# CALANGO: an annotation-based, phylogeny-aware comparative genomics framework for exploring and interpreting complex genotypes and phenotypes

**AUTHORS**

Jorge Augusto Hongo[1∞], Giovanni Marques de Castro[2∞], Agnello César Rios Picorelli[2], Thieres Tayroni Martins da Silva[2], Eddie Luidy Imada[3], Luigi Marchionni[3], Anderson Vieira Chaves[2], Gabriel Magno de Freitas Almeida[4], Felipe Campelo[2,5], Francisco Pereira Lobo[2*]

1 – Instituto de Computação, Universidade Estadual de Campinas, Brazil.

2 - Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

3 – Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, New York.

4 - Department of Biological and Environmental Science, University of Jyvaskyla, Jyvaskyla, Finland.

5 - Aston University, Birmingham, England.

∞ - Both authors contributed equally to this work.

* - Corresponding Author (franciscolobo@ufmg.br, franciscolobo@gmail.com)

**ABSTRACT**

The increasing availability of high-quality genomic, annotation and phenotypic data for different species contrasts with the lack of general software for comparative genomics that integrates these data types in a statistically sound framework in order to produce biologically meaningful knowledge. In this work, we present CALANGO (Comparative AnaLysis with ANnotation-based Genomic cOmponentes), a first-principles comparative genomics tool to search for annotation terms, such as GO terms or Pfam domain IDs, associated with a quantitative variable used to rank species data, after correcting for phylogenetic relatedness. This information can be used to annotate genomes at any level, including protein domains, genes, or promoters, allowing comparative analyses of genomes at several resolutions and from distinct functional and evolutionary angles. CALANGO outputs a set of HTML5 files that can be opened in any conventional web browser, featuring interactive heatmaps, scatter plots, and tables, stimulating scientific reproducibility, data sharing, and exploratory analysis. Detailed results and a reproducibility-focused data structure are also returned after each run of the tool. CALANGO provides classic association statistics used in comparative genomics, including correlation coefficients, probabilities, and phylogeny-aware linear models. To illustrate how CALANGO can be used to produce biologically meaningful, statistically sound knowledge, we present a case study of the co-evolution of *Escherichia coli* and their integrated bacteriophages (prophages). Through controlled *in silico* experiments, we demonstrate that terms from a functional annotation are both more prevalent across genomes and more abundant than the homology-based annotation terms commonly used in traditional comparative genomics studies. This result demonstrates how GO-based annotation captures information of non-homologous sequences fulfilling the same biological roles. Most homologous regions positively associated with prophage occurrence are found in genes of viral origin (e.g. capsids, lysozymes, and integrases), as expected, while the second most abundant category is virulence factors. The removal of viral genes demonstrated that most of the virulence factors associated with prophage density are located outside viral genes, suggesting a more complex biological

process than the archetypal bacteriophage-mediated horizontal gene transfer of virulence factors. The functional annotation performed by CALANGO revealed several GO terms describing general and specific aspects of viral biology (e.g. “viral life cycle”, “DNA integration”). We also found an association of the GO term “pathogenicity”, used to annotate several non-homologous virulence factors, as well as terms describing several known virulence mechanisms in pathogenic *E. coli* (e.g. “Type III secretion system”). Moreover, CALANGO also detected previously unknown associations that unveil a richer scenario of the bacteriophage-host biological interaction. An interesting example is the association of GO term “response to stress”, used to annotate several classes of non-homologous genes components of distinct stress response mechanisms (e.g. peroxidases, DNA repair enzymes, heat shock proteins), indicating that the horizontal transfer of such genes may be adaptive for host cells and, consequently, advantageous for the integrated prophages as well. CALANGO is provided as a fully operational, out-of-the-box R package that can be freely installed directly from CRAN. Usage examples and longer-format documentation are also available at https://fcampelo.github.io/CALANGO/.

**KEYWORDS:** comparative genomics, exploratory analysis, complex phenotypes, comparative methods, functional comparative genomics.

**ABBREVIATIONS:** QVAL - quantitative values across lineages; MHT - multiple hypothesis testing

# 1 - BACKGROUND

Life is one of the most complex and diverse physical phenomena in the known universe, displaying a wide range of variation across phenotypes and genotypes. Living organisms span a mass range of over 20 orders of magnitude, from acellular, self-replicating biomolecules to complex single and multicellular organisms. The number of protein-coding genes and genome sizes also vary by several orders of magnitude [1, 2]. Several phenotypes and genotypes can be described as quantitative variables. These may directly represent a variable under interest (e.g., body mass, genome size, or number of protein-coding genes) or act as proxy variables that capture the feature under analysis (e.g., number of distinct cell types as a proxy for organism complexity). In this work, we will refer to such quantitative phenotypes/genotypes variables as Quantitative Values Across Lineages (QVAL).

Over centuries of research, vast amounts of quantitative data describing genotypic and phenotypic variation across species have been collected, being currently available as searchable databases and other structured and unstructured data sources [3, 4]. Nowadays, even relatively modest research budgets allow the use of massively parallel DNA sequencing to generate high-quality genomic data for non-model organisms, providing an increasing number of complete genomes across the tree of life [5]. Therefore, both QVAL information and high-quality genomic sequences are increasingly available for several major groups of interest.

Consequently, the main challenge of extracting biologically meaningful knowledge from phenotypic and genomic data for different species lies no longer lies in obtaining this data, but rather in analyzing these heterogeneous data types in a biologically meaningful manner under a comparative and evolutionary genomics framework, to gain insights into the putative genomic mechanisms associated with - and eventually playing causal roles on - the evolution of inherited quantitative complex phenotypes [6].

The abundance of genomic and phenotypic information for several species contrasts with the absence of a general comparative genomics workflow

to consolidate both data sources for the extraction of context-relevant, statistically sound results and successful generation of new biological insights and hypotheses [6]. Several factors have contributed to the lack of software dedicated to such analyses. Possibly the most important relates to the challenge of developing a data-modeling schema and a statistical workflow that accounts for controllable sources of bias when dealing with genomic and phenotypic data from phylogenetically related species.

Several comparative genomic strategies have been developed to search for an association between homologous genomic regions and complex phenotypic/genotypic properties [6]. Despite their success in detecting gene-phenotype associations, we argue these methods fail to take into account current genomic knowledge when searching for possible functional molecular convergences associated with complex phenotypes. Genomes are composed of distinct functional and structural components, such as genes, regulatory regions, repeats, operons and, transposable elements, each having interdependent biological roles that may contribute to the emergence of a given phenotype [7].

Current strategies to search for an association between complex phenotypes and sequence-level information (such as homologous genomic regions shared across genomes) fail to incorporate the huge amount of available annotation information of distinct genomic components [6]. For instance, genes coding for distinct virulence factors in pathogenic organisms may not share a common evolutionary origin but nevertheless perform similar biological functions that contribute to a pathogenicity phenotype. Yet, homology-based comparative genomic analyses would fail to recognize such non-homologous factors that share the same biological role. In comparative genomics studies, the association of genomic elements to their biological roles currently relies on laborious *ad hoc* classification schemas of genomic elements to arbitrary functional categories [2].

A common approach to capturing additional layers of biological knowledge, beyond sequence-level information, is the annotation of genomic components using controlled dictionaries of biological roles, such as Gene

Ontology (GO) terms, followed by the analysis of the distribution of biological roles in different experimental conditions [8]. This annotation methodology allowed, for instance, the development of statistical procedures such as enrichment analysis and similar strategies, which represent the current state-of-the-art in supporting the interpretation of several categories of "omics" data to gain biologically meaningful insights [9].

Another major issue when dealing with genomic or phenotypic data from phylogenetically related species is the lack of proper control for non-independence due to relatedness by common ancestry [10]. Most statistical procedures commonly used in comparative studies, such as linear model building and correlation analysis, rely on the distributional assumption of independence, which is violated when analyzing species that evolved from a common ancestor [10, 11]. Despite some interesting studies and tools that use phylogeny-aware models to test for the individual association between phenotypic and genotypic variables [12-14], we are not aware of any general tool to compute and explore genomic-scale, phylogenetic-aware models to search for putative associations between QVAL and genomic components annotated using controlled dictionaries.

The development of such a tool is warranted, from the computational and statistical perspectives, by the need to adequately correct significance levels to account for the large number of hypotheses that are inevitably tested in parallel [15]; and data quality control, to mitigate the effects of differences in genomic data quality arising from common bioinformatics procedures used to generate it (such as genome assembly, gene prediction and annotation [16]) on downstream analysis of this data [17]. The presentation of results in the form of a consolidated, standardized, and easily accessible format is another requirement to facilitate the analysis and reproducibility of results.

Although there are tools that deal with some of the aforementioned issues [6], we are not aware of any existing software capable of simultaneously tackling all of them when analyzing genomic, annotation, and phenotypic data in a comparative manner, to provide biological knowledge and genomic

interpretation for complex QVAL in a user-friendly and fully reproducible environment. In this article, we report CALANGO (Comparative AnaLysis with ANnotation-based Genomic cOmponentes), a first-principles, general comparative genomics tool that is explicitly designed to use annotation information and phylogeny-aware protocols, from data processing and modeling to visualization. As mentioned earlier, traditional comparative genomics studies rely heavily on homologous elements shared across genomes as a common set of elements when comparing genomes. In contrast, CALANGO works by searching for biologically meaningful and formally defined annotation terms associated with some quantitative variable used to rank distinct species. This tool can therefore be used not only to search for homologous elements associated with QVAL values using annotation schemas that reflect sequence homology, but also to explore higher-order biological knowledge, using a function-based dictionary as provided, e.g., by GO, that models the functional convergences of non-homologous genes that perform the same biological roles or are components of the same biological processes.

To demonstrate how CALANGO can be used to extract statistically sound, biologically meaningful knowledge from publicly available data, we present a case study on the co-evolution of pathogenic and commensal *Escherichia coli* and their integrated bacteriophages (prophages). We found that terms from GO-based annotation are both more prevalent across genomes and more abundant than homology-based annotation terms commonly used in traditional comparative genomics studies, therefore providing strong evidence that a GO-based annotation effectively captures the information of non-homologous sequences fulfilling the same biological roles.

After verifying that pathogenic *E. coli* have a greater occurrence of prophages in their genomes, we observed that most of the homologous regions positively associated with prophage occurrence are found in genes of viral origin (e.g., capsids, lysozymes, and integrases), as expected. The second most abundant category is virulence factors. A controlled *in silico* experiment consisting of the removal of viral genes demonstrated that most of the virulence factors associated with prophage density are located outside these regions, suggesting a

more complex biological process than the bacteriophage-mediated horizontal gene transfer of virulence factors. The GO-based annotation found several GO terms that describe both general and specific aspects of viral biology (e.g. "viral life cycle", "DNA integration"), the most general GO term to describe a virulence factor ("pathogenicity") and several virulence mechanisms known to play important roles in pathogenic *E. coli* (e.g. "Type III secretion system"). CALANGO also detected previously unknown associations that provide deeper insights into the bacteriophage-host biological interaction. For instance, we found the association of GO terms describing several classes of stress response genes with prophage occurrence, indicating that the bacteriophage-mediated horizontal transfer of such genes may be adaptive for host cells.

The results of this comprehensive case study showcase how CALANGO can leverage annotation information and phylogeny-aware protocols to enable the investigation of sophisticated biological questions. CALANGO is freely available as an R package, which can be installed directly from CRAN (https://cran.r-project.org/package=CALANGO) and also from the project website on Github (https://fcampelo.github.io/CALANGO/), where usage examples and long-format documentation can be found.

# 2 - RESULTS

## 2.1 - CALANGO - a brief overview

CALANGO provides a flexible and user-friendly platform for interpreting the variation of QVAL in a phylogeny using current genomic knowledge (Figure 1; Supplementary File 1, section "CALANGO development"). To address the challenges in performing this task, our software package was inspired by several approaches previously used in fields as diverse as enrichment analysis [18] and comparative methods in biology [10]. Starting with a set of species and their associated genomic, phylogenetic, and phenotypic data, CALANGO searches for annotation terms associated with a quantitative phenotypic/genotypic value used to rank species, using both classical statistical methods to detect the

association between variables in large-scale searches as well as comparative methods to deal with the hierarchical, non-independent data structures that emerge when analyzing phylogenetically related organisms [11].

CALANGO outputs a dynamic HTML5 report containing graphical elements and tables with association statistics and other useful quantities, intended to instigate users to interact and actively interpret the results. The full set of user-defined input parameters is also returned alongside all computed results as a *list* object, which can be easily integrated into other bioinformatics pipelines to expand the potential applicability and reproducibility of the results and analyses executed by CALANGO.

## 2.2 – Case study: Homologous regions and biological roles associated with prophage density in pathogenic and free-living *Escherichia coli*

To demonstrate how CALANGO can be used to extract biologically meaningful insights as well as to produce new, testable hypotheses, we applied the tool to investigate annotation terms associated with a complex genomic property, namely the prophage density, in pathogenic and non-pathogenic *Escherichia coli*. *E. coli* is arguably one of the most well-characterized model organisms, with a wealth of phenotypic, genomic, and annotation information readily available. *E. coli* lineages also possess a remarkable genomic diversity, sharing as little as 11% of their core genes [19]. This genomic diversity is reflected in the distinct ecological niches occupied by this bacterium, which is commonly found in several body niches of animal hosts as a commensal or pathogen, and also as a free-living organism in the environment. Interestingly, up to 26% of accessory genes in *E. coli* lineages are found within integrated bacteriophages (prophages), indicating a major role of virus-mediated horizontal gene transfer in *E. coli* genomic composition [20].

Bacteriophage infections are not always deleterious to their bacterial hosts. On one side there are obligate lytic phages, agents of cell death and population control, and on the other, there are persistent lysogenic phages, responsible for gene transfer and mutualism. In a microbial population the lysis-

lysogeny events are dynamic and extremes of a continuum comprising antagonistic and beneficial biological interactions with contrasting effects to the bacterial host fitness [21]. During a lytic cycle, viruses control the host's cell machinery to produce new viral particles that are released into extracellular media by cell lysis. Consequently, successful lytic cycles decrease the fitness of susceptible bacterial populations. Conventional lysogenic cycles result in the integration of bacteriophage genomes into host chromosomes, producing new biological entities known as prophages. A prophage can remain inside a bacterial genome for many generations, often providing beneficial traits to the bacterium, such as superinfection exclusion, increased virulence, or novel metabolic options [22].

Stress signals such as starvation, DNA damage, or quorum sensing trigger prophage excision, initiating a lytic cycle that ends in cell death and spread of the lysogenic phage, eventually resulting in new cells containing the prophage. Prophage insertion and excision events may also bring fragments from a previous host's genomes through transduction, which can eventually contain genes that increase the host's fitness, besides the ones coded by the lysogenic phage itself. Therefore, bacteriophages play contrasting roles in bacterial evolution, contributing to bacterial fitness but also eventually causing host cell death [23]. Virulence factors represent an archetypal example of genes horizontally transferred across bacteria by bacteriophages that can increase fitness in new hosts, contributing to pathogenicity phenotypes in several unrelated bacteria [24]. Despite being a well-known phenomenon, we are not currently aware of any systematic evaluation of the association between prophage occurrence and the abundance of virulence factors across bacterial genomes.

*E. coli* evolution and its interaction with bacteriophages comprise a rich evolutionary scenario to evaluate the analytic capabilities of CALANGO. In brief, *E. coli* 1) has a wealth of genomic and phenotypic information readily available; 2) occupies two distinct niches regarding host interaction (pathogenic *versus* non-pathogenic); and 3) has high genome plasticity, with prophages playing a major role in this variation. Therefore, we were interested in evaluating the

hypothesis that the frequency of virulence factors in *E. coli* genomes is associated with prophage occurrence, after removing possible biases due to common ancestry. Furthermore, we were also interested in evaluating whether CALANGO would detect previously unknown biologically relevant phenomena associated with prophage abundance in *E. coli* genomes.

To demonstrate CALANGO's flexibility when analyzing genomes at different levels regarding their genomic components and annotation schemas, we proceeded by performing two distinct analyses (Figure 2A): 1) a domain-level analysis (from now on referred to as *domain2Pfam*), in which individual protein domains predicted by the Pfam database were annotated using a traditional homology-based annotation schema (Pfam IDs, where each ID represents a domain and, consequently, a different homologous region) to emulate the results traditionally obtained through standard comparative genomics analyses; 2) the same domain-level genomic components of step "1" annotated using GO terms (from now on referred as *domain2GO*), used to highlight how CALANGO may identify functional convergences of non-homologous sequences not detectable in the first analysis. It is worth mentioning that, using the same genomic elements (protein domains as predicted by Pfam) annotated using two distinct schemas (either Pfam or GO) allows us to objectively evaluate the influence of annotation schema in our results.

### 2.2.1 - Pathogenic *E. coli* lineages have a greater occurrence of integrated viruses compared with non-pathogenic strains

After a thorough literature review and analysis, we selected 80 high-quality *E. coli* genomes (chromosomes and plasmids) that could be unambiguously classified as either pathogenic (51) or non-pathogenic (29) lineages, excluding laboratory strains (Supplementary Table 1; see also "*Acquiring genomic, phylogenetic and phenotypic data*" under the Methods section). We predicted the numbers and location of prophages in *E. coli* chromosomes using PHASTER [25] and found 729 prophages, with counts varying from one to 22 integrated viruses per *E. coli* genome.

We found pathogenic lineages to have a significantly higher count of prophages when compared with non-pathogenic ones (Wilcoxon rank-sum test, p = 4.60e-05, Figure 2B, "Prophage count"). We also observed pathogenic lineages to possess genome sizes significantly larger than non-pathogenic lineages (Wilcoxon rank-sum test, p = 1.36e-13, Figure 2B, "Genome length"). To exclude the possibility that pathogenic *E. coli* have more integrated prophages due to their larger genomes we proceeded by computing prophage densities for each genome, defined as the number of prophages divided by genome length, which was also found to be significantly higher in pathogenic lineages (Wilcoxon rank-sum test, p = 1.17e-04, Figure 2B, "Prophage density"). We also detected the association between higher prophage densities and a pathogenic phenotype while taking into account phylogeny (Supplementary File 1, section 2.1.2 - "Using CALANGO to survey downstream hypotheses"). CALANGO provided evidence that pathogenic *E. coli* have an increased prophage occurrence in their genomes even after accounting for genome size and phylogeny as a possible source of bias. We proceeded by using phage density as a QVAL to survey the *E. coli* dataset for biological functions associated with prophage occurrence.

### 2.2.2 - Functional annotations are more prevalent across genomes and more abundant than homology-based ones

From the set of 403,571 protein-coding genes observed in the 80 *E. coli* genomes under analysis, InterProScan annotation provided Pfam IDs for 351,234 of them (~87%, Supplementary Table 2). A total of 3,335 distinct Pfam IDs were observed more than five times in our dataset and were further surveyed by CALANGO for potential associations between their raw count values and prophage densities (see Supplementary File 1, section "Evaluating annotation term frequencies and counts" and Supplementary Figures 1B-C for the rationale of our decision to work with raw count data instead of normalized frequencies).

We found 230 Pfam IDs (6,8%) to be significantly associated with prophage density (corrected q-values for both phylogeny-aware models and Pearson's correlation < 0.1 and an occurrence of five or greater). Of these, 207

presented positive correlation (with values between 0.26 and 0.85) and 23 showed negative correlation (values between -0.26 and -0.50). (Supplementary Table 2, sheet "domain2PfamCount"). As for the *domain2GO* annotation, from the set of 1963 GO terms found to annotate at least five distinct protein domains as predicted by Pfam, CALANGO found 219 (11%) to be significantly associated with prophage density, with 193 positively correlated terms (correlations between 0.25 and 0.86) and 26 negatively associated (correlations between -0.27 and -0.54) (Supplementary Table 2, sheet "domain2GOCount"). As higher prophage densities are also associated with pathogenic *E. coli*, we used CALANGO output to demonstrate how archetypal examples of annotation terms from both Pfam and GO databases found to be associated with prophage density are also associated with the pathogenicity phenotype (Supplementary File 1, section 2.1.2 - "Using CALANGO to survey downstream hypotheses").

We investigated whether a GO-based annotation effectively integrates information from non-homologous domains that fulfill the same biological role by evaluating the prevalence of annotation terms, defined as the percentage of genomes where they were observed. This is a useful statistic to evaluate whether a GO-based annotation, which is expected to capture non-homologous functional similarities, is more observed across genomes when compared with a homology-based annotation schema. The sum of annotation term occurrence in all genomes also provides a proxy for the overall abundance of annotation terms in distinct annotation schemas. If a GO-based annotation of a set of genomic components integrates the information from non-homologous coding regions that fulfill the same biological roles, we expect it to have both a greater prevalence and a greater sum value than a Pfam-based annotation of the same set of components. Additionally, we also expect these initial differences in the occurrence of annotation terms to be reflected in the annotation terms found to be significantly associated with prophage density.

The distribution of prevalence values in all the 3729 annotation terms from *domain2Pfam* analysis observed in at least one genome revealed a highly skewed distribution, where most annotation terms are observed in the vast majority of genomes (Figure 2C, "Prevalence" chart, "domain2Pfam all"). We

found that 65.7% of the Pfam IDs (2450 out of 3729) are observed in more than 90% of the *E. coli* proteomes. A minor, second distribution peak is observed in Pfam IDs with the smallest prevalence values, comprising 11.6% of the Pfam IDs with prevalence < 0.1 (432 out of 3729 Pfam IDs).

The distribution of GO terms used to annotate the same set of genomic components (domain2GO experiment) has an even more biased distribution regarding terms with high prevalence values, where 88.8% (1768 out of 1992 terms found annotating at least one Pfam entry) have prevalence values greater than 0.9 (Figure 2C, "Prevalence" chart, "domain2GO_all"). Furthermore, only 2.3% of the GO terms have prevalence values smaller than 0.1 (46 out of 1992). This indicates that, even though the majority of terms from both functional- and homology-based annotation schemas are observed in most genomes, the terms from the functional-based schema are significantly more prevalent than those from a homology-based one (Wilcoxon test, p-value < 2.22e-16). The higher prevalence observed in our GO-based annotation suggests this annotation schema is indeed capable of better representing common biological themes shared across genomes than a homology-based one.

The higher values of prevalence observed in annotation terms from *domain2GO* appear to be reflected in significantly associated terms from this annotation schema (Figure 2C, "Prevalence" chart, "domain2GO_all" and "domain2GO_sig"), which were also found to be highly biased towards high prevalence values (80.8%, 177 out of 219). The observed difference in the median prevalence from all terms from *domain2GO* annotation and the significant terms from the same annotation was not found to be statistically significant (Wilcoxon test, p-value = 0.057).

In contrast to *domain2GO* annotation, the significant annotation terms from the *domain2Pfam* experiment have much lower prevalence values than the distribution observed in all *domain2Pfam* annotation terms (Figure 2C, "Prevalence" chart, "domain2Pfam_all" and "domain2Pfam_sig"). Only 25,7% (59 out of 230) of the significant annotation terms have prevalence values greater than 0.9. We also found the median prevalence value for the associated terms to

be significantly lower than both the one observed for all annotation terms from the *domain2Pfam* experiment and the significant terms from *domain2GO* experiment (Wilcoxon test, p-values < 2.22e-16).

We further evaluated annotation term occurrence in homology- and functional-based annotation schemas by computing the sum of occurrences of each annotation term in each annotation schema (Figure 2C, "Sum" chart). We initially observed that the homologous regions have a significantly smaller sum of occurrences than GO terms (medians of 80 and 227, respectively; Wilcoxon test, p-value < 2.22e-16). The distribution of significant annotation terms from homology-based annotations schema was also found to be more similar to the distribution of sums of all *domain2Pfam* terms (medians of 108.5 and 80, respectively; Wilcoxon test, p-value = 3.6e-04) than the significant associations of functional-based annotations when compared with the sum of term occurrences for *domain2GO* experiment (medians of 1125 and 227, respectively for associated GO IDs and all GO terms; Wilcoxon test, p-value < 2.22e-16). Furthermore, associated GO terms have significantly higher sum values than associated Pfam IDs (medians of 1125 and 108.5, respectively; Wilcoxon test, p-value < 2.22e-16).

The scenario where annotation terms from a GO-based annotation schema are both more prevalent and more abundant than terms from the Pfam-based annotation is compatible with the several classes of genes non-homologous genes of viral origin that play complementary roles in the viral life cycle found to be associated with prophage densities. These observations are also compatible with the integration of biological knowledge at the function level, where non-homologous Pfam domains observed in distinct genomes are annotated using the same GO term, therefore increasing both the prevalence and the abundance of GO terms in comparison to Pfam IDs. As we demonstrate below, GO-based annotation fulfills both expectations.

### 2.2.3 - Protein domain expansions associated with prophage density in *E. coli*

Figure 3 shows an annotated version of the heatmap produced by CALANGO using *domain2Pfam*, representing all Pfam domain IDs associated with prophage density.[1] This figure demonstrates how our tool integrates phylogenetic, phenotypic, and annotation data to allow downstream exploratory analysis. Species clustering is done using the phylogenetic tree, together with visualization of user-defined groups (pathogenic and non-pathogenic *E. coli* lineages, in our case), allowing the detection of interesting distribution patterns of annotation terms while considering, in our case, phylogenetic and ecological information. Pathogenic and non-pathogenic lineages are distributed with no clear grouping pattern, suggesting that both phenotypes emerged and/or were lost several times during the evolution of these lineages. Two pathogenic *E. coli* lineages have the highest count of most Pfam domains (Figure 3, red arrows). These lineages comprise enterohemorrhagic (EHEC) and enteropathogenic (EPEC) pathotypes, including all O157:H7 lineages (larger group), a Shiga-like toxin-producing serotype, and an important source of foodborne disease [26].

We found the domain clustering to comprise mostly two major functional classes: groups of non-homologous regions of viral origin that play several roles in the viral life cycle, such as integrases, structural proteins, lysozymes, and DNA metabolism enzymes, together with three main classes of virulence factors known to play important roles in EHEC and EPEC disease etiology: Type III secretion systems, urease, adhesins, hemolysins, NFkB-degrading protease, attaching and effacing virulence factors, and Shiga-like toxin [26-28]. The two groups of pathogenic *E. coli* with the highest count of associated Pfam IDs also contain most of the genomes where the virulence factor clusters were observed.

---

[1] Live versions of all HTML documents are available at CALANGO's Github examples page, https://fcampelo.github.io/CALANGO/articles/examples-page.html

#### 2.2.3.1 – Positive association – viral life cycle genes

To better characterize protein domains associated with prophage density, we manually curated all 230 protein domains associated with prophage density to their major biological roles. More than half of the positively associated domains with prophage density are found in genes that play several roles in the viral life cycle (125 out of 217, 57.6%), such as bacteriophage lysozymes (PF00959 - *Phage lysozyme*, Figure 4A), structural proteins of viral particles, holins, lysis proteins and integrases (Supplementary Figure 3A-D). Again, we used CALANGO's HTML output to further demonstrate how domains found in genes of viral origin are enriched in the associated domains (Supplementary File 1, section 2.1.2 – *Using CALANGO to survey downstream hypotheses*).

Since our case study uses prophage density as QVAL when ordering bacterial lineages, and integrated viruses are themselves genomic elements with defined genomic coordinates within bacterial genomes, it was possible to perform a controlled *in silico* experiment to evaluate the ability of our software to support the investigation of a possible causal relationship, defined here as annotation terms associated with phage density because they annotate protein-coding genes of viral origin (found within predicted prophages).

This experiment consisted of re-executing CALANGO while keeping the QVAL values fixed (i.e., as calculated previously) while removing all genes predicted as being of viral origin from the data (Supplementary File 1, section "Annotation terms associated with prophage density after removal of genes of viral origin"). This effectively blocks out possible effects of viral genes on the CALANGO output, allowing us to test whether the significant associations detected earlier between the annotation terms and the QVAL are indeed due to genes of viral origin, or whether other extraneous variables may be playing a role.

As expected, the vast majority of protein domains annotated as being of viral origin were no longer significantly associated with prophage density after the removal of genes of viral origin: only two out of the 125 domains annotated as playing a role in bacteriophage biology (~ 1.57%) were detected in the second

experiment (Supplementary Figure 2; Supplementary Table 2 contains all annotation terms associated with prophage density in these experiments). It is worth noting that one of these two domains (PF06316 - *Enterobacterial Ail/Lom protein*, Supplementary Figure 3E) is involved in the lysogenic life cycle [29], but also contributes to a pathogenicity phenotype in gram-negative bacteria [30]. The second domain (PF07799) codes for a DUF with no known biological function that may correspond to an unknown player in bacteria/bacteriophage biological interactions, and is consequently an interesting target for further experimental studies.

Together, these data indicate that the mechanism that associates the annotation terms previously classified through manual curation as having roles in the viral life cycle and found to be associated with prophage density is indeed the presence of genes of viral origin, rather than by other potential mechanisms. This *in silico* manipulative experiment showcases the ability of CALANGO to support the investigation of basic causal relationships by enabling a level of counterfactual investigation of observed associations in the data. While this is still short of a fully developed causal inference package for genomic data, the ability to uncover some causal relationships from data by *in* silico isolation and testing of the influence of possible confounders can already provide valuable insights into biologically meaningful phenomena, as illustrated by this case study.

#### 2.2.3.2 – Positive association – virulence factors and insertion elements

The second most abundant category of protein domains found to be associated with prophage density were several classes of virulence factors (58 domains out of 217, ~26.7%, Figure 3, Figure 4B). Interestingly, we found that 44 out of the 58 (75,9%) of domains annotated by us as virulence factors are still significantly associated with prophage density even after the removal of viral genes, suggesting that a considerable fraction of such genomic elements is not located within predicted prophage genomes (Supplementary Figure 2, Supplementary File 1, section "Annotation terms associated with prophage

density after removal of genes of viral origin"). In fact, from the 4379 domains coding for virulence factors, 3122 (71.3%) are located outside detectable regions of viral origin (Supplementary Table 3).

We observed that 7 out of the 14 virulence factor domains not associated with prophage density after the removal of the genes of viral origin are components of pathogenicity mechanisms long known to be horizontally transferred by bacteriophage integration, such as Shiga-like toxins and effectors of Type III secretion system [26, 31] (*e.g.* all 22 copies of domain PF02258 - *Shiga-like toxin beta subunit*, Supplementary Table 3). This provides additional evidence that our strategy to remove the effect of genes of viral origin was successful.

The unexpected observation that most virulence factors are located outside predicted prophages and remain associated with the QVAL after the removal of genes of viral origin suggests a more complex relationship between virulence factors and prophage occurrence. One possibility is that the presence of these domains is a consequence of previous bacteriophage integration events that, over time, resulted in 1) prophage degeneration by the loss of genes needed for viral replication, as a result of selective pressure for smaller bacterial genomes with lower replication times; and 2) the maintenance of genes contributing to a pathogenicity phenotype due to a fitness increase provided by such genes in pathogenic *E. coli* lineages [32] - a scenario compatible with the association of domain PF06316 after the removal of viral genes (Supplementary Figure 3E). A non-excluding alternative hypothesis is that some of these virulence domains may have arrived in bacterial genomes by non-phage-mediated horizontal transfer events and, together with virulence factors acquired through bacteriophage integration, contribute to a pathogenic phenotype. More complex hypotheses are also reasonable, such as the occurrence of synergism between the presence of a prophage and the external virulence factor, resulting in increased fitness to the bacteria when both are present.

We found five domains coding for insertion elements to be positively associated with prophage density, even after the removal of genes of viral origin (Figure 3, Supplementary Figure 2, Supplementary Figure 3J, PF13007 - *Transposase C of IS166 homeodomain*). Despite having been initially described as classic deleterious genomic parasites, insertion elements have been increasingly documented as important agents of bacterial genome evolution by providing the horizontal acquisition of genomic components that confer adaptation to new niches, including virulence factors, and, interestingly, may provide resistance to bacteriophage infection [33].

This data appears to suggest that an interplay of symbiosis and competition between insertion elements, integrated bacteriophages, and their hosts' genomes may be responsible for the associations observed, which does not appear to have been quantitatively evaluated prior to this study. Although there is plenty of evidence that bacteriophage integration contributes to pathogenic phenotypes in bacterial species through horizontal gene transfer of virulence factors [32, 36], we found no previous studies demonstrating a significant association between prophage density in bacterial genomes and the occurrence of virulence factors, while correcting for possible confounders and known biases, such as phylogeny and genome quality. Furthermore, our results demonstrate that the majority of virulence factors remain associated with prophage density even after the removal of genes of viral origin, therefore excluding the contribution of all known genes of viral origin for these associations. These unforeseen results showcase how CALANGO can be used to enable controlled *in silico* experiments for the investigation of biologically meaningful questions.

#### 2.2.3.3 – Positive association – other biological roles

Besides protein domains associated with phage biology, virulence factors, and insertion elements, CALANGO detected 36 additional domains positively associated with prophage density that fulfill several biological roles. Remarkably, these include four domains that code for anti-viral components of bacterial

immunity, such as putative components of the restriction-modification and CRISPR-Cas systems (e.g. PF13588 - *Type I restriction enzyme R protein N terminus (HSDR_N)*, Figure 4C). It may be tempting at this point to assume that integrated prophages constitute a selective pressure for the occurrence and maintenance of such mechanisms. However, these four domains are not associated with prophage density after removal of genes of viral origin, therefore suggesting that at least a fraction of them occur within prophage genomes. We hypothesize these could represent viral genes related to the evasion of bacterial immunity, but further studies are needed for confirmation. Furthermore, some components of anti-viral mechanisms were also found to be negatively associated with prophage density, as we shall see, suggesting a more complex interplay between integrated prophages and the occurrence of anti-viral mechanisms (Supplementary File 1, section "Distinct anti-viral mechanisms are positively and negatively associated with prophage density").

We also found 10 positively associated DUFs that may either comprise domains of viral origin playing roles in viral life cycles, or genes performing other biological functions found to be associated with prophage density, such as previously unknown virulence factors (Supplementary Table 2). Three out of the 10 DUFs remain significantly associated with the QVAL even after the removal of all predicted prophage genes, indicating that at least a fraction of them is located outside detectable integrated viral genomes. This property is shared with several of the known virulence factors we observed, making these domains interesting targets for downstream functional evaluation.

#### 2.2.3.4 – Protein domains negatively associated with prophage density

A total of 23 domains were found to have significant negative associations with prophage density (Figure 3), and 20 of them are associated after the removal of genes of viral origin. Most of these domains suggest a scenario where *E. coli* lineages with fewer integrated prophages, which tend to be non-pathogenic, have a set of genes enabling a greater diversity of lifestyles at several levels. At the molecular/metabolic level, we found domains coding for metabolic

functions (e.g. PF02615 - *Malate/L-lactate dehydrogenase*, Figure 4D) and transmembrane transporters involved in the uptake of ions and organic molecules (e.g. PF00950 - *ABC 3 transport family*, Supplementary Figure 3K).

We found two domains representing distinct facets of bacterial diversity at the population level, the first coding for a putative glycosyl hydrolase described with a known role in biofilm formation (PF14883 - *Hypothetical glycosyl hydrolase family 13* Supplementary Figure 3L) [34]. Bacteriophages can modulate biofilm formation in several bacterial species, and certain phages are efficient biofilm destroyers [35]. However, inside biofilms, susceptible bacterial populations could be protected from phage infections by their resistant counterparts [36]. In this case, the presence of the PF14883 domain in the bacterium could be a protective factor against phages, and explain the negative association detected by CALANGO. The second domain (PF17508 - *Microcin V bacteriocin*, Supplementary Figure 3M) codes for a bacteriocin, an umbrella term used to describe protein toxins produced by bacteria that inhibit the growth of closely related species [37]. The death of related competitors could be a protective measure against phages, allowing a bacteriocin-producing population to keep in check potential viral spreaders in its surroundings during kill-the-winner dynamics [38].

Up to this point, we observed a greater number of survival strategies in *E. coli* with less integrated viruses - ranging from molecular/metabolic to cellular/ecological – possibly reflecting a greater niche diversity in non-pathogenic *E. coli*, which tend to possess less integrated bacteriophages (Figure 2A). In contrast, lineages with greater prophage density tend to be pathogenic species specialized in occupying a particular niche (the bodies of their vertebrate hosts).

Phage-bacteria interactions inside a metazoan have been shown to differ from those happening in the environment or in laboratory conditions. At the metazoan-environment interface (mucosal layers), the pressure from phages might be greater than elsewhere, considering the impact of enriched mucosal-associated phages [39]. Also, the mucosal layer could influence lysis-lysogeny

decisions, and prophage-containing bacteria have been speculated to be protected in this environment by superinfection exclusion [40]. This correlates well to CALANGO findings that pathogenic *E. coli* tends to have more associated prophages, as the inserted viral genes could provide protection against phages colonizing the metazoans during the bacterial invasion process. This situation could also explain why some virulence factors coded by the bacteria are positively associated with prophages: the prophage-mediated protection could add up to the virulence gain, synergistically increasing the bacterial fitness in pathogenic situations.

We also observed a negative association of a component of an anti-viral mechanism, even after removing genes of viral origin (PF04313 - *Type I restriction enzyme R protein N terminus (HSDR_N)*), suggesting this domain may play some role in preventing viral infections in the lineages with a smaller prophage occurrence (Supplementary File 1, section "Distinct anti-viral mechanisms are positively and negatively associated with prophage density").

Some negatively associated domains may be evidence of previously unknown aspects of the ecology of *E. coli* and its phages related to the emergence of the virulence phenotype. Three interesting examples are the negative associations for two glycosyltransferase domains (PF08437 and PF01501) that are components of the lipopolysaccharide (LPS) biosynthesis pathway, and for one peptidoglycan binding domain (PF01471) found in genes involved in cell wall processes (Supplementary Figure 3N-P).

LPS are a major component of cell walls of gram-negative bacteria and, together with peptidoglycan, are also receptors of several bacteriophages during adsorption to the bacterial cell surface when starting the infection process [41]. Therefore, the decrease of these domains may be a consequence of the selective pressure caused by bacteriophage infection, resulting in surface modifications leading to the loss or change of molecules used by the phages as viral receptors. Acquisition of phage resistance through surface modification is detrimental to bacterial virulence for different pathogenic bacterial species [42, 43].

However, LPS molecules have also been long recognized as major activators of the innate branch of the vertebrate immune system [44]. Consequently, the decrease of these domains may also represent evidence of the selective pressures of vertebrate immune systems on pathogenic bacteria, resulting in a decrease in *E. coli* immunogenicity through the loss of bacterial cell wall components. It is possible that both effects may be happening concomitantly, with the selective pressure induced by prophage infection resulting in the decrease of genes coding for components of LPS and other components of the cell wall, but also causing an exaptation scenario where this selection enables pathogenic *E. coli* to better adapt to its niche by partially evading host immune system. This illustrates another powerful aspect of our software, namely its ability to elicit relevant testable hypotheses from the data. In the particular case of this study, it would be possible to evaluate the relative fitness contribution of these domains through targeted knockout in distinct *E. coli* lineages, followed by controlled experiments investigating vertebrate immune system evasion or prophage infection rates.

### 2.2.4 - Biological roles associated with prophage density in *E. coli*

The *domain2Pfam* analysis described in the last section required extensive curation of domain functions to provide proper biological context to our findings. Furthermore, by using Pfam IDs to annotate genomic components, we could only evaluate protein domains and annotation terms in a 1-1 relationship where each genomic component is annotated to a single Pfam ID. Finally, protein domain IDs represent homologous conserved regions within proteins and do not capture biological knowledge such as functional convergences of non-homologous genomic components.

As CALANGO models genomic components independently from annotation terms, it is possible to objectively evaluate the influence of distinct annotation schemas used to annotate the same set of genomic components. For that purpose, we annotated the set of protein domains from the earlier *domain2Pfam* experiment using their GO terms, when available, as provided by

InterProScan output files (experiment *domain2GO*, Supplementary Table 2, Figure 2B). At this point, we are interested in evaluating if GO annotation would detect the same major biological themes observed during our manual curation of *domain2Pfam* results. Additionally, we also evaluated if the information integration at the function level, made possible by GO annotation, allows the detection of biological functions associated with prophage density that was not immediately discernible in our *domain2Pfam* analysis.

#### 2.2.4.1 – Positive association – viral biology and pathogenicity

We found the majority of the 194 GO terms positively associated with prophage density to represent major aspects of bacteriophage biology and life cycle, including both general concepts and more specific components of the lytic and lysogenic cycles (Figure 5, Supplementary Table 2, sheet "domain2GOCount"). These findings contrast with the Pfam2Domain analysis, which found individual components of the viral life cycle that required extensive manual curation for the proper description of their biological context. As examples of more general GO concepts, we mention GO:0019058 (viral life cycle) and GO:0051704 (multi-organism process) (Figure 5A, Supplementary Figure 4A). Protein domains annotated to these GOs comprise non-homologous sequences that play complementary roles in bacteriophage life cycles and their relationships with its bacterial host, such as in the several structural components of viral particles (e.g., portal proteins, head-to-tail joining proteins, tail proteins, head proteins, and capsids) and in the proteins needed to cleave bacterial cell walls during lytic cycles (e.g. endolysins and spanins). Together, these results demonstrate how CALANGO integrates information from non-homologous sequences at the function level to automatically provide the same functional roles found in *domain2Pfam* analysis through manual curation and literature review.

CALANGO also detected several GO terms describing specific steps of bacteriophage life cycles. As examples of terms describing the initial steps of a lytic cycle (viral genome replication and viral particle assembly), we mention

GO:0006323 (DNA packaging; Figure 5B), GO:0006260 (DNA replication), GO:0019068 (virion assembly). and GO:0019069 (viral capsid assembly; Supplementary Figure 4B-D). After the assembly of new viral particles, the lytic cycle progresses by causing host cell lysis, resulting in the release of viral particles into the extracellular medium. As examples of GO terms describing aspects of this step of the lytic cycle, we mention GO:0003796 (lysozyme activity), GO:0061783 (peptidoglycan muralytic activity), and GO:0019076 (viral release from host cell; Figure 5C, Supplementary Figure 4E-F).

Another possibility of bacteriophage replication is the lysogenic cycle, which starts with the repression of the lytic cycle by the action of repressor proteins, followed by the integration of bacteriophage genomes into bacterial chromosome through recombination, a process represented by the association of GO:0008907 (*integrase activity*) (Figure 5D). A common theme of both bacteriophage life cycles is the occurrence of distinct domains that rely on DNA interaction, such as gene expression regulation, integration in host genomes, or transposition. The GO term GO:0003677 (DNA binding), also observed to be associated with prophage density (Supplementary Figure 5G), annotates several bacteriophage domains of distinct evolutionary origins annotated to this GO, such as repressor proteins, transposases, and integrases.

CALANGO also detected associated GO terms representing general and specific aspects of *E. coli* pathogenicity to be the second-largest group of annotation terms positively associated with prophage density. The GO term GO:0009405 (pathogenesis), found to be positively associated with prophage density, annotates 23 protein domains comprising several groups of non-homologous coding regions playing roles in several pathogenicity mechanisms in *E. coli*, such as cell invasion and adhesion, toxins, hemolysins, colicins and components of secretion systems (Figure 5E). In addition to the more general "pathogenesis" concept, CALANGO also detected other GO terms representing specific pathogenicity modules and mechanisms known to play important roles in distinct *E. coli* pathotypes, such as type III secretion and urease activity [46, 47] (Supplementary Figure 4H-I).

#### 2.2.4.2 – CALANGO uncovers new biological knowledge – association of prophage density and stress response mechanisms

Up to this point, GO annotation analysis detected the same major trends found after the manual curation of Pfam analysis results, mostly confirming its usefulness to detect biological knowledge associated with a quantitative variable across species, a procedure usually achieved through laborious manual curation in comparative genomics studies. Additionally, we were also interested to verify if our analysis found new, unexpected biological functions associated with prophage density, which would provide additional evidence that using CALANGO as a function-based annotation schema captures biological knowledge not detectable in traditional, homology-dependent comparative genomics analysis. We argue that a compelling example of this ability was demonstrated by the detection of GO:0006950 (*response to stress*) (Figure 5F), as detailed below.

This GO term annotates genes responsible for changes in the state or activity of a cell or organism, as a result of a disturbance in homeostasis such as temperature extremes, ionizing radiation, or infection by pathogens. Bacteriophages are a classical stress source in bacterial species and are known to trigger several stress response pathways such as the SOS and chaperone expression [35].

It may be appealing to assume that an increase in genes fulfilling this biological role is the consequence of an adaptation process of host cells to the presence of integrated prophages. However, even though we found some genomic elements likely to be cellular response mechanisms to viral infections, such as components of restriction-modification systems and DNA repair enzymes, a large fraction of these genes code for response mechanisms to other stress sources. These include peroxidases (response to redox stress), tellurite resistance (response to chemical stress), colicins (bacteriocins), and heat-shock proteins (response to unfolded protein stress), among others (Supplementary Table 5, Supplementary File 1, section "Stress response genes associated with *E. coli*").

In addition to the classical host-parasite biological interaction expected to occur between bacteriophages and their hosts, integrated prophages may also contribute to the fitness increase of bacteria specifically under conditions of stress, such as the host's immune response against pathogenic bacterial lineages [48] and adverse environmental conditions [45]. Here we provide the first genomic evidence that several classes of stress-response elements, found in both host and viruses genomes, contribute to the association of the biological process of stress response with prophage densities, and conclude that this association is likely to be a reflex of their ecological interactions in complex environments that go beyond the classical host-parasite biological interaction.

#### 2.2.4.3 – GO terms negatively associated with prophage density

We found 26 GO terms negatively associated with prophage density (Supplementary Table 2, sheet "domain2GOCount") that largely reflect the conclusions achieved at the domain level analysis (*domain2Pfam*) after extensive manual curation. We found GO terms reflecting protein-membrane transporters (e.g., GO:0043190 - *ATP-binding cassette (ABC) transporter complex*; Figure 5G), metabolic pathways (e.g., GO:0072330 - *monocarboxylic acid biosynthetic process*), and LPS biosynthesis (e.g., GO:0008918 - *lipopolysaccharide 3-alpha-galactosyltransferase activity*, GO:0035250 - *UDP-galactosyltransferase activity*; Figure 5H). Together, these results provide further evidence that starting from the same set of genomic elements annotated to distinct dictionaries of biological roles, CALANGO can automatically detect biologically meaningful associations that are comparable to our manual curation of the results and equivalent to traditional comparative genomics analysis.

## 3 - Conclusion

The post-genomic era has brought a plethora of high-quality sequenced genomes at several phylogenetic scales, ranging from previously underrepresented early-branching lineages of cellular organisms to thousands of

genomes of a single bacterial species. In contrast to this abundance of genomic data, there are currently no standardized methods in computational statistics for extracting genomic properties associated with a quantitative genotype/phenotype of interest. As recently reviewed by Lázló *et al.* [6], several tools are already available to search for associations between the patterns of occurrence of homologous genomic components and a binary phenotype of interest. The vast majority of these methods search for gene-phenotype associations, where the occurrence patterns of groups of homologous genes are evaluated for possible associations with a binary phenotypic trait.

Despite being highly successful to detect gene losses associated with the emergence of complex phenotypes in several eukaryotic groups, none of these methods uses current genomic knowledge at the function level, as all of them evaluate the distribution of sets of homologous elements across genomes. By dissociating the genomic components from their functional annotation, CALANGO provides the required flexibility to survey the distribution of several classes of genomic components, such as protein domains or entire genes. Furthermore, as we demonstrated, distinct annotation schemas allow both classic comparative genomics analysis - using annotations that reflect homology relationships - and comparative genomics at the function level – using annotations that capture functional molecular convergence. As exposed in our case study, both strategies are feasible in our tool and, together, deliver a richer, more biologically meaningful interpretation of the results.

Another important distinction is the class of phenotypic data accepted by these tools. Most of the methods for searching gene/phenotype associations are based on binary phenotypes (presence/absence) which, while useful to describe several phenotypes, cannot be used to survey quantitative data. Our tool is, to the best of our knowledge, the first one to search for associations of the occurrence of annotation terms of genomic components with a quantitative phenotypic/genotypic value across species.

From a computational and data analysis perspective, CALANGO was inspired by the enrichment analysis tools that evaluate thousands of gene sets

sharing common attributes, such as the same GO term, to provide biological meaning to large-scale omics experiments. Several of the most computationally demanding execution steps can take advantage of multicore parallelization, allowing users to survey large-scale association data in a feasible time. CALANGO outputs a list object containing all different quantities and statistics calculated in the analysis, as well as the parameters used in the analysis call, which advanced users can use to survey downstream hypothesis, integrate into bioinformatics pipelines, or share with colleagues for easy reproducibility and auditing of analyses. Our tool also generates a full HTML5 site containing interactive summaries, including association and distribution plots, heatmaps, and tables of statistics for each annotation term. This output is meant to simplify and stimulate data sharing, exploration, and interpretation, in addition to the full reproducibility enabled by the list object.

Here we demonstrated how CALANGO can be used for help interpreting complex biological interactions, finding both known phenomena and also providing new biological insight. More importantly, our tool also provided several interesting targets for testable hypotheses, ranging from potential virulence factors to new players in phage biology. We believe CALANGO represents a considerable step towards the establishment of an annotation-based, phylogeny-aware comparative genomics framework to survey genomic data beyond sequence level and search for associations between quantitative variables across lineages sharing a common ancestor and the multiple layers of biological knowledge coded in their genomes.

## 4- METHODS

### General overview of CALANGO

The CALANGO package is designed to be an open-source phylogeny-aware, annotation-based tool for comparative genomics, that searches for annotation terms associated with a quantitative genotypic/phenotypic variable used to rank lineages from a phylogeny.

## Genomic data modeling

The CALANGO package is designed to process distinct classes of genomic components, such as protein-coding genes, protein domains, and promoter regions, as well as their respective annotations, to extract biologically meaningful patterns of variation. CALANGO explicitly models a genome and the biological functions coded within it by using two associated concepts (Supplementary Figure 1). The first is a genomic component, which is defined as a common element type that can be discriminated in the group of genomes under analysis (Supplementary Figure 1; genomic component IDs in this case could be protein-coding genes, their translated protein sequences, promoters, or protein domains coded within a genome). The second is a set of controlled terms used to annotate the genomic components, which are expected to reflect some specific biological aspect thought to help answer the biological question under investigation. These can be, e.g., Pfam domain IDs or GO terms (Supplementary Figure 1, annotation term IDs). By explicitly dissociating genomic components from their annotation terms, CALANGO makes it possible to use annotation schemas designed to represent the functional similarities of genomic components that may not share a common ancestor, enabling comparative genomics at the function level.

## Obtaining compatible data for CALANGO

CALANGO was developed to be focused on flexibility and simplicity of use and distribution. Genomic features, annotation, and dictionary files are simple tabular files containing textual information used to describe genomic elements and their annotations (Figure 1A).

**Annotation data:** A Perl script (*calanguize_genomes.pl*) that can be used to parse Genbank files into high-quality genomic annotation compatible with CALANGO input data (Figure 1B) is distributed as part of the package. The current version of *calanguize_genomes.pl* annotates protein-coding genes and their conserved regions as predicted by InterProScan. In brief, the script performs the following steps:

A) Downloads genomic data for the species/individual to be analyzed;

B) Extracts the protein-coding genes described;

C) Provides a single coding sequence per locus, reporting only the longest coding sequence per locus to avoid possible biases introduced due to the larger number of isoforms described for model organisms [2];

D) Executes BUSCO for genome completeness evaluation [17];

E) Annotates all valid protein sequences using InterProScan [46];

F) Generates CALANGO-compatible files for annotated genomes and dictionaries of annotation terms;

**Phylogenetic tree data:** CALANGO currently supports fully dichotomous, ultrametric trees in the newick or nexus formats (Figure 1C). If users provide a tree with multichotomies, our tool uses the *multi2di* function from R package *ape* [47] to produce a dichotomous tree with branches of length zero.

**Metadata:** CALANGO expects a metadata file containing QVAL values and other information needed for proper execution (Figure 1A; see also CALANGO's documentation).

**Analyzing data using CALANGO**

CALANGO starts its analysis with (i) a set of genomic components from distinct genomes annotated using a common, controlled set of annotation terms; (ii) a dictionary file defining each annotation term in a biologically meaningful way (Figure 1D); and (iii) a metadata file containing genome-centered information, such as values for optional normalization of annotation count values in each genome (e.g. total count of annotation terms per genome), and the quantitative phenotype/genotype used to rank genomes. A full description of the initial files needed to execute CALANGO is available in Supplementary File 1, and

also on the project repository page. Users must also provide (iv) an ultrametric phylogenetic tree containing all lineages in a given analysis, which allows CALANGO to correct for phylogeny-related dependencies in the values of annotation terms in distinct genomes and QVAL values (Figure 1C and E, enabling further investigation of associations using traditional linear models that work under the assumption of independence of residuals [11] (Figure 1G).

After selecting a suitable type of genomic component to be surveyed and a dictionary containing annotation terms assumed to contain useful information for the extraction of biological insights related to the biological question, CALANGO proceeds by computing different classes of association statistics between the phenotype and annotation terms (Figure 1G). Our tool computes three commonly used correlation statistics (Pearson's, Spearman's, and Kendall's correlation values and their respective significance p-values) and a phylogeny-aware linear model constructed using Phylogenetic Independent Contrasts (PICs) [18] as implemented in the R package *ape* [51]. To account for the multiple hypothesis scenario of simultaneously searching for associations between QVAL and thousands of annotation terms, CALANGO reports FDR-corrected q-values for each association statistic. In addition to these, CALANGO also computes common statistics that measure annotation term variability (variance / standard deviation) as well as two customized statistics that summarize how abundant an annotation term is and how frequently it is observed in distinct genomes: the sum of annotation terms and their prevalence (the fraction of genomes where an annotation term is observed).

CALANGO provides two main output structures (Figure 1H-L). The first is an R *list* object containing all computed results, which can be used to survey specific downstream hypotheses (see section “Functional annotations are more prevalent across genomes and more abundant than homology-based annotations” for an example). This object also contains all input parameters used to generate the results, therefore providing a simple and convenient way to share results as well as all necessary parameters required for full reproducibility.

The second output is an interactive HTML5 website that can be easily shared, hosted online or browsed locally using any modern web browser. The CALANGO outputs were designed to intentionally facilitate a more transparent reporting of results and sharing of raw data and code, in line with a more general push towards better reproducibility of results in the scientific literature [42,43]. This user-friendly output facilitates the critical evaluation of all statistics provided by CALANGO in a dynamic tabular and visual manner. For association-based comparisons, the HTML5 website features searchable, sortable, and customizable tables with statistical outputs for each annotation term (correlation values and their corrected q-values, sum, prevalence, variance and standard deviation); as well as scatterplots and heatmaps, which allow users to easily survey several results for distinct annotation terms. These visualizations provide a good starting point to detect non-trivial patterns of association between genomic and phenotypic data, which can indicate putative answers to the questions of interest or suggest new follow-up hypotheses for further investigation.

## CALANGO development

CALANGO was developed as a CRAN-compliant R package [48] and makes full use of the existing R package ecosystem to implement its internal routines using code that has been validated by other researchers and developers. A complete listing of the CALANGO package dependencies is provided in Supplementary File 1 (section 1.1.3, "CALANGO package dependencies").

## Acquiring genomic, phylogenetic, and phenotypic data for *Escherichia coli*

A thorough literature review was performed to select *E. coli* genomes that could unambiguously be classified as either pathogenic or non-pathogenic (Supplementary Table 1). We evaluated genome assembly quality based on the expected gene content using BUSCO 3.0 and the *Enterobacteriales* order [17]. All genomes have BUSCO values of single-copy orthologs greater than 0.95.

To obtain an ultrametric phylogenetic tree for the 80 *E. coli* genomes we initially computed groups of homologous genes using OrthoMCL v2 [49] with default parameters and DIAMOND [50] for the sequence alignment step. At this point, we included two species from the same family as external groups: *Escherichia fergusonii* (ATCC 35469, NC_011740) and *Shigella dysenteriae* (Sd197, NC_007606) (Roy R. Chaudhuri e Ian R. Henderson 2012; H. Hata *et al.,* 2016). We proceeded by randomly selecting 25 universal 1-1 orthologs from OrthoMCL output file to be used for downstream phylogenetic analyses (Supplementary Table 4).

The global alignment of coding regions was done using MUSCLE [51], followed by the manual curation of alignments using MEGA X [52]. We used SequenceMatrix [53] to concatenate individual alignments in a single matrix using, and MEGA X to select the best nucleotide substitution models for individual gene alignments. We conducted Bayesian inference using MRBAYES [54] as provided by the CIPRES SCIENCE platform to estimate phylogenetic relationships based on the Bayesian information criterion (BIC) and assuming a minimal mutation rate (μ) of $2x10^{-10}$ mutations per nucleotide site per cell division, as described by [55]. The parameters of the posterior probability model, tree branch lengths, and tree topology were obtained using the *Metropolis-coupled Markov Chain Monte Carlo* (MCMCMC) algorithm. We executed two independent runs of two simultaneous chains for 20 million generations for each of the four MCMCMC, collecting tree samples every 500 generations and discarding the 20% first runs obtained during the burn-in step. Tracer v1.6 was used to evaluate MCMCMC convergence and the conjoint distributions of parameters and trees [56]. The final summarized tree topology was saved in *newick* format and used as input for our analyses.

**Manual curation of Pfam domains associated with prophage density**

We used a combination of individual description of Pfam IDs provided by InterProScan, together with the Pfam ID entries from the Pfam website [57] , to classify all Pfam IDs associated with prophage density to the broadest categories

that functionally encompasses them. Specifically, we used the following categories: "Unknown function", "Bacteriocin", "Biotin binding", "Biofilm formation", "DNA biology", "DNA metabolism", "DNA repair", "Energetic metabolism", "Flagellar motor", "LPS biosynthesis", "Membrane biology", "Multi-function", "Nitrogen metabolism", "Anti-viral immunity", "Cell wall metabolism", "Dissulfite bond formation", "Insertion element", "Membrane transporter", "Pathogenesis", "Phage biology", "Plasmid stability", "Stress response to oxidizing agents", "Toxin-antitoxin system", "Transcriptional regulator", "Zinc uptake" (Supplementary Table 2, sheet "domain2PfamCount").

## Data analysis/visualization

Unless explicitly stated, all figures presented were either direct CALANGO outputs or produced using CALANGO code and exported as vectorial images to generate higher quality files for publication. Venn diagrams were produced using Interactivenn [55]. All figures were aesthetically edited and annotated using GIMP for publication purposes and following the best scientific practices.

## Acknowledgments

We would like to thank Prof. Daniel dos Santos Mansur and Prof. Glória Regina Franco for the insightful discussions during the elaboration of this research. We also would like to thank the Graduate Program in Genetics, the Graduate Program in Bioinformatics, and the Vice Dean for Research from Universidade Federal de Minas Gerais, Brazil, for the financial support for this research and publication fees. This work was partially funded by CAPES/Brazil (Grant 001).

## Figure Legends

**Figure 1 - General schema of CALANGO.** A) Species-centered data files. B) Genomic/annotation data; C) phylogenetic data; D) Reading input data; E) Building of annotation vectors (number of occurrences of annotation terms per genome); F) Optional data normalization (e.g., relative frequencies of annotation terms); G) Statistical model building; H) Production of output files based on user-defined cutoffs; I) Heatmaps; J) Scatterplots of annotation term statistics; K) Interactive table and graphical outputs for individual annotation terms.

**Figure 2 – *Escherichia coli* data.** A) Experimental design. 1) 1) domain2Pfam (domain, as predicted by Pfam, are annotated using Pfam IDs to emulate a traditional, homology-based comparative genomics approach); 2) domain2GO (the same domains as predicted by Pfam in domain2Pfam experiment are annotated using GO terms to emulate a functional-based comparative genomics approach). B) Comparison of the number of prophages, genome length, and prophage density in pathogenic (green) and non-pathogenic (pink) *E. coli*. C) Prevalence and sum of annotation terms for distinct annotation schemas and annotation terms associated with prophage density. Red lines denote median prevalence values for each database.

**Figure 3 – Heatmap of protein domains associated with prophage density in *E. coli* as produced by CALANGO.** Species clustering was performed using the *E. coli* phylogeny produced in this study, annotation terms clustering was performed using Manhattan distance and average clustering. Red arrows indicate *E. coli* lineages with some of the highest counts of both viral proteins and virulence factors.

**Figure 4 – Pfam domains associated with prophage density in *E. coli*.** For each Pfam domain ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

**Figure 5 – GO terms associated prophage density in *E. coli*.** For each GO term ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

## Supplementary Figure Legends

**Supplementary Figure 1. Genomic components, annotation terms, and feature normalization.** A) Genomic components and their annotation terms: 1) Protein-centered annotation to Gene Ontology terms. A single protein-coding gene may be annotated to as many GO terms as needed. 2) Domain-centered annotation of protein domains to domain IDs, such as the ones available in Pfam. 3) Promoter-centered annotation of distinct binding sites of specific DNA binding proteins. 4) relationship between annotation term IDs and their definitions (in this case, hypothetical protein domain IDs and their biological roles). B) Hypothetical genomes and their associated metadata for normalization. The figure represents annotation terms with either the same number of copies in all genomes (orange boxes) or with a variable number of copies in all genomes (blue boxes). The table contains QVAL for these genomes, the normalization factor for annotation counts (sum of all annotation terms), raw annotation term counts, and their relative frequencies (raw counts of each annotation term in a genome normalized by the sum of all annotation terms in a genome). C) Correlation of QVAL, count, and frequency data. Each plot contains the correlation for the hypothetical variables in Supplementary Figure 1B. It is possible to observe that, the annotation term represented by blue boxes is associated with QVAL (both count and normalized values). The orange term, however, is not associated when considering count data, as it is present as a single-copy region in all genomes. As the relative frequencies of the orange annotation terms in each genome are dependent on the relative frequency of blue annotation terms, a relative increase in the frequency of blue annotation terms causes the relative decrease of orange annotation terms in the same genome.

**Supplementary Figure 2 - Heatmap of protein domains associated with prophage density in *E. coli* after the removal of genes of viral origin as produced by CALANGO.** Species clustering was performed using the *E. coli* phylogeny produced in this study, annotation terms clustering was performed using Manhattan distance and average clustering. Red arrows indicate *E. coli*

lineages with some of the highest values of domains of viral origin (before removal of genes of viral origin) and virulence factors.

**Supplementary Figure 3 – Additional examples of Pfam domains associated with prophage density.** For each Pfam domain ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

**Supplementary Figure 4 – Additional examples of GO terms associated with prophage density.** For each GO term ID, from left to right, CALANGO provides two plots for traditional association statistics (a scatterplot, together with a linear model, for direct data visualization, and a scatterplot of ranked data, together with a locally estimated scatterplot smoothing (LOESS)) and a plot containing the contrast data for phylogeny-aware linear regression of phylogenetically independent contrasts.

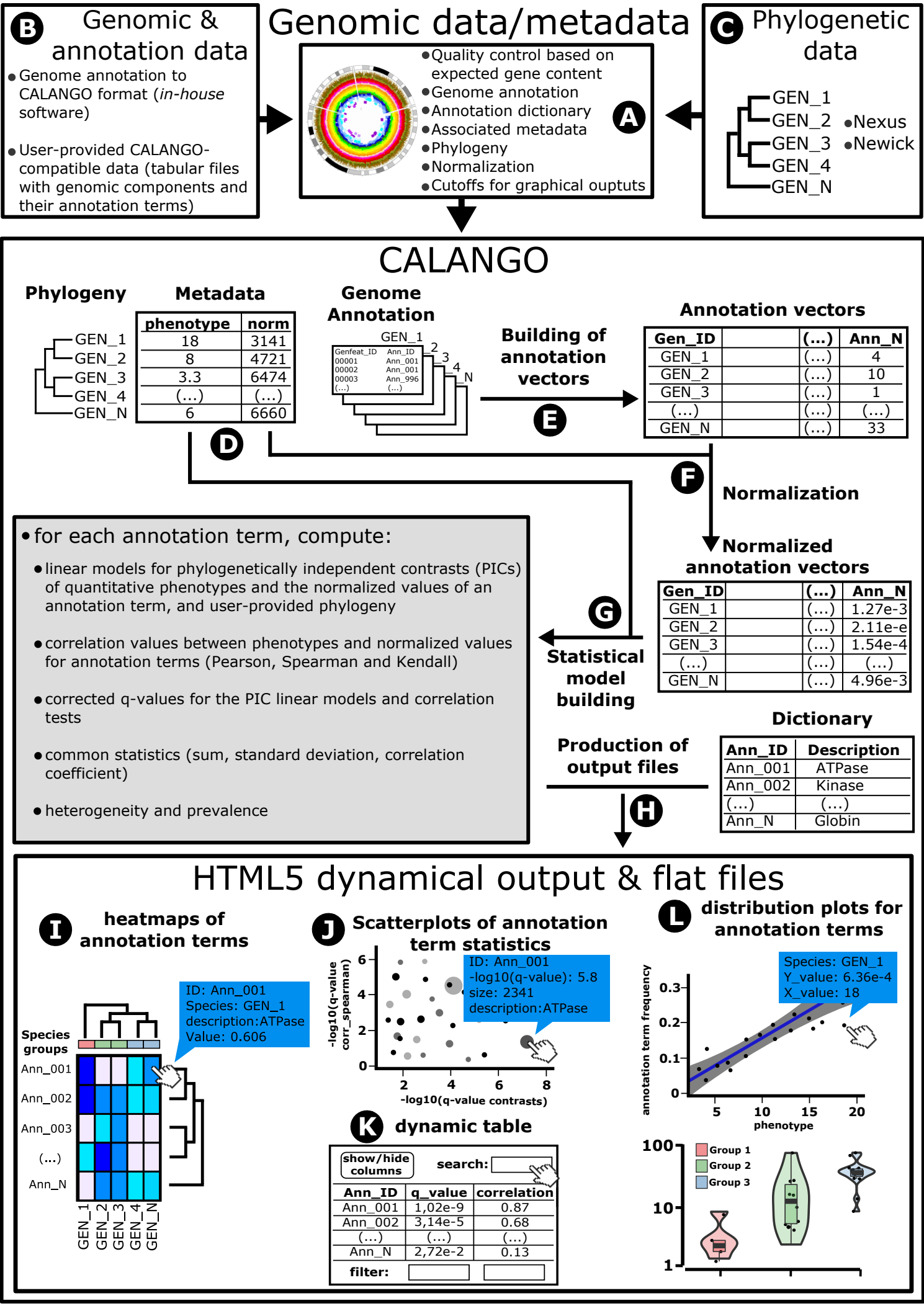
## References

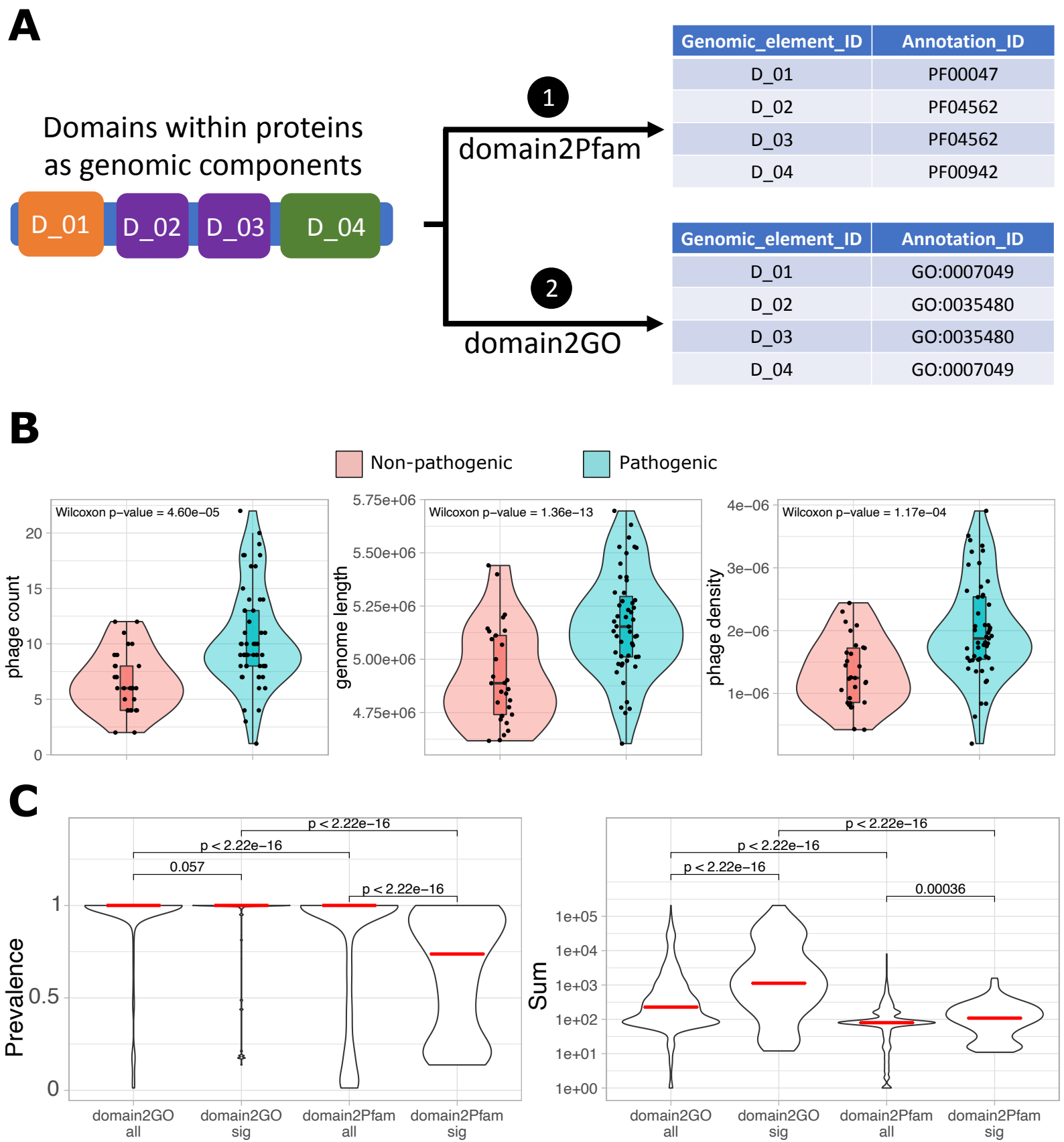

1. West GB, Brown JH: **The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization.** *J Exp Biol* 2005, **208:**1575-1592.
2. Vogel C, Chothia C: **Protein family expansions and biological complexity.** *PLoS Comput Biol* 2006, **2:**e48.
3. Groth P, Pavlova N, Kalev I, Tonov S, Georgiev G, Pohlenz HD, Weiss B: **PhenomicDB: a new cross-species genotype/phenotype resource.** *Nucleic Acids Res* 2007, **35:**D696-699.
4. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2010, **38:**D346-354.
5. Ellegren H: **Genome sequencing and population genomics in non-model organisms.** *Trends Ecol Evol* 2014, **29:**51-63.
6. Nagy LG, Merenyi Z, Hegedus B, Balint B: **Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing.** *Nucleic Acids Res* 2020.
7. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
8. The Gene Ontology C: **The Gene Ontology Resource: 20 years and still GOing strong.** *Nucleic Acids Res* 2019, **47:**D330-D338.
9. Marco-Ramell A, Palau-Rodriguez M, Alay A, Tulipani S, Urpi-Sarda M, Sanchez-Pla A, Andres-Lacueva C: **Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data.** *BMC Bioinformatics* 2018, **19:**1.
10. Cornwell W, Nakagawa S: **Phylogenetic comparative methods.** *Curr Biol* 2017, **27:**R333-R336.
11. Felsenstein J: **Phylogenies and the Comparative Method** *The American Naturalist* 1985, **125:**15.
12. Touchon M, Bernheim A, Rocha EP: **Genetic and life-history traits associated with the distribution of prophages in bacteria.** *ISME J* 2016, **10:**2744-2754.
13. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al: **The evolution of gene expression levels in mammalian organs.** *Nature* 2011, **478:**343-348.
14. Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G: **A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species.** *Cell Rep* 2012, **2:**817-823.
15. Hung JH, Yang TH, Hu Z, Weng Z, Delisi C: **Gene set enrichment analysis: performance evaluation and usage guidelines.** *Brief Bioinform* 2012, **13:**281-291.
16. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nat Rev Genet* 2012, **13:**329-342.
17. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO Applications from Quality**

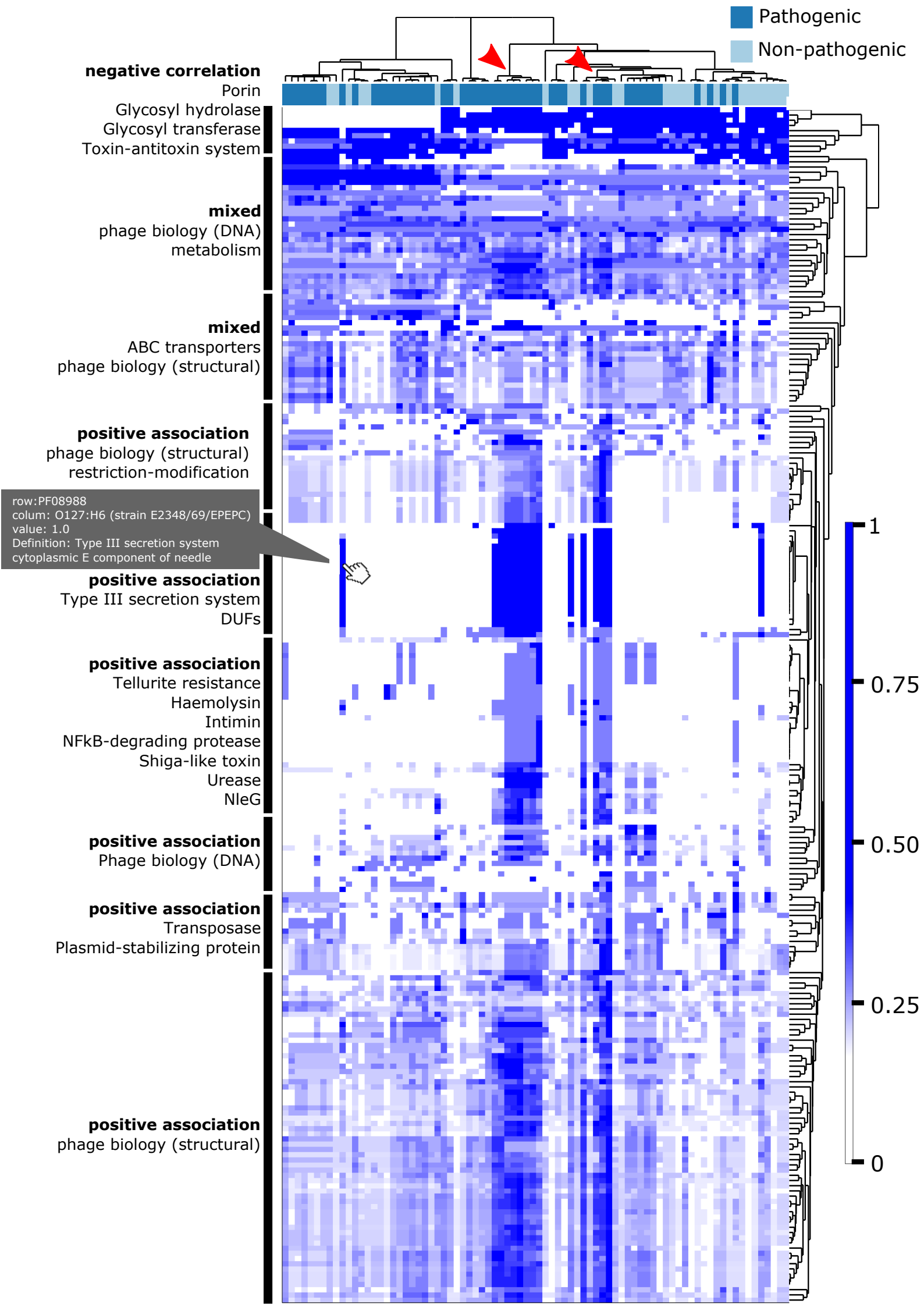
**Assessments to Gene Prediction and Phylogenomics.** *Mol Biol Evol* 2018, **35:**543-548.
18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
19. Kaas RS, Friis C, Ussery DW, Aarestrup FM: **Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes.** *BMC Genomics* 2012, **13:**577.
20. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.
21. Correa AMS, Howard-Varona C, Coy SR, Buchan A, Sullivan MB, Weitz JS: **Revisiting the rules of life for viruses of microorganisms.** *Nat Rev Microbiol* 2021.
22. Bondy-Denomy J, Davidson AR: **When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness.** *J Microbiol* 2014, **52:**235-242.
23. Lai JYH, Zhang H, Chiang MHY, Lun CHI, Zhang R, Lau SCK: **The putative functions of lysogeny in mediating the survivorship of Escherichia coli in seawater and marine sediment.** *FEMS Microbiol Ecol* 2018, **94**.
24. Boyd EF, Brussow H: **Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved.** *Trends Microbiol* 2002, **10:**521-529.
25. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS: **PHASTER: a better, faster version of the PHAST phage search tool.** *Nucleic Acids Res* 2016, **44:**W16-21.
26. Steyert SR, Kaper JB: **Contribution of urease to colonization by Shiga toxin-producing Escherichia coli.** *Infect Immun* 2012, **80:**2589-2600.
27. Jarvis KG, Giron JA, Jerse AE, McDaniel TK, Donnenberg MS, Kaper JB: **Enteropathogenic Escherichia coli contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation.** *Proc Natl Acad Sci U S A* 1995, **92:**7996-8000.
28. Wu B, Skarina T, Yee A, Jobin MC, Dileo R, Semesi A, Fares C, Lemak A, Coombes BK, Arrowsmith CH, et al: **NleG Type 3 effectors from enterohaemorrhagic Escherichia coli are U-Box E3 ubiquitin ligases.** *PLoS Pathog* 2010, **6:**e1000960.
29. Pulkkinen WS, Miller SI: **A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein.** *J Bacteriol* 1991, **173:**86-93.
30. Cirillo DM, Heffernan EJ, Wu L, Harwood J, Fierer J, Guiney DG: **Identification of a domain in Rck, a product of the Salmonella typhimurium virulence plasmid, required for both serum resistance and cell invasion.** *Infect Immun* 1996, **64:**2019-2023.

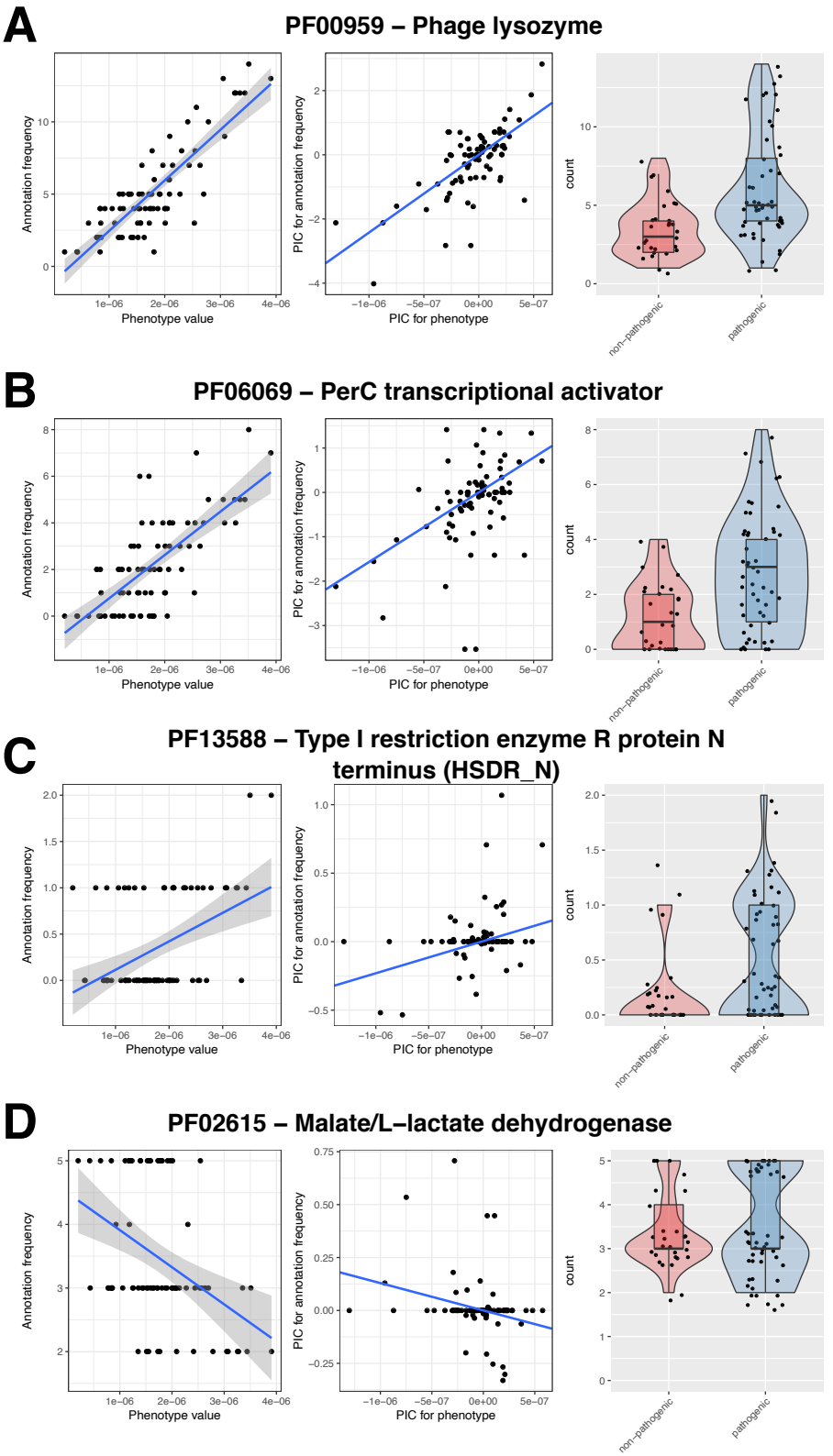
31. Ehrbar K, Hardt WD: **Bacteriophage-encoded type III effectors in Salmonella enterica subspecies 1 serovar Typhimurium.** *Infect Genet Evol* 2005, **5:**1-9.
32. Ramisetty BCM, Sudhakari PA: **Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation.** *Front Genet* 2019, **10:**65.
33. Vandecraen J, Chandler M, Aertsen A, Van Houdt R: **The impact of insertion sequences on bacterial genome plasticity and adaptability.** *Crit Rev Microbiol* 2017, **43:**709-730.
34. Nishiyama T, Noguchi H, Yoshida H, Park SY, Tame JR: **The structure of the deacetylase domain of Escherichia coli PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family.** *Acta Crystallogr D Biol Crystallogr* 2013, **69:**44-51.
35. Fernandez L, Rodriguez A, Garcia P: **Phage or foe: an insight into the impact of viral predation on microbial communities.** *ISME J* 2018, **12:**1171-1179.
36. Simmons EL, Bond MC, Koskella B, Drescher K, Bucci V, Nadell CD: **Biofilm Structure Promotes Coexistence of Phage-Resistant and Phage-Susceptible Bacteria.** *mSystems* 2020, **5**.
37. Cotter PD, Ross RP, Hill C: **Bacteriocins - a viable alternative to antibiotics?** *Nat Rev Microbiol* 2013, **11:**95-105.
38. Maslov S, Sneppen K: **Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems.** *Sci Rep* 2017, **7:**39642.
39. Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS, et al: **Bacteriophage adhering to mucus provide a non-host-derived immunity.** *Proc Natl Acad Sci U S A* 2013, **110:**10771-10776.
40. Silveira CB, Rohwer FL: **Piggyback-the-Winner in host-associated microbial communities.** *NPJ Biofilms Microbiomes* 2016, **2:**16010.
41. Bertozzi Silva J, Storms Z, Sauvageau D: **Host receptors for bacteriophage adsorption.** *FEMS Microbiol Lett* 2016, **363**.
42. Leon M, Bastias R: **Virulence reduction in bacteriophage resistant bacteria.** *Front Microbiol* 2015, **6:**343.
43. Gordillo Altamirano F, Forsyth JH, Patwa R, Kostoulias X, Trim M, Subedi D, Archer SK, Morris FC, Oliveira C, Kielty L, et al: **Bacteriophage-resistant Acinetobacter baumannii are resensitized to antimicrobials.** *Nat Microbiol* 2021, **6:**157-161.
44. Park BS, Lee JO: **Recognition of lipopolysaccharide pattern by TLR4 complexes.** *Exp Mol Med* 2013, **45:**e66.
45. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, Wood TK: **Cryptic prophages help bacteria cope with adverse environments.** *Nat Commun* 2010, **1:**147.
46. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.

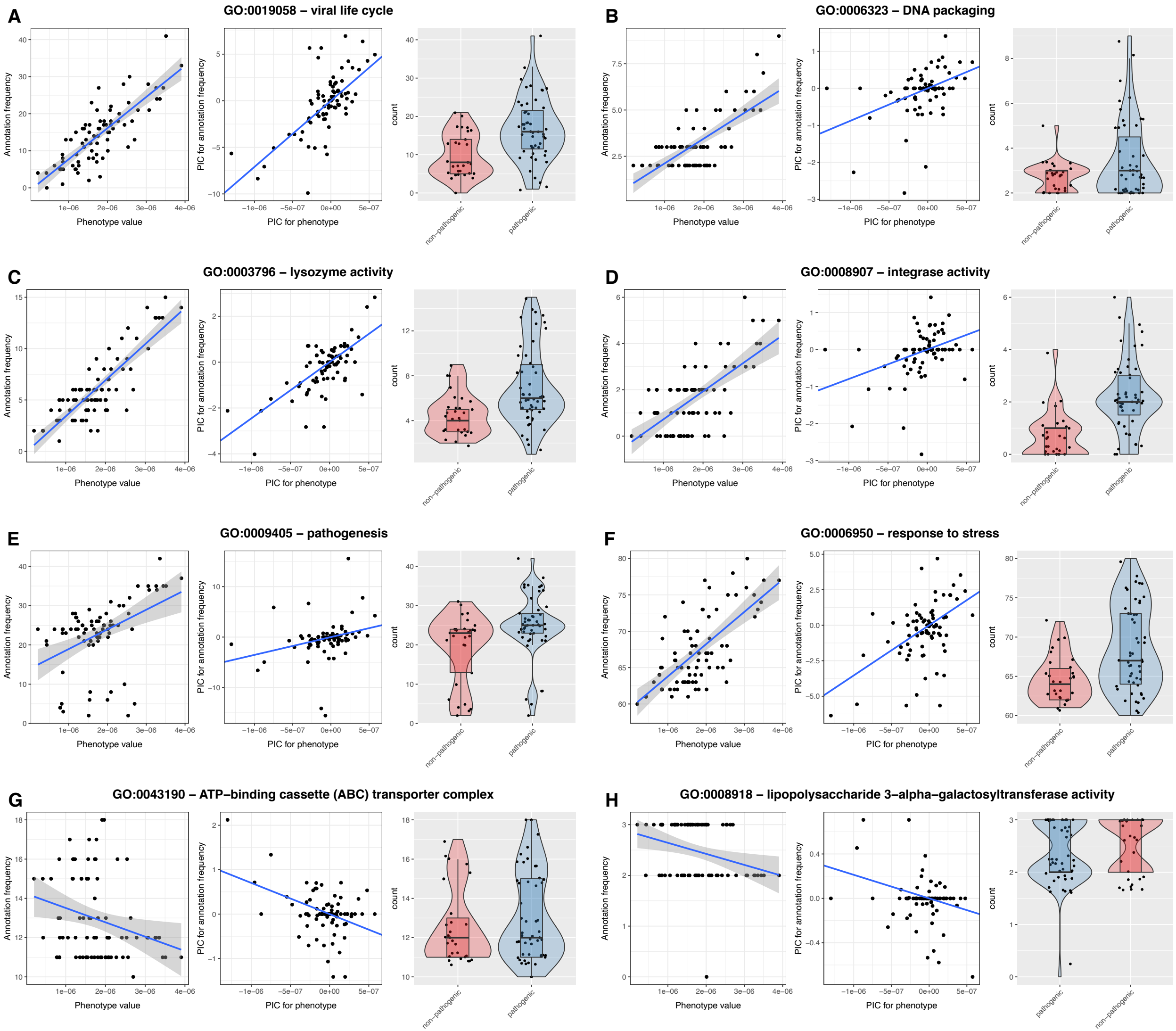
47. Paradis E, Schliep K: **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.** *Bioinformatics* 2019, **35:**526-528.
48. **R: A Language and Environment for Statistical Computing** [http://www.R-project.org]
49. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ, Jr.: **Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups.** *Curr Protoc Bioinformatics* 2011, **Chapter 6:**Unit 6 12 11-19.
50. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND.** *Nat Methods* 2015, **12:**59-60.
51. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
52. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.** *Mol Biol Evol* 2018, **35:**1547-1549.
53. Vaidya G. LD, Meier R.: **SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information.** *Cladistics* 2011, **27:**9.
54. Pang S, Stones RJ, Ren MM, Liu XG, Wang G, Xia HJ, Wu HY, Liu Y, Xie Q: **GPU MrBayes V3.1: MrBayes on Graphics Processing Units for Protein Sequence Data.** *Mol Biol Evol* 2015, **32:**2496-2497.
55. Lee H, Popodi E, Tang H, Foster PL: **Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing.** *Proc Natl Acad Sci U S A* 2012, **109:**E2774-2783.
56. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA: **Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.** *Syst Biol* 2018, **67:**901-904.
57. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al: **The Pfam protein families database in 2019.** *Nucleic Acids Res* 2019, **47:**D427-D432.


Genomic data/metadata
B Genomic & annotation data
Genome annotation to CALANGO format (in-house software)
User-provided CALANGO-compatible data (tabular files with genomic components and their annotation terms)
Quality control based on expected gene content
Genome annotation
Annotation dictionary
Associated metadata
Phylogeny
Normalization
Cutoffs for graphical ouputs
A
C Phylogenetic data
GEN_1
GEN_2
GEN_3
GEN_4
GEN_N
Nexus
Newick
CALANGO
Phylogeny
Metadata
phenotype
norm
18
3141
8
4721
3.3
6474
(...)
(...)
6
6660
Genome Annotation
GEN_1
Genfeat_ID
Ann_ID
00001
Ann_001
00002
Ann_001
00003
Ann_996
Building of annotation vectors
E
Annotation vectors
Gen_ID
(...)
Ann_N
GEN_1
4
GEN_2
10
GEN_3
1
33
D
F
Normalization
Normalized annotation vectors
1.27e-3
2.11e-e
1.54e-4
4.96e-3
G
Statistical model building
for each annotation term, compute:
linear models for phylogenetically independent contrasts (PICs) of quantitative phenotypes and the normalized values of an annotation term, and user-provided phylogeny
correlation values between phenotypes and normalized values for annotation terms (Pearson, Spearman and Kendall)
corrected q-values for the PIC linear models and correlation tests
common statistics (sum, standard deviation, correlation coefficient)
heterogeneity and prevalence
Production of output files
H
Dictionary
Description
Ann_001
ATPase
Ann_002
Kinase
Globin
HTML5 dynamical output & flat files
I heatmaps of annotation terms
Species groups
ID: Ann_001
Species: GEN_1
description:ATPase
Value: 0.606
J Scatterplots of annotation term statistics
-log10(q-value corr_spearman)
-log10(q-value contrasts)
ID: Ann_001
-log10(q-value): 5.8
size: 2341
description:ATPase
K dynamic table
show/hide columns
search:
q_value
correlation
Ann_001
1,02e-9
0.87
Ann_002
3,14e-5
0.68
2,72e-2
0.13
filter:
L distribution plots for annotation terms
annotation term frequency
phenotype
Species: GEN_1
Y_value: 6.36e-4
X_value: 18
Group 1
Group 2
Group 3


A
Domains within proteins as genomic components
D_01
D_02
D_03
D_04
1
domain2Pfam
2
domain2GO
Genomic_element_ID
Annotation_ID
D_01 PF00047
D_02 PF04562
D_03 PF04562
D_04 PF00942
Genomic_element_ID
Annotation_ID
D_01 GO:0007049
D_02 GO:0035480
D_03 GO:0035480
D_04 GO:0007049
B
Non-pathogenic
Pathogenic
Wilcoxon p−value = 4.60e−05
phage count
Wilcoxon p−value = 1.36e−13
genome length
Wilcoxon p−value = 1.17e−04
phage density
C
p < 2.22e−16
p < 2.22e−16
0.057
p < 2.22e−16
Prevalence
p < 2.22e−16
p < 2.22e−16
p < 2.22e−16
0.00036
Sum
domain2GO all
domain2GO sig
domain2Pfam all
domain2Pfam sig


Pathogenic
Non-pathogenic
negative correlation
Porin
Glycosyl hydrolase
Glycosyl transferase
Toxin-antitoxin system
mixed
phage biology (DNA)
metabolism
mixed
ABC transporters
phage biology (structural)
positive association
phage biology (structural)
restriction-modification
row:PF08988
colum: O127:H6 (strain E2348/69/EPEPC)
value: 1.0
Definition: Type III secretion system
cytoplasmic E component of needle
positive association
Type III secretion system
DUFs
positive association
Tellurite resistance
Haemolysin
Intimin
NFkB-degrading protease
Shiga-like toxin
Urease
NleG
positive association
Phage biology (DNA)
positive association
Transposase
Plasmid-stabilizing protein
positive association
phage biology (structural)
1
0.75
0.50
0.25
0


A
PF00959 – Phage lysozyme
Annotation frequency
Phenotype value
PIC for annotation frequency
PIC for phenotype
count
non−pathogenic
pathogenic
B
PF06069 – PerC transcriptional activator
Annotation frequency
Phenotype value
PIC for annotation frequency
PIC for phenotype
count
non−pathogenic
pathogenic
C
PF13588 – Type I restriction enzyme R protein N terminus (HSDR_N)
Annotation frequency
Phenotype value
PIC for annotation frequency
PIC for phenotype
count
non−pathogenic
pathogenic
D
PF02615 – Malate/L−lactate dehydrogenase
Annotation frequency
Phenotype value
PIC for annotation frequency
PIC for phenotype
count
non−pathogenic
pathogenic


A
GO:0019058 – viral life cycle
B
GO:0006323 – DNA packaging
C
GO:0003796 – lysozyme activity
D
GO:0008907 – integrase activity
E
GO:0009405 – pathogenesis
F
GO:0006950 – response to stress
G
GO:0043190 – ATP–binding cassette (ABC) transporter complex
H
GO:0008918 – lipopolysaccharide 3–alpha–galactosyltransferase activity
Annotation frequency
Phenotype value
PIC for annotation frequency
PIC for phenotype
count
non–pathogenic
pathogenic