environmental microbiology

Environmental Microbiology (2015) **17**(10), 3964–3975 doi:10.1111/1462-2920.12908

Identification and analysis of the bacterial endosymbiont specialized for production of the chemotherapeutic natural product ET-743

Michael M. Schofield,[1,2†] Sunit Jain,[3†] Daphne Porat,[2] Gregory J. Dick[3,4,5] and David H. Sherman[1,2,6,*]

[1]*Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA.*

[2]*Life Sciences Institute, Departments of* [3]*Earth and Environmental Sciences,* [5]*Ecology and Evolutionary Biology,* [6]*Medicinal Chemistry and Chemistry and* [4]*Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, USA.*

Summary

Ecteinascidin 743 (ET-743, Yondelis) is a clinically approved chemotherapeutic natural product isolated from the Caribbean mangrove tunicate *Ecteinascidia turbinata*. Researchers have long suspected that a microorganism may be the true producer of the anti-cancer drug, but its genome has remained elusive due to our inability to culture the bacterium in the laboratory using standard techniques. Here, we sequenced and assembled the complete genome of the ET-743 producer, *Candidatus* Endoecteinascidia frumentensis, directly from metagenomic DNA isolated from the tunicate. Analysis of the ~631 kb microbial genome revealed strong evidence of an endosymbiotic lifestyle and extreme genome reduction. Phylogenetic analysis suggested that the producer of the anti-cancer drug is taxonomically distinct from other sequenced microorganisms and could represent a new family of Gammaproteobacteria. The complete genome has also greatly expanded our understanding of ET-743 production and revealed new biosynthetic genes dispersed across more than 173 kb of the small genome. The gene cluster's architecture and its preservation demonstrate that the drug is likely essential to the interactions of the microorganism with its mangrove tunicate host. Taken together, these studies elucidate the lifestyle of a unique, and pharmaceutically important microorganism and highlight the wide diversity of bacteria capable of making potent natural products.

Received 10 April, 2015; revised 15 May, 2015; accepted 16 May, 2015. *For correspondence. E-mail davidhs@umich.edu; Tel. (+1) 734 615 9907; Fax (+1) 734 764 1247. †These authors contributed equally to this work.

Introduction

Natural products are a critical source of pharmaceuticals and lead compounds in drug discovery efforts (Newman and Cragg, 2012). Over the last several decades, scientists have isolated thousands of biologically active metabolites from terrestrial and marine macroorganisms, including plants and animals. Mounting evidence suggests that microbial symbionts may be the actual producers of many of these natural products (Piel, 2009).

Currently, the vast majority of drug-producing symbiotic microbes remain uncharacterized. Most fall into the > 99% of prokaryotic species currently incapable of being cultured in the laboratory, hindering their study (Staley and Konopka, 1985; Piel, 2009). Identifying these symbionts and understanding their genetic, biochemical and metabolic characteristics is critical for advancing fundamental knowledge and potential applications. Many symbiont-derived secondary metabolites can only be isolated in low yields from their hosts, making large-scale production for pharmaceutical purposes unsustainable from both an economic and environmental perspective. Although total synthesis can sometimes solve the supply problem, it can be costly and fails to address our understanding of the unique biosynthetic processes that are mediated by these elusive microbes. Sequencing and analysis of symbiont genomes could provide insight into the lifestyles of these poorly understood bacteria, illuminate possible host-free cultivation methods, and provide a route to economical and sustainable large-scale production with the opportunity for genetic manipulation to produce novel drug analogs.

The chemotherapeutic compound ET-743 (**1**, Yondelis, Trabectedin) is one of the most important natural products suspected to be of symbiotic origin. Isolated directly from the mangrove tunicate *Ecteinascidia turbinata* (Fig. 1A and B), the biological activity of the drug against cancer cells has inspired over 40 years of research (Lichter *et al.*, 1975; Rinehart *et al.*, 1990). Currently, ET-743 is clinically approved in Europe against soft tissue sarcoma and

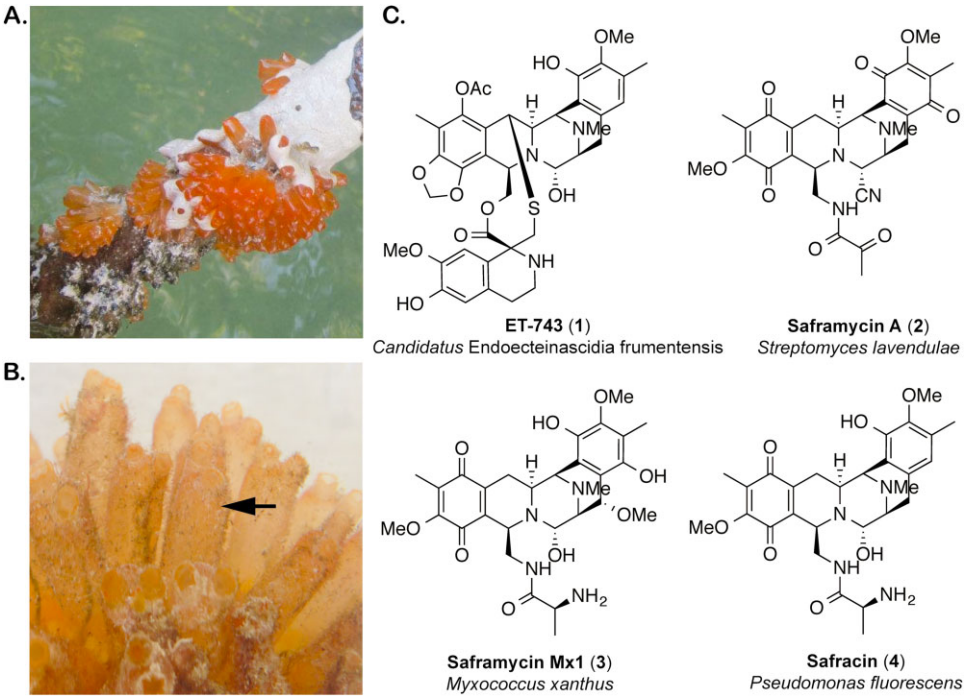

Fig. 1. A. Tunicate colonies growing on the root of a mangrove tree in the Florida Keys.
B. A tunicate colony composed of individual zooids (indicated by arrow). In this study, we sequenced the metagenomic DNA from four zooids.
C. The chemotherapeutic compound ET-743 (**1**) and three natural products from cultivable bacteria that share a similar tetrahydroisoquinoline core.

relapsed ovarian cancer and is currently in phase III trials as an anticancer therapeutic in the United States (McLaughlin, 2015).

The tetrahydroisoquinoline alkaloid natural products saframycin A (**2**), saframycin Mx1 (**3**) and safracin (**4**) are derived from three distinct cultivable bacteria and are structurally similar to ET-743, supporting a prokaryotic origin for the drug (Fig. 1C). Studies of the mangrove tunicate over a decade ago identified the potential intracellular Gammaproteobacterium *Candidatus* Endoecteinascida frumentensis to be the most prevalent member of the host microbial consortium (Moss *et al*., 2003; Pérez-Matos *et al*., 2007) and the only microorganism consistently associated with tunicates in both the Mediterranean and Caribbean seas (Pérez-Matos *et al*., 2007). A metagenomically derived contig containing a partial ET-743 biosynthetic gene cluster was later indirectly linked to a separate contig bearing the 16S rRNA gene sequence for *Ca*. E. frumentensis through analysis of %G+C content and codon usage (Rath *et al*., 2011). Cultivation of the producing bacterium has so far been unsuccessful (Moss *et al*., 2003; Pérez-Matos *et al*., 2007), and aquaculture (Carballo *et al*., 2000) of the host tunicate and total synthesis (Corey *et al*., 1996) have also failed to provide sustainable access to the drug for clinical applications. ET-743 is therefore currently generated by a lengthy semisynthetic process starting from fermentation-derived cyanosafracin B (Cuevas and Francesch, 2009).

In this study, we utilized next generation sequencing technologies to expand our understanding of ET-743 biosynthesis and uncover the complete genome of the microorganism responsible the drug's production. Analysis of phylogenetic markers and protein coding genes suggests that the microbe belongs to a novel family of Gammaproteobacteria. In-depth genomic analysis also provides initial insights into the endosymbiotic lifestyle of *Ca.* E. frumentensis, the ecological role of its sole secondary metabolic pathway and key information that may provide access to host-cell free growth in the laboratory.

Results and discussion

Overview of samples and dataset

The colonies of *E. turbinata* consist of thick bundles of individual zooids connected by a network of stolons that enable adherence of the animal to a stable surface. Our laboratory previously isolated metagenomic DNA from individual zooids and uncovered a 35 kb gene cluster responsible for ET-743 biosynthesis using 454 pyrosequencing (Rath *et al*., 2011). In the present study, we isolated additional metagenomic DNA from four zooids obtained from two colonies (Fig. S1). We shotgun sequenced the resulting DNA samples using Illumina

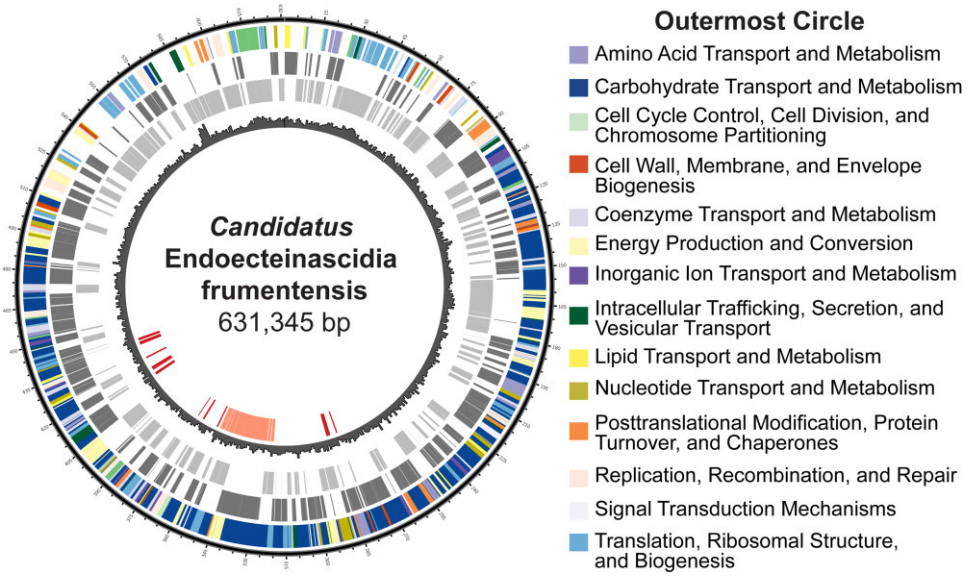

Fig. 2. A circular map of the closed genome of *Candidatus* E. frumentensis. The outermost circle displays protein-coding genes assigned to Pfam categories (see key). The dark grey and light grey circles display protein-coding genes on the plus strand and minus strands respectively. The fourth circle depicts a histogram of G+C content throughout the genome. The innermost circle represents ET-743 biosynthetic genes. Genes previously identified are depicted in light red while putative new genes are shown in dark red.

HiSeq technology and assembled the data into contigs. The four zooids provided metagenome datasets each containing over 800 Mbp of sequence (Table S1).

We assigned the assembled contigs to taxonomic bins using tetranucleotide frequency with emergent self-organizing maps (tetra-ESOM) as previously described (Fig. S2) (Dick *et al.*, 2009). Each of the four metagenomic samples possessed a single bin containing both the previously identified partial ET-743 biosynthetic gene cluster and the 16S rRNA gene for *Ca.* E. frumentensis (Table S1). The four bins containing the ET-743 producing microorganism were further assembled into a consensus genome containing three contigs. PCR amplification closed a 200 bp gap between two of the contigs to create a 630 kb scaffold. Additional PCR amplification closed a final 1.5 kb gap in the scaffold to create the closed genome for *Ca.* E. frumentensis (Fig. 2, Table 1, Fig. S1).

The coverage depth for the endosymbiotic genome averaged 721× between the four samples (Table S2). However, one contig that consistently binned with the ET-743 producer and was retrieved in all four samples was not incorporated into the genome. This much smaller ~18 kb contig encodes a DNA primase and two protein-coding genes with ambiguous functions that repeat throughout the stretch of the sequence. Unlike the circular genome, the shorter contig has a coverage depth of only ~74× and reads could not be mapped to the sequence with confidence (Table S2). The excluded contig may be an extrachromosomal element that is present in only a subset of the *Ca.* E. frumentensis population or an artifact of the assembly and binning process. Given that the rest of the genome was closed and displayed even and deep coverage, we focused our analysis on the closed *Ca.* E. frumentensis genome in this study.

Very few other genomic bins were detected in the metagenomic datasets, despite prior evidence that the tunicate housed a complex microbial consortium (Table S1) (Moss *et al.*, 2003; Rath *et al.*, 2011). However, previous studies indicated *Ca.* E. frumentensis was one of the most abundant microorganisms in the consortium (Moss *et al.*, 2003; Pérez-Matos *et al.*, 2007; Rath *et al.*, 2011) and the only microorganism found to be consistently associated with the tunicate host in both the Mediterranean and Caribbean marine habitats (Pérez-Matos *et al.*, 2007). Further, metagenomic assembly of the sym-

Table 1. General features of the *Candidatus* E. frumentensis genome.

Candidatus E. frumentensis	
Genome size (bp)	631,345
Taxonomy	New family of Gammaproteobacteria
GC content (%)	
Total	23.3
Coding regions	24.2
Noncoding regions	12.7
Coding density (%)	91.5
Intergenic pseudogenes	10
Protein-coding genes	586
With functional annotation	559 (95.4%)
With ambiguous function	27 (4.6%)
rRNA genes	3
tRNA genes	32

biont population was likely facilitated due to its low genomic diversity compared to populations that are non-specifically associated. Thus, it is likely that the eukaryotic host and *Ca*. E. frumentensis monopolized the sequencing data, especially the large assembled contigs, despite the presence of a complex but lower abundance microbial community. The only other notable bin after tetra-ESOM was a cyanobacterium from the order *Oscillatoriales* that was present in two of the four metagenomic DNA samples (Table S1, Fig. S2).

Genome reduction in the symbiont

Previous *in situ* hybridization analysis provided an initial indication that *Ca*. E. frumentensis could be a bacterial endosymbiont (Shigenobu *et al*., 2000; Wernegreen, 2002). Assembly and analysis of the microbe's complete genome provides further convincing evidence of an intracellular lifestyle and long-term evolution with the tunicate host, *E. turbinata*. *Ca.* E. frumentensis possesses many of the hallmarks of genome reduction, which is thought to be driven by a small bacterial population size and an inherent deletion bias (Moran, 1996; Moran *et al*., 2008; McCutcheon and Moran, 2012). The circular genome for *Ca.* E. frumentensis is quite small, totaling only 631 345 bp (Fig. 2). The small size of the genome rivals those of the model obligate endosymbionts *Buchnera aphidicola* in aphids and *Wigglesworthia glossinidia* in tsetse flies (Table S3). The functions maintained by *Ca*. E. frumentensis are also consistent with the minimal gene sets observed in these and other obligate symbionts (Fig. S3). For example, *Ca.* E. frumentensis appears to have lost a number of genes involved in DNA replication and repair mechanisms (Fig. S3). The loss of DNA repair mechanisms is thought to be a crucial turning point during the evolution of an endosymbiont (Moran *et al*., 2008; McCutcheon and Moran, 2012). Loss of these genes is frequently accompanied by increased mutation rates, an A + T DNA sequence bias, and the loss of additional non-essential genes.

Indeed, the exceptionally low total G+C content (23.3%) of *Ca.* E. frumentensis genomic DNA supports a mutational bias and an obligate endosymbiotic lifestyle. The G+C content disparity between the coding (24.2%) and noncoding (12.7%) regions of the genome (Table 1) further exemplifies this bias. Bacterial lineages that only recently became restricted to a host organism also often have higher numbers of pseudogenes within these noncoding regions and a consequently low overall coding density (Kuo *et al*., 2009). However, as bacteria continue to co-evolve with their hosts, pseudogenes gradually shrink and become unrecognizable through deletions while genomes become more compact (Moran, 1996; Kuo and Ochman, 2009). The noncoding regions of the *Ca.* E. frumentensis genome have only 10 pseudogenes whose predicted translation products show amino acid sequence similarity to known proteins (Table S4). The genome also has a higher overall coding density of 90.7% (Table 1), similar to *B. aphidicola*, *W. glossinidia*, and other obligate endosymbionts that co-evolved with their hosts along the order of millions of years (Moran and Munson, 1993; Moran *et al*., 2008). Taken together, these data provide strong support that *Ca*. E. frumentensis is an obligate endosymbiont that has undergone long-term co-evolution with the tunicate host, *E. turbinata*.

Phylogenetic analysis and novelty of Ca*. E. frumentensis*

The genome of *Ca*. E. frumentensis also appears to be remarkably distinct from other studied microorganisms. Analysis of conserved markers provided the first evidence that *Ca.* E. frumentensis may be phylogenetically distant from characterized bacterial species. The closest homologues for genes encoding the 16S rRNA gene, *rpoB*, and *recA* had 86.1%, 69.0% and 74.8% sequence identities respectively (Fig. S5).

Phylogenetic markers can be useful for microorganisms that have many well-studied and cultivable close relatives. However, in microorganisms with fewer obvious relatives, the average amino acid identity (AAI) of shared genes can be more revealing (Konstantinidis and Tiedje, 2005). To further explore the phylogenetic novelty of *Ca.* E. frumentensis, we compared the AAI and 16S rRNA gene of the microorganism to other bacterial species selected from a taxonomic profile of the *Ca*. E. frumentensis genome. This analysis confirmed that *Ca.* E. frumentensis is taxonomically distinct from many of its originally predicted relatives and likely represents a new family of Gammaproteobacteria (Fig. S5) (Yarza *et al*., 2014).

Primary metabolism

Analysis of the endosymbiont's primary metabolism provided further insight into the lifestyle of *Ca*. E. frumentensis (Fig. 3). The small genome appears to have portions of all three components of central metabolism, including the tricarboxylic acid cycle (TCA cycle), the non-oxidative branch of the pentose phosphate pathway and most of the glycolytic pathway (Fig. 3). Although the genome is missing genes involved in early glucose catabolism, it does encode several sugar phosphate transporters. Sugar phosphates may therefore represent an important carbon source for the endosymbiont, similar to other microorganisms living in an intracellular environment (Munoz-Elias and McKinney, 2006).

Like most obligate endosymbionts and many intracellular pathogens, *Ca*. E. frumentensis is also missing a

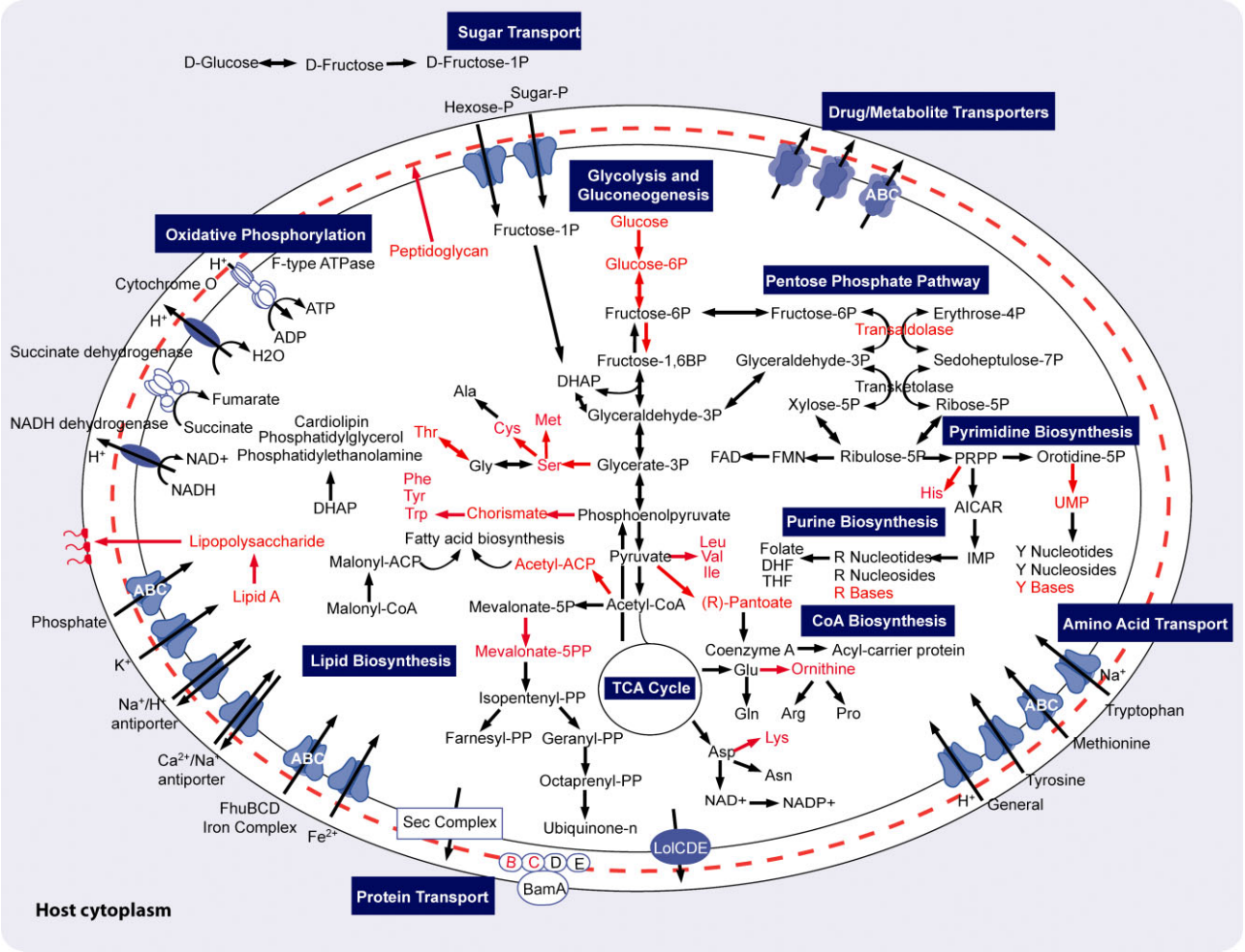

Fig. 3. Overview of the metabolism of *Ca.* E. frumentensis deduced from genomic analysis. Reaction products depicted in red have either missing or partially missing biosynthetic pathways. ACP, acyl carrier protein; AICAR, 5-aminoimidazole carboxamide ribonucleotide; CoA, coenzyme A; DHAP, Dihydroxyacetone phosphate; DHF, dihydrofolate; DMAPP, dimethylallyl pyrophosphate; FAD, flavin adenine dinucleotide; FMN, flavin mononucleotide; IMP, inosine monophosphate; NAD, nicotinamide adenine dinucleotide; PRPP, phosphoribosyl pyrophosphate; THF, tetrahydrofolate; UMP, uridine monophosphate.

number of key amino acids and cofactors (Fig. 3). The genome only has the machinery to generate asparagine, aspartic acid, glutamate and glutamic acid *de novo*. There are only partial gene sets for the remaining amino acids and several cofactors, including coenzyme A (CoA). It is likely that the endosymbiont acquires some of these essential metabolites or their precursors from the tunicate host. Indeed, the endosymbiont encodes 71 genes putatively linked to transporter function, including several involved in amino acid import (Fig. 3).

The *Ca.* E. frumentensis genome also has gene sets for the biosynthesis of lipids commonly incorporated into bacterial membranes, including phosphatidylethanolamine, cardiolipin and phosphatidylglycerol (Fig. 3). However, the genome is missing a number of genes involved in the biosynthesis of peptidoglycan and lipid A biosynthesis. The vast majority of bacteria incorporate some level of peptidoglycan into their cell walls and most Gram-negative bacteria possess lipid A-containing lipopolysaccharides in their outer membrane. However, some microorganisms undergoing genome reduction have been known to lack both of these usually standard components (Pérez-Brocal *et al.*, 2006; Wu *et al.*, 2006; Moran *et al.*, 2008; Nakabachi *et al.*, 2013). The absence of the majority of these genes within *Ca.* E. frumentensis further highlights the extent of its genome reduction.

Secondary metabolism

We previously identified a 35 kb contig containing many of the genes involved in the biosynthesis of the chemotherapeutic natural product ET-743 (Rath *et al.*, 2011). However, close examination of ET-743, its previously isolated precursors (Rinehart *et al.*, 1990) and other

well-studied tetrahydroisoquinoline natural products (Pospiech *et al*., 1995; Velasco *et al*., 2005; Lei *et al*., 2008; Hiratsuka *et al*., 2013) led us to suspect that we were still missing a number of key biosynthetic genes (Rath *et al*., 2011). Expanding the 35 kb gene cluster to a complete genome for *Ca*. E. frumentensis has enabled us to identify many of these previously missing genes and improved our understanding of ET-743 biosynthesis. Key genes involved in production of the chemotherapeutic drug are dispersed over 173 kb of the small 631 kb genome (Fig. 2). Biosynthetic genes are split into three distinct regions within this expansive genomic range (Fig. 4A, Table S5). Newly detected gene products include the acetyltransferase EtuY and EtuM4, likely involved in acetylation and N-methylation to make **7** and ET-597 (**9**) respectively. We also identified three new flavoproteins in addition to the FAD-dependent monooxygenase (EtuO1) contained within the original ET-743 biosynthetic gene cluster (Rath *et al*., 2011).

We additionally identified a gene encoding the E3 component of the pyruvate dehydrogenase complex (EtuP3, Fig. 4). The reactions catalyzed by this enzyme system typically provide the TCA cycle with acetyl-CoA (Patel *et al*., 2014). However, the primary metabolic enzymes were recently shown to also contribute to the biosynthesis of quinocarcin and naphtyridinomycin natural products (Peng *et al*., 2012). The enzyme complex can work with an acyl carrier protein (ACP) to provide a glycolicacyl-S-ACP extender unit (**5**) for a non-ribosomal peptide synthetase (NRPS). Both of these gene clusters in addition to SF-1739 (Hiratsuka *et al*., 2013) and the original ET-743 (Rath *et al*., 2011) biosynthetic gene cluster contain the E1 and E2 components for the enzyme complex. Although the E3 component has been absent in previously studied clusters, purified exogenous E3 does seem necessary for complete product conversion (Peng *et al*., 2012). The presence of the E3 component in *Ca*. E. frumentensis and its proximity to other ET-743 biosynthetic genes further exemplifies its importance in the biosynthesis of tetrahydroisoquinoline natural products.

Another genomic feature that may set the ET-743 biosynthesis apart from other natural products is the placement of the ACP that operates with the pyruvate dehydrogenase complex. The ACP is located in the main biosynthetic gene clusters for quinocarcin, naphtyridinomycin and SF-1739. However, the only ACP in the entire *Ca*. E. frumentensis genome is located within a region containing fatty acid biosynthetic genes 61 kb downstream of the original ET-743 gene cluster (EtuF9, Fig. 4A). The location of the ACP and the presence of other fatty acid biosynthetic genes (EtuF1 and EtuF2) within the original ET-743 biosynthetic gene further supports potential interaction between primary and secondary metabolism during ET-743 biosynthesis. This ACP most likely functions in concert with EtuP1, EtuP2 and EtuP3 to provide the glycolicacyl-S-ACP extender unit (**5**) to EtuA1 (Fig. 4B).

Despite these new discoveries, we may still be missing some genes involved in ET-743 biosynthesis. For example, gene candidates for enzymes that catalyze formation of the thioether ring (**8**) and transamination to make ET-596 (**10**) remain to be identified. We cannot rule out that these genes may be located elsewhere in the *Ca*. E. frumentensis genome or that the microbe works together with its host to complete construction of the chemotherapeutic compound. Similar host–endosymbiont cooperation has been observed during the biosynthesis of parasitic plant fungus natural product rhizoxin (Lackner *et al*., 2011). Symbiotic bacteria have also been known to cooperatively biosynthesize compounds. However, previous findings suggesting that *Ca*. E. frumentensis is the only microorganism consistently associated with the tunicate (Pérez-Matos *et al*., 2007) make other symbionts a less likely source for additional biosynthetic genes.

The *Ca*. E. frumentensis genome also contains several widely dispersed genes found within the biosynthetic gene clusters of other tetrahydroisoquinoline natural products. For example, the gene encoding the excision nuclease subunit UvrA is found within the saframycin A, and SF-1739 and quinocarcin gene clusters, perhaps playing a role in repairing damage induced by these potent natural products. However, the gene in the *Ca*. E. frumentensis genome is located several hundred base pairs upstream from the original ET-743 gene cluster. The saframycin A gene cluster also contains a complete gene set for the recycling of S-adenosyl methionine, a coenzyme essential for methyltransferase activity during the biosynthesis of all tetrahydroisoquinoline natural products. The complete gene set for the recycling system is still present in the *Ca*. E. frumentensis genome, but the genes are located both upstream and downstream of the original 35 kb gene cluster.

The semi-dispersed nature of ET-743 biosynthetic genes is notable as microbial secondary metabolite systems are typically tightly clustered in bacteria with clearly identifiable boundaries (Malpartida and Hopwood, 1984; Walton, 2000; Chu *et al*., 2011). However, genes involved in ET-743 biosynthesis are located in different points throughout the genome, interspersed with genes involved in primary metabolism (Fig. 4). The fragmented nature of *Ca*. E. frumentensis secondary metabolism could be a consequence of horizontal gene transfer (Lawrence and Roth, 1996) and co-regulation of gene expression within operons (Price *et al*., 2005), which are two important forces thought to encourage selection and formation of gene clusters. However, the endosymbiont lifestyle provides few opportunities for horizontal gene

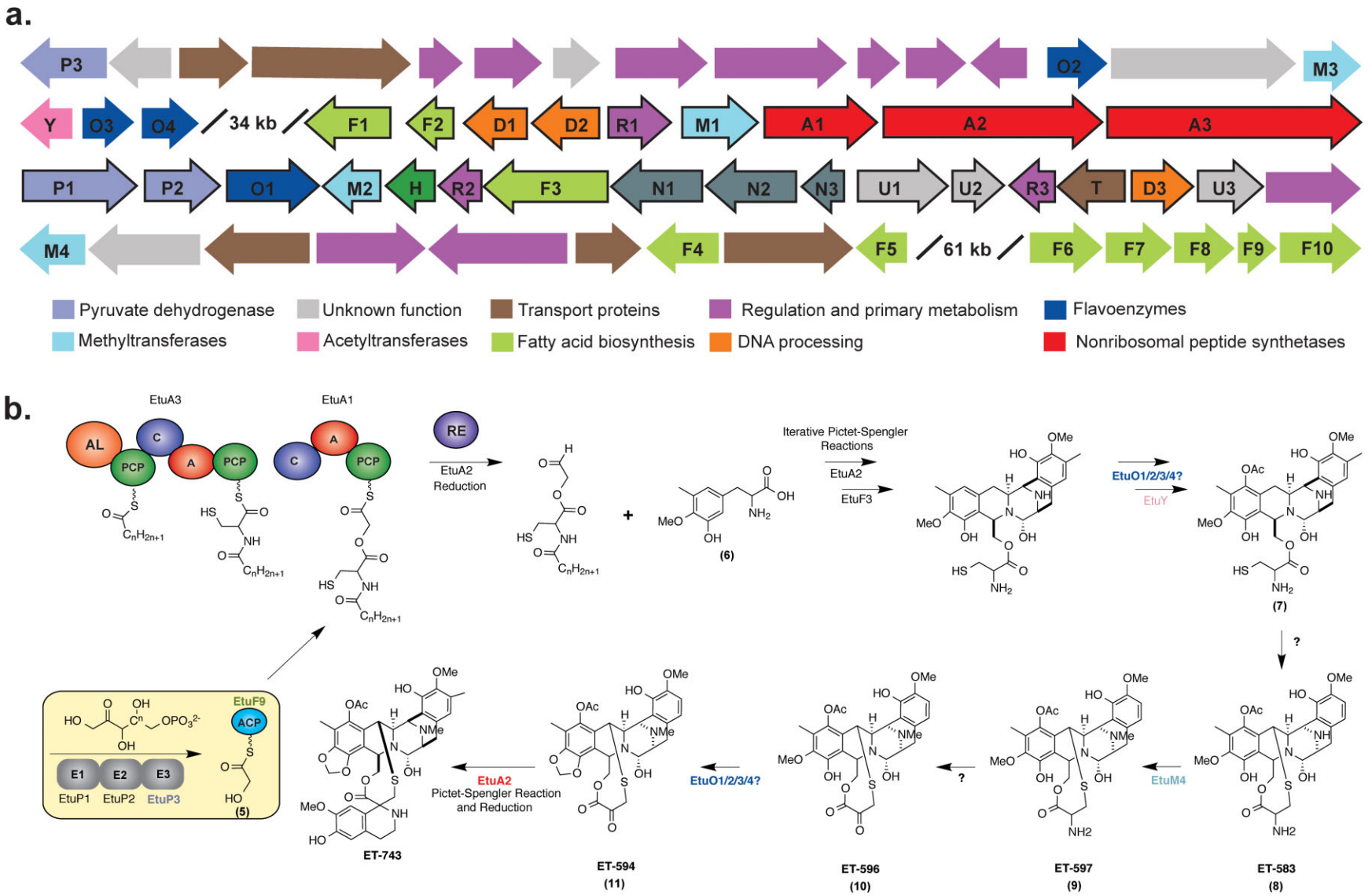

Fig. 4. The identification of new genes with suspected involvement in ET-743 biosynthesis. The genes and their putative roles are also depicted in Table S5.
A. New ET-743 biosynthetic genes were identified upstream and downstream of the original ET-743 biosynthetic gene cluster (outlined in black). Gene products are classified according to the corresponding colour key.
B. A condensed ET-743 biosynthetic pathway illustrating proposed new steps based on analysis of the complete genome. Coloured steps represent new enzymes or new roles for previously identified enzymes. An updated proposal for the complete biosynthesis of ET-743 is depicted in Fig. S6.

transfer, and regulatory mechanisms are often among the first genetic elements lost during genome reduction (Moran *et al*., 2008; McCutcheon and Moran, 2012). The lack of selective pressure to retain clusters is thought to contribute to fragmentation of biosynthetic genes in other endosymbionts (Kwan *et al*., 2012), and likely also plays a role in the organization of genes involved in ET-743 production. The genome no longer possesses a canonical gene cluster, but instead contains scattered biosynthetic genes that may function *in trans*.

Analysis of the *Ca*. E. frumentensis genome has also improved our understanding of the importance of ET-743 biosynthesis in the relationship between the endosymbiont and the tunicate host, *E. turbinata*. In long-term co-evolution, bacterial genes that are useful to the host are retained despite ongoing genome erosion (Moran *et al*., 2008; McCutcheon and Moran, 2012). The survival of ET-743 biosynthetic genes despite clear evidence of extreme genome reduction is indicative of an important role for the secondary metabolite to the host. A query of the endosymbiont genome against the full complement of bioinformatics tools revealed that ET-743 was the only natural product gene cluster found within the genome, further exemplifying its ecological value to the tunicate. Adult ascidians such as *E. turbinata* are sessile marine invertebrates with soft bodies, making them particularly vulnerable to predation. Their large larvae are released during daylight hours, making them similarly susceptible to predators. The secondary metabolite ET-743 could serve as a defense mechanism for the host. Many other ascidians and sponges are thought to produce secondary metabolites and inorganic acids that make them unpalatable (Lindquist *et al*., 1992). Indeed, ecological studies have already demonstrated that taste and orange colouring of larvae from *E. turbinata* protects the animal against predators (Young and Bingham, 1987). If ET-743 is the chemical deterrent responsible for protecting the host, it provides a driving force to assure the survival of ET-743 biosynthetic genes despite millions of years of genome reduction.

Conclusions

We have assembled a complete genome for *Ca*. E. frumentensis, an endosymbiont responsible for production of the chemotherapeutic drug ET-743. Microbial symbionts like *Ca*. E. frumentensis have long been thought to be the source of many natural products isolated from terrestrial and marine invertebrates. However, very little is known about the majority of these microbes due to our current inability to culture them in the laboratory.

The complete genome of *Ca*. E. frumentensis has enriched our understanding of ET-743 biosynthesis. The discovery of new ET-743 biosynthetic genes will enable future biochemical studies to confirm the roles of individual enzymes. A better understanding of its biosynthesis can facilitate future *in vitro* and heterologous expression efforts to engineer sustainable production of the drug and related analogs. Analysis of the complete genome has also highlighted the importance of ET-743 to the host–symbiont relationship. The lack of genomic evidence for other secondary metabolites, the survival of the gene cluster despite extreme genome reduction and the dispersal of ET-743 genes across the small genome suggests the microbe has become specialized for production of the drug. The chemotherapeutic natural product is therefore likely crucial to the microorganism's relationship with the tunicate host and its continued survival. This is intriguing since secondary metabolites are traditionally thought to be nonessential for microbial life (Williams *et al*., 1989) despite their prevalence in microbial genomes and ability to confer competitive advantages (Stone and Williams, 1992). However, improved sequencing technologies and metagenomic pipelines now permit more detailed studies of genomes undergoing reduction. Full genome studies on the endosymbionts found in macroorganisms like insects (Nakabachi *et al*., 2013), tunicates (Kwan *et al*., 2012; Kwan and Schmidt, 2013) or even fungi (Lackner *et al*., 2011) provide increasing evidence that natural products may sometimes play essential roles. When these secondary metabolites benefit a host organism, their preservation may ensure a microorganism's survival and even facilitate co-evolution with a host. The drastically reduced genome of *Ca*. E. frumentensis presented here further supports this theory.

A better understanding of symbiont genomes along with their primary and secondary metabolism could provide new routes to economical and sustainable large-scale production of bioactive natural products. Analysis of the drastically reduced genome of *Ca*. E. frumentensis provides unique insight into the microorganism's lifestyle and clues to possible host-free cultivation. Previous attempts to grow the microorganism in the laboratory were unsuccessful. However, our ability to culture elusive microorganisms is continually improving. Recent advances in host-cell free growth of *Coxiella burnetii* (Omsland *et al*., 2009) or the facultative symbionts *Burkholderia* spp., *Rhodococcus rhodnii* and *Wolbachia* spp. (Kikuchi, 2009) motivate future efforts to develop suitable growing conditions and techniques to access the uncultivable majority of bacteria. Genome analysis in particular has proven a powerful method to pinpoint nutrient and oxygen requirements for microbial growth (Kikuchi, 2009; Omsland *et al*., 2009). The loss of key primary metabolic pathways in *Ca*. E. frumentensis suggests that the microorganism could not live independently of the host using standard media and cultivation techniques. The loss of genes involved in amino acid, CoA and glucose biosynthesis indicates that

media enhanced with nutrients, cofactors and alternative carbon sources may be necessary. However, genomic evidence for aerobic respiration and transporters for key metabolites indicates that the right environmental conditions might lead to host-cell free growth.

Experimental procedures

Sample collection and isolation of metagenomic DNA

Two tunicate colonies were collected off the coast of the Florida Keys. Animals were immediately frozen on dry ice after collection and stored at −80°C until processing. Metagenomic DNA was isolated from single zooids plucked from each colony (Fig. S1) following the protocol outlined for mouse tails in the Wizard Genomic DNA Purification Kit (Promega).

Genome sequencing, assembly, binning and annotation

The four metagenomic samples were shipped on dry ice to the Joint Genome Institute (JGI) for immediate sequencing. Gene calling and annotation of the assembled metagenome was then completed through JGI IMG/M (Markowitz *et al*., 2013). The JGI Submission IDs and Taxon Object IDs for these four samples are listed in Table S1. Individual contigs from each assembly were assigned to taxonomic groups through binning with tetranucleotide frequency with ESOM as described previously (Dick *et al*., 2009). Since the metagenomes had an excess of sequences belonging to the eukaryotic host tunicate, iterative rounds of ESOM were required to hone in on microbial communities present in the sample.

Genes from the previously identified ET-743 biosynthetic gene cluster (Rath *et al*., 2011) and the 16S gene for *Ca*. E. frumentensis (Moss *et al*., 2003; Pérez-Matos *et al*., 2007; Rath *et al*., 2011) were used as BLAST queries to identify the bin containing the ET-743 producer in each of the four metagenomic samples. The four resulting bins were manually evaluated for completeness through the analysis of the distribution of conserved phylogenetic markers (Ciccarelli, 2006). Contigs from the four bins were assembled into a consensus genome with Geneious (v. 7.1.3).

Closing genomic gaps

We designed primers upstream of any suspected genomic gaps and carried out PCR using KOD Xtreme™ Hot Start DNA Polymerase (Novagen). Reactions contained 0.02 U μl^{-1} polymerase, 1X of the supplied buffer, 0.3 μM custom primers, 0.4 mM each dNTP, and 100 ng of metagenomic DNA. Reactions consisted of a hot start (94°C, 2 min), followed by 35 cycles of denaturing (98°C, 10 s), annealing (variable temperatures for 30 s), and extension (68°C for variable times). Since we were unsure about the size of genomic gaps, we began with a longer extension time of 5 min. If we saw a DNA band after running reactions on a 1% agarose gel, we repeated PCR and tailored the extension time to the size of the band (1 min kbp^{-1}) to limit any nonspecific amplification. Amplified DNA was then isolated from agarose gels using the standard protocol from the Wizard® SV Gel and PCR Clean-Up Kit (Promega).

Samples were submitted for Sanger sequencing with the primers used in the PCR reactions. Primer walking along the DNA strand then provided the missing sequence within both gaps. The complete consensus genome was submitted to JGI IMG (Markowitz *et al*., 2014) for gene calling and annotation (Taxon Object ID: 2616645016; Analysis Project ID: Ga0072939). The final genome was reassessed for the completeness and accuracy through analysis of the distribution of conserved phylogenetic markers (Ciccarelli, 2006).

Genome analysis

The common genes included in Fig. S3 were compiled from other studies examining genome reduction in endosymbionts and intracellular pathogens (Moran *et al*., 2008; Kwan *et al*., 2014). Analysis of primary metabolic pathways was completed using the KEGG and MetaCyc annotations provided through JGI/IMG. To confirm the absence of any missing genes, protein sequences from a model organism (typically from *Escherichia coli* E12) were used as queries in a BLASTP search against the *Ca*. E. frumentensis annotated genome.

To detect pseudogenes, all intergenic regions larger than 100 bp were used as BLASTX queries against the entire NR database using default settings. Any hits with e-values lower than 1×10^{-3} against nonhypothetical proteins were considered pseudogenes.

Visualization of the complete genome (Fig. 2) was constructed using Circos (Krzywinski *et al*., 2009). Data for circles displaying Pfam categories for protein-coding genes, genes on the plus strands, and genes on the minus strands were provided directly through JGI IMG annotations and analysis.

To detect natural product gene clusters, the full genome was analysed with a host of previously described bioinformatics tools, including antiSMASH 2.0 (Blin *et al*., 2013), NP.searcher (Li *et al*., 2009), CLUSEAN (Weber *et al*., 2009), BAGEL3 (van Heel *et al*., 2013) and 2metdb (Bachmann and Ravel, 2009).

Phylogenetic analysis

The gene sequences for conserved phylogenetic markers (16S rRNA, *rpoB* and *recA*) were used as BLASTN queries against the NT database. Trees were constructed with Geneious (v. 7.1.3) after ClustalW multiple alignments with an IUB cost matrix (default settings). Neighbour-joining trees were constructed with the Jukes–Cantor genetic distance model (default settings). Top hits for cultivable or well-studied uncultivable microorganisms were included in the phylogenetic tree for 16S rRNA gene sequences. All unique hits for *rpoB* and *recA* were used in respective genetic trees.

To further explore taxonomic uniqueness (Fig. S5), the complete or draft genomes of the top hits from phylogenetic analysis were used in a two-way BLAST against *Ca*. E. frumentensis to acquire average amino acid identity (AAI) as previously described (Konstantinidis and Tiedje, 2005). Thresholds for unique taxonomic rankings were based on

16S rRNA gene sequence identity as previously described (Yarza *et al*., 2014).

Acknowledgements

We thank Erich Bartels for use of facilities at Mote Marine Laboratories in the Florida Keys and for assistance with field collecting of *E. turbinata*. We also thank Tijana Glavina del Rio and Susannah Tringe at the Joint Genome Institute for their assistance. This research was supported by the International Cooperative Biodiversity Groups initiative (U01 TW007404) at the Fogarty International Center, the NSF under the CCI Center for Selective C–H Functionalization, CHE-1205646 and the Hans W. Vahlteich Professorship (D.H.S.). This project was also supported in part by the University of Michigan Water Center, which is supported by the Erb Family Foundation and University of Michigan Provost. G.J.D. was supported as an Alfred P. Sloan Research Fellow. Support for M.M.S was provided by the NSF Graduate Research Fellowship Program (1256260). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

Bachmann, B.O., and Ravel, J. (2009) Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* **458:** 181–217.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013) antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* **41:** W204–W212.

Carballo, J.L., Naranjo, S., Kukurtzü, B., Calle, F., and Hernández Zanuy, A. (2000) Production of *Ecteinascidia turbinata* (Ascidiacea: Perophoridae) for obtaining anticancer compounds. *J World Aquacult Soc* **31:** 481–490.

Chu, H.Y., Wegel, E., and Osbourn, A. (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J* **66:** 66–79.

Ciccarelli, F.D. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283–1287.

Corey, E.J., Gin, D.Y., and Kania, R.S. (1996) Enantioselective total synthesis of ecteinascidin 743. *J Am Chem Soc* **118:** 9202–9203.

Cuevas, C., and Francesch, A. (2009) Development of Yondelis (trabectedin, ET-743). A semisynthetic process solves the supply problem. *Nat Prod Rep* **26:** 322–337.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10:** R85.

van Heel, A.J., de Jong, A., Montalbán-López, M., Kok, J., and Kuipers, O.P. (2013) BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* **41:** W448–W453.

Hiratsuka, T., Koketsu, K., Minami, A., Kaneko, S., Yamazaki, C., Watanabe, K., *et al*. (2013) Core assembly mechanism of quinocarcin/SF-1739: bimodular complex nonribosomal peptide synthetases for sequential mannich-type reactions. *Chem Biol* **20:** 1523–1535.

Kikuchi, Y. (2009) Endosymbiotic bacteria in insects: their diversity and culturability. *Microbes Environ* **24:** 195–204.

Konstantinidis, K.T., and Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187:** 6258–6264.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645.

Kuo, C.-H., and Ochman, H. (2009) Deletional bias across the three domains of life. *Genome Biol Evol* **1:** 145–152.

Kuo, C.-H., Moran, N.A., and Ochman, H. (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19:** 1450–1454.

Kwan, J.C., and Schmidt, E.W. (2013) Bacterial endosymbiosis in a chordate host: long-term co-evolution and conservation of secondary metabolism. *PLoS ONE* **8:** e80822.

Kwan, J.C., Donia, M.S., Han, A.W., Hirose, E., Haygood, M.G., and Schmidt, E.W. (2012) Genome streamlining and chemical defense in a coral reef symbiosis. *Proc Natl Acad Sci USA* **109:** 20655–20660.

Kwan, J.C., Tianero, M.D.B., Donia, M.S., Wyche, T.P., Bugni, T.S., and Schmidt, E.W. (2014) Host control of symbiont natural product chemistry in cryptic populations of the tunicate *Lissoclinum patella*. *PLoS ONE* **9:** e95850.

Lackner, G., Moebius, N., Partida-Martinez, L.P., Boland, S., and Hertweck, C. (2011) Evolution of an endofungal lifestyle: deductions from the Burkholderia *rhizoxinica* genome. *BMC Genomics* **12:** 210.

Lawrence, J.G., and Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143:** 1843–1860.

Lei, L., Deng, W., Song, J., Ding, W., Zhao, Q.-F., Peng, C., *et al*. (2008) Characterization of the saframycin A gene cluster from Streptomyces lavendulae NRRL 11002 revealing a nonribosomal peptide synthetase system for assembling the unusual tetrapeptidyl skeleton in an iterative manner. *J Bacteriol* **190:** 251–263.

Li, M.H., Ung, P.M.U., Zajkowski, J., Garneau-Tsodikova, S., and Sherman, D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics* **10:** 185.

Lichter, W., Lopez, D.M., Wellham, L., and Sigel, M.M. (1975) *Ecteinascidia turbinata* extracts inhibit DNA synthesis in lymphocytes after mitogenic stimulation by lectins. *Exp Biol Med* **150:** 547–568.

Lindquist, N., Hay, M.E., and Fenical, W. (1992) Defense of ascidians and their conspicuous larvae: adult vs. larval chemical defenses. *Ecol Monogr* **62:** 547–568.

McCutcheon, J.P., and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10:** 13–26.

McLaughlin, K. (2015) U.S. FDA grants priority review for YONDELIS® (trabectedin) for the treatment of patients with advan.

Malpartida, F., and Hopwood, D.A. (1984) Molecular cloning of the whole biosynthetic pathway of a Streptomyces

antibiotic and its expression in a heterologous host. *Nature* **309:** 462–464.
Markowitz, V.M., Chen, I.M.A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., *et al*. (2013) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* **42:** D568–D573.
Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., *et al*. (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* **42:** D560–D567.
Moran, N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci USA* **93:** 2873–2878.
Moran, N.A., and Munson, M.A. (1993) A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc R Soc Lond B* **253:** 167–171.
Moran, N.A., McCutcheon, J.P., and Nakabachi, A. (2008) Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* **42:** 165–190.
Moss, C., Green, D.H., Pérez, B., Velasco, A., and Henríquez, R. (2003) Intracellular bacteria associated with the ascidian *Ecteinascidia turbinata*: phylogenetic and in situ hybridisation analysis. *Mar Biol* **143:** 99–110.
Munoz-Elias, E.J., and McKinney, J.D. (2006) Carbon metabolism of intracellular bacteria. *Cell Microbiol* **8:** 10–22.
Nakabachi, A., Ueoka, R., Oshima, K., Teta, R., Mangoni, A., Gurgui, M., *et al*. (2013) Defensive bacteriome symbiont with a drastically reduced genome. *Curr Biol* **23:** 1478–1484.
Newman, D.J., and Cragg, G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* **75:** 311–335.
Omsland, A., Cockrell, D.C., Howe, D., Fischer, E.R., Virtaneva, K., Sturdevant, D.E., *et al*. (2009) Host cell-free growth of the Q fever bacterium Coxiella burnetii. *Proc Natl Acad Sci USA* **106:** 4430–4434.
Patel, M.S., Nemeria, N.S., Furey, W., and Jordan, F. (2014) The pyruvate dehydrogenase complexes: structure-based function and regulation. *J Biol Chem* **289:** 16615–16623.
Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J.M., *et al*. (2006) A small microbial genome: the end of a long symbiotic relationship? *Science* **314:** 312–313.
Pérez-Matos, A.E., Rosado, W., and Govind, N.S. (2007) Bacterial diversity associated with the Caribbean tunicate *Ecteinascidia turbinata*. *Antonie Van Leeuwenhoek* **92:** 155–164.
Peng, C., Pu, J.-Y., Song, L.-Q., Jian, X.-H., Tang, M.-C., and Tang, G.-L. (2012) Hijacking a hydroxyethyl unit from a central metabolic ketose into a nonribosomal peptide assembly line. *Proc Natl Acad Sci USA* **109:** 8540–8545.
Piel, J. (2009) Metabolites from symbiotic bacteria. *Nat Prod Rep* **26:** 338–362.
Pospiech, A., Cluzel, B., Bietenhader, J., and Schupp, T. (1995) A new *Myxococcus xanthus* gene cluster for the biosynthesis of the antibiotic saframycin Mx1 encoding a peptide synthetase. *Microbiology* **141:** 1793–1803.
Price, M.N., Huang, K.H., Arkin, A.P., and Alm, E.J. (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15:** 809–819.
Rath, C.M., Janto, B., Earl, J., Ahmed, A., Hu, F.Z., Hiller, L., *et al*. (2011) Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chem Biol* **6:** 1244–1256.
Rinehart, K.L., Holt, T.G., and Fregeau, N.L. (1990) Ecteinascidins 729, 743, 745, 759A, 759B, and 770: potent antitumor agents from the Caribbean tunicate *Ecteinascidia turbinata*. *J Org Chem* **55:** 4512–4515.
Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* **407:** 81–86.
Staley, J.T., and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39:** 321–346.
Stone, M.J., and Williams, D.H. (1992) On the evolution of functional secondary metabolites (natural products). *Mol Microbiol* **6:** 29–34.
Velasco, A., Acebo, P., Gomez, A., Schleissner, C., Rodríguez, P., Aparicio, T., *et al*. (2005) Molecular characterization of the safracin biosynthetic pathway from *Pseudomonas fluorescens* A2-2: designing new cytotoxic compounds. *Mol Microbiol* **56:** 144–154.
Walton, J.D. (2000) Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis. *Fungal Genet Biol* **30:** 167–171.
Weber, T., Rausch, C., Lopez, P., Hoof, I., and Gaykova, V. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol* **140:** 13–17.
Wernegreen, J.J. (2002) Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* **3:** 850–861.
Williams, D.H., Stone, M.J., Hauck, P.R., and Rahman, S.K. (1989) Why are secondary metabolites (natural-products) biosynthesized? *J Nat Prod* **52:** 1189–1208.
Wu, D., Daugherty, S.C., Van Aken, S.E., Pai, G.H., Watkins, K.L., Khouri, H., *et al*. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol* **4:** e188.
Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., *et al*. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12:** 635–645.
Young, C.M., and Bingham, B.L. (1987) Chemical defense and aposematic coloration in larvae of the ascidian *Ecteinascidia turbinata*. *Mar Biol* **96:** 539–544.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Origin of the four metagenomic DNA samples used to compile the consensus microbial genome of the ET-743 producer. The metagenomic DNA from zooids isolated from two separate tunicate colonies was sequenced. Bins corresponding to the ET-743 producer in each metagenomic dataset were combined to create a consensus genome.
Fig. S2. An Emergent Self Organized Map of one of the four metagenomic DNA samples (Metagenomic sample 4) from

Ecteinascidia turbinata. Each data point represents a 5-kb sequence window, generated computationally from assembled contigs. Green data points are from unidentified contigs putatively assigned to the eukaryotic host. Data points shown in purple are from the *Candidatus* E. frumentensis bin, with circled data points contigs containing the 16S rRNA gene for E. frumentensis and the ET-743 biosynthetic gene cluster. Those in red are from a cyanobacterium present in only two of the four samples (Metagenomic sample numbers 3 and 4). The background represents Euclidean distance of tetranucleotide frequencies between data points; grey and dark colours indicate larger distances, which are used to visualize the borders between genomic bins. Borders defined for *Ca.* E. frumentensis and the cyanobacterium are outlined in blue. Clustered data points in yellow and dark blue represent bins from unknown bacteria.

Fig. S3. The gene content of drastically reduced genomes. Shaded boxes represent the presence of a gene in the genome while white boxes represent its absence. The minimal gene content of *Ca*. E. frumentensis more closely resembles the reduced obligate symbiont genomes of *B. aphidicola* (NC_011834) and *W. glossinidia* (CP003315) than the intracellular pathogen *C. Burnetii* (NC_011528) or the free-living microorganisms *F. hongkongensis* (GCA_000379445.1) and *E. coli* (NC 000913).

Fig. S4. Phylogenetic reconstruction using conserved markers. Reconstruction of (A) 16S rRNA gene, (B) RpoB and (C) RecA first suggested that *Ca*. E. frumentensis might have a novel taxonomic rank higher than the species level. Genes analysed in this study for *Ca.* E. frumentensis are depicted in bold. The sequence identity between listed genes and the corresponding gene in *Ca*. E. frumentensis is also included. Branch labels on the bootstrapped trees represent consensus support (%).

Fig. S5. Relatedness between the 16S rRNA gene and the average amino acid identity for *Ca*. E. frumentensis and similar microorganisms. Previously described taxonomic thresholds for phylum, class, order, family, genus (Yarza *et al*., 2014) and species are shown in dotted lines. Genomes were selected based on phylogenetic analysis of conserved genes and the taxonomic profile of the bin of the ET-743 producer. Genomes for the listed Pseudomonas *fluorescens*, *Streptomyces lavendulae*, and Myxococcus *xanthus* strains were included because similar strains are associated with the tetrahydroisoquinoline natural products safracin, saframycin A, and saframycin Mx1 respectively.

Fig. S6. A proposed updated pathway for ET-743 biosynthesis. A condensed pathway depicting only new biochemical steps is also presented in Fig. 4 (main paper).

Table S1. An overview of the four metagenomic DNA sequence datasets isolated from *Ecteinascidia turbinata*. A single bin containing the ET-743 biosynthetic gene cluster and the 16S rRNA gene for *Ca.* E. frumentensis was present in every metagenomic sample. Samples 3 and 4 also contained a bin with an rRNA marker for an *Oscillatoriales* species. Sample 4 contained an additional prokaryotic bin from an unknown microorganism.

Table S2. Coverage comparison between the complete *Ca*. E. frumentensis genome and the discarded contig. The consistently lower coverage for the ~18 kb contig that binned with the ET-743 producer caused us to exclude it from our analysis.

Table S3. An overview of the differences in genomes of endosymbionts, intracellular pathogens and free-living microorganisms. The features of the complete genome of *Ca.* E. frumentensis correspond with those of obligate endosymbionts.

Table S4. Pseudogenes identified in the noncoding regions of the *Ca*. E. frumentensis genome.

Table S5. New genes proposed to be involved in ET-743 biosynthesis in *Ca*. E. frumentensis.