased |
|              |             |           | 2106 | 0.058* | 0.0026 |
|              |             |           | 2107 | 0.004 | 0.0002 |
| Type III     | Phage unknown | 0.0005 | 0956 | 0.001 | <0.0005 | Decreased |
|              |             |           | 0957 | 1* | 1 |
|              |             |           | 0958 | 0.001 | <0.0005 |
|              |             |           | 0961 | 0.006 | 0.0001 | Decreased |
|              |             |           | 0962 | 0.001 | <0.0005 |
|              |             |           | 0963 | 0.001 | <0.0005 |
|              |             |           | 0965 | 0.001 | <0.0005 |

aGenes representing the first and last genes in the cluster (Spy start–stop) and genes included within clusters are indicated by Spy numbers followed by common names in parentheses (when available), as designated in the annotated SF370 genome. Gene function designations are also from annotated genome. For type II clusters, the function of the cluster is based on the known, annotated functions of all gene members. For type I clusters, only putative function can be assigned for the entire cluster, and is based solely on the function of the known members of the cluster.

bStatistically significant clusters (\(P_E\) value < 0.05) are distinguished by qualitative type (Type I, II, or III).

cFold-change \(P\) values \((P_F)\), calculated by Cyber-T and WY permutation algorithm as described in Methods, are listed for each member of a given cluster (asterisks indicate individual genes not scored as statistically significant, as \(P_F > 0.05\)).

dExpression \(P\) values \((P_F)\) as described in Methods.

eChange in expression ratios between streptococci adhered to pharyngeal cells compared with control streptococci that are associated with, but not adhered to, pharyngeal cells.

doi:10.1371/journal.pcbi.0030132.t003

The contribution of each gene resulted in a significant cluster. For example, the nine-gene operon that spans genes spy0738–0746 encodes streptolysin S, a potent cytolytic toxin that promotes internalization and host tissue dissemination [25,44]. Though the differential expression of the individual genes was not significant following our initial statistical analysis, GenomeCrawler identified a significant downregulated cluster containing all nine genes (Table 3). Adherence-induced downregulation of streptolysin S is consistent with its previously determined role in host cell internalization [25]; however, without neighbor clustering, expression of this operon was not evident immediately.

Although individual gene members of Type I clusters may not be statistically significant as a result of technical variability within experiments [17], the genetic structure of certain Type I operons may provide an alternative explanation. For example, the streptolysin operon encodes an internal terminator downstream of the sagA gene (the first gene in the operon), which modulates the abundance of particular mRNA species (e.g., sagA mRNA versus the polycistronic message for all nine genes) under different environmental conditions [45]. If transcription is internally disrupted by such a terminator, the abundance of the sagA transcript may be much greater than the polycistronic message; such disproportionate transcript levels would affect log-fold change values and impact the statistical significance of individual genes within these types of clusters. Thus, in addition to helping resolve clusters that would not be easily recognized because of experimental technical variability, the neighbor clustering method may help to resolve operons with such internal terminators and regulators.

These results demonstrate that neighbor clustering effectively reconstructed a number of complete pathways and loci from processed array data. Importantly, because functional
gene data are not incorporated into its algorithms, GenomeCrawler is not biased toward identifying “expected” clusters. Curating the dataset following its application may make the algorithms less user-friendly; however, the elimination of such bias is essential for this type of analysis.

Type II Clusters

Based on the Type I cluster results, we speculated that genes contained in Type II clusters might be related by function or regulation. Type II groupings contain a combination of both known and unknown gene members and could provide preliminary clues about the function of unknown genes within a particular cluster by associating their expression with neighboring genes of known and defined function. Alternatively, co-expression of genes results from common regulation, and Type II associations may suggest shared regulatory mechanisms for clustered genes. We note, however, that despite the statistical framework with which groupings are assigned, experimental evidence is necessary to confirm functional or regulatory relatedness. We do not suggest simply assigning either based on cluster membership; rather, cluster associations may provide some preliminary functional or regulatory clues for gene members.
A total of 18 (90%) of 20 Type II clusters (Table 3 and Figure S3) may not have been identified without neighbor clustering; eight (44.4%) of 18 gained additional gene members; the remaining ten comprised genes that demonstrated significant differential expression only after applying GenomeCrawler. Only two clusters (spy0127–0130 and spy1701–1704) could have been identified without neighbor clustering; however, a number of these genes were initially annotated as hypothetical proteins, so a potential relationship between the gene members may not have been readily apparent.

The upregulated spy0127–0130 cluster is part of a larger genomic region known as FCT (for fibronectin- and collagen-binding proteins and T antigen-encoding loci), which spans spy0123–0136 in the SF370 genome and encodes surface proteins and transcriptional regulators [46]. A search of both the PFAM database [47] (http://pfam.wustl.edu) and sortase database (http://www.doe-mbi.ucla.edu/Services/Sortase) predicted that spy0129 encodes a sortase enzyme, which are transpeptidases that cleave protein substrates at conserved C-terminal motifs (often LPXTG) and then anchor these proteins to the bacterial cell wall [48,49]. Recently, it was reported that the four genes spanning spy0127–0130 encode, and are responsible for, the formation of surface-localized, trypsin-resistant pili that induce protective immunity against a lethal dose of group A streptococci in a mouse model of infection [50]. This same report provided the first experimental evidence supporting the sortase prediction, indicating that the gene product of spy0129 is responsible for the cell-wall sorting of the proteins encoded by both spy0128 (annotated as a Cpa homolog [50]) and spy0130 (annotated as a protein F homolog [14]). Furthermore, the spy0128-encoded protein is the structural backbone of the pili, and the gene product of spy0130 may be involved in stabilizing the structure [36]. Together with the identification of this cluster by GenomeCrawler, these results prompted us to study this cluster and the contributions of the gene products to pharyngeal cell adherence.

We determined experimentally that cluster spy0127–0130 is an operon, verifying both related function and regulation of the gene members. Reverse transcription of SF370 RNA, with primer combinations that spanned all four genes, produced cDNA fragments of sizes that could only result from a polycistronic mRNA template (Figure 3). In silico sequence inspection identified a single putative promoter sequence upstream of spy0127 (see Table S6). Although GenomeCrawler is not an operon-identifying algorithm, these results show that it could (1) identify this commonly regulated gene cluster and (2) define the cluster boundaries, excluding other proximate genes, such as an additional sortase-encoding gene, spy0135.

Allelic Replacement of spy0129

We created a spy0129 deletion mutant in strain SF370 (SF370Δspy0129) to determine if genes contained within the spy0127–0130 cluster were directly involved in adherence to pharyngeal cells. We posited that a deletion in the spy0129 sortase gene may have the greatest overall effect on the production and processing of the gene products of this cluster, since both the spy0128 and spy0130 gene products do not localize to the cell-wall surface in the absence of the sortase enzyme [36]. Allelic replacement created two putative deletion mutants; however, RT-PCR analysis (Figure 4A) revealed that only one such clone (SF370Δspy0129/2) was a true knock-out for the spy0129 gene and useful for further study. Because the gene cluster is also an operon, expression of the downstream gene spy0130, encoding the protein F homolog/plus protein, was also eliminated in this mutant (Figure 4A). In vitro pharyngeal cell adherence assays revealed that the SF370Δspy0129 mutant was approximately 66% less adherent than the parental control strain, SF370 (Figure 4B; p = 0.03 as determined by the Student’s t-test). These results suggest that either the spy0130 gene product is involved directly in adherence, or that due to the elimination of the sortase, the pili, which may function in their entirety as adhesins, were not assembled on the surface of the mutant. Because the spy0129 gene product is not expected to be found on the streptococcal surface (i.e., it lacks a cell-wall anchoring motif), it is not likely to be involved directly in adherence. We are working to produce an in-frame deletion of spy0128 and a spy0130 single knock-out mutant to delineate the contribution of each individual clustered gene product to adherence.

These results show that neighbor clustering is able to identify biologically relevant gene clusters. This attribute may be particularly important for datasets in which the relationship between clustered genes is not obvious, and may facilitate the organization of larger datasets into more manageable packages.

Additional Type II Cluster Example

Another cluster, spy1725–1719, contained six genes that together (though not individually) exhibited significant downregulation. The genes spy1724, spy1722, spy1721, and spy1719 share transcriptional order and predicted function with homologs in the nusA-infB protein biosynthesis operon of Bacillus subtilis and Escherichia coli [51]. We examined the spy1725 and spy1723 gene products (annotated as hypothetical proteins [14]) for similarities with known proteins that might indicate a role for these gene products in protein biosynthesis. BlastP analysis aligned the spy1725 gene product, which has homologs in all sequenced streptococcal genomes, with the SP14.3 protein from S. pneumoniae [52] (80% sequence similarity; 67% identity). Based on structural characterization, SP14.3 is a predicted RNA-binding protein. The spy1723 gene product has similar domain structure to the YksR protein of S. pneumoniae, an RNA-binding protein implicated in transcription termination [53]. These results indicate that both genes likely encode RNA-binding proteins, in agreement with their functionally defined cluster members. Although domain and homology searches yielded the functional predictions, their membership within a protein biosynthetic cluster provided the initial indication of common function or regulation.

Neighbor Clustering and Operons

Although neighbor clustering is not an operon-predicting method, we wanted to identify additional putative operons among the groupings since neighbor clusters by definition share certain operon characteristics (tandemly arranged genes, separated by <300 bp, with similar expression patterns). Although operon-modeling methods exist [54,55], we inspected clusters in silico for upstream regulatory elements and identified 17 candidates, including clusters such as streptolysin S that have been previously confirmed as...
operons [56]; the spy0127–0130 grouping, which was confirmed as an operon in this study; and others that have yet to be verified (Table S6). Experimental confirmation of each candidate is beyond the scope of this study, but Northern blot and RT-PCR analyses could provide such information.

Analysis of Previously Published Array Data
We applied the statistical analysis and the GenomeCrawler algorithms to data from a recently published streptococcal microarray study that is relevant for comparison to our own data (same streptococcal strain, similar array platform) [57]. In this study, the transcriptomes of S. pyogenes strain SF370 and an isogenic mutant deficient for the Mga regulon were compared during exponential growth in culture broth. The Mga regulator is a growth-phase mediator of a number of surface-exposed molecules and secreted proteins involved in colonization and immune evasion during infection [58]. Although the authors of that study did not provide a statistical analysis of their data, we compared the published results for the magnitude and direction of fold-changes for each gene reported in this study with those obtained from our initial significance analysis of this dataset (presented as Table S7). A total of 256 genes reported in this study were also detected by our analysis, and the magnitude and log2-fold change were found to be in agreement for 81% of the genes.

Figure 3. Neighbor Cluster spy0127–0130
(A) Schematic representation of the spy0127–0130 cluster. Gene size (bp) is indicated under the gene name, and the position of primers for the reverse transcription of streptococcal mRNA are indicated by colored arrows (forward primers, F; reverse primers, R). Eight different combinations of primers were used, and the expected sizes of the eight resulting cDNAs that would be produced if genes compose an operon are indicated in the numbered, colored boxes. cDNA box colors correspond to colors of the primer pairs that would generate each fragment.
(B) A 1% agarose gel of the cDNA fragments amplified from mRNA with each primer combination. Lane numbers correspond to the numbering of the predicted cDNAs from (A). First lane of gel (MW) contains 1 kb Plus DNA ladder, and the sizes of the relevant DNA fragments in the ladder are indicated (bp).

doi:10.1371/journal.pcbi.0030132.g003
group of genes (107 versus four) contained within 36 statistically significant clusters ($P_k < 0.05$; Table S9). These groupings included clusters of genes that have been shown previously in streptococci to be functionally related, indicating that the algorithms were performing as expected. Two of the identified upregulated clusters (spy0209–2010 and spy2039–2040) encoding the well-studied virulence factors, C5a peptidase and SpeB, respectively, showed consistently large log$_2$-fold changes of the genes across replicates [57]. GenomeCrawler confirmed these results by identifying both groupings as statistically significant neighbor clusters.

GenomeCrawler also identified a number of clusters that contained genes known to share common function or regulation; however, they were not as apparent in the dataset without its application. For example, the algorithm identified a significant neighbor cluster spanning spy0711–0712. This grouping encodes two known virulence factors, pyrogenic exotoxin SpeC and the MF2 DNase, previously shown to be commonly regulated as an operon [11]. The algorithm also identified other neighbor clusters containing genes known to be functionally related, including spy0098–0100 (encoding the $\beta$ and $\beta'$ subunits of DNA-dependent RNA polymerase), spy2159–2160 (encoding the 50S ribosomal subunit proteins L32 and L33), and spy0741–0746 (six of the nine streptolysin S–encoding genes) [14].

Although the analysis of this previously published dataset did not reveal as many intact biological pathways as were identified from the pharyngeal cell adherence data, the inclusion of more replicates in the analysis to increase statistical power could resolve such loci. However, these results provided further supporting evidence that the GenomeCrawler algorithms can identify (1) a larger group of genes than a rigorous statistical analysis alone and (2) biologically relevant groupings in other microarray datasets, even if they contain fewer replicates than presented in our study.

Concluding Remarks

Although GenomeCrawler improves bacterial array analyses, it has limitations: it cannot identify regulons comprising genes dispersed throughout the genome by virtue of its design, it does not specifically interrogate single-gene operons, and it only applies to genomes with available and accurate experimental information (expression data and gene annotations). We recognize that incorporating intergenic distance and transcription direction into the algorithms would reduce processing time. Adding available clusters of orthologous groups (COG) information into a downstream processing step could decrease errors by minimizing clustering of unrelated genes.

Nonetheless, neighbor clustering provided a more comprehensive view of the transcriptome of group A streptococci during adherence to human pharyngeal cells, a critical step in the infection program of this organism. We found that even a rigorous statistical analysis of well-replicated microarray data produced a dataset that was somewhat limited, although certainly more informative than assigning arbitrary thresholds for significance. As described in other microarray reports, we had initially identified a number of incomplete biological pathways in which we did not detect the differential expression of a number of known pathway members.
Neighbor clustering was able to extend the results by identifying more differentially expressed genes and reconstructing more intact biological pathways.

Neighbor clustering, despite the statistical framework with which it assigns groupings, would be valuable to microarray data analysis only if it produced biologically relevant data. Although biological testing of every identified gene or cluster is unrealistic, we provided evidence, through the creation and testing of isogenic deletion mutants and through the identification of clusters of known, functionally related genes from a published streptococcal array study, that the algorithms produce results that are pertinent to the biology of streptococci. This may be of particular importance for data in which the relationship between clustered genes is not obvious, and may facilitate the organization of larger datasets into more meaningful packages. It is also possible that GenomeCrawler (in its current form) could be used to interrogate intergenic portions of the genome (such as those encoding small noncoding RNAs or sRNAs), if probes representing such regions were included on the microarray, and experimental conditions were designed to promote their differential expression. Finally, because of the common architecture of bacterial chromosomes, the neighbor clustering algorithms may be applicable to microarray datasets from other prokaryotes.

Methods

**Spotted oligonucleotide microarrays.** Sense strand oligonucleotides (primarily 55-mers), representing the 1,769 open reading frames in the genome of *S. pyogenes* strain SF370 (M1 serotype) [14] were designed and produced by Illumina (http://www.illumina.com). Oligonucleotide spots were spotted using a Biobotics Tas II 6100 arrayer (http://biobotics.org) onto Corning UltraGAPS (gamma amino propyl silane-coated) slides (Corning Life Sciences, http://www.corning.com/lifesciences), and slides were post-processed and blocked according to the manufacturer’s instructions. Each oligonucleotide was spotted four times in a well-spaced configuration to generate in-line replicates.

**Bacterial cultures.** For adherence assays, *S. pyogenes* strain SF370 (kindly provided by J. Ferretti, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, United States) was grown to late log phase (OD600 = 0.7) in Todd Hewitt broth (BD Biosciences, http://www.bdbiosciences.com) containing 0.2% yeast extract (THY; BD Biosciences). Bacterial cells were washed in 0.1 M phosphate-buffered saline (PBS; pH 7.4), resuspended in minimal essential medium (MEM; Invitrogen, http://www.invitrogen.com), and incubated at 37 °C for 1 h. Glycerol (20% vol/vol) was added, and cultures were flash frozen in liquid N2 and stored at −80 °C. To minimize culture-to-culture variability, these stock cultures were used for all subsequent adherence and association experiments.

**Pharyngeal cell association and adherence assay.** Assays on streptococcal adherence to human pharyngeal cell line Detroit 562 were performed as described previously [15] with the following modifications. Streptococcal stock cultures were pre-incubated at 37 °C for 1 h, and 2-ml aliquots (2 × 10^8 CFUs) were added to confluent monolayers of Detroit 562 cells grown in wells of six-well tissue culture plates (1 × 10^5 cells/well). Cultures were incubated for 2.5 h at 37 °C, and the monolayers were then washed with PBS to recover associated (nonadherent) streptococci. Pharyngeal cell association and adherence were performed using 0.005% trypsin-0.004% EDTA for 15 min at 37 °C to desorb adherent streptococci (90% recovery) without disrupting the eukaryotic monolayer. Trypsin treatment does not affect gene expression in adherent streptococci as measured using oligonucleotide microarray analysis.

**RNA isolation.** Recovered streptococci were washed twice in PBS and lysed with the amidase enzyme lysis [59]. Lysin was added to the bacterial samples (2 U/10^8 CFUs) and incubated for 15 min at room temperature, which in preliminary experiments was determined to be optimum for complete streptococcal lysis. RNA was isolated immediately after lysis with a modified phenol-chloroform protocol as described previously [60]. RNA was digested with DNase I (Invitrogen), and RNA quality was assessed with the Nucleic Acid Bioanalyzer 2100 (Agilent Technologies, http://www.agilent.com).

**Synthesis of cDNA and labeling.** DNase-treated streptococcal RNA (5 μg) was reverse-transcribed using the Atlas Glass Fluorescent Labeling kit (BD Biosciences Clontech, http://www.bdbiosciences.com) using Random hexamers (Invitrogen) primer. cDNA synthesis reaction that incorporated a 5′-(3-aminolevulinyl)-dUTP into the first synthesized cDNA strand. cDNAs from associated streptococci and from adherent streptococci were indirectly labeled with the N-fluorescein isothiocyanate (FITC) dye cyanine 3 (Cy3) and cyanine 5 (Cy5), respectively, as outlined in the Atlas kit. Labeled cDNA samples were purified following Atlas kit instructions.

**Microarray hybridization and image acquisition.** Four biological replicate experiments incorporating dye swaps [17] were performed to account for both biological and technical variability. Labeled cDNA samples were hybridized to the anchored hybridization buffer (Ambion, http://www.ambion.com) for 16 h at 55 °C using a GeneTAC hybridization station (Genomic Solutions, http://www.genomicsolutions.com). Slides were washed twice in 0.1 × SSC, dried, and then scanned with a ScanArray 4000 scanner (GSI Lumonics, http://www.gslumonics.com) at 10 μm per pixel resolution. The resulting images were processed using the GenePix Pro program (version 4.0; Axon Instruments, http://www.axon.com).

**Data filtering, normalization, statistical significance analysis, and selection of P-values for individual genes.** For microarray analysis, low-level processing of microarray data included probe and array quality filtering to remove probes that were saturated, that displayed a low signal-to-noise ratio, and/or that produced signal only in one dye channel. Lowess normalization [19] was performed, and robust summary statistics were applied to the standardized log2 fold-change data for each sample (Hierarchical and unbiased MAD estimator) [20]. A Bayesian-derived regularized t-test was implemented with the Cyber-T program for control of variance artifacts associated with low sample size [20–22]. Calculation of the p-value of the log2-fold change for each gene (P(t)) uses a Westfall–Young stepdown permutation test [18,19] for multiplicity adjustment in place of the Bonferroni correction typically implemented in Cyber-T. Although more computationally intensive, we chose Westfall–Young over the Bonferroni correction because: (1) Bonferroni assumes independence between tests and since genes can be regulated in conjunction with one another, we preferred to avoid the assumption of independence; (2) Westfall–Young, which is based on permutation, calculates p-values (from t-test statistics) based on the actual distribution of the data itself, and no assumption of independence is required; and (3) the power of coupling a permutation algorithm with a t-test is that one can take advantage of the sensitivity associated with a t-test, while using the distribution-free nature of a randomization test. We used the t-test statistics and P values generated in this analysis (referred to throughout the text as initial statistical significance analysis) to rank genes (endogenous reference/control gene) for downstream analysis (fold-change expression (Pf < 0.05) during adherence to pharyngeal cells compared with the associated control (Table 1). Datasets resulting from each processing step are available for download at www.rocketfeller.edu/afsl/fslpreparatory.php.

**Real-time qRT-PCR primers, probes, and plasmid standards.** We performed real-time qRT-PCR analysis (TaqMan) on 11 different genes to verify the fold-change in gene expression estimated by microarray analysis. Five of these genes exhibited statistically significant fold-changes in expression (Pf < 0.05) during adherence (two downstream genes [16] and three demonstrated decreased expression), and the remaining six selected genes were scored as statistically significant only when included in a significant neighbor cluster (P < 0.05). The list of genes, as well as the oligonucleotide primers and fluorogenic (TaqMan) probes designed using Primer Express Software (Applied Biosystems, http://www.appliedbiosystems.com) and purchased from Sigma-Genosys (http://www.sigmaaldrich.com), are provided in Table S2. Each of the 11 genes, as well as the endogenous reference/control gene, was amplified in its entirety from SF370 genomic DNA by PCR and cloned into pCR-TOPO (Invitrogen). spy0929 was chosen as control due to equivalent expression between adherent and associated SF370 mutants.

**Real-time qRT-PCR and TaqMan analysis.** We used a two-step RT-PCR procedure to reverse transcribe RNA samples from two independent sets of two biological replicates for each transcriptional assay, which were prepared as those for microarray analysis. Using SuperScript II First Strand Synthesis System for RT-PCR (Invitrogen), DNase I-treated RNA preparations (2 μg each) were separately
converted to cDNA preparations with 50 ng random hexamers (Invitrogen; 45 °C, 50 min, 20 μl reactions) according to manufacturer instructions. RNA samples were reverse-transcribed in separate reactions, and no pooling of samples occurred. Control reactions without reverse transcriptase were included to confirm that genomic DNA was not present. TaqMan analysis was performed (in duplicate) with 1.25 μl of the previous reaction mixtures (10-fold dilutions ranging from 10 copies to 10 copies) and standards corresponding to the previous reaction mixtures, prepared in triplicate and included on each reaction plate. Results were normalized with C_{v} values for the control, spv0929. We averaged data from duplicate reactions to produce a single value for each gene and log-transformed the fold difference in the number of cDNA molecules present in adherent streptococcal samples relative to associated streptococcal samples. This created a dataset of 11 paired values from RT-PCR and microarray analyses for each gene. We performed linear regression analysis and regressed qRT-PCR data on the microarray data.

RT-PCR of spv0127–0130 cluster. The sequences of forward (F) and reverse (R) primers for each of the four genes contained within the spv0127–0130 neighbor cluster are provided in Table S2. RT–PCR generation of amplicons was performed with the SuperScript III One-Step RT–PCR system with Platinum Taq DNA polymerase (Invitrogen) in reaction mixtures containing 50 ng genomic DNA (50 ng containing 0.1 μg of DNA). PCR was performed using the primer sets listed in Table S2. RT–PCR analysis of total RNA extracted (as described) and treated, purified total RNA from late-log phase cultures (OD_600nm = 0.7) of strain SF370. All remaining components were added as per manufacturer specifications. We included control reactions, in which Taq DNA polymerase was substituted for the reverse transcription enzyme mixture, to confirm that genomic DNA was not present in the RNA preparations. RNA was converted to cDNA (50 °C for 30 min), which was then PCR amplified in the same tube (45 cycles of the following conditions: 94 °C for 15 s, 52 °C for 30 s, and 68 °C for 2 min). Resulting DNA fragments were separated on 1% agarose gels in TAE buffer and visualized by ethidium bromide staining.

Allelic replacement of the speH and spy0129 genes in SF370. The strategy for allelic replacement of speH and spy0129 genes was followed as previously described [62]. Briefly, upstream and downstream DNA regions flanking both genes were separately amplified using the primer sets listed in Table S2. PCR products were treated with the appropriate restriction enzymes (New England Biolabs, http://www.neb.com) and used according to manufacturer instructions. We included control reactions, in which Taq DNA polymerase was substituted for the reverse transcription enzyme mixture, to confirm that genomic DNA was not present in the RNA preparations. PCR was performed using the primer sets listed in Table S2. RT–PCR analysis of total RNA extracted (as described) and treated, purified total RNA from late-log phase cultures (OD_600nm = 0.7) of strain SF370. All remaining components were added as per manufacturer specifications. We included control reactions, in which Taq DNA polymerase was substituted for the reverse transcription enzyme mixture, to confirm that genomic DNA was not present in the RNA preparations. RNA was converted to cDNA (50 °C for 30 min), which was then PCR amplified in the same tube (45 cycles of the following conditions: 94 °C for 15 s, 52 °C for 30 s, and 68 °C for 2 min). Resulting DNA fragments were separated on 1% agarose gels in TAE buffer and visualized by ethidium bromide staining.

Pharyngeal cell adherence. We tested late-logarithmic phase SF370spelH and SF370spel29 genes were the lowest clusters of all putative resulting neighbor clusters (PK value), using a permutation algorithm with the sum of the t-test statistics (generated by Cyber-T) from each gene within a given cluster as the metric for comparison. We then inspected the output visually and disqualified groupings that violate the neighborhood criteria on extracted expression patterns. Genes that were involved directly in the adherence of strain SF370 to pharyngeal cells. The parental strain SF370 served as control. We did not associate genes with functional annotations to prevent biasing the formation of clusters toward those that were “expected.” The GenomeCrawler algorithm, written in the statistical language R (http://www.R-project.org), steps through the expression data and identifies adjacent gene groupings that exhibit similar expression fold changes. The algorithm varies window size and applies a gap penalty for including in a cluster those genes that we did not observe experimentally to be present or genes that did not exhibit differential expression between sample and control.

GenomeCrawler calculates statistical significance of all putative neighbor clusters (PK value), using a permutation algorithm with the sum of the t-test statistics (generated by Cyber-T) from each gene within a given cluster as the metric for comparison. Neighboring cluster are provided in Table S2. RT–PCR analysis of total RNA extracted (as described) and treated, purified total RNA from late-log phase cultures (OD_600nm = 0.7) of strain SF370. All remaining components were added as per manufacturer specifications. We included control reactions, in which Taq DNA polymerase was substituted for the reverse transcription enzyme mixture, to confirm that genomic DNA was not present in the RNA preparations. RNA was converted to cDNA (50 °C for 30 min), which was then PCR amplified in the same tube (45 cycles of the following conditions: 94 °C for 15 s, 52 °C for 30 s, and 68 °C for 2 min). Resulting DNA fragments were separated on 1% agarose gels in TAE buffer and visualized by ethidium bromide staining.

||
generates the lowest \( P_c (M_c \mid t_c, K) \) for the specified gene determines the reported value of \( P_c (M_c \mid t_c, K) \).

Calculation of \( P_c (t_c \mid K) \) relies on Bayes’ Theorem (Equation 2) in which \( P_c (t_c) \) is the prior, \( P_c (K \mid t_c) \) is the likelihood, and \( P_c (K) \) is the probability associated with the cluster.

\[
P_c(t_c \mid K) = \frac{P_c(t_c) P_c(K \mid t_c)}{P_c(K)}
\]  

(2)

All of the right-hand side probabilities are readily calculated using the following general equation:

\[
P_c(\Lambda) = \frac{\sum_{g \in \Lambda} \left( \sum_{j \in J(g)} I(\Lambda) \geq \sum_{j \in J(g)} |\tau_j| \right)}{B}
\]  

(3)

in which \( \Lambda \subseteq K \) and \( \Lambda = \{t_j \mid j \in J \} \) in which \( \Gamma \subseteq J \) and represents a set of genes defined in the parenthesis of Equation 2. Since \( P_c (K) \) is the measure for the statistical significance of the specified cluster in our analysis, for this probability \( \Gamma = J \). For the prior, \( \Gamma = \{t_j' \mid j' \in J, j' \neq j, \exists \; g \in J \} \), whereas for the likelihood, \( \Gamma = \{j' \in J, j' \neq g, \exists \; g \in J \} \). \( B \) is the total number of iterations of permutation resampling performed, with \( b \) representing a resampled value of the 8th iteration. The indicator function \( I(\Lambda) \) equals 1 when the condition in parentheses is satisfied, and 0 when it is not.

**Relationship between \( P_c \) and \( P_c \) :** We define the relationship between \( P_c (M_c) \) and \( P_c (t_c \mid K) \), as both use \( t_c \) for their respective calculations. For \( t_c \simeq t_p, P_c (M_c) \rightarrow 0 \) and \( 0 < P_c (t_c \mid K) < 1 \). Therefore, the analysis ensures that even the most significant gene with respect to \( P_c (M_c) \) can theoretically have \( P_c (t_c \mid K) = 1 \). For example, when a gene is a member of a cluster in which the other members are insignificant on a genome scale, \( P_c (t_c \mid K) \rightarrow 1 \), since \( P_c (K \mid t_c) \rightarrow 1 \) and \( P_c (t_c \mid K) \rightarrow P_c (K) \). Conversely, for \( t_c \simeq t_p, P_c (M_c) \rightarrow 1 \) and \( 0 < P_c (t_c \mid K) < 1 \). Here, there is a strong dependency between \( P_c (M_c) \) and \( P_c (t_c \mid K) \). This prevents a gene with a relatively low \( t_c \) value from being scored as significant due to a pure circumstantial association with a gene of \( P_c (M_c) \rightarrow 0 \). Hence, this analysis exhibits the required dynamic relationship between \( P_c (M_c) \) and \( P_c (t_c \mid K) \) and, more important, is consistent with the criterion that, on its own, a gene with a lower \( P_c (M_c) \) should not generate an informative \( P_c (t_c \mid K) \) (i.e., \( P_c (t_c \mid K) \ll 0 \)). \( P_c (t_c \mid K) \), therefore, reflects a group context derived from a cluster of genes and is not merely the recapitulation of the \( P_c (M_c) \) of an individual gene.

**Identification of putative operons:** The published SF370 genome does not contain promoter annotations, so we examined the entire genome in 100,000-bp segments (available for download at ftp://ftp.ncbi.nih.gov/pub/strrep/ and used the Vector NTI advance 9.0 sequence analysis suite (Invitrogen) to identify sequences that were similar (75% identity threshold) to the consensus (FASTA Crawler) promoter sequences [64]. We cross-referenced the clusters containing a single upstream putative promoter sequence with a list of rhinoceros-independent terminator sequences, previously identified in the SF370 genome by TransTerm (www.tigr.org/software/transterm.html).

**Analysis previously published streptococcal microarray data:** We analyzed recently published microarray data from \( S. pyogenes \) strain SF370 [57] in the same manner as the adherence data presented in this study to assess the overall reliability of our analytical methods. We applied the initial statistical package to assess the differential expression of individual genes, followed by the GenomeCrawler algorithms. We compared the results of this analysis, when applicable, to the published analysis of the array data.

**Software and microarray datasets:** For MIAME (Minimum Information About a Microarray Experiment) compliance, all microarray datasets (pre- and post-processing) have been deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus. We cross-referenced the clusters containing a single upstream putative promoter sequence with a list of rhinoceros-independent terminator sequences, previously identified in the SF370 genome by TransTerm (www.tigr.org/software/transterm.html).

**Supporting Information**

**Figure S1.** Comparison of Microarray and TaqMan Analyses

Found at doi:10.1371/journal.pcbi.0030132.s001 (876 KB EPS).

**Figure S2.** Statistically Significant Type I Neighbor Clusters in the SF370 Genome

Found at doi:10.1371/journal.pcbi.0030132.s002 (1.7 MB EPS).

**Figure S3.** Statistically Significant Type II Neighbor Clusters in the SF370 Genome

Found at doi:10.1371/journal.pcbi.0030132.s003 (1.7 MB EPS).

**Figure S4.** Statistically Significant Type III Neighbor Clusters in the SF370 Genome

Found at doi:10.1371/journal.pcbi.0030132.s004 (1.0 MB EPS).

**Table S1.** Summary of All Streptococcal Genes Exhibiting Differential Expression during Adherence to Pharyngeal Cells as Compared with Associated Streptococcal Control

Found at doi:10.1371/journal.pcbi.0030132.s001 (158 KB XLS).

**Table S2.** DNA Primers Used in This Study

Found at doi:10.1371/journal.pcbi.0030132.s002 (25 KB XLS).

**Table S3.** Fold Changes in Gene Expression Estimated Using Microarray and Real-Time Quantitative PCR

Found at doi:10.1371/journal.pcbi.0030132.s003 (16 KB XLS).

**Table S4.** All Putative Neighbor Clusters Generated by GenomeCrawler

Found at doi:10.1371/journal.pcbi.0030132.s004 (144 KB XLS).

**Table S5.** Putative Neighbor Clusters Following Visual Inspection

Found at doi:10.1371/journal.pcbi.0030132.s005 (38 KB XLS).

**Table S6.** Summary of Neighboring Clusters Identified as Putative Operons through In Silico Inspection

Found at doi:10.1371/journal.pcbi.0030132.s006 (21 KB XLS).

**Table S7.** Comparison of Log-Fold Changes of Genes Reported in a Previously Published Microarray Study of \( S. pyogenes \) SF370

Found at doi:10.1371/journal.pcbi.0030132.s007 (40 KB XLS).

**Table S8.** Summary of All Genes in a Mg-Deficient SF370 Mutant Exhibiting Differential Expression during Growth in Culture Broth Compared with the Parental SF370 Control

Found at doi:10.1371/journal.pcbi.0030132.s008 (188 KB XLS).

**Table S9.** Putative Neighbor Clusters Identified by GenomeCrawler in Published Microarray Dataset from \( S. pyogenes \) SF370

Found at doi:10.1371/journal.pcbi.0030132.s009 (28 KB XLS).

**Accession Numbers**

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession number for \( S. pyogenes \) strain SF370 (serotype M1) is AE004092.

The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) accession number for all microarray datasets in this paper is GSE 7620.

**Acknowledgments**

We thank J. Ferretti for providing \( S. pyogenes \) strain SF370 and G. Khitrov for numerous microarray supplies. We are grateful to A. Farnsworth, J. Loeffler, and M. Collins for insightful discussions and review of the manuscript; M. Zeegers for review of the statistical analysis and algorithms; and all members of the Fischetti lab, especially T. Broudy, B. Juncosa, A. Raz, A. Daniel, and D. Nelson for helpful comments and suggestions. We especially thank Shiwei Zhu for expert technical support and assistance in cloning and PCR. We are indebted to Tshaka Cunningham for technical advice for qRT-PCR assays and analysis. We sincerely thank V. Pancholi, A. Khitrov, A. Pearlman, and D. Nayduch for invaluable discussions and FDA grant number 18225 from the National Institutes of Health (VAF).

**Competing interests:** The authors have declared that no competing interests exist.
References

1. Quackenbush J (2005) Microarrays—Guilt by association. Science 302: 240–241.
2. Cayon T, Schoolnik GK (2005) Microarray expression profiling: Capturing a genome-wide portrait of the transcriptome. Mol Microbiol 47: 879–889.
3. Quackenbush J (2002) Computer analysis of microarray data. Nat Rev Genet 2: 418–427.
4. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
5. Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 9042–9046.
6. Cunningham MW (2000) Pathogenesis of Group A streptococcal infections. Clin Microbiol Rev 13: 470–491.
7. Voyich JM, Sturdevant DE, Braughton KR, Kobayashi SD, Lei B, et al. (2002) Genome-wide protective response used by Group A streptococci to evade destruction by human polymorphonuclear leukocytes. Proc Natl Acad Sci U S A 100: 1996–2001.
8. Banks DJ, Lei B, Musser JM (2003) Prophage induction and expression of prophage-encoded virulence factors in Group A Streptococcus serotype M3 strain SF370. Proc Natl Acad Sci U S A 100: 4958–4963.
9. Ryan PA, Pancioli V, Fischetta VA (2001) Induction of lyogenic bacteriophage and phage-associated toxin from Group A streptococci during coculture with human pharyngeal cells. Infect Immun 69: 1440–1443.
10. Broudy TB, Pancioli V, Fischetta VA (2002) The in vitro interaction of Streptococcus pyogenes with human pharyngeal cells induces a phage-encoded extracellular DNAse. Infect Immun 70: 2805–2811.
11. Banks DJ, Lei B, Musser JM (2003) Prophage induction and expression of prophage-encoded virulence factors in Group A Streptococcus serotype M3 strain MGAS3153. Infect Immun 71: 7079–7086.
12. Chaussee MS, Cole RL, van Putten JP (2000) Streptococcal erythrogenic toxin B interacts with human fibronectin: A novel mechanism of bacterial adhesion. Infect Immun 68: 1211–1213.
13. Riccardi VM, Rognoni P, De’Lellis C, Sesenna R, Barbareschi M, et al. (2002) The T cell recognition repertoire of human streptococcal exotoxins. J Immunol 168: 980–987.
14. Nyberg P, Rasmussen MK, Lund M, Bjorck L, et al. (2004) Expression profile, binding characteristics, biological functions, and regulation of the streptococcal superantigen streptolysin O. J Biol Chem 279: 33301–33308.
15. Rognoni P, Riccardi VM, Rossolini MG, De’Lellis C, et al. (2002) A genome-wide portrait of the transcriptome in host-streptococcal interactions. Infect Dis 191: 2121–2129.
16. Karsli SU, Kansal R, Aziz RK, Hooshranian M, Norrby-Teglund A, et al. (2001) Reciprocal, temporal expression of SpeA and SpeB in invasive M1 Group A streptococcal isolates in vivo. Infect Immun 69: 4988–4995.
17. Aziz RK, Fabijan MJ, Jiang J, Kansal R, Low DE, et al. (2004) Invasive M1 Group A streptococci undergo a phase-shift in vivo to prevent the proteolytic degradation of multiple virulence factors by SpeB. Microbiol Mol Biol Rev 68: 1251–1253.
18. Proctor RM, Mora M, Bensi G, Capo S, Falugi F, et al. (2005) Growth of Group A and Lancefield T antigens. Proc Natl Acad Sci U S A 102: 15641–15646.
19. Aziz RK, Fabijan MJ, Jiang J, Kansal R, Low DE, et al. (2004) Invasive M1 Group A streptococci undergo a phase-shift in vivo to prevent the proteolytic degradation of multiple virulence factors by SpeB. Microbiol Mol Biol Rev 68: 1251–1253.
20. Proctor RM, Mora M, Bensi G, Capo S, Falugi F, et al. (2005) Growth of Group A and Lancefield T antigens. Proc Natl Acad Sci U S A 102: 15641–15646.
21. Aziz RK, Fabijan MJ, Jiang J, Kansal R, Low DE, et al. (2004) Invasive M1 Group A streptococci undergo a phase-shift in vivo to prevent the proteolytic degradation of multiple virulence factors by SpeB. Microbiol Mol Biol Rev 68: 1251–1253.
22. Proctor RM, Mora M, Bensi G, Capo S, Falugi F, et al. (2005) Growth of Group A and Lancefield T antigens. Proc Natl Acad Sci U S A 102: 15641–15646.
23. Aziz RK, Fabijan MJ, Jiang J, Kansal R, Low DE, et al. (2004) Invasive M1 Group A streptococci undergo a phase-shift in vivo to prevent the proteolytic degradation of multiple virulence factors by SpeB. Microbiol Mol Biol Rev 68: 1251–1253.
24. Proctor RM, Mora M, Bensi G, Capo S, Falugi F, et al. (2005) Growth of Group A and Lancefield T antigens. Proc Natl Acad Sci U S A 102: 15641–15646.
25. Aziz RK, Fabijan MJ, Jiang J, Kansal R, Low DE, et al. (2004) Invasive M1 Group A streptococci undergo a phase-shift in vivo to prevent the proteolytic degradation of multiple virulence factors by SpeB. Microbiol Mol Biol Rev 68: 1251–1253.
26. Proctor RM, Mora M, Bensi G, Capo S, Falugi F, et al. (2005) Growth of Group A and Lancefield T antigens. Proc Natl Acad Sci U S A 102: 15641–15646.
transcriptomes analysis reveals both direct and indirect regulation by Mga in the group A streptococcus. Mol Microbiol 62: 491–508.

58. Kreikemeyer B, McIver KS, Podbielski A (2003) Virulence factor regulation and regulatory networks in Streptococcus pyogenes and their impact on pathogen–host interactions. Trends Microbiol 11: 224–252.

59. Nelson D, Loomis L, Fischetti VA (2001) Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. Proc Natl Acad Sci U S A 98: 4107–4112.

60. Philbrick JB, Diner BA, Zilinskas BA (1991) Construction and characterization of cyanobacterial mutants lacking the manganese-stabilizing polypeptide of photosystem II. J Biol Chem 266: 13570–13576.

61. Klebanov L, Qiu S, Welle S, Yakovlev A (2007) Statistical methods and microarray data. Nat Biotechnol 25: 25–26.

62. Euler CW, Ryan PA, Martin JM, Fischetti VA (2007) M.SpyI, a DNA methyltransferase encoded on a mefA chimeric element, modifies the genome of Streptococcus pyogenes. J Bacteriol 189: 1044–1054.

63. Podbielski A, Spellberg B, Woiwodik M, Pohl B, Latteken R (1996) Novel series of plasmid vectors for gene inactivation and expression in group A streptococci (GAS). Gene 177: 137–147.

64. Provvedi R, Maggi T, Oggioni MR, Manganelli R, Pozzi G (2005) Selection and characterization of a promoter for expression of single-copy recombinant genes in Gram-positive bacteria. BMC Biotechnol 5: 3.