Published in final edited form as:
Biol Bull. 2012 August ; 223(1): 21–29.

Making the Most of Omics for Symbiosis Research

J. Chaston and **A.E. Douglas**
Department of Entomology, Comstock Hall, Cornell University, Ithaca, NY14853, USA

Abstract

Omics, including genomics, proteomics and metabolomics, enable us to explain symbioses in terms of the underlying molecules and their interactions. The central task is to transform molecular catalogs of genes, metabolites etc. into a dynamic understanding of symbiosis function. We review four exemplars of omics studies that achieve this goal, through defined biological questions relating to metabolic integration and regulation of animal-microbial symbioses, the genetic autonomy of bacterial symbionts, and symbiotic protection of animal hosts from pathogens. As omic datasets become increasingly complex, computationally-sophisticated downstream analyses are essential to reveal interactions not evident to visual inspection of the data. We discuss two approaches, phylogenomics and transcriptional clustering, that can divide the primary output of omics studies – long lists of factors – into manageable subsets, and we describe how they have been applied to analyze large datasets and generate testable hypotheses.

Keywords

Gaussian graphical models; transcriptomics; metabolomics; phylogenomics; proteomics; symbiosis; regulatory network prediction; symbiosis

INTRODUCTION

In symbioses, the irreducible complexity of an organism is compounded by persistent symbiotic interactions with one, two or many phylogenetically-different organisms, each of which is adapted to function in the context of its partner(s). Until recently, the molecular basis of symbiosis could only be studied on the basis of one or a few genes and their products at a time. For example, in their research on interactions between a *Legionella*-like bacterium and its host *Amoeba proteus*, Jeon and colleagues were able to correlate the reduced expression of a single host gene product, S-adenosyl methionine synthetase (SAMS), with the rapid evolution from a pathogenic to a mutualistic relationship (Jeon and Jeon, 2003). It is very likely that this dramatic evolutionary transition involved multiple coevolved changes in the metabolic and regulatory networks of the two organisms, but a systematic analysis of these putative changes accompanying the change in host SAMS expression was technically unrealistic at that time. Today, just nine years later, the association in *A. proteus* and other fascinating symbioses can be interrogated by a range of high throughput methods that reveal the total (or near-total) complement of a particular class of biological molecules: genes, transcripts, proteins, lipids, metabolites etc.

The omics revolution of the last decade has transformed our capacity to understand symbioses at the molecular level. It is now possible, for example, to construct an inventory of the genes coded by each partner, to quantify patterns of transcription under different

Author for correspondence: aes326@cornell.edu.
Co-author: chaston@cornell.edu

environmental conditions, to establish the relationship between transcript and protein abundance for every protein-coding gene, and to determine the metabolite set that make up the metabolic pool of the interacting symbiotic partners. Symbiosis researchers have adopted these techniques with great alacrity, providing many novel insights. In the first part of this review, we have selected four studies as exemplars of how to "make the most" of omics approaches for symbiosis research. The common theme of these studies is that the omics approaches with the greatest impact are driven by important biological questions.

Nevertheless, omics biology brings challenges, as well as opportunities. The minimal output of omics is lists of genes, proteins, metabolites etc. that is a partial or near-complete molecular catalog of an organism or symbiosis. To use omics methods to answer biological questions requires great care in experimental design and interpretation. In the second part of this review, we discuss informatics routes that can help the researcher to "make the most" of omics data for symbiosis research, especially in relation to genomic and transcriptomic approaches.

EXEMPLARS THAT "MAKE THE MOST" OF OMICS APPROACHES IN SYMBIOSIS RESEARCH

In this section, we describe four sets of experiments, each conducted on a single type of animal-microbial symbiosis. In their very different ways, using each of genomic, transcriptomic, proteomic and metabolomic data, these studies illustrate how omics approaches can be applied to answer specific questions in symbiosis research.

Inferring Metabolic Interactions from Bacterial Genome Sequences

The complete genome sequence of an organism defines its biological capabilities, but our capacity to interpret genome sequence data depends on the quality of gene annotation. The large proportion of conserved hypothetical genes (i.e. genes predicted *in silico* but without evidence of expression *in vivo*) and lineage-specific genes of unknown function in most sequenced genomes demonstrate the limitations to our capacity to interpret genomic data. Arguably, the simplest genomes to interpret are the smallest; and these are found among the bacterial symbionts of insects. The insight into symbioses that can be obtained from genomics is illustrated by recent studies on the genome sequences of bacterial endosymbionts of insects (McCutcheon and Moran, 2012).

Symbioses involving bacteria with very small genomes (<1 Mb) have evolved independently in multiple insect groups with the common trait of feeding through the lifecycle on nutritionally poor or unbalanced diets. Examples include blood-feeding insects (e.g. bedbugs, lice, tsetse flies), plant-sap feeding insects (whitefly, aphids, cicadas etc.), and scavengers, such as the cockroaches. The inference that these symbioses have a nutritional basis (Buchner, 1965) has been confirmed amply by modern nutritional and physiological studies indicating that these insects derive specific nutrients, including essential amino acids and vitamins, from their microbial symbionts (Douglas, 2009). The bacteria are intracellular, restricted to one cell type, generically known as the bacteriocyte, which apparently functions exclusively to house and maintain the symbiosis. The bacteria are vertically transmitted, usually directly from the bacteriocyte to the cytoplasm of the eggs in the female ovary (Buchner, 1965). The small genome size of the bacterial symbionts is interpreted as the evolutionary consequence of obligate vertical transmission. Gene loss can be attributed to relaxed selection on genes not required in the symbiosis and to genomic decay, caused by the small effective population size of the vertically-transmitted bacteria and resultant accumulation of mildly deleterious mutations (Moran, 1996); and these genomes appear not to be susceptible to horizontally acquired genes.

Our exemplar of genomics in symbiosis research is a set of papers by McCutcheon and Moran describing the genomes of the symbiotic bacteria in three related groups of xylem-feeding insects: the Cercopidae (spittlebugs), Cicadoidae (cicadas) and Cicadellinae (sharpshooters) (McCutcheon *et al.*, 2009; McCutcheon and Moran, 2007, 2010). All three insect groups bear two morphologically-distinct bacterial symbionts: a common primary symbiont (*Sulcia mulleri*) and a distinct auxiliary symbiont (Fig. 1). Plant xylem sap lacks the 10 essential amino acids that animals cannot synthesize but require for proteinsynthesis. Genomic inspection of the primary and auxiliary symbionts revealed that *Sulcia* has the genetic capacity to synthesize either 7 (in spittlebugs) or 8 (in cicadas and sharpshooters) essential amino acids, and that the auxiliary symbionts encode the biosynthetic pathways for the remaining essential amino acids; thus, the various auxiliary symbionts and the cohabiting *Sulcia* have perfect complementarity in their genetic capacity for essential amino acid synthesis. These studies demonstrate how the genome of each bacterial symbiont is shaped by coevolutionary interactions with symbiotic partners. Furthermore, they generate very specific predictions about the three-way transfer of multiple metabolites (including essential amino acids) among the host and symbionts.

As these studies illustrate, metabolic interactions in symbioses can be inferred by visual inspection of genomic data. Nevertheless, the metabolic networks are inherently complex, even in bacteria with reduced genomes, and metabolic modeling based on the inventory of metabolism genes offers a valuable route to identify and quantify the nutritional resources utilized by the symbiotic bacteria and their metabolic adaptations for the release of specific nutrients to the host. These methods have been applied with success, for example, to the endosymbionts of aphids (MacDonald *et al.*, 2011; Thomas *et al.*, 2009), sharpshooters (Cottret *et al.*, 2010) and cockroaches (Gonzalez-Domenech *et al.*, 2012). This approach is ideally suited to bacteria with much reduced genomes, in which all metabolism-related genes are expressed. For many bacteria, however, the metabolic phenotype under any one set of conditions is underpinned by a subset of the metabolism-related genes. For these bacteria, it is essential to complement the genomic information with gene expression data.

Transcriptomics and the Regulation of Symbiotic Bacteria

Our exemplar of a transcriptomics study is the analysis by Wier *et al.*(2010) of gene expression patterns in the symbiosis between the bobtail squid *Euprymna scolopes* and the luminescent bacterium, *Vibrio fischeri*. This study concerns a central issue in symbiosis research: the regulation of microbial symbionts, specifically the mechanisms by which the numbers and location of symbionts in an animal host are controlled. Valuable insights have come from studying how symbioses respond to environmental perturbations, such as changes in temperature or nutrient supply. *E. scolopes* juvenile squid are born devoid of *Vibrio* bacteria but acquire these by passing water over the entrance to a specialized ventral squid organ called the light organ. Through a series of selective events (Nyholm, 2004), *V. fischeri* alone gain access to the organ, where they grow to high densities [$> 10^8$ CFU/ml; (Boettcher and Ruby, 1990)] at night and generate light, providing camouflage by counterillumination for their host (Jones 2004). At dawn, >90% of the bacterial cells are expelled from the light organ, and the population of residual bacterial population subsequently proliferates to regenerate the dense, luminescent population by nightfall. Thus, the squid symbiosis has a particular experimental advantage that the symbiosis is perturbed naturally on a daily basis, yielding regulatory changes in the symbiosis that are highly reproducible across individuals and over time.

Fluctuations in the light organ bacterial population are associated with dramatic reorganization of host tissues and gene expression. The greatest changes in gene expression occurred at dawn, the time of symbiont expulsion. The expression of > 50 host genes with annotated cytoskeletal function and various symbiont genes mediating anaerobic respiration

of glycerol was elevated uniquely at this time. Simultaneously, the microvilli of the apical membrane were lost, with associated membrane blebbing. The inference that the bacteria utilized host lipids derived from the membrane reorganization was supported by the close similarity in fatty acid composition of host tissues and symbiont cells. Through the subsequent day, the symbiosis "re-assembled". The apical microvilli on the host epithelial cells generated anew and the residual symbionts initiated cell growth and division, processes orchestrated by changes in the expression of the host cytoskeletal genes, and a shift in expression of the symbiont metabolism genes from glycerol fermentation to the utilization of chitin as the preferred respiratory substrate, respectively.

Experimental design was key to the success of this analysis. Importantly, the transciptomes of host and symbiont were analyzed in parallel, enabling the interactions between the gene expression patterns of the partners to be analyzed. Furthermore, the host samples exclusively comprised the core of the light organ, including the epithelial cells that interact directly with the symbiotic bacteria, minimizing the incidence of host transcriptional responses that vary over the diel cycle for reasons unrelated to the symbiosis. Further, because the spatial and temporal interactions between host and symbiont were already well characterized it was possible to link gene expression patterns to specific symbiotic phenomena, such as bacterial expulsion, diel variation in bacterial resource acquisition patterns, and ultrastructural host cell membrane reorganization.

Proteomes and Symbiont Autonomy

Proteomics, the global analysis of proteins, is technically more demanding and costly than transcriptomics; it is less sensitive than transcriptomics; and it requires a protein sequence database derived from the genome or extensive cDNA libraries, preferably of the same species. For these multiple reasons, transcriptomics is widely adopted as the method of choice to study the expression of protein-coding genes, on the assumption that transcript and protein abundance are correlated. In broad terms, this assumption is justified. Multiple studies have demonstrated a moderate, positive correlation between transcript and protein abundance. Even so, the slope of the relationship is significantly less than unity, suggesting that the proteins contributing to a cell, tissue or organism vary less than transcripts in abundance (Bonaldi *et al.*, 2008; Sun *et al.*, 2010). An additional complication for time-course studies is that the temporal pattern of transcript and protein abundance can differ for a single gene, and that the pattern of this difference can vary widely among genes (Wang *et al.*, 2010a). Where the focus of interest is protein-coding genes and post-transcriptional regulation is known (or anticipated) to be important, proteomics is the method of choice. Proteomics is also essential for large-scale analysis of the spatial distribution of proteins within cells and among organisms in symbioses.

Our exemplar of a proteomics study concerns the transfer of proteins between host and symbiont, and relates, specifically, to the question whether intracellular symbionts with very small genomes are analogous to bacterial-derived organelles. For example, animal proteins of nucleocytoplasmic origin are targeted to mitochondria, subsidizing the limited functional capabilities coded by the mitochondrial genome. The focus of the study is a vertically-transmitted symbiont housed in bacteriocytes of an insect (as described above in relation to sharpshooters and their allies): specifically, the symbiotic bacterium *Buchnera aphidicola* in the pea aphid *Acyrthosiphon pisum*. The *Buchnera* genomeis just 0.64 Mb, less than 20% of the genome of the related bacterium *E. coli*, and it lacks genes for metabolic functions that are also coded by the aphid genome (IAGC, 2010). To investigate whether host proteins, including metabolic enzymes, are transported to the *Buchnera* cells, (Poliakov *et al.*, 2011) conducted a quantitative analysis of the proteome of multiple aphid samples in which the *Buchnera* cells were progressively enriched: from the whole insect body, through isolated host cells, and partially-purified *Buchnera* cells to *Buchnera* cells purified on a Percoll

gradient. Those proteins that co-purify with the *Buchnera* cells through the enrichment series are predicted to be associated with the *Buchnera* cells (Fig. 2). Overall, >1,900 aphid proteins and 400 *Buchnera* proteins were detected. Cluster analysis revealed that proteins coded by the *Buchnera* genome were selectively enriched in *Buchnera* cells, with no evidence for selective transfer of any proteins in either direction between host and symbiont. This study indicates that metabolic integration between the partners is mediated by the transfer of small metabolites, and not proteins, and generates the specific hypothesis that certain metabolic pathways are shared between the host and symbiont, with the transfer of intermediate metabolites between the partners.

The genome sequence of the pea aphid host is congruent with these results. As for all other eukaryotes, the genome includes various genes that can be assigned to the bacterial ancestor of the mitochondrion and that code for proteins which are targeted to the mitochondrion. By contrast, the only genetic material of likely *Buchnera* origin is two highly truncated pseudogenes (ψDnaE and ψAtpH) (Nikoh *et al.*, 2010). It appears that genome reduction in *Buchnera* has not involved the net transfer of intact genes to the host nucleus.

The conclusion that *Buchnera* lacks genetic subsidy by the host, a cardinal feature of a bacterial-derived organelle, raises the question whether other insect endosymbionts with genomes even smaller than *Buchnera* and comparable to mitochondria and plastids are also genetically autonomous. Quantitative proteomics, as conducted for *Buchnera*, can resolve this issue. We should not presume that all symbioses involving bacteria with small genomes have solved the functional problems posed by genomic erosion in the same way.

Metabolomics and Symbiotic Protection against Pathogens

The metabolome, i.e. the global set of metabolites in a biological system, differs fundamentally from the genome, transcriptome and proteome in that it cannot be deduced directly from the genome sequence. Metabolomics poses a unique set of challenges. In particular, different techniques are required to analyze different classes of metabolites; and many of the metabolites detected by mass spectrometry, NMR spectroscopy and related methods cannot realistically be identified. For some purposes, important information can be gleaned from analysis of the metabolite differences between samples (e.g. hosts bearing and lacking symbionts) without substantial investment in identification of the compounds. This type of metabolite analysis is often known as metabolite fingerprinting or metabonomics. For other experimental designs, where identification is crucial, access to high quality spectral library databases is essential(Tohge and Fernie, 2009).

Of particular interest for symbiosis research is the study of Fukuda *et al.*(2011), which used metabolomics to pinpoint a single metabolite that conferred resistance against a pathogenic bacterium. Their system comprised mice, symbiotic bacteria of the genus *Bifidobacterium* that colonize the mouse colon, and the pathogen *E. coli* strain O157, which produces the proteinaceous Shiga toxin. Fukuda *et al.* (2011) demonstrated that when mice bearing either the gut bacterium *Bifidobacterium longum* or *B. adolescentis* were infected with the pathogenic *E. coli* O157, the O157 cells proliferated, but only the mice with *B. adolescentis* died. They hypothesized that metabolites released by *B. longum* were important in mediating protection against O157. Their ^{1}H-^{13}C NMR metabolomic study revealed striking differences in the sugar profiles of feces produced by mice bearing the two *Bifidobacterium* species. This metabolomic analysis set Fukuda *et al.* onto the scientific “trail” to identify the active compound. Recognizing that Bifidobacteria ferment sugars to short chain fatty acids, Fukuda *et al.*(2011) then demonstrated that the concentration of one SCFA, acetic acid, was significantly elevated in the feces of mice bearing *B. longum*; and that acetic acid enhanced the barrier function of the colon epithelial cells, such that the translocation of O157 cells and the Shiga toxin across the epithelium was inhibited. The elevated production of acetic acid

by *B. longum* could be linked to its expression of genes for fructose-transporters and high rates of fructose uptake *in vitro*.

The study of Fukuda *et al.*(2011) provides an important lesson in omics: that omics methods are a discovery tool that can be used to construct testable hypotheses. Although omics lend themselves to cataloguing the molecular composition of living organisms, their power is most evident when harnessed to answering defined biological questions. This vital point is the basis for the following section of this review: a consideration of approaches that have been used to gain useful information from the large dataset outputs of omics studies.

MAKING THE MOST OF OMICS APPROACHES

As considered in the Introduction, the crude output of an omics study is a catalog of the genes, proteins or small metabolites that constitute the biological sample studied. Visual inspection of the data is very important for understanding the results, but the datasets are often so large and complex that supplementary computational methods are essential for full interpretation of the data. For example, these approaches can protect against inadvertent "cherry-picking" of data that conform to preconceived expectations while ignoring potentially important genes or metabolites with no known function or functions apparently unrelated to the experimental treatment. The burgeoning field of systems biology offers a great diversity of strategies and tools to analyze omics datasets. We will discuss two approaches – phylogenomics and regulatory gene network discovery – and focus on their use in inferring gene function by identifying genes of unknown function that have similar distribution patterns or patterns of expression as genes of known function.

Phylogenomics allocates genes according to their evolutionary history with the rationale that genes with a similar evolutionary history will cluster according to function (Srinivasan *et al.*, 2005). This technique is particularly valuable to generate candidate functions for genes lacking functional annotation. Gene evolutionary history is predicted by creating a coinheritance matrix of all the proteins in the genome sequence of interest (rows) against a library of genomes (for example, all of the genome sequences available on NCBI; one column per genome). From this matrix, a similarity matrix is created that clusters proteins according to the similarity of their coinheritance pattern. The similarity matrix is then transformed into a 2-D plot with each point representing a protein in the input genome sequence. Clusters of points tend to represent shared functions (e.g. as measured by mapping with gene ontology categories), and the role of genes with unknown function can be predicted from their co-clustering with genes of known function. Additional information can be obtained where particular functions, as identified by gene ontology, are over-represented in a gene cluster, although functional inference can be complicated for large clusters (comprising hundreds of genes) and clusters in which multiple gene ontology categories are represented.

Phylogenomics is particularly well-suited for functional inference of genes in bacteria because many sequenced bacterial genomes are available to support the analysis. Phylogenomics has been applied recently in combination with filtering approaches to identify putative symbiosis-related genes of *Xenorhabdus nematophila* and *X. bovienii*, bacterial symbionts of entomopathogenic nematodes (Chaston *et al.*, 2011).*Steinernema* nematode species carry specific *Xenorhabdus* species at the anterior of the nematode intestine in a specialized structure called the receptacle(Poinar, 1966; Wouts, 1980; Bird and Akhurst, 1983; Flores-Lara *et al.*, 2007). The nematodes actively seek out and penetrate soil-dwelling insect hosts. Once inside the insect, the symbiotic bacteria are released and kill the insect host via immune system suppression and production of potent effectors, and the bacteria provide nutrition to the nematode, which reproduces through several generations

inside the insect(Kaya, 1993; Forst*et al.*, 1997; Goodrich-Blair and Clarke, 2007). When nutrients are spent, the nematodes acquire a complement of colonizing bacteria and leave the nutrient depleted cadaver in search of a new insect host (reviewed in (Richards *et al.*, 2009).The *X. nematophila* and *X. bovienii* genomes were studied to identify candidate genes that contribute to the maintenance of this symbiosis. *X. nematophila* genes that were conserved in other Enterobacteriaceae and specific to each of the two *Xenorhabdus* species were discarded, and the remaining genes were divided into two groups based on their conservation in other bacterial symbionts of entomopathogenic nematodes (Fig. 3). Each of the two gene groups was analyzed by phylogenomics separately, resulting in assignment of 533 genes to 24 clusters(12% of the *X. nematophila* genome). To focus on clusters containing genes with predicted symbiotic functions, the genes from each cluster were assessed for enrichment in proteins found in host-associated microbes (Fig. 3E). Inferred symbiosis clusters included genes that function in toxin production and secretion, and resistance to heavy metal stress. This analysis offered a rich, but manageable, catalog of 221 genes (5% of the total gene complement) with candidate symbiotic function for empirical analysis that could not have been achieved by visual inspection of the *Xenorhabdus* genome sequences alone.

A valuable approach for interpreting transcriptome data is provided by transcriptional networks. Specifically, computational models can be applied to generate gene networks that identify functional groups of genes responding to the same regulatory factors. Gene network construction is a particularly useful tool for predicting gene function because genes of unknown and known functions allocated to the same regulatory cluster are inferred to function in similar regulatory hierarchies and respond to similar stresses and stimuli. One approach for gene network creation is to identify direct interactions between genes using Gaussian graphical models (GGMs) (Schäfer, 2005; Dobra, 2004).GGMs use partial correlations of transcriptional data as a metric to identify direct interactions of two genes. (Standard, i.e. Pearson, correlations are not suitable because they do not discriminate correlations resulting from other factors, e.g. indirect gene-gene interactions, regulation by a common gene). One area of current focus is to increase the predictive power of the causality of identified interactions, i.e. discriminate causal from reactive interactions (Schadt*et al.*, 2005), by integrating transcriptomic data with other omics approaches that identify expression quantitative trait loci (eQTLs: loci that cause changes in gene transcription across individuals)or protein-protein interactions (Zhu *et al.*, 2008). However, networks constructed only from transcriptomic data produce similar connectivity profiles as networks created by integrating multiple omics methods; they are just reduced in their ability to infer causality {e.g. (Zhu *et al.*, 2008). Thus, gene network creation can still be applied to microarray and RNAseq datasets, even when researchers may not have access to extensive strain libraries for eQTL mapping, or to other relevant omics information such as protein-protein interaction data.

To date, the symbiosis community appears to have made little use of transcriptional network analysis, but we anticipate that this will change rapidly in the next few years. One recent study that does apply gene networks to the study of symbiosis investigated the transcriptional responses of humans to probiotic *Lactobacillus*(van Baarlen *et al.*, 2011). The same human subjects were separately exposed to each of three probiotic *Lactobacillus* species. The microarray data were highly variable across individuals: transcriptional responses were more similar for different treatments on the same individual than for the same treatments on different individuals. Nevertheless, network analysis of the transcriptomes identified a number of networks (e.g. blood pressure regulation, wound healing) which responded to the different *Lactobacillus* species, and certain networks (e.g. ion homeostasis) that responded to more than one of the probiotic bacteria. Although the actual transcriptional levels varied among individuals across the study, the network

responses appeared to be conserved, allowing interrogation of what otherwise appeared to be a near-uninterpretable dataset.

Although the regulatory networks described in van Baarlen *et al.*(2011) were not created from transcriptomic data (instead, networks were created by combing the literature for experimental data), they demonstrate the power of gene network discovery for analysis of large transcriptomic datasets in symbiotic systems. Network creation from transcriptomic data has recently been used to study mammalian gastrointestinal symbiotic systems (Shulzehnko*et al.*, 2011; Greenblum*et al.*, 2012), and has also been used successfully, sometimes with experimental verification, with transcriptomic data in a variety of organisms (Ayroles *et al.*, 2009; Chang *et al.*, 2010; Guan *et al.*, 2008; Lee *et al.*, 2008; Logsdon and Mezey, 2010).

OUTLOOK

Omics approaches are taking our understanding of symbioses to a new level of molecular sophistication. Indeed, the monumental efforts to define and characterize human-associated microbial communities by initiatives such as the Human Microbiome Project are only attainable by implementation of omics methods. Some types of interactions, such as regulation of the nutritional status and cell proliferation patterns of the host (Buchon *et al.*, 2009; Shin *et al.*, 2011), make intuitive sense, but interpreting other results of omics experiments will require further research on the function of certain gene classes. For example, one-third of the proteins that differ in abundance between pea aphids bearing and experimentally-deprived of their *Buchnera* bacteria are cuticular proteins (Wang *et al.*, 2010b), a result that cannot be related to any currently known aphid-*Buchnera* interaction. As this result illustrates, omics experiments have great potential to spur efforts to understand molecular function in symbiosis.

From a symbiotic perspective, gene classes of particular potential interest are those that respond specifically to perturbation of the symbiosis and have no known function. Conserved genes of this class may represent the deep history that defines the predisposition of animals for symbioses with microorganisms, while recently-evolved, lineage-specific genes that underpin the unique functions of individual associations. Elucidation of these patterns will contribute not only to our understanding of symbiosis, but also to the resolution of central problems posed by conserved and lineage-specific genes of no known function.

Acknowledgments

We thank Tomas Lazo for the drawing of the aphid in Figure 2. This work was supported by NIH grant 1R01GM095372-01, NSF grant IOS-0919765, and the Sarkaria Institute for Insect Physiology and Toxicology.

LITERATURE CITED

Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RR, et al. Systems genetics of complex traits in *Drosophilamelanogaster* . Nat. Genet. 2009; 41:299–307. [PubMed: 19234471]

Bird AF, Akhurst RJ. The nature of the intestinal vesicle in nematodes of the family Steinernematidae. Int. J. Parasitol. 1983; 13:599–606.

Boettcher KJ, Ruby EG. Depressed light emission by symbiotic Vibrio fischeri of the sepiolid squid Euprymna scolopes. J. Bacteriol. 1990; 172:3701–3706. [PubMed: 2163384]

Bonaldi T, Straub T, Cox J, Kumar C, Becker PB, Mann M. Combined use of RNAi and quantitative proteomics to study gene function in *Drosophila* . Mol. Cell. 2008; 31:762–772. [PubMed: 18775334]

Buchner, P. Endosymbioses of Animals with Plant Microorganisms. Chichester, UK: John Wiley and Sons; 1965.

Buchon N, Broderick NA, Chakrabarti S, Lemaitre B. Invasive and indigenous microbiota impact intestinal stem cell activity through multiple pathways in *Drosophila* . Genes Dev. 2009; 23:2333–2344. [PubMed: 19797770]

Chang X, Liu S, Yu YT, Li YX, Li YY. Identifying modules of coexpressed transcript units and their organization of Saccharopolyspora erythraea from time series gene expression profiles. PLoS One. 2010; 5:e12126. [PubMed: 20711345]

Chaston JM, Suen G, Tucker SL, Andersen AW, Bhasin A, Bode E, Bode HB, Brachmann AO, Cowles CE, Cowles KN, et al. The entomopathogenic bacterial endosymbionts xenorhabdus and photorhabdus: convergent lifestyles from divergent genomes. PLoS One. 2011; 6:e27909. [PubMed: 22125637]

Cottret L, Milreu PV, Acuna V, Marchetti-Spaccamela A, Stougie L, Charles H, Sagot MF. Graph-based analysis of the metabolic exchanges between two co-resident intracellular symbionts, *Baumanniacicadellinicola and Sulciamuelleri* with their insect host, *Homalodiscacoagulata* . PLoS Comput. Biol. 2010; 6:e1000904. [PubMed: 20838465]

Dobra A. Sparse graphical models for exploring gene expression data. J.Multivariate Anal. 2004; 90:196–212.

Douglas AE. The microbial dimension in insect nutritional ecology. Funct. Ecol. 2009; 23:38–47.

Flores-Lara Y, Renneckar D, Forst S, Goodrich-Blair H, Stock P. Influence of nematode age and culture conditions on morphological and physiological parameters in the bacterial vesicle of *Steinernema carpocapsae* (Nematoda: Steinernematidae). J. Invertebr. Pathol. 2007; 95:110–118. [PubMed: 17376477]

Forst S, Dowds B, Boemare N, Stackebrandt E. *Xenorhabdus* and *Photrhabdus* spp.: bugs that kill bugs. Annu. Rev. Microbiol. 1997; 51:47–72. [PubMed: 9343343]

Fukuda S, Toh H, Hase K, Oshima K, Nakanishi Y, Yoshimura K, Tobe T, Clarke JM, Topping DL, Suzuki T, et al. *Bifidobacteria* can protect from enteropathogenic infection through production of acetate. Nature. 2011; 469:543–547. [PubMed: 21270894]

Gonzalez-Domenech CM, Belda E, Patino-Navarrete R, Moya A, Pereto J, Latorre A. Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two strains, primary endosymbionts of cockroaches. BMC Microbiol. 2012; 12:S5. [PubMed: 22376077]

Goodrich-Blair H, Clarke DJ. Mutualism and pathogenesis in *Xenorhabdus* and *Photorhabdus* two roads to the same destination. Mol. Microbiol. 2007; 64:260–268. [PubMed: 17493120]

Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Natl Acad. Sci. USA. 2012; 109:594–599. [PubMed: 22184244]

Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG. A genomewide functional network for the laboratory mouse. PLoS Comp.Biol. 2008; 4:e1000165.

International Aphid Genomics Consortium. Genome sequence of the pea aphid*Acyrthosiphon pisum* . PLoS Biol. 2010; 8:e1000313. [PubMed: 20186266]

Jeon TJ, Jeon KW. Characterization of *sams* genes of *Amoeba proteus* and the endosymbiotic X-bacteria. J. Eukaryot. Microbiol. 2003; 50:61–69. [PubMed: 12674481]

Jones BW, Nishiguchi MK. Counterillumination in the hawaiian bobtail squid *Euprymna scolopes* Berry (Mollusca : Cephalopoda). Mar. Biol. 2004; 144:1151–1155.

Kaya HK, Gaugler R. Entomopathogenic nematodes. Annu. Rev. Entomol. 1993; 8:181–206.

Lee L, Lehner B, Crombie B, Wong W, Fraser AG, Marcotte EM. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans* . Nat. Genet. 2008; 40:181–188. [PubMed: 18223650]

Logsdon BA, Mezey J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. PLoS Comput. Biol. 2010; 6:e1001014. [PubMed: 21152011]

MacDonald SJ, Thomas GH, Douglas AE. Genetic and metabolic determinants of nutritional phenotype in an insect-bacterial symbiosis. Mol. Ecol. 2011; 20:2073–2084. [PubMed: 21392141]

McCutcheon JP, McDonald BR, Moran NA. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc. Natl Acad. Sci. USA. 2009; 106:15394–15399. [PubMed: 19706397]

McCutcheon JP, Moran NA. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl Acad. Sci. USA. 2007; 104:19392–19397. [PubMed: 18048332]

McCutcheon JP, Moran NA. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol. Evol. 2010; 2:708–718. [PubMed: 20829280]

McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 2012; 10:13–26. [PubMed: 22064560]

Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl Acad. Sci. USA. 1996; 93:2873–2878. [PubMed: 8610134]

Nikoh N, McCutcheon JP, Kudo T, Miyagishima SY, Moran NA, Nakabachi A. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. 2010; 6:e1000827. [PubMed: 20195500]

Nyholm SV, McFall-Ngai MJ. The winnowing: establishing the squid-vibrio symbiosis. Nat. Rev. Genet. 2004; 2:632–642.

Poinar GO. The presence of *Achromobacter nematophilus* in the infective stage of a *Neoaplectana* sp. (Steinernematidae: Nematoda). Nematologica. 1966; 12:105–108.

Poliakov A, Russell CW, Ponnala L, Hoops HJ, Sun Q, Douglas AE, van Wijk KJ. Large-scale label-free quantitative proteomics of the pea aphid-*Buchnera* symbiosis. Mol. Cell. Proteomics. 2011; 10:M110 007039. [PubMed: 21421797]

Richards GR, Goodrich-Blair H. Masters of conquest and pillage: *Xenorhabdus nematophila* global regulators control transitions from virulence to nutrient acquisition. Cell Microbiol. 2009; 11:1025–1033. [PubMed: 19374654]

Schadt E, Lamb J, Yang X, Zhu J, Edwards S, Guhathakura D, Sieberts SK, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 2005; 37:710–717. [PubMed: 15965475]

Schäfer J. Learning large-scale graphical Gaussian models from genomic data. AIP Conf. Proc. 2005; 776:263–276.

Shin SC, Kim SH, You Y, Kim B, Kim AC, Lee KA, Yoon JH, Ryu JH, Lee WJ. *Drosophila* microbiome modulates host developmental and metabolic homeostasis via insulin signaling. Science. 2011; 334:670–674. [PubMed: 22053049]

Shuylzhenko N, Morgun A, Hsiao W, Battle M, Yao M, Gavriolva O, Orandle M, Mayer L, MacPherson AJ, McCoy KD, Fraser-Liggett C, Matzinger P. Crosstalk between B lymphocytes, microbiota and the intestinal epithelium governs immune versus metabolism in the gut. Nat. Med. 2011; 17:1585–1593. [PubMed: 22101768]

Srinivasan BS, Caberoy NB, Suen G, Taylor RG, Shah R, Tengra F, Goldman BS, Garza AG, Welch RD. Functional genome annotation through phylogenomic mapping. Nat. Biotech. 2005; 23:691–698.

Sun N, Pan C, Nickell S, Mann M, Baumeister W, Nagy I. Quantitative proteome and transcriptome analysis of the archaeon *Thermoplasma acidophilum* cultured under aerobic and anaerobic conditions. J. Proteome Res. 2010; 9:4839–4850. [PubMed: 20669988]

Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE. A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola* . BMC Syst. Biol. 2009; 3:24. [PubMed: 19232131]

Tohge T, Fernie AR. Web-based resources for mass-spectrometry-based metabolomics: a user's guide. Phytochemistry. 2009; 70:450–456. [PubMed: 19285697]

van Baarlen P, Troost F, van der Meer C, Hooiveld G, Boekschoten M, Brummer RJ, Kleerebezem M. Human mucosal in vivo transcriptome responses to three lactobacilli indicate how probiotics may modulate human cellular pathways. Proc. Natl Acad. Sci. USA. 2011; 108(Suppl 1):4562–4569. [PubMed: 20823239]

Wang H, Wang Q, Pape UJ, Shen B, Huang J, Wu B, Li X. Systematic investigation of global coordination among mRNA and protein in cellular society. BMC Genomics. 2010a; 11:364. [PubMed: 20529381]

Wang Y, Carolan JC, Hao F, Nicholson JK, Wilkinson TL, Douglas AE. Integrated metabonomic-proteomic analysis of an insect-bacterial symbiotic system. J. Proteome Res. 2010b; 9:1257–1267. [PubMed: 19860485]

Wier AM, Nyholm SV, Mandel MJ, Massengo-Tiasse RP, Schaefer AL, Koroleva I, Splinter-Bondurant S, Brown B, Manzella L, Snir E, et al. Transcriptional patterns in both host and bacterium underlie a daily rhythm of anatomical and metabolic change in a beneficial symbiosis. Proc. Natl Acad. Sci. USA. 2010; 107:2259–2264. [PubMed: 20133870]

Wouts WM. Biology, life cycle and redescription of *Neoaplectana bibionis* Bovien 1937 (Nematoda; Stenernematidae). J. Nematol. 1980; 12:62–72. [PubMed: 19300673]

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat. Genet. 2008; 40:854–861. [PubMed: 18552845]

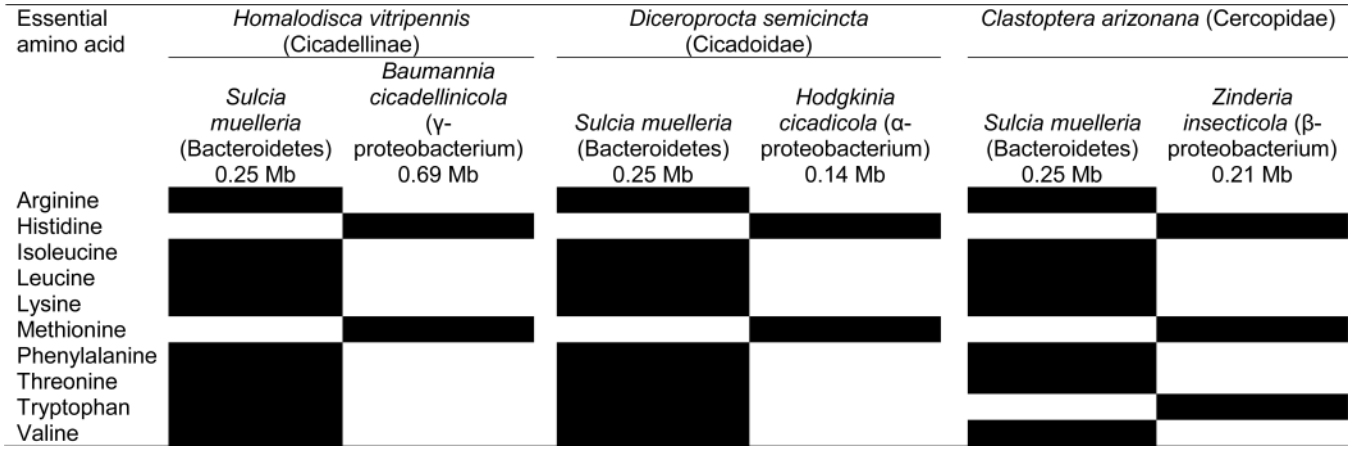

Figure 1.

Complementary genetic capacity for essential amino acid synthesis by the primary symbiont (*Sulcia*) and secondary symbiont in three groups of xylem-feeding insects. Solid bars, genes for biosynthetic pathway present on genome. Data collated from McCutcheon *et al.* (2009); McCutcheon and Moran(2007 & 2010)

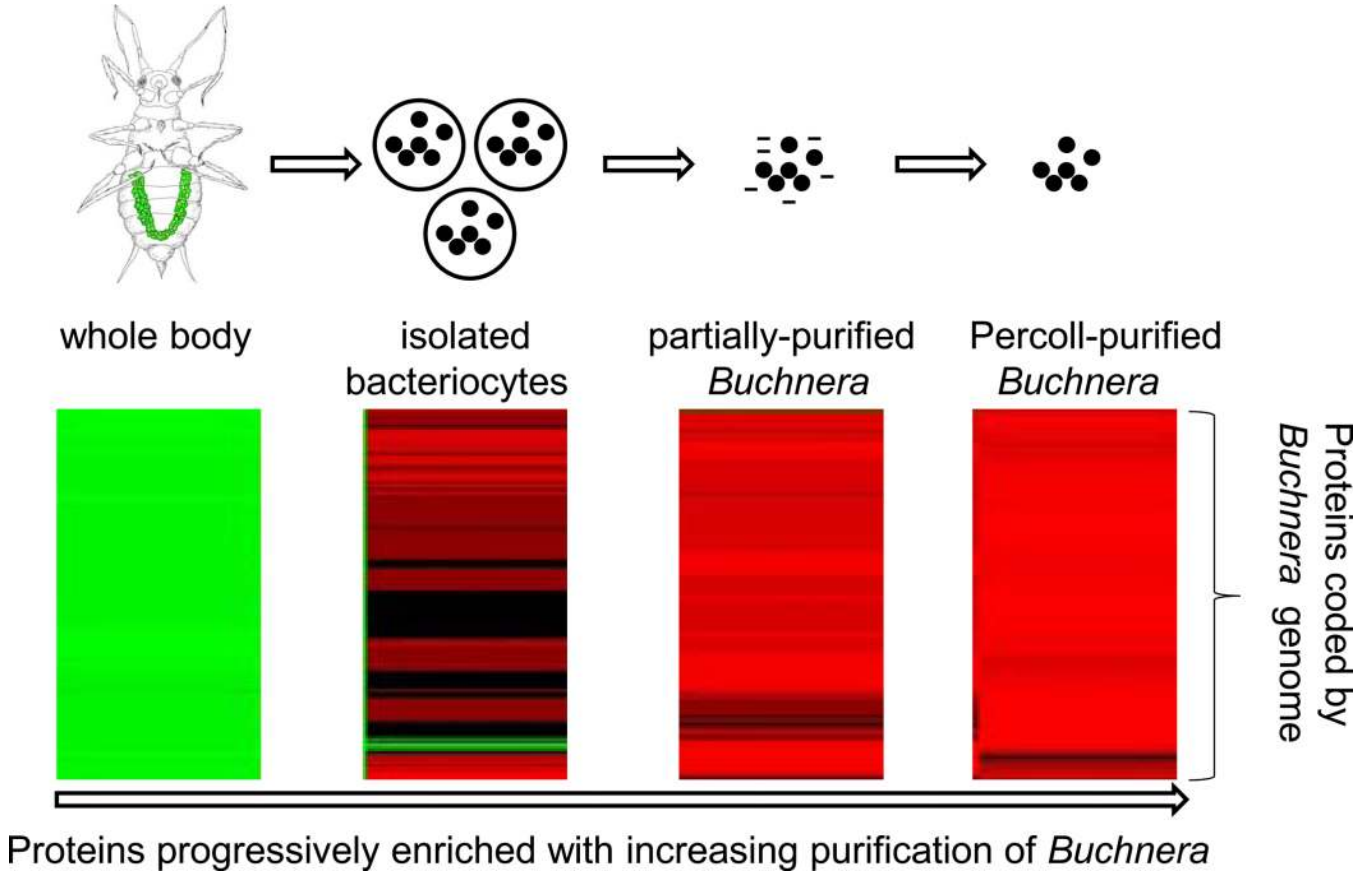

Figure 2.
Quantitative proteomic analysis of tissue fractions of the pea aphid-*Buchnera* symbiosis identifies proteins coded by the *Buchnera* genome enriched in *Buchnera* cells by hierarchical clustering (red, enriched; green, depleted). Data from Poliakov *et al.*(2011).

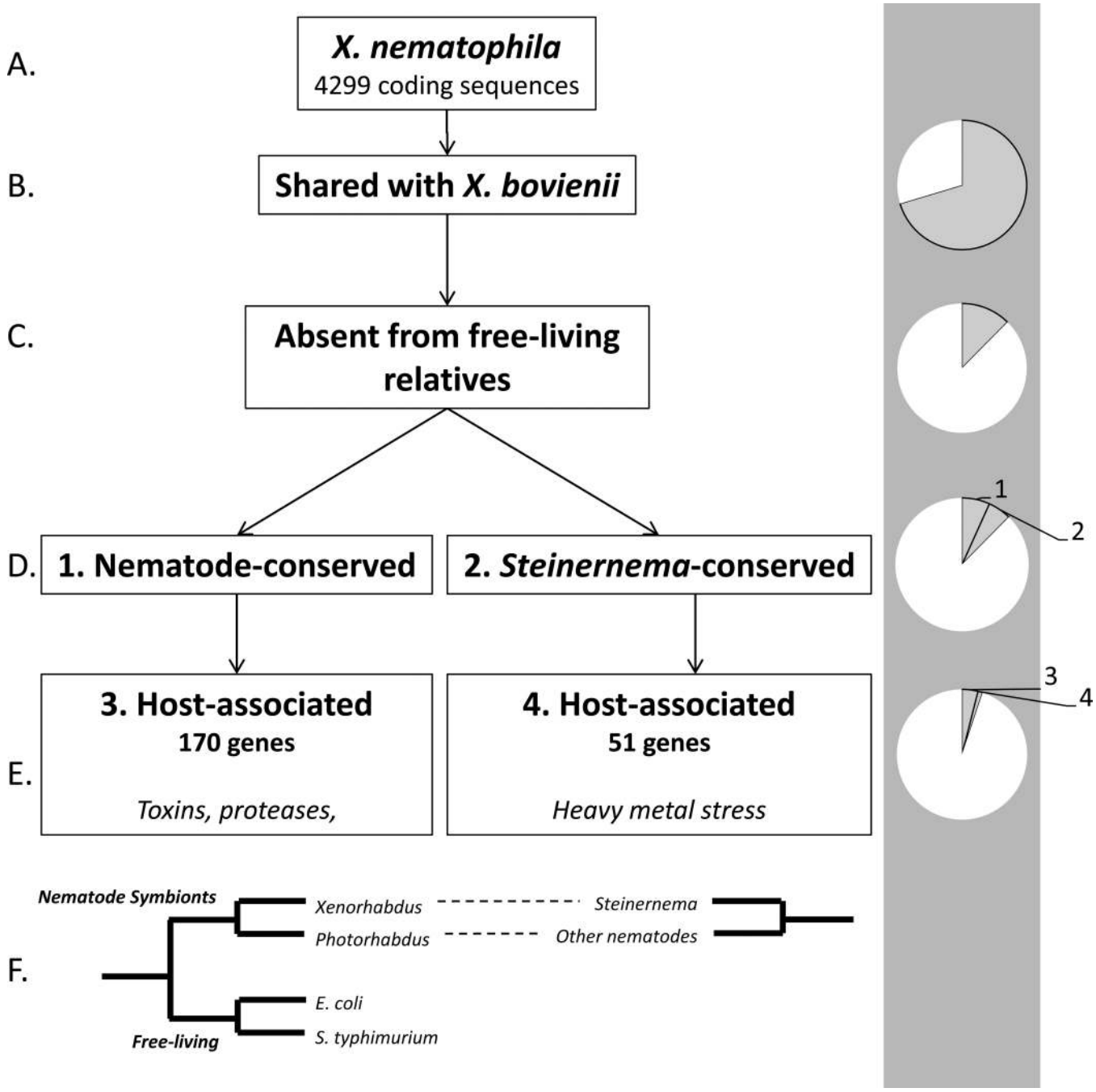

Figure 3.

Use of phylogenomics to identify candidate symbiosis-related genes in the bacterium *Xenorhabdus nematophila* (Enterobacteriaceae), which associates with entomopathogenic nematodes of the genus *Steinernema*.A) The pool of 4299 coding genes in the *X. nematophila* genome was reduced by B) subtracting all 1275 genes absent from the congeneric symbiont *X. bovienii* and C) subtracting 2491 non-symbiotic genes (shared with the free-living Enterobacteriaceae *Salmonella enterica* Typhimurium LT2 and *Escherichia coli* K12). D) The remaining 533 genes were divided as "nematode symbiosis-conserved" if they were shared in two con-familial *Photorhabdus* species that are also nematode symbionts (290 genes), or specific to the nematode host *Steinernema* ("*Steinernema*-conserved") (243

genes) if they were absent in the *Photorhabdus* species. E) Phylogenomic analyses performed on the two gene groups identified 15 and 9 clusters, respectively, of which 6 and 4 clusters were enriched in genes shared in plant and animal symbionts in the NCBI database (identified by custom metadata mining), yielding 170 and 51 genes for final analysis. F) Schematic of the phylogenetic relationship between *Xenorhabdus* and other bacteria that provided reference genomes used in the study, all of which are related in the Family Enterobacteriaceae. Data from Chaston *et al.* (2011).