Encyclopedia of bacterial gene circuits whose presence or absence correlate with pathogenicity – a large-scale system analysis of decoded bacterial genomes

Maksim Shestov, Santiago Ontañón and Aydin Tozeren

Abstract

Background: Bacterial infections comprise a global health challenge as the incidences of antibiotic resistance increase. Pathogenic potential of bacteria has been shown to be context dependent, varying in response to environment and even within the strains of the same genus.

Results: We used the KEGG repository and extensive literature searches to identify among the 2527 bacterial genomes in the literature those implicated as pathogenic to the host, including those which show pathogenicity in a context dependent manner. Using data on the gene contents of these genomes, we identified sets of genes highly abundant in pathogenic but relatively absent in commensal strains and vice versa. In addition, we carried out genome comparison within a genus for the seventeen largest genera in our genome collection. We projected the resultant lists of ortholog genes onto KEGG bacterial pathways to identify clusters and circuits, which can be linked to either pathogenicity or synergy. Gene circuits relatively abundant in nonpathogenic bacteria often mediated biosynthesis of antibiotics. Other synergy-linked circuits reduced drug-induced toxicity. Pathogen-abundant gene circuits included modules in one-carbon folate, two-component system, type-3 secretion system, and peptidoglycan biosynthesis. Antibiotics-resistant bacterial strains possessed genes modulating phagocytosis, vesicle trafficking, cytoskeletal reorganization, and regulation of the inflammatory response. Our study also identified bacterial genera containing a circuit, elements of which were previously linked to Alzheimers disease.

Conclusions: Present study produces for the first time, a signature, in the form of a robust list of gene circuitry whose presence or absence could potentially define the pathogenicity of a microbiome. Extensive literature search substantiated a bulk majority of the commensal and pathogenic circuitry in our predicted list. Scanning microbiome libraries for these circuitry motifs will provide further insights into the complex and context dependent pathogenicity of bacteria.

Keywords: Bacteria, Pathogen, Ortholog, Gene circuits, Virulence factors
Background

Microbiology experiments identified a large number of bacterial virulence mechanisms conserved through evolution [1]. Pathogenic bacterial strains could be defined as those with capacity to harm the host and cause disease [2]. A bacterial strain may appear as asymptomatic or pathogenic depending on the state of the immune system of the host [3], composition of the microbiome [4], presence and absence of elicitors [5], and other environmental factors [6]. Synthesis of secreted virulence factors are transcriptionally regulated by environmental stimuli [7]. Production of enzymes that degrade host cytoskeleton and cause damage is linked to bacterial strain density and quorum sensing [8]. It is clear that pathogenicity is context dependent.

The large research literature on bacterial pathogenicity was recently curated into an open access web platform, the Virulence Factor Database or VFDB [2, 6, 9]. The database presents genes and gene groups (virulence factors) associated with infectious disease. Since the metagenome of a human microbiome contains over three million genes, the extent of our knowledge on virulence genes may not be close to saturation. Nevertheless, not only the VFDB database is valuable for identifying genes linked to pathogenicity in specified bacteria (even strains), but it also serves as a benchmark for studies predicting pathogenicity linked gene clusters.

Virulence factors may cause disease in multiple hosts [1]. Yet, specificity of virulence factors to certain bacterial and strains is poorly understood. A recent study involving 50 bacterial genomes found some virulence factors to be exclusive to pathogenic bacterial strains in this small sample [10]. Virulence genes abundant in pathogenic as well as nonpathogenic strains included those facilitating coding of translocation proteins, apparatus proteins, and chaperons. The findings of the study suggest the presence of pathogenic gene circuits in which some genes belong exclusively to pathogenic strains whereas others can be abundant in bacterial strains synergistic or commensal to the host.

The drivers of pathogenicity may not only be virulence factors but also the absence of antivirulence factors in bacterial genomes [11–13]. Some bacterial species contain strains known as commensal to a host and other strains that exhibit pathogenicity. Transformation to pathogenic state maybe due to exchange of DNA fragments between bacteria [14]. Literature points out to the acquisitions of gene clusters and pathogenicity islands via horizontal gene transfer [15–17]. Deletion of genes or loss of gene function through mutation appears to be part of the adaptation to newly acquired pathogenesis [18].

Recent identification of the elicitors for activating antibiotics-synthesizing bacterial gene circuits [5] provides a new dimension in our understanding of the environmental forces affecting pathogenicity. Although, the research literature contains many examples of bacterial genes linked to synthesis of antibiotics [19–21], a large-scale study exposing diversity of such genes and gene circuits is yet to be carried out. The same is true for pathogenic gene circuits. That is the reason we set out in this study a systematic approach to identify and annotate prokaryotic ortholog gene circuits whose presence or absence are linked to phenotypes of antagonism to the host. Comparative genomics has been utilized before in the identification of drug and vaccine targets in Staphylococcus aureus [22] and mycobacterial peptidoglycan remodeling enzymes linked to pathogenicity [23]. Ours, however, takes it one step further, employing comparisons among 2527 distinct genomes. Restriction of genome comparison to ortholog groupings reduced dimension of this meta-scale analysis. It also expanded the reach of findings within the context of evolution. Aspects of pathogenicity, which are universal across eukaryotic hosts, will likely emerge in this approach [24].

Our methodology has multiple steps. First, we annotate existing genome sequences of bacterial strains as pathogenic or otherwise, based on literature curation. Next, we identify the presence and absence of orthologs in the genomes of these bacterial strains. Thirdly, we determine the relative abundance or absence of these orthologs in the pathogenic and non-pathogenic strains within and across genera. Projection onto cellular pathways result in annotation of gene circuits linked to pathogenicity. Extensive comparison with experimental literature provides biological context to our findings. Our study in effect creates an encyclopedia for pathogenicity, built on big data on genome sequences and literature on phenotypes of bacterial strains.

Results

A. Orthologs linked either to pathogenicity or synergy with the host

Annotation of pathogenic bacterial strains

Our literature search detailed in the Methods section identified 949 bacterial strains as pathogenic, meaning they had been reported in the literature as pathogenic to an animal host at least once. The label pathogenic, in the way we use it, doesn’t mean that the bacterial strain will cause disease to an animal host under all circumstances.

Additional file 1 presents the list of bacterial strains deemed as pathogenic, with evidence provided in the file in the form of references or database citations. This supplemental file also contains labeling of pathogenic strains found as antibiotic resistant in the Antibiotic Resistance Genes Database (ARDB) [25]. Table 1 shows a sample of the bacterial genera possessing both pathogenic and non-pathogenic strains. Strains of the same bacterial
genus often separated into pathogenic and nonpathogen- 
cic clusters.

There was no KEGG- or literature-recorded evidence 
of pathogenicity for the remaining 1578 decoded bacter-
ial strain genomes in the KEGG database. Thus, they 
were deemed for this study as nonpathogenic. Additional 
file 2 presents the list of the 1578 bacterial strains 
deemed nonpathogenic in our study.

Orthology contents of pathogenic and nonpathogenic 
strains

Next, we created two sets of matrices for ortholog genes, 
one set for the genomes of pathogenic and the other for 
nonpathogenic strains. The columns of each matrix iden-
tified the bacterial strain whereas rows identified whether 
an ortholog was present (1) or absent (0) in that strain. It 
turned out that 7194 different orthologs were present in 
at least one of the genomes of the 2527 bacterial strains 
under study. These large matrices are used for the abun-
dance computations presented in this study and hence in-
cluded as Additional file 3.

For a given ortholog, fractions of pathogenic and non-
pathogenic bacterial strains presenting the ortholog in 
their genomes are represented by symbols $A_p$ and $A_{np}$, 
respectively. The scatter diagram shown in Fig. 1 pre-
sents the $A_p$ and $A_{np}$ values for the 7194 orthologs 
present in bacterial strains. It appears that most of the 
orthologs have comparable presence in all bacterial 
strains whereas a small portion (green and red dots for 
pathogenic and nonpathogenic strains, respectively) is 
biased towards one of the two phenotypes.

The histogram shown in Fig. 2 is another view of the 
data presented in the scatter diagram in Fig. 1. Here, we 
plotted the frequency of occurrence against the patho-
gen abundance score $\log PA$ for all the orthologs under 
consideration. The parameter $PA = A_p / (A_{np} + 0.0001)$ 
is a measure of relative abundance of the ortholog in 
pathogenic strains. In cases where $A_{np}$ equalled zero, the 
equation still enables division due to the presence of 
0.0001 in the denominator. The two tail ends of the dis-
tribution indicate those orthologs abundant in patho-
genic but rarely found in nonpathogenic (red) and vice 
versa (green). The cutoff values we used ($PA > 4$, and 
$PA < 0.5$), although somehow arbitrary, were placed at the 
inner edges of the tails of the histograms.

Shown in Fig. 3a are the overall characteristics of the 
PA distribution among the orthologs. In brief, there 
were 229 pathogenic only and an additional 379 ortho-
logs for which $PA > 4$, representing about 8 percent of 
the orthologs found in our bacterial strain library. 
Taken together, we deemed this group as pathogen 
abundant or pathogen-linked. Total number of genes in 
the 608 pathogen-abundant orthologs was 18,982, indi-
cating their presence in a diverse set of bacterial spe-
cies. Pathogen exclusive orthologs comprised only 
1,518 of this set of genes, suggesting most genes previ-
ously linked to pathogenicity is not exclusive to 
disease-causing bacterial strains.

| Genus          | Total | Pathogenic | Genus          | Total | Pathogenic |
|----------------|-------|------------|----------------|-------|------------|
| Streptococcus  | 120   | 77         | Rickettsia     | 41    | 19         |
| Chlamydia      | 98    | 81         | Burkholderia   | 39    | 17         |
| Bacillus       | 81    | 22         | Listeria      | 34    | 25         |
| Candidatus     | 71    | 4          | Campylobacter  | 26    | 13         |
| Mycobacterium  | 65    | 34         | Vibrio        | 25    | 20         |
| Mycoplasma     | 65    | 54         | Brucella      | 20    | 18         |
| Escherichia    | 64    | 39         | Acinetobacter | 19    | 12         |
| Helicobacter   | 63    | 47         | Francisella   | 19    | 13         |
| Pseudomonas    | 53    | 12         | Yersinia      | 19    | 17         |
| Staphylococcus | 51    | 42         | Borrelia      | 18    | 14         |
| Corynebacterium| 50    | 34         | Neisseria     | 18    | 18         |
| Clostridium    | 49    | 22         | Treponema     | 17    | 13         |
| Salmonella     | 45    | 36         | Haemophilus   | 16    | 13         |

The "a" identifies those genera for which we have also conducted within genus comparisons.

Table 1 List of bacterial genera highlighting the number of strains that have been associated with the pathogenic state, as well as the strains associated with the commensal state in the present study

| Total bacterial strain genomes | 2527 |
|-------------------------------|------|
| Total pathogenic bacterial strains | 949  |
| KEGG based pathogenic strains | 767  |
| Additional literature based pathogenic strains | 182  |

The "a" identifies those genera for which we have also conducted within genus comparisons.
The number of orthologs exclusive to nonpathogenic strains was much larger at 879, and an additional 485 had \( PA < \frac{1}{4} \). The rest, a total of 5222 orthologs, were commonly present among pathogenic and nonpathogenic strains. It is expected that these numbers will change as the number of available bacterial genomes in the literature increase from thousands to tens of thousands.

Next, we identified those orthologs in our list, a total of 1308, which were also present in the Virulence Factor Database, VFDB, by matching either the gene names or gene descriptions. As indicated in Fig. 3b, VFDB orthologs are significantly biased towards pathogen abundant orthologs. Additional file 4 lists the pathogen-abundant and non-pathogen-abundant orthologs in accordance with the \( PA \) ranking, along with VFDB labeling if present in that database. Overall, our study indicates the absence of one-to-one match between known virulence factors and pathogen-abundant orthologs.

**Orthologs enriched in pathogenic, antibiotic resistant, and nonpathogenic bacterial strains**

Statistical enrichment of KEGG pathways was conducted based on the hypergeometric test for ortholog sets \( PA > 4 \), and \( PA < \frac{1}{4} \), respectively. Results are shown in Fig. 4. Orthologs abundant in pathogenic strains crowd pathogen-linked cellular pathways: *Staphylococcus aureus*, *Legionellosis*, *Pertussis*, *Salmonella*, *Shigellosis*, and *Escherichia coli* infections, as well as epithelial signaling in *H. pylori* infection. Pathogen abundant orthologs are also found in pathways
involving bacterial secretion systems, Nod like receptor signaling, bacterial invasion of epithelial cells, and plant-pathogen interactions.

Orthologs found exclusively in nonpathogenic strains occupy nodes in metabolic pathways (Fig. 4). These pathways include biosynthesis of peptidoglycans, microlides, carotenoids, ansamycins, and nonribosomal peptides. The GO cell compartment investigations not shown in the figure indicate that pathogen abundant orthologs are enriched in crosstalk positions of contact with the host whereas nonpathogen exclusive orthologs code proteins involving in events in the cell interior.

Next, we looked at molecular function enrichments of pathogenic associated orthologs and compared the results with corresponding enrichments obtained using the VFDB database. Figure 5 shows that both our annotation
and VFDB contain roughly equal amounts of orthologs in secretion, toxins, peptidase, and pilin categories. However, pathogen abundant list of the present study has significantly more abundance in enzyme categories such as oxidoreductases, transferases, hydrolases, lyases, and ligases. Some of the orthologs in our list are also enriched in regulatory function, particularly involving transcription and translation.

Pathway enrichment was also conducted within the population of pathogenic strains for the subset identified as antibiotic resistant using ARDB Database [25]. The hypergeometric test revealed the pathways shown in Table 2 as particularly enriched in antibiotic resistant strains. These included sphingolipid metabolism, producing bioactive metabolites that regulate cell function [26], PI3K-Akt signaling pathway, an intracellular pathway important in apoptosis [27], and Aminoacyl-tRNA biosynthesis [28]. Some of the modules in the enriched pathways also appear in eukaryotic processes for drug resistance against chemotherapy. One must caution, however, that the results could potentially change with the updating of ARDB, even if the p values in these enrichments are vanishingly small.

Gene clusters linked either to pathogenicity or synergy
This section presents results on genetic circuits statistically enriched in pathogenic and nonpathogenic bacteria. We used two different types of comparison to achieve our results: a) analyzing the entire set of genomes partitioned into pathogenic and nonpathogenic strains; and b) conducting the same operation within genera for the 17 genera identified in Table 1 with a star. In the first approach, we mapped the list of ortholog genes linked to pathogenicity (PA > 4) and non-pathogenicity onto KEGG reference pathways and identified, based on KEGG repository, those multiply connected clusters of genes (gene circuits) containing at least three pathogen-linked or nonpathogen-linked orthologs. Results are shown in Tables 3 and 4, respectively. Additional file 5 presents corresponding results for within-genera comparisons, both for pathogen- and nonpathogen-linked orthologs. Results are shown in Tables 3 and 4, respectively. Additional file 5 presents corresponding results for within-genera comparisons, both for pathogen- and nonpathogen-linked orthologs. Results are shown in Tables 3 and 4, respectively. Additional file 5 presents corresponding results for within-genera comparisons, both for pathogen- and nonpathogen-linked orthologs. Results are shown in Tables 3 and 4, respectively.

| Pathway                                    | p value via hypergeometric test |
|--------------------------------------------|----------------------------------|
| D-Arginine and D-ornithine metabolism      | 0.00572                          |
| Amino sugar and nucleotide sugar metabolism| 0.0273                           |
| Peptidoglycan biosynthesis                 | 0.00135                          |
| Sphingolipid metabolism                    | 0.00524                          |
| Aminoacyl-tRNA biosynthesis                | 0.00076                          |
| Phosphatidylinositol signaling system      | 0.04909                          |
| PI3K-Akt signaling pathway                 | 0.00384                          |
| Amoebiasis                                 | 0.03099                          |

Gene clusters more common in pathogenic strains
The Table 3 presents a set of gene circuits with the ortholog genes linked to pathogenicity and also indicates the pathway to which the gene circuits belong.
Along with Table 3, comes Fig. 6, in which the wiring diagrams for the gene circuitry are shown in the form of cutouts from the KEGG Reference pathways. The circuits in the figure have the same ordering number used in Table 3. Note also that the actual circuits contain orthologs not only pathogen-linked (shown in pink and orange) but also others, some preferentially found in pathogenic strains and others more ubiquitous. The p value through hypergeometric test for a bacterial strain containing at least one pathogen-ortholog in a circuit shown in Table 3 was less than 0.01.

The circuitry in Table 3 and Fig. 6 falls into the following categories:

Gene circuits for bacterial secretion and invasion pathways:

The type III bacterial secretion system pathway mediates toxin and protein delivery to host cells. Table 3 shows the existence of multiple clusters of pathogen-linked orthologs in this pathway. Consistent with our findings, the type III pathway is listed in the literature as modulating pathogenic interactions with host organisms including animals and plants [29–32]. Also shown in Fig. 6 are examples of pathogen-linked circuits in the secretion system. One such circuit contains pathogen-linked orthologs yscF, yscO, yscP, yscX, yscC, and yscW. Subsets of pathogen-linked orthologs of the cluster are present in 242 pathogenic and 106 nonpathogenic strains, resulting in vanishing p values in hypergeometric test. Moreover, the bias towards pathogenicity increases dramatically with increasing number of pathogen-linked orthologs in this cluster in the genome of the bacterial strain. Another secretion-linked pathway is that of type IV gene circuit, for which some of its genes exist in both pathogenic and nonpathogenic strains. The circuit functions in translocation of DNA and protein substrates to target cells via direct cell-to-cell contact. In our study, the complete circuit is preferentially present in pathogenic strains. Consistent with these observations, recent investigations uncovered a role for pathogenicity for this circuit [33–35].

Pathogen-linked gene clusters in the two-component System:

The two-component regulatory system is a stimulus–response coupling pathway, which enables bacteria to sense and respond to changes in its environment [36–40]. Membrane-bound histidine kinases are major building blocks of the pathway. These signal transduction systems modulate crosstalk between species within the microbiome. The Table 3 contains multiple gene clusters (circuits) in the two-component system containing orthologs linked to pathogenicity: cluster (devS, nreA, nreB, narT) involved in hypoxia, oxygen, and nitrogen assimilation; cluster (uhpC, uhpA, uhpT) modulating hexose phosphate uptake; and the cluster (pagC, pagO, pagD, pagK, pgkE) involved in Mg2+ starvation, and others. See also Fig. 6 for the wiring diagrams of these clusters. Elements of the metabolite assimilation cluster have been linked in the literature to pathogenicity of mycobacterium tuberculosis [41, 42]. The second cluster in the list in the two-component system, mediating hexose phosphate uptake, plays an important role in the sodium-dependent D-glucose transport protein of Helicobacter pylori [43]. This gene circuit is involved in Mg2+ starvation and was shown to play a role in the pathogenicity of Salmonella enterica [44]. Mg2+ starvation is also involved in quorum sensing of Pseudomonas fluorescens [45] and in biosynthesis of complex lipids needed for virulence of mycobacterium tuberculosis [46].

Metabolic circuits linked to pathogenicity:

A metabolic gene circuit whose genes are commonly found in pathogenic strains is the CMP-Pse metabolism circuit cluster belonging to the amino nucleotide sugar mechanism. Pathogen-linked ortholog genes in this circuit consist of pscC, pscH, pseF, and ILAP (Table 3). This circuit is linked to the synthesis of glycoconjugates, which are typically expressed on the surfaces of pathogenic bacteria. The protein products of the circuit have already been identified as virulence factors in the VFDB database and in the literature [47–49].

Nodal elements of the Peptidoglycan biosynthesis circuit cluster shown in Table 3 are also preferentially present in pathogenic strains. Pathogen-linked orthologs in this gene circuit consist of the genes sgtA, sgtB, femA, phbA, femB, psh3, femX, and fnhb. Peptidoglycans are polymers consisting of sugars and amino acids forming a mesh scaffold external to the plasma membrane. Recent studies in the literature point to the role of peptidoglycans in the pathogen phenotype of different bacteria [50–52]. Sorbose to Sorbose 1-phosphate circuit of the Phosphotransferase (PTS) system also shown in Table 3 contains pathogen-linked genes PTS-Sor-EIIC, sorA, PTS-Sor-EIID, sorM, PTS-Sor-EIIA, sorF, PTS-Sor-EIIB, sorB. PTS circuit codes a group translocation process present in many bacteria, transporting sugars from the environment into the bacterial cell. The circuit has been linked in the literature to Streptococcus invasion [53].

Our statistical computations based on hypergeometric test indicate that the likelihood of pathogenic identification of a strain increases dramatically with increasing numbers of the circuit.
| Pathway | Circuit | Orthologs | Genera |
|---------|---------|-----------|--------|
| 1 Two Component system | Mg2+ starvation, antimicrobial peptide | pagC, pagO, pagD, pagK, pgtE | S. enterica*, B. aphidicola, E. coli*, P. ananatis, Y. pseudotuberculosis |
| 2 Two Component system | Hexose phosphate uptake | uhpC, uhpA, uhpT | C. trachomatis, E. coli*, S. enterica*, S. aureus, L. monocytogenes |
| 3 Two Component system | Competence Stimulating Peptide | ABC-BA-COMA, comA, comB, comC, comD, comE | S. pneumoniae, L. casei, L. rhamnosus, S. equi, L. plantarum |
| 4 Two Component system | Hypoxia Oxygen, Nitrogen assimilation | devS, nreA, nreB, narT | S. aureus*, M. tuberculosis, M. bovis, M. canettii, P. polyomysa |
| 5 Two Component system | Phosphoglycerate transport | pgtC, pgtB, pgtA, pgtP | S. enterica*, C. jejuni, V. cholerae, K. pneumoniae, E. coli |
| 6 Biosynthesis of siderophore group nonribosomal peptides | Yersiniabactin | irp1, HMWP1, irp5, ybtE, irp3, ybtU | E. coli*, Y. pestis, R. solanacearum, C. diphtheriae, K. pneumoniae |
| 7 Biosynthesis of siderophore group nonribosomal peptides | Pyochelin | pchF, pchG, pchD, pchE | P. aeruginosa, B. pseudomallei, B. cenocepacia, A. dieselolei, A. lipoferum |
| 8 Bacterial secretion system Type III | yscF, yscO, yscP, yscX, yscC, yscW | C. trachomatis*, E. coli*, C. psittaci, P. aeruginosa |
| 9 Bacterial secretion system Type V | vacA, yadA, yadB_C | H. pylori*, Y. pestis, Y. pseudotuberculosis, Y. enterocolitica, H. cetonatum |
| 10 Peptidoglycan biosynthesis | Peptidoglycan branch | sgtA, sgtB, femA, pbpA, femB | S. aureus, C. Anthromitus, E. sp., S. aureus |
| 11 Carotenoid biosynthesis | Terpenid backbone to Staphyloxanthin | crtM, crtP, crtQ, K10212, crtO | S. aureus*, B. megaterium, C. sp., S. lugdunensis, B. infantis |
| 12 Salmonella infection | Translocon | sipB, ipaB, ipaC, ipaC, ipaD, ipaD, ipaD, sipD, sipD, sipE, sseB, sseC, sseD | S. enterica*, B. pseudomallei, E. coli*, S. baltica, B. mallei |
| 13 Salmonella infection | Type III secretion system/SP2-2 encoded | ssl, sflA, sflF, sseG, pipB2, sspH2, sseH, srfH, spB | S. enterica*, S. baltica, E. tarda, Y. enterocolitica, C. violaceum |
| 14 Pentose phosphate pathway | D-Glucosamine metabolism | dgaA-dgaE, PTS-Dga-EI, PTS-Dga-EII, PTS-Dga-EIII | S. enterica*, E. coli*, E. faecalis, C. difficile, L. casei |
| 15 Bacterial invasion of epithelial cells | ECM-receptor interaction | pfba, sfb1, fnbA, fnbB | S. aureus, S. pyogena, S. pneumoniae, S. dysgalactiae, S. equi |
| 16 Phosphotransferase system | Sorbose to Sorbose 1-phosphate | PTS-Sor-EIIC, sorA, PTS-Sor-EII, sorM, PTS-Sor-EII, sorF, PTS-Sor-EII, sorB | E. coli*, K. pneumoniae, L. casei, L. rhamnosus, S. flexneri |
| 17 Phosphotransferase system | D-Glucosamin to D-Glucosamine 6-P | dgaA-dgaE, PTS-Dga-EI, PTS-Dga-EII, PTS-Dga-EIII, PTS-Dga-EVa | S. enterica*, E. coli*, E. faecalis, C. difficile, L. casei |
| 18 Amino nucleotide sugar metabolism | CMP-Pse metabolism | pseC, pseH, pseF, UAP1 | S. enterica*, E. coli*, S. pyogenes, V. cholerae, C. difficile, L. casei |
| 19 Plant-pathogen interaction | Bacterial secretion system | hopAB, avrPtoB, avrPto1, avrRpm1, avrXcc, avrB | P. syringae, X. campestris, X. campestris |
| 20 Bacterial invasion of epithelial cells | Type III - Salmonella | sipA-sipD, ipaA-ipaD, ipaB, pipB, sipC, sipC, sipC, sipD, sipD, sipD, sipE, sseB, sseC, sseD | S. enterica, B. pseudomallei, E. coli, B. mallei, B. thailandensis |
| 21 Bacterial invasion of epithelial cells | Type III - Shigella | sipC, ipaA-ipaC, ipaC, sipC, sipC, sipB, sipB, sipB, sipB, sipB, sseB, sseC, sseD | S. enterica, E. coli, B. pseudomallei, B. mallei, B. thailandensis |
| 22 Vibrio cholerae infection | Type II secretion system | ctxA, ctxB, ace, rtxA | V. cholerae*, E. coli*, V. vulnificus, Y. enterocolitica, C. hydrophila |
| 23 Vibrio cholerae infection | Type IV pilus | tcpA, tcpB, tcpC, tcpD, tcpE, tcpF | V. cholerae*, C. rodentium, E. cloaca, R. aquitilis, R. sp. |
| 24 Vibrio cholerae pathogenic cycle | Quorum Sensing | luxQ, luxU, qcr, tcpB, K10917, apha, tcpA, tcpH, tcpP | V. cholerae*, V. vulnificus, V. fisheri, V. parahaemolyticus, V. sp. |
| 25 Epithelial cell signaling in Helicobacter pylori infection | Type IV secretion system | cag1, cag2, cag1, cag2, cag2, virB11, vrbB11, cag7-cag25 | H. pylori*, P. rouszewksi, L. pneumophila, R.ickettsii, S. milliotti |
| 26 Epithelial cell signaling in Helicobacter pylori infection | Adhesins | holC, holA, holB, holP, holQ, K15846, hpaA, sdbA, sdbA | H. pylori*, L. pneumophila*, H. cetonatum, H. acinaroxys, H. bizzozeronii |
| 27 Pathogenic Escherichia coli infection | EspG, virA, tir, espF, map, eae, nleA, nleH, espH, tcpP, espG2 | E. coli*, S. flexneri, S. sonnei, C. rodentium, S. boydii |
genes expressed in the strain’s genome. Reflecting this finding, the Table 3 contains 15 gene circuits for which bacterial strains containing at least 75 percent of the circuit elements are always pathogenic. Hence a signature for pathogenicity may be derived from the study of clusters of pathogen-linked orthologs in bacterial strains.

Figure 6 presents other examples of ortholog groupings listed in Table 3 and acting in tandem in pathogenic processes. One such circuitry shown in the figure is involved in the biosynthesis of siderophore group of nonribosomal peptides. These are high affinity iron binding compounds [54] and were found to play an important role in virulent bacterial infections [55]. Also shown in the Figure is pathogen associated ortholog circuit clusters crowding the bacterial secretion system not discussed above in detail. As noted in the literature, the secretion system facilitates transport, injection, and release of effector compounds including enzymes, and toxins in bacteria [56, 57].

Additional pathogen-linked circuitry identified through comparisons of genomes belonging to the same genera: Additional circuits linked to pathogenicity could be identified using within-genera genome comparisons. We have conducted comparisons of ortholog contents of strains belonging to the same genera for the seventeen genera with most number of strains in our dataset, shown in Table 1. Again, the clusters of pathogen-linked orthogs forming on KEGG reference pathways were identified. However, in this case, we relaxed the pathogen-linkage evaluation from PA > 4 to PA > 2 since genomes belonging to the same genera are more or less similar. In addition, we are looking here for circuitry common across genera.

Results of these computations are presented in Additional file 5, identified in rows 1 to 21 for pathogen-linked circuits. The table shows not only the circuitry but also the genera associated with the specified circuitry. The circuit clusters most common across genera in this Table lie in the pathways for glycine, serine and threonine metabolism and sulfur metabolism (Fig. 7). These pathways have been implicated in playing important roles in pathogenicity [58–61]. Also in this category, is the gene circuit in Additional file 5 row 1 linked to Alzheimer’s disease via Amyloid B and Mitochondrial Dysfunction.

Another pathogen-linked gene circuitry that comes out in genera-specific comparisons is the Amyloid B and Mitochondrial Dysfunction circuitry in the KEGG Alzheimer’s pathway (Additional file 5). The pathogen-linked orthologs in this circuitry (UQCRCFSI, RIP1, petA, MME, IDE, ide, CALM, NDUFV2) are found in 16 of the 17 genera under consideration. This observation suggests the diversity of a bacterial infection that could be linked as a possible modulator of the Alzheimer’s disease [62–64].

**Table 3** Gene circuits linked to pathogenic phenotype. Gene symbols in the table indicate orthologs in the circuit, which are abundant in pathogenic but rarely found in nonpathogenic strains (Continued)

| 28 | Shigellosis | Intimate adhesion/ Type III secretion system | icsA, virG, espG, virA, icsB, bopA | E. coli, B. pseudomallei, B. mallei, S. flexneri, S. sonnei |
| 29 | Shigellosis | Type III Secretion/ Downstream | ipgB1-2, ipaH, espE-G, espG, mkaD, sipA-D, ipaA-D, bipB-D, ipgD, sopB | S. enterica*, E. col*, B. pseudomallei*, B. mallei*, E. tarda |
| 30 | Salmonella infection | Type III secretion system/SP-1 encoded | sipB, ipaB, bipB, sopE, ipgD, sopB, sipA, ipaA, sipD, yopJ | S. enterica*, Y. pestis, B. pseudomallei, E. col*, B. mallei* |
| 31 | Pertussis | Type IV secretion system | ptxD, ptxB, ptxE, ptxC, ptxA | S. enterica*, B. pertussis, Y. enterocolitica, B. bronchiseptica, B. parapertussis |
| 32 | Legionellosis | Adhesion/Cell entry | sdeA, laiA, rtxA, rtxB, ehNC, ipnE | L. pneumophila*, C. burnetii, L. longbeachae, L. pneumophila*, L. longbeachae, |
| 33 | Legionellosis | Dot/icm secretion system | ralF, lidA, sidC, legK1, lgtl_2_3, sidLM, FleP8, vipA,DE, sidBA, sidC, dnsA, lnxK | S. aureus*, C. pseudotuberculosis, S. epidermidis, S. pseudintermedius, E. casseliflavus |
| 34 | Staphylococcus aureus infection | Colonization/ MSCRAMMs | clfB, isdA, sdrC_D_E, sasG | S. aureus*, S. lugdunensis*, S. pseudintermedius, B. thuringiensis, S. carnosus |
| 35 | Staphylococcus aureus infection | Surface proteins | spa, sbl, clfA, sak, scn, scin, fib, efB, chap, chips, flr, filp | The symbol * identifies those genera for which the circuit shown was also linked to pathogenicity via within-genus comparison |
Table 4 Nonpathogen-linked gene circuits in bacterial strains. The columns identify pathways, circuits, nonpathogen-linked orthologs within the circuit, and genera expressing the circuit.

| Pathway                          | Circuit                                      | Orthologs                                                                 | Genera                                                                 |
|----------------------------------|----------------------------------------------|---------------------------------------------------------------------------|------------------------------------------------------------------------|
| 1 Steroid biosynthesis           | Squalene to 24,25-Dihydro-lanosterol          | SQLE, ERG1, E5.4.99.7, LSS, ERG7, DHC2R4                                   | C. coralloides*, F. taffensis, M. capsulatus*, M. alcaliphilum, S. aurantia |
| 2 Arginine and proline metabolism| N-Acetyl-glutamate - N-Acetyl-ornithine       | argB, ARG56, argC, lysY, ARG56, E2.6.1.11, argD, argD, lysJ                | E. coli*, S. enterica*, S. aureus*, L. monocytogenes, M. tuberculosis*  |
| 3 Biosynthesis of 12-, 14- and 16-membered macrolides | Propanoyl-CoA to Erthromycin A | E2.3.1.94, eryF, eryBV, eryCII, eryClI, eryG, eryK | F. alni, S. erythrea, |
| 4 Lysine biosynthesis           | L-2-Amino adipate to Pyrolysin              | lysX, lysZ, lysY, lysJ, E2.4.1.173, pylB, pylC, pylD          | T. thermophilus*, P. mucilaginosus, C. sp., D. hafniense, M. ruber   |
| 5 Glycine, serine and threonine metabolism | Choline - Glycine                           | codA, gbsB, BHMT, DMGKD, SARDH                                          | B. amylophilus, B. subtilis, S. melloti, R. sphaeroides, S. licheniformis |
| 6 Arginine and proline metabolism | Creatine Pathway                            | GAMT, E2.7.3.2, E3.5.3.3, E3.5.1.99, hyuA, hyuB, E3.5.4.21            | H. pylori*, P. putida, C. sp., A. mediterranei, G. sp.                  |
| 7 Type I polyketide structures   | Rifamycin B                                 | rifA, rif14, rifB, rif20, rifC_D, rifE, rifF, asm9, rif19               | A. mediterranei*, F. sp., A. mirum, M. aurantia, M. sp.                 |
| 8 Biosynthesis of 12-, 14- and 16-membered macrolides | 2-Methylbutanoyl to Avermectin | aveA, aveE, aveF, aveD, aveBl                                               | S. avermitilis.                                                   |
| 9 Type I polyketide structures   | Erythromycin A                              | eryK, E2.3.1.94, eryF, E2.3.1.94, eryCII, eryClI, eryG, eryK              | F. alni, S. erythrea.                                                   |
| 10 Type I polyketide structures  | Avermectin A1a                              | aveA, aveBl, aveE, aveF, aveD                                            | S. avermitilis.                                                   |
| 11 Type I polyketide structures  | Myalamid S                                  | mxaB, mxaE_D, mxaF                                                         | G. violaceus, H. ochraceum, H. aurantiacus, M. xanthus, N. punctiforme |
| 12 Biosynthesis of type II polyketide backbone | 7,9,12-Octaketides                          | actI, actI2, actI3, actII, actVII                                      | A. mediterranei*, F. sp., A. missouriensis, C. acidiphila*, C. epipsammum |
| 13 Biosynthesis of type II polyketide products | 7,9,12-Octaketides - Actinorhodin           | actVI1, RED1, actVI3, actVI2, actIV, actVI4, actVA6, actVA4, actV, actVB | A. mediterranei*, F. sp., M. abscissus, S. cattleya, A. ferrooxidans |
| 14 Serotonergic synapse          | Signal transduction                         | PLA2G4, CPLA2, ALOX5, PTGS2, COX2, PTGS2, COX2                         | R. sphaeroides*, C. coraloides, G. obscurus, M. nodulans, M. sp.        |
| 15 Insulin signaling pathway     | Glycogenesis/antipolysos                    | GYS, CALM, PHKA_B, PRKAR                                                  | S. sp., C. sp., A. sp., B. coagulans, C. sp.                          |
| 16 Chemical carcinogenesis       | Azo dyes/Liver cancer/Bladder cancer        | PTGS2, COX2, CYP1A1                                                      | R. sphaeroides*, G. obscurus, M. mediterranea, M. nodulans, M. sp.     |

The symbol * indicates those genera for which the circuitry was linked to synergy via within genus comparison.

within genera comparisons. Our detailed results shown in Table 4 and Fig. 8 are summarized below.

Antibiotics and metabolite producing circuits:

Presence of clusters of nonpathogen-linked orthologs in metabolic circuits such as steroid biosynthesis, arginine and proline metabolism, and the Insulin signaling pathway indicate the importance of these pathways in establishing synergy with the host in all the major genera considered (Table 4). Some of the orthologs in the nonpathogen-linked bacterial gene circuits have orthologs in the human. Other synergy circuits in Table 4 are involved in radiation survival [65].

KEGG reference metabolic pathways contain large numbers of nodes creating thousands of clusters for bacterial species, and hence the relative lack of literature for some of the clusters shown in Table 4.

The polyketide circuit shown in Fig. 8 and presented in Table 4 facilitates the synthesis of common antibiotics [68]. Polyketides are complex organic compounds, which are highly active biologically. Many pharmaceuticals are derived from or inspired by polyketides. In addition to the polyketide circuits, the circuit shown in Fig. 8 composed of scyllo-inosamine orthologs (stsE, strB1, E2.4.2.27, strK, K12570, aphD, strA) is involved in the biosynthesis of streptomycin and similar anti-mycobacteria antibiotics. Recent studies show bacterial virulence factors in type 3-secretion pathway as targeted by virulence inhibitors such as those illustrated in Fig. 8 [69]. Also shown in the figure is the one-carbon pool by folate pathway, activating one-carbon units for biosynthesis [70]. It plays a major role in amino acid metabolism [71]. It has been shown to affect proof reading of DNA replication, DNA
Fig. 6 Examples of gene circuitry containing pathogen-linked ortholog clusters in KEGG reference pathways. Orthologs with $PA > 4$ but not present in VFDB were shaded in pink whereas orthologs with $PA > 4$ and also in VFDB in orange. The numbers indicating specific circuitry correspond to their identification numbers in Table 3.
methylation, and chromatin structure [72–74]. The list for commensal circuitry presented in Table 4 is not complete, but representative of the diversity of commensal circuits found in bacterial strains. Genera-specific genome comparisons reveal additional circuitry clusters found almost exclusively in nonpathogenic bacterial strains. Shown in Additional file 5 in rows 22 to 24 are clusters for benzoate degradation, and the cluster for dopamine circuitry in Isoquinoline alkaloid biosynthesis. Benzoate degradation is an important factor in reducing drug-induced toxicity [66]. It is not clear how dopamine inducing bacterial gene circuits drive synergy with the host, yet modulations in dopamine circuitry in bacteria was previously linked to Alzheimer’s disease stage progressions via Borrelia infection [67].

Discussion

Pathogenicity is context dependent. A survey of the literature indicates many dimensions of complexity in defining and investigating pathogenic behavior of bacterial strains and microbiomes in relation to host [75–78]. Consider a microbiome composed of thousands of bacterial strains. The combined genome is large, containing millions of genes. It is clear in this scenario the need for obtaining a signature for pathogenic or commensal phenotypes for dimension reduction. This study begins with such a task, namely deducing a list of genes which are highly abundant
in pathogenic and relatively absent in commensal strains and vice versa. Two distinct datasets were used in our analysis: (a) list of bacterial strains observed to be pathogenic at least in a context dependent manner and (b) bacterial genomes annotated with ortholog genes. The literature search yielded 949 decoded bacterial genomes deemed as pathogenic and another 1578 as nonpathogenic or commensal due to lack of evidence in the literature for pathogenicity. Additional file 1, presenting the list of pathogenic strains along with literature-curated evidence for each, will prove to be a useful resource for microbiology.

The ortholog contents of genomes in KEGG Orthology (KO) Database were recently used in developing an algorithm (PICRUSt) for analysis of genes in microbiomes in health and disease [79]. In the present case, KO database yielded two large matrices with columns identifying pathogenic or nonpathogenic strains and rows representing 7194 orthologs present in at least one bacterial strain. We created similar matrices for within genera comparison of pathogenic and nonpathogenic strains. From this point on, we could compare the two matrices and identify, for each ortholog, the relative abundance in pathogenic and nonpathogenic strains. Some of the orthologs associated in the literature with antibiotic resistance [80] were indeed preferentially abundant in antibiotic resistant

Fig. 8 Examples of gene circuitry containing nonpathogen-linked ortholog clusters in KEGG reference pathways. Orthologs with PA < 1/4 are shaded in pink. The numbers indicating specific circuitry correspond to their identification numbers in Table 4.
strains. Others were not but they co-localized with orthologs highly enriched in antibiotic resistant strains in cellular pathways. The pathways enriched by orthologs abundant in antibiotic resistant strains included P13K-Akt signaling pathway of eukaryotic hosts. The effector bacterial orthologs in this pathway divert host cell signaling pathways to the benefit of the pathogen and target kinase signaling cascades present in P13K-Akt [81], resulting in pathogenic infections [82]. The other most enriched pathway in antibiotics resistant bacterial strains was peptidoglycan biosynthesis, previously linked to biofilm production [83]. The pathway mediating the metabolism of Sphingolipids, a class of lipids, was occupied by orthologs found in antibiotic resistant bacteria. Sphingolipids play significant roles in membrane biology and provide many bioactive metabolites that regulate cell function [84]. It has already been linked in the literature to antibiotic resistance [85, 86].

Pathway enrichment protocols provided us with an overall portrait of ortholog sets linked either to commensal or pathogenic behavior to the host. Additional annotation was needed however in order to get a signature for synergy or pathogenicity. Visually, when we projected pathogen- or synergy-linked ortholog lists onto KEGG pathways, we could see orthologs from our lists forming multiply connected clusters along the bacterial pathways. Thus, our next task was identifying circuits along the pathways that contained such clusters and then to conduct extensive literature searches to annotate our discoveries.

The circuits identified included toxin-producing units for pathogenic strains and the gene circuits synthesizing antibiotics for commensal strains. We observed that the list of circuitry linked to synergy contained among others, biosynthesis modules of type II polyketide products, such as Erythromycin, and Doxycycline. Other examples of nonpathogen-linked ortholog clusters in gene circuits could be found in VEGF signaling pathway and in metabolic pathways.

The ortholog modules highly abundant in pathogenic bacteria included gene circuits found in bacterial secretion mechanism type III. Also in this category were gene circuits involved in peptidoglycan biosynthesis. In general, pathogen abundant orthologs were statistically enriched in KEGG pathways for pathogen interactions and in signaling pathways. Comparison of pathogenic strains with nonpathogenic strains of the same genera identified one carbon pool by folate pathway as highly abundant in pathogenic strains. This pathway mediates the activation of one-carbon units in the biosynthesis and metabolism of amino acids. It affects DNA methylation as well as DNA replication. Recent research implicates it in folic acid-mediated degeneration of the brain–blood barrier of the host [87]. In addition, our system approach identified a bacterial gene circuit whose genes are preferentially found in pathogenic strains and linked in the literature to Alzheimer’s disease. We find the gene clusters in this circuit to be abundant in a diverse set of genera, possibly providing new avenues of research for linkage between bacterial infections and Alzheimer’s disease.

The work presented here is a first draft of an encyclopedia for bacteria – host interactions. The study uncovers identity of ortholog clusters as possible signatures of pathogenicity or synergy in a mixture of bacteria. We show that gene clusters formed on bacterial pathways to be much stronger determinants of pathogenicity than a list of virulence and/or anti-virulence factors. The signature we derive in the form of gene circuits may not be complete as our results relied on the decoded genomes of bacterial strains currently available in the literature. Moreover, our reliance on the KEGG reference pathways in identifying clusters of orthologs preferentially found in pathogenic and nonpathogenic bacteria introduces additional limitations due to incompleteness of the KEGG pathway models. Nevertheless, the gene circuitry signature we discovered for synergy and pathogenicity is comprehensive enough to derive from it biomarker clusters identifying pathogenic phenotypes in bacterial strains isolated or in a mixture.
Methods

Annotation of bacterial strains exhibiting pathogenicity

The list of pathogenic bacterial strains was obtained via literature search using three complementary and partially redundant approaches. First, we text mined the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [91, 92] for pathogenicity label for all decoded bacterial strains in KEGG. Second, we identified as pathogenic those bacterial strains with pathogenic citation in at least one of the following web tools: Virulence Factor Database (VFDB) [2, 6, 9], High-quality Automated and Manual Annotation of Proteins (HAMAP) [93], and the Interactive Atlas for Exploring Bacterial Genomes (BacMap) [94]. Third, we text mined PUBMED article abstracts with a) the name of the bacterial strain and b) one of the pathogen-related code words (pathogen, virulence, pathogenic, virulent). All hits were then verified for pathogenicity by reading the articles. We labeled bacterial strains as nonpathogenic if the process described above did not yield evidence for pathogenesis. We have used the Antibiotic Resistance Database (ABRD) [25] to identify antibiotic resistant bacterial strains within the set of pathogenic bacterial strains.

Orthology content of decoded bacterial strains

KEGG programming interface was used to obtain gene orthology information for the bacterial organisms found in KEGG. The individual organism orthology content was translated into a logical content vector that describes the orthology information for that particular organism in binary form, “1” meaning that the ortholog is present and “0” meaning it is absent, with respect to all collective orthologs found in the KEGG repository. The individual feature vectors were than accumulated into a logic matrix, with rows representing the presence/absence of orthologs for each decoded bacterial strain. The resulting array contained 2527 sequenced bacterial strain genomes potentially expressing 7194 orthologs. The number of genes that were accounted for in this array totaled to over 3.2 million. The logical array was then split into pathogenic vs. non-pathogenic logical arrays with equal number of columns representing the orthologs.

Abundance scores for orthologs in bacterial strain genomes

An abundance score for each ortholog was generated for both the pathogenic and non-pathogenic bacterial strain arrays based on the sum of all elements in the ortholog column vector divided by vector length (number of bacterial strains). These scores were denoted as Ap and Anp for pathogenic and non-pathogenic strains, respectively. They represent the fraction of genomes expressing the ortholog in pathogenic and non-pathogenic logical matrices. Next we defined a pathogenicity abundance score as $PA = Ap / (Anp + 0.0001)$. If Anp turned out to be “0” for that ortholog, meaning that ortholog was absent in non-pathogenic bacteria, the equation yields PA as equal to $Ap / 10^{-4}$. The list of orthologs could then be ordered with respect to the PA value, creating a histogram.

The comparisons of contents of genome orthology of bacterial strains belonging to the same genera are identical to the one for all the 2527 content comparison described above. For this operation, we used only strains belonging to the genera shown in Table 2, one genus at a time. Also, in this case we reduced the cutoffs for pathogen association and defined orthologs as pathogen-abundant for $PA > 2$, and nonpathogen abundant for $PA < 1/2$. This was necessitated by the relative similarity of genomes within a genus. However, to compensate, we studied only those clusters in KEGG Reference pathways abundant in pathogenic or nonpathogenic strains in multiple genera.

Statistical enrichment of cellular pathways with pathogenic and nonpathogen-abundant orthologs

We identified orthologs as pathogen exclusive if they were expressed in some of the pathogenic strains but not at all in nonpathogenic strains. Nonpathogen exclusive orthologs were similarly defined. We defined orthologs as pathogen-abundant for $PA > 4$, and nonpathogen abundant for $PA < 1/4$. These were also called pathogenic orthologs and vice versa.

For KEGG cellular pathway enrichments, we created a score matrix, similar to that described in the making of the orthology database. Next, the enrichment was carried out via hypergeometric test [63, 95] using total number of orthologs in bacterial strains as the population size, pathogenic orthologs as the number of success states in the population; number of orthologs in the pathways as the number of draws; and the number of pathogenic orthologs in the pathways as the number of successes in the draw.

Subpathways containing pathogenic ortholog circuits were identified via manual KEGG orthology mapping and screening for clustering. Cellular pathways were drawn using the KEGG web tool for pathways. Although the cutoff values for PA for generating pathogenic and nonpathogenic ortholog lists appear arbitrary, they capture the tails of the PA distribution for orthologs in the thousands of decoded genomes under consideration. Preliminary studies involving the perturbation of the cutoff did not yield variation in the enriched pathways.

Ortholog clusters relevant to pathogenicity in KEGG reference cellular pathways

For this purpose, we mapped the list of pathogen-abundant and nonpathogen abundant orthologs to all available KEGG Reference pathways, and manually curated the clusters (subcircuitry) they form. The criterion for cluster was the
minimum of three orthologs from the list to be multiply-connected with each other and have connections to other neighboring orthologs. Results shown in Table 3 to Additional file 5 indicate only the gene symbols of those abundant in pathogenic or vice versa. As presented in these tables, most clusters contained many more orthologs then three. However, the Reference pathways typically contain genes not found in all strains and hence in almost all cases we studied 75% presence of pathogene-abundant orthologs in a strain as indicating that the strain was deemed pathogenic in the literature. The p values listed in these tables were obtained using hypergeometric test based on the following scheme: a) there are 949 strains deemed pathogenic in a pool of 2527 strains, and b) among M strains, m cluster orthologs were found only on N strains. In this set up, M, N, and m are positive integers.

Additional files

Additional file 1: List of 949 decoded bacterial genomes present in the KEGG web platform deemed as pathogenic along with literature citations for pathogenicity and antibiotic resistance phenotype. (XLSX 51 kb)

Additional file 2: List of 1578 decoded bacterial genomes present in the KEGG web platform with no present citation implicating them as pathogenic. (XLSX 54 kb)

Additional file 3: Bacterial strain orthology logic array indicating the presence of given orthologs (7194 of them) in the 2527 bacterial strain genomes considered in this study. (RAR 1397 kb)

Additional file 4: List of 7194 orthologs ordered according to pathogen association scorePA. The orthologs shaded in pink represent ones present in VFDB. (XLSX 570 kb)

Additional file 5: Pathogen- and nonpathogen gene circuits identified via within genera comparisons. The genera exhibiting a pathogen-linked circuit are painted in pink whereas nonpathogen-linked circuits are painted in green.

Abbreviations

AROB: Antibiotics resistance genes database; KEGG: Kyoto encyclopedia of genes and genomes; PA: Pathogenicity index; VEGF: Vascular endothelial growth factor; VFDB: Virulence factor database.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

AT and MS conceptualized and designed the study. MS wrote the algorithms for the study and implemented the research with continuous input and discussions with AT and SO. AT and MS co-wrote the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

Not applicable.

Author’s Information

Not applicable.

Acknowledgements

MS was supported by the US Department of Education Doctoral Training Fellowship Program (GAANN), entitled Bioinformatics of Infectious Diseases. AT is the PI and the Director of the GAANN project.

Ethics

This study utilized open access data, exclusively. No human or animal subjects were involved. Supporting files provide all the information required to reproduce our results.

Author details

1School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, USA. 2College of Computing and Informatics, Drexel University, Philadelphia, PA, USA.

Received: 8 January 2015 Accepted: 28 September 2015
Published online: 13 October 2015

References

1. Rahme LG et al. Plants and animals share functionally common bacterial virulence factors. Proc Natl Acad Sci U S A. 2000;97(16):8885–21.
2. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2012;40(Database issue):D641–5.
3. Long TM, Nisa S, Donnenberg MS, Hassel BA. Enteropathogenic Escherichia coli inhibits type I interferon- and RNaseL-mediated host defense to disrupt intestinal epithelial cell barrier function. Infect Immun. 2014;82(7):2802–14.
4. Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. MBio. 2014;5(2):e01012–4.
5. Seyedsayamdost MR. High-throughput platform for the discovery of elicitors of silent bacterial gene clusters. Proc Natl Acad Sci U S A. 2014;111(20):7266–71.
6. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(Database issue):D325–8.
7. Pascador L, Cook JM, Gambello MJ, Rust L, Igleswki BH. Expression of Pseudomonas aeruginosa virulence genes requires cell-to-cell communication. Science. 1993;260(11):1127–30.
8. de Kievit TR, Kakai Y, Register JK, Pisci EC, Igleswki BH. Role of the Pseudomonas aeruginosa las and rhl quorum-sensing systems in rhl regulation. FEMS Microbiol Lett. 2002;212(1):101–6.
9. Yang J, Chen L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 2008;36(Database issue):D339–42.
10. Niu C, Yu D, Wang Y, Ren H, Jin Y, Zhou W et al. Common and pathogen-specific virulence factors are different in function and structure. Virulence. 2013;4(6):473–82.
11. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome analysis of a transmissible lineage of pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS Genet. 2013;9(9):e1003741.
12. Koskinen S, Gibbons HS, Sandegren L, Anwar N, Ouellette G, Broomall S et al. Pathoadaptive mutations in Salmonella enterica isolated after serial passage in mice. PLoS One. 2013;8(7):e70147.
13. Prosessa G, Di Martino ML, Campilongo R, Fioravanti R, Michel J, Casalino M et al. Shedding of genes that interfere with the pathogenic lifestyle: the Shigella model. Res Microbiol. 2012;163(6):679–90.
14. Juhász J, Kendés-farkas A, Szabó D, Porog S. Emergence of collective territorial defense in bacterial communities: horizontal gene transfer can stabilize microbiomes. PLoS One. 2014;9(4):e95511.
15. Maurelli AT. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. FEMS Microbiol Lett. 2007;267(1):1–8.
16. Swan BK, Chaffin MD, Martinez-garcia M, Morrison HG, Field EK, Poulton NJ et al. Genomic and metabolic diversity of marine group I thaumarchaeota in the mesopelagic of two subtropical gyres. PLoS One. 2014;9(6):e95380.
17. Yang X, Brisbin J, Yu H, Wang Q, Yin F, Zhang Y et al. Selected lactic Acid-producing bacterial isolates with the capacity to reduce salmonella infection in the mesopelagic of two subtropical gyres. PLoS One. 2014;9(4):e95380.
18. Miskinyte M, Sousa A, Ramiro RS, de Sousa JA, Kotlinowski J, Caramalho I et al. The genetic basis of Escherichia coli pathoadaptation to macrophages. PLoS Pathog. 2013;9(12):e1003802.
19. Podoll JD, Liu Y, Chang L, Walls S, Wang W, Wang X. Bio-inspired synthesis yields a tricyclic indoline that selectively rezesistates methicillin-resistant Staphylococcus aureus (MRSA) to beta-lactam antibiotics. Proc Natl Acad Sci U S A. 2013;110(39):15573–8.

Page 16 of 18

Shestov et al. BMC Genomics (2015) 16:773
20. Uda N, Matoba Y, Kumagai T, Oda K, Noda M, Sugiyama M. Establishment of an in vitro D-cycloserine-synthesizing system by using O-uereido-L-serine synthase and D-cycloserine synthase found in the biosynthetic pathway. Antimicrob Agents Chemother. 2013;57(6):2603–12.

21. Paredes D, Ortz C, Torres R. Synthesis, characterization, and evaluation of antibacterial effect of Ag nanoparticles against Escherichia coli O157:H7 and methicillin-resistant Staphylococcus aureus (MRSA): Int. J. Nanomedicine. 2014;9:1717–29.

22. Ghosh S, Prava J, Samal HB, Saur M, Mahapatra PK. Comparative genomics study for the identification of drug and vaccine targets in Staphylococcus aureus: MurA ligase enzyme as a proposed candidate. J Microbiol Methods. 2014;101:1–8.

23. Machowski EE, Senzari S, Esland C, Kana BD. Comparative genomics for mycobacterial peptidoglycan remodelling enzymes reveals extensive genetic multiplicity. BMC Microbiol. 2014;4:175.

24. Wichmann F, Udzikovic-kolic N, Andrew S, Handelsman J. Diverse antibiotic resistance genes in dairy cow manure. MBio. 2014;5(2):e01017.

25. Liu B, Pop M. ARDB–Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009;37(Database issue):D443–7.

26. Gouzeau-anderson V, Cabot MC. Sphingolipid metabolism and drug resistance in hematological malignancies. Anticancer Agents Med Chem. 2011;11(9):891–903.

27. West KA, Castillo SS, Dennis PA. Activation of the P2X7R pathway and chemotherapeutic resistance. Drug Resist Updat. 2002;5(6):234–48.

28. Hurdle JG, Omenn AJ, Chopra I. Prospects for aminoglycosid-β-nucleotidase inhibitors as a new antimalarial antibiotic. Antimicrob Agents Chemother. 2005;49(12):5781–3.

29. Macho AP, Zipfel C. Targeting of plant pattern recognition receptor-triggered immunity by bacterial type-III secretion system effectors. Curr Opin Microbiol. 2014;23:C14–22.

30. Dai W, Li Z. Conserved type III secretion system exerts important roles in Chlamydia trachomatis. Intr. J Clin Exp Pathol. 2014;7(9):5044–4.

31. Patel S, McCormick BA. Mucosal Inflammatory Response to Salmonella typhimurium infection. Front Immunol. 2014;5:311.

32. Galán JE, Lara-tejero M, Marlovits TC, Wagner S. Bacterial type III secretion systems: specialized nanomachines for protein delivery into target cells. Annu Rev Microbiol. 2014;68:415–38.

33. Christie PJ, Whitaker N, Gonzalez-ivea C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843(8):1578–91.

34. Wang Y, Wei X, Hao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.

35. Cabezon E, Ripoll-rozada J, Peña A, De La Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2014;38:145–66.

36. Stephenson K, Hoch JA. Two-component and phosphorelay signal-transduction pathways. Annu Rev Microbiol. 2014;68:1–23.

37. Galán JE, Lara-tejero M, Marlovits TC, Wagner S. Bacterial type III secretion systems: specialized nanomachines for protein delivery into target cells. Annu Rev Microbiol. 2014;68:415–38.

38. Christie PJ, Whitaker N, Gonzalez-ivea C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843(8):1578–91.

39. Wang Y, Wei X, Hao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.

40. Cabezon E, Ripoll-rozada J, Peña A, De La Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2014;38:145–66.

41. Saiyi F, Asif A, Shah G, Iqbal A, Khaliq A, Ahmad W et al. Bacterial type IV secretion system effectors: specialized nanomachines for protein delivery into target cells. Annu Rev Microbiol. 2014;68:415–38.

42. Christe PJ, Whitaker N, Gonzalez-ivea C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843(8):1578–91.

43. Wang Y, Wei X, Hao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.

44. Cabezon E, Ripoll-rozada J, Peña A, De La Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2014;38:145–66.

45. Saiyi F, Asif A, Shah G, Iqbal A, Khaliq A, Ahmad W et al. Bacterial type IV secretion system effectors: specialized nanomachines for protein delivery into target cells. Annu Rev Microbiol. 2014;68:415–38.

46. Christe PJ, Whitaker N, Gonzalez-ivea C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843(8):1578–91.

47. Wang Y, Wei X, Hao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.

48. Cabezon E, Ripoll-rozada J, Peña A, De La Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2014;38:145–66.

49. Saiyi F, Asif A, Shah G, Iqbal A, Khaliq A, Ahmad W et al. Bacterial type IV secretion system effectors: specialized nanomachines for protein delivery into target cells. Annu Rev Microbiol. 2014;68:415–38.

50. Christe PJ, Whitaker N, Gonzalez-ivea C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843(8):1578–91.

51. Wang Y, Wei X, Hao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50.
72. Kamen B. Folate and antifolate pharmacology. Semin Oncol. 1997;24(Suppl 18):S18–30. S18-39.
73. Crider KS, Yang TP, Berry RJ, Bailey LB. Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate’s role. Adv Nutr. 2012;3(1):21–38.
74. Zhou HR, Zhang FF, Ma ZY, Huang HW, Jiang L, Cai T et al. Folate polyglutamylation is involved in chromatin silencing by maintaining global DNA methylation and histone H3K9 dimethylation in Arabidopsis. Plant Cell. 2013;25(7):2545–59.
75. Lijek RS, Weiser JN. Co-infection subverts mucosal immunity in the upper respiratory tract. Current Opin Immunol. 2012;24:417–23.
76. Branchu P, Matrat S, Vareille M, et al. NsrR, GadE, and GadX interplay in repressing expression of the Escherichia coli O157:H7 LEE pathogenicity island in response to nitric oxide. PLoS Pathog. 2014;10(1):e1003874.
77. Marullo S, Coureuil M. Arrestins in host-pathogen interactions. Handb Exp Pharmacol. 2014;219:361–74.
78. Tack AJ, Horns F, Laine AL. The impact of spatial scale and habitat configuration on patterns of trait variation and local adaptation in a wild plant parasite. Evolution. 2014;68(1):76–89.
79. Langille MG, Zaneveld J, McDonald D, Knights D, Reyes JA et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.
80. Nesme J, Cécillon S, Delmont TO, Monier JM, Vogel TM, Simonet P. Large-scale metagenomic-based study of antibiotic resistance in the environment. Current Biol. 2014;24:1–5.
81. Krachler AM, Woolery AR, Orth K. Manipulation of kinase signaling by bacterial pathogens. J Cell Biol. 2011;195(7):1083–92.
82. Oviedo-boyso J, Cortés-vieyra R, Huante-mendoza A, Yu HB, Valdez-Alarcón JJ, Bravo-Patio A et al. The phosphoinositide-3-kinase-Akt signaling pathway is important for Staphylococcus aureus internalization by endothelial cells. Infect Immun. 2011;79(1):4569–77.
83. Li L, Zhu J, Yang K, Xu Z, Liu Z, Zhou R. Changes in gene expression of Actinobacillus pleuropneumoniae in response to anaerobic stress reveal induction of central metabolism and biofilm formation. J Microbiol. 2014;52(6):673–81.
84. Gault CR, Obeid LM, Hannun YA. An overview of sphingolipid metabolism: from synthesis to breakdown. Adv Exp Med Biol. 2010;688:176–89.
85. Morad SA, Cabot MC. Tamoxifen regulation of sphingolipid metabolism—Therapeutic implications. Biochim Biophys Acta. 2015;1851(9):1134–45.
86. Hill JA, Ammar R, Torti D, Nislow C, Cowen LE. Genetic and genomic architecture of the evolution of resistance to antifungal drug combinations. PLoS Genet. 2013;9(4):e1003390.
87. Kolani A, Kamar PK, Givirmani S, Brown K, Metreveli N, Tyagi SC et al. Nutri-epigenetics ameliorates blood-brain barrier damage and neurodegeneration in hyperhomocysteinemia: role of folic acid. J Mol Neurosci. 2014;52(2):202–15.
88. Freeman S, Bhatt A, Pedamallu C, King S, Duke F, Jung J et al. Al21: in search of infectious triggers of periodic fever, aphthous stomatitis, pharyngitis and adenitis syndrome. Arthritis Rheumatol. 2014;66(11):S158.
89. Xu G, Strong MJ, Lacey MR, Barbaud C, Perimington EK, Taylor CM. RNA CoMPASS: a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. PLoS One. 2014;9(2):e89445.
90. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A. 2013;110(30):12450–5.
91. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
92. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42:D199–205.
93. Pedrazzoli R, Rivoire C, Aucuncloss AH, Coudert E, Keller G, de Castro E et al. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 2013;41(Database issue):D584–9.
94. Cruz J, Liu Y, Liang Y, Zhou Y, Wilson M, Dennis JJ, et al. BacMap: an up-to-date electronic atlas of annotated bacterial genomes. Nucleic Acids Res. 2012;40(Database issue):D599–604.
95. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 2007;23(4):401–7.