HAL
open science

Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans.

Gregorio Iraola, Gustavo Vazquez, Lucía Spangenberg, Hugo Naya

To cite this version:

Gregorio Iraola, Gustavo Vazquez, Lucía Spangenberg, Hugo Naya. Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans.. PLoS ONE, 2012, 7 (8), pp.e42144. 10.1371/journal.pone.0042144 . pasteur-00839680

HAL Id: pasteur-00839680

https://riip.hal.science/pasteur-00839680

Submitted on 28 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reduced Set of Virulence Genes Allows High Accuracy Prediction of Bacterial Pathogenicity in Humans

Gregorio Iraola[1,2], Gustavo Vazquez[3], Lucía Spangenberg[1], Hugo Naya[1,4]*

1 Unidad de Bioinformática, Institut Pasteur Montevideo, Montevideo, Uruguay, **2** Sección Genética Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay, **3** Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Bahía Blanca, Argentina, **4** Departamento de Producción Animal y Pasturas, Facultad de Agronomía, Universidad de la República, Montevideo, Uruguay

Abstract

Although there have been great advances in understanding bacterial pathogenesis, there is still a lack of integrative information about what makes a bacterium a human pathogen. The advent of high-throughput sequencing technologies has dramatically increased the amount of completed bacterial genomes, for both known human pathogenic and non-pathogenic strains; this information is now available to investigate genetic features that determine pathogenic phenotypes in bacteria. In this work we determined presence/absence patterns of 814 different virulence-related genes among more than 600 finished bacterial genomes from both human pathogenic and non-pathogenic strains, belonging to different taxonomic groups (i.e: *Actinobacteria*, *Gammaproteobacteria*, *Firmicutes*, etc.). An accuracy of 95% using a cross-fold validation scheme with in-fold feature selection is obtained when classifying human pathogens and non-pathogens. A reduced subset of highly informative genes (120) is presented and applied to an external validation set. The statistical model was implemented in the BacFier v1.0 software (freely available at *http://bacfier.googlecode.com/files/Bacfier_v1_0.zip*), that displays not only the prediction (pathogen/non-pathogen) and an associated probability for pathogenicity, but also the presence/absence vector for the analyzed genes, so it is possible to decipher the subset of virulence genes responsible for the classification on the analyzed genome. Furthermore, we discuss the biological relevance for bacterial pathogenesis of the core set of genes, corresponding to eight functional categories, all with evident and documented association with the phenotypes of interest. Also, we analyze which functional categories of virulence genes were more distinctive for pathogenicity in each taxonomic group, which seems to be a completely new kind of information and could lead to important evolutionary conclusions.

Citation: Iraola G, Vazquez G, Spangenberg L, Naya H (2012) Reduced Set of Virulence Genes Allows High Accuracy Prediction of Bacterial Pathogenicity in Humans. PLoS ONE 7(8): e42144. doi:10.1371/journal.pone.0042144

Editor: Ramy K. Aziz, Cairo University, Egypt

Received March 21, 2012; **Accepted** July 2, 2012; **Published** August 6, 2012

Copyright: © 2012 Iraola et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors acknowledge the financial support provided by Agencia Nacional de Investigación e Innovación - Uruguay (PhD. grant to LS), Comisión Sectorial de Investigación Científica - Uruguay (MSc grant to GI) and Consejo Nacional de Investigación Científica y Técnica - Argentina (research grant to GV). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: naya@pasteur.edu.uy

Introduction

Several factors, including globalization and sanitation conditions, have been shaping the world's landscape of infectious diseases over the years. In developed countries, 90 percent of documented infections in hospitalized patients are caused by bacteria. These cases probably show only a small proportion of the actual number of bacterial infections occurring in the entire population, and they usually represent the most severe cases. In developing countries, a variety of bacterial infections often provoke a devastating effect on the inhabitants' health. The World Health Organization (WHO) has estimated that each year, 1.3 million people die of tuberculosis, 0.2 million die of pertussis and 0.1 million die of syphilis. Diarrheal diseases, many of which are of bacterial etiology, are the second leading cause of death in the world (after cardiovascular diseases), killing 2.5 million people annually (WHO, 2008). This scenario evidences that even today, infectious diseases are a permanent threat for human health around the world.

Understanding the biology of the causative agents of these diseases has been a permanent challenge since the beginning of bacteriology. Nowadays, the mechanisms involved in the virulence (defined as the relative capacity of a microbe to cause damage in a host) of pathogenic bacteria are widely studied in clinical bacteriology, but the advent of new technologies has enabled their study from different perspectives. In this context, bacterial genomics have greatly contributed to the better understanding of pathogenicity due to the possibility of generating and comparing whole genome sequences. The onset of this discipline started with the automation of Sanger sequencing chemistry and the completion of *Haemophilus influenzae* and *Mycoplasma genitallium* genomes [1,2] in the mid-1990 s; since then, projects to sequence the genomes of a large number of organisms were undertaken by means of this method [3–5]. However, during the last decade, to cover the increasing sequencing demands, new non-Sanger high-throughput sequencing systems have been developed under the name of "second generation" or "next-generation" sequencing technologies [6,7]. These developments have significantly reduced the cost and simultaneously increased the speed of DNA sequencing. In this sense, the great majority of organisms whose genomes have been sequenced so far are bacteria, with 1505 complete and published genome sequences and 6037 ongoing

projects (http://www.genomesonline.org/cgi−bin/GOLD/bin/gold.cgi).

Comparative genomics, including comparison at the DNA, transcriptome, and proteome levels, have emerged as a key to give a biological sense to all this massive information. Focused on improving the knowledge on pathogenicity determinants two bioinformatic approaches have been used, based on two complementary explanations for bacterial pathogenesis. On the one hand, pathogenicity has been related to amino acid substitutions which lead to modified protein structures, and probably modified functions [8–10]. In this case, a particular gene shared by a human pathogenic species and a non-pathogenic species, could be causing a pathogenic phenotype in the first one, determined by non-synonymous mutations that modify key aminoacids and alter protein function. Based on this, our group has recently published a method that detects variable regions inside protein sequences which can be potentially related to pathogenicity [11].

On the other hand, trying to give an integrative view of bacterial pathogenicity prediction from a bioinformatic's perspective, in this work we exploit an alternative explanation for bacterial pathogenicity. Pathogenicity has been attributed to the presence or absence of genes which confer particular pathogenic phenotypes, like toxins [12]. In this case, these genes would be present in pathogenic species but absent in non-pathogenic ones. The most widely spread approach to evaluate this is the pairwise comparison between genomes of pathogenic and non-pathogenic bacteria or even multiple comparisons between different strains of the same species [13–15]. These kinds of approaches can give information regarding the presence or absence of genes involved in pathogenicity of a particular species or even a genus. However, it is difficult to extrapolate this information to higher taxonomic levels, which keeps us from drawing conclusions about general features that are determining bacterial pathogenicity.

For this reason, our motivation was: i) try to identify presence/absence patterns of virulence-related genes which could explain the pathogenic phenotype of bacteria at higher taxonomic levels than species or genus, ii) discuss the biological significance of those genes giving an integrative view of genetic determinants of bacterial pathogenicity, iii) use this information to develop a machine learning model to classify bacterial genomes into human pathogens and non-pathogens and iv) implement this model in a software that can be used to predict pathogenicity in the upcoming sequenced bacterial genomes. The last two points are particularly interesting because a statistical model implemented in an easy-to-use software, capable of predicting bacterial pathogenicity based on genomic information, can be helpful for practical purposes. For example, in food or pharmaceutical industries it is essential to know the pathogenic potential of bacterial strains used in bioengineering.

Results and Discussion

The idea that bacterial species can be effectively grouped into human pathogens and non-pathogens based on their virulence-genes composition, arises from preliminary results that indicated differential patterns in presence or absence of these kind of genes among both groups (human pathogens and non-pathogens).

All finished and annotated genomes of human pathogenic and non-pathogenic bacteria were used to perform a presence/absence analysis over 814 groups of orthologous genes belonging to 8 functional categories (toxins, two-component systems, ABC transporters, motility, flagellar assembly, LPS biosynthesis, secretion systems and chemotaxis), in order to determine which ones are strongly related to pathogenicity in different bacterial taxonomic groups (*Actinobacteria*, *Alphaproteobacteria*, *Betaproteobacteria*, *Bacteroidetes/Chlorobi*, *Chlamydiae/Verrucomicrobia*, *Deltaproteobacteria*, *Epsilonproteobacteria*, *Firmicutes*, *Gammaproteobacteria*, *Spirochaetes*, etc.). Figure 1 shows phylogenetic relations and the proportion of pathogenic and non-pathogenic organisms in studied taxa.

The analysis was accomplished by calculating the frequency of genes belonging to each functional category in pathogenic and non-pathogenic species of each taxon. The assumed null hypothesis was that, if a certain gene is not related to pathogenicity, its frequency would not be biased towards pathogenic or non-pathogenic species; furthermore, it would be almost equally distributed within both classes. Genes presenting a high frequency among pathogens and a low frequency in non-pathogens are probably contributing to a pathogen-related phenotype, for example genes coding for toxins. Conversely, a gene that presents low frequency among pathogens and high frequency in non-pathogens could be indicating the loss of genes coding for redundant functions. For example, proteins that transport certain molecules across membranes, which are essential for a free-living style, are often dispensable when bacteria are well-adapted to the environment inside their hosts. The frequency distribution of ABC transporter genes in *Alphaproteobacteria* and *Gammaproteobacteria* clearly exemplifies this situation. Figure 2 shows the frequency of each gene in pathogenic and non-pathogenic organisms. Points falling on the diagonal line represent genes whose frequency is balanced between pathogens and non-pathogens. Points closer to the Y axis are more represented in non-pathogens and points closer to the X axis are more frequent in pathogens. As it is shown in this figure, ABC genes are strongly related to non-pathogenic species in *Alphaproteobacteria*, while there are overrepresented in pathogenic species in *Gammaproteobacteria* (Figure 2).

As shown in Figure 3 the number of present genes is highly variable among classes (pathogens and non-pathogens) and even between taxonomic groups. Moreover, a great number of these present genes, belonging to the 8 functional categories, presented a frequency bias towards either pathogenic or non-pathogenic species (Figure 4), deviating from the proposed null hypothesis. These findings supported the idea that presence/absence patterns of virulence-related genes are informative enough to discriminate between human pathogenic and non-pathogenic bacterial species (Table 1), indicating that this data can be used to construct a classification model based on highly significant biological information.

Classification Model

We used a machine learning approach based on a cross-validation scheme with in-fold feature selection together with a linear Support Vector Machine (SVM) classifier. Preliminary models were constructed using the whole 814 set of genes, but the number of genes was systematically reduced by means of a feature selection process. The definitive model included the first 120 genes ranked by their significance for classification (Table S1). However, since the number of variables is still high, problems associated with chance correlation might arise. For these reason a y-randomization test was implemented. Figure S1 shows the performance obtained in the test (50% accuracy), indicating the absence of chance correlation. Section Model construction further explains these methodologies.

The number of correctly/incorrectly classified genomes in the complete set was 618/30, obtaining an accuracy of 95.4% (Table S2). Table 2 describes the classification performance related to all bacteria taxonomy considered in the dataset. The last column of the table indicates the classification success rate for each group

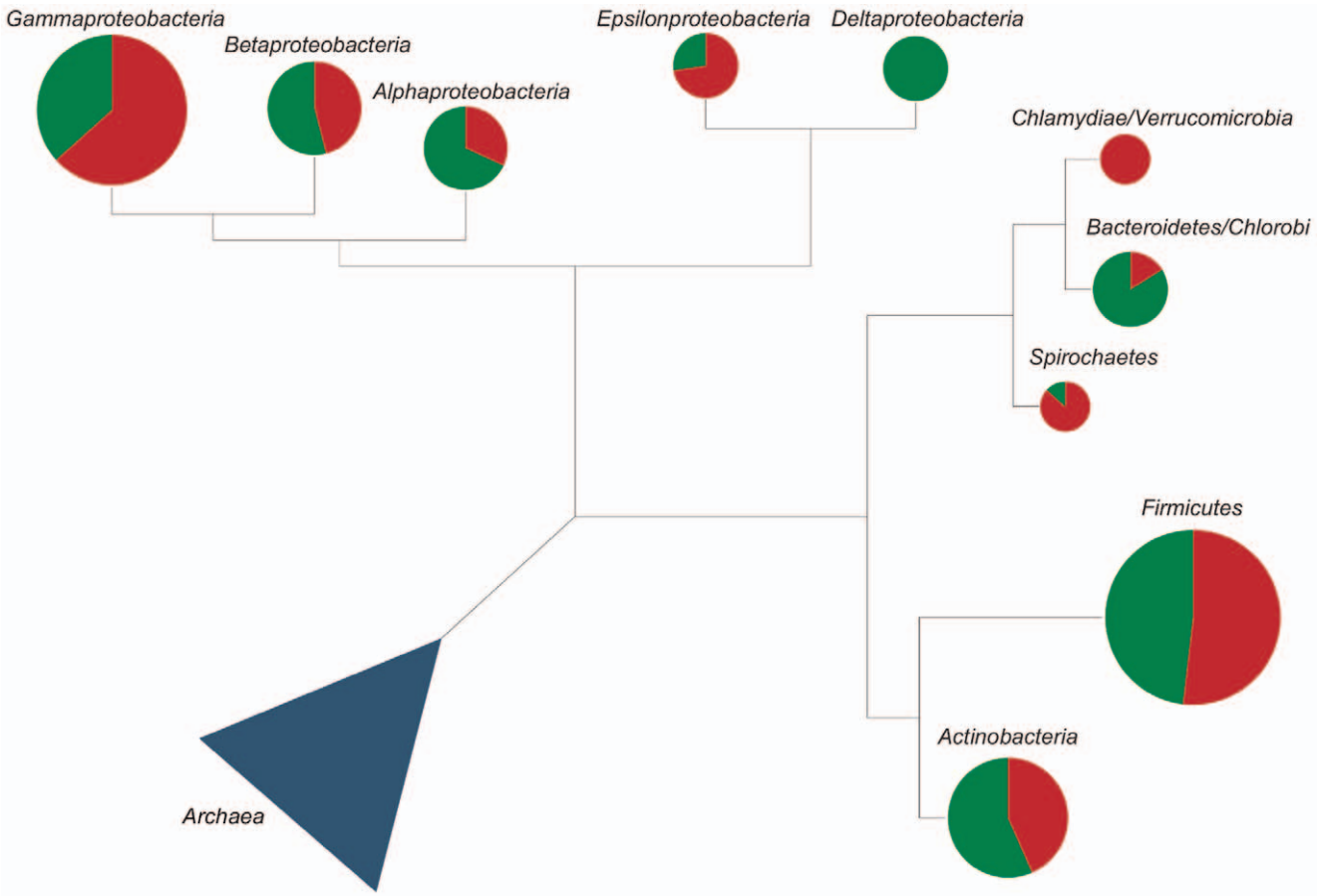

Figure 1. Phylogenetic relations of bacterial groups used in this work. Chart sizes are proportional to the number of genomes present in each taxonomic group. The precentage of pathogenic organisms is shown in red and green is used for non-pathogenic.
doi:10.1371/journal.pone.0042144.g001

considered in the taxonomy; all values were obtained using the 10-fold cross validated SVM model, not by retraining the model using only organisms of the particular taxon. The performance is preserved across the whole taxonomy, ranging from 91% in *Epsilonproteobacteria*, up to 100% in *Bacteroidetes/Chlorobi*. Mid-size groups like *Betaproteobacteria*, *Actinobacteria* and *Alphaproteobacteria* showed a prediction success rate similar or better than the general performance rate. Finally the *Firmicutes*, the biggest group, showed an excellent classification level of 97.4%. Classification performance according to class labels is shown in Table 3, the general error rate is almost equal for false positives and negatives and the general success rate is also equal for pathogens and non-pathogens.

Model Testing and Comparison

To further test the SVM model we evaluated its performance by analyzing genomes originally not included in the dataset used to construct the model. On the one hand, we defined a Group I of 124 genomes with known labels for human pathogen or non-pathogen, originally excluded from the dataset due to reduced number of genomes per group or misrepresentation of one of the two classes. On the other hand, we defined a Group II of 232 "blind" genomes without previous information for pathogenicity.

Group I genomes were classified with an accuracy of 98% (Table 4), even better than the average 95.4% obtained during cross-validation procedure using the original dataset. Only in two taxonomic groups (*Chlamydiae/Verrucomicrobia* and *Fusobacteria*) the model showed an accuracy lower than 100%, and in each case only one genome was misclassified. Group II genomes were previously subjected to an exahustive bibliographic search in order to assign them to human pathogens or non-pathogens (Table S3). Application of SVM model over this group resulted in 92% of average accuracy (Table 4), ranging from 87% in *Epsilonproteobacteria* to 100% in *Deltaproteobacteria*, *Bacteroidetes*, etc. The fact that accuracy is preserved in both test groups reaffirms the results obtained when performing the cross-validation scheme, indicating that our model is robust and the high performance in classification and prediction of human pathogens and non-pathogens is independent of the dataset used to build the model.

The SVM model was also compared to a method developed by Andreatta et al. [16], which is the unique tool reported so far with the same purpose of predicting bacterial pathogenicity. Andreatta et al. proposed a classifier for the prediction of pathogenicity restricted only to *Gammaproteobacteria*, considering a dataset of 155 organisms and obtaining an accuracy of 87%. This is lower than the 96.5% achieved for the same taxonomic group (using 172 organisms) with our SVM model, and even worse than the general performance of our classifier (95.4%). Furthermore, in the particular case of *Gammaproteobacteria*, our method presented a lower error rate in misclassifying human pathogens as non-pathogens (only $\frac{1}{50}$), than the other way around ($\frac{1}{15}$ non-pathogens classified as pathogens). This is of crucial importance in practical applications (such as for clinical or industrial purposes), since the

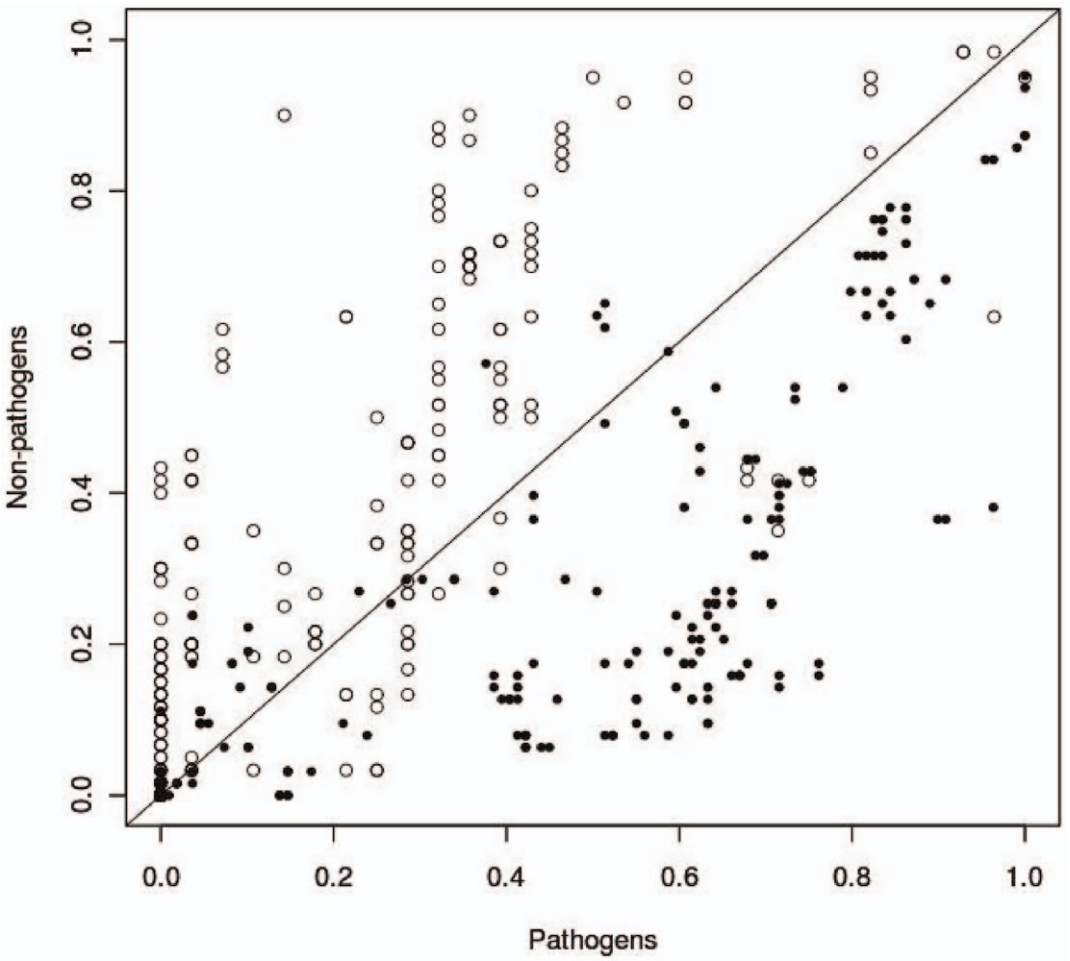

Figure 2. Frequency distribution of ABC transporter genes in *Alphaproteobacteria* and *Gammaproteobacteria*. For each gene, abcisse value is the number of pathogenic strains inside a certain taxonomic group in which it is present, divided by the total number of pathogenic strains inside the taxonomic group. The ordinate value is the same but for the non-pathogenic strains inside the group. White circles show that genes coding for ABC transporters are more frequent in pathogenic species of *Gammaproteobacteria* than in non-pathogenic species of this group. The opposite pattern is observed for *Alphaproteobacteria* in black circles.
doi:10.1371/journal.pone.0042144.g002

social costs of misclassifying a pathogenic strain as non-pathogenic are usually higher than the opposite scenario.

Biological Interpretation

The eight pathogenicity-related functional categories investigated in this work were represented in the set of 120 genes selected for the classifier. Forty genes belonged to ABC transporters, 41 corresponded to two-component systems and chemotaxis proteins, 11 corresponded to toxins, 6 belonged to the LPS biosynthesis pathway and 22 coded for flagellar assembly proteins, motility proteins and proteins from secretion systems. We selected from each group the most distinctive genes and discussed their biological meaning considering their implications in bacterial pathogenesis (Table 5).

ABC transporters. ABC transporters are specialized proteins that function as either importers, which bring nutrients and other molecules into cells, or as exporters, which pump toxins, drugs and lipids across membranes [17]. Based on the kind of substrate ABC transporters are specific for: i) metallic cations, iron-siderophore and vitamin B12, ii) phosphate and amino acids, iii) oligosaccharides and polyol, iv) monosaccharides, v) mineral and organic ions, vi) peptides and nickel and vii) others (ABC-2). Our classification model selected those ABC transporters related to transport of metallic cations, vitamin B12, phosphate and amino acids as the most important.

It is widely known that metallic ions, are essential for prokaryotic cell physiology. The amount of these ions is not constant inside the hosts of pathogenic bacteria, and their concentration is sometimes considerably lower than needed [18]. The presence of systems implied in metallic cations scavenging is mandatory for bacterial survival inside host cells, and it is a key feature for downstream processes like the development of pathogenic phenotypes [19].

The emergence of most pathogenic species is associated with an evolutionary transition from a free-living to a host dependent lifestyle, to a certain extent. Bacterial genomes, and especially those from pathogens, abide by the maxim "use it or leave it", where genes or even whole gene pathways are lost if their products are not essential for cell maintenance, or can be taken from the environment [20]. Two examples are amino acid and vitamin biosynthesis pathways, which have been lost in most pathogens [21]. In this sense, the high representation of these types of ABC systems support the idea that it is more convenient for pathogens to incorporate these compounds from the host environment than to produce them *de novo*.

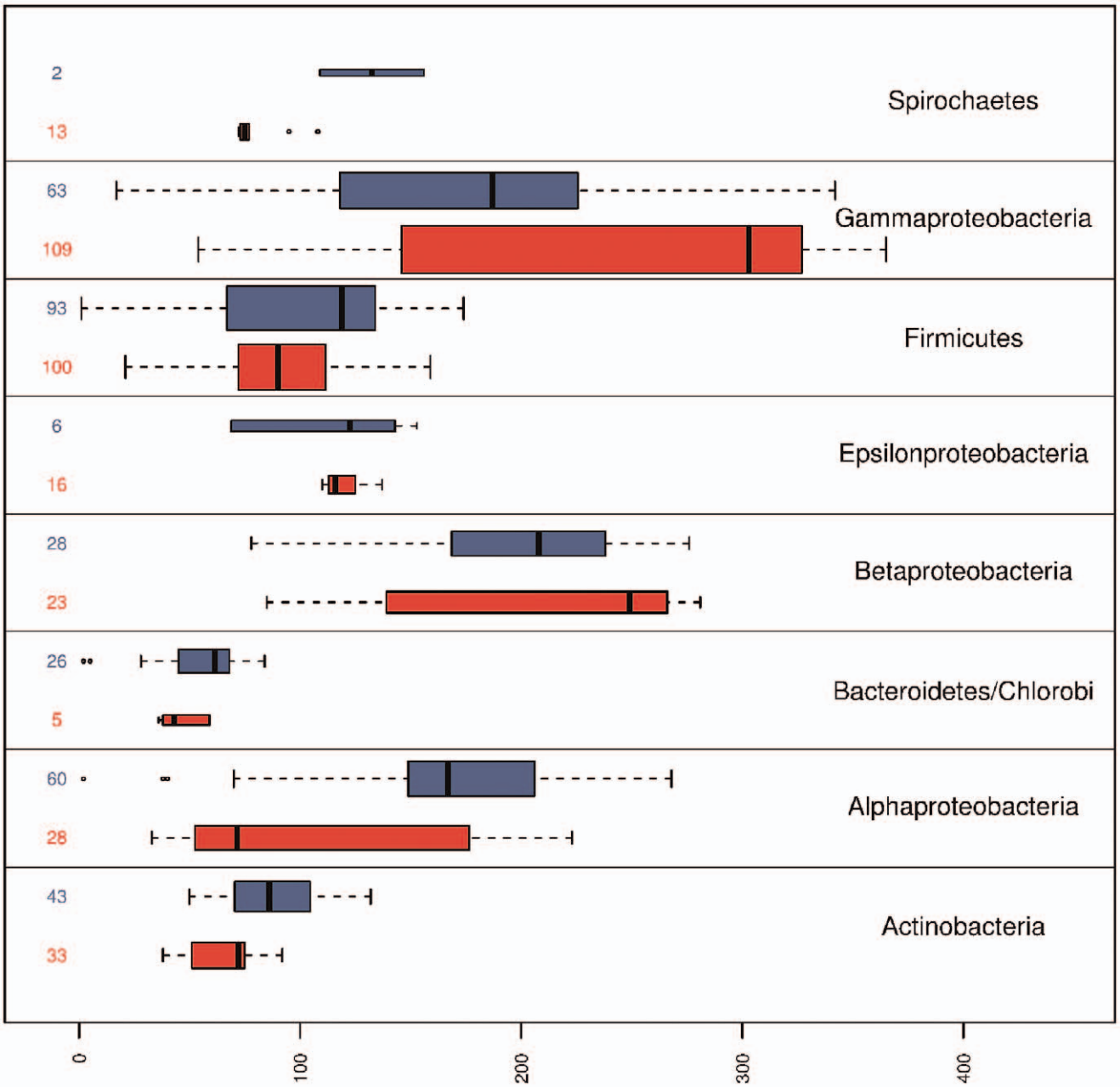

Figure 3. Boxplot representing the presence of genes per taxonomic group. The length of each box represent the number of genes present in both pathogenic (red) and non-pathogenic (blue) organisms for each taxonomic group considered. The number of organisms inside each group are shown leftside, this number is proportional to box width. Dark vertical lines show the median for the amount of present genes per group, box limits represent quartiles and whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range.
doi:10.1371/journal.pone.0042144.g003

Two component systems and chemotaxis. Two-component systems (TCS) are widespread signal transduction pathways among bacteria, which play a crucial role in adaptation to fluctuating surroundings by sensing changes in environmental conditions [21], like those experimented during process of entry, colonization and spread [21]. Genes belonging to 9 TCS families were selected by the classifier as most informative, being OmpR and NtrC the families with the highest TCS representation.

Osmolarity sensors EnvZ-OmpR and CpxA-CpxR (OmpR family) regulate the expression of outer membrane porins in Gram-negative bacteria. Porins control osmolar pressure in response to environmental changes, like from a free-living context to inside a host cell [22].

Gene *vicK* is part of *Bacillus subtilis* VicR-VicK system (also a member of OmpR family). It has been widely related to exopolysaccharide biosynthesis, biofilm formation and virulence factors expression in Gram-positives [23,24]. Gene *vicK* is absent in an important group of non-pathogenic *Firmicutes*, including most non-pathogenic species of genus *Clostridium*. Seemingly, this feature allows the correct classification of these species and is also indicating a certain importance of the VicR-VicK system in some point of *Clostridium* pathogenesis.

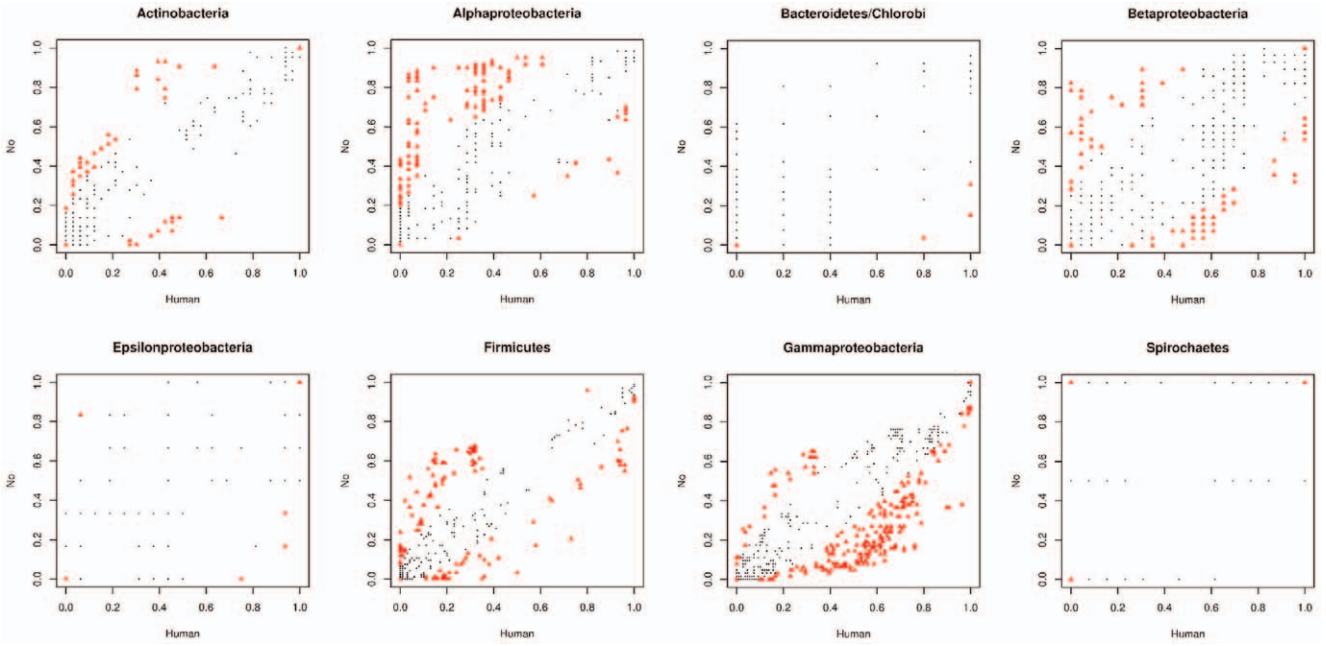

Figure 4. Frequencies of each of 814 genes per bacterial taxonomic group. Frquency calculation was performed for each gene as in Figure 2. Red triangles show significative genes that apart from the null distribution (same frequency in pathogens and non-pathogens) by exact Fisher test, black circles are non significative genes.
doi:10.1371/journal.pone.0042144.g004

The QseB-QseC system is involved in regulation of motility proteins [25], which are key virulence factors of many bacterial pathogens. Often, this system has pleiotropic effects over phenotypes including chemotaxis, adherence, host cell invasion, colonization and innate immune signaling [26]. It was identified in most distinctive pathogenic members of *Gammaproteobacteria*, including *Salmonella*, *Escherichia*, *Vibrio*, and *Shigella*. Surprisingly, it was absent in *Yersinia pestis*' genomes.

Genes representing 5 TCS for NtrC family were selected. Among them we found PilS-PilR, another TCS involved in adherence and cell invasion. This system is essential for type IV secretion systems induction in *Neisseriaceae* species, like *Kingella kingae* an increasingly common cause of septic arthritis, bacteremia, and osteomyelitis in young children [27]. Interestingly, orthologous genes of *pilR* were found in a small group of *Gammaproteobacteria*, including *Pseudomonas aeruginosa*, *Acitnetobacter baumanii* and *Legionella pneumophila*.

Toxins. Pathogenic bacteria have been developing a variety of strategies to manipulate host cell functions, often involving toxins [12]. These proteins have a wide range of action, causing different effects, like host cells deregulation, protein synthesis interruption or membrane damage [28–30]. A total of 76 different bacterial toxins were included in this work. Feature selection analysis selected 11 toxins for the model.

Streptolysin O (SLO) is a thiol-activated cytolysin, the effect of this pore-forming toxin is more subtle than simple lysis of host

Table 1. Statistical overview of data distribution among taxonomic groups.

			Class NP					Class HP			
Taxon	**Purpose[1]**	**n**	**median**	**IQR**	**min**	**max**	**n**	**median**	**IQR**	**min**	**max**
Actinobacteria	M	43	17.0	7.00	9	28	33	12.0	4.00	5	20
Alphaproteobacteria	M	60	28.5	13.50	0	49	28	10.0	23.00	5	37
Bacteroidetes/Chlorobi	M	26	10.5	3.75	0	15	5	8.0	4.00	7	11
Betaproteobacteria	M	28	29.5	11.50	11	47	23	39.0	25.00	14	49
Epsilonproteobacteria	M	6	17.5	9.75	7	20	16	14.0	0.25	13	20
Firmicutes	M	93	20.0	10.00	0	30	100	16.0	10.00	3	30
Gammaproteobacteria	M	63	25.0	15.00	1	47	109	43.0	24.00	9	51
Spirochaetes	M	2	20.0	6.00	14	26	13	9.0	1.00	8	14
Chlamydiae/Verrucomicrobia	T	–	–	–	–	–	14	11.0	0.00	10	12
Deltaproteobacteria	T	28	22.0	5.25	6	31	–	–	–	–	–

Statistical variation is measured as the interquartile range (IQR) in human pathogens (HP) and non-pathogens (NP).
[1]M: used in model construction and testing, T: used only in model testing.
doi:10.1371/journal.pone.0042144.t001

Table 2. Classication performance for each taxonomic groups used to construct the model.

		Class NP		Class HP		
	Number	Predicted as NP	Predicted as HP	Predicted as NP	Predicted as HP	correct classif. rate
Actinobacteria	76	42	1	1	32	97.4%
Alphaproteobacteria	88	54	6	0	28	93.2%
Bacteroidetes/Chlorobi	31	26	0	0	5	100%
Betaproteobacteria	51	27	1	0	23	98.1%
Epsilonproteobacteria	22	6	0	2	14	91%
Firmicutes	193	91	2	3	97	97.4%
Gammaproteobacteria	172	59	4	2	107	96.5%
Spirochaetes	15	2	0	0	13	100%

Inside each class the number of correct and incorrect classified genomes are shown.
doi:10.1371/journal.pone.0042144.t002

cells, and may include interference with immune cell function [31]. SLO is synthesized by more than 20 species of Gram-positive bacteria [32], and it is intimately involved in pathogenesis of *Arcanobacterium pyogenes*, *Clostridium perfringens*, *Listeria monocytogenes* and *Streptococcus pneumoniae* [31]. In this work, SLO was identified in pathogenic *Firmicutes* and absent in non-pathogenic species of this group. This gene is present in most pathogenic strains of *S. pyogenes*, *S. pneumoniae* and those species described by Billington et al. [31], but it is also present in pathogenic *Bacillus cereus*, *Streptococcus dysgalactiae* and *Gardnerella vaginalis*, the latter belonging to *Actinobacteria*.

Hemolysin II and thermolabile hemolysin are also pore-forming toxins selected by the model. The first is produced by pathogenic species of genus *Bacillus*, [33,34] although, in this work, genes extremely similar to hemolysin II were also identified in all pathogenic strains of *Staphylococcus aureus*. Thermolabile hemolysin is characteristic of *Vibrio* species [35] as confirmed by the identification of this gene exclusively in *V. cholerae* and *V. vulnificus* strains.

Cytolethal distending toxin is able to block the host cell cycle between G2 and mitosis [28]. As described in previous works it was identified in a broad range of pathogenic bacteria including *Campylobacter* spp., *Salmonella enterica*, *Haemophilus ducreyi* and *Actinobacillus actinomycetemcomitans* [31]. A/B toxins have similar effects in cell-cycle deregulation, affecting migration, morphogenesis, cell division [36] and membrane trafficking [37]. These were identified in *Clostridium difficile* and in many pathogenic strains of *Escherichia coli*, including O157:H7, O55:H7, O127:H6 and O103:H2. In addition to the contribution for classification, the presence of A/B toxin in these phylogenetically distant groups of possibly indicates horizontal gene transfer events between them.

LPS biosynthesis. Lipopolysaccharides (LPS) are major components of the outer membrane of Gram-negative bacteria, which can be recognized by the host's toll-like receptor 4 (involved in inflammatory response). High concentrations of LPS can induce fever, increase heart rate, and lead to septic shock and death [38].

The model selected six (*lpxK*, *wapR*, *rgpA*, *gmhB*, *rfe* and *rfbP*) out of 94 genes, which code for proteins comprising different steps of typical Gram-negative LPS biosynthesis. Tetraacyldisaccharide 4′-kinase (*lpxK*) catalyzes one of the last steps for Lipid A biosynthesis [39]. Genes *wapR* and *rgpA* produce rhamnosyltransferases, which add rhamnose to the polysaccharide backbone. In particular cases,

Table 3. Confusion matrix showing average classification performance across all taxonomic groups.

Classified as	Pathogenic	Non-pathogenic
Pathogenic	313 (95.2%)	15 (4.8%)
Non-pathogenic	15 (4.9%)	308 (95.1%)

doi:10.1371/journal.pone.0042144.t003

Table 4. Classification performance for Group I and Group II.

	Taxon	Correctly classified	Wrongly classified	Accuracy
	Chlamydiae	14	0	100%
	Deltaproteobacteria	26	0	100%
	Planctomycetes	3	0	100%
	Deinococcus-Thermus	3	0	100%
	Acidobacteria	3	0	100%
Group I	*Deltaproteobacteria*	4	1	80%
	Chloroflexi	8	0	100%
	Cyanobacteria	27	1	96.4%
	Thermotogae	9	0	100%
	Other bacteria	19	0	100%
	Actinobacteria	26	4	87%
	Alphaproteobacteria	24	2	92%
	Bacteroidetes	13	0	100%
	Betaproteobacteria	22	2	91%
	Deltaproteobacteria	5	0	100%
Group II	*Epsilonproteobacteria*	8	1	89%
	Firmicutes	42	4	91%
	Gammaproteobacteria	38	4	90.5%
	Chloroflexi	6	0	100%
	Cyanobacteria	11	1	91%
	Deinococcus-Thermus	7	0	100%
	Other bacteria	13	0	100%

doi:10.1371/journal.pone.0042144.t004

Table 5. Summary of the biological relevance for pathogenicity of a reduced subset of the selected 120 genes.

Functional category	Genes	Comment
ABC	*sitC, hrtB, btuD, gluD*	Strong association between pathogens and the presence of transporters for metallic cations, vitamin B12, phosphate and amino acids
TCS&CH	*vicK, qseC*	VicK absent in most non-pathogenic *Firmicutes*. QseC is present in most pathogenic *Gammaproteobacteria*, but absent in *Yersinia*
LPS	*lpxK, wapR, rgpA, rfbP*	Genes involved in LPS biosynthesis did not show differences in presence/absence patterns between pathogens and non-pathogens
FLA&MOT	*flbP, fimH, fimI, pilA*	FlbP is found in pathogenic *Spirochaetes*. FimH and FimI are found in *Enterobacteraceae*. PilA is present in pathogens of a group of families inside *Gammaproteobacteria*
SS	*tatA, yscC, ppkA*	TatA is found in pathogenic *Epsilonproteobacteria*. YscC is part of T3SS from *Y. pestis* and many other pathogens. PpkA is part of T6SS from *Pseudomonas*
TOX	*slo, tlh, cdtC*	SLO is present in more than 20 pathogenic Gram-positive bacteria, including Firmicutes. Thermolabile hemolysin is exclusive from Vibrio. CdtC is present in a wide broad of pathogens including *Campylobacter*

The functional categories are described in Methods section.
doi:10.1371/journal.pone.0042144.t005

the incorporation of L- or R-rhamnose determines different glycoforms of the core region, leading to LPS variability, hence virulence [40]. Two genes are involved in O-antigen biosynthesis: *rfbP* codes for a glycosyltransferase responsible for the first step in O-antigen biosynthesis [41], while *rfe* (*wecA*) catalyzes the first membrane step of O-antigen and enterobacterial common antigen biosynthesis in *E. coli*. Its involvement in the virulence of Gram-negative bacteria has also been reported [42].

In spite of being selected by the model as relevant for classification, none of these genes showed a clear presence/absence pattern among pathogenic and non-pathogenic species. However, this does not mean they are not informative; on the contrary, these genes may be contributing to classification by an additive effect, being their individual inputs restricted to more particular groups.

Flagellar assembly and motility. Bacterial motility is a major factor in pathogenesis. This feature is involved in processes like biofilm formation, host cell colonization and bacterial spread inside the host [43]. Flagellar macromolecular machinery is the paradigm of bacterial motility, being present in a wide range of human pathogens, including *E. coli*, *S. enterica* and *P. aeruginosa* [44–46]. In the present work, 34 different genes involved in flagellum formation were investigated. Additionally, other 137 genes involved in different mechanisms related to bacterial motility (fimbrial proteins, adhesins, chemosensory proteins and regulatory proteins) were included.

Five genes directly involved in flagellar biosynthesis (*fliA*, *fliD*, *fliK*, *fliL* and *fliW*) were selected by the model. Gene *fliA* codes for σ^{28}, responsible for the regulation of flagellin biosynthesis. Inactivation experiments of *fliA* in *P. aeruginosa* cause non-motility, due to inability of expressing the flagellin gene [47]. The *fliD* gene codes for a structural component of the flagellar cap, which is important in host cell adhesion and colonization [48]. Gene *fliL* is dispensable for swimming in pathogenic species like *E. coli* and *S. enterica* [49], but it is essential for swarming (flagellar-dependent motility in solid medium) in these species. Gene *fliK* is responsible for controlling flagellar hook length, which directly affects the performance of the flagella in producing translational motion [50]. Gene *fliW* codes for a new flagellin assembly protein in *Treponema pallidum* which has orthologous in many related species [51].

Gene *flbB* is part of the flagellar motor exclusively in *Spirochaetes* sp. [52]. In this work, this gene was found in pathogenic *Spirochaetes* and was absent in many other genomes, suggesting its importance for the correct classification of this group. Nevertheless, *flbB* homologues were also found in *Thermoanaerobacter* (*Firmicutes*). Independently of its role in the classification of pathogens, this finding questions the evolutionary origin of this flagellar motor, apparently exclusive for *Spirochaetes*.

Bacterial motility and host-cell adhesion are intimately related processes. Fimbria (type I pili) are filamentous proteinaceous surface appendages present in many Gram-negative bacteria [53,54] that aid the adhesion process. In *E. coli*, fimbria are made of a repeating monomer, FimA, encoded by *fimA*. This gene is almost exclusively present in pathogenic *Gammaproteobacteria* and *Betaproteobacteria*, like *Escherichia*, *Salmonella*, *Acinetobacter* and *Burkhordelia*. FimH protein (encoded by *fimH*) is the most common adhesin located on the tip of type I fimbriae [55,56]. Its expression, hence pilus formation, is regulated by gene *fimI*, which is essential for fimbriated phenotype. Specific mutations in *fimI* lead to pilus-negative phenotype in *E. coli* and *S. enterica* [57]. Both genes, *fimH* and *fimI*, were found exactly in the same group of species belonging to *Enterobacteraceae* family: *Salmonella*, *Escherichia*, *Proteus*, *Shigella* and *Klebsiella*. This supports the functional relationship of both genes and also denotes the importance of them for classification of this family of pathogenic *Gammaproteobacteria*.

Another relevant pili aparatus is the type IV system. This macromolecular machinery is present in Gram-negative bacteria and in at least one Gram-positive [58]. Type IV pili are highly pleiotropic, being involved in bacterial motility, adhesion, immune escape, biofilm formation, secretion and phage transduction. The most relevant selected gene for this pili system was *pilA*, which codes for pilin, the major component of filament. It is present in most pathogenic *Clostridium* (*C. perfringes*, *C. tetani*, *C. difficile* and *C. botullinum*). PilA is also present in pathogenic members of a group of families belonging to *Gammaproteobacteria* (*Vibrionaceae*, *Pseudomonadaceae*, *Francisellaceae*, *Moraxellaceae*). Interestingly, *pilA* is absent in pathogenic *Enterobacteraceae*, so the combination of three genes (*pilA*, *fimH* and *fimI*) seems to explain the discrimination of most pathogenic *Gammaproteobacteria* with respect to the rest of non-pathogenic bacteria and even distinguishing between two enormous phyolgenetic groups inside this taxon.

Secretion systems. Several differences in secretion systems exist between Gram-positive and Gram-negative bacteria. Protein secretion across the inner membrane of both kinds of organisms generally involves the same Sec-dependent pathway, although

other routes have been identified, i.e. Twin-arginine translocation (Tat) [59–61]. Translocation across Gram-negatives inner membrane results in release of products into the periplasmic space. Hence, these bacteria have developed several types of secretion systems which carry molecules from the periplasmic space to the cell surface or extracellular matrix. These secretory pathways of Gram-negatives can be classified into six different groups: type I to VI secretion systems (T1SS–T6SS). The presence/absence of 73 different genes coding for both shared secretory pathways (like Sec or Tat) and for T1SS–T6SS was tested. The model selected 13 genes as the most relevant to explain class differences.

Genes for Sec system were not selected by the model. For Tat system the *tatA* gene was selected; it codes the major pore-forming subunit for translocation complex [62]. Homologues of *tatA* have been identified in a wide range of human pathogens, including *E. coli* O:157, *Vibrio cholerae*, *Mycobacterium tuberculosis*, *Listeria monocytogenes* and *Staphylococcus aureus* [63]. Moreover, this gene has orthologous in all *Epsilonproteobacteria* analyzed in this work, except for the non-pathogenic *Sulfurovum* sp. NBC37-1. Even though *tatA* was selected as an important feature for classification, a clear presence/absence pattern between pathogenic and non-pathogenic species was not observed.

Gene *yscC* encodes a key protein of the archetypical T3SS of *Yersinia pestis*, the infective agent of human plague. YscC orthologs are now identified in more than a dozen of pathogens [64], including *Salmonella enterica*, *Shigella flexneri* [65] and enteropathogenic *E. coli* [66]. Beyond these well-known examples, we identified the presence of *yscC* orthologs only in species belonging to *Gammaproteobacteria* and *Betaproteobacteria*, being absent in a great number of non-pathogenic species.

T4SS have been described in several organisms including *Bordetella pertussis* [67], *Legionella pneumophila* [68], *Brucella suis* [69], *Bartonella henselae* [59], and *Helicobacter pylori* [70]. VirB2, coded by *virB2*, is major component of T4SS pilus and has an important role in secretion [71]. Beyond its identification in the species mentioned above, *virB2* is present in some genomes of well-known pathogens with different taxonomic context: *Campylobacter jejuni* subsp. *jejuni* 81–176 (*Epsilonproteobacteria*), *Klebsiella pneumoniae* subsp. *pneumoniae* NTUH-K2044 (*Gammaproteobacteria*), *Neorickettsia sennetsu* str. Miyayama (*Alphaproteobacteria*) and three *Burkholderia* sp. species (*Betaproteobacteria*). This suggests an important role of T4SS in pathogenic processes, even in species with different pathogenic mechanisms.

T6SS have been found in species from a wide taxonomic range [72], comprising most bacterial groups included in this work. Two T6SS genes were selected: *ppkA* codes for a serine/threonine-protein kinase that phosphorylates protein FHA (encoded by *fha1*). The phosphorylation initiates a signal transduction cascade that results in T6SS assembly and function. Mutation of *P. aeruginosa fha1* gene resulted in defective secretion of Hcp1, an essential protein for pathogenesis as demonstrated by attenuated virulence phenotype observed *in vivo* [73]. Both *fha1* and *ppka* were identified in *P. fluorescens* and *P. mendocina* and all strains of *P. aeruginosa*. Interestingly, the absence of these genes in other genomes shows the great importance of their presence for the classification of these organisms exclusively. Moreover, the high correlation in the presence of both genes in the same genomes evidences their functional relationship.

Phylogenetic Distribution of Virulence Genes

In the sections above we disscused the biological meaning of some genes selected by the model, emphasizing their presence/absence patterns among pathogens and non-pathogens and their importance in the development of pathogenic phenotypes. Here we give an integrative overview of virulence genes distribution along bacterial phylogeny, taking into account their frequency bias among pathogenic and non-pathogenic organisms. Fisher exact test (p-value <0.001) was used to select genes with significant differences in their presence/absence patterns for each functional category inside each taxonomic group. Then, gene frequency was calculated among pathogens and non-pathogens for those selected genes, separated by functional category. Finally, individual genes frequencies were added inside each group and normalized over the total number of genes belonging to each functional category.

Figure 5 shows normalized frequency values for genes belonging to each functional category, taking into account the phylogenetic relationships between studied taxonomic groups. Some expected patterns arise from these results, for example toxins are exclusively overrepresented in pathogenic species. This is expectable taking into account the biological purpose of toxins; it would be highly improbable that pathogenicity in a certain species was determined by the absence of a toxin that is present in the non-pathogenic species of the group. ABC transporters seem to be the most variable functional category along the phylogeny, it is positive (associated to pathogenic organisms) in *Gammaproteobacteria*, *Betaproteobacteria* and *Firmicutes*, and negative (associated to non-pathogenic organisms) in *Alphaproteobacteria* and *Actinobacteria*. This is coherent with the wide range of functions that ABC transporters can perform; for example the presence of aminoacid importers can be essential for pathogenesis of species that have lost biosynthetic genes, however, it is not contradictory with the presence of these kind of transporters in non-pathogenic species.

The most powerful association between pathogens and high gene frequencies is observed in *Gammaproteobacteria*, evidencing the importance of these kinds of genes for pathogenic species of this group, which is mainly composed of enteropathogens. The most striking result of this analysis is the pattern observed for *Alphaproteobacteria*, totally opposite to the phylogenetically related *Gammaproteobacteria*. The first question that rises is why genes previously thought of as mostly present in pathogenic species, are highly frequent in non-pathogenic species of this taxon. Marine environments contain the major component of non-pathogenic *Alphaproteobacteria* biodiversity. A recent study [74] showed that out of 119 marine bacteria, 60 had homologues to known virulence genes from pathogenic bacteria. Interestingly, new insights in host-pathogen interactions propose a wider ecological and evolutionary perspective to better understanding the life strategy of pathogenic bacteria [75], suggesting that functions have evolved over a long time in nature and then recruited through horizontal gene transfer to perform similar or different functions in more recently emerging pathogenic species. This hypothesis opens a three-step way of thinking about how natural selection plays a role in the emergence of bacterial pathogens. First, the random appearance and fixation of new genes in bacteria colonizing inaccessible environments generate a reservoir of species carrying potentially virulent genes. Second, these bacteria can contact human hosts by movement through intermediate hosts in which they live as commensals or they can transfer virulent genes horizontally to other human-adapted bacteria. Third, positive selection over the most successful species determines the fixation of virulence genes that let bacteria to damage or survive inside human cells. The high frequency of virulence-related genes in non-pathogenic *Alphaproteobacteria* might be explained by the emergence of these kinds of genes in common ancestors for *Gammaproteobacteria* and *Alphaproteobacteria*. Then, the branch that originated *Alphaproteobacteria* conserved these genes in both pathogenic and non-pathogenic species. In contrast, *Gammaproteobacteria* could have acquired these functions by

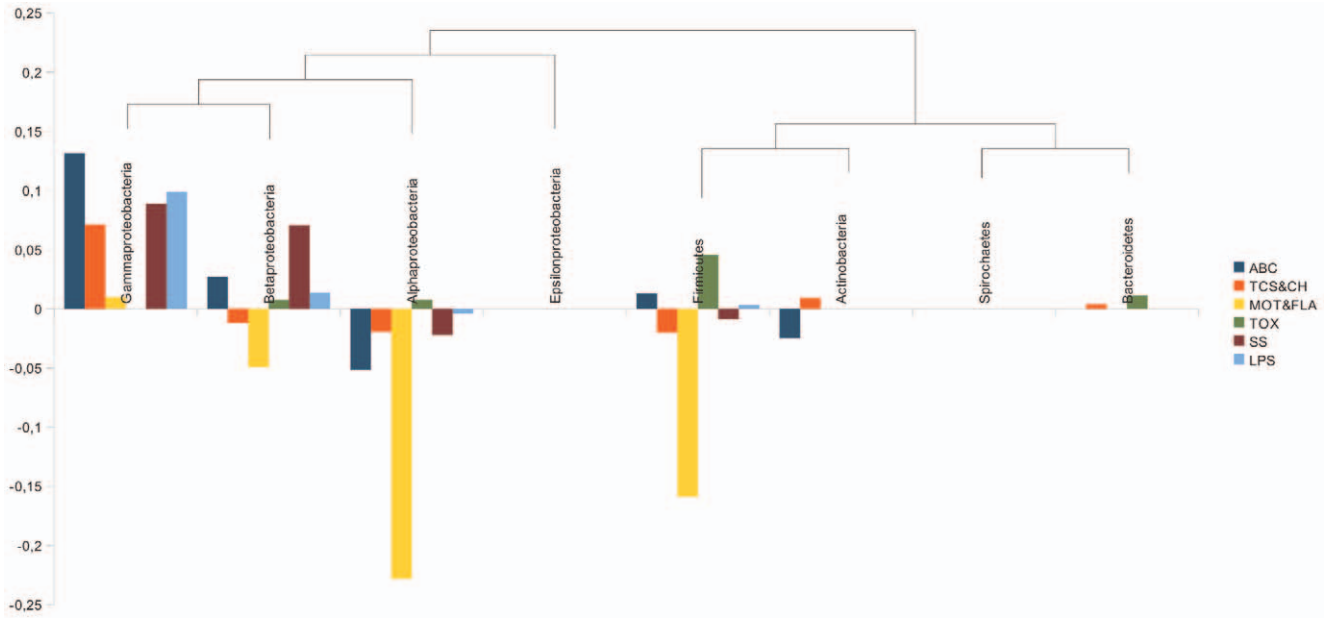

Figure 5. Phylogenetic distribution of virulence genes. Each functional category of virulence-related genes is represented as a vertical bar. Positive values denote association of a particular functional category with pathogenic species of a certain taxonomic group, while negative values with non-pathogenic species. Taxons are grouped according to phylogenetic relationships. In graph legend: ABC: ABC transporters, TCS&CH: two-component systems and chemotaxis, MOT&FLA: motility and flagellar assembly, TOX: toxins, SS: secretion systems, LPS: LPS biosynthesis.
doi:10.1371/journal.pone.0042144.g005

horizontal gene transfer, to produce the actual scenario of high frequency in pathogenic species and low frequency in non-pathogenic ones.

Two groups (*Spirochaetes* and *Epsilonproteobacteria*) showed very few genes with significant differences according to Fisher exact test. This reveals that for these two taxonomic groups there are no clear presence/absence patterns among genes of pathogenic and non-pathogenic species but, in spite of this, our model is able to assign each organisms to the correct class with high accuracy. This is particularly interesting because our model is using information coded in high-dimensional spaces, leaving behind the simple presence/absence patterns. Moreover, here we could identify only some particular associations between phylogeny topology and functional categories, suggesting that, in general, the functional importance of these genes varies along bacterial taxonomy. The lack of general patterns between the presence of functional categories and phylogenetically related groups supports the notion that most virulence-related genes are spread among bacteria by horizontal gene transfer. Probably our method is taking benefit of this scenario, being able to correctly classify organisms independently of their taxonomic context, based on widely spread genes along bacterial phylogeny.

Misclassified Organisms

A group of 28 out of the 648 genomes tested were systematically misclassified by the model. We defined a genome to be misclassified if it was assigned to the wrong class, at least in 50% of 20 consecutive classifications (Table S4). Ten out of these 28 are labeled as human pathogens but the model returned them as non-pathogenic, while 18 out of 28 are labeled as non-pathogenic but were classified as human pathogens. Most cases of misclassification are observed in species with a big number of sequenced genomes of different strains. This is the case of *Staphylococcus aureus*, an important human pathogen. Thirteen out of the 14 genomes of different strains of this species were well classified as human pathogens. Nevertheless, the strain *S. aureus* subsp. *aureus* MRSA252 was assigned to the non-pathogenic class. Comparison of present/absent genes for all *S. aureus* genomes showed that gene *hlyII* (coding for hemolysin II) was absent in *S. aureus* subsp. *aureus* MRSA252 while present in the rest. This was the only difference between these genomes; moreover gene *hlyII* was one of the 11 toxin-coding genes selected as more informative during the feature selection process. On the one hand, this fact shows that for a particular species even the presence of a single feature is determining the classification of the genome as pathogenic or non-pathogenic, indicating a great power of some genes in determining the class assignment by the model. On the other hand, it is possible to misclassify genomes due to a particular gene loss, especially in those cases of high genetic variability among strains of certain species.

For misclassified genomes that do not have other well-classified strains belonging to the same species, it is not possible to assess the present/absent comparison to find differences in gene patterns. In these cases, misclassification can be explained by inherent errors of SVM model construction or because the features (groups of orthologous genes) originally used to determine the presence/absence matrix, might not be informative enough to reach a 100% classification performance. However, in some cases it is possible to propose a biological explanation for misclassification, based on the particular ecological and genetic features of some species.

The first example is *Bordetella petrii* (*Betaproteobacteria*) which is originally labeled as non-pathogenic, but the model classifies it as pathogenic. This could be primarily seen as a classification error, but there is strong evidence that supports this species is an emerging human pathogen. Though being an environmental isolate, the sequenced *B. petrii* DSM12804 strain also encodes proteins related to virulence factors of the pathogenic *Bordetellae*, including the filamentous hemagglutinin, which is a major colonization factor of *B. pertussis*. The genomic analysis of *B. petrii* suggests an evolutionary link between free-living environmental

bacteria and the host-restricted obligate pathogenic *Bordetellae* [76]. Moreover, clinical isolates of *B. petrii* have been recently described to cause, for example, mandibular osteomyelitis [77] or supurative mastoiditis [78].

Other example comprises a group of 6 marine non-pathogenic *Alphaproteobacteria* (*Rhodobacter capsulatus*, *Erythrobacter litoralis*, *Rhodopseudomonas palustris*, *Novosphingobium aromaticivorans*, *Parvularcula bermudensis* and *Sphingobium japonicum*), wrongly classified as pathogenic. As explained in the section above, *Alphaproteobacteria* have the highest frequency of virulence-related genes in non-pathogenic species. The 6 misclassified species shared the presence of 9 genes involved in secretion processes, supporting the findings of Persson et al. [74] regarding the extensive appearence of these kinds of genes in marine bacteria. Despite this, only 6 out of 88 *Alphaproteobacteria* were misclassified, indicating that the classification model can deal with unexpectedly biased gene frequencies towards non-pathogenic organisms without compromising classification performance.

Model Sensitivity

A simple approach to evaluate the sensitivity of the constructed model is to assess the propensity of label shift (pathogens to non-pathogens and vice versa). This experiment was implemented for each taxonomic group in the dataset by artificially modifying presence/absence vectors. For each genome those present genes were systematically "turned off" one at a time, running the classification model each time and recording in which cases a category shift occurred. The same strategy was used to "turn on" those genes which were originally absent.

The change from non-pathogen to pathogen was lead by a group of 14 genes, which were mainly toxin-coding genes (5) and TCS (5). These two functional categories together comprise $\frac{2}{3}$ of the genes that influence the category shifting in the mentioned direction, evidencing a great importance of these features as exclusive determinants of bacterial pathogenicity. Individually, the presence of any of these genes is able to change a number of organisms ranging from 78 to 153, depending on the gene. The most extreme is the case of SLO toxin, whose presence determines that 153 species change from non-pathogens to pathogens.

Changing from pathogen to non-pathogen is mainly determined by gene "turn off". A group of 9 genes are responsible for category shifting in this direction, changing the classification of 10 to 96 species. It is worth mentioning that the gene coding for the SLO toxin is one of the most influential; this makes sense, since the gain of this gene provoked a label change to pathogen, it is expectable that losing it defines a label change to non-pathogen.

Software Development: The BacFier

BacFier v1.0 was implemented as a Java software, and hence platform independent, in order to make it easier for the common user to work with the model. A simple interface allows the user to upload the genome sequence (finished or unfinished) of the organism of interest. The genome is used as query to perform BLAST against the final set of 120 orthologous groups (selected as explained in section Model construction) creating a presence/absence vector for the genome. The vector is evaluated with a SVM model, and an outcome (pathogen/non-pathogen) is produced associated to a probability.

Moreover, the sensitivity analysis described in the previous section can be automatically performed with the software, this is assessed by selectively "turning off" or "turning on" desired genes in the presence/absence vector and re classifying the result. This might indicate genes that are likely to change the label of the organism, so that one can pay more attention to them and corroborate their status of presence/absence. Furthermore, this strategy becomes crucial when inputing an unfinished genome. In this situation, the absence of some genes important for pathogenicity could be determined by the unfinished status of the genome, so if prediction result is non-pathogenic, the user can sistematically "turn on" those absent genes until the model shift to pathogenic. Then, the real presence of genes that determined the shift can be investigated by a more refined search or by other methods, like PCR.

BacFier v1.0 is freely available under http://bacfier.google code.com/files/Bacfier_v1_0.zip.

Conclusions

The constructed SVM model classifies bacterial genomes in human pathogens and non pathogens with 95.4% of average accuracy. To the best of our knowledge, this is the statistical model with this purpose that achieves the highest accuracy reported so far. Moreover, our method classifies bacterial genomes independently of their taxonomic context, in contrast to other similar approaches that only take into account a certain part of bacterial diversity, being useful only to classify specific taxa [6]. Our statistical learning approach is grounded on the biological meaning of the selected genes and supported by the fact that bacterial pathogenicity can be explained by the presence or absence of a set of specific genes that code for virulence determinants. The application of BacFier v1.0 may be useful for clinical or industrial purposes, for example to determine if a new sequenced strain could be pathogenic for humans.

Methods

Data Selection and Matrix Construction

Complete genome sequences from all available bacteria were downloaded from the National Center for Biotechnology Information (NCBI). Over 1000 genomes were obtained and from those organisms, we originally kept 848 that were labeled as human pathogens or non-pathogens. This set of bacteria comprehends 22 taxonomic groups. In this work, we focused only on human pathogens; if a certain species was a multi-host pathogen including humans, it was considered human pathogen. By the contrary, if a certain species was a multi-host pathogen or a pathogen of other host different from human, it was excluded from the dataset considered.

Eight gene functional categories that we considered related to pathogenicity were determined. These are toxins, chemotaxis proteins, ABC transporters, motility proteins, LPS biosynthesis, two-component systems, flagellar assembly and secretion systems. Orthologous groups from proteins coded by genes belonging to these categories were downloaded from KEGG Orthology database (http://www.genome.jp/kegg/ko), all the categories together resulted in 814 orthologous groups. With this data, we built a presence/absence table showing which orthologous groups (genes/proteins) were present or absent in the organisms considered. We selected local protein BLAST [79] searches to perform orthologous genes determination. Not only does this approach absolve us from using a refined orthologous search method (which can be much more laborious and time-consuming), but it also provides good enough accuracy in orthologous determination. In this case, our method must be robust and tolerant enough to identify possible false positive or false negative orthologs.

BLAST searches were performed formatting the 814 orthologous groups and querying the organisms. If an alignment between

an organism and a gene (member of an orthologous group) was "good enough" (see below), then we considered the gene (orthologous group) as present in the organism, otherwise as absent. This, is represented as a 0/1-matrix with dimensions $|\text{organisms}| \times |\text{orhologousgroups}|$. We defined "good" alignments as the ones having a percentage of identity higher than 90%, length of the alignment larger than 90% of the gene's length and an e-value smaller than 0.001. Further analyses were made on 648 genomes belonging to 8 of the 22 taxonomic groups: *Actinobacteria*, *Alphaproteobacteria*, *Bacteroidetes/Chlorobi*, *Betaproteobacteria*, *Epsilonproteobacteria*, *Firimicutes*, *Gammaproteobacteria* and *Spirochaetes*, since there were not enough genomes available for the other groups. However, these excluded genomes were then used as part of external groups to further test the constructed model.

Model Construction

In this work a machine learning approach based on a cross-fold validation with in-fold feature selection was developed. This technique ensures that particular predictions are not biased by overselected features or overfitting since each prediction is performed without using the sample in neither the feature selection nor the classifier building process. Algorithm 1 shows the methodology.

Algorithm 1. General overview of cross-fold validation with in-fold feature selection (be X = whole set of samples).

for $i \leftarrow 1 \rightarrow$ *nfolds* **do**
  Define validation set VS $\leftarrow$ *samples in fold i*
  Define training set TS $\leftarrow X - VS$
  Perform feature selection over TS samples
  Train classifier using TS
  Perform prediction of VS samples with previous classifier
end for

The number of folds (nfold) was set to 10 and the feature selection routine was SVMAttributeEval from Weka [80]. Regarding the classification algorithm, a Support Vector Machine (SVM) was employed. The SVM method performs the classification by constructing an N-dimensional hyperplane that optimally separates the data into two classes. In this case classes are labeled as human pathogens and non-pathogens. The raw dataset of variables is defined by the presence/absence of orthologous groups in the genomes of the organisms considered. It is important to note that the taxonomy is not used as another variable in the model since it would introduce an artificial separation in the SVM model training.

Following the spirit of Occam's razor, in this work a linear SVM model is proposed. Although the number of genes looks relatively large, it is worth to mention that the model variables encode low level information related to gene presence/absence in each organism. Also, it is well known that linear SVM models benefit from using these kinds of variables since higher dimensions allow easier class separation. The subroutine libsvm in Weka was also employed [80].

A final analysis was done in order to determine an appropiate number of features to retain. Experiments were carried out considering 30, 60, 90, 120, 150, 200 and 841 (entire set of genes) features. The accuracy obtained in each case was 90%, 93.5%, 94.4%, 95.4%, 95.5%, 94.9% and 92.1% respectively. A set of 120 genes was then considered, as they represent a reasonable tradeoff between accuracy prediction and the number of genes used for prediction.

From Algorithm 1 is clear that a different set of features can be selected in each loop of the cross-validation procedure. However, it is necessary to find a final set of genes to build a classification model and check and external validation set (for practical purpose) or predict pathogenicity of new sequenced bacteria. A common solution is to employ a voting scheme that sums how many times a feature is selected in each loop of Algorithm 1. In this particular case, the list of genes selected is available in Table S1.

Y-randomization test. Since in this work a binary occurrence matrix is used to represent the presence/absence of genes in a set of organisms, the number of calculated variables is high, as expected. In this particular case, the number of genes is 814. A feature selection technique further reduced the set to the 120 most significant variables. Although this meets the rule of thumb that states the ratio between number of samples (648 organisms) and variables (120) must be greater than 5 [81], problems associated with chance correlation could still arise. This is a major concern when the prediction model is expected to be reliable in terms of generalizability.

The y-randomization validation method tries to observe the influence of chance when fitting any given data. This is done by deliberately destroying the relationship between the target y and the independent variables x (genes, in this case). This is done by randomly shuffling the y data, preserving all x data untouched, and retraining the learning algorithm. A common pitfall is to apply the y-randomization procedure but using the same set of variables resulting from the feature selection process. Following the good-practice procedures, in this work the test was carried out using the full set of variables, so there was no "overestimation" (in the sense of chance correlation).

In this work we have two classes, so the expected behavior was to obtain an accuracy of roughly 50% in the y-randomization test (since 50% is the probability of a "good" prediction when no relation is found between variables and targets, the same as a random assignment of predicted labels). In this work the y-randomization procedure was carried out 100 times (Figure S1).

Genes Significance and Frequency Calculation

In order to weight the importance of each functional category for each taxonomic group, we selected those genes with statistically significant presence/absence patterns inside pathogens and non-pathogens. Fisher exact test was applied to genes belonging to each functional category for each taxonomic group. Those genes with p-value <0.001 were taken into account. Then, the frequency of those genes was calculated for pathogenic and non-pathogenic species of each taxonomic group, as the number of presences over the total number of organisms inside the group. Finally, for a certain functional category, the significance value was calculated as the accumulated frequency of those genes significant to the category, and normalized over the total number of genes belonging to it. For a better graphical visualization of Figure 5, frequencies in non-pathogenic organisms were multiplied by -1, in this way positive values are associated with pathogenic organisms while negative with non-pathogenic ones.

Supporting Information

Figure S1 Y-randomization performance over 100 runs.
(TIF)

Table S1 Description of the subset of 120 selected genes.
(XLS)

Table S2 Classification results for each tested genome in 10-fold cross validation.
(XLS)

Table S3 Prediction for each organism belonging to test Group II. These organisms were previously subjected to a bibliography revision to determine their assignation to human pathogens or non-pathogens. When organism resulted to be pathogen, citation is reported. Column Prediction shows the result after model prediction, correctly classified organisms are highlighted in green while wrongly classified are in red.
(XLS)

Table S4 List of misclassified organisms.
(XLS)

Acknowledgments

We thank Natalia Rego for thorough reading and insightful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: GI GV LS HN. Performed the experiments: GI GV LS HN. Analyzed the data: GI. Contributed reagents/materials/analysis tools: GI GV LS HN. Wrote the paper: GI GV LS HN.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus inuenzae Rd. Science 269: 496–512.
2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of Mycoplasma genitalium. Science 270: 397–403.
3. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388: 539–547.
4. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, et al. (2001) Whole genome sequencing of meticillin-resistant Staphylococcus aureus. Lancet 357: 1225–1240.
5. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, et al. (2001) Genome sequence of entero-haemorrhagic Escherichia coli O157:H7. Nature 409: 529–533.
6. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
7. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
8. Sokurenko EV, Chesnokova V, Dykhuizen DE, Ofek I, Wu XR, et al. (1998) Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. Proc Natl Acad Sci USA 95: 8922–8926.
9. Conenello GM, Zamarin D, Perrone LA, Tumpey T, Palese P (2007) A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 inuenza A viruses contributes to increased virulence. PLoS Pathog 3: 1414–1421.
10. Marjuki H, Scholtissek C, Franks J, Negovetich NJ, Aldridge JR, et al. (2010) Three amino acid changes in PB1-F2 of highly pathogenic H5N1 avian inuenza virus affect pathogenicity in mallard ducks. Arch Virol 155: 925–934.
11. Spangenberg L, Battke F, Grana M, Nieselt K, Naya H (2011) Identifying associations between amino acid changes and meta information in alignments. Bioinformatics 27: 2782–2789.
12. Oswald E, Nougayrede JP, Taieb F, Sugai M (2005) Bacterial toxins that modulate host cell-cycle progression. Curr Opin Microbiol 8: 83–91.
13. Lanie JA, Ng WL, Kazmierczak KM, Andrzejewski TM, Davidsen TM, et al. (2007) Genome sequence of Avery's virulent serotype 2 strain D39 of Streptococcus pneumoniae and comparison with that of unencapsulated laboratory strain R6. J Bacteriol 189: 38–51.
14. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K (2008) Genome sequence of Staphylococcus aureus strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. J Bacteriol 190: 300–310.
15. Ho Sui SJ, Fedynak A, Hsiao WW, Langille MG, Brinkman FS (2009) The association of virulence factors with genomic islands. PLoS ONE 4: e8094.
16. Andreatta M, Nielsen M, Moller Aarestrup F, Lund O (2010) In silico prediction of human pathogenicity in the -proteobacteria. PLoS ONE 5: e13680.
17. Rees DC, Johnson E, Lewinson O (2009) ABC transporters: the power to change. Nat Rev Mol Cell Biol 10: 218–227.
18. Rohmer L, Hocquet D, Miller SI (2011) Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. Trends Microbiol 19: 341–348.
19. West SA, Buckling A (2003) Cooperation, virulence and siderophore production in bacterial parasites. Proc Biol Sci 270: 37–44.
20. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108: 583–586.
21. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annu Rev Biochem 69: 183–215.
22. Shikuma NJ, Yildiz FH (2009) Identification and characterization of OscR, a transcriptional regulator involved in osmolarity adaptation in Vibrio cholerae. J Bacteriol 191: 4082–4096.
23. Dubrac S, Boneca IG, Poupel O, Msadek T (2007) New insights into the WalK/WalR (YycG/YycF) essential signal transduction pathway reveal a major role in controlling cell wall metabolism and biofilm formation in Staphylococcus aureus. J Bacteriol 189: 8257–8269.
24. Senadheera MD, Guggenheim B, Spatafora GA, Huang YC, Choi J, et al. (2005) A VicRK signal transduction system in Streptococcus mutans affects gtfBCD, gbpB, and ftf expression, biofilm formation, and genetic competence development. J Bacteriol 187: 4064–4076.
25. Sperandio V, Torres AG, Kaper JB (2002) Quorum sensing Escherichia coli regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of agella and motility by quorum sensing in E. coli. Mol Microbiol 43: 809–821.
26. Josenhans C, Suerbaum S (2002) The role of motility as a virulence factor in bacteria. Int J Med Microbiol 291: 605–614.
27. Kehl-Fie TE, Porsch EA, Miller SE, St Geme JW (2009) Expression of Kingella kingae type IV pili is regulated by sigma54, PilS, and PilR. J Bacteriol 191: 4976–4986.
28. Whitehouse CA, Balbo PB, Pesci EC, Cottle DL, Mirabito PM, et al. (1998) Campylobacter jejuni cytolethal distending toxin causes a G2-phase cell cycle block. Infect Immun 66: 1934–1940.
29. Marches O, Ledger TN, Boury M, Ohara M, Tu X, et al. (2003) Enteropathogenic and enterohaem-orrhagic Escherichia coli deliver a novel effector called Cif, which blocks cell cycle G2/M transition. Mol Microbiol 50: 1553–1567.
30. Gebert B, Fischer W, Weiss E, Hoffmann R, Haas R (2003) Helicobacter pylori vacuolating cytotoxin inhibits T lymphocyte activation. Science 301: 1099–1102.
31. Billington SJ, Jost BH, Songer JG (2000) Thiol-activated cytolysins: structure, function and role in pathogenesis. FEMS Microbiol Lett 182: 197–205.
32. Alouf JE (1980) Streptococcal toxins (streptolysin O, streptolysin S, erythrogenic toxin). Pharmacol Ther 11: 661–717.
33. Sinev MA, Budarina ZhI, Gavrilenko IV, Tomashevski AIu, Kuz'min NP (1993) [Evidence of the existence of hemolysin II from Bacillus cereus: cloning the genetic determinant of hemolysin II]. Mol Biol (Mosk) 27: 1218–1229.
34. Budarina ZI, Sinev MA, Mayorov SG, Tomashevski AY, Shmelev IV, et al. (1994) Hemolysin II is more characteristic of Bacillus thuringiensis than Bacillus cereus. Arch Microbiol 161: 252–257.
35. Zhang XH, Austin B (2005) Haemolysins in Vibrio species. J Appl Microbiol 98: 1011–1019.
36. Ja_e AB, Hall A (2005) Rho GTPases: biochemistry and biology. Annu Rev Cell Dev Biol 21: 247–269.
37. Ridley AJ (2001) Rho proteins: linking signaling with membrane trafficking. Traffic 2: 303–310.
38. Yamamoto M, Sato S, Hemmi H, Uematsu S, Hoshino K, et al. (2003) TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway. Nat Immunol 4: 1144–1150.
39. Raetz CR, Guan Z, Ingram BO, Six DA, Song F, et al. (2009) Discovery of new biosynthetic pathways: the lipid A story. J Lipid Res 50 Suppl: S103–108.
40. Poon KK, Westman EL, Vinogradov E, Jin S, Lam JS (2008) Functional characterization of MigA and WapR: putative rhamnosyltransferases involved in outer core oligosaccharide biosynthesis of Pseudomonas aeruginosa. J Bacteriol 190: 1857–1865.
41. Murray GL, Attridge SR, Morona R (2006) Altering the length of the lipopolysaccharide O antigen has an impact on the interaction of Salmonella enterica serovar Typhimurium with macrophages and complement. J Bacteriol 188: 2735–2739.
42. Al-Dabbagh B, Mengin-Lecreulx D, Bouhss A (2008) Purification and characterization of the bacterial UDP-GlcNAc:undecaprenyl-phosphate GlcNAc-1-phosphate transferase WecA. J Bacteriol 190: 7141–7146.
43. Hibbing ME, Fuqua C, Parsek MR, Peterson SB (2010) Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol 8: 15–25.
44. Yim L, Betancor L, Martinez A, Bryant C, Maskell D, et al. (2011) Naturally occurring motility-defective mutants of Salmonella enterica serovar Enteritidis isolated preferentially from non-human rather than human sources. Appl Environ Microbiol.
45. O'Toole GA, Kolter R (1998) Flagellar and twitching motility are necessary for Pseudomonas aeruginosa biofilm development. Mol Microbiol 30: 295–304.
46. Wood TK, Gonzalez Barrios AF, Herzberg M, Lee J (2006) Motility inuences biofilm architecture in Escherichia coli. Appl Microbiol Biotechnol 72: 361–367.
47. Starnbach MN, Lory S (1992) The iA (rpoF) gene of Pseudomonas aeruginosa encodes an alternative sigma factor required for agellin synthesis. Mol Microbiol 6: 459–469.
48. Tasteyre A, Barc MC, Collignon A, Boureau H, Karjalainen T (2001) Role of FliC and FliD agellar proteins of Clostridium difficile in adherence and gut colonization. Infect Immun 69: 7937–7940.

49. Schoenhals GJ, Macnab RM (1999) FliL is a membrane-associated component of the agellar basal body of Salmonella. Microbiology (Reading, Engl) 145 (Pt 7): 1769–1775.
50. Waters RC, O'Toole PW, Ryan KA (2007) The FliK protein and agellar hook-length control. Protein Sci 16: 769–780.
51. Titz B, Rajagopala SV, Ester C, Hauser R, Uetz P (2006) Novel conserved assembly factor of the bacterial agellum. J Bacteriol 188: 7700–7706.
52. Liu J, Lin T, Botkin DJ, McCrum E, Winkler H, et al. (2009) Intact agellar motor of Borrelia burgdorferi revealed by cryo-electron tomography: evidence for stator ring curvature and rotor/C-ring assembly exion. J Bacteriol 191: 5026–5036.
53. Soto GE, Hultgren SJ (1999) Bacterial adhesins: common themes and variations in architecture and assembly. J Bacteriol 181: 1059–1071.
54. Aprikian P, Interlandi G, Kidd BA, Le Trong I, Tchesnokova V, et al. (2011) The bacterial fimbrial tip acts as a mechanical force sensor. PLoS Biol 9: e1000617.
55. Jones CH, Pinkner JS, Roth R, Heuser J, Nicholes AV, et al. (1995) FimH adhesin of type 1 pili is assembled into a fibrillar tip structure in the Enterobacteriaceae. Proc Natl Acad Sci USA 92: 2081–2085.
56. Hahn E, Wild P, Hermanns U, Sebbel P, Glockshuber R, et al. (2002) Exploring the 3D molecular architecture of Escherichia coli type 1 pili. J Mol Biol 323: 845–857.
57. Valenski ML, Harris SL, Spears PA, Horton JR, Orndorff PE (2003) The Product of the fimI gene is necessary for Escherichia coli type 1 pilus biosynthesis. J Bacteriol 185: 5007–5011.
58. Craig L, Li J (2008) Type IV pili: paradoxes in form and function. Curr Opin Struct Biol 18: 267–277.
59. Schulein R, Dehio C (2002) The VirB/VirD4 type IV secretion system of Bartonella is essential for establishing intraerythrocytic infection. Mol Microbiol 46: 1053–1067.
60. Robinson C, Bolhuis A (2001) Protein targeting by the twin-arginine translocation pathway. Nat Rev Mol Cell Biol 2: 350–356.
61. Mori H, Ito K (2001) The Sec protein-translocation pathway. Trends Microbiol 9: 494–500.
62. Muller M (2005) Twin-arginine-specific protein export in Escherichia coli. Res Microbiol 156: 131–136.
63. Dilks K, Rose RW, Hartmann E, Pohlschroder M (2003) Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey. J Bacteriol 185: 1478–1483.
64. Cornelis GR, Van Gijsegem F (2000) Assembly and function of type III secretory systems. Annu Rev Microbiol 54: 735–774.
65. Sansonetti PJ (2001) Microbes and microbial toxins: paradigms for microbial-mucosal interactions III. Shigellosis: from symptoms to molecular pathogenesis. Am J Physiol Gastrointest Liver Physiol 280: G319–323.
66. Celli J, Deng W, Finlay BB (2000) Enteropathogenic Escherichia coli (EPEC) attachment to epithelial cells: exploiting the host cell cytoskeleton from the outside. Cell Microbiol 2: 1–9.
67. Farizo KM, Huang T, Burns DL (2000) Importance of holotoxin assembly in Ptl-mediated secretion of pertussis toxin from Bordetella pertussis. Infect Immun 68: 4049–4054.
68. Zink SD, Pedersen L, Cianciotto NP, Abu-Kwaik Y (2002) The Dot/Icm type IV secretion system of Legionella pneumophila is essential for the induction of apoptosis in human macrophages. Infect Immun 70: 1657–1663.
69. Boschiroli ML, Ouahrani-Bettache S, Foulongne V, Michaux-Charachon S, Bourg G, et al. (2002) Type IV secretion and Brucella virulence. Vet Microbiol 90: 341–348.
70. Backert S, Churin Y, Meyer TF (2002) Helicobacter pylori type IV secretion, host cell signaling and vaccine development. Keio J Med 51 Suppl 2: 6–14.
71. Schroder G, Dehio C (2005) Virulence-associated type IV secretion systems of Bartonella. Trends Microbiol 13: 336–342.
72. Bingle LE, Bailey CM, Pallen MJ (2008) Type VI secretion: a beginner's guide. Curr Opin Microbiol 11: 3–8.
73. Potvin E, Lehoux DE, Kukavica-Ibrulj I, Richard KL, Sanschagrin F, et al. (2003) In vivo functional genomics of Pseudomonas aeruginosa for high-throughput screening of new virulence factors and antibacterial targets. Environ Microbiol 5: 1294–1308.
74. Persson OP, Pinhassi J, Riemann L, Marklund BI, Rhen M, et al. (2009) High abundance of virulence gene homologues in marine bacteria. Environ Microbiol 11: 1348–1357.
75. Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449: 835–842.
76. Gross R, Guzman CA, Sebaihia M, dos Santos VA, Pieper DH, et al. (2008) The missing link: Bordetella petrii is endowed with both the metabolic versatility of environmental bacteria and virulence traits of pathogenic Bordetellae. BMC Genomics 9: 449.
77. Fry NK, Duncan J, Malnick H, Warner M, Smith AJ, et al. (2005) Bordetella petrii clinical isolate. Emerging Infect Dis 11: 1131–1133.
78. Stark D, Riley LA, Harkness J, Marriott D (2007) Bordetella petrii from a clinical sample in Australia: isolation and molecular identification. J Med Microbiol 56: 435–437.
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
80. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479–2481.
81. Dearden JC, Cronin MT, Kaiser KL (2009) How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ Res 20: 241–266.