7-20-2016

Intein Epidemiology: A Study of the Lifestyle, Distribution, Phylogenetics, and Dynamics of Inteins

Shannon M. Soucy
shannon.soucy@uconn.edu

Intein Epidemiology:
A Study of the Lifestyle, Distribution, Phylogenetics, and Dynamics of Inteins

Shannon Margaret Soucy, PhD
University of Connecticut, 2016

Horizontal gene transfer (HGT) is an important process experienced by all cellular life on some level to enable rapid adaptation to changes in both the extracellular and intracellular environments. Quantifying the extent of HGT, as well as identifying conditions that both foster and deter HGT within both natural and artificial environments. As a solution we propose using mobile genetic elements to track the interactions between organisms, as well as the evolution of these mobile genetic elements to learn about HGT within various cellular communities. We use inteins to identify networks of HGT events within various collections of organisms that share a specific characteristic. Inteins are found in highly conserved proteins that are often part of the replication or recombination machinery. Though they are associated with highly conserved proteins in very slowly evolving sites (ATP or DNA binding) inteins have a surprisingly high substitution rate. The homing cycle, the canonical model for intein spread in populations, posits that this substitution rate can eventually lead to loss of the homing endonuclease domain generating a mini-intein. Mini-inteins are incapable of spreading into un-invaded exteins, and are passed down through vertical evolution only. Thus we can use the presence of the homing endonuclease domain as an indication of how the intein was acquired by each organism.

We started by using a group of organisms that have an intein allele in common, then progressed to a group of organisms that are phylogenetically related, and finally used organisms that were isolated from the same environment to investigate networks of gene transfer and intein dynamics within these communities. We learned inteins are most often transferred between closely related organisms, especially those found within the same environments. We also show

that inteins promote recombination during mating. We find that all three states of intein invasion (full-size, mini, and un-invaded) can co-exist within the same environment, debunking the homing cycle model, at least in prokaryotic communities.

Intein Epidemiology:
A Study of the Lifestyle, Distribution, Phylogenetics, and Dynamics of Inteins

Shannon Margaret Soucy

B.S., Central Connecticut State University, 2008

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Intein Epidemiology:
A Study of the Lifestyle, Distribution, Phylogenetics, and Dynamics of Inteins

Presented by
Shannon Margaret Soucy, B.S.

Major Advisor

_____

J. Peter Gogarten

Associate Advisor

_____

R. Thane Papke

Associate Advisor

_____

Dan Gage

University of Connecticut 2016

Acknowledgements

"I've taken my bows,
And my curtain calls,
You brought me fame and fortune and everything that goes with it,
I thank you all"
-Freddie Mercury, We are the Champions

I would like to acknowledge my major advisor for his unfailing optimism throughout my time in his lab. I am a born worrier, Peter really made sure to constantly remind me to make sure I was having fun. I would also like to thank all of my lab mates for their helpful insight and discussion pertaining to research. I cannot imagine what my time here would have been like without friends like Matthew Fullmer, Thiberio Rangel, Nikhil Ram Mohan, Kunica Asija, Israela Turgemen, Neta Altman-Price, and too many more to list.

I also want to thank my committee members Dr. R. Thane Papke, Dr. Joerg Graf, Dr. Jonathan Klassen, and Dr. Dan Gage, as well as Dr. Spencer Nyholm who agreed to stand in last minute in case there was a scheduling conflict. I have had thoughtful discussions with each of you and have grown so much in my time here in a large part because of your wisdom.

I have to thank my family for their never ending support, I cannot imagine where I would be without such strength and determination behind me in the form of my parents, brothers, sisters, brother-in-laws, sister-in-laws, nieces, and nephew. From crayon drawings of encouragement from the kids, cheers and rallying cries from my sisters, my youngest brother's interest in science and evolution, my oldest brothers inspiring chats reassuring me I am capable of more than I think, and the feeling of absolute acceptance and pride from my parents and step-parents, no matter what direction I choose.

I cannot forget to thank the people who made ALL the coffee I drank in the last six years, especially Sheri and Christine at Dunkin Donuts, who remember my order even when I was too exhausted to remember my name. I always say there is no one that has more power to make or break my day than the person making my coffee.

Last but not least, Fred my partner in life who supports me in all things, and is the best souvenir I'm taking with me from UConn.

Table of Contents

Chapter 1. General Introduction and Chapters Overview.

Horizontal gene transfer (HGT) is an important force in the evolution of organisms and communities. HGT is enables the assembly of innovative pathways like oxygenic photosynthesis(1), and even the emergence of an entirely new family of organisms, the *Halobacteria*(2, 3). However, more still needs to be done to describe the boundaries and timeline of this process. Most gene exchange occurs between closely related organisms(4), yet there are many examples of genes that have been transferred between domains. Selfish mobile genetic elements (MGEs) in particular frequently experience HGT (transposons, integrases, inteins, and plasmids), by using phylogenetics, as well as the distribution of MGEs we can build networks of gene transfer among groups of organisms. Furthermore because MGEs are so dynamic, we can use them to generate profiles that may help to unravel relationships between closely related organisms not distinguishable through reference gene phylogenies.

MGEs generally impose a fitness cost on their host organisms. Many selfish elements balance the fitness cost they impose by associating with (or capturing) genes that provide useful functions to their host. Antibiotic resistance genes in particular are often found in association with compound MGEs, for example plasmids and integrative and conjugative elements (ICEs) (5). The majority of the genetic material associated with these mobile elements imposes a fitness cost on the host species; however, the potential of even one particularly useful phenotype can cause selection to favor lineages that carry these selfish elements. Studying patterns of HGT using these elements is difficult because they often loose and gain genes as they evolve. Also their distribution is likely to be biased by the traits they associate with, for example a plasmid

that carries antibiotic resistance genes will likely be transferred amongst networks that experience selective pressure through exposure to antibiotics frequently. For these reasons gene exchange networks based on compound mobile element dynamics would not be suitable for investigating the barriers of HGT.

Some truly selfish MGEs are much smaller than the compound mobile elements like transposons, group II introns, and inteins. These are still multi-domain proteins or multifunctional RNAs, but they are much simpler than the compound elements. Group II introns are fairly sparse within prokaryotic genomes, and thus would provide a limited view of HGT networks. Transposons, though ubiquitous are difficult to use for a network of HGT because they propagate within genomes as well as between genomes (6) and are found sporadically among the genome, thus the fitness cost of each insertion is variable. Inteins, however, are an ideal tool to build networks of HGT.

Inteins are selfish MGEs, and can be found in all three domains of life and viruses. Inteins are made up of two independent domains, the splicing domains flank the N and C-termini of the intein, and the homing endonuclease (HEN) domain is in the middle. The splicing domains excise the intein from the host extein after translation, and the homing endonuclease enables horizontal propagation of the intein upon encountering an un-invaded extein sequence. Inteins are unique MGEs in that inteins form alleles, a cohesive group of inteins that reside in the same location in the same protein (extein), and the insertion sites for inteins are usually in slowly evolving sites in highly conserved proteins (7). The consistency of the insertion site normalizes the fitness cost of intein invasion, at least compared to the cost of transposon insertion. Intein insertion sites (IIS) are often involved in ATP or DNA binding and most extein functions are those related to replication, recombination, repair, and nucleotide metabolism (8). The

conservation of these proteins across all domains of life ensures that most organisms that the intein encounters will have a highly similar IIS. Furthermore, the importance of the ATP and DNA binding sites in the exteins ensures that frequent substitutions at this site that could enable the extein to escape intein invasion would likely also render the protein non-functional.

Inteins invasion occurs when a new host is encountered by an invaded extein. Briefly, the homing endonuclease domain recognizes the uninterrupted insertion site and makes a double strand break. The double strand break recruits the repair machinery of the host protein, which then uses the intein containing copy of the extein as a template for repair. Interestingly, inteins do not have a mechanism to penetrate the cell wall, and thus they must rely on mechanisms of HGT that are already occurring within the community to spread to a new host.

Another unique characteristic of inteins as MGEs is the modularity, inteins are found with and without the HEN domain intact. When the HEN domain is missing the intein is called a mini-intein, as these inteins tend to be a couple of hundred amino acids shorter than their full-size counterparts. Mini-inteins cannot invade new exteins through homing, thus their presence indicates that the intein was most likely acquired through vertical gene transfer. The canonical model of intein invasion, the homing cycle, posits that the HEN domain is lost because of reduced selective pressure for its function. In other words, there were not enough opportunities for homing to occur and thus over time substitutions occurred that lead to a loss of function. The absence of a functional HEN domain can be used to find inteins that were passed on by ancestors, and those that might have been newly acquired.

Like most MGEs inteins have a much higher substitution rate relative to the extein protein they reside in. Additionally these substitution events seem to be focused on the HEN domain of the inteins. This could be because substitution events in the splicing domains could

disrupt the splicing function and improper splicing can lead to loss of function in the host protein. Substitutions in the intein splicing domain that disrupt their functions would most likely kill the cell, as most exteins perform functions related to important processes like replication, recombination, and repair.

This work uses intein epidemiology to examine HGT interaction networks between many different organisms, with a strong focus on the haloarchaea. Several distinctive traits make inteins ideal candidates to create interaction networks between organisms. The conservation of IIS ensures that transfers between both closely related and distantly related organisms will be included. The patchy distribution of inteins within closely related organisms indicates that HGT is an important mode of transmission for these sequences. The high substitution rates provide more phylogenetically informative sites to build phylogenies with. The modular structure of the intein sequence in addition to the plethora of phylogenetically informative sites enables us to infer the most likely mode of transmission for each intein sequence. Lastly the dependence on mechanisms of HGT already in place in the community ensures that the epidemiology of inteins can reveal interaction networks within clusters of organisms analyzed together.

Chapter 2 is a collection of several review articles I was involved in writing that investigate the impact of HGT on innovation and complexity in cellular life, both prokaryotic and eukaryotic. The articles each explore the impact of reticulate processes like HGT. The first article focuses on reticulate processes creating innovative pathways, or extending pathways that already exist. The article also explores HGT as a neutral process involving genes like inteins that do not appear to improve their host's fitness. The second article focuses on the specific consequences of HGT in the human holobiont. Humans rely on their associated microbiome for functions ranging from immunity to metabolism. Most HGT occurring in association with the

human holobiont is between organisms that share a body site, and affected pathways are usually related to metabolism. The third article explores the levels of selection that act on gene content within an organism or community. The last article discusses the implications of gene sharing on a community of prokaryotic organisms. The themes of these articles largely overlap, and the major message is that HGT is an important process to ensure that a community is able to respond to perturbations in the environment.

The rest of this dissertation focuses on intein epidemiology to determine interaction networks between organisms. Chapter 3 is published work, where we compare the phylogeny of the intein and extein sequences of all organisms known to have the *vma*1-b intein (in 2011). We find that most inteins are shared among closely related organisms, but one bacterium shares this intein with a diverse group of archaea, indicating that inteins can be transferred across long phylogenetic distances. The rest of this work uses intein epidemiology in the Halobacteria, colloquially referred to as the haloarchaea to make networks and observations on the interactions between various groups of organisms. Chapter 4 examines interaction networks among a diverse group of 118 haloarchaea from several different families and many different geographic locations. All intein sequences were concatenated along with a presence/absence matrix in order to look for highways of gene sharing within the haloarchaea. The next chapter (chapter 5) expands on this work but rather than considering all intein alleles together; each intein allele is broken down into pairwise comparisons among all members. These distances are compared to an expected distance based on a reference set of pairwise distances to identify gene transfer events within each allele.

Chapter 6 is work that was recently re-submitted to PNAS after one round of reviews. In collaboration with the Gophna lab at Tel Aviv University we use bench work, environmental

sequences, and simulations of intein invasions to show that inteins increase the frequency of recombination during mating, compared to isogenic strains under the same conditions. Also we propose an equation that predicts the intein invasion rate given several measureable parameters. Importantly this equation can be used with intein distribution information from environmental samples to try and predict the same parameters in a given environment.

In chapter 7 I characterized the intein population in Deep Lake, Antarctica using metagenomic reads along with four completed genomes from the same location (9). Using metagenomic data allowed us to consider each species as a population, and the raw reads enabled us to look for sequence variation in the inteins in each species' population.

In chapter 8 I describe the intein distribution within a collection of environmental isolates of haloarchaea from two locations along the Mediterranean coastline. This dataset contains eight newly described species, and several clusters of highly related organisms that can be distinguished by their intein distribution. This work also shows that many different versions of an intein allele can exist within a single environment.

Lastly chapter 9 revisits bench work analysis with hybrid strains of *Haloferax volcanii*/ *Haloferax mediterranei* generated through mating experiments with subsequent selection. This work is in good agreement with the environmental data in that it illustrates that some strains are more prone to intein invasion than others. Interestingly all strains that had a *Hfx. mediterranei* background acquired the *polB*-c intein, and all except one strain (75IS) maintained their inteins in positions *cdc21*-a, *cdc21*-b, and *pol-II*-a. In contrast *Hfx. volcanii* background strains did not gain any new inteins from *Hfx. mediterranei*.

The work described in this thesis resulted in the reporting of 376 new haloarchaeal intein sequences, and made use of intein distributions and phylogenies, as well as phylogenetically

independent methods to build networks of gene transfer. Additionally this data was used to distinguish closely related environmental isolates, and learn more about the life cycle of these selfish genetic elements.

Chapter 2. Innovation and Complexity through Horizontal Gene Transfer, across the Web of Life.

This chapter consists of four publications. The first three were invited reviews and the last is a peer-reviewed opinion piece. The first paper (10) is a review of HGT events that lead to innovative pathways, enhanced existing pathways, or enabled niche expansion for the recipient. This article was primarily written by Dr. Kristen Swithers, though I was the main author for section 5, "Parasitic HGTs can lead to Innovation and Complexity". Peter Gogarten supervised the research and writing.

I was the primary author for the second publication (11). This review is concerned mainly with HGT events that have affected humans. The paper asserts that genome dynamics of humans should consider the gene content of associated microbes as well as the human genome. The complexity of the human body plan and sequestration of the germ cells from somatic cells makes passing on changes in gene content difficult for humans. However, studies have shown that adaptations in the microbiome can respond very quickly to environmental changes, thus human genomes should be considered only in the context of the holobiont - the human genome, plus the gene content of associated microbes. This work was done in collaboration with Dr. Lorraine Olendzenski, and was supervised by Peter Gogarten.

I was also the primary author on the third publication in this chapter, titled "Horizontal Gene Transfer: Building the Web of Life". This paper explores examples of gene transfers that occur between prokaryotes, prokaryotes and eukaryotes, and even between multicellular eukaryotes. There are many examples of HGT events that have led to important changes in body plan construction, immune cell development, and metabolism. This paper was written in collaboration with Dr. Jinling Huang, who wrote most of the sections of HGT to plants, and supervised by Peter Gogarten.

The last paper included in this chapter is a peer-reviewed opinion piece that was written primarily by Matthew S. Fullmer and Peter Gogarten. I contributed to the development of the ideas in this paper, and helped with the editing. This paper proposes a role for both HGT and gene loss in environments as part of a balance that maintains the number of cooperating cells and limits the number of cheaters that a community can sustain. Gene content will mosaic especially considering genes whose products are common goods that are exported into the extracellular environment. Also the amount of mosaicism within the genomes can be a measure of cooperation within the community. Conversely if there is not mosaicism with respect to gene content in an environment this indicates that social cheaters are not tolerated in the environment.

2.1 The Role of Reticulate Evolution in Creating Innovation and Complexity.

Hindawi Publishing Corporation International
Journal of Evolutionary Biology Volume 2012,
Article ID 418964, 10 pages
doi:10.1155/2012/418964

Review Article

The Role of Reticulate Evolution in Creating Innovation and Complexity

Kristen S. Swithers, Shannon M. Soucy, and J. Peter Gogarten

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3125, USA

Correspondence should be addressed to J. Peter Gogarten, gogarten@uconn.edu Received

3 February 2012; Revised 8 May 2012; Accepted 10 May 2012 Academic Editor:

Wen Wang

Reticulate evolution encompasses processes that conflict with traditional Tree of Life efforts. These processes, horizontal gene transfer (HGT), gene and whole-genome duplications through allopolyploidization, are some of the main driving forces for gene- rating innovation and complexity. HGT has a profound impact on prokaryotic and eukaryotic evolution. HGTs can lead to the invention of new metabolic pathways and the expansion and enhancement of previously existing pathways. It allows for organismal adaptation into new ecological niches and new host ranges. Although many HGTs appear to be selected for because they provide some benefit to their recipient lineage, other HGTs may be maintained by chance through random genetic drift. Moreover, some HGTs that may initially seem parasitic in nature can cause complexity to arise through pathways of neutral evolution. Another mechanism for generating innovation and complexity, occurring more frequently in eukaryotes than in prokaryotes, is gene and genome duplications, which often occur through allopolyploidizations. We discuss how these different evolutionary processes con- tribute to generating innovation and complexity.

1. Introduction

Reconstruction of the Tree of Life attempts to represent the organismal histories of all of life on earth on a single bifur- cating tree. Since the dawn of the molecular age, and, more so recently, with the numerous whole-genome sequences that are now available, it has become apparent that reticulate evo- lutionary processes such as horizontal gene transfer (HGT), genome fusion, and incomplete lineage sorting have a pro- found impact on microbial and eukaryotic evolution. These processes dissolve or embed the lines of vertical descent that are a hallmark of the tree of life into net-like relationships between genomes and organisms. To more accurately des- cribe the complexity of organismal histories many groups have proposed net-like reconstructions of life's history [1] to account for the lines of vertical descent and lateral lines created from reticulate processes; the "rooted net of life" [2], the "forest of life" [3, 4], and the "rhizome of life" [5, 6] are a few examples.

HGT is the nonvertical transmission of genetic mate- rial, that is, the exchange of genetic information between organisms not in an ancestor descendant relationship. HGT causes individual genes in a genome to have vastly different evolutionary histories. Studies show HGT occurs more frequently between closely related organisms than in divergent organisms [7, 8]. Closely related organisms tend to have similar sequences and intracellular environments. These similarities allow for more opportunity for homologous recombination and for an easier integration of the transferred gene into the metabolic and regulatory networks of the recipient. However, there are increasing examples of HGTs between divergent species, even across domain boundaries, revealing that barriers to HGT can occasionally be overcome. Examples include the highways of HGTs [9] that exist bet- ween divergent organisms: members of the Thermotogae phylum share about half of their genes with both the Fir- micutes and the Archaea [10], and the Aquificae share many genes with the Epsilonproteobacteria [11]. Many of these successful HGTs allow for innovations in metabolism and body plan that provide a selective advantage to the organisms involved and allow expansion into new ecological niches.

Table 1: Categories of HGTs leading to innovation and complexity.

Type	"Beneficial" HGTs	"Neutral" HGTs	"Parasitic" HGTs
Definition	HGTs that provide an initial selective advantage to the recipient	HGTs are maintained by random genetic drift	HGTs do not provide an initial selective advantage to the recipient but over time may adapt to have a beneficial function or be maintained via pathways to neutral complexity in the recipient
Examples	(i) Metabolic pathway expansion and invention (ii) Adaptation to new ecological niches	(i) Many ORFan genes and genes of limited distribution and with unknown function may be in this category [14, 15]	(i) Inteins (ii) Group I Introns (iii) Group II Introns

Transferred genes can be distinguished based on their long- and short-term impact on the fitness of the recipient (Table 1). Genes that provide an adaptation create a selective advantage for the recipient and have a higher chance to persist over longer periods of time. As their frequency in the population increases over time these genes will become fixed. Examples of these "beneficial" HGTs are those that allow the recipient to expand into a previously empty ecological niche. These provide a huge increase in fitness to the recipient, even if the transferred gene has not yet adapted perfectly to the genomic and regulatory environment of the recipient [12]. Many of the genes that extend, enhance, or create new metabolic pathways fall into this category. These genes may be selfish in Dawkins' [13] original definition, but they cooperate with the other genes in the organism's genome and provide a selective advantage for the organism.

Many other, and possibly most, transferred genes that can be identified in the pan-genome [16] of bacterial or archaeal populations may be selectively neutral or nearly neutral to their carriers [14]. Many of these genes will be lost after a few generations; however, a few may be fixed through ran- dom genetic drift. It could be argued that most of the endo- symbiant *Wolbachia* to host transfers are selectively neutral or nearly neutral. Almost all of the *Wolbachia* genes are found in the host genome and their transcript levels are very low [17]. This low transcript level may indicate that these genes do not provide a function to the host and supports the notion that many genes transferred from the symbiont are only transiently present in the host nuclear genome. Although the majority of these transferred genes are transcribed at very low level, two hypothetical proteins in the *Aedyes aegypti* originating from *Wolbachia* have been maintained in the nuclear genome for a long period of time and are transcribed at higher levels than background suggesting these genes were fixed in the population [18].

Some transferred genes initially are like infections in that their survival and spread is through a mechanism that decouples the genes propagation from host replication and host fitness. Although the propagation of these selfish genetic elements is decoupled from the host's genetic machinery, the element does utilize the host's resources to propagate through a population. In this sense these genetic elements can be considered parasitic. To more clearly distinguish them from the selfish gene concept in Dawkins' gene-centered view of evolution, which considers all genes as selfish, we term these elements as parasitic genetic elements and their trans- fers "parasitic HGTs"; examples include inteins and self- splicing introns. Initially, a self-splicing molecular parasite may provide little or no advantage to the host but may later adapt a function to benefit the host. Many inteins and group I introns contain a homing endonuclease (HE) that provides mobility to the element and allows them to follow a life cycle known as the homing cycle [19]. Briefly, the homing cycle begins when an allele with an HE is horizontally trans- ferred to a recipient in a new population or species that before the invasion harbored only alleles without HE [20]. Through faster than Mendalian inheritance the HE contain- ing parasite spreads through the population, leaving little or no detrimental effects on the host. However, once all the members of the population have the HE containing element the HE containing genetic element starts to degrade. To escape this cycle, over time the parasites may adapt to provide a beneficial function for the host [7] or are maintained through neutral pathways to complexity as discussed below for the case of the *dnaE* intein [21, 22].

Transferred genes can be integrated into the recipient genome by homologous recombination or through illegiti- mate recombination [23]. The former process requires stret- ches of similar sequences; however, the stringency of this requirement depends on the activity of the mismatch repair system [24]. The similarities necessary for homologous recombination can be due to the presence of a homolog in the recipient genome or can be created through transposable elements present in the recipient that jump into the trans- ferred extrachromosomal genetic material [25]. Transferred DNA also can be integrated independent of sequence simi- larity through double-strand break repair pathways, such as nonhomologous end joining, allowing for the integration of DNA from divergent organisms [7]. Transposable elements can also facilitate transfer and integration into recipient DNA. One such example is the integrative and conjugative elements (ICEs). ICEs have been implicated in transfer of genes involved in antibiotic and heavy metal resistance, nitrogen fixation, virulence, biofilm formation, and the deg- radation of aromatic compounds (for reviews see [26, 27] and references therein), providing another example for mul- tiple levels of selection, in this case benefiting both the trans- ferred genes and the recipient.

Although HGT appears to be more prevalent in prokary- otes, more and more examples of HGT are being documen- ted in single-celled and even multicellular eukaryotes (see [28] and below for examples of transfer from bacteria to eukaryotes). Related driving forces in creating innovation and complexity in eukaryotic lineages are gene and whole genome duplications. Genome fusion resulting from hybri- dization between members of related species, a frequent pathway towards polyploidization, is akin to HGT in that it results in mosaic genomes and that the resulting gene family expansion is due to reticulate evolution. Observed in plants [29], animals [30, 31], and fungi [32, 33] whole-genome duplication followed by neofunctionalization and/or sub- functionalizations have been implicated in providing the building blocks for more complex developmental and meta- bolic pathways.

Gene, genome duplication, and HGT, regardless of the type of selection, beneficial, neutral, or parasitic, are all reti- culate processes that affect evolution across all domains of life. Here we explore how the process of HGTs can expand metabolic pathways, allow for microorganisms to adapt to new host ranges, expand environmental niches, and even influence multicellular eukaryotes. We also explore how "parasitic HGTs" can ultimately lead to innovation and increased complexity. Additionally, we discuss how gene and whole-genome duplications can give rise to novel pathways that are important for development.

2. HGT and Expansion Metabolic Pathways

HGTs can lead to the enhancement, expansion, and con- struction of more complex metabolic pathways. About two- thirds of the annual biogenic methane is produced from the acetoclastic methanogenesis pathway, which is exclusively carried out by the methanogenic eurarchaeal order Metha- nosarcinales [34]. Most members of this group carry out the conversion of acetate to acetyl-coenzyme A using the acetyl- CoA synthesis pathway. However, members of the more widely distributed *Methanosarcina* use a variation on this pathway, which uses the enzymes acetate kinase (AckA) and phosphoacetyl treasferase (Pta) [34]. Both the *ackA* and *pta* genes were shown through multiple phylogenetic methods to be transferred in one event from the cellulolytic clostridia, where the encoded enzymes are used to produce acetate as a product of fermentation, to *Methanosarcina* [35], where the same enzymes are used to produce acetyl-CoA.

Another example of an expanded pathway created by HGT is found in the Thermotogae phylum. Some of the lower-temperature lineages are able to produce vitamin B_{12} using the cobinamide salvage pathway [36] (Figure 2). In this pathway a partial B_{12} molecule is scavenged from the envi- ronment and subsequently modified to produce an active B_{12} molecule. This method of B_{12} production was shown to be the ancestral pathway for the Thermotogae lineage by presence and absence of the genes in the phylum (Figure 2). A later HGT allowed the *Thermosipho* genus to synthesize B_{12} *de novo* from glutamate, through transfer of twenty-one genes from the Firmicutes.

An enhancement of a pathway is observed in HGT events between eukaryotic species of grasses. Some members of the *Alloteropsis* grasses have acquired highly functional genes for C_4 photosynthesis from the Cenchrinae and Melinidinae: phosphoenolpyruvate carboxylases (ppc) were likely trans- ferred from both the Cenchrinae and Melinidinae, and phosphoenolpyruvate carboxykinase (pck) was transferred from the Cenchrinae. Christin et al. hypothesize that before the arrival of these genes the *Alloteropsis* may have had a subfunctional C_4 CO_2-fixation pathway, as in the case of the extant *A. semialata* subsp. *semialata* grass, which did not receive these HGTs. This enhancement of the C_4 pathways allows for adaptation of the grass to warm and arid climates [37].

The metabolic pathways expanded and enhanced through HGT allow for an occupation of a new ecological niche. The *Thermosipho* can now produce B_{12} and thrive in an environment where no partial B_{12} derivatives are present, while members of the genus *Methanosarcina* are able to produce most of the world's methane from acetate and the *Alloteropsis* grasses can thrive in warm and arid climates.

3. HGT and Metabolic Innovations

Members of at least six different bacterial phyla use chloro- phyll-based photosynthesis to gain energy from light [38, 39]. Comparative phylogenetic analysis revealed that hori- zontal gene transfer played an important role in evolution and distribution of bacterial photosynthesis [40, 41]. The assembly of the electron transport chain that allows the use of water as electron donor likely represents the gene trans- fer event that most changed Earth's biosphere [42, 43]. Chloroflexi (green filamentous bacteria) and purple bacteria possess a photosynthetic reaction center similar to photo- system II of the cyanobacteria; whereas the reaction centers in Chlorobi (green sulfur bacteria) and Heliobacteria (Fir- micutes) are similar to the photosynthetic reaction center I in cyanobacteria [39, 44]. However, in the cyanobacteria photosystem I and photosystem II are present, and only when the two divergent types of reaction centers work in series do the harvested photons provide sufficient energy to lift electrons over the electrochemical potential difference between water and NADP. It is theoretically possible that photosystems I and II arose through a within-lineage gene duplication, diverged within the cyanobacteria, and subse- quently individual photosystems were transferred to other bacteria. A more likely scenario is that the two photosystems diverged from an ancestral photosystem in diverging lineages (Figure 1(b)), which each used a single photosystem, and that the two distinct photosystems were brought together in the cyanobacterial ancestor through HGT.

The recently described methylaspartate cycle in Haloar- chaea [45] provides another example for the creative power of HGT. This cycle provides an alternative to the glyoxylate cycle and the ethylmalonyl-CoA pathway for acetyl CoA to enter central carbon metabolism to synthesize cellular build- ing blocks. According to analyses reported in [45] the key enzymes of the methylaspartate cycle were acquired by

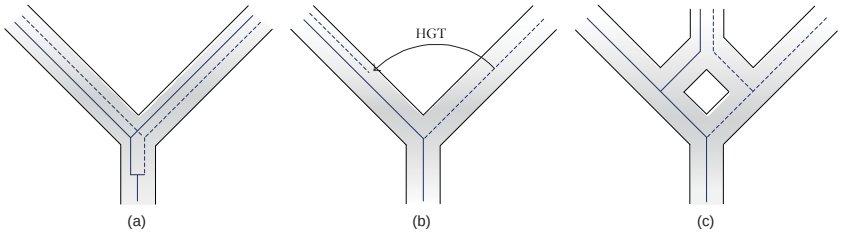


Figure 1: Types of genetic duplications. (a) Shows an autochthonous duplication, which can happen either through tandem duplication, segmental duplication, chromosomal duplication, genome duplications, or retro-transposition. (b) Shows gene family expansion through HGT. Following the divergence of two lineages orthologous genes diverge in sequence and possibly in function. These orthologs can be brought together in a single genome through HGT or allopolyploidization (c). The scenarios depicted in (c) and (b) explain an apparent duplication through reticulated evolution.

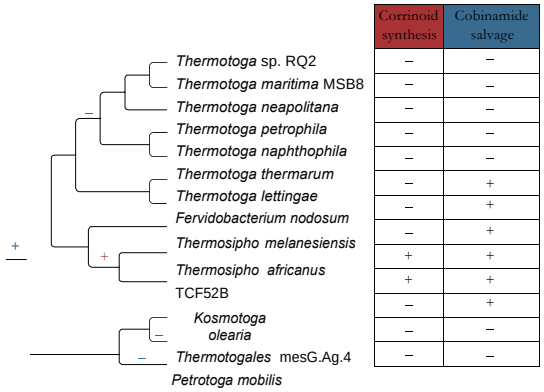


Figure 2: Distribution of the two gene clusters involved in vitamin B_{12} biosynthesis among the Thermotogae phylum. The corrinoid synthesis gene cluster contains genes for the first part of the *de novo* B_{12} synthesis pathway and the cobinamide salvage gene cluster contains genes that synthesize vitamin B_{12} from cobinamides, incomplete B_{12} molecules. Together these two gene clusters complete the *de novo* B_{12} biosynthesis pathway. Presence of a gene cluster is denoted by (+) and absence is denoted by (–). The most parsimonious explanation for the extant presence/absence patterning for the cobinamide salvage gene cluster is one gain at the root of the phylum and three losses marked by blue and (+) and (–) and for the corrinoid synthesis gene cluster one gain marked by a red (+). This suggests the cobinamide salvage pathway was present in the ancestor of the Thermotogae phylum and the genes for complete *de novo* synthesis were gained in a later event by the *Thermosipho* lineage.

the Haloarchaea through gene transfer from different bac- teria. Furthermore, before the transfer, these enzymes were part of different pathways in the donor organisms, such as propionate assimilation or glutamate fermentation. The methylaspartate cycle thus represents a metabolic patchwork of enzymes acquired from different donors and combining fragments of different pathways into a novel enzymatic cycle.

4. HGT and Innovations in Communities

The human microbiome provides an opportunity to under- stand a complex community of microorganisms and how HGT has facilitated innovation within a large community of microorganisms. Many traits, such as antibiotic resistance, and xenobiotic metabolism observed in the human gut microbiota are a consequence of HGT. One study showed that antibiotic resistance genes can be transferred to the gastrointestinal microbiome from food sources [46]. Volun- teers were fed chicken, which had a strain of vancomycin- resistant *Enterococcus faecium*, and vancomycin resistance was transferred to *E. faecium* in the human gut. Other studies in Japanese individuals showed that genes for porphyranases, agarases, and alginases, which facilitate the breakdown of red and brown algae (seaweed) in the human gut, were likely transferred from marine bacteria to Japanese gut symbiont

Bacteroidetes [47, 48]. These HGTs not only allow the gut bacteria to utilize seaweeds as a novel carbon source, but confer secondary benefits to the human host, which can now utilize seaweed as a nutrient source. The act of introducing foreign material to the gut microbiota (consuming a food source) facilitates interactions between the microbiome and the microorganisms on that food source. This interaction encourages possible HGTs from microorganisms outside the gut and allows for constant innovation and evolution of our microbiome to cope with the frequent changes in the gut environment, reinforcing the "you are what you eat" saying. These findings also confirm that the holobiont (host plus symbiont) can evolve and gain new adaptations without changes in the host's genome, simply by acquiring new symbionts with novel metabolic capabilities [49].

5. "Parasitic HGTs" Can Lead to Innovation and Complexity

"Parasitic HGT" involving molecular parasites, such as inteins and group I introns, are HGTs that confer no imme- diate selective advantage to the host but over time adapt to benefit the host. These inteins and group I introns are self-splicing genetic elements that are made mobile by hom- ing endonucleases, an endonuclease that recognizes target sequences of 12–40 bps [50]. They can evade purifying selec- tion on the organismal level as they cause little or no harm to their host [51]. These HE containing parasites have their own life cycle described by the homing cycle [20, 50, 52]. A possible escape route from this cycle presents itself, if the HE or the intein/intron evolves a beneficial function in the host. One such example of this is found in the mating type switching HO endonuclease in yeast [53]. This endonuclease is left over from what once was a close relative to the large intein in the yeast vacuolar ATPase catalytic subunit, but now facilitates genetic recombination from one mating type to another. This innovation is beneficial to the organism in that it expands the reproductive capabilities of the yeast cell. Another example where an intein may have been retained and adapted to benefit its host is found in bacterial intein-like (BIL) domains. These are degenerated remnants of the HINT domain intein family, which are now thought to function to facilitate rearrangements in hypervariable surface proteins [54, 55]. Over time the HEs of some group I introns are maintained as functional maturases to aid in the folding and splicing of the intron they reside in or other introns that may have lost their self-splicing ability [21, 56]. In these cases parasitic HGTs have facilitated beneficial innovations; however, most of these innovations evolved after a long period of neutral or nearly neutral association between the parasite and host.

Although many "parasitic HGTs" eventually provide some benefit for the host, there are other cases where they are maintained via selectively neutral pathways, which also can lead to higher complexity. The *dnaE* gene, of some cyano- bacterial species, is split on two parts (*dnaE1* and *dnaE2*), and each portion has part of an N-terminal or C-terminal intein [57]. An autocatalytic mechanism allows the split inteins to find each other after translation and splice the split protein together, resulting in a functional DNA polymerase III. Deletion or mutation of the intein portions of the split gene results in a nonfunctional DNA polymerase III, a major selective disadvantage for the organism and even possibly detrimental. This intein likely never supplied a selective advantage for the host. Through a series of intermediate steps, each of them neutral or nearly neutral to the organism, a complex processing system emerged that places the intein under strong purifying selection, because the self-splicing reaction of the intein now is necessary to synthesize a func- tioning DNA polymerase III [22]. The wide distribution of the split intein in *dnaE* in cyanobacteria [58] suggests that this rather complex gene structure is an evolutionarily stable arrangement.

Another mobile genetic element that is frequently trans- ferred and creates novelties and complexity is group II introns. They are thought to be the predecessors of both the eukaryotic spliceosomal introns and non-LTR retrotranspo- son [59–61]. These self-splicing elements are found in all domains of life; they are made mobile either via retro- homing, using an endonuclease [62], or retrotransposition mechanisms, using a reverse transcriptase [63]. Evidence for group II introns being the ancestors of the spliceosomal intron in eukaryotes includes similar splicing mechanisms, comparable boundary sequences, and secondary structure similarities [64–66]. One hypothesis suggests the group II intron originated in the bacteria and were horizontally trans- ferred from the alphaproteobacterial endosymbiont ancestor of the mitochondria to the genome of the ancestor of the eukaryotic nucleocytoplasm. The presence of introns in most transcripts might have necessitated a separation between transcription and translation, facilitating the emergence of a nucleus [67]. Some of the original introns may have lost their self-splicing activity and relied on other introns and their associated proteins to catalyze the splicing reaction in trans, evolving over time into the spliceosomal machinery. In this scenario, the introns initially proliferated as molecular para- sites; however, on the long run they allowed for exon shuf- fling, alternative splicing, and the nonsense mediated decay pathway to evolve. Interestingly, extant bacterial group II introns maintain self-splicing and mobility, while most mito- chondrial and chloroplast group II introns are not mobile and have lost the ability to self-splice. For example, about 20 group II introns present in the organelles of plants have lost their ability to self-splice [68, 69]. However, to maintain functional genes, they must be spliced out thus their maintenance is dependent on the complex interactions with nuc- lear and plastid splicing factors. Group II introns have also been implicated in genome rearrangements and gene con- version events [70], both of which can cause innovations in gene function and structure.

6. Interdomain HGT and Innovation

One of the benefits of HGT is that it can provide a selective advantage for organisms to occupy new niches and expand host ranges. Many interdomain transfers from bacteria to

single-celled eukaryotes provided for innovations and adaptation to new environments [28, 71]. In many instances these genes were subsequently transferred between divergent single-cell eukaryotes [28]. One example is the parasitic pro- tozoan *Blastocystis*, which is found in many different animal gut environments and causes gastrointestinal diseases, and has acquired genes for energy metabolism, adhesion, and osmotrophy from various bacterial donors. These transfers have allowed the successful adaptation of Blastocystis to the gut environment[72].

Surprisingly many genes were transferred from bacteria into multicellular eukaryotes. The ancient bacterivorous nematodes acquired cell wall degrading enzymes from several bacterial lineages via HGT [73–75]. The cell wall deg- rading genes are required for the initial stages in plant patho- genesis, without them plants would be an unavailable niche for the nematode [76]. Therefore, the transfer of those genes allowed the transition of the nematode from a free living state to a plant parasite [77]. Other examples of innovative interdomain HGTs can be found in the tunicates. A cellulose synthase gene (*cesA*) is proposed to have been transferred to the ancestor of the tunicates from a bacterial lineage [78]. Following a gene duplication, CesA1 produces cellulose for the larval tail and CesA2 synthesizes cellulose for the complex filter-feeding house of the ascidians and larvaceans [78]. This HGT played a role in body plan development in tunicates.

Examples of bacteria to animal transfers also reveal the adaptive benefits. The HhMAN1 gene in the coffee berry borer, *Hypothenemus hampei*, was likely transferred from a bacterial lineage [79]. The gene encodes a secreted manna- nase that allows the coffee berry borer access the primary seed storage polysaccharide in the coffee plant and ultimately confers an adaptive advantage because *H. hampei* uses the coffee berry as a specific host [79]. The spider mite *Tetrany- chus urticae* has several genes likely transferred from bacterial lineages; those are genes that encode a secreted fructosidase and a cyanate lyase-encoding gene that may be involved in feeding on cyanogenic plants [80]. These acquisitions have allowed the spider mite to utilize different plants for feeding thereby expanding its host range [80].

The aphid genome, *Acyrthosiphon pisum*, encodes for multiple carotenoids transferred from fungal lineages. These genes allow the aphid to synthesize its own carotenoids rather than to acquire them from food sources as many other ani- mals do [81]. These are only a few of the current examples of interdomain HGTs. As more and more genomes from multicellular organisms become available more interdomain transfers are likely to be revealed.

7. Gene Duplication and Gene Transfer

The emergence of new genes from previously noncoding DNA is a rare event (e.g., [82, 83]). Most new genes are believed to originate through gene duplication [84]. In Eukaryotes gene duplications frequently occur in an auto- chthonous fashion within a single lineage (Figure 1(a)). Mechanisms include tandem, segmental, and chromosomal duplication, retrotransposition, and genome duplications [85]. Of the two genes created, most frequently one accumu- lates mutations and is no longer maintained under purifying selection and decays [86]. There are two mechanisms by which the duplicated gene can be maintained, subfunction- alization or neofunctionalization. In subfunctionalization, functions of the parent gene are divided among the dupli- cated genes; in neofunctionalization, after duplication one copy diverges to create a new function. The creation of new functions from duplicated genes appears to be a rare event [87].

Ancient genome duplications have played an important role in vertebrate, plant, and fungi evolution (see [88] for review). In these ancient duplications it is difficult to decide if the whole genome duplication resulted from an autochthonous autopolyploidization or an allopolyploidiza- tion following a between-species hybridization (Figure 1(c)). The latter process is particularly important in plant evolution and breeding [89]. Many of these whole-gene duplications are followed by neofunctionalization and subfunctionaliza- tions of various genes throughout the genome. However the above example of the cellulose synthatase genes in the larvacean lineage of tunicates is an example of a gene dup- lication leading to neofunctionalization in a eukaryote.

The whole-genome duplication of the fungus *Saccha- romyces cerevisiae* followed by neofunctionalization of vari- ous genes led to the emergence of viral defense mechanisms from translation elongation and the emergence of gene silencing from origin of replication binding proteins [33]. Subfunctionalization events after gene or genome duplica- tions can also arise and create novel regulatory pathways. For example, the maize genome arose from an allotetraploidiza- tion between two grass species [90–93]. In the extant maize lineage the *ZAG1* and *ZMM2* genes are necessary for the development of stamens and carpals in the plant. The *ZAG1* gene is expressed throughout carpal development, and the *ZMM2* gene is expressed in maize stamen but not in the immature carpal [94]. It is thought that these genes were expressed in both developing stamens and carpals in the allotetraploid ancestor shortly after the polyploidization event [95]. Over time mutations affecting the regulation of *ZAG1* decrease expression of ZAG1 in stamens but not carpals and mutations affecting the regulation of *ZMM2* eliminated expression in the early carpal but not in stamens [95].

In Bacteria and Archaea autochthonous gene duplica- tions appear to be rare [42, 96]. The typical pathway for gene family extension is through HGT followed by non- homologous recombination in the recipient. Following the divergence of two lineages, orthologous genes experience substitutions. These might be associated with altered prop- erties of the encoded protein; for example, mutations in an ion translocating subunit of an ATP synthase/ATPase might increase its specificity for protons, thereby changing the specificity for the transported ion from Na^+ to H^+ [97], allowing the organism to use the proton motive force for ATP synthesis. When subsequently the two genes end up in the same cell following horizontal gene transfer, they have diverged so much that homologous recombination between the divergent forms is no longer possible (Figure 1(b)).

As both genes have different functions, both can be maintained in the recipient through purifying selection. For example, one ATPase might function as ATP synthase driven by a Na^+ gradient, and the homolog might function in controlling the cellular pH.

8. Conclusions

The processes of reticulate evolution lead to innovations and complexity. Horizontal gene transfer whether beneficial or parasitic in nature can lead to innovations and increased complexity. "Beneficial" HGTs provide an immediate selec- tive advantage to the recipient, which increases fitness and guarantees that the transferred gene will be fixed in the recipient's population. Such benefits include but are not limited to innovations in metabolic pathways, expansion of niche adaptations, and in the case of the human gut micro- biome can have important secondary implications for the human. "Parasitic" HGTs can also provide innovation, although innovation is more likely to be formed through neutral or nearly neutral pathways to complexity. Gene and genome duplications are another way to spawn innovation and complexity, more so in Eukaryotes than in prokaryotic lineages. In both cases, the horizontal transfer of genetic material and gene and genome duplications are a driving fac- tor in organismal evolution.

Acknowledgment

This work was supported through the National Science Foundation (DEB 0830024) and the NASA Exobiology pro- gram (NNX08AQ10G).

References

[1] T. Dagan, Y. Artzy-Randrup, and W. Martin, "Modular net- works and cumulative impact of lateral transfer in prokaryote genome evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 29, pp. 10039–10044, 2008.

[2] D. Williams, G. P. Fournier, P. Lapierre et al., "A rooted net of life," *Biology Direct*, vol. 6, article 45, 2011.

[3] P. Lopez and E. Bapteste, "Molecular phylogeny: reconstruct- ing the forest," *Comptes Rendus Biologies*, vol. 332, no. 2-3, pp. 171–182, 2009.

[4] P. Puigb, Y. I. Wolf, and E. V. Koonin, "Search for a "Tree of Life" in the thicket of the phylogenetic forest," *Journal of Bio- logy*, vol. 8, no. 6, article 59, 2009.

[5] V. Merhej, C. Notredame, M. Royer-Carenzi, P. Pontarotti, and D. Raoult, "The rhizome of life: the sympatric *Rickettsia felis* paradigm demonstrates the random transfer of DNA sequences," *Molecular Biology and Evolution*, vol. 28, no. 11, pp. 3213–3223, 2011.

[6] D. Raoult, "The post-Darwinist rhizome of life," *The Lancet*, vol. 375, no. 9709, pp. 104–105, 2010.

[7] O. Popa, E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan, "Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes," *Genome Research*, vol. 21, no. 4, pp. 599–609, 2011.

[8] C. P. Andam and J. P. Gogarten, "Biased gene transfer in microbial evolution," *Nature Reviews Microbiology*, vol. 9, no. 7, pp. 543–555, 2011.

[9] R. G. Beiko, T. J. Harlow, and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14332–14337, 2005.

[10] O. Zhaxybayeva, K. S. Swithers, P. Lapierre et al., "On the chimeric nature, thermophilic origin, and phylogenetic place- ment of the Thermotogales," *Proceedings of the National Acad- emy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5865–5870, 2009.

[11] B. Boussau, L. Gue´guen, and M. Gouy, "Accounting for hori- zontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria," *BMC Evolutionary Biology*, vol. 8, no. 1, article 272, 2008.

[12] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, "Prokaryotic evolution in light of gene transfer," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002.

[13] R. Dawkins, *The Selfish Gene*, Oxford University Press, 1976. [14] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 679–687, 2005.

[15] P. Lapierre and J. P. Gogarten, "Estimating the size of the bac- terial pan-genome," *Trends in Genetics*, vol. 25, no. 3, pp. 107– 110, 2009.

[16] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[17] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira et al., "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes," *Science*, vol. 317, no. 5845, pp. 1753–1756, 2007.

[18] L. Klasson, Z. Kambris, P. E. Cook, T. Walker, and S. P. Sinkins, "Horizontal gene transfer between *Wolbachia* and the mosquito Aedes aegypti," *BMC Genomics*, vol. 10, article 33, 2009.

[19] M. R. Goddard and A. Burt, "Recurrent invasion and extinc- tion of a selfish gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13880–13885, 1999.

[20] J. P. Gogarten and E. Hilario, "Inteins, introns, and homing endonucleases: recent revelations about the life cycle of para- sitic genetic elements," *BMC Evolutionary Biology*, vol. 6, article 94, 2006.

[21] D. Mo, L. Wu, Y. Xu et al., "A maturase that specifically stabi- lizes and activates its cognate group I intron at high tempera- tures," *Biochimie*, vol. 93, no. 3, pp. 533–541, 2011.

[22] K. S. Swithers and J. P. Gogarten, "Introns and Inteins," in *Bac- terial Integrative Mobile Genetic Elements*, chapter 4, Landes Bioscience, Austin, Tex, USA, 2012.

[23] J. De Vries and W. Wackernagel, "Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 4, pp. 2094–2099, 2002.

[24] M. Vulic´, R. E. Lenski, and M. Radman, "Mutation, recombi- nation, and incipient speciation of bacteria in the laboratory," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 13, pp. 7348–7351, 1999.

[25] H. Ochman, J. G. Lawrence, and E. A. Grolsman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000.

[26] A. P. Roberts and P. Mullany, "A modular master on the move: the Tn916 family of mobile genetic elements," *Trends in Microbiology*, vol. 17, no. 6, pp. 251–258, 2009.

[27] R. A. F. Wozniak and M. K. Waldor, "Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow," *Nature Reviews Microbiology*, vol. 8, no. 8, pp. 552–563, 2010.

[28] J. O. Andersson, "Gene transfer and diversification of micro- bial eukaryotes," *Annual Review of Microbiology*, vol. 63, pp. 177–193, 2009.

[29] A. H. Paterson, M. Freeling, H. Tang, and X. Wang, "Insights from the comparison of plant genome sequences," *Annual Review of Plant Biology*, vol. 61, pp. 349–372, 2010.

[30] O. Jatllon, J. M. Aury, F. Brunet et al., "Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early verte- brate proto-karyotype," *Nature*, vol. 431, no. 7011, pp. 946–957, 2004.

[31] J. M. Aury, O. Jaillon, L. Duret et al., "Global trends of whole- genome duplications revealed by the ciliate *Paramecium tetra- urelia*," *Nature*, vol. 444, no. 7116, pp. 171–178, 2006.

[32] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997.

[33] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolu- tionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[34] J. G. Ferry, "Methane from acetate," *Journal of Bacteriology*, vol. 174, no. 17, pp. 5489–5495, 1992.

[35] G. P. Fournier and J. P. Gogarten, "Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene trans- fer from cellulolytic Clostridia," *Journal of Bacteriology*, vol. 190, no. 3, pp. 1124–1127, 2008.

[36] J. D. Woodson, C. L. Zayas, and J. C. Escalante-Semerena, "A new pathway for salvaging the coenzyme B12 precursor cobi- namide in archaea requires cobinamide-phosphate synthase (CbiB) enzyme activity," *Journal of Bacteriology*, vol. 185, no. 24, pp. 7193–7201, 2003.

[37] P.-A. Christin, E. J. Edwards, G. Besnard et al., "Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer," *Current Biology*, vol. 22, no. 5, pp. 445–449, 2012.

[38] R. E. Blankenship, "Molecular evidence for the evolution of photosynthesis," *Trends in Plant Science*, vol. 6, no. 1, pp. 4–6, 2001.

[39] J. Raymond, "Coloring in the tree of life," *Trends in Microbiol- ogy*, vol. 16, no. 2, pp. 41–43, 2008.

[40] J. Raymond, O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes, and R. E. Blankenship, "Whole-genome analysis of photosynthetic prokaryotes," *Science*, vol. 298, no. 5598, pp. 1616–1620, 2002.

[41] J. Xiong and C. E. Bauer, "Complex evolution of photosynthesis," *Annual Review of Plant Biology*, vol. 53, pp. 503–521, 2002.

[42] D. Williams, C. P. Andam, and J. P. Gogarten, "Horizontal gene transfer and the formation of groups of microorganisms," in *Molecular Phylogeny of Microorganisms. Hethersett*, A. Oren and R. T. Papke, Eds., Caister Academic Press, Norwich, UK, 2010.

[43] J. Raymond, "The role of horizontal gene transfer in photosynthesis, oxygen production, and oxygen tolerance," *Methods in Molecular Biology*, vol. 532, pp. 323–338, 2009.

[44] S. Sadekar, J. Raymond, and R. E. Blankenship, "Conserva- tion of distantly related membrane proteins: photosynthetic reaction centers share a common structural core," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 2001–2007, 2006.

[45] M. Khomyakova, O. Bukmez, L. K. Thomas, T. J. Erb, and I. A. Berg, "A methylaspartate cycle in haloarchaea," *Science*, vol. 331, no. 6015, pp. 334–337, 2011.

[46] C. H. Lester, N. Frimodt-Møller, T. L. Sørensen, D. L. Monnet, and A. M. Hammerum, "In vivo transfer of the vanA resistance gene from an *Enterococcus faecium* isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers," *Antimicrobial Agents and Chemotherapy*, vol. 50, no. 2, pp. 596–599, 2006.

[47] J. H. Hehemann, G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, and G. Michel, "Transfer of carbohydrate-active enzy- mes from marine bacteria to Japanese gut microbiota," *Nature*, vol. 464, no. 7290, pp. 908–912, 2010.

[48] F. Thomas, T. Barbeyron, T. Tonon, S. Genicot, M. Czjzek, and G. Michel, "Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine Flavobacteriia to their independent transfers to marine Proteobacteria and human gut Bacteroides," *Environmental Microbiology*. In press.

[49] E. Rosenberg, G. Sharon, and I. Zilber-Rosenberg, "The hologenome theory of evolution contains Lamarckian aspects within a Darwinian framework," *Environmental Microbiology*, vol. 11, no. 12, pp. 2959–2962, 2009.

[50] B. S. Chevalier and B. L. Stoddard, "Homing endonucleases: structural and functional insight into the catalysts of intron/ intein mobility," *Nucleic Acids Research*, vol. 29, no. 18, pp. 3757–3774, 2001.

[51] L. Olendzenski and J. P. Gogarten, "Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer," *Annals of the New York Academy of Sciences*, vol. 1178, pp. 137–145, 2009.

[52] B. Dujon, "Group I introns as mobile genetic elements: facts and mechanistic speculations—a review," *Gene*, vol. 82, no. 1, pp. 91–114, 1989.

[53] V. Koufopanou and A. Burt, "Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site- directed mutagenesis," *Molecular Biology and Evolution*, vol. 22, no. 7, pp. 1535–1538, 2005.

[54] M. Dori-Bachash, B. Dassa, O. Peleg, S. A. Pineiro, E. Jurke- vitch, and S. Pietrokovski, "Bacterial intein-like domains of predatory bacteria: a new domain type characterized in *Bdel- lovibrio bacteriovorus*," *Functional and Integrative Genomics*, vol. 9, no. 2, pp. 153–166, 2009.

[55] G. Amitai, O. Belenkiy, B. Dassa, A. Shainskaya, and S. Pietrokovski, "Distribution and function of new bacterial intein-like protein domains," *Molecular Microbiology*, vol. 47, no. 1, pp. 61–73, 2003.

[56] O. G. Wikmark, C. Einvik, J. F. De Jonckheere, and S. D. Johansen, "Short-term sequence evolution and vertical inher- itance of the *Naegleria* twin-ribozyme group I intron," *BMC Evolutionary Biology*, vol. 6, article 39, 2006.

[57] B. Dassa, N. London, B. L. Stoddard, O. Schueler-Furman, and S. Pietrokovski, "Fractured genes: a novel genomic arran- gement involving new split inteins and a new homing endo- nuclease family," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2560–2573, 2009.

[58] J. Caspi, G. Amitai, O. Belenkiy, and S. Pietrokovski, "Distribution of split DnaE inteins in cyanobacteria," *Molecular Microbiology*, vol. 50, no. 5, pp. 1569–1577, 2003.

[59] T. R. Cech, "The generality of self-splicing RNA: relationship to nuclear mRNA splicing," *Cell*, vol. 44, no. 2, pp. 207–210, 1986.
[60] P.A. Sharp, "On the origin of RNA splicing and introns," *Cell*, vol. 42, no. 2, pp. 397–400, 1985.
[61] S. Zimmerly, H. Guo, P. S. Perlman, and A. M. Lambowitz, "Group II intron mobility occurs by target DNA-primed reverse transcription," *Cell*, vol. 82, no. 4, pp. 545–554, 1995.
[62] B. Cousineau, D. Smith, S. Lawrence-Cavanagh et al., "Retro-homing of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombi-nation," *Cell*, vol. 94, no. 4, pp. 451–462, 1998.
[63] N. Toro, J. I. Jime´ nez-Zurdo, and F. M. Garc´ıa-Rodr´ıguez, "Bacterial group II introns: not just splicing," *FEMS Microbi- ology Reviews*, vol. 31, no. 3, pp. 342–358, 2007.
[64] G. C. Shukla and R. A. Padgett, "A catalytically active group II intron domain 5 can function in the U12-dependent spliceo- some," *Molecular Cell*, vol. 9, no. 5, pp. 1145–1150, 2002.
[65] H. D. Madhani and C. Guthrie, "A novel base-pairing inter- action between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome," *Cell*, vol. 71, no. 5, pp. 803–817, 1992.
[66] K. S. Keating, N. Toor, P. S. Perlman, and A. M. Pyle, "A struc- tural analysis of the group II intron active site and implications for the spliceosome," *RNA*, vol. 16, no. 1, pp. 1–9, 2010.
[67] E. V. Koonin, "The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?" *Biology Direct*, vol. 1, article 22, 2006.
[68] T. S. Kroeger, K. P. Watkins, G. Friso, K. J. Van Wijk, and A. Barkan, "A plant-specific RNA-binding domain revealed through analysis of chloroplast group II intron splicing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 11, pp. 4537–4542, 2009.
[69] K. P. Watkins, M. Rojas, G. Friso, K. J. van Wijk, J. Meurer, and A. Barkan, "APO1 promotes the splicing of chloroplast group II introns and harbors a plant-specific zinc-dependent RNA binding domain," *Plant Cell*, vol. 23, no. 3, pp. 1082–1092, 2011.
[70] S. Leclercq, I. Giraud, and R. Cordaux, "Remarkable abun- dance and evolution of mobile group II introns in *Wolbachia* bacterial endosymbionts," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 685–697, 2011.
[71] J. Huang and J. P. Gogarten, "Ancient horizontal gene transfer can benefit phylogenetic reconstruction," *Trends in Genetics*, vol. 22, no. 7, pp. 361–366, 2006.
[72] F. Denoeud, M. Roussel, B. Noel et al., "Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite," *Genome Biology*, vol. 12, no. 3, article R29, 2011.
[73] S. E. Kalla, D. C. Queller, A. Lasagni, and J. E. Strassmann, "Kin discrimination and possible cryptic species in the social amoeba *Polysphondylium violaceum*," *BMC Evolutionary Biol- ogy*, vol. 11, no. 1, article 31, 2011.
[74] J. P. McCarter, "Nematology: terra incognita no more," *Nature Biotechnology*, vol. 26, no. 8, pp. 882–884, 2008.
[75] M. Mitreva, G. Smant, and J. Helder, "Role of horizontal gene transfer in the evolution of plant parasitism among nema- todes," *Methods in Molecular Biology*, vol. 532, pp. 517–535, 2009.
[76] G. Smant, J. P. W. G. Stokkermans, Y. Yan et al., "Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes," *Proceed- ings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 4906–4911, 1998.
[77] C. Dieterich and R. J. Sommer, "How to become a parasite—lessons from the genomes of nematodes," *Trends in Genetics*, vol. 25, no. 5, pp. 203–209, 2009.
[78] Y. Sagane, K. Zech, J. M. Bouquet, M. Schmid, U. Bal, and E. M. Thompson, "Functional specialization of cellulose synthase genes of prokaryotic origin in chordate larvaceans," *Development*, vol. 137, no. 9, pp. 1483–1492, 2010.
[79] R. Acun˜a, B. E. Padilla, C. P. Flo´ rez-Ramos et al., "Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 11, pp. 4197–4202, 2012.
[80] M. Grbic´, T. Van Leeuwen, R. M. Clark et al., "The genome of *Tetranychus urticae* reveals herbivorous pest adaptations," *Nature*, vol. 479, no. 7374, pp. 487–492, 2011.
[81] N. A. Moran and T. Jarvik, "Lateral transfer of genes from fungi underlies carotenoid production in aphids," *Science*, vol. 328, no. 5978, pp. 624–627, 2010.
[82] M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun, "Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9935–9939, 2006.
[83] D. G. Knowles and A. McLysaght, "Recent de novo origin of human protein-coding genes," *Genome Research*, vol. 19, no. 10, pp. 1752–1759, 2009.
[84] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, Germany, 1970.
[85] M. Long, E. Betra´n, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.
[86] M. Lynch and J. S. Conery, "The evolutionary fate and con-sequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.
[87] M. W. Hahn, "Distinguishing among evolutionary models for the maintenance of gene duplicates," *Journal of Heredity*, vol. 100, no. 5, pp. 605–617, 2009.
[88] Y. Van De Peer, S. Maere, and A. Meyer, "The evolutionary significance of ancient genome duplications," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 725–732, 2009.
[89] J. P. Gogarten and L. Olendzenski, "Orthologs, paralogs and genome comparisons," *Current Opinion in Genetics and Deve- lopment*, vol. 9, no. 6, pp. 630–636, 1999.
[90] M. M. Goodman, C. W. Stuber, K. Newton, and H. H. Weissinger, "Linkage relationships of 19 enzyme loci in maize," *Genetics*, vol. 96, pp. 697–710, 1980.
[91] T. Helentjaris, D. Weber, and S Wright, "Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms," *Genetics*, vol. 118, pp. 353–363, 1988.
[92] B. S. Gaut and J. F. Doebley, "DNA sequence evidence for the segmental allotetraploid origin of maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6809–6814, 1997.
[93] S. White and J. Doebley, "Of genes and genomes and the origin of maize," *Trends in Genetics*, vol. 14, no. 8, pp. 327–332, 1998.
[94] M. Mena, B. A. Ambrose, R. B. Meeley, S. P. Briggs, M. F. Yanofsky, and R. J. Schmidt, "Diversification of C-function activity in maize flower development," *Science*, vol. 274, no. 5292, pp. 1537–1540, 1996.
[95] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait, "Preservation of duplicate genes by comple-mentary, degenerative mutations," *Genetics*, vol. 151, no. 4, pp. 1531–1545, 1999.

[96] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes," *PLoS Genetics*, vol. 7, no. 1, Article ID e1001284, 2011.

[97] J. Dzioba, C. C. Ha¨se, K. Gosink, M. Y. Galperin, and P. Dibrov, "Experimental verification of a sequence-based prediction: F_1F_o-type ATPase of *Vibrio cholerae* transports protons, not Na^+ ions," *Journal of Bacteriology*, vol. 185, no. 2, pp. 674–678, 2003.

2.2 Orthologues, Paralogues, and Horizontal Gene Transfer in the Human Holobiont.

Orthologues, Paralogues and Horizontal Gene Transfer in the Human Holobiont

Shannon Soucy, University of Connecticut, Storrs, Connecticut, USA
Lorraine Olendzenski, St. Lawrence University, Canton, New York, USA
J Peter Gogarten, University of Connecticut, Storrs, Connecticut, USA

Based in part on the previous version of this eLS article 'Orthologues, Paralogues and Xenologues in Human and Other Genomes' (2008) by Olga Zhaxybayeva, J Peter Gogarten and Lorraine Olendzenski.

Advanced article

Article Contents

Online posting date: 15th March 2013

Evolution is commonly measured using comparative phylogenetic analysis. Comparisons of orthologous characters and sequences from different species are used to infer organismal evolution. Analyses of duplicated genes can be used to root phylogenetic trees and infer ancestral groups. The expansion of gene families through gene and genome duplications allowed more complex regulatory and developmental pathways to evolve in multicellular eukaryotes. In prokaryotes and single-celled eukaryotes, the acquisition offoreign genes by horizontal gene transfer is the main mechanism for gene family expansion; it allows genomes to evolve new traits quickly and facilitates the assembly of new metabolic pathways. Additionally, prokaryotic organisms with short generation times will accumulate genetic adaptations at a much faster rate than organisms with longer generation times (e.g. humans). In multicellular animals where somatic cells and gametes are separate, acquisition offoreign genes is rare, leading to high levels of similarity in gene content. However, multicellular eukaryotes have evolved in close association with prokaryotic symbionts that impact development, physiology and ecology of the association. To understand the evolution of the complex human systems, we must consider the genomes of the associated microbiota, known as the microbiome. We must therefore consider the human as a holobiont, a complex ecosystem, whose evolutionary fitness is determined by the host, the symbionts and their interactions.

eLS subject area: Evolution & Diversity of Life

How to cite:
Soucy, Shannon; Olendzenski, Lorraine; and Gogarten, J Peter (March 2013) Orthologues, Paralogues and Horizontal Gene Transfer in the Human Holobiont. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005298.pub3

Orthologous, Paralogous and Xenologous Genes

Sequences or structures that evolved from a single ancestral structure or sequence are homologous. To classify the different types of homology, Fitch (1970) introduced the terms orthology and paralogy. Orthologous structures or sequences in two organisms are homologues that evolved from the same feature in their most recent common ancestor; however, orthologues do not necessarily retain their ancestral function. Because the evolution of orthologues reflects organismal evolution, molecular systematics has been concerned traditionally with comparing orthologous sequences. By contrast, paralogues are homologues whose evolution reflects gene duplication events. For example, the b-chain of haemoglobin is a paralogue of both the haemoglobin a-chain and myoglobin because they each evolved from the same ancestral globin gene through repeated gene duplication events. Only the deepest split in a phylogenetic tree relating homologous proteins determines orthology versus paralogy (Fitch, 1970). If the deepest split between two genes corresponds to a speciation event, those genes are orthologues. If the split corresponds to a gene duplication event, then those genes would be considered paralogues. To clearly distinguish whether two genes are orthologues or paralogues, a rooted phylogeny is necessary.

Using genes that encode resistance to antibiotics as a model, the term xenology was coined for homologues that were acquired by an organism through horizontal gene transfer (Fitch, 1970). Synology denotes homologues that

arose from the fusion of complete genomes (Gogarten, 1994), such as bacterial genes brought into the eukaryotic cell through the mitochondrial endosymbiont. See also: Homologous, Orthologous and Paralogous Genes

Gene Number and Genome Organisation

In the human genome, protein-coding genes tend to exist in nonrandom clusters that are separated by large stretches of nonprotein-coding deoxyribonucleic acid (DNA), referred as gene-poor 'deserts'. Regulatory elements also appear to be clumped into regulatory rich and poor regions (Zhang et al., 2007). Up to 1% of the total human genome comprises exons – the regions of genes that encode proteins. Introns, the regions in genes that are spliced out during the creation of messenger ribonucleic acid (mRNA), make up approximately 24% of the genome. The number of protein coding genes is estimated to be approximately 22 000 (Pertea and Salzberg, 2010). More than 90% of multiexon-coded proteins undergo alternative splicing, allowing more than a single protein to be translated from a region of exons. This further complicates estimates of protein-coding gene number, although the function of the vast majority of alternatively spliced transcripts remains unknown, and some may reflect transcriptional noise rather than a distinct function (Pertea, 2012). See also: Clustering of Highly Expressed Genes in the Human Genome ; Gene Clustering in Eukaryotes ; Genome Organization of Vertebrates ; Isochores

Additionally, individual humans can differ slightly in genome content with variation related primarily to deletions or regions of segmental duplication. Comparison of two different human genomes, one from Africa and one from Asia, with the reference genome at NCBI showed 5 Mb of unique DNA in each of the new genomes. Estimates suggest a human pan-genome would include up to 40 Mb more DNA (or 5 0.01%) than the reference genome (Pertea and Salzberg, 2010). This is a very small difference compared to the estimated 90% difference in size between the pan-genome and a single reference genome of the bacterium Escherichia coli. Comparison of 60 E. coli genomes suggests that less than 10% of the genes in the E. coli pan-genome are present in all of the 61 E. coli genomes analysed (Lukjancenko et al., 2010).

In addition to protein-coding mRNAs, a new class of transcripts, collectively called noncoding RNAs (ncRNA), has been identified. These transcripts do not code for proteins, and they originate from intergenic regions, introns or from sequences antisense to known transcripts. Currently, their function is not well understood (Johnson et al., 2005). These ncRNA encoding regions expand the traditional definition of a gene to include a myriad of nonprotein coding sequences and hint towards complex patterns of expression and regulation during development of different cell types. The growing list of functional RNAs increases the number of estimated genes from 22 000 to 30 000–40 000 genes (Pertea, 2012). ncRNAs include small interfering RNA and microRNA, which play a central role in RNA interference by binding to specific mRNA molecules to increase or decrease the amount of protein translated. Other ncRNAs that have a role in gene silencing include the PIWI-interacting RNAs, which bind PIWI proteins during spermatogenesis and are thought to be involved in silencing transposons in the genome. Promoter-associated RNAs, transcription initiation RNAs, X-inactivation RNAs and various other classes of ncRNAs are also suggested to have functional roles (Pertea, 2012). The long ncRNAs (lncRNAs), defined as ncRNAs longer than 200 bp, undergo splicing with similar frequency to protein-coding mRNAs and are probably the least well-understood transcripts. Ponjavic et al. (2007) analysed a set of more than 3000 lncRNAs and found the substitution pattern and indel distribution in comparison of mouse, and human homologues suggest that these macro RNAs are under purifying selection. Although the function of many of these nonprotein-coding RNAs is still to be determined, they may be the key regulators of epigenetic gene regulation in mammalian cells (Pertea, 2012).

Based on the differential expression, localisation and patterns of conservation in ncRNAs, it is likely that the portion of the human genome that is functional has previously been underestimated. Analysis of substitution rates suggests that 6.5–10% of the genome appears to be under selective constraint (Meader et al., 2010). At the other extreme of functional DNA estimates, the EnCODE project combined data from a variety of analyses to map RNA transcribed regions, protein-coding regions, transcription-factor binding sites, chromatin structure and DNA methylation sites. The study encompassed 1640 genome wide datasets from 147 different human cell types. The non-translated regions that may have regulatory functions included elements such as enhancers, promoters and regions that contribute to the structure of chromatin. The sum of these data were interpreted to suggest that up to 80% of the genome contains elements that participate in at least one of these functions (Zhang et al., 2007).

The ratio of constrained (and therefore likely functional) nonprotein-coding bases to coding bases in Drosophila is 2, whereas in humans, it is between 5 and 8. Much of the apparent differences in complexity between species may be due to a varying amount of noncoding regulatory sequence, regulating a fairly stable core of protein-coding genes (Meader et al., 2010). This is compatible with the notion that much of the organismal complexity and interspecific differences of mammals are encoded in the nonprotein-coding functional complement rather than in protein-coding sequence (Ponting and Hardson, 2011).

Duplicated and Repetitive DNA

Repeated sequences derived from transposable elements comprise 43–45% of the genome (Li et al., 2001) and

include long interspersed nuclear elements (LINEs or L1 elements), short interspersed nuclear elements (SINEs), DNA transposons and long-terminal repeat elements. SINEs include Alu family of repeats (a distinct class of retrotransposon-amplified repeat DNA that arose with the evolution of primates) and comprise approximately 10% of the genome (Li et al., 2001). Transposable elements can have a significant role in gene duplication through the formation of pseudogenes that lack introns (Kazazian, 2004). L1 elements can mobilise transcribed DNA and are involved in exon reshuffling. Many known proteins incorporate truncated L1 or Alu family elements in their transcripts through alternative splicing events (Li et al., 2001). Only 35–40 subfamilies of transposable elements remain actively mobile in the human genome (Mills et al., 2007). Comparison of human and chimpanzee genomes indicates that since the two species diverged, human endogenous retrovirus K (HERV-K) and L1 elements are active in both species, whereas Alu family of elements show approximately 3-fold higher activity in humans (The Chimpanzee Sequencing and Analysis Consortium, 2005). See also: Centromeric Sequences and Sequence Structures; Long Interspersed Nuclear Elements (LINEs) ; Retroviral Repeat Sequences; Transposable Elements: Evolution ; Transposons

Several ancient genome duplications occurred in the evolution of the vertebrate, plant and fungal lineages (Van De Peer et al., 2009). It is difficult to decide if these whole genome duplication resulted from an autochthonous autopolyploidisation or as a consequence of a between-species hybridisation (an allopolyploidization). The latter process is particularly important in plant evolution and breeding. Following these whole-genome duplications, many duplicated genes undergo pseudogenisation – a few duplicates acquire new functions following sub- or neo-functionalisation (Van De Peer et al., 2009).

Duplicated segments in the human genome are generally enriched in protein-coding genes (Zhang et al., 2005), and hence they have the potential to evolve novel transcripts, either as whole-gene duplications or through the creation of mosaic genes. For example, 11 new transcripts have been identified in the 10% of chromosome 22 that originated through segmental duplication (Bailey et al., 2002). One region of chromosome 16 contains a newly evolved, unique family of repeats. This gene family, named 'morpheus', consists of highly similar genes evolving so rapidly that they show no sequence similarity to known genes from other organisms and seem to be under positive selection (Johnson et al., 2001). See also: Chromosome 16; Chromosome 22; Segmental Duplications and Genetic Disease

Gene duplications can be either DNA or RNA mediated. RNA-mediated duplication results in genes that have lost introns and regulatory regions of the original gene; consequently, the rate with which duplicated genes turn into pseudogenes is much higher for RNA than for DNA-mediated duplications. However, because the former occur at a much higher rate, about half of the functional duplicated copies in mammals were determined to have originated through RNA intermediates (Jun et al., 2009). Using comparative genome analyses, Ciccarelli et al. (2005) identified 22 primate-specific gene duplications that are maintained as a single copy in other metazoan genomes. Eighty-two percent of these duplications are part of genome regions that underwent recent segmental duplications.

Recent variations in the number of paralogues in the lineage leading to humans and within the human populations are considered to reveal genomic regions under selective pressures (Gokcumen et al., 2011; Han et al., 2011). In asymmetric evolution after duplication, one duplicate evolves or degrades faster than the other and often becomes functionally or conditionally specialised. In a study on asymmetrically duplicated genes, confirmed duplicated gene sets identified across 13 vertebrate genomes were enriched in functional categories related to neuron differentiation and response to external stimuli (Prosdocimi et al., 2012).

Human Orthologues in Other Genomes

Many proteins evolved early in the metazoan lineage and have orthologues in invertebrate genomes, in fact only 7% of the protein motifs in humans are vertebrate specific, and it appears most of the protein complexity is due to shuffling of existing domains. The initial human genome sequence contained detectable homologues to 61% of proteins found in Drosophila, 43% of proteins found in Caenorhabditis elegans and 46% of proteins found in Saccharomyces cerevisiae It was found that 1308 groups of proteins, encompassing 3129 human proteins, contain at least one orthologue in each of the four species (human, fruitfly, nematode and yeast). These groups of proteins represent basic housekeeping functions in the cell, including respiration, transcription, translation and membrane functions. Of these groups, 564 contained only one orthologue (and no paralogues) from human, fruitfly, nematode and yeast (Lander et al., 2001) representing genes that had not undergone duplication or modification. This is a small percentage of the complete gene complement and indicates the extensive occurrence of gene duplication in the evolution of lineages. The large number of duplicated genes poses a challenge for identification of orthologues amongst eukaryotic genomes. Consequently, the numbers of orthologous gene sets vary with detection method. An analysis using four different methods for orthologue detection found 7663 orthologues shared between humans and*C. elegans* (or $\sim$38% of *C. elegans* proteins) and illustrates the complexity of finding orthologues between two species (Shaye and Greenwald, 2011). In the analysis of the chimpanzee genome, 13 454 pairs of human and chimpanzee genes were designated as orthologues with high-quality alignments, whereas addition of rat and mouse sequences reduced the number of unambiguously orthologous genes to 7043 (The Chimpanzee Sequencing and Analysis Consortium, 2005). See also: Alignment: Statistical Significance; Sequence Similarity; Similarity Search

Orthologous regions between the genomes are not limited to coding regions of a genome. The conserved noncoding portions of the genome (so-called 'dark matter') have been analysed by comparative genome analyses (Johnson et al., 2005). Multiple stretches of the human genome are identified as being extraordinarily conserved across large evolutionary distances (called 'ultra conserved elements' or UCEs). The 481 regions of the human genome that are more than 200 bp in length are 100% identical between human, rat and mouse genomes, and many of them are also highly conserved in chicken, dog and fish (Bejerano et al., 2004). Most of these UCEs lie outside of exons are under stronger purifying selection than nonsynonymous sites in protein-coding genes (Katzman et al., 2007) and still await functional assignment and explanation for such remarkable sequence conservation. The 0.14% of the human genome consisting of regions ofless striking conservation, but still of high similarity, are found in human and four other vertebrate genomes (mouse, rat, chicken and Fugu rubripes) (Siepel et al., 2005). These highly conserved elements (HCEs) are longer than UCEs and only 42% of them overlap with known exons. The reason for conservation of HCEs is unknown, as in the case of UCEs, but roles of control in gene expression and post-transcriptional regulation are suggested based on individual examples.

In addition to studying conserved regions, which may provide a hint towards functionality, it is also interesting to look into the fastest evolving (compared to other vertebrates) regions of human genome. 34 498 genomic regions that are $\geq$96% identical in chimpanzee, mouse and rat genome, but show changes in the human genome, were examined (Pollard et al., 2006). Only approximately 20% of these regions overlap with exons and 202 show evidence for accelerated evolution in the human genome. Many of the 202 human-accelerated regions are located either in introns of the genes related to transcription and DNA binding or adjacent to such genes.

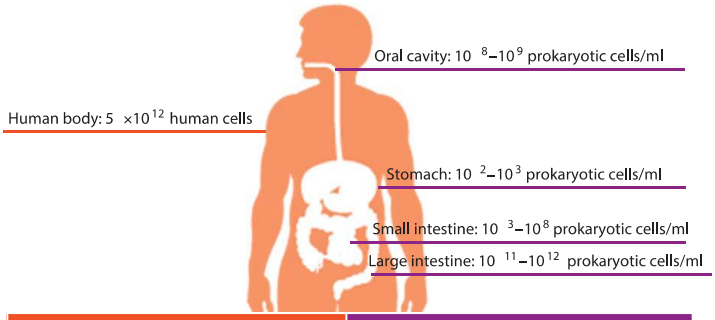
Xenologues in the Human Genome

Initial analyses of the draft human genome were interpreted to suggest that the human genome contained 113–223 genes that probably originated from horizontal gene transfer from bacteria directly into the human lineage (Lander et al., 2001). Given a close association between a prokaryotic symbiont and a eukaryotic host, gene transfer into the nucleus of the eukaryotic host, even in case of multicellular animals, is possible (e.g. Kondo et al., 2002). (A well-studied example for bacteria-to-eukaryote transfer is the many mitochondrial genes that now reside in the nucleus.) However, few, if any, of the postulated bacteria to human transfers have upheld closer scrutiny (Andersson et al., 2001; Salzberg et al., 2001). A reanalysis by Salzberg et al. (2001) has shown that the number suggested initially was affected by a species-sampling effect (i.e. by the number of nonvertebrate genomes that were included in the analyses). Differential gene loss might also produce similar results (Andersson et al., 2001). In addition, the direction of potential horizontal gene transfer remains unclear. Thus, the existence of putatively transferred genes directly from bacteria to the human lineage remains unconfirmed and requires additional analyses with more genomic data. See also: Bacterial DNA in the Human Genome ; Homologous, Orthologous and Paralogous Genes

The Human Microbiome

Although no recent xenologues were confirmed to be present in the human genome, humans are home to a complex coevolved microbial community. The small generation times, lack of nucleus and unicellular life cycle of bacteria make them conducive to relatively rapid evolution compared to humans. As eukaryotes evolved over time and developed more complex body plans, prokaryotes adapted to inhabit these newly developing niches. The limitations on the effective population size in these developing niches, imposed by host number, cell number, cell space and population bottlenecks during host transmission resulted in selective sweeps and specialisation in colonising different eukaryotic tissues (Toft and Andersson, 2010). Currently, each of us has approximately 100 trillion bacterial cells found in various locations from the skin to the lining of the alimentary canal and urogenital tract (Figure 1). The heaviest colonisation on the human body occurs in the gut or large intestine where densities approach 10^{11}–10^{12} cells per gram of colon contents (Walter and Ley, 2011). Humans are born sterile and are colonised during development with organisms from the environment, initially during passage through the birth canal and through subsequent contact with the primary caregiver. Factors such as breast feeding and vaginal birth increase the similarity between maternal and infant microbiome until the age of 2.5 years when the microbiota of the children becomes more unique, more stable and more like that of an adult (Parfrey and Knight, 2012). The collective number of different species associated with human intestine is $\sim$1000–1500, whereas the number of species associated with any single individual is $\sim$160, suggesting distinct and adaptable symbiotic populations relative to the environmental parameters specific to each individual. At a larger taxonomic scale microbiota cluster with respect to the host diet (herbivores, omnivores and carnivores) (Fraune and Bosch, 2010); however, within primates the composition of the gut microbiota tracks the evolutionary history of the host organism (Ochman et al., 2010), revealing a tight coevolutionary relationship.

Human genomes are 99.9% similar between individuals; however, the genetic material of the microbiota between even closely related individuals is 70–90% different (Parfrey and Knight, 2012). The microbiota contain approximately 150 times more nonredundant genes than in the human genome, suggesting functional flexibility as an important role of the microbiota (Qin et al., 2010).



Human body	Large intestine
5 × 10 12 Human cells	10 12 Prokaryotic cells/ml
2.1 × 10 4 Human genes	3.3 × 10 6 Non redundant prokaryotic genes

Figure 1 To the left of the body, underlined in red, are the number of human cells which make up the average human body, to the right underlined in purple are the number of prokaryotic cells associated with different locations of the body. The units are bacterial cells per ml, and thus the cumulative amount of prokaryotic cells in each organ is much greater. Additionally, beneath the body is a chart comparing the number of human cells and genes in the human genome (in red) to the number of cells per ml of bacteria in the large intestine, one of the most heavily colonised areas of the human body, and the number of nonredundant prokaryotic genes isolated from the large intestine (in purple).

Metagenomic analysis offaecal samples collected from 124 individuals were pooled and revealed 3.3 million non-redundant genes across all samples, 8% of these were genes shared between at least 50% of subjects, whereas 72% were rare genes present in less than 20% of subjects (Qin et al., 2010). Despite differences in composition between individuals, microbiomes appear largely functionally equivalent (Walter and Ley, 2011). Thus, the genetic information present in humans is a composite of Homo sapiens genes and genes present in the genomes of the trillions of microbes that colonise our adult bodies (Turnbaugh et al., 2006). When functional categories of genes were compared between the gut microbiota and human genome using odds ratios, the gut microbiota showed a significant enrichment in genes involved in metabolism, which were under-represented in the human genome (Gill et al., 2006). Metabolic specialisation encourages high species diversity and niche partitioning related to substrate preference (Spor et al., 2011). 'Our' microbial genomes (the microbiome) encode metabolic capacities that we have not had to evolve in our nuclear genome. See also: Endosymbionts ; Metagenomics and Microbial Communities

Possible Roles of the Microbiota

Human bacterial symbionts contribute to the absorption of carbohydrates, lipids and micronutrients, as well as metabolism of xenobiotics and toxins (Gill et al., 2006). It is difficult to gauge the extent of the impact of microbiota in human physiology, however, faecal transplants of microbiota from healthy subjects have been used to alleviate chronic Clostridium difficile infections in patients where antibiotics are ineffective; in 95% of cases colitis caused by dysbiosis was alleviated after transplantation (Gough et al., 2011). There is a well-demonstrated correlation between states of dysbiosis and diseases in humans such as, Crohn's, IBD, allergies, celiac's disease, gastric cancer, autism, obesity, anorexia, type II diabetes, type I diabetes, multiple sclerosis and rheumatoid arthritis (Clemente et al., 2012); however, it is not known if correlation can be attributed to causality as many of the diseases where a dysbiosis is noted are autoimmune disorders, and the tight link between the microbiota and immune function is difficult to tease apart. Much has been learned about the relationship between endogenous microbiota and the host organisms by using mouse gnotobiotic models (i.e. animals that harbour only a defined set of microorganisms). Studies using gnotobiotic mice have shown that microbes are involved in the development ofinnate immunity through mucosal fortification and additionally play a definitive role in development of the adaptive immune systems. Such studies illustrate a profound difference in physiology, especially relating to host defence, between animals with and without microbiota, implying coevolution between the

host and its symbionts for the purpose of collaboration against infectious agents (Lee and Mazmanian, 2010). Supporting this hypothesis, studies of specific members of the microbiota, such as Lactobacillus have shown such organisms have a protective effect against many forms of intestinal dysbiosis by inducing protective modifications to both the mucin and epithelial barrier, secreting antimicrobial substances, and replenishing suppressed beneficial microbiota (Mattar et al., 2001). Additionally 'parasitic' organisms such as helminthes may modulate the immune system and elicit a protective effect against certain types of dysbiosis, alleviating symptoms of arthritis, multiple sclerosis, type I diabetes and Crohn's disease (Rook, 2012).

These examples suggest long-term coevolution towards the currently established and delicately balanced relationship between the host and microbiota, which is essential to maintain homoeostasis (Rook, 2012).

Horizontal Transfers in the Human Microbiome

The field of horizontal gene transfer (HGT) within the human microbiome has exploded recently with the Human Microbiome project funded by NIH. In one study, a total of 13 514 high-confidence HGT genes were identified in the genomes of 308 human microbes (Liu et al., 2012). Most of the genes were involved in either catalysis or metabolism, again highlighting the important role of the microbiota in metabolic functions. In another study, a screen of 2235 human-associated bacterial genomes from different body sites showed a network of 10 770 unique, recently transferred genes, in most of which the HGT occurred between isolates from ecologically similar but geographically and phylogenetically distinct environments (Smillie et al., 2011). Bacteria involved in transfers often share similar body sites, oxygen tolerance or ability to cause disease, indicating an important role for ecology (environment) in driving these networks of gene sharing. A classic example of HGT is the transfer of antibiotic resistance genes; such genes have a selective advantage in the gut environment and can be transferred from the outside environment to the gut microbiota via food sources. In a study by Lester et al. (2006), volunteers were fed a strain of vancomycin resistant Enterococcus faecium isolated from a chicken; subsequently, vancomycin resistance was transferred to the human gut E. faecium in these volunteers, providing evidence for food as a reservoir for possible HGTs. An example of HGT impacting the nutrients available to the holobiont was recently discovered in the human microbiome of Japanese individuals. Genes for porphyranases, alginases and agarases, enzymes which facilitate the breakdown of carbohydrates in algal cell walls, were transferred from marine algal parasites to the gut organism Bacteroidetes plebeius (Hehemann et al., 2010). These HGTs allow the gut bacterium to utilise seaweed as a carbon source, and confers a secondary benefit to the human host, who can now utilise metabolites released by the bacterium after the food source is broken down. In these examples HGTs increase the fitness primarily of the microbes associated with the human gut and have a secondary benefit on the human host.

The Hologenome Theory of Evolution

The preceding observations support the hologenome theory of evolution (Fraune and Bosch, 2010; Rosenberg et al., 2009): the unit of selection is the holobiont, but in this case, the human host and microbial symbionts. The microbial symbionts profoundly affect the fitness of the host organism; in turn, the evolutionary trajectory of the microbiota is impacted by the health and well-being of the human host. Selection acting at either level, the microbiota or the human host, will act on the collective set of genes, the hologenome, in such a way that genes maintained and expressed by any organism present in the holobiont will have an effect on the holobiont as a whole. The holobiont can adapt to changing environmental conditions through acquiring new symbionts, or symbionts already present may acquire new genes and properties through HGT. The interactions between host and symbiont (including commensals and parasites) have a long evolutionary history. Disturbance of these long-established interactions may have surprising consequences for human health and well-being (Rook, 2012).

References

Andersson JO, Doolittle WF and Nesbo CL (2001) Genomics. Are there bugs in our genome? Science 292: 1848–1850.

Bailey JA, Yavor AM, Viggiano L et al. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. American Journal of Human Genetics 70: 83–100.

Bejerano G, Pheasant M, Makunin I et al. (2004) Ultraconserved elements in the human genome.Science 304: 1321–1325.

Ciccarelli FD, von Mering C, Suyama M et al. (2005) Complex genomic rearrangements lead to novel primate gene function. Genome Research 15: 343–351.

Clemente JC, Ursell LK, Parfrey LW and Knight R (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148: 1258–1270.

Fitch WM (1970) Distinguishing homologous from analogous proteins. Systematic Zoology 19: 99–113.

Fraune S and Bosch TCG (2010) Why bacteria matter in animal development and evolution. Bioessays 32: 571–580.

Gill SR, Pop M, DeBoy RT et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359.

Gogarten JP (1994) Which is the most conserved group of proteins? Homology–orthology, paralogy, xenology, and the fusion of independent lineagesJournal of Molecular Evolution 39: 541–543.

Gokcumen O, Babb PL, Iskow RC et al. (2011) Refinement of primate copy number variation hotspots identifies candidate

genomic regions evolving under positive selection. Genome Biology 12(5): R52.
Gough E, Shaikh H and Manges AR (2011) Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection. Clinical Infectious Diseases 53(10): 994–1002.
Han K, Lou DI and Sawyer SL (2011) Identification of a genomic reservoir for new TRIM genes in primate genomes. PLoS Genetics 7(12): e1002388.
Hehemann JH, Correc G, Barbeyron T et al. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 464: 908–912.
Johnson JM, Edwards S, Shoemaker D and Schaadt AA (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends in Genetics 21: 93–102.
Johnson ME, Viggiano L, Bailey JA et al. (2001) Positive selection of a gene family during the emergence of humans and African apes. Nature 413: 514–519.
Jun J, Ryvkin P, Hemphill E, Mandoiu I and Nelson C (2009) The birth of new genes by RNA- and DNA-mediated duplication during mammalian evolution. Journal of Computational Biology 16: 1429–1444.
Katzman S, Kern AD, Bejerano G et al. (2007) Human genome ultraconserved elements are ultraselected.Science 317: 915.
Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303: 1626–1632.
Kondo N, Nikoh N, Ijichi N, Shimada M and Fukatsu T (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proceedings of the National Academy of Sciences of the USA 99: 14280–14285.
Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Lee YK and Mazmanian SK (2010) Has the microbiota played a critical role in the evolution of the adaptive immune system? Science 330: 1768–1773.
Lester CH, Frimodt-Moller N, Sorenson TL, Monnet DL and Hammerum AM (2006) In vivo transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers. Antimicrobial Agents and Chemotherapy50: 596–599.
Li WH, Gu Z, Wang H and Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409: 847–849.
Liu L, Chen X, Skogerbø G et al. (2012) The human microbiome: a hot spot of microbial horizontal gene transfer. Genomics 100: 265–270.
Lukjancenko O, Wassenaar TM and Ussery DW (2010) Comparison of 61 sequenced Escherichia coli genomes. Microbial Ecology 60: 708–720.
Mattar AF, Drongowski RA, Coran AG and Harmon CM (2001) Effect of probiotics on enterocyte bacterial translocation in vitro. Pediatric Surgery International 17: 265–268.
Meader S, Ponting CP and Lunter G (2010) Massive turnover of functional sequence in human and other mammalian genomes. Genome Research 20: 1335–1343.
Mills RE, Bennett EA, Iskow RC and Devine SE (2007) Which transposable elements are active in the human genome?Trends in Genetics 23: 183–191.
Ochman H, Worobey M, Kuo CH et al. (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biology 8(11): e1000546.
Parfrey LW and Knight R (2012) Spatial and temporal variability of the human microbiota. Clinical Microbiology and Infection 18: 5–7.
Pertea M and Salzberg SL (2010) Between a chicken and a grape: estimating the number of human genes.Genome Biology 11: 206.
Pertea M (2012) The human transcriptome: an unfinished story. Genes (Basel) 3(3): 344–360.
Pollard KS, Salama SR, King B et al. (2006) Forces shaping the fastest evolving regions in the human genome.PLoS Genetics 2: e168.
Ponjavic J, Ponting CP and Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Research 17: 556–565.
Ponting CP and Hardison R (2011) What fraction of the human genome is functional? Genome Research 21: 1769–1776.
Prosdocimi F, Linard B, Pontarotti P, Poch O and Thompson JD (2012) Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics 13: 5.
Qin J, Li R and Raes J (2010) A human gut microbial gene catalogue established by metagenomic sequencing.Nature 464: 59–65.
Rook GAW (2012) A Darwinian view of the hygiene or "old friends" hypothesis. Microbe 7(4): 173–180.
Rosenberg E, Sharon G and Zilber-Rosenberg I (2009) The hologenome theory of evolution contains Lamarckian aspects within a Darwinian framework. Environmental Microbiology 12: 2959–2962.
Salzberg SL, White O, Peterson J et al. (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science 292: 1903–1906.
Shaye DD and Greenwald I (2011) OrthoList: a compendium of C. elegans genes with human orthologs. PLoS One 6: e20085.
Siepel A, Bejerano G, Pedersen JS et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15: 1034–1050.
Smillie CS, Smith MB and Friedman J (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241–244.
Spor A, Koren O and Ley R (2011) Unravelling the effects of the environment and host genotype on the gut microbiome.Nature Reviews 9: 279–290.
The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome.Nature 437: 69–87.
Toft C and Andersson SEG (2010) Evolutionary microbial genomics insights into bacterial host adaptation. Nature Reviews 11: 465–475.
Turnbaugh PJ, Ley RE, Mahowald MA et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444(7122): 1027–1031.
Van De Peer Y, Maere S and Meyer A (2009) The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10(10): 725–732.
Walter J and Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. Annual Review of Microbiology 65: 411–429.
Zhang L, Lu HHS, Chung WY, Yang J and Li WH (2005) Patterns of segmental duplication in the human genome.Molecular Biology and Evolution 22: 135–141.

Zhang ZD, Paccanaro A, Fu Y et al. (2007) Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome Research 17: 787–797.

Further Reading

Chow J, Lee SM, Shen Y, Khosravi A and Mazmanian SK (2010) Host-bacterial symbiosis in health and disease. Advances in Immunology 107: 243–274.

ENCODE Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Human Microbiome Project (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes.Science 290(5494): 1151–1155.

Subramanian G, Adams MD, Venter JC and Broder S (2001) Implications of the human genome for understanding human biology and medicine. Journal of the American Medical Association 286: 2296–2307.

Thomas F, Barbeyron T, Tonon T et al. (2012) Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine flavobacteria to their independent transfers to marine proteobacteria and human gut bacteriodetes. Environmental Microbiology 14(9): 2379–2394.

Wolff MJ, Broadhurst MJ and Loke P (2012) Helminthic therapy: improving mucosal barrier function. Trends in Parasitology 28(5): 187–194.

2.3 Horizontal Gene Transfer: Building the Web of Life.

Horizontal gene transfer: building the web of life

Shannon M. Soucy[1], Jinling Huang[2] and Johann Peter Gogarten[1,3]

Abstract | Horizontal gene transfer (HGT) is the sharing of genetic material between organisms that are not in a parent–offspring relationship. HGT is a widely recognized mechanism for adaptation in bacteria and archaea. Microbial antibiotic resistance and pathogenicity are often associated with HGT, but the scope of HGT extends far beyond disease-causing organisms. In this Review, we describe how HGT has shaped the web of life using examples of HGT among prokaryotes, between prokaryotes and eukaryotes, and even between multicellular eukaryotes. We discuss replacement and additive HGT, the proposed mechanisms of HGT, selective forces that influence HGT, and the evolutionary impact of HGT on ancestral populations and existing populations such as the human microbiome.

Selfish genetic element
A gene or group of genes that enhance their own transmission and reproductive success without making a positive contribution to the host's fitness.

[1]*Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, Connecticut 06269–3125, USA.*
[2]*Department of Biology, East Carolina University, Greenville, North Carolina 27858, USA.*
[3]*Institute for Systems Genomics, University of Connecticut, Connecticut 06269–3125, USA.*
Correspondence to J.P.G.
e-mail: gogarten@uconn.edu
doi:10.1038/nrg3962

Horizontal gene transfer (HGT) was first described in microorganisms in the late 1940s[1], and around 20 years later it was speculated to have a role in the adaptation of multicellular eukaryotes — specifically plants[2]. Since then, methods to detect HGT have improved, and these have revealed the surprising extent and relevance of HGT to the variation of viral, prokaryotic and eukaryotic gene content. Many apparent gene duplications, for example, are now known to be the result of HGT, not autochthonous gene duplication, resulting in a 'web of life' rather than in a steadily bifurcating tree[3,4].

For a transferred gene to survive in the recipient lineage for long periods of time, the gene usually needs to provide a selective advantage either to itself (in the case of a selfish genetic element) or to the recipient, and research on HGT initially focused on such genes. However, it is now known that many of the genes that have been identified as transferred through comparative genomics between close relatives have neutral or nearly neutral effects in the recipient in both prokaryotic and eukaryotic organisms[5]. One rule for transferred genes seems to be 'first do no harm' — genes that are successfully integrated into a recipient are often expressed at low levels and encode functions at the periphery of metabolism[6]. These neutral acquisitions, however, can later provide novel combinations of genetic material for selection to act on — in some cases, the transferred material becomes domesticated over time and produces a beneficial phenotype. In other cases, when the imported genes remain neutral and there is no obvious benefit associated with their retention, the genes are likely to be lost over time.

HGT has long been recognized as an important force in the evolution of bacteria and archaea. However, the exchange of genetic information between prokaryotic symbionts and their eukaryotic hosts, and even between eukaryotes, signifies that HGT in eukaryotes occurs more frequently than previously thought[7,8]. Often these transfers involve gene donations to unicellular eukaryotes[9] and are frequently associated with bacterial endosymbionts[10] (known as endosymbiotic gene transfer (EGT) or intracellular gene transfer (IGT)). However, bacterial genes can also be transferred to multicellular eukaryotes[8]. Recent interest in the human microbiome has reinvigorated the search for HGTs from symbionts into the human genome. Although transfers of bacterial genes into the human germ line[11,12] have not be confirmed, evidence is accumulating of HGT from bacteria to human somatic cells[13]. These findings demonstrate the enduring influence of HGT on the evolution of all parts of the web of life, eukaryotes included.

In this Review, we present an overview of how HGT has contributed to innovation throughout the web of life by providing novel combinations of gene sequences for selection to act upon, thus shaping the evolution of species ranging from single-celled microorganisms to multicellular eukaryotes. Advances in the understanding of mechanisms of HGT, methods of identifying HGT events and the growth of genome databases have facilitated these insights.

Mechanisms of HGT

The three most recognized mechanisms of HGT in prokaryotes are conjugation, transformation and

transduction (FIG. 1). Conjugation requires physical contact between a donor and a recipient cell via a conjugation pilus, through which genetic material is transferred. Conjugation is canonically restricted to bacterial cells as the donor and recipient, however, *Agrobacterium* spp. is an exception and uses its conjugation machinery for HGT into plant cells[14,15]. Transformation is the uptake of exogenous DNA from the environment and has been reported in both archaea and bacteria[16,17]. Transduction is the delivery of genetic material through phage predation owing to the integration of exogenous host genetic material into a phage genome, and this phenomenon has been observed in both bacteria and archaea. There are two types of transduction: generalized, in which a random piece of the host DNA is incorporated during cell lysis; and specialized, in which a prophage imprecisely excises itself from a host genome and incorporates some of the flanking host DNAs.

Other mechanisms of gene transfer, such as gene transfer agents (GTAs) and cell fusion, have more recently been described. GTAs are gene delivery systems that are integrated into a host chromosome and are sometimes under host regulatory control. GTAs carry small random pieces of host genome in capsids for delivery to nearby hosts. GTAs are found in both bacteria and archaea. The GTA-encoding genes do not provide an obvious benefit to the host, which donates its DNA to others, nor is the benefit to the GTA-encoding genes obvious, because the GTA does not preferentially transfer the GTA-encoding genes. The question of how these genes remain under selection for function remains enigmatic[18]. One study found that GTAs from *Rhodobacter capsulatus* were able to transfer antibiotic resistance to bacteria from different phyla; however, other studies have shown that not all bacteria, including those with the genes encoding GTAs, are able to receive gene donations via GTAs[18]. GTAs have evolved from prophages that have lost the ability to target their own DNA for packaging[18]. Most GTAs cannot package a long enough segment of DNA to transfer all the genes that are necessary to produce GTAs — that is, in contrast to phages, GTAs cannot transfer all of the genes that encode them to a new host. This is an important distinction from transduction.

Cell fusion has been observed in both Euryarchaeota (*Haloferax* spp.) and Crenarchaeota (*Sulfolobus* spp.)[19,20]. Experimentally, cell fusion has been observed on solid media where *Haloferax volcanii* forms aggregates and cells become physically joined by several small bridges of fused cell membrane[21]. Bidirectional gene transfer that is mediated through cell fusion has also been observed between different *Haloferax* species[22]. The bidirectionality of this method of gene exchange means that it is more similar to sexual reproduction in eukaryotes than it is to conjugation in prokaryotes.

Circumstances that facilitate HGT in eukaryotes. The development of the nucleus sequestered genetic material in eukaryotes made gene exchange a more complicated process, although physical association over extended periods of time can facilitate HGT. Obligate endosymbiosis as a stable form of physical association often leads to the presence of foreign genes in eukaryotic genomes, as is the case for mitochondria and plastids, which are eukaryotic organelles that evolved from bacterial endosymbionts[10], and many other endosymbionts that have donated genetic material to their host genomes[23]. In the absence of an endosymbiotic partner, a congruent phylogenetic signal from multiple foreign genes has also been used to infer the presence of obsolete endosymbionts in plants and other photosynthetic eukaryotes[25,30]. Notably, however, genes of endosymbiotic origin are either absent or not obviously enriched in several eukaryotes that harbour endosymbionts[24,26], suggesting that proximity alone is not enough to ensure successful HGT.

Feeding activities are also frequently linked to gene acquisition. The mechanism of the 'you are what you eat' gene transfer ratchet proposed by W. Ford Doolittle suggests that many protists acquire genes through phagotrophy[27]. This mechanism is consistent with the findings that phagotrophic microbial eukaryotes often harbour many foreign genes[28,29].

The recently proposed weak-link model suggests that weakly protected unicellular or early developmental stages, especially in oviparous species, might constitute potential entry points for foreign genes into multicellular eukaryotes[8]. These foreign genes could then be spread through mitosis to germline cells, and thus to offspring. This model could potentially explain the fact that genes are frequently acquired in plants and animals that have eggs associated with endosymbionts or exposed to exterior environments (for example, mosses, *Drosophila* spp. and nematodes)[23,31,32].

One way that genes can be exchanged between related species is through introgression — that is, gene flow due to interspecies hybridization followed by repeated backcrosses to one of the parent species. This mechanism is a major concern in transgenic crops that are grown in proximity to non-domesticated relatives[33]. Introgression of adaptive genes is not limited to plants. For example, introgression was inferred to have introduced an allele that is important in brain development from archaic to modern humans, and this transferred allele shows signs of being under positive selection in human populations[34].

Detecting HGT

Methods for detecting HGT generally rely on phylogenetic conflict, that is, conflicting branching patterns between two gene trees; usually one of these trees is considered to be an accepted species or a reference tree. Often the reference tree is assumed to represent the vertical evolution of the organisms that are being analysed; however, detecting conflict between a gene tree and the reference tree that is not due to uncertainty in phylogenetic reconstruction is sufficient to infer the transfer of either the gene or the markers used to calculate the reference tree[35]. Deviations from the branching pattern of the reference tree identify potential HGT events, and provide information about the organisms between which genes were exchanged. Species trees are often built using well-conserved housekeeping or informational genes, such as ribosomal proteins. These genes are

Microbiome
Following a definition ascribed to Joshua Lederberg this term is most often used to denote the collective genome of the indigenous microorganisms of a multicellular or unicellular host. However, the term has also been used by Lederberg and others to signify an ecological community of commensal, symbiotic and pathogenic microorganisms.

Phylogenetic conflict
Differences between the evolutionary history of a species and the evolutionary history of its genes are embodied by discrepancies in branching order between the species and the gene tree.

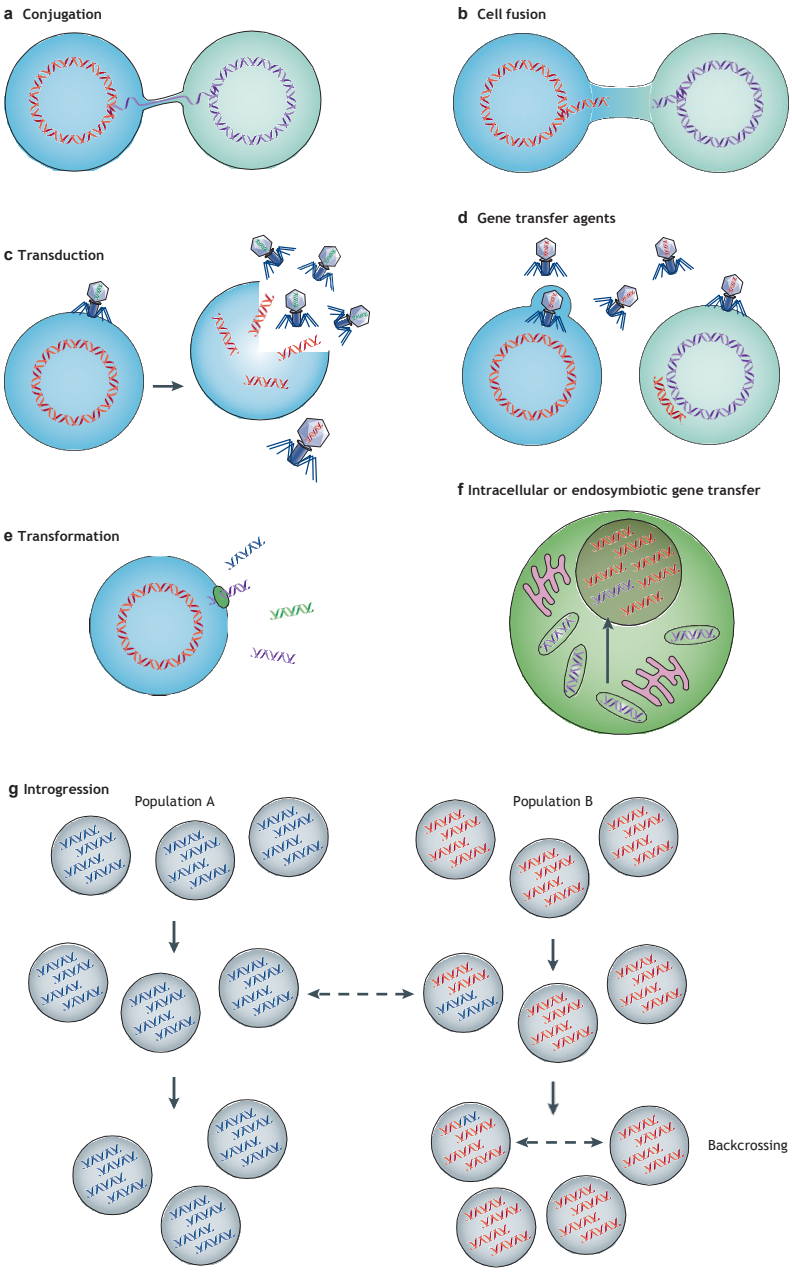

a Conjugation
b Cell fusion
c Transduction
d Gene transfer agents
e Transformation
f Intracellular or endosymbiotic gene transfer
g Introgression
Population A
Population B
Backcrossing

Figure 1 | **Mechanisms of gene transfer.** Each panel represents a method of gene transfer. Conjugation (part **a**) occurs through donor–recipient cell contact, and single-stranded DNA is transferred from the donor cell to the recipient cell. Cell fusion (part **b**) differs from conjugation in that DNA is exchanged bi-directionally after cell contact and bridge formation between two cells. Gene transfer mediated by phage is known as transduction (part **c**). In the case of generalized transduction, any piece of genomic DNA may be loaded into the phage head; a general transducing phage is shown with host DNA (red). Specialized transduction occurs when an activated prophage loads a piece of genomic DNA neighbouring the prophage genome into the phage head together with the phage DNA (not shown). Gene transfer agents (GTAs) (part **d**) are phages that no longer recognize their own DNA and only carry random fragments of host DNA. Like prophage, they reside in the host cell genome. During transformation (part **e**) DNA is taken up from the surrounding environment; in the picture the DNA is depicted as entering the cell in the double stranded form, though many DNA uptake systems degrade one of the strands upon cell entry. Intracellular or endosymbiotic gene transfer (part **f**) occurs when genetic material from an endosymbiont or organelle (such as a chloroplast or mitochondrion) is incorporated into the host genome, this mainly pertains to eukaryotes. Introgression (part **g**) occurs when a hybridization event occurs between two diverging species (orange and blue populations). Backcrosses with one of the parent populations (orange) can lead to only a small piece of the divergent genome (blue) remaining in the recipient.

transferred less frequently between divergent organisms and can thus provide a good measure of vertical ancestry. Historically, the small subunit rRNA gene (SSU rRNA) has been used to determine the prokaryotic phylogeny. This practice was suggested to be problematic because several organisms have multiple divergent rRNA operons, and it was reported that homologous recombination can occur between them (see REF. 36 for a review). Multi-locus sequence analysis (MLSA) has emerged as a supplementary method for determining prokaryotic phylogeny. The aim is to minimize the phylogenetic conflict that results from the transfer of one or more of the genes by concatenating many genes. However, if the individual genes are not screened for phylogenetic conflict caused by HGT between divergent organisms, the resulting MLSA tree might not represent either a single gene tree or the organismal evolutionary history[5]. Careful screening of genes used in an MLSA data set for significant phylogenetic conflict, and using a large number of genes (such as the suite of 50 ribosomal proteins), can help to mitigate this problem. Generally, within a phylum, phylogenetic trees that are generated using MLSA are in good agreement with those made using SSU rRNA and also provide better resolution at the species level[37,38].

Quantification of bacterial and archaeal HGT is difficult because most transfers occur between closely related organisms and are difficult to distinguish owing to the genetic similarity of the host and the recipient genomes[39–41]. As mentioned above, the canonical method for detecting HGT events uses phylogenetic conflict comparing the gene history to the species history. Substantial and statistically supported conflict in the branching patterns of the gene and species trees can identify possible gene donors or the gene exchange partners if the direction of transfer cannot be interpreted. Gene duplication followed by differential gene loss is an alternative to HGT[5]; however, the more genome sequences become available, the more independent gene loss events need to be postulated and the less parsimonious the differential gene loss scenario becomes compared with an HGT explanation. Gene composition (codon usage and oligonucleotide composition) provides a tool to identify HGT candidates[42]. Composition that is different from the genome average performs especially well to identify recent transfers from distantly related donors or from phages, which have a composition that is distinct from that of the recipient[43]. Generally, the sets of identified HGTs using each of these methods (composition or phylogenetic based) are complementary rather than redundant[44].

The comparison of genomes from closely related organisms has identified large variation in gene content within a single species, especially in prokaryotic species. This variation in genome content reflects the ongoing process of gene gain and loss. Pan-genomes have been useful for studying the evolution of gene content in both prokaryotic species and genera. The pan-genome is defined as the set of all genes present in a taxon; the accessory genome contains genes that are present in only one or a few members of the taxon; and the core genome is the set of genes present in every member of the taxon. Each individual genome thus represents a sample from the pan-genome (BOX 1). An analysis of 61 *Escherichia coli* genomes revealed that only 6% of gene families were present in all genomes[45]. Pan-genomes were originally developed to explore the fluidity of prokaryotic genomes[46]; however, because HGT is more frequent between close relatives, the pan-genome also represents the set of genes that is potentially available via HGT to any member of the group. The eukaryotic pan-genome has been less extensively studied than the prokaryotic pan-genome, possibly because the impact of HGT is less well understood and the genomes are much larger. However, the pan-genome of *Emiliania huxleyi*, a globally distributed haptophyte phytoplankton species, has been studied. Although the accessory genome accounts for approximately one-third of genes present in the reference genome *E. huxleyi* CCMP1516, much of the variation in the pan-genome is related to intron tandem repeats and exon swapping, rather than HGT[47]. These data suggest that HGTs may be less frequent or that transferred genes may be less likely to persist in eukaryotes.

HGT in evolution

Mobile selfish genetic elements promote HGT. HGT enables innovations that evolved in one group of organisms to be shared across the web of life. Many HGTs provide a selective advantage to the recipient but, as described above, some transferred genes seem to be initially neutral or nearly neutral to the recipient. HGT of self-splicing selfish genetic elements such as introns and inteins provide examples of nearly neutral mobile genetic elements. Although the self-splicing activity minimizes the cost to the host organism, the additional DNA, RNA and protein synthesis associated with the selfish genetic element provide an additional burden to the host[48]. These elements persist because their success in invading new hosts compensates for the fitness cost to the host. Once

Genome streamlining
The reduction of genome size through relaxed selection and eventual loss of loci that are superfluous to the niche occupied by the organism.

Mobilome
The aggregate of mobile genetic elements in a genome, population or environment of interest.

Genome architecture imparting sequences
Strand-biased sequence motifs that are enriched towards the termini of replication; thought to direct proteins towards the termini.

established, these elements can provide material for variation, increased complexity and innovations. For example, in *Saccharomyces cerevisiae* the HO endonuclease, which evolved from an intein, functions as a mating-type switch cleaving at the MAT locus. Split inteins have become an integral part of synthesizing the DNA polymerase in marine picocyanobacteria. The group 2 introns evolved into spliceosomal introns, which now enable alternative splicing and fine-tuned regulation in most eukaryotes (see REF. 4 for a review). Thus, HGT disseminates beneficial, neutral and nearly neutral genes; subsequent selection can act on the variations that occur in the transferred genes, leading in some cases to their integration into cellular regulatory and metabolic networks.

Selfish genetic elements are commonly involved in promoting HGT and genome rearrangements, as well as facilitating the acquisition of genes that provide a selective advantage for recipients[49]. One example is the localization of antibiotic resistance genes in compound selfish elements such as plasmids, integrative conjugative elements (ICEs) and even group 2 introns[50]. These compound structures can contain a large repertoire of genes with unrelated functions. Compound selfish elements are often associated with toxin resistance genes, metabolic genes, virulence factors and a wide range of secreted factors[50]. The acquisition of a useful gene repertoire could offset the cost of maintaining and transferring a large selfish element such as a conjugal plasmid. The traits carried on compound mobile elements can be used as a gene reservoir in times of adversity[50,51]. Genome streamlining is common in prokaryotic populations, and thus the mobility of adaptive genes associated with the mobilome becomes an important evolutionary strategy. Studies of the mobilome in different populations might provide information about the selective pressures (FIG. 2) that act on these populations and that influence gene distribution via HGT.

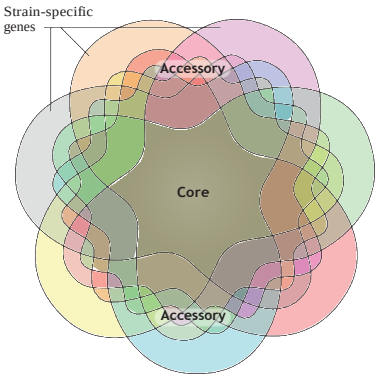
Selfish genetic elements are common in large multicellular eukaryotic genomes. Long terminal repeats often flank selfish elements and have been frequently co-opted to either increase or decrease gene expression in different tissues[52]. Syncytin genes that have a key role in trophoblast cell fusion during placenta development were repeatedly derived from retroviral envelope protein genes[52,53]. In organisms with distinct somatic and germline cells, phenotypic ingenuity often depends on the result of changes in the copy number or expression of a gene, which are often the result of selfish element dynamics in the germ line[54]. These changes can lead to divergence among or within species.

Biased gene transfer and highways of HGT. Successful HGTs frequently occur between closely related organisms[55], and the compositional similarity between the donor and the recipient genomes promotes homologous recombination that leads to homologous replacement with divergent alleles from close relatives. Additionally, the similarity between genome architecture imparting sequences in closely related organisms (same species or genera) leads to streamlined integration of the imported material[56]. In an analysis of 21 haloarchaeal genomes, over 90% of the HGTs identified through phylogenetic conflict were integrated into the recipient genome through homologous recombination[39]. The frequency of successful HGTs between pairs of Haloarchaea was shown to decrease exponentially with the phylogenetic distance (FIG. 3), probably due to the reduced efficiency of homologous recombination between genetically divergent organisms.

It was long thought that orthologous replacement through homologous recombination would be limited to the exchange of very similar gene sequences; however, the discovery of divergent isofunctional genes (known as homeoalleles) that can replace a divergent homologue in the recipient genome illustrated that homologous replacement can occur through homologous recombination in the conserved region flanking the divergent homeoalleles[40]. Divergent homeoalleles

Box 1 | **Pan-genome**

This depiction (see the figure) of the pan-genome and core genome is based on Edward's Venn cogwheel[104], and was designed by O. Zhaxybayeva, Dartmouth College, USA. The pan-genome of a group refers to the sum of all the genes that are present in members of the group. Pan-genomes comprise the core genome, which comprises the genes found in all members of a group of interest, and the accessory genome — genes that are present in only one or a few members of the group.
The concept of a pan-genome has led to the idea that steps in metabolic pathways may be distributed over several individuals within a community. The Black Queen hypothesis[105] suggests that the combination of leaky functions — genes that produce a product that is shared with others in the community — combined with a selection for small genomes, will lead to a situation in which leaky functions are encoded in the genomes of only a fraction of community members that produce this function as a common good. The pan-genomes of many taxa seem to be open (that is, of an unlimited size)[106–108], although the combination of limited population size and limited time of divergence from a common ancestor certainly limits the numbers of genes actually present in a given taxon. Estimated pan-genome sizes taking population size and divergence time into consideration can be large; for example, the *Prochlorococcus* pan-genome has been estimated to contain approximately 58,000 genes[109], whereas the individual genomes of the members of this genus encode only about 2,000 genes each.



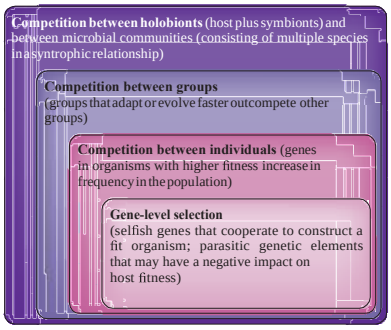


Figure 2 | **Nested levels of selection on gene content.** Each coloured box represents a different level of selection that can act on gene content.

of aminoacyl tRNA synthetases (aaRSs) provide an example of gene transfer that would go undetected by phylogenetic and compositional HGT detection methods. For many aaRSs, divergent forms evolved early in bacterial and archaeal evolution, and thus the diversity among aaRSs is easy to detect. The two or three forms with the same amino acid specificity frequently replace one another among both archaeal and bacterial species; however, because the transfers occur between related species, the gene tree of each type of aaRS remains in good agreement with the species tree[40]. Only the patchy distribution of each type reveals gene transfers and losses. Surprisingly, replacement with the divergent form was found to sometimes occur through homologous recombination in the more conserved flanking regions[40].

The frequency and bias of HGT makes it difficult to understand how adaptations might be maintained in ecological niches that are in close physical proximity[41]. At least during the initial divergence of ecotypes, genes are transferred between organisms that are adapted to different niches. It is possible that the higher frequency of within-ecotype HGT than between-ecotype HGT maintains ecotype adaptation. However, genes that adapt an organism to a particular niche are also transferred between niche boundaries[57], and such HGTs might help recipients to integrate into a new ecological niche (FIG. 4).

HGT enables key metabolic innovations. The enormous pan-genome size of many microbial species illustrates the importance of additive gene transfer, which is the process of the integration of novel genetic material into a genome. Integration into the genome can occur by non-homologous recombination or through homologous recombination involving the genes neighbouring the transferred gene (for example, see REF. 41). An additive transfer from a close relative of a gene that has an orthologue in the recipient genome leads to two similar copies being present in the recipient genome, an outcome that is similar to a gene duplication[4]. The methylaspartate cycle, for example, combines genes from several bacterial metabolic pathways that were transferred to the haloarchaeal ancestor from different bacterial donors and incorporated into a novel pathway for carbon assimilation[58]. Other examples of HGT contributing to the assembly or extension of metabolic pathways are acetoclastic methanogenesis in *Methanosarcina* spp. and the assembly of two photosystems functioning in series in oxygen-producing photosynthesis (see discussion in REF. 4 for details). In addition to frequently exchanging genes within and between genera, Haloarchaea also exchange genes with bacteria[39,59]. Haloarchaea are aerobic heterotrophs, although they evolved from methanogens — an anaerobic chemolithotrophic lineage. More than 1,000 genes were identified as imports from bacteria into Haloarchaea, including those for carbon assimilation, respiratory chain complexes, membrane transporters and cofactor biosynthesis[59]. The influx of these bacterial genes allowed the haloarchaeal ancestor to move into an aerobic environment. Similarly, the influx of bacterial genes to the ancestors of 12 other major archaeal clades is thought to have provided the key innovations to the origin of these groups[60]. Debate continues about whether the transferred genes originated from one or a few donors over a short period of time, or whether these transfers involved diverse bacterial donors[112,113]. The limited distribution of these genes within single groups of archaea indicates that 'highways' of gene sharing between archaea and bacteria have promoted archaeal diversity.

HGT and the evolution of the holobiont

Many organisms rely on a complex network of symbionts for functions ranging from defence and immunity to metabolism. The symbiotic communities that are associated with larger macro-organisms provide an initial interface with the environment, thus new properties and physiological responses often occur through HGT involving these communities. The holobiont[61] is used as a collective term for the host and its associated microbiota. For many multicellular eukaryotes, the number of genes in the microbiome[62] (genes that are present in the microbiota) dwarfs the number of genes in the nuclear genome of the host and provides an important source of genetic diversity.

The composition of human gut microbiota is affected by the diet and ecology of the human host, and by competition between members of the microbiota[62]. For example, bacteria in the gut of Japanese people can break down polysaccharides from the cell walls of seaweeds that are commonly present in the Japanese diet. The genes encoding the polysaccharide-digesting enzymes were transferred from parasites of marine algae to the gut bacteria[63,64]. This HGT has enabled Japanese people to use carbohydrates from algal cell walls as a nutrient source, whereas other populations cannot. It is tempting to interpret this as selection acting on the holobiont; however, it is more likely to reflect gut bacteria evolving to fill an available ecological niche (FIG. 2).

Ecotypes
Genetically distinct subsets of organisms within a population or species, usually genetic differences correspond to niche adaptation.

Holobiont
A multicellular or unicellular host and its collective symbionts.

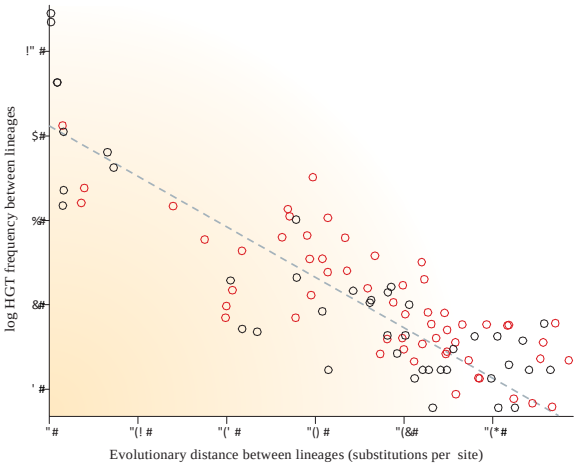


Figure 3 | **HGT is more frequent between closely related species.** The frequency of horizontal gene transfer (HGT) events in haloarchaea is plotted against evolutionary distance. Gene transfers were detected through phylogenetic conflict between the gene's phylogeny and the reference phylogeny calculated from ribosomal proteins. HGTs between terminal edges of the reference phylogeny are shown in black and those between internal edges are shown in red. Similar inverse log-linear relationships between recombination rate and divergence were also observed for bacterial genera. Reprinted from Williams, D., Gogarten, J. P. and Papke, R. T. Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol. Evol. (2012). **4**, 1223–1244 by permission of Oxford University Press.

The results of recent research on the human microbiome have demonstratedthe importance of the microbiota in nutrient acquisition and immune defence in humans. In an analysis that investigated recent gene transfers in the human microbiome, HGT was shown to be 25-fold more frequent between pairs of human-associated organisms than between pairs of organisms in different hosts or in aquatic or terrestrial environments[49]. Moreover, HGT between pairs of human-associated organisms isolated from the same body site are 50-fold more likely to exchange genes than pairs from other environments[49]. The surprising extent of gene transfer in human microbiota compared with other environments could indicate that environmental fluctuations that promote frequent adaptive changes are more prevalent in holobiont ecology, especially in the human holobiont. Notably, however, quantification of HGT is difficult, and sampling bias between environments (in that particular study, for example, 53% of the samples were of the human holobiont and the remaining 47% were split between aquatic, terrestrial and other host-associated environments[49]) could falsely inflate the rate of detection of HGT in well-sampled environments (humans) compared with that in environments with less available data.

HGT in eukaryotic evolution

Although still fragmented, the available data indicate that HGT is widespread in all major eukaryotic groups and has been ongoing throughout evolutionary time[7,8,65]. As stated above, the sequestration of genetic material to the nucleus requires distinct mechanisms for HGT in eukaryotes. Nevertheless, HGT is important in conferring beneficial phenotypes that may lead to the origin of major lineages. Furthermore, changes brought about by HGT may prompt the adaptive radiation of other groups through organismal interactions and genetic integration in a co-evolving web of life.

HGT in the origin of plastids and Plantae. The plant lineage is ripe with examples of HGTs that have conferred novel functions (FIG. 5). Plastids, the hallmark of photosynthetic eukaryotes, are derived from cyanobacterial endosymbionts in a eukaryotic host. With the only exception of chromatophores in amoeboid *Paulinella* spp., the well-founded belief is that all other photosynthetic eukaryotes trace their plastids to a single cyanobacterial endosymbiosis[66]. The transformation of a free-living cyanobacterium into a permanent organelle required both genetic and metabolic integration between the two partners. Several analyses identified 20–50 genes from chlamydiae, a group of obligate intracellular bacteria, in various photosynthetic eukaryotes[30,67,68]. These findings led to the suggestion that cyanobacterial and chlamydial endosymbionts coexisted in an early eukaryotic host cell, and that this tripartite relationship was responsible for the transformation of cyanobacterial endosymbionts into modern-day plastids[30,67,69,70]. Although it has been argued that these chlamydiae-related genes could have resulted from phylogenetic artefacts or could have existed in the cyanobacterial progenitor of plastids[71–73], some of these genes are only adaptive in parasitic or heterotrophic bacteria and are not found in extant cyanobacteria, suggesting that chlamydial involvement in plastid establishment is plausible[30,67,68,74]. Non-cyanobacterial prokaryotes other than chlamydiae also contributed genes for plastid genesis and functionality[69,75–77].

The establishment of cyanobacterial endosymbionts or plastids triggered the origin of Plantae: red algae, glaucophytes and green plants. Recent investigations have indicated that all three of these lineages have been affected by HGT during their evolution[69,78–80]. The glaucophyte *Cyanophora paradoxa* acquired more than 400 genes from bacteria[69]. In red algae, HGTs contributed to at least 5% of protein-coding genes in *Galdieria sulphuraria* and many others in *Porphyridium purpureum*[78,80]. Evidence of HGT has also been found in green algae[79] and land plants[81,82] (see below). For example, the moss *Physcomitrella patens* acquired genes from various sources, including fungi, bacteria, viruses and aquatic animals[32,83,84]. In most of these cases, acquired genes expanded the metabolic capabilities of recipients and had a key role in their adaptation to new environments, such as those with high salinity or acidity, extreme temperatures, or toxic substances.

HGT between plants and other eukaryotes. The origin of plastids and Plantae also spawned the emergence of other photosynthetic eukaryotes through secondary or higher-level endosymbioses. In addition, Plantae, which are rich in complex carbohydrates, generated new niches and resources for other organisms to exploit. Particularly, plant cell walls are the most abundant biomass on earth. Both the prevalence and novelty of this insoluble stored energy enhanced adaptive pressure to take advantage of novel resources free of competition. To effectively utilize plant biomass, other organisms often share genes or metabolic capabilities. For example, numerous soil bacteria reside in the rhizosphere and rely on root exudes as their primary nutrient source. An increase in exude production leads to active bacterial growth and thus more frequent plasmid transfer among rhizobacteria[85]. Choanoflagellates and rotifers, both of which live in aquatic environments, acquired numerous genes from plants and miscellaneous algae[86,87], frequently related to complex carbohydrate degradation[28]. In rumen ciliates, 46 genes related to the degradation of complex carbohydrates, such as plant biomass, were acquired by HGT, many of them from the gut bacteria of ruminant animals[88]. Beyond choanoflagellates and rumen ciliates, the ability to degrade plant biomass has been independently acquired by many other eukaryotic groups such as oomycetes, fungi and nematodes[89,90]. The widespread and diverse mechanisms for degrading complex carbohydrates in plants in so many different lineages highlight the convergent evolution through HGT for adaptation.

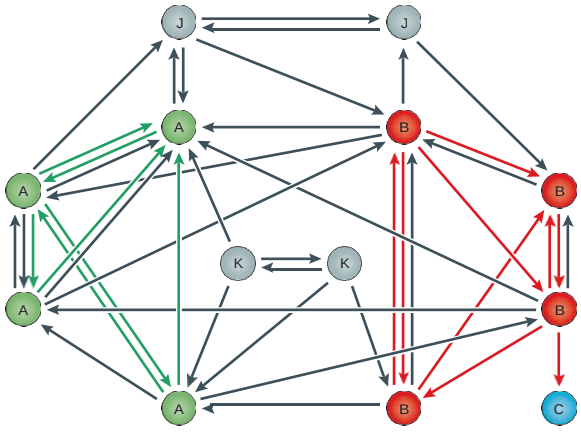


Figure 4 | **Structured exchange community.** Prokaryotic members of two distinct niches are shown as green and red circles (A and B); grey circles (K and J) are related species occupying different niches. Genes that enable the adaptation of their hosts to these niches are mostly exchanged between members of the same niche (green and red arrows), but they might also be shared with recent niche invaders (blue circle; C), accelerating the adaptation of the invader to a new habitat. Adapted with permission from REF. 57, (AAAS).

Lepidopterans are the largest group of plant-feeding insects, and their diversification coincided with the emergence of flowering plants. In an analysis of HGT in lepidopteran insects, most of the acquired genes were shown to be distributed in multiple lepidopteran groups and related to nutritional metabolism and detoxification[91]. The production of toxins by plants and the corresponding genes for detoxification in lepidopterans, and other phytophagous arthropods, exemplifies a genetic 'arms race' fuelled by HGT. Many plants can produce cyanogenic glucosides, which can be converted to highly toxic hydrogen cyanide as a defence against herbivores. Conversely, phytophagous arthropods not only sequester hydrogen cyanide as a defence against their own predators, but also counteract cyanide poisoning through detoxification genes that were originally recruited from bacteria[92].

HGT between multicellular eukaryotes. Many cases of HGT were reported between parasitic plants and their hosts[93–96]. In almost all of these cases, the direction of HGT is consistent with the direction of nutrient transfer from the host to the parasitic plant. HGT also occurs between multicellular eukaryotes with less obvious physical associations. For example, the moss *P. patens* acquired an actinoporin gene that is involved in desiccation resistance from metazoans[83]. *Alloteropsis* grasses switched to C_4 photosynthesis at least four times in the past 10 million years through the acquisition of genes from other C_4 grasses[97]. A photoreceptor gene was transferred from hornworts to ferns, allowing modern ferns to thrive in low-light conditions under the canopy[98]. Sturgeons, lampreys, which have been known to feed on sturgeons, and paddle fishes all share a transposable element, probably the result of HGT mediated by the exchange of fluids during lamprey feeding[99]. The sporadic distribution of type II antifreeze protein (AFP) genes in herring, smelt and sea raven was also mediated by HGT, allowing these fish to adapt to icy water[31].

For a long time, mitochondria were considered uniparentally inherited and subject to Muller's ratchet[100]. For many groups of organisms, this assumption seems to be correct[101]; however, plant, algal and fungal mitochondrial genomes are known to be dynamic and promiscuous, varying greatly among species in structure and gene content[102]. The transfer of mitochondrial genes between plant species can be massive and widespread. In an extreme case, *Amborella trichopoda*, a basal flowering plant, acquired at least four whole mitochondrial genomes from mosses and green algae, as well as many mitochondrial and, to a lesser degree, plastidal fragments from other flowering plants[103]. This example of HGT is not known to be associated with an adaptive benefit and is instead an important example of neutral or nearly neutral gene transfer in eukaryotes.

The mode of HGT between multicellular eukaryotes remains controversial. Are individual genes transferred, or are the transfers the consequence of between-species hybridization followed by backcrosses to one of

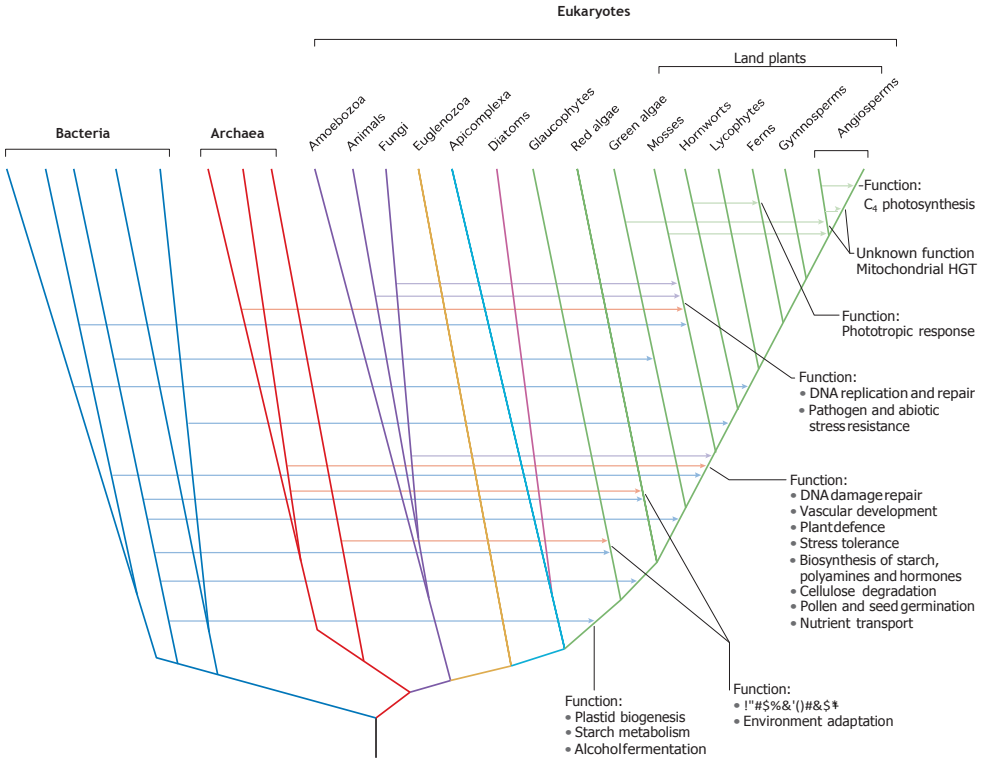


Figure 5 | **HGT to the plant lineage.** Arrows are coloured based on the origin of the gene transferred. Lines at the tips of the arrows indicate the gain of function for the plant lineage that acquired the genetic material. HGT, horizontal gene transfer. Figure modified from REF. 32, Nature Publishing Group.

the parents[7]? In many instances, such as the transfer of AFP genes from herring to smelt[31], donor and recipient diverged more than 200 million years ago, making hybridization an unlikely scenario. The conservation of introns between donor and recipient argues against independent transfers from bacterial symbionts. Sperm-mediated gene transfer between fish is one possible scenario[31]. In the moss *P. patens*, eggs and embryos that are exposed to bacteria and fungi in the environment might have facilitated gene acquisition. The large-scale acquisitions of mitochondrial genes in *Amborella trichopoda* probably occurred through mitochondrial genome fusion mediated by regenerated meristems from wounded areas.

Perspective

In this Review, we have discussed examples that illustrate how HGT shapes gene content in bacteria, archaea and unicellular eukaryotes (see Supplementary information S1 (table)). Even in multicellular eukaryotes, HGT from symbionts and between mitochondria occurs frequently and can have an important impact on gene content. Currently, we have a good understanding of the mechanisms by which prokaryotes exchange genes, including through GTAs and cell fusion in archaea; however, the mechanisms by which multicellular eukaryotes exchange genes with one another and with prokaryotes are less clear. The weak-link model, sperm-mediated gene transfer and introgression are possible gene transfer pathways, but more work is needed to explore the specific mechanisms involved. Importantly, comparisons between closely related strains will lead to a more accurate characterization of HGTs. Improvements in HGT detection based on the growing collection of sequence data will result in a more realistic estimation of HGT rates. However, accounting for false negatives and various types of transfer over different phylogenetic distances remains a challenge. Nevertheless, the surprising density of the web of life woven through genetic exchange is becoming visible.

1. Tatum, E. L. & Lederberg, J. Gene recombination in the bacterium *Escherichia coli*. *J. Bacteriol.* **53**, 673–684 (1947).
2. Went, F. W. Parallel evolution. *Taxon* **20**, 197–226 (1971).
3. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
4. Swithers, K. S., Soucy, S. M. & Gogarten, J. P. The role of reticulate evolution in creating innovation and complexity. *Int. J. Evol. Biol.* **2012**, 418964 (2012).
5. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3**, 679–687 (2005).
6. Park, C. & Zhang, J. High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4**, 523–532 (2012).
This paper examines the impact of expression level on the transferability of a gene in both environmental and laboratory populations of *!'#$%&'*
7. Boto, L. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. Biol. Sci.* **281**, 20132450 (2014).
8. Huang, J. Horizontal gene transfer in eukaryotes: the weak-link model. *Bioessays* **35**, 868–875 (2013).
This letter proposes a model for ongoing HGT in eukaryotes involving unicellular and early developmental stages to overcome the barrier of genome sequestration in eukaryotes.
9. Andersson, J. O. Gene transfer and diversification of microbial eukaryotes. *Annu. Rev. Microbiol.* **63**, 177–193 (2009).
10. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135 (2004).
11. Koonin, E. V. Darwinian evolution in the light of genomics. *Nucleic Acids Res.* **37**, 1011–1034 (2009).
12. Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. & Micklem, G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol.* **16**, 50 (2015).
13. Riley, D. R. *et al.* Bacteria–human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS Comput. Biol.* **9**, e1003107 (2013).
14. Norman, A., Hansen, L. H. & Sørensen, S. J. Conjugative plasmids: vessels of the communal gene pool. *Phil. Trans. R. Soc. B* **364**, 2275–2289 (2009).
15. Kyndt, T. *et al.* The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: an example of a naturally transgenic food crop. *Proc. Natl Acad. Sci. USA* **112**, 201419685 (2015).
16. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J.-P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).
17. Chimileski, S., Dolas, K., Naor, A., Gophna, U. & Papke, R. T. Extracellular DNA metabolism in *Haloferax volcanii*. *Front. Microbiol.* **5**, 57 (2014).
18. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* **10**, 472–482 (2012).
19. Naor, A. & Gophna, U. Cell fusion and hybrids in Archaea: prospects for genome shuffling and accelerated strain development for biotechnology. *Bioengineered* **4**, 126–129 (2013).
20. Schleper, C., Holz, I., Janekovic, D., Murphy, J. & Zillig, W. A multicopy plasmid of the extremely thermophilic archaeon *Sulfolobus* effects its transfer to recipients by mating. *J. Bacteriol.* **177**, 4417–4426 (1995).
21. Mevarech, M. & Werczberger, R. Genetic transfer in *Halobacterium volcanii*. *J. Bacteriol.* **162**, 461–462 (1985).
22. Naor, A., Lapierre, P., Mevarech, M., Papke, R. T. & Gophna, U. Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr. Biol.* **22**, 1444–1448 (2012).
23. Dunning Hotopp, J. C. *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).
This paper describes HGT between *(%&)*5+'*#spp",#* an intracellular bacterial symbiont, and its multicellular eukaryotic insect hosts.
24. Nikoh, N. *et al.* Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genet.* **6**, e1000827 (2010).
25. Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724–1726 (2009).
26. Chapman, J. A. *et al.* The dynamic genome of *Hydra*. *Nature* **464**, 592–596 (2010).
27. Doolittle, W. F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
28. Yue, J., Sun, G., Hu, X. & Huang, J. The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate *Monosiga brevicollis*. *BMC Genomics* **14**, 729 (2013).
29. Grant, J. R. & Katz, L. A. Phylogenomic study indicates widespread lateral gene transfer in entamoeba and suggests a past intimate relationship with parabasalids. *Genome Biol. Evol.* **6**, 2350–2360 (2014).
30. Huang, J. & Gogarten, J. P. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol.* **8**, R99 (2007).
This paper discusses a complex tripartite relationship between a eukaryotic host, a cyanobacterium and a chlamydia that may have facilitated the establishment of modern plastids.
31. Graham, L. A., Li, J., Davidson, W. S. & Davies, P. L. Smelt was the likely beneficiary of an antifreeze gene laterally transferred between fishes. *BMC Evol. Biol.* **12**, 190 (2012).
32. Yue, J., Hu, X., Sun, H., Yang, Y. & Huang, J. Widespread impact of horizontal gene transfer on plant colonization of land. *Nat. Commun.* **3**, 1152 (2012).
33. Stewart, C. N., Halfhill, M. D. & Warwick, S. I. Transgene introgression from genetically modified crops to their wild relatives. *Nat. Rev. Genet.* **4**, 806–817 (2003).
34. Evans, P. D., Mekel-Bobrov, N., Vallender, E. J., Hudson, R. R. & Lahn, B. T. Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage. *Proc. Natl Acad. Sci. USA* **103**, 18178–18183 (2006).
35. Williams, D. *et al.* A rooted net of life. *Biol. Direct* **6**, 45 (2011).
36. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
37. Colston, S. M. *et al.* Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *mBio* **5** e02136-14 (2014).
38. Delamuta, J. R. M., Ribeiro, R. A., Menna, P., Bangel, E. V. & Hungria, M. Multilocus sequence analysis (MLSA) of *Bradyrhizobium* strains: revealing high diversity of tropical diazotrophic symbiotic bacteria. *Braz. J. Microbiol.* **43**, 698–710 (2012).
39. Williams, D., Gogarten, J. P. & Papke, R. T. Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* **4**, 1223–1244 (2012).
40. Andam, C. P. & Gogarten, J. P. Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* **9**, 543–555 (2011).
41. Polz, M., Alm, E. & Hanage, W. Horizontal gene transfer and the evolution of bacterial. **29**, 170–175 (2015).
This paper investigates the interplay between HGT, population structure and lineage divergence in bacteria and archaea.
42. Langille, M. G. I., Hsiao, W. W. L. & Brinkman, F. S. L. Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* **8**, 373–382 (2010).
43. Daubin, V. & Ochman, H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* **14**, 1036–1042 (2004).
44. Ragan, M. A. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* **201**, 187–191 (2001).
45. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708–720 (2010).
46. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
47. Read, B. A. *et al.* Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**, 209–213 (2013).
48. Barzel, A., Obolski, U., Gogarten, J. P., Kupiec, M. & Hadany, L. Home and away — the evolutionary dynamics of homing endonucleases. *BMC Evol. Biol.* **11**, 324 (2011).
49. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
This letter investigates the frequency of HGT in the human microbiome across body sites and across continents.
50. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why? *Hered. (Edinb.).* **106**, 1–10 (2011).
This paper investigates the types of traits that are associated with compound selfish genetic elements and investigates the ecological scenarios that would select for specific types of traits.
51. Broaders, E., Gahan, C. G. M. & Marchesi, J. R. Mobile genetic elements of the human gastrointestinal tract: potential for spread of antibiotic resistance genes. *Gut Microbes* **4**, 271–280 (2013).
52. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).
53. Cornelis, G. *et al.* Ancestral capture of *syncytin-Car1*, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proc. Natl Acad. Sci. USA* **109**, E432–E441 (2012).
54. Schaack, S., Gilbert, C. & Feschotte, C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25**, 537–546 (2010).
55. Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli–Shigella* genetic exchange communities. *Open Biol.* **2**, 120112 (2012).
56. Hendrickson, H. & Lawrence, J. G. Selection for chromosome architecture in bacteria. *J. Mol. Evol.* **62**, 615–629 (2006).
57. Papke, R. T. & Gogarten, J. P. Ecology. How bacterial lineages emerge. *Science* **336**, 45–46 (2012).
58. Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J. & Berg, I. A. A methylaspartate cycle in haloarchaea. *Science* **331**, 334–337 (2011).
59. Nelson-Sathi, S. *et al.* Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of haloarchaea. *Proc. Natl Acad. Sci. USA* **109**, 20537–20542 (2012).
60. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2014).
This paper suggests that acquisitions of genes from bacteria lead to the evolution of the major clades in archaea.
61. Guerrero, R., Margulis, L. & Berlanga, M. Symbiogenesis: the holobiont as a unit of evolution. *Int. Microbiol.* **16**, 133–143 (2013).
62. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837–848 (2006).
63. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
64. Thomas, F. *et al.* Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine Flavobacteriia to their independent transfers to marine Proteobacteria and human gut *Bacteroides*. *Environ. Microbiol.* **14**, 2379–2394 (2012).
65. Hirt, R. P., Alsmark, C. & Embley, T. M. Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. *Curr. Opin. Microbiol.* **23**, 155–162 (2015).
66. McFadden, G. I. Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6**, a016105 (2014).
67. Ball, S. G. *et al.* Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell* **25**, 7–21 (2013).
68. Moustafa, A., Reyes-Prieto, A. & Bhattacharya, D. Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS ONE* **3**, e2205 (2008).
69. Price, D. C. *et al. Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* **335**, 843–847 (2012).
70. Cenci, U. *et al.* Transition from glycogen to starch metabolism in archaeplastida. *Trends Plant Sci.* **19**, 18–28 (2014).
71. Deschamps, P. Primary endosymbiosis: have cyanobacteria and Chlamydiae ever been roommates? *Acta Soc. Bot. Pol.* **83**, 291–302 (2014).

72. Ku, C. *et al.* Endosymbiotic gene transfer from prokaryotic pangenomes: inherited chimerism in eukaryotes. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1421385112 (2015).
73. Domman, D., Horn, M., Embley, T.M. & Williams, T.A. Plastid establishment did not require a chlamydial partner. *Nat. Commun.* **6**, 6421 (2015).
74. Ball, S. G. *et al.* Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. *Biochim. Biophys. Acta* **1847**, 495–504 (2015).
75. Huang, J. & Gogarten, J. P. Concerted gene recruitment in early plant evolution. *Genome Biol.* **9**, R109 (2008).
76. Suzuki, K. & Miyagishima, S. Y. Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. *Mol. Biol. Evol.* **27**, 581–590 (2010).
77. Qiu, H. *et al.* Assessing the bacterial contribution to the plastid proteome. *Trends Plant Sci.* **18**, 680–687 (2013).
78. Schonknecht, G. *et al.* Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207–1210 (2013).
79. Blanc, G. *et al.* The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* **13**, R39 (2012).
80. Bhattacharya, D. *et al.* Genome of the red alga *Porphyridium purpureum*. *Nat. Commun.* **4**, 1941 (2013).
81. Yue, J., Hu, X. & Huang, J. Origin of plant auxin biosynthesis. *Trends Plant Sci.* **19**, 764–770 (2014).
82. Yang, Z. *et al.* Ancient horizontal transfer of transaldolase-like protein gene and its role in plant vascular development. *New Phytol.* **206**, 807–816 (2015).
83. Hoang, Q. T. *et al.* An actinoporin plays a key role in water stress in the moss *Physcomitrella patens*. *New Phytol.* **184**, 502–510 (2009).
84. Maumus, F., Epert, A., Nogue, F. & Blanc, G. Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **5**, 4268 (2014).
85. Molbak, L., Molin, S. & Kroer, N. Root growth and exudate production define the frequency of horizontal plasmid transfer in the rhizosphere. *FEMS Microbiol. Ecol.* **59**, 167–176 (2007).
86. Sun, G., Yang, Z., Ishwar, A. & Huang, J. Algal genes in the closest relatives of animals. *Mol. Biol. Evol.* **27**, 2879–2889 (2010).
87. Boschetti, C. *et al.* Biochemical diversification through foreign gene expression in Bdelloid Rotifers. *PLoS Genet.* **8**, e1003035 (2012).
88. Ricard, G. *et al.* Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* **7**, 22 (2006).
89. Paganini, J. *et al.* Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes. *PLoS ONE* **7**, e50875 (2012).
90. Richards, T. A. *et al.* Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl Acad. Sci. USA* **108**, 15258–15263 (2011).
91. Li, Z. W., Shen, Y. H., Xiang, Z. H. & Zhang, Z. Pathogen-origin horizontally transferred genes contribute to the evolution of Lepidopteran insects. *BMC Evol. Biol.* **11**, 356 (2011).
92. Wybouw, N. *et al.* A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *eLife* **3**, e02365 (2014).
93. Yoshida, S., Maruyama, S., Nozaki, H. & Shirasu, K. Horizontal gene transfer by the parasitic plant *Striga hermonthica*. *Science* **328**, 1128 (2010).
94. Xi, Z. *et al.* Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics* **13**, 227 (2012).
95. Zhang, Y. *et al.* Evolution of a horizontally acquired legume gene, *albumin 1*, in the parasitic plant *Phelipanche aegyptiaca* and related species. *BMC Evol. Biol.* **13**, 48 (2013).
96. Zhang, D. *et al.* Root parasitic plant *Orobanche aegyptiaca* and shoot parasitic plant *Cuscuta australis* obtained Brassicaceae-specific *strictosidine synthase*-like genes by horizontal gene transfer. *BMC Plant Biol.* **14**, 19 (2014).
97. Christin, P.-A. *et al.* Adaptive evolution of Ç photosynthesis through recurrent lateral gene transfer. *Curr. Biol.* **22**, 445–449 (2012).
98. Li, F.-W. *et al.* Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proc. Natl Acad. Sci. USA* **111**, 6672–6677 (2014).
99. Zhang, H.-H., Feschotte, C., Han, M.-J. & Zhang, Z. Recurrent horizontal transfers of Chapaev transposons in diverse invertebrate and vertebrate animals. *Genome Biol. Evol.* **6**, 1375–1386 (2014).
100. Blanchard, J. L. & Lynch, M. Organellar genes. *Trends Genet.* **16**, 315–320 (2000).
101. Lynch, M. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol. Biol. Evol.* **13**, 209–220 (1996).
102. Palmer, J. D. *et al.* Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl Acad. Sci. USA* **97**, 6960–6966 (2000).
103. Rice, D. W. *et al.* Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* **342**, 1468–1473 (2013).
This paper reports the acquisition of several mitochondrial genomes by the mitochondria in -.)%/0&&*##1/'$+%2%3#1a basal flowering plant.
104. Edwards, A. W. F. *Cogwheels of the Mind: The Story of Venn Diagrams* (JHU Press, 2004).
105. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036-12 (2012).
This paper explains how the interplay between cheating and selection for streamlined genomes can give rise to shared genomic resources.
106. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Estimation of prokaryotic supergenome size and composition from gene frequency distributions. *BMC Genomics* **15**, S14 (2014).
107. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).
108. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* **12**, 66 (2014).
109. Baumdicker, F., Hess, W. R. & Pfaffelhuber, P. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.* **4**, 443–456 (2012).
110. Hooper, L. V. & Gordon, J. I. Commensal host–bacterial relationships in the gut. *Science* **292**, 1115–1118(2001).
111. Lederberg, J. & McCray, A. 'Ome sweet 'omics — a genealogical treasury of words. *Scientist* **15**, 8 (2001).
112. Becker, E, A. *et al.* Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response.. *PLoS Genet.* **10**, e1004784 (2014).
113. Groussin, M *et al.* Origins of major archaeal clades do not correspond to gene acquisitions from bacteria. *BioRxiv* http://dx.doi:10.1101/019851(2015).

Acknowledgements
Work in the authors' laboratories was supported through grants from the National Science Foundation Grant (DEB 0830024), NASA Exobiology (NNX13AI03G), Binational Science Foundation (BSF 2013061), NSFC Oversea, Hong Kong, Macao collaborative grant (31328003), and the CAS/SAFEA International Partnership Program for Creative Research Teams. The authors would also like to thank K. Swithers for providing insightful comme nts and discussion pertaining to the body of this text.

Competing interests statement
The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
See online article: S1 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF

2.4 The Pan-Genome as a Shared Genomic Resource: Mutual Cheating, Cooperation and the Black Queen Hypothesis.

frontiers in Microbiology

OPINION
published: 21 July 2015
doi: 10.3389/fmicb.2015.00728

The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis

Matthew S. Fullmer [1], Shannon M. Soucy [1] and Johann Peter Gogarten [1,2]*

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA, [2] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

Keywords: pan-genome, black queen hypothesis, red queen hypothesis, social cheating, gene transfer

OPEN ACCESS

Edited by:
Luis Delaye,
Centro de Investigación y de Estudios Avanzados - Unidad Irapuato, Mexico

Reviewed by:
Luis David Alcaraz,
Universidad Nacional Autónoma de México, Mexico
Luisa I. Falcon,
Universidad Nacional Autónoma de México, Mexico

*Correspondence:
Johann Peter Gogarten,
gogarten@uconn.edu;
jpgogarten@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 27 April 2015
Accepted: 03 July 2015
Published: 21 July 2015

Citation:
Fullmer MS, Soucy SM and Gogarten JP (2015) The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. Front. Microbiol. 6:728. doi: 10.3389/fmicb.2015.00728

Cells without Complete Genomes

Cells have long been recognized as life's building blocks (e.g., Virchow's dictum "*omnis cellula e cellula,*" Virchow, 1860). Specifically, a cell's genome is considered the repository of genetic information that pairs with the cellular machinery to determine the organism's phenotype. Except for rare circumstances, the majority of a genome is passed on from ancestor to descendant, although the acquisition of genes from organisms that are not direct ancestors is recognized to play an important role in evolution (Swithers et al., 2012).

Jeffrey Lawrence, in discussing minimal genome size proposed a meta-cell model (Lawrence, 1999), in which many micelles (small vesicles containing resources, products, and genes) exchange genes frequently. Genes temporarily reside in a micelle and direct the synthesis of compounds important for replication. A micelle only can replicate when all compounds necessary for division have been generated. However, at each point in time only a fraction ofthe necessary genes are present in an individual micelle. This model relies on gene transfer being so frequent that each of the genes that encode necessary functions visits the individual micelles often enough to allow for sufficient synthesis ofthe necessary gene products for future micelle divisions. The meta-cell can be considered an organism, whose genome is divided into a network of micelles. Lawrence's meta-cell model is reminiscent of Woese's progenote (Woese, 1998) and Kandler's pre-cell populations (Kandler, 1994) that were postulated to have existed early in evolution before genes coalesced into genomes.

The Pan-Genome as a Shared Genomic Resource

For most bacterial and archaeal species different strains contain non-overlapping gene sets. The pan-genome of a taxon or group refers to the sum of all genes that are present in members ofthe group (Tettelin et al., 2005; Lapierre and Gogarten, 2009). Pan-genomes comprise the core genome, i.e., the genes that are found in all members, and the accessory genome, i.e., genes that are present in only one or a few members ofthe group. Welch et al. (2002) provided the first illustration that genome content in bacteria changes rapidly. Comparing three *Escherichia coli* strains they found the shared core to be less than 40% ofthe gene families present in all three genomes. More recently the size ofthis core was further reduced to only 6% of gene families present in 61 *E. coli* genomes (Lukjancenko et al., 2010). Baumdicker et al. (2012) estimate that the *Prochlorococcus* pan-genome contains about 58,000 genes, whereas the individual genomes encode only about 2000 genes each.

The pan-genome concept was originally developed to explore the fluidity of prokaryotic genomes (Tettelin et al., 2005). Because HGT is more frequent between close relatives (Andam and Gogarten, 2011), the pan-genome may also represent the set of genes that is potentially available via HGT to any member ofthe group. The function ofthe pan-genome may then be thought of as a shared resource. This is supported by the observation that genes encoding weakly selected functions are frequently lost from bacterial genomes, when they do not provide selective advantages, only to be re-acquired through HGT, when new conditions provide a selective advantage to carriers (Lawrence and Roth, 1996). The idea ofthe pan-genome of a population as a shared genomic resource is similar to the description of meta-cells and pre-cell populations. In particular, these concepts have in common that the individual genome of a cell or micelle does not represent a sufficient description ofth e genomic resources ofthe population. The following paragraphs discuss some factors that contribute to the large size of pan-genomes.

The Strong Black Queen Hypothesis

The black queen hypothesis proposed by Morris et al. (2012) is built on the premise of "leaky" common good functions, which cannot be restricted to benefit only the producer. The hypothesis suggests that these functions combined with selection for small genomes may lead to a situation in which these leaky functions are encoded in only a fraction ofthe genomes comprising the community. Under the black queen hypothesis a cell's evolution can follow one oftwo pathways (see **Figure 1**): (1) the cell can retain all genes encoding leaky functions (in the game ofhea rts, from which the name for the black queen hypothesis derives, this strategy is known as "shooting the moon"). The cost is a large genome maintaining and expressing many genes that are not essential to central metabolism, growth, and reproduction. Consequently, maintaining these genes and expression of extra proteins competes for cellular resources that could be put toward replication and results in a lower growth rate (Dong et al., 1995; Scott et al., 2010; Weiße et al., 2015). The advantage of the "shooting the moon" strategy is that following a population bottleneck all genes encoding leaky functions are available in the genome. These members of a community following this strategy may be thought of as analogous to a keystone species. (2) The cell looses some or all ofits leaky functions and increases i ts growth rate (in hearts, this represents the usual strategy oft aking as few point cards as possible). Traditionally this is described as cheating, as the second strategy relies upon other cells in the population for the leaky functions it has lost. If a bottleneck occurs, a single cheating cell is unlikely to survive on its own. A possible outcome of all cells in a population following strategy

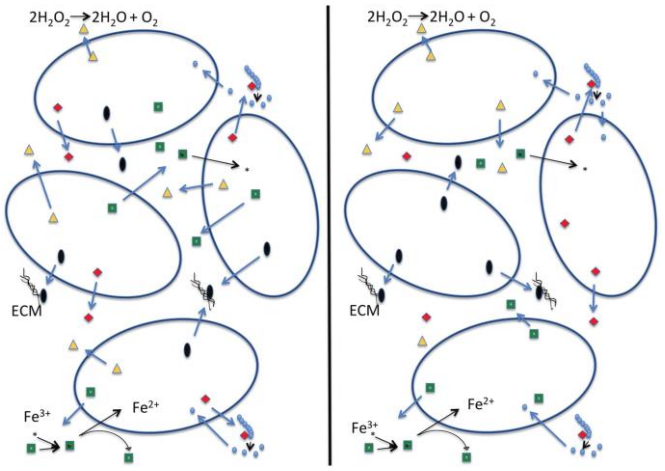


FIGURE 1 | Schematic depiction of a strong version of the black queen hypothesis (Morris et al., 2012). Each cell in the left panel contains all the genes that produce the four depicted common goods. Ce lls on the right, due to selection to minimize genome size, each produc e only one of the common goods, and rely on the presence of other cells to pr oduce the other ones. We consider this a strong version of the black que en hypothesis because it no longer contains exclusive helper strains; rat her all individuals in the population described in the right hand panel are simultaneously helpers and beneficiaries. In this example dark open ovals represent in dividual cells; red diamonds: hydrolases that digest an extracellular poly mer (e.g., phosphatase, sialidase, glucanase); green squares: sider ophores that allow cells to acquire iron; black ovals: building blocks of the ex tracellular matrix (ECM) or enzymes that produce and assemble these building bl ocks; yellow triangles: enzymes that destroy oxygen radicals (e.g., cat alase—reaction given; or peroxidase—reaction: ROOR′ + electron donor (2 e $^-$) + $2H^+$ ROH + R′OH).

#2 is that all members of a population cheat on some leaky functions. The members in the population then become mutually dependent on one another (**Figure 1**). In this scenario there are no keystone members providing all ofthe leaky functions. For the population to establish itselfin a new environment several members ofthe population are required for the migrati on to be successful, as no single cell has all the components necessary to sustain itself. We term this the "strong" version ofthe black queen hypothesis. If all members of a population follow the second strategy, this may under some conditions lead to instability, the tragedy ofthe commons, and extinction o f the population; however, experimental work by Morris et al. (2014) has shown that partitioning of a leaky common-goods function can enable the stable co-existence oftwo very simi lar organisms that use the same resources. Additionally, under natural conditions cells do rarely exist in homogeneous mixture (Davey and O'toole, 2000). Cells existing in biofilms or small aggregates are likely to be proximal to cells with which they share recent ancestry, and therefore proximal cells will have the same genotype with respect to shared functions. Drescher et al. (2014) show that *Vibrio cholera* can avoid the public goods dilemma by strengthening relationships between cells ofthe same genotype through creation of a thick biofilm, thereby providing a local selective advantage to producers of a particular common good in case this good becomes an overall limiting resource. It seems likely that genes encoding common goods are under frequency dependent selection, leading to local feedback loops that contribute to a long-term co-existence ofthe different t ypes of cheaters.

Black vs. the Red Queen

Bacteria are under severe predation by phage (Thurber, 2009). They need to constantly change to evade predation, hence the analogy to the red queen from Lewis Carroll's (Carroll and Gardner, 1999) *Through the Looking-Glass*, who needs to run as fast as she can just to stay in place (Van Valen, 1973). The analysis of phage metagenomes and rank abundance curves indicated that the phage predation follows the *kill the winner* strategy (Hoffmann et al., 2007), where successful strains are targeted more frequently. The surprising long term stability of species composition despite phage predation suggests that cycling between different susceptible target cells occurs within a population and not between populations from different species (Rodriguez-Brito et al., 2010). Consequently, within a population, host genes that encode receptors utilized by phage and virus to enter the cell are expected to turn over quickly, creating within population diversity (Chaturongakul and Ounjai, 2014).

Random Acquisition of Genes

Genes are constantly acquired by genomes, and many of the transferred genes do not find a long term home in the recipient genome (Lawrence and Ochman, 1997). Among these genes are parasites (prophages) and selfish genetic elements. Most, but certainly not all (Lobkovsky et al., 2013), ofthe transferred genes are selectively neutral or nearly neutral to the recipient (Gogarten and Townsend, 2005; Baumdicker et al., 2010; Haegeman and Weitz, 2012). Though these genes may not find long term homes in the genomes they "visit," selfish genes especially can affect the rates of gene sharing and thus the size ofthe pan-genome in a population. Furthermore, many selfish elements induce genome rearrangements that can promote the loss and gain of genes, and thus may have a significant impact on the initiation ofthe loss ofleaky functions.

Generation of paralogs may play a role in facilitation of loss ofleaky functions. Additional copies increase gene dos age, ameliorating the loss offunction in other members ofthe population by providing more ofthe common good. However, the pressure to delete genes from genomes is much stronger than to duplicate them (Mira et al., 2001) and an increase in gene transcription can have a similar or greater effect on the overall expression level (Weiße et al., 2015). Regardless of whether the increased production comes from paralogy or regulation it would need to be countered by a greater decrease in production from other common good functions to overcome the cost ofincreased protein expression.

Conclusion

Random acquisition of genes and selfish genetic elements, selection by predators, and cheating on common goods, all undoubtedly play a role in generating diversity within populations ofbacteria and archaea. The conjecture ofthe strong black queen hypothesis is that mutual cheating leads to mutual dependencies and therefore cooperation. Under this hypothesis individual cells would be integrated into a meta-organism, whose genome is the pan-genome ofthe population, similar to Lawrence's meta-cells whose genome is distributed over individual micelles.

The pan-genome of a population as shared genomic resource could explain part ofthe "genome of Eden" paradox (Doolittle et al., 2003), where estimations of ancestral genomes are far larger and more complex than those of any extant individual genome. Large estimates of archaeal ancestors' genome sizes (Csurös and Miklós, 2009; Wolf et al., 2012) could actually represent the pan-genome ofthe ancestral population rather than any single cell. Ifthis is the case, the complexity ofthe progenitor cells in a lineage/population might often be at a similar level of complexity as their extant relatives. We hypothesize the large estimates of progenitor genome size might in part reflect a "strong" black queen scenario where genome variation creates a large pan-genome, but no single cell contains a "keystone genome" with all genes in the population represented. More extensive studies ofindividual and population genomes, and rates of within population transfer are needed to confirm that master genomes, encoding all the leaky functions needed for survival ofthe population, can be and often are absent from a population.

Ifthe hypothesis ofthe population pan-genome as a shared genomic resource is borne out, then the scientific community will need to continue to increase its appreciation for the import

of pan- and meta-genomes. Likewise, we may need to more seriously consider populations as the operative units in which genes are selected in rather than exclusively individual organisms. Similar to how Richard Dawkins (1976) advocated thinking of an organism as a collection of generally agreeable, but selfish, genes perhaps we should be thinking oflineages and populations as the collections of genes, i.e., pan-genomes, rather than the individual cells.

Acknowledgments

We thank Michael Stephens, Joerg Graf, and Thane Papke for Discussion. Work in the Gogarten-lab was supported through grants from the National Science Foundation (AToL DEB0830024), the National Aeronautics and Space Agency (Exobiology NNX13AI03G), and the Bi-national Science Foundation (BSF 2013061).

References

Andam, C. P., and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* 9, 543–555. doi: 10.1038/nrmicro2593

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. (2010). The diversity of a distributed genome in bacterial populations. *Ann. Appl. Probab.* 20, 1567–1606. doi: 10.1214/09-AAP657

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome ofbacteria. *Genome Biol. Evol.* 4, 443–456. doi: 10.1093/gbe/evs016

Carroll, L., and Gardner, M. (1999). *The Annotated Alice: The Definitive Edition*. New York, NY: Norton.

Chaturongakul, S., and Ounjai, P. (2014). Phage-host interplay: examples from tailed phages and Gram-negative bacterial pathogens. *Front. Microbiol.* 5:442. doi: 10.3389/fmicb.2014.00442

Csurös, M., and Miklós, I. (2009). Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.* 26, 2087–2095. doi: 10.1093/molbev/msp123

Davey, M. E., and O'toole, G. A. (2000). Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.* 64, 847–867. doi: 10.1128/MMBR.64.4.847-867.2000

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Dong, H., Nilsson, L., and Kurland, C. G. (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *J. Bacteriol.* 177, 1497–1504.

Doolittle, W. F., Boucher, Y., Nesbø, C. L., Douady, C. J., Andersson, J. O., and Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos. Trans. R. Soc. B Biol. Sci.* 358, 39–58. doi: 10.1098/rstb.2002.1185

Drescher, K., Nadell, C. D., Stone, H. A., Wingreen, N. S., and Bassler, B. L. (2014). Solutions to the public goods dilemma in bacterial biofilms. *Curr. Biol.* 24, 50–55. doi: 10.1016/j.cub.2013.10.030

Gogarten, J. P., and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687. doi: 10.1038/nrmicro1204

Haegeman, B., and Weitz, J. S. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* 13:196. doi: 10.1186/1471-2164-13-196

Hoffmann, K. H., Rodriguez-Brito, B., Breitbart, M., Bangor, D., Angly, F., Felts, B., et al. (2007). Power law rank-abundance models for marine phage communities. *FEMS Microbiol. Lett.* 273, 224–228. doi: 10.1111/j.1574-6968.2007.00790.x

Kandler, O. (1994). "The early diversification oflife," in *Early Life on Earth*, ed S. Bengston (New York, NY: Columbia University Press), 152–509.

Lapierre, P., and Gogarten, J. P. (2009). Estimating the size ofth e bacterial pan-genome. *Trends Genet.* 25, 107–110. doi: 10.1016/j.tig.2008.12.004

Lawrence, J. G. (1999). "Gene transfer and minimal genome size," in *Size Limits of Very Small Microorganisms*, eds A. Knoll, M. J. Osborn, J. Baross, H. C. Berg, N. R. Pace, and M. Sogin (Washington, DC: National Research Council), 32–38.

Lawrence, J. G., and Ochman, H. (1997). Amelioration ofbacterial ge nomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397. doi: 10.1007/PL00006158

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

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2013). Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* 5, 233–242. doi: 10.1093/gbe/evt002

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* 60, 708–720. doi: 10.1007/s00248-010-9717-3

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution ofbacterial genomes. *Trends Genet.* 17, 589–596. doi: 10.1016/S0168-9525(01)02447-7

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The Black Queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* 3:e00036-12. doi: 10.1128/mbio.00036-12

Morris, J. J., Papoulis, S. E., and Lenski, R. E. (2014). Coexistence of evolving bacteria stabilized by a shared black queen function. *Evolution* 68, 2960–2971. doi: 10.1111/evo.12485

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J.* 4, 739–751. doi: 10.1038/ismej.2010.1

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. *Science* 330, 1099–1102. doi: 10.1126/science.1192588

Swithers, K. S., Soucy, S. M., and Gogarten, J. P. (2012). The role of reticulate evolution in creating innovation and complexity. *Int. J. Evol. Biol.* 2012:418964. doi: 10.1155/2012/418964

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Thurber, R. V. (2009). Current insights into phage biodiversity and biogeography. *Curr. Opin. Microbiol.* 12, 582–587. doi: 10.1016/j.mib.2009.08.008

Van Valen, L. (1973). A new evolutionary law. *Evol. Theory* 1, 1–30.

Virchow, R. L. K. (1860). *Cellular Pathology (Google eBook)*. London: John Churchill. Available online at: https://play.google.com/store/books/details/Rudolf_Ludwig_Karl_Virchow_Cellular_Pathology?id=JUth7ntb0_AC

Weiße, A. Y., Oyarzún, D. A., Danos, V., and Swain, P. S. (2015). Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1038–E1047. doi: 10.1073/pnas.1416533112

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., et al. (2002). Extensive mosaic structure revealed by the

complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 99, 17020–17024. doi: 10.1073/pnas.252529799

Woese, C. (1998). The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* 95, 6854–6859. doi: 10.1073/pnas.95.12.6854

Wolf, Y. I., Makarova, K. S., Yutin, N., and Koonin, E. V. (2012). Updated clusters of orthologous genes for Archaea: a complex ancestor ofthe Archae a and the byways ofhorizontal gene transfer. *Biol. Direct* 7:46. doi: 10.1186/1745-6150-7-46

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict ofinterest.

Chapter 3. Evolution and Distribution of the mobile *vma1*-b Intein.

This chapter consists of a paper that I was second author on. I was involved in the discussion that lead to most of the analyses, built the extein and intein phylogenies, and performed the GARD analysis to look for recombination breakpoints. Dr. Kristen Swithers primarily wrote this paper, and Peter Gogarten supervised the work and writing. This paper is the first time that an intein sequence was used to examine networks of gene transfer, and the methods in this paper are expanded on in later analyses. This paper shows that inteins are primarily shared between closely

related sequences, but can be shared across domains as well.

This is a pre-copied author produced pDF of an article accepted for publication in MBE following peer review. The definitive publisher-authenticated version of Swithers K S, Soucy S M, Lasek-Nesselquist E, Lapierre P, and Gogarten JP Distribution and evolution of the Mobile *vma*-1b Intein (2012) Molecular Biology and Evolution is available online at http://mbe.oxfordjournals.org/content/30/12/2676.long.

MBE Advance Access published October 18, 2013

Distribution and Evolution of the Mobile *vma-1*b Intein

Kristen S. Swithers,[y,1] Shannon M. Soucy,[1] Erica Lasek-Nesselquist,[z,1] Pascal Lapierre,[§,2] and Johann Peter Gogarten*[,1]

[1]Department of Molecular and Cell Biology, University of Connecticut

[2]University of Connecticut Biotechnology Center, University of Connecticut

[y]Present address: Department of Cell Biology, Yale University Medical School, New Haven, CT

[z]Present address: Department of Biology, University of Scranton, Scranton, PA

[§]Present address: New York State Department of Health, Wadsworth Center, Albany, NY

*Corresponding author: E-mail: jpgogarten@gmail.com, gogarten@uconn.edu.

Associate editor: Tal Pupko

Abstract

Inteins are self-splicing parasitic genetic elements found in all domains of life. These genetic elements are found in highly conserved positions in conserved proteins. One protein family that has been invaded by inteins is the vacuolar and archaeal catalytic ATPase subunits (*vma-1*). There are two intein insertion sites in this protein, "a" and "b." The b site was previously thought to be only invaded in archaeal lineages. Here we survey the distribution and evolutionary histories of the b site inteins and show that the intein is present in more lineages than previously annotated, including a bacterial lineage, *Mahella australiensis* 50-1 BON. We present evidence, through ancestral character state reconstruction and substitution ratios between host genes and inteins, for several transfers of this intein between divergent species, including an interdomain transfer between the archaea and bacteria. Although inteins may persist within a single population or species for long periods of time, transfer of the *vma-1*b intein between divergent species contributed to the distribution of this intein.

Key words: intein, parasitic gene, homing endonuclease, gene transfer, coevolution of genes.

Introduction

Inteins are self-splicing, parasitic, mobile genetic elements that splice out of their flanking proteins sequence (extein) after translation via an autocatalytic mechanism (Perler et al. 1994; Cooper and Stevens 1995). Inteins can be divided into two different groups based on their size—large and mini inteins—which are differentiated based on the presence or absence of a homing endonuclease (HE) domain (Liu 2000; Elleuche and Poggeler 2010). The large inteins contain a HE domain, an enzyme with endonuclease activity that has a 14–40 bp recognition site (Chevalier and Stoddard 2001), and N- and C-terminal splicing domains (Pietrokovski 1998). This HE domain provides mobility to the intein and the ability to move into new inteinless target sites through a process called homing (Chevalier and Stoddard 2001; Gogarten and Hilario 2006). Transfer of the intein to hosts that contain an inteinless target site is necessary for the HE to remain under purifying selection (Goddard and Burt 1999). The transfer of the HE-containing inteins can occur either within or between species (Koufopanou et al. 2002; Okuda et al. 2003; Gogarten and Hilario 2006; Yahara et al. 2009; Barzel et al. 2011; Clerissi et al. 2013; Macgregor et al. 2013). The mini intein is simply composed of the N and C-terminal splicing domains with a linker region between them.

Inteins are found in all three domains of life and in viruses and tend to be present in conserved regions of proteins (exteins) with high sequence similarity (Swithers et al. 2009). Targeting conserved sites in conserved proteins is advantageous for two reasons. First, it provides selective pressure on the splicing domains, which ensures the intein is spliced out exactly to maintain a functional extein. Mutations to the splicing domains of the intein will result in improper splicing, resulting in a nonfunctional extein that could be detrimental to the cell. Second, targeting the most conserved regions of the most conserved genes will facilitate transfer to inteinless alleles across species or even domain boundaries.

One such conserved protein family invaded by inteins is the vacuolar and archaeal catalytic ATPase subunits (*vma-1*) (Hirata et al. 1990; Kane et al. 1990; Senejani et al. 2001). The vacuolar, archaeal, and bacterial ATPases (or ATP synthases) are a family of multisubunit proteins that are present in all three domains of life and share a common ancestor (Zimniak et al. 1988; Gogarten et al. 1989). The eukaryotic version, called the vacuolar ATPase (V-type), is involved in the intravesicular acidification of the endomembrane system (vacuoles, lysosome, endosomes, and trans Golgi network), and it is also found in the plasma membrane of specialized cells of transport epithelia (Harvey and Nelson 1992). The archaeal (A-type) and bacterial (F-type) counterparts are used for ATP generation and/or ion transport. Several cases of horizontal gene transfer of the A-type ATPase to bacteria have been documented (Hilario and Gogarten 1993; Lapierre et al. 2006; Lapierre 2007).

Two distinct sites in this family of ATPases have been invaded by inteins; the "a" site in the vacuolar ATPases in yeasts and the "b" site in the archaea. These two insertion sites have been shown to be in the most conserved parts of the protein (Swithers et al. 2009). The intein in the a site was frequently transferred between different yeast species (Koufopanou et al. 2002; Okuda et al. 2003). The intein in the b site has a wide phylogenetic distribution among archaea. To date, the intein database (InBase) annotates seven *vma-I* b inteins in the *Thermoplasma* and *Pyrococcus* (Perler 2002) genera. Here we report on additional *vma-I* b inteins, show that they are found in more diverse species, and discuss their evolutionary histories.

Results

Distribution of *vma-I* b Inteins

Databases at the National Center for Biotechnology Information (NCBI) were surveyed for *vma-I* exteins and inteins that reside in the b insertion position (*vma-I* b). To date, we identified 15 *vma-I* b inteins within complete flanking extein sequences present in these databases (table 1). Thirteen are found within the Euryarchaeota, one is present in Clostridia and another is found in a deep-sea hypersaline metagenomic sequencing project. Additionally, we identified two inteins in contigs from a marine metagenome (gi:129952466) and a hot spring metagenome (gi:290482983) that only encode partial extein sequences. These partial sequences did not contain sufficient information for phylogenetic placement of the host organisms. The multiple sequence alignment shows the *vma-I* b intein family is composed of both mini inteins and larger inteins with HE domains (see supplementary data, Supplementary Material online, alignment of intein sequences). The larger inteins all have the LAGLIDADG HE domain, suggesting these may be active HEs. However, additional experimental evidence is required to verify their activity. Compared with the currently annotated inteins in InBase, we report 10 additional inteins at this insertion site in three new classes of prokaryotes.

Breakpoints within the Extein/Intein Alignment

The genetic algorithm for recombination detection (GARD) determined breakpoints in the extein/intein alignment. This algorithm analyzes fragments of the overall alignment and compares the goodness of fit of phylogenies from these smaller alignments under a maximum likelihood (ML) framework using the corrected Akaike information criterion. Significant breakpoints were identified three amino acids before the insertion site (position 245 $P < 0.01$) and at three amino acids before the end of the intein (position 891 $P < 0.01$) (supplementary fig. S1, Supplementary Material online). The fact that the GARD algorithm detects breakpoints three amino acids upstream of the insertion site might be due to the conserved, phylogenetically uninformative nature of the splice site. The last three amino acids at the carboxy terminal end of the intein are conserved in all the inteins included in this study. This level of conservation is more typical for the extein than for the intein sequences (see histograms for conservation in supplementary fig S8, Supplementary Material online). Regardless of the slight misplacement of the breakpoint, the GARD analysis does suggest that extein and intein have different evolutionary histories, which is also corroborated by the intein and extein trees (fig. 1).

Archaeal-Type ATPase Distribution

A survey of the bacterial domain reveals that all major clades of bacteria have representatives containing an archaeal-type ATPase (uninvaded extein sequence). A phylogenetic analysis of the extein protein using representatives from all clades shows three bacterial clusters interspersed within the archaea (supplementary fig. S2, Supplementary Material online). The phylogeny of the archaeal exteins is in good agreement with the ribosomal protein phylogeny (see later). Thus, the finding that the bacteria are interspersed within the archaea suggests that the three distinct bacterial groups (supplementary fig. S2, Supplementary Material online) represent independent gene transfer events from archaea to bacteria.

Archaeal Extein and Ribosome Trees

Overall, the extein and ribosomal protein reference phylogeny display a high degree of topological similarity, particularly at the ordinal level. Many clades show identical branching orders, such as the Methanosarcinales, Methanococcales, and Methanobacteriales (supplementary fig. S4, Supplementary Material online). Deeper phylogenetic relationships were also similar, including the sister relationship between the Archaeoglobi and a Halobacteria-Methanomicrobia clade. Less resolved positions on the tree tended to stem from taxa on long branches, prone to artifact, such as *Micrarchaeum* and *Nitrosopumilis* in the ribosomal phylogeny.

The reduced extein and ribosomal protein trees were congruent (supplementary fig. S5, Supplementary Material online) with the exception of the "misplacement" of *Candidatus* Nanosalinarum sp. in the ribosome phylogeny due to Long Branch Attraction (LBA). Both Shimodaira-Hasegawa (SH) and approximately unbiased (AU) tests indicated that the extein and ribosomal topologies were not significantly different from each other ($P = 0.28$ and 0.73 for AU and SH tests, respectively). This suggests that the archaeal exteins follow ribosomal evolution (considered to be vertical evolution) and have not been transferred between divergent archaeal lineages. This finding also implies that transfers of the intein occurred independently of any extein transfers.

Ancestral Character State Reconstruction.

Maximum parsimony and ML ancestral state reconstructions (ASRs) were performed using the phyletic patterning of the intein relative to its host gene across the prokaryotes (fig. 2). There was only one most parsimonious reconstruction with nine steps given the data set, which is validated by the MLASR analysis. The reconstructions suggest seven independent gains and two losses of the intein. According to the parsimony and the ML reconstruction, the Thetis sea sequence (probability from MLASR $P = 100\%$), the ancestor of the four *Pyrococcus* spp. ($P = 99\%$), *Thermococcus litoralis* ($P = 100\%$),

Table 1. Distribution of *vma-1*b

Organisms	Lineage	AA Length of Intein	Isolated from	Reference
Pyrococcus furiosus DSM 3638	Archaea; Euryarchaeota; Thermococci	426	Shallow marine solfatara at Vulcano Island off southern Italy	(Fiala and Stetter 1986)
P. abyssi GE5	Archaea; Euryarchaeota; Thermococci	428	Hydrothermal vent in the North Fiji Basin in the Pacific Ocean	(Erauso et al. 1993)
Pyrococcus sp. NA2	Archaea; Euryarchaeota; Thermococci	428	Hydrothermal vent in Papua New Guinea-Australia-Canada-(PACMANUS) field	(Lee et al. 2011)
P. horikoshii OT3	Archaea; Euryarchaeota; Thermococci	375	Hydrothermal vent in the Okinawa Trough in the Pacific Ocean	(Gonzalez et al. 1998)
Thermococcus litoralis DSM 5473	Archaea; Euryarchaeota; Thermococci	428	Shallow submarine hot spring at Lucrino Beach near Naples	(Neuner et al. 1990)
Thermoplasma acidophilum 122-1B2 ATCC 25905	Archaea; Euryarchaeota; Thermoplasmata Indiana,	174	Self-heating coal refuse pile from the Friar Tuck mine in US A	(Searcy 1975; Senejani et al. 2001)
Thermop. acidophilum DSM 1728 Indiana,	Archaea; Euryarchaeota; Thermoplasmata	173	Self-heating coal refuse pile from the Friar Tuck mine in US A	(Darland et al. 1970)
Thermop. volcanium GSS1	Archaea; Euryarchaeota; Thermoplasmata	185	Solfataric field	(Segerer et al. 1988)
Thermoplasmatales archaeon I-plasma Iron	Archaea; Euryarchaeota; Thermoplasmata	179	Ultraback A location (UBA site) in the Richmond Mine, Mountain, CA	(Dick et al. 2009)
Ferroplasma sp. Type II Iron	Archaea; Euryarchaeota; Thermoplasmata	535	5-way CG acid mine drainage site in the Richmond Mine, Mountain, CA	(Tyson et al. 2004)
Picrophilus torridus DSM 9790	Archaea; Euryarchaeota; Thermoplasmata	332	Dry solfataric field in northern Japan	(Schleper et al. Á̂Ñ̂erson et al.
Methanothermus fervidus DSM 2088	Archaea; Euryarchaeota; Methanobacteria	357	Hot solfataric spring from Iceland	(Ñarasingarao et al.
Candidatus Nanosalinarum sp.	Archaea; Euryarchaeota;	522	Hypersaline lake NW Victoria, Australia	2012)
Thetis Sea contig00086	Metagenome gi number 354855569	521	Deep-sea hypersaline anoxiclake Thetis	(Ferrer et al. 2012)
Mahella australiensis 50-1 BON	Bacteria; Firmicutes;	522	Riversiea oil field in the Bowen-Surat basin, Australia	(Salinas et al. 2004)

FIG. 1. Extein and intein phylogenetic trees. Phylogenetic trees depicting the extein (left) and intein (right). Support values (from left to right) were determined by posterior probability, approximate likelihood ratio test, and bootstrap replicates. Supported incongruencies are found between the extein and intein tree in the Pyrococcus/Thermococcus clade.

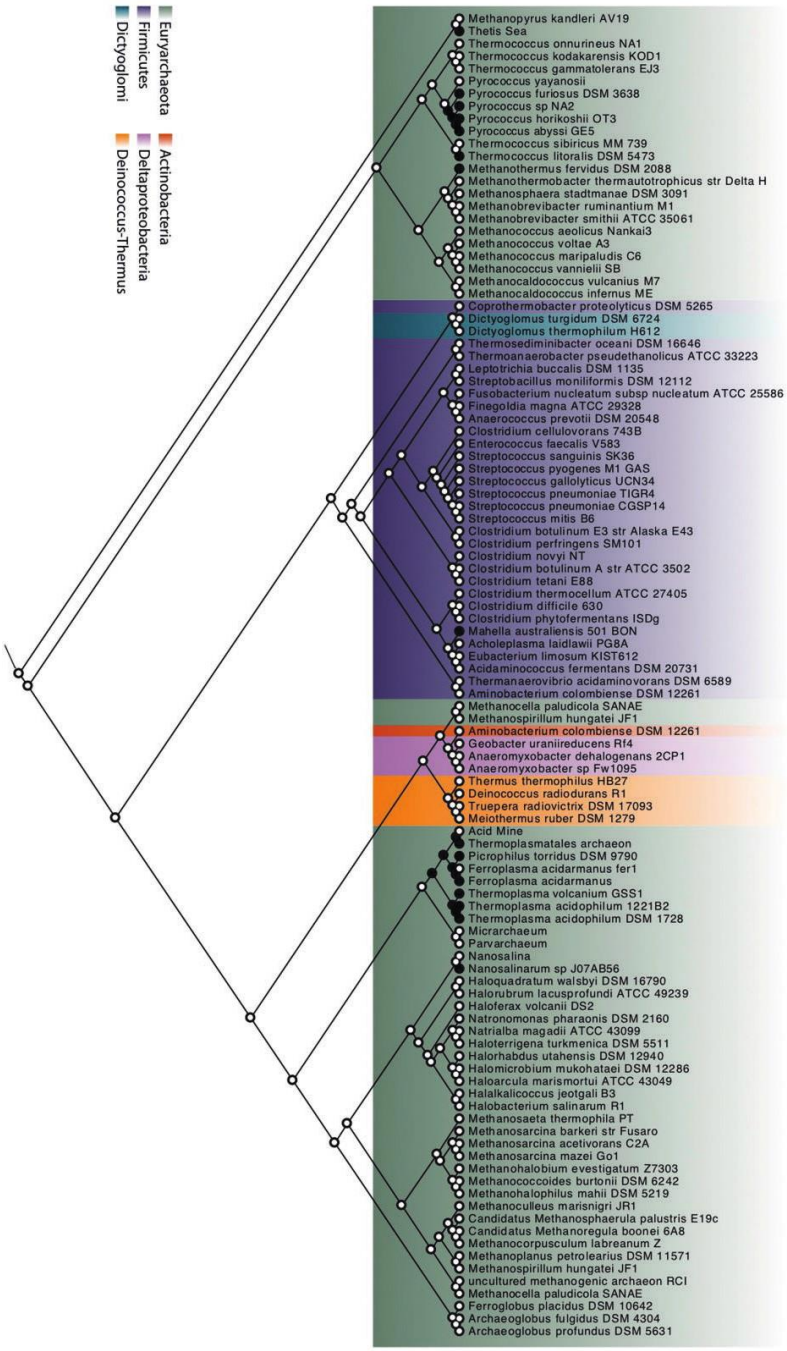


Fig 2. Maximum parsimony ancestral character state reconstruction. The depicted cladogram represents the part of the VMA-1 tree where inteins are present (see supplementary fig. S2, Supplementary Material online, for full tree). Presence and absence of the intein are represented at the tips of the tree; black circles correspond to the presence of intein and white circles correspond to the absence of the intein. The single most parsimonious reconstruction with nine steps indicated seven independent gains and two losses of the intein. The gains are once for the Thetis sea sequence, once for the *Pyrococcus* spp., once for the *Thermococcus litoralis*, once for *Methanothermus fervidus*, once for *Mahella australiensis*, once for the Thermoplasmatales, and once for *Candidatus* Nanosalinarum sp. J07AB56. The two losses are in the acid sequence and *Ferroplasma acidarium* fer1. A maximum likelihood reconstruction using the MK1 model (Lewis 2001) produced nearly identical results (see supplementary fig. S3, Supplementary Material online).

M. australiensis (*P* = 100%), *Candidatus* Nanosalarium (*P* = 100%), *M. fervidus* (*P* = 100%), and the ancestor of the *Thermoplasma*/*Ferroplasma*/*Picrophilus*/Thermoplasmatales clade (*P* = 70%) all independently gained the intein. See the Discussion for consideration of the overestimation of intein absence in the species represented by the leaves of the cladogram in figure 2.

Multiple Horizontal Gene Transfers of the *vma-I* b Intein

Phylogenetic reconstructions were performed for the extein and intein splicing domains. These trees are incongruent with each other by both the AU and SH tests (*P* : 0.01). This significant incongruence is due to a well-supported transfer of the intein from within the *Pyrococcus* spp. to *T. litorialis* (fig. 1). Unfortunately, none of the other discrepancies between the extein and intein trees show high bootstrap support. Including the HE domain in reconstruction of the intein's phylogeny did not improve the placement of the bacterial intein sequence from *Mahella australiensis* relative to the *Methanothermus* and *Picrophilus*/*Ferroplasma* inteins (supplementary fig. S6, Supplementary Material online). The ratio of extein to intein substitution rates provides additional insight on intein transfers relative to exteins.

Differences in divergence rates (Novichkov et al. 2004) and higher than expected sequence similarity (e.g., Podell and Gaasterland 2007) are frequently used to infer horizontal gene transfer (HGT) events. Immediately after an intein has been transferred between two species, the intein sequences are identical between donor and recipient and more conserved than the extein. This unexpected high sequence similarity of two inteins is an indcation of transfer of an intein (Liu et al. 1997). Moreover, the extein and intein evolve under very different selection pressures. Although the ATPase catalytic subunit belongs to one of the most conserved protein families (Gogarten 1994; Swithers et al. 2009), the inteins evolve much faster making sequence-based phylogenetic reconstruction difficult (Perler et al. 1997; Gogarten et al. 2002). This is reflected in the conservation profiles of the extein and intein (supplementary fig. S7, Supplementary Material online). The intein has a significantly higher average sequence variation score than the extein, 3.42 and 2.85, respectively ($P < 0.01$, *t* test). The higher score also indicates a higher substitution rate, because it represents the number of different amino acids in a sliding window. Among site rate variation for the extein sequences is more extreme than for the intein sequences, that is, the exteins contain many sites under strong purifying selection (supplementary fig. S8, Supplementary Material online). To assess the divergence rate of the inteins relative to extein sequences, we used the well-aligned splicing domains of the intein, omitting the less conserved linker and HE domains (see Alignments, supplementary data, Supplementary Material online). The phylogenetic reconstructions of the extein and intein data sets are the same for the *Thermoplasma* genus and the Thermoplasmatales I-plasma archaeon, suggesting that in this group (in the following designated as *Thermoplasma* group) the extein and intein were inherited together. The inteins in these sequences (table 1) also have lost their HE domain, which suggests they are no longer mobile and further strengthens the notion that the intein was inherited together with the extein for this clade. Therefore, the pairwise ratios of the substitution rates of extein to intein in this clade should represent ratios expected under coinheritance of extein and intein. Ratios significantly greater than these (i.e., the intein is less divergent than expected) likely indicate horizontal transfer of the intein relative to the extein (see table 2 for ratios). The chosen cutoff level of 0.7 corresponds to a significance level of $P < 0.002$ given the rate ratios within the *Thermoplasma* group. To assess false-positive rates obtained using this ratio approach, we simulated sequence evolution along the extein tree using the parameters estimated for the intein and extein sequences. Using the rate ratio of 0.31, corresponding to the ratio of the lengths of intein and extein subtree in the *Thermoplasma* group, the rate of false positives was smaller than 0.000005. Incorporating uncertainty of the extein/intein ratio into the simulation, the false-positive rate increases to 0.0025, corresponding to an expectation of 0.5 false identifications of transfer in table 2.

The most parsimonious explanation for the ratios is mapped on the extein phylogeny in figure 3. Although these values cannot resolve direction of transfer they do show the *vma-I* b inteins are mobile genetic elements that have undergone several transfers throughout the evolution of the gene family encoding the archaeal ATPase catalytic subunit, including several transfers between divergent organisms.

Discussion

Compared with the currently annotated sequences in InBase, we have identified 10 additional inteins belonging to the *vma-I* b family of inteins. Three of which are in new taxonomic classes that were not previously described, the Methanobacteria, Nanohaloarchaea, and Clostridia. This is the first time the *vma-I* b intein is found in a bacterial lineage.

Similar to the PRP8 inteins (Bokor et al. 2012), the *vma-I* b family of inteins is composed of mini inteins and inteins with HE domains. These likely represent inteins in different states of the homing or life cycle. In the homing cycle, an exogenous intein with a HE comes in contact with a population that does not contain an intein and spreads through the population through super-Mendelian rates of inheritance (Goddard and Burt 1999; Gogarten et al. 2002; Gogarten and Hilario 2006). Once all target sites in a population are occupied by inteins containing HEs, there is reduced selection on the HE and it starts to degrade over time, followed by loss of the intein. The mini inteins of the Thermoplasmatales order may be representatives of the later part of the homing cycle, indicating a more ancient invasion of the *vma-I b* site in this clade. Even in cases where the homing cycle does not go to complete fixation in the intein-containing allele, the individual intein insertion sites progress through the same life history from empty target site, occupation by an intein with functioning HE, and then eventual decay of the HE activity (Yahara et al. 2009; Barzel et al. 2011).

Table 2. Pairwise Extein to Intein Substitution Ratios.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Thermoplasmatales archaeon I-														
2. *Thermoplasma volcanium* GSS1	0.39													
3. *Thermop. acidophilum* DSM 1728	0.35	0.14												
4. *Thermop. acidophilum* DSM 122-	0.39	0.17	0.53											
5. *Ferroplasma acidarmanus* Type II	0.49	0.40	0.37	0.41										
6. *Picrophilus torridus* DSM 9790	0.56	0.33	0.44	0.48	0.63									
7. *Mahella australiensis* 50-1 BON	0.55	0.44	0.42	0.46	0.75	0.73								
8. *Candidatus* Nanosalinarum so.	0.59	0.57	0.59	0.72	0.71	0.75	0.64							
9. Thetis_Sea_contig_00086	0.60	0.51	0.49	0.58	0.60	0.59	0.56	1.22						
10. *Methanothermus fervidus* DSM	0.71	0.50	0.48	0.50	0.86	0.84	0.74	0.48	0.37					
11. *Thermococcus litoralis* DSM 5473	0.45	0.37	0.31	0.34	0.53	0.44	0.34	0.48	0.42	0.32				
12. *Pyrococcus abyssi* GE5	0.48	0.38	0.32	0.35	0.54	0.42	0.36	0.53	0.47	0.38	0.41			
13. *Pyrococcus* sp. NA2	0.47	0.38	0.33	0.36	0.51	0.42	0.35	0.53	0.46	0.36	0.41	0.50		
14. *P. furiosus* DSM 3638	0.45	0.38	0.32	0.35	0.53	0.44	0.37	0.53	0.45	0.34	1.29	0.27	0.30	
15. *P. horikoshii* OT3	0.43	0.38	0.31	0.35	0.55	0.42	0.35	0.50	0.43	0.32	0.76	0.09	0.13	0.59

NOTE.—Number of amino acid substitutions were calculated for extein and intein sequences and corrected for multiple substitutions. Ratios greater than inside the *Thermoplasma* group (the top four entries in the table) are taken to be deviations from a ratio of extein to intein that represents coinheritance of intein and extein. Underlined values indicate ratios significantly greater than the ones within the *Thermoplasma* group.

The evolutionary history of the vacuolar, archaeal, and bacterial ATPase catalytic subunit is characterized by several gene transfer events, including transfers between the archaeal and the bacterial domain. Because of the sequence conservation of the catalytic ATPase subunit and that ATPase phylogeny can be rooted by an ancient gene duplication event (Gogarten et al. 1989), these transfer events have been well established (Hilario and Gogarten 1993; Olendzenski et al. 2000; Lapierre et al. 2006; Lapierre 2007).

Although there were multiple transfers of the host gene, all of our analyses were preformed relative to the host gene in order to distinguish the intein transfers from the transfers of the extein. Breakpoint analyses, consideration of pairwise substitution rates, and ancestral character state reconstructions all suggest that the inteins repeatedly invaded the *vma-I* host gene in independent transfer events. In particular, these phylogenetic methods suggest that the intein in *M. australiensis* was gained independently from the between domain transfer of the host A-ATPase operon. The alternate explanation that the intein was gained only once in the ancestor of the *vma-I* protein family and that the extant presence/absence pattern is a result of vertical descent and loss of the intein is incompatible with the evidence listed earlier. In addition, under the loss-only scenario, there would have to be a minimum of 28 steps to explain the patterning, 1 gain and 27 losses, as opposed to only nine steps to explain the intein distribution through independent gains and losses of the intein. The scenario with the minimum number of events is depicted in figure 2. It involves seven intein gains and two losses. Study of inteins and HEs in natural populations as well as theoretical considerations suggest that alleles with and without inteins can coexist in populations for long periods of time (Butler et al. 2006; Gogarten and Hilario 2006; Yahara et al. 2009; Barzel et al. 2011). Given that the sequences used in this study only represent one member of a larger population, assigning a species branch a character state of intein absence ignores the possibility that other members of the species might contain the intein. Thus, only taking ASR into consideration, one needs to entertain the possibility that the loss events inferred for branches represent only absence in the particular sampled genome and not absence in the population or species. Likewise, the inference that the ancestor of the *Pyrococcus* and *Thermococcus* genera did not harbor the intein is not strongly supported through the parsimony reconstruction. The most parsimonious scenario depicted in figure 2 corresponds to two gains; the presence of the intein in the *Pyrococcus–Thermococcus* ancestor corresponds to three loss events within the *Pyrococcus* and *Thermococcus* group. Given that intein absence possibly is overestimated, three apparent loss events versus two gains do not providea strong argument for intein gain in *T. litoralis*. However, the gain of the intein in *T. litoralis* is also supported by the phylogeny of the intein-splicing domain, which groups *T. litoralis* inside the cluster of homologs from the *Pyrococcus* species, supporting the hypothesis that the intein in *T. litoralis* was recently acquired from a *Pyrococcus* species. The general loss-only scenario is incompatible with a comparison of intein and extein phylogenies. If the phyletic pattern could be explained with one gain and subsequent losses, the data sets of the host extein phylogenetic tree and the intein tree should be compatible. However, the AU and SH tests and the GARD analysis all confirm these trees are incompatible.

The GARD algorithm identifies breakpoints three amino acids upstream of the extein intein junctures. This slight deviation from the true intein extein boundary likely reflects a limitation of the approach due to the insertion site and the last three amino acids of the intein being highly conserved. Consequently, little phylogenetic information is contained in the amino acids directly neighboring the intein/extein boundary. However, alternative or additional explanations are possible: the positions in the extein upstream of the intein participate in catalyzing the splicing reaction (Paulus 2000;

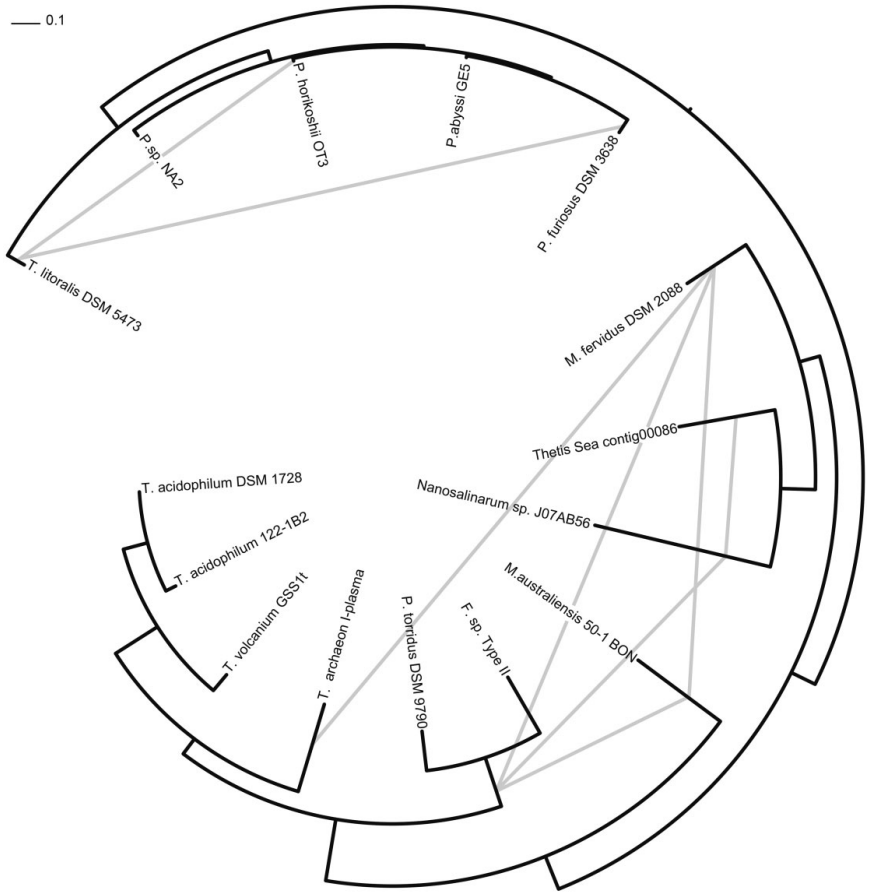


FIG. 3.—Pairwise substitution rate ratios mapped onto the extein tree. Lines connect taxa with extein to intein substitution ratios greater than 0.7. Although direction of intein movement cannot be determined using these ratios, the ratios reveal several putative transfers of the *vma-I* b intein.

Saleh and Perler 2006), thus it is not surprising that they have a signature similar to the intein's splicing domain; during homing, the sequence surrounding the target site is also copied from the invading template containing the intein (Chevalier and Stoddard 2001).

The incongruence between the extein and intein trees in conjunction with the extein to intein divergence ratio data and the ASR data all suggest that the inteins in the *vma-I* b position were horizontally acquired several times. The phylogenetic incongruence between the extein and intein tree provides strong support for the horizontal transfer of the *T. litoralis* intein from within the *Pyrococcus* spp. The extein to intein divergence ratios suggest several additional transfers between *Pyrococcus horikoshii* and *P. furious*; *T. litorialis* and *P. horikoshii*; *T. litorialis* and *P. furious*. The most likely scenario to explain these findings would be a within-group transfer between the two *Pyrococcus* spp. followed by a transfer to *T. litorialis*. The phylogenetic incompatibility and the extein and intein substitution rate ratios taken together are a strong support for a transfer from a *Pyrococcus* sp. to *T. litorialis*.

This is the first report of the *vma-I* b intein found in a bacterial lineage. Although the region where *M. australiensis* falls on the intein tree is not well resolved, the extein to intein ratios suggest several transfers of the intein between the *Candidatus* Nanosalinarium and Thetis sea lineages, the *Methanothermus* and *Mahella* lineages, the *Methanothermus* and *Ferroplasma/Picrophilus* clade, and the *Mahella* and *Ferroplasma/Picrophilus* clade. The phylogenetic distribution

relative to its host gene and the ratio data of the *vma-I* b intein taken together suggest an interdomain transfer between *M. australiensis* and the archaea. However, we cannot conclude that the intein was directly transferred between specific lineages; it is possible that both lineages recently received the intein from a third, currently unsampled lineage.

Conclusions

Here we surveyed the distribution of the *vma-I* b intein and showed that this family of inteins is found in more diverse species than previously reported. These inteins were transferred between divergent organisms in several independent horizontal transfer events, including an HGT from within the *Pyrococci* to *T. litorialis,* a likely transfer to *Candidatus* Nanosalinarum and a transfer across the domain boundaries between the archaea and bacteria. Although an intein may persist within a lineage for a long time without transfer between species (Butler et al. 2006; Gogarten and Hilario 2006; Yahara et al. 2009; Barzel et al. 2011), our data convincingly show that transfer of the intein between divergent species did occur and contributed to the observed modern day distribution of this intein.

Materials and Methods

Sequences

For the large archaeal type, ATPase tree sequences were gathered from the NCBI's nonredundant database. Sequences for the smaller 15 sequence data sets were gathered from either the NCBI nonredundant, whole genome shotgun, or metagenome databases. The metagenomics-rapid annotations using subsystems technology (MG-RAST) database was also surveyed for additional inteins but did not reveal any. All databases were queried in February 2012. The A-ATPases sequence from the Thetis sea metagenome (Ferrer et al. 2012) was translated from contig 00086 (gi:3548555569) after correction of frame shifts.

Phylogenetic Trees

For most data sets (except the ribosomal reference), sequences were aligned with SATe 2.03 (Liu et al. 2009) using the following options: MAFFT for the initial alignment, MUSCLE for the merger, RAxML with a gamma plus invariant sites LG substitution model (PROTGAMMAILGF) for tree estimation. Intein splicing domain and extein sequences were also aligned using MUSCLE (Edgar 2004) and PRANK (Loytynoja and Goldman 2010). ML phylogenies calculated from these alignments (WAG Gamma + F + I) for the exteins were identical to the one given in figure 1. For both the muscle and the PRANK alignments, the *T. litoralis* intein sequence grouped as sister to *P. horikoshii* within the clade formed by the *Pyrococcus* homologs. For both alignments, the grouping of the *T. litoralis* intein with *P. furiosus* and *P. horikoshii* inside the *Pyrococcus* clade was supported by a bootstrap support value of 96%. In the intein ML phylogeny calculated from the PRANK alignment, the *Ferroplasma* and *Picrophilus* inteins group with the *Thermoplasma* homologs but without significant support (52%). The phylogenetic trees were reconstructed using PhyML v3.0, with the best parameters determined by ProtTest3 ML phylogenies, approximate likelihood ratio test (aLRT), and bootstrap support values were determined using PhyML v3.0. under the WAG + G + F + I model. Posterior probabilities were calculated in MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003; Huelsenbeck and Ronquist 2001) using the WAG + G + F + I substitution model (Whelan and Goldman 2001) and two runs. After 70,000 generations, the standard deviation of split frequencies were 0.004317 and 0.008746 for the extein and intein data sets, respectively (small standard deviation of split frequencies indicate adequate convergence of the two runs, the MrBayes tutorial recommends smaller than 0.01 as stop criterion). After inspection of the likelihood trace, the first 700 and 450 generations were discarded as burnins (beginning phase of the tree search that has not reached stationarity) for the extein and intein data sets, respectively.

Archaeal Ribosomal Phylogeny

We generated an archaeal ribosomal reference tree from a concatenated alignment of 62 ribosomal proteins (supplementary multiple sequence alignments) extracted from genomes downloaded from the NCBI ftp site. The ribosome tree included all archaeal taxa from the extein data set, except those that did not have genome representation, such as Thetis Sea contig from a metagenome project (Ferrer et al. 2012). This led to a total of 81 archaea. The number of taxa in each alignment ranged from 19 to 81 (see supplementary table 1, Supplementary Material online, for the complete taxon set). BLASTP (Altschul et al. 1997) searches with reference sets of bacterial and archaeal ribosomal proteins against each archaeal genome and an *e* value cutoff of 1E-10 identified homologs. In most cases, each sequence from a reference set returned the same best Basic Local Alignment Search Tool (BLAST) hit from a genome; interpreted as a strong evidence of homology. We also accepted instances where the majority of reference sequences returned the same best BLAST hit as evidence of homology (such as ratios of 80:1, where 80 references sequences returned the same best BLAST hit and only one differed). Muscle (Edgar 2004) aligned each protein set and Trimal v.1.2 (Capella-Gutierrez et al. 2009) with the "automated1" option trimmed ambiguous regions from all alignments, which were concatenated into a supermatrix with an in-house Python script. PhyML generated a ML phylogeny with 100 bootstrap replicates under the same parameters as those used to generate the extein genealogy.

Reduced Taxa Genealogies

We generated an extein genealogy and corresponding ribosomal phylogeny exclusively for organisms that contained inteins involved in HGT to highlight the correspondence between gene and reference tree. PhyML constructed ML trees under the same parameters as those employed for the full archaeal trees with 100 bootstrap replicates. PhyloBayes v.3.3e (Lartillot et al. 2009) performed Bayesian analyses under the

CAT mixture model (Lartillot and Philippe 2004) with global exchange rates estimated by WAG and two chains for each analysis. The bpcomp option estimated the convergence of the two chains from the maximum difference of the bipartition frequencies. A total of 72,787 and 14,650 cycles were run for the reduced extein and ribosomal trees, respectively. The maximum bipartition frequency difference for the extein genealogy was 0.02 and 0.06 for the ribosomal phylogeny. We employed the CAT model for its efficacy in reducing or eliminating the effects of long-branch attraction. Three taxa were susceptible to LBA artifacts in the ribosomal tree due to amino acid composition and incomplete ribosomal data sets from incomplete genomes—*Micrarchaeum acidiphilium*, *Nanosalinarum* sp., and *Nanosalina* sp. The bacterial sequence from *M. australiensis* (which served as an outgroup) attracted *Nanosalinarum* to the base of the archaea in the reduced ribosomal phylogeny and even the CAT model failed to overcome this LBA artifact.

Topology Tests

CONSEL v.0.2 (Shimodaira and Hasegawa 2001) performed both AU and SH tests to determine whether trees and their bootstrap replicates were significantly different from each other. Site likelihoods were calculated in RAxML v.7.3.5 (Stamatakis et al. 2008).

Breakpoint Analysis

The GARD (Kosakovsky Pond et al. 2006) as implemented on the Datamonkey web server (Delport et al. 2010) was used to determine potential points of recombination in the fifteen *vma*-1 proteins that harbor inteins. Sequences were aligned using SATe´ 2.1 (Liu et al. 2012), and sequences for the HEs were deleted prior to the analysis.

Substitution Rate Ratios

Pairwise ML distances were calculated for each extein and each intein protein splicing domain alignment using TREE-PUZZLE 5.2 (Schmidt et al. 2002), using the WAG + F + I model for substitution and a Gamma distribution approximated through four discrete rate categories, such that the model was the same for all other phylogenic reconstructions. The extein and the intein splicing domain were extracted from the SATe alignment (see earlier). The estimated alpha shape parameter for the intein splicing domains was 3.01 (Standard Error [SE], 0.11), whereas the extein sequences showed a stronger among site rate variation (ASRV) with an estimated alpha of 1.37 (SE, 0.04). To estimate the substitution rate ratio between extein and intein sequences under vertical inheritance, we used the distances estimated with TreePuzzle between *Thermoplasma acidophilum* DSM 122-1B2, *Thermop. volcanium* GSS1, and the Thermoplasmatales archaeon I-plasma (see table 1 and the spreadsheet in supplementary material S1, Supplementary Material online). The inteins in these Thermoplasmatales do not contain a HE domain, indicating they are no longer mobile and were likely inherited together with the extein (see fig. 1). To obtain a phylogenetically independent estimate of the rate ratio, trees were calculated from the intein and extein distances within the *Thermoplama* using FITCH from the PHYLIP package (Felsenstein 2011), and rate ratios were calculated from the branch lengths. Ratios that were significantly larger than the within *Thermoplasma* ratios (mean ratio, 0.30; standard error of the mean [SEM], 0.13), which represent coinheritance of intein and extein, suggest that the inteins are more similar to each other than under the assumption of coinheritance with the exteins, indicating an intein transfer relative to the extein. Although direction of the intein transfer cannot be determined using these ratios, recent transfers can be detected. Values greater than 0.70 were considered to be significantly greater than the *Thermoplasma* ratios (P _: 0.002). The high cutoff level was chosen to avoid false positives.

False-Positive Rates for the Substitution Ratio Test

We performed two simulation studies to estimate the rate of false positives associated with the rate ratio cutoff used to identify putative transfer events. An intein tree, under the assumption of vertical inheritance only, was derived from the extein tree by dividing the branch lengths in the extein tree by 0.3145. This value was obtained from the ratio of the tree lengths within the *Thermoplama* group between the actual extein and intein trees. For the first simulation, we used the aforementioned trees and simulated 10,000 sequence sets for the intein and extein tree using Evolver (Yang 2007). The sequences were simulated under the WAG model, and alpha values were taken from the actual trees, which were 3.01 and 1.37 for the intein and extein, respectively. The substitution ratios were calculated. Extein distances less than 0.5 were excluded, and the number of ratios above 0.7 was counted as false positives.

We also estimated the false-positive rates incorporating the uncertainty in determining the rate ratio. This was done by randomly sampling 1,000 rate ratios from a folded normal distribution with mean 0.3145 and standard deviation of 0.1343 (the SEM for the determined rate ratio). The sampled ratios were used to generate new intein trees. The simulated intein and extein trees were evaluated as described earlier.

Maximum Parsimony ASR

A parsimony ASR and MLASR were both performed to determine where among the *vma-I* protein family the inteins gained. The presence/absence patterns of the inteins were converted to a binary form, and the maximum parsimony ASR was calculated using the ordered model and the MLASR was calculated under the MK1 model (Lewis 2001). Both reconstructions were implemented in Mesquite (Maddison WP and Maddison DR 2011).

Supplementary Material

Supplementary table S1, figures S1–S8, and files are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Aeronautics and Space Administration Astrobiology, Exobiology and Evolutionary Biology (grant numbers NNX08AQ10G and NNX13AI03G) and the NSF Assembling the Tree of Life (ATOL) (DEB0830024) programs. The authors thank Matthew Fullmer, David Williams, Tim Harlow, Ofir Cohen, and an anonymous reviewer for providing insightful discussions on the manuscript.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.

Anderson I, Djao OD, Misra M, et al. (41 co-authors). 2010. Complete genome sequence of *Methanothermus fervidus* type strain (V24S). *Stand Genomic Sci.* 3(3):315–324.

Barzel A, Obolski U, Gogarten JP, Kupiec M, Hadany L. 2011. Home and away—the evolutionary dynamics of homing endonucleases. *BMC Evol Biol.* 11(1):324.

Bokor AA, Kohn LM, Poulter RT, van Kan JA. 2012. PRP8 inteins in species of the genus *Botrytis* and other ascomycetes. *Fungal Genet Biol.* 49(3):250–261.

Butler MI, Gray J, Goodwin TJ, Poulter RT. 2006. The distribution and evolutionary history of the PRP8 intein. *BMC Evol Biol.* 6:42.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

Chevalier BS, Stoddard BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* 29(18):3757–3774.

Clerissi C, Grimsley N, Desdevises Y. 2013. Genetic exchanges of inteins between prasinoviruses (phycodnaviridae). *Evolution* 67(1):18–33.

Cooper AA, Stevens TH. 1995. Protein splicing: self-splicing of genetically mobile elements at the protein level. *Trends Biochem Sci.* 20(9):351–356.

Darland G, Brock TD, Samsonoff W, Conti SF. 1970. A thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. *Science* 170(3965):1416–1418.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26(19):2455–2457.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10(8):R85.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.

Elleuche S, Poggeler S. 2010. Inteins, valuable genetic elements in molecular biology and biotechnology. *Appl Microbiol Biotechnol.* 87(2):479–489.

Erauso G, Reysenbach A-L, Godfroy A, et al. (11 co-authors). 1993. *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *Archiv Microbiol.* 160(5):338–349.

Felsenstein J. 2011. PHYLIP (Phylogeny Inference Package) version 3.67. Distributed by the author. Seattle (WA): University of Washington, Department of Genome Sciences.

Ferrer M, Werner J, Chernikova TN, et al. (12 co-authors). 2012. Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study. *Environ Microbiol.* 14(1):268–281.

Fiala G, Stetter KO. 1986. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Archiv Microbiol.* 145(1):56–61.

Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci U S A.* 96(24):13880–13885.

Gogarten JP. 1994. Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. *J Mol Evol.* 39(5):541–543.

Gogarten JP, Hilario E. 2006. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol Biol.* 6:94.

Gogarten J, Kibak H, Dittrich P, et al. (13 co-authors). 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A.* 86(17):6661–6665.

Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. 2002. Inteins: structure, function, and evolution. *Annu Rev Microbiol.* 56:263–287.

Gonzalez JM, Masuchi Y, Robb FT, Ammerman JW, Maeder DL, Yanagibayashi M, Tamaoka J, Kato C. 1998. *Pyrococcus horikoshii* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. *Extremophiles* 2(2):123–130.

Harvey WR, Nelson N, editors 1992. V-ATPases. *J Exp Biol.* 172:1–485.

Hilario E, Gogarten JP. 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. *Biosystems* 31(2-3):111–119.

Hirata R, Ohsumk Y, Nakano A, Kawasaki H, Suzuki K, Anraku Y. 1990. Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of *Saccharomyces cerevisiae*. *J Biol Chem.* 265(12):6726–6733.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.

Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH. 1990. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* 250(4981):651–657.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22(24):3096–3098.

Koufopanou V, Goddard MR, Burt A. 2002. Adaptation for horizontal transfer in a homing endonuclease. *Mol Biol Evol.* 19(3):239–246.

Lapierre P. 2007. The impact of horizontal gene transfers on prokaryotic genome evolution [PhD thesis]. Storrs (CT): University of Connecticut.

Lapierre P, Shial R, Gogarten JP. 2006. Distribution of F- and A/V-type ATPases in *Thermus scotoductus* and other closely related species. *Syst Appl Microbiol.* 29(1):15–23.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21(6):1095–1109.

Lee HS, Bae SS, Kim MS, Kwon KK, Kang SG, Lee JH. 2011. Complete genome sequence of hyperthermophilic *Pyrococcus* sp. strain NA2, isolated from a deep-sea hydrothermal vent area. *J Bacteriol.* 193(14):3666–3667.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol.* 50(6):913–925.

Liu L, Laufer H, Gogarten PJ, Wang M. 1997. cDNA cloning of a mandibular organ inhibiting hormone from the spider crab *Libinia emarginata*. *Invert Neurosci.* 3(2-3):199–204.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324(5934):1561–1564.

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol.* 61(1):90–106.

Liu XQ. 2000. Protein-splicing intein: genetic mobility, origin, and evolution. *Annu Rev Genet.* 34:61–76.

Loytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinf.* 11: 579.
Macgregor BJ, Biddle JF, Teske A. 2013. Mobile elements in a single-filament orange Guaymas Basin Beggiatoa ("*Candidatus* Maribeggiatoa") sp. draft genome: evidence for genetic exchange with cyanobacteria. *Appl Environ Microbiol.* 79(13):3974–3985.
Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. [cited 2013 Oct 5]. Available from: http://mesquiteproject.org.
Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* 6(1):81–93.
Neuner A, Jannasch HW, Belkin S, Stetter KO. 1990. *Thermococcus litoralis* sp. nov.: a new species of extremely thermophilic marine archaebacteria. *Archiv Microbiol.* 153(2):205–207.
Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, Koonin EV. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *J Bacteriol.* 186(19):6575–6585.
Okuda Y, Sasaki D, Nogami S, Kaneko Y, Ohya Y, Anraku Y. 2003. Occurrence, horizontal transfer and degeneration of VDE intein family in Saccharomycete yeasts. *Yeast* 20(7):563–573.
Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP. 2000. Horizontal transfer of archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. *J Mol Evol.* 51(6):587–599.
Paulus H. 2000. Protein splicing and related forms of protein autoprocessing. *Annu Rev Biochem.* 69:447–496.
Perler FB. 2002. InBase: the intein database. *Nucleic Acids Res.* 30(1):383–384.
Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J, Belfort M. 1994. Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res.* 22(7):1125–1127.
Perler FB, Olsen GJ, Adam E. 1997. Compilation and analysis of intein sequences. *Nucleic Acids Res.* 25(6):1087–1093.
Pietrokovski S. 1998. Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci.* 7(1):64–71.
Podell S, Gaasterland T. 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* 8(2):R16.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
Saleh L, Perler FB. 2006. Protein splicing in cis and in trans. *Chem Rec.* 6(4):183–193.
Salinas MB, Fardeau ML, Thomas P, Cayol JL, Patel BK, Ollivier B. 2004. *Mahella australiensis* gen. nov., sp. nov., a moderately thermophilic anaerobic bacterium isolated from an Australian oil well. *Int J Syst Evol Microbiol.* 54(Pt 6):2169–2173.
Schleper C, Piihler G, Kuhlmorgen B, Zillig W. 1995. Life at extremely low pH. *Nature* 375(6534):741–742.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3):502–504.
Searcy DG. 1975. Histone-like protein in the prokaryote *Thermoplasma acidophilum*. *Biochim Biophys Acta.* 395(4):535–547.
Segerer A, Langworthy TA, Stetter KO. 1988. *Thermoplasma acidophilum* and *Thermoplasma volcanium* sp. nov. from Solfatara Fields. *Syst Appl Microbiol.* 10(2):161–171.
Senejani AG, Hilario E, Gogarten JP. 2001. The intein of the thermoplasma A-ATPase A subunit: structure, evolution and expression in *E. coli*. *BMC Biochem.* 2:13.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol.* 57(5):758–771.
Swithers KS, Senejani AG, Fournier GP, Gogarten JP. 2009. Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol Biol.* 9:303.
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37–43.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18(5):691–699.
Yahara K, Fukuyo M, Sasaki A, Kobayashi I. 2009. Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer. *Proc Natl Acad Sci U S A.* 106(44):18861–18866.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24(8):1586–1591.
Zimniak L, Dittrich P, Gogarten JP, Kibak H, Taiz L. 1988. The cDNA sequence of the 69-kDa subunit of the carrot vacuolar H + - ATPase. Homology to the beta-chain of F0F1-ATPases. *J Biol Chem.* 263(19):9102–9112.

Chapter 4. Intein Distribution and Evolution in the Haloarchaea.

This chapter contains two publications and an addendum to the work that has already been published. For the first paper I was the second author and helped with the analysis of the intein presence/absence in the genomes, as well as helped with the writing. Matthew S. Fullmer primarily wrote the first paper, I performed the intein analysis; Peter Gogarten and Thane Papke supervised the research and writing.

I was the primary author for the second publication, where the intein distribution in the Haloarchaea was surveyed. This paper examines the distribution of inteins in 118 haloarchaeal genomes. Additionally the inteins that were found were used to search the non-redundant database hosted by NCBI (December 2013) to find organisms that share inteins with the Haloarchaea that are not represented by the 118 Haloarchaeal genomes in this work. The haloarchaeal inteins were examined all together to look for highways of gene transfer. The supported clusters of the concatenated intein sequences show that highways of gene transfer are most strongly supported within environments. Though there is, as shown in previous research, a strong phylogenetic bias within these environmental gene exchange networks.

4.1 Population and Genomic Analysis of the Genus Halorubrum

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 April 2014
doi: 10.3389/fmicb.2014.00140

Population and genomic analysis of the genus *Halorubrum*

Matthew S. Fullmer[1], Shannon M. Soucy[1], Kristen S. Swithers[1,2], Andrea M. Makkay[1], Ryan Wheeler[1], Antonio Ventosa[3], J. Peter Gogarten[1] and R. Thane Papke[1]*

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[2] Department of Cell Biology, Yale School of Medicine, Yale University, New Haven, CT, USA
[3] Department of Microbiology and Parasitology, University of Seville, Seville, Spain

Edited by:
Jesse Dillon, California State University, Long Beach, USA

Reviewed by:
Jesse Dillon, California State University, Long Beach, USA
Federico Lauro, University of New South Wales, Australia

***Correspondence:**
R. Thane Papke, Microbiology Program, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Storrs, CT 06269-3125, USA
e-mail: thane@uconn.edu!

The Halobacteria are known to engage in frequent gene transfer and homologous recombination. For stably diverged lineages to persist some checks on the rate of between lineage recombination must exist. We surveyed a group of isolates from the Aran-Bidgol endorheic lake in Iran and sequenced a selection of them. Multilocus Sequence Analysis (MLSA) and Average Nucleotide Identity (ANI) revealed multiple clusters (phylogroups) of organisms present in the lake. Patterns of intein and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) presence/absence and their sequence similarity, GC usage along with the ANI and the identities of the genes used in the MLSA revealed that two of these clusters share an exchange bias toward others in their phylogroup while showing reduced rates of exchange with other organisms in the environment. However, a third cluster, composed in part of named species from other areas of central Asia, displayed many indications of variability in exchange partners, from within the lake as well as outside the lake. We conclude that barriers to gene exchange exist between the two purely Aran-Bidgol phylogroups, and that the third cluster with members from other regions is not a single population and likely reflects an amalgamation of several populations.

Keywords: Halobacteria, Multilocus Sequence Analysis (MLSA), Average Nucleotide Identity (ANI), intein, CRISPR

INTRODUCTION

Besides an obligate requirement for high concentrations of NaCl, a unifying trait of Halobacteria (often referred to colloquially as the haloarchaea)—a class within the archaeal phylum Euryarchaeota, is their propensity for horizontal gene transfer (HGT) (Legault et al., 2006; Rhodes et al., 2011; Nelson-Sathi et al., 2012; Williams et al., 2012). Although HGT occurs continuously, events that provide an adaptive advantage and are maintained in modern lineages can be detected. For instance, HGTs from bacterial lineages into the Halobacteria occurred before their last common ancestor and brought respiration and nutrient transport genes that transformed them from a methanogen to their current aerobic heterotrophic state (Nelson-Sathi et al., 2012). Other examples including rhodopsins (Sharma et al., 2006), tRNA synthetases (Andam et al., 2012), 16S rRNA genes (Boucher et al., 2004), membrane proteins (Cuadros-Orellana et al., 2007), and genes allowing the assembly of novel pathways (Khomyakova et al., 2011) have been reported for this group and reflect the adaptive benefit of acquiring these genes.

HGT into the Halobacteria has profoundly impacted their evolution; however, understanding this contribution is only part of their evolutionary picture. The study of recombination frequency among this class has been utilized to address population genetics questions that address whether they are clonal (i.e., linked alleles at different loci) or "sexual" in the sense that alleles at different loci are randomly associated. Several studies have addressed those questions by assessing the impact of frequent HGT on Halobacteria. Homologous replacement of loci was inferred within and between phylogenetic clusters (phylogroups) using Multilocus Sequence Analysis (MLSA) on closely related strains (Papke et al., 2004) and comparative analyses of genomes (Williams et al., 2012). Within phylogroups where genetic diversity was less than one percent divergent for protein coding genes, alleles at different loci were randomly associated whereas between phylogroups they were not (Papke et al., 2007) indicating haloarchaea are highly sexual. Measurements of frequency across the breadth of halobacterial diversity indicates no absolute barrier to homologous recombination; rather between relatives, there is a log-linear decay in recombination frequency relative to phylogenetic distance (Williams et al., 2012).

Laboratory experiments also support these results. Mating experiments measuring the rate of recombination using *Haloferax (Hfx) volcanii* and *Hfx. mediterranei* auxotrophs demonstrated the degree of genetic isolation between species was much lower than expected. The observed rate of exchange between species suggested that given an opportunity over time these species would homogenize, indicating strong barriers to recombination would have to exist for speciation to occur, and for lineages to be maintained (Naor et al., 2012). Further, mating experiments demonstrated that enormous genomic fragments (i.e., 300–500 kb, ~18% of the chromosome size) could be exchanged in a single event (Naor et al., 2012). Similar large fragment exchange events were recently observed in natural isolates from Deep Lake (Antarctic hypersaline lake): Distantly related strains

(<75% average nucleotide identity) shared up to 35 kb with nearly 100% sequence identity (DeMaere et al., 2013).

The Halobacteria have clearly been shaped by gene transfer and are actively engaged in substantial genetic exchange. However, little is known about genomic diversity within populations, and the impact of gene flow is unknown at these scales. In this study we report the intra and inter population sequence diversity of *Halorubrum* spp. strains cultivated from the same location and compare them to the genomic diversity of type strains from the same genus. Our results lead to insights on the genomic diversity that comprises haloarchaeal species.

METHODS

GROWTH CONDITIONS AND DNA EXTRACTION

Halorubrum spp. cultures were grown in Hv-YPC medium (Allers et al., 2004) at 37°C with agitation. DNA from Halobacteria was isolated as described in the Halohandbook (Dyall-Smith, 2009). Briefly, stationary-phase cells were pelleted at 10,000 × *g*, supernatant was removed and the cells were lysed in distilled water. An equal volume of phenol was added, and the mixture was incubated at 65°C for 1 h prior to centrifugation to separate the phases. The aqueous phase was reserved and phenol extraction was repeated without incubation, and followed with a phenol/chloroform/iso-amyl alcohol (25:24:1) extraction. The DNA was precipitated with ethanol, washed, and re-suspended in TE (10 mM tris, pH 8.0, 1 mM EDTA).

MULTILOCUS SEQUENCE ANALYSIS (MLSA)

Five housekeeping genes were amplified using PCR. The loci were *atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB* and the primers used for each locus are listed in **Table 1**. To more efficiently sequence PCR products, an 18 bp M13 sequencing primer was added to the 5′ end of each degenerate primer (**Table 1**). Each PCR reaction was 20 µl in volume. The PCR reaction was run on a Mastercycler Ep Thermocycler (Eppendorf) using the following PCR cycle protocol: 30 s initial denaturation at 98°C, followed by 40 cycles of 30 s at 98°C, 5 s at the annealing temperature for each set of primers and 15 s at 72°C. Final elongation occurred at 72°C for 1 min. **Table 2** provides a detailed list of reagents and the PCR mixtures for each amplified locus. The PCR products were separated by gel electrophoresis with agarose (1%). Gels were stained with ethidium bromide. An exACTGene mid-range plus DNA ladder (Fisher Scientific International Inc.) was used to estimate the size of the amplicons, which were purified using Wizard SV gel and PCR cleanup system (Promega). The purified amplicons were sequenced by Genewiz Inc. using Sanger sequencing technology.

GENOME SEQUENCING

DNA purity was analyzed with a Nanodrop spectrophotometer, was quantified using a Qubit fluorometer (Invitrogen) and then prepared for sequencing using the Illumina Nextera XT sample preparation kit as described by the manufacturer. Fragmented and amplified libraries were either normalized using the normalization beads and protocol supplied with the kit, or manually as described in protocols for the Illumina Nextera kit. Libraries were loaded onto 500 cycle MiSeq reagent kits with a 5% spike-in PhiX control, and sequenced using an Illumina MiSeq benchtop sequencer. The genomes to be sequenced were selected based upon the results of the initial PCR MLSA data analysis (see Results).

Table 1 | Degenerate primers used to PCR amplify and sequence the genes for MLSA.

	MLSA primer sequence 5′–3′	
Locus	**Forward**	**Reverse**
atpB	tgt aaa acg acg gcc agt aac ggt gag scv ats aac cc	cag gaa aca gct atg act tca ggt cvg trt aca tgt a
ef-2	tgt aaa acg acg gcc agt atc cgc gct bta yaa stg g	cag gaa aca gct atg act ggt cga tgg wyt cga ahg g
glnA	tgt aaa acg acg gcc agt cag gta cgg gtt aca sga cgg	cag gaa aca gct atg acc ctc gcs ccg aar gac ctc gc
ppsA	tgt aaa acg acg gcc agt ccg cgg tar ccv agc atc gg	cag gaa aca gct atg aca tcg tca ccg acg arg gyg g
rpoB	tgt aaa acg acg gcc agt tcg aag agc cgg acg aca tgg	cag gaa aca gct atg acc ggt cag cac ctg bac cgg ncc

Table 2 | PCR conditions for each locus.

	atpB	ef-2	glnA	ppsA	rpoB
Water (µl)	11.6	8.2	11.8	7.9	11.9
5× phire reaction buffer (µl)	4.0	4.0	4.0	4.0	4.0
DMSO (µl)	0.6	0	0.4	0.6	0.6
Acetamide (25%, µl)	0	4.0	0	4.0	0
dNTP mix (10 mM, µl)	0.4	0.4	0.4	0.4	0.4
Forward primer (10 mM, µl)	1.0	1.0	1.0	1.0	1.0
Reverse primer (10 mM, µl)	1.0	1.0	1.0	1.0	1.0
Phire II DNA polymerase (µl)	0.4	0.4	0.4	0.4	0.4
Template DNA (20 ng/µl, µl)	1.0	1.0	1.0	0.7	0.7
Annealing temperature (° C)	60.0	61.0	69.6	66.0	63.7

GENOME ASSEMBLY

Type strain genomes were obtained from the NCBI ftp repository. *Halorubrum lacusprofundi* and the non-*Halorubrum* genomes (*Haloarcula marismortui* ATCC 43049 and *Har. hispanica* ATCC 33960 as well as *Haloferax volcanii* DS2 and *Hfx. mediterranei* ATCC 33500) are completed projects. The other *Halorubrum* genomes are drafts, also obtained from the NCBI ftp repository. New draft genomes were sequenced using an Illumina MiSeq platform. Assembly on strain Ga2p was carried out using the ngopt A5 pipeline(Tritt et al., 2012) while all others were assembled via the CLC Genomics Workbench 6.0.5 suite with a trim and merge workflow with scaffolding enabled.

To ensure equal gene calling across the genomes all genomes, including the 19 draft and completed *Halorubrum*, *Haloferax,* and *Haloarcula* genomes available on the NCBI ftp site as of June 2013, were reannotated using the rapid annotation using subsystem technology (RAST) server (Aziz et al., 2008). Assembled contigs were reconstructed from the RAST-generated genbank files for all genomes using the seqret application of the emboss package (Rice et al., 2000).

PHYLOGENETIC METHODOLOGY

Top scoring BLASTn hits for each MLSA target gene (*atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB*) in each genome were identified. Multiple-sequence alignments (MSAs) were generated by translating the genes to protein sequences in SeaView (Gouy et al., 2010), aligning the proteins using MUSCLE (v.3.8.31) (Edgar, 2004) and then reverting back to the nucleotide sequences. In-house scripts created a concatenated alignment of all five genes. The best model of evolution was determined by calculating the Akaike Information Criterion with correction for small sample size (AICc) in jModelTest 2.1.4 (Guindon et al., 2010; Darriba et al., 2012). The best-fitting model was GTR + Gamma estimation + Invariable site estimation. A maximum likelihood (ML) phylogeny was generated from the concatenated MSA and individual gene phylogenies from the individual gene MSAs using PhyML (v3.0_360-500M)(Guindon et al., 2010). PhyML parameters consisted of GTR model, estimated p-invar, 4 substitution rate categories, estimated gamma distribution, subtree pruning, and regrafting enabled with 100 bootstrap replicates.

PAIRWISE SEQUENCE IDENTITY CALCULATION

Calculation of pairwise identities was carried out using Clustal Omega on the EMBL-EBI webserver (http://www.ebi.ac.uk/Tools/msa/clustalo/). The alignments were uploaded and percent identity matrices calculated (Sievers et al., 2011).

INTEIN METHODOLOGY

To retrieve haloarchaeal intein sequences Position-Specific Scoring Matrices (PSSMs) were created using the collection of all inteins from InBase, the Intein database, and registry (Perler, 2002). A custom database was created with all inteins, and each intein was used as a seed to create a PSSM using the custom database. These PSSMs were then used as a seed for PSI-BLAST (Altschul et al., 1997) against each of the halobacterial genomes available from NCBI. A size exclusion step was then performed to remove false positives. Inteins were then aligned using MUSCLE (Edgar, 2004) with default parameters in the SeaView version 4.0 software package (Gouy et al., 2010). Insertions, which passed the size exclusion step but did not contain splicing domains, were filtered out and the previous steps were repeated using the resulting dataset on this study's dataset. Once the collection of haloarchaeal inteins was complete, sequences were re-aligned using SATé v2.2.2 (Liu et al., 2012) to generate a final alignment.

INTEIN PHYLOGENETIC METHODOLOGY

Intein protein sequences were retrieved using in house scripts. Each intein allele was aligned separately using MUSCLE (v.3.8.31) (Edgar, 2004). In-house scripts created a concatenated alignment from the allele alignments. ProtTest v3.4 (Darriba et al., 2011) evaluated the protein sequences for an optimal model using the AICc and returned WAG_I+G+F. A presence-absence matrix of zeros and ones was amended to each taxon's alignment data. The presence-absence data allows for grouping of taxa by sharing or lacking an allele. This complements the protein data, and allows the resolution of taxa with few inteins from those lacking them entirely or possessing many. To accommodate the two different formats of data simultaneously MrBayes v3.2.2 (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012) was employed for the phylogenetic reconstruction.

AVERAGE NUCLEOTIDE IDENTITY/TETRAMER ANALYSIS

JSpecies1.2.1 (Richter and Rosselló-Móra, 2009) was used to analyze the genomes for Average Nucleotide Identity (ANI) and tetramer frequency patterns. As the relationships of interest for this study are within the same genus only the nucmer and tetra algorithms were used. The BLAST-based ANI was not used as we were primarily interested in understanding the degree of relatedness between closely related organisms, which the nucmer method is equally capable of (Richter and Rosselló-Móra, 2009). Additionally, the increased rate of drop-off between moderately divergent sequences (<90%) the nucmer method yields relative to the BLAST method (Richter and Rosselló-Móra, 2009) was useful in highlighting when organisms were dissimilar. The default settings for both algorithms were used (Richter and Rosselló-Móra, 2009).

CODON POSITION GC CONTENT

Complete sets of nucleotide sequences for all called ORFs were downloaded from RAST. In house scripts confirmed that all ORF calls were divisible by three and thus could be taken as in-frame. In house scripts were used to calculate the GC percentages for each codon position in each genome. Two-tailed *t*-tests were calculated using the StatsPlus software package (AnalystSoft, 2009).

CRISPRs

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) presence/absence patterns were determined using the CRISPR Recognition Tool (CRT) v1.2 (Bland et al., 2007) with minimum repeat and minimum spacer parameters set to 30 nucleotides. All other parameters were the CRT defaults.

RESULTS

ASSEMBLED GENOMES

The assembled genomes ranged in size from 2.3 to 4.2 Mb. The median assembled genome size is 3.6 Mb. The median N50 (the size of the contig where 50% of the basepairs in the assembly are part of a contig that size or larger. N75 and N90 are similar but use 75 and 90% cutoffs) was 47.5 kb with a range from 1.86 to 80.3 kb (see **Table 3**, for statistics on the assembled genomes). Plasmids were not identified during assembly. As such, if some isolates possess differing numbers or types of plasmids then some of the genome-to-genome size variability may be attributable to this. A list of genomes used in this study can be found in **Table 4**.

PHYLOGENETIC ASSIGNMENT OF PHYLOGROUPS

Initial MLSA analysis (5-genes: *atpD*, *ef-2*, *glnA, radA, rpoB*) revealed the presence of three well-supported clusters [hereafter referred to as phylogroups *in sensu* (Papke et al., 2007)] within the canonical *Halorubrum* population of Aran-Bidgol (**Figures 1**, **2**). A phylogroup was initially defined as a cluster of isolates with very low sequence divergence across the sequenced (MLSA) loci (<~1%). Seventeen of these isolates were then selected for genome sequencing for a higher resolution assessment. Selection criteria were biased toward the two larger phylogroups (A and B) to facilitate comparison between clusters. Only a single genome

Table 3 | Assembly statistics for the genomes sequenced in this study.

	C191	C3	C49	Cb34	E3	E8	Ea1	Ea8	Eb13	Ec15	Fb21	G37	Ga2p	Ga36	Hd13	Ib24	LD3	LG1	
N75 (kb)	18.9	2.3	23.2	24.7	1.1	1.3	30.0	25.1	25.4	42.7	25.3	27.2	41.1	23.8	32.1	23.2	21.4	8.4	
N50 (kb)	54.9	4.4	56.3	42.9	1.9	2.3	43.8	51.6	51.6	80.3	42.7	68.1	74.9	51.2	64.4	43.4	39.6	32.1	
N25 (kb)	97.3	7.8	99.8	73.4	3.5	4.0	77.5	95.4	95.7	131.8	90.3	118.4	118.9	91.9	83.0	68.2	76.0	67.9	
Minimum (kb)	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.6	0.5	0.5	0.3	0.5	0.5	0.5	0.5	0.4	
Maximum (kb)	180.2	40.5	183.6	123.4	26.7	25.0	203.3	169.6	268.1	412.4	174.7	230.0	246.3	145.6	122.0	190.3	145.8	153.4	
Average (kb)	16.6	2.9	22.5	23.1	1.5	1.8	24.7	22.6	23.3	44.3	20.6	25.7	40.3	21.0	27.9	19.6	17.5	4.4	
Contig count	233	1165	159	145	2764	1278	159	166	156	74	176	138	83	160	137	189	213	1090	
Length (Mb)	3.87	3.33	3.58	3.35	4.21	2.26	3.93	3.75	3.63	3.28	3.63	3.55	3.35	3.36	3.82	3.70	3.73	4.79	
Base composition (GC%)	66.0	65.8	65.8	67.6	65.5	66.3	67.0	67.6	67.5	67.6	66.6	67.1	67.8	67.7	67.6	67.6	66.2	66.0	
Number of coding sequences	3908	3379	3529	3323	4147	2187	3977	3672	3544	3245	3600	3617	3400	3382	3718	3612	3724	4615	
Number of RNAs	57	37	49	54	51	31	50	49	48	47	65	48	49	47	51	48	56	69	

from phylogroup C was sequenced. Once genomic data were available, the PCR amplicons were replaced with the full-length genes from the assemblies. Further analysis made use of only these genomic sequences. The addition of the 19 NCBI genomes was made to provide context to the placement of the phylogroups within the genus and to determine their relationship with each other. The phylogenetic reconstruction including the type strains sequences revealed the presence of a fourth phylogroup (designated D) composed of three isolates from Aran-Bidgol and five type strains isolated from Central Asia and China (Figure 2).

PHYLOGROUPS A AND B ARE WELL-SUPPORTED AS DISCRETE AND COHESIVE ENTITIES

The bootstrap values provided by the phylogenetic reconstruction strongly supported both phylogroups A and B. Individual gene trees and the concatenated gene tree returned support values of 99% or higher for all of the clusters (Figures 1, 2) and the trees showed no paraphyly with other taxa. Both phylogroups also displayed sequence divergence below 1% across the five loci (Table 5). Further, genome-level analysis (ANI) demonstrated similar results to the MLSA data (Figure 3). Additional support for these phylogroups came from the tetramer frequency analysis, which found no discordance amongst the members of either group, and each phylogroup displayed an intra-group ANI >98%. An analysis of G+C composition in the protein coding ORFs found that the strains within phylogroups A and B had a statistically different content in overall coding G+C and at the third codon position ($P < 0.05$ for both, Figure 4). Analyses of the inter-phylogroup differences showed the two phylogroups were quite different from each other and all other examined taxa. Both clusters were less than 97% similar in their pairwise MLSA distance to any other taxon in this study. Additionally phylogroups A and B were different from each other in tetramer frequency (below the 0.9900 correlation of Richter and Rosselló-Móra, 2009), ANI (only ~87% identity), and G+C content in the third codon position ($P < 0.05$; two-tailed t-test, Figure 4). Taken together these data support the notion that these phylogroups are discrete entities within a single environment, and that the individual phylogroups are cohesive.

To further evaluate the cohesion of the phylogroups a survey of inteins was performed. Inteins are molecular parasites that invade new hosts through horizontal transmission (Okuda et al., 2003; Swithers et al., 2013). Their patterns of presence and absence have been used as a barometer for horizontal transfer between closely and distantly related lineages (Swithers et al., 2013). Analysis of intein distributions supported earlier findings of cohesion within phylogroups and major distinctions between the phylogroups (Figure 5). Phylogroup A contains three non-fixed intein alleles that are present in more than half of the isolates (cdc21-a, cdc21-b, and pol-II-a. Phylogroup B contains four non-fixed intein alleles also present in half or more of its isolates (rir1-b, rfc-a, polB-a, and polB-b but are absent from phylogroup A. Closer examination of the two shared alleles reveals that these inteins are not the same between the phylogroups. The pol-II-a inteins in phylogroup B are 515 aa long while those in phylogroup A are 494 aa long, indicating an insertion or deletion event occurred in one of the phylogroups before the intein spread through the population. The preservation

Table 4 | List of genomes used in this study.

Organism name	NCBI identifier	Sequence source	Isolation site	Environment	Status
Haloarcula hispanica ATCC 33960	PRJNA72475	NCBI	Alicante, Spain	Solar saltern	Complete
Haloarcula marismortui ATCC 43049	PRJNA57719	NCBI	Dead Sea, Israel	Saline lake/sea	Complete
Haloferax mediterranei ATCC 33500	PRJNA167315	NCBI	Alicante, Spain	Solar saltern	Complete
Haloferax volcanii DS2	PRJNA46845	NCBI	Dead Sea, Israel	Saline lake/sea	Complete
Halorubrum sp. T3	PRJNA199598	NCBI	Yunnan, China	Solar saltern	Draft
Halorubrum aidingense JCM 13560	PRJNA188616	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum arcis JCM 13916	PRJNA188617	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum californiensis DSM 19288	PRJNA188618	NCBI	California, United States	Solar saltern	Draft
Halorubrum coriense DSM 10284	PRJNA188619	NCBI	Geelong, Australia	Solar saltern	Draft
Halorubrum distributum JCM 10118	PRJNA188621	NCBI	Turkmenistan	Saline soils	Draft
Halorubrum distributum JCM 9100	PRJNA188620	NCBI	Turkmenistan	Saline soils	Draft
Halorubrum hochstenium ATCC [illegible]	PRJNA188622	NCBI	California, United States	Solar saltern	Draft
Halorubrum kocurii JCM 14978	PRJNA188615	NCBI	Inner Mongolia, China	Saline lake	Draft
Halorubrum lacusprofundi ATCC 49239	PRJNA58807	NCBI	Deep Lake, Antarctica	Saline lake	Complete
Halorubrum lipolyticum DSM 21995	PRJNA188614	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum litoreum JCM 13561	PRJNA188613	NCBI	Fujian, China	Solar saltern	Draft
Halorubrum saccharovorum DSM 1137	PRJNA188612	NCBI	California, United States	Solar saltern	Draft
Halorubrum tebenquichense DSM [illegible]	PRJNA188611	NCBI	Atacama, Chile	Solar saltern	Draft
Halorubrum terrestre JCM 10247	PRJNA188610	NCBI	Turkmenistan	Saline soils	Draft
Hrr. Cb34	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. C49	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ea1	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Eb13	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ib24	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ea8	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Hd13	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. C3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. E8	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. E3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. LG1	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Fb21	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ga2p	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. G37	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. LD3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ec15	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ga36	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft

of the insertion or deletion within the phylogroups indicates that gene flow is occurring more readily within phylogroups than between, even when the same intein allele is shared. In accordance with earlier evidence, within phylogroups the intein sequence similarity is much higher than between phylogroups. It is unlikely that intein lengths are the result of sequencing or assembly artifacts, as they are constant within phylogroups.

The phylogenetic reconstruction derived from the combined presence-absence data and intein sequence data (**Figure 6**) shows clustering among phylogroup A and B of their constituent taxa. None of the taxa placed anywhere else but with the other members of its phylogroups and the posterior probabilities for these placements are high (0.991 for A and 0.923 for B). These results indicate that inteins are diverging mainly along cluster boundaries, as phylogroups A and B are distinct and separate, which further suggests that it is more challenging for the inteins to migrate outside compared to inside their phylogroups.

Another genetic element that serves to distinguish phylogroups A from B is the relative presence of CRISPRs. CRISPRs are a type of microbial innate immunity that provides a record of MGEs previously encountered by the lineage that carries them. This record serves the organism by recognizing and destroying sequences that resemble previously encountered MGEs. CRISPRs have been reported in 90% of surveyed archaeal genomes (Kunin et al., 2007), thus the presence and similarity of CRISPR loci provides a means for comparing the phylogroups. The distribution of CRISPRs was surprisingly patchy in phylogroup A and the genus as a whole; however, even more surprisingly was that putative CRIPSRs were absent in phylogroup B indicating its members may be devoid of them entirely (**Figure 5**). To assess if the absence

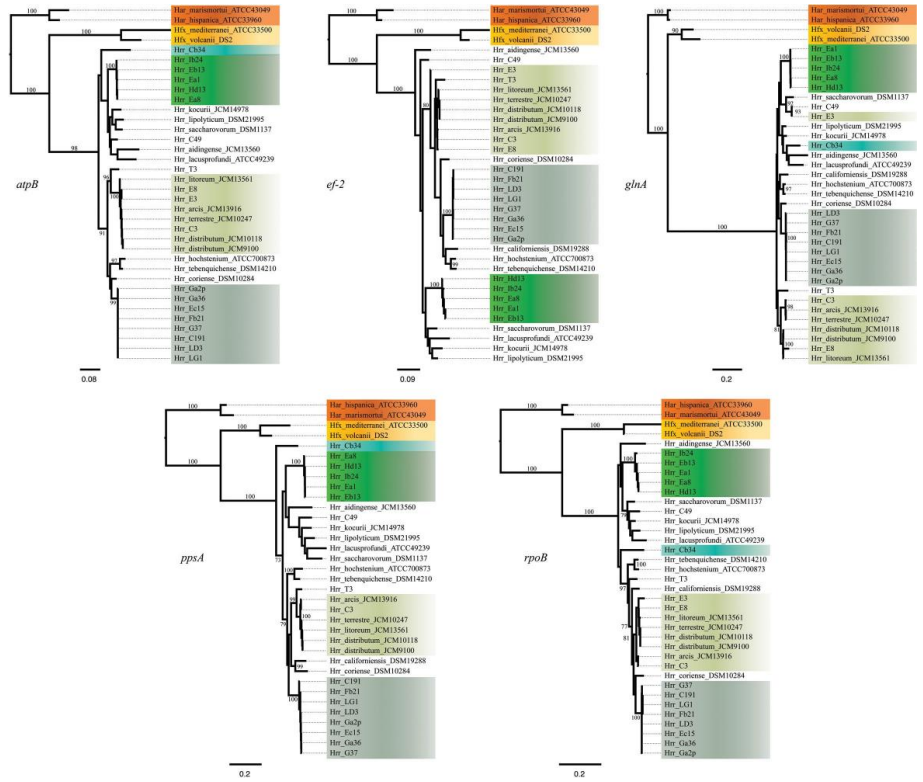


FIGURE 1 | Maximum-likelihood gene trees made from the DNA sequences of *atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB*. Support values on branches are bootstrap replicates. Bootstrap values below 70 are not displayed.

of CRISPRs was an artifact of using draft genome assemblies, we tested for a correlation by relating N50 to CRISPR counts per genome and found there to be no correlation ($R^2 = 0.105$, $P > 0.05$). Therefore, the CRISPR absences do not appear to be a result of genome assembly.

PHYLOGROUP D IS NOT A COHESIVE AND DISCREET ENTITY

Phylogroup D appeared in the phylogenetic reconstructions of MLSA genes after the inclusion of the NCBI *Halorubrum* genomes. It includes five genomes representing four previously described *Halorubrum* species (*Hrr. arcis*, *Hrr. terrestre*, *Hrr. Distributum,* and *Hrr. litoreum*). It was surprising that multiple named species formed such a unit, but evidence suggests it is not discreet and cohesive like phylogroups A and B: much of the data conflict leading to an ambiguous demarcation of its boundary (see below).

The phylogenetic reconstruction of this cluster is supported by the bootstrap values, with exceptions. The concatenated phylogeny has a bootstrap value of 100 at its base and the individual gene trees each support the cluster with bootstrap value of greater than 80 (**Figures 1**, **2**). Pairwise identity between the MLSA genes shows phylogroup D meets the initial criterion of <1% sequence divergence (**Table 5**). While high, the intra-cluster sequence identity is statistically lower than both phylogroup A and B values ($P < 0.05$, two-tailed *t*-test). ANI gives similar results to the pairwise identity (**Figure 3**): the intra-cluster value is ~97%. However some members of the group do not meet the 96% threshold identity, such as E3. Tetramer analysis shows good cohesion within the group, as all but one genome (E3) passed the cutoff. Both E3 and *Hrr. litoreum*'s tetramer frequency patterns are poorly correlated and are below the 0.99 coefficient cutoff advocated by the JSpecies 1.2.1 (Richter and Rosselló-Móra, 2009) package.

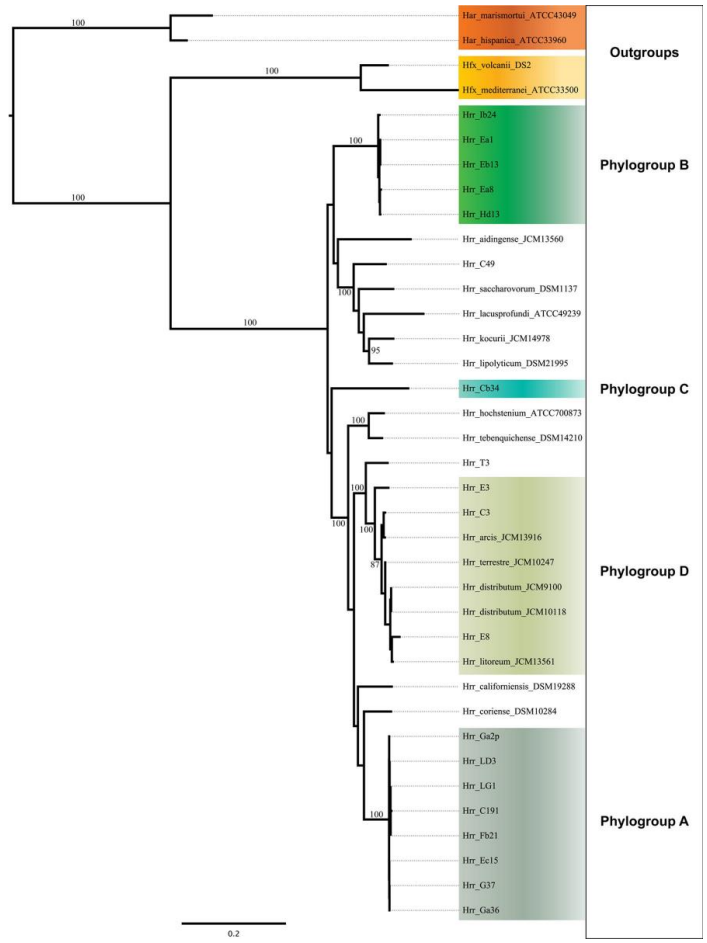
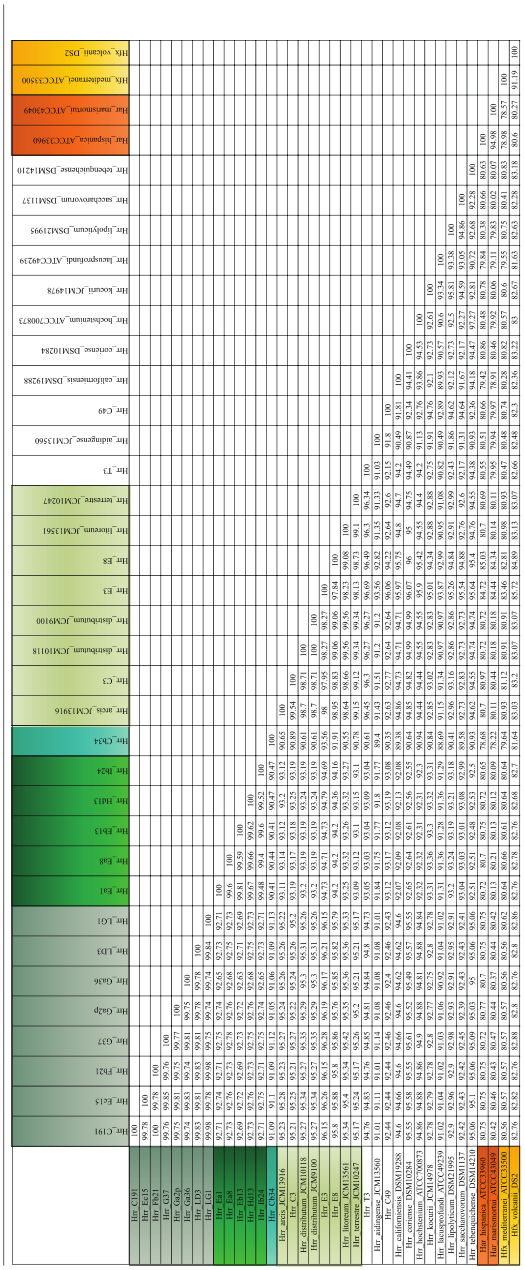


FIGURE 2 | Maximum-likelihood tree made from the concatenated DNA sequences of five housekeeping genes (*atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB*). Support values on branches are bootstrap replicates. Bootstraps values below 70 are not displayed.

As tetramer patterning is largely a granular filter, it strongly suggests that E3 and *Hrr. litoreum* may be distantly related, which is further supported by the ANI analysis.

The phylogroup D intein distribution patterns and sequences identities are dissimilar to phylogroup A and B (**Figure 5**). The intra-phylogroup identity of *pol-II*a is quite low in D compared to phylogroups A and B (~78 vs. ~99% and ~89%, respectively). The inter-group identities are much higher between B and D than in any other phylogroup relationship (~71%). These relationships are partly explained by *Hrr. terrestre*, which features an intein of much greater length and sequence divergence than the other alleles. This intein shares no more than 55% identity with any other phylogroup D *pol-II*a allele. If it is removed from consideration, the phylogroup D intra-cluster identity increases to ~99%. The relatedness to phylogroup A rises to ~53% while the value to phylogroup B is 76%. Intra-phylogroup D *cdc21b*

Table 5 | Pairwise distances of the concatenated alignment of the five MLSA genes.



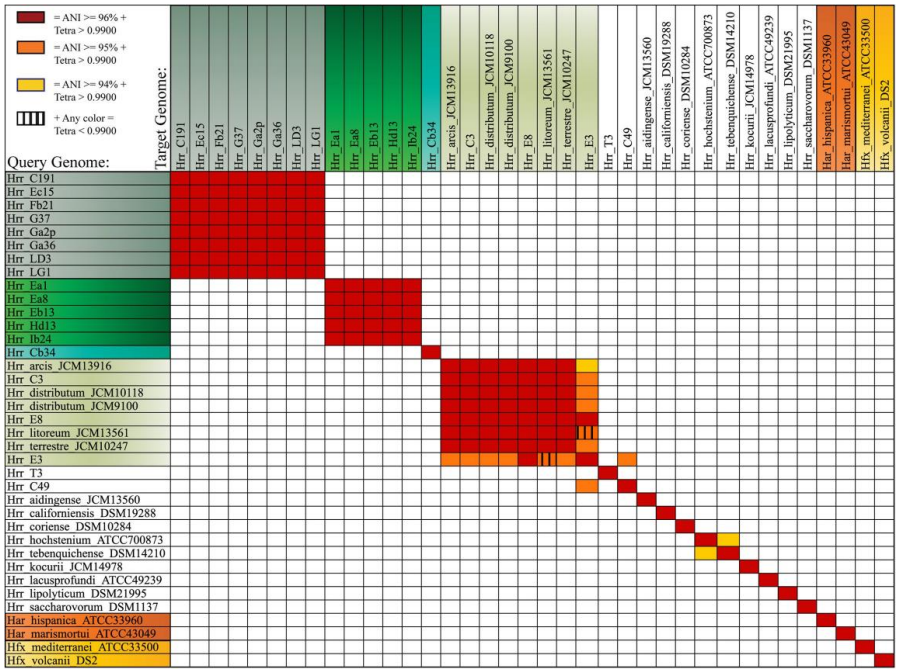


FIGURE 3 | Average Nucleotide Identity (ANI) and tetramer frequency correlation analysis. Color coding reflects three described ANI cutoffs for species delineation. Red squares represent ANI values of 96% or greater, Orange 95% or greater, and yellow represents 94% or greater. The vertical stripes indicate tetramer regression coefficients lower than 0.9900.

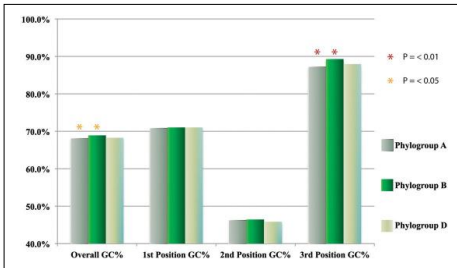


FIGURE 4 | GC usage of all annotated ORFs within and between phylogroups.

diversity is nearly the same as its inter-phylogroup D diversity, which further indicates phylogroup D is a fuzzy entity. The intra-phylogroup identity for the *cdc21b* intein is ∼91% (as compared to ∼100% for A and ∼99% for B) and its inter-phylogroup values are not much lower with D vs. B at ∼83% and D vs. A at ∼87%. However, the remaining taxa (*Hrr. arcis*, *Hrr. litoreum*, *Hrr. distributum*, *Hrr. terrrestre*, E8, and C3), including the named species appear to form a stable phylogroup. These data suggest that phylogroup D as constructed in our analysis is an amalgamation of populations that resembles other analyzed phylogroups but is not a cohesive unit upon additional investigation. The phylogenetic reconstruction derived from the combined presence-absence data and intein sequence data (**Figure 6**) shows that phylogroup D does not retain monophyly. Members place at four locations in the tree. The phylogroup displays high identities for core members, but "fringe" members are at the edge of inclusion.

Hrr. T3 and E3 presented significant challenges to defining the boundary of phylogroup D. As mentioned above, *Hrr.* T3 placed directly sister to the phylogroup in three of five gene phylogenies and inside the group in a fourth (**Figure 1**). In the fifth phylogeny it placed several nodes away from the cluster. The concatenation also places it sister to the cluster with maximum bootstrap support. However, its branch is long relative to the phylogroup. As noted, the pairwise identities and ANI values (**Figure 3**) both

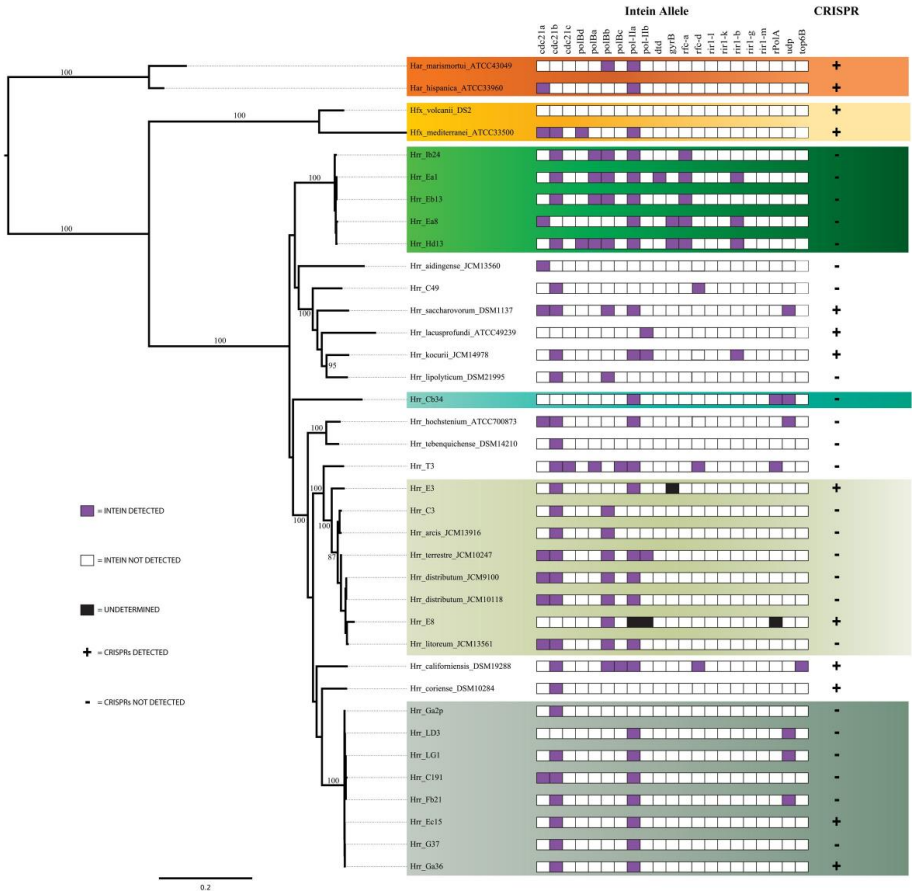


FIGURE 5 | Assessment of the presence of inteins and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs). For inteins, purple boxes indicates the presence of an intein allele, white indicates its absence and black indicates an undetermined result. For CRISPRs a (+) indicates the presence of one or more CRISPRs and a (–) indicates the absence of CRISPRs.

place it below the values seen inside the cluster. These notably lower values were used to exclude this taxon from the phylogroup. *Hrr.* E3 is less of a clean-cut case. Its *glnA* gene is outside of the phylogroup. It also falls on a branch by itself at the base of the cluster with rest of the phylogroup supported by an 87% bootstrap score. However, its intra-cluster pairwise and ANI values are several percent higher than *Hrr.* T3 and only a percent or two below most of the other members of the phylogroup. Overall, the ANI support was on the edge of current cutoffs for species delineation (95% or 96%) (Konstantinidis et al., 2006; Richter and Rosselló-Móra, 2009). Its genome had ANIs ~95% to most of the others in the phylogroup and was only 94% to *Hrr. arcis*. Further, E3's tetramer frequency was also substantially different from *Hrr. litoreum*. A possible explanation for some of these differences is that C49 and E3 show a high degree of sequence identity (95% ANI). It is also C49 with which E3's *glnA* gene associates. Finally, the combined presence-absence and intein phylogeny places these taxa together (**Figure 6**). These data suggest that the two lineages may have engaged in a recent round of genetic exchange, which might explain why E3 is on the periphery

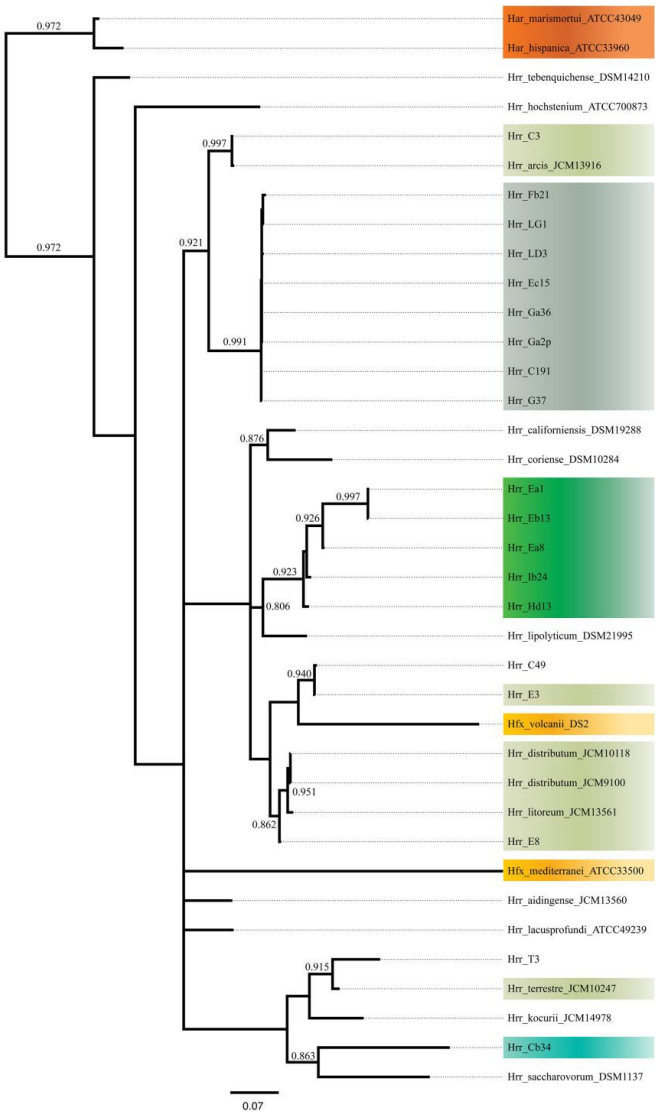


FIGURE 6 | Bayesian tree made from presence-absence of intein alleles and protein sequences of present alleles. Support values on branches are posterior probabilities. Posteriors below 0.8 are not displayed.

of phylogroup D. Ultimately, it was concluded to include E3 as a member of the phylogroup with the acceptance that it was probably an arbitrary distinction in either direction. It was this difficulty in defining the border that resulted in closer examination of phylogroup D and the ultimate rejection of it representing the same sort of entity that phylogroups A and B are.

DISCUSSION

ARE PHYLOGROUPS SPECIES?

The data presented here raise the question: are phylogroups species? We use the term "phylogroup" because a polyphasic analysis (currently defined for the Halobacteria by Oren and Ventosa, 2013) for species description has yet been published on any of the clusters. Still, an evaluation of the data strongly suggests that at least some phylogroups will be eventually described as new species. From the phylogenetic data the perspective provided by the type strain sequences would indicate that phylogroups A and B are unique species. The ANI data support the idea of phylogroups A and B belonging to separate, novel species as several studies advocate cutoffs for species delineation (Konstantinidis and Tiedje, 2005; Konstantinidis et al., 2006; Richter and Rosselló-Móra, 2009) and phylogroupsA andB meet all of them. Additionally, both phylogroups form a cohesive cluster with no particular affinity for other clusters, as evidenced by the strong bootstrap support at the base of each cluster. Also, phylogroups A and B are separated from the others by multiple type strains that place between them. Despite many of these branches being poorly supported, their placement and the strong cohesion within the phylogroups argue that the clusters indicate meaningful phylogenetic splits. These splits likely represent barriers that affect the frequency of gene flow between phylogroups, but not within.

Despite the phylogroups' seemingly species-like attributes, each gene analyzed demonstrates a different topological relationship for them, which means species cannot be viewed as a group of individuals that have a common ancestor, as would be expected from eukaryotic species. While the individual organisms in a prokaryotic species do not share a common ancestor, some of their genes will. For instance, analysis of marine *Vibrio* strains showed that ~1% of the genes within populations shared a common heritage (Shapiro et al., 2012), thus the term species in prokaryotes reflects a process of homogenization, but not heritage, the assumption of Darwinian tree-like speciation. A model that could explain the data is that genes are recombined frequently within *Halorubrum* populations and less so between them. Within the high frequency recombination background new genes that confer selective advantage constantly enter phylogroups from outside the population. These advantageous genes/alleles rise rapidly in frequency throughout the recombining population causing them to diverge in comparison to other phylogroups, yet remaining homogenized within. Like continental drift gives the appearance of discreet units yet are comprised of parts derived from other continents, so too are these two *Halorubrum* phylogroups.

Phylogroup D demonstrates further the model above, as recombination from outside the group is causing divergence, and disallowing a clean species prediction compared to phylogroups A or B. Therefore, phylogroups D is unlikely to be a single species because it is less cohesive in other measurements, which reflects that it contains several previously described species and also that it has engaged in numerous gene exchanges with not-to-distantly-related organisms. Alternatively, since species assignment is a pragmatic endeavor it could be argued from our data and analyses that phylogroup D is a single species with more genetic diversity than found in A and B. The ambiguous relationships of *Hrr.* T3 and E3 suggest there are different recombination partners available to the cluster members. Such differential exchange partners are key elements in microbial speciation (Papke and Gogarten, 2012) and it could be that T3 and E3 are in the process of speciation from the other members of D, but is incomplete. Tetramer frequency data, which has been demonstrated to convey phylogenetic information (Bohlin et al., 2008a,b) casts doubt on the phylogroup representing a single species. It is less stringent than ANI, being more inclusive with the clusters it forms at typical cutoff values (Richter and Rosselló-Móra, 2009). For this reason, when tetramer frequencies are in disagreement it is likely that the two sequences being compared are not closely related. Thus, the tetramer frequency difference between E3 and *Hrr. litoreum* is also strong evidence for those two taxa not belonging to the same species. Interestingly, if T3 and E3 belong to different species and are removed from consideration, the remaining members of phylogroup D would be a single species by all measurements and cutoffs, and yet are still comprised of four named species. However, these strains were isolated from three different geographic regions of Asia at three different time points (Zvyagintseva and Tarasov, 1987; Ventosa et al., 2004; Cui et al., 2007; Xu et al., 2007), from Chinese solar salterns to Turkmenistani saline soils. While the role of geography and ecology in haloarchaeal speciation is unsettled (Oh et al., 2010; DeMaere et al., 2013; Dillon et al., 2013; Zhaxybayeva et al., 2013) all four of the named species have undergone polyphasic characterization, including DNA-DNA hybridization (Ventosa et al., 2004; Cui et al., 2007; Xu et al., 2007). Presumably, if these taxa lived in the same environments and exchanged genes with each other in a positively biased manner like phylogroups A and B, they would be homogenized and indistinguishable by current polyphasic description processes. What sets phylogroup D apart in our analysis is that we do not have population data on members from the same site, and cannot compare equivalently: if we had more data from natural populations like we do for phylogroups A and B, it might be possible to detect reliable differences that separate the named species into different MLSA phylogroups. For example, dozens of *Sulfolobus* strains isolated from geographically distant sites were less than 1% divergent across multiple loci, yet population data analysis demonstrated they fall into discreet clusters associated with geography (Whitaker et al., 2003) While the taxonomy of the Halobacteria is in flux (for example: McGenity and Grant, 1995; Oren and Ventosa, 1996) it seems unlikely that these four separate species will be merged into one. Recent work has served to split *Hrr. terrestre* from *Hrr. distributum* (Ventosa et al., 2004). Thus, it is challenging to conceive of phylogroup D as a single species, which serves as a strong example of the limits

to MLSA and ANI in regards to being the defining measurements of species.

CRISPR DISTRIBUTION MAY BE THE RESULT OF SELECTION

It is important to acknowledge that the patchy CRISPR distribution may be in part an artifact of genome assembly. Repeats can prove a challenge to assembly of short read data (Miller et al., 2010; Magoc et al., 2013) and CRISPRs are repeat heavy. However, false negatives that may exist are unlikely to be directly correlated with assembly quality, and no significant correlation is found between N50 score and the number of CRISPR arrays detected ($P > 0.05$). Additionally, the use of a different CRISPR detector, Crass v0.3.6 (Skennerton et al., 2013), which analyzes raw sequencing reads, rather than finding them in assemblies, supported the CRISPRs reported and found only slight evidence for three additional taxa possessing CRISPRs (data not shown). This would only represent individual CRISPR repeats no larger than about three spacers. While CRISPRs this size have been reported (Kunin et al., 2007) the evidence is inconclusive and if these three taxa do possess CRISPRs their distribution would remain sparse. Only seven of the 18 genomes sequenced in this study would possess them.

CRISPRs have been reported to be very common in the archaea (Jansen et al., 2002; Godde and Bickerton, 2006; Kunin et al., 2007; Held et al., 2010) with reported incidence as high as 90% (Koonin and Makarova, 2009). The incidence in bacteria is closer to 50%. The higher incidence in the archaea may be due to the underrepresentation of archaeal genomes in databases. With viruses and other MGEs so common (for discussion of haloviruses see Dyall-Smith et al., 2003; Porter et al., 2007) and horizontal transfer of CRISPRs a frequent occurrence (Kunin et al., 2007; Sorek et al., 2008), why does selection ever conjure a no-CRISPR lineage? One possibility is that the benefit provided is not strong enough to outweigh the costs, as CRISPR systems require precise matches with their target, and a "protospacer" with one or two mismatches can eliminate functionality (Deveau et al., 2008). The loss of cassettes in CRISPR arrays is not uncommon (Deveau et al., 2008; Díez-Villaseñor et al., 2010; Touchon and Rocha, 2010), while loss of an entire array is less so (Held et al., 2010; Touchon and Rocha, 2010). Possession of large CRISPR arrays may not offer extra protection against the viruses in an environment (Díez-Villaseñor et al., 2010). It might be that if predation level by MGEs rise and fall then the value of the CRISPR system might follow those trends. *Escherichia* and *Salmonella* CRISPR arrays do not appear to deteriorate rapidly enough to be lost entirely and they show a high rate of transfer and loss of the *cas* proteins that form the machinery of the functional system (Touchon and Rocha, 2010). This might suggest that the need for the system may not be constant. Another reason for degradation of the system could be related to it behaving in an auto-immune fashion. When challenged by artificial constructs including a proto-spacer and a gene complementing an autotrophic defect in the strain, *Sulfolobus* cells developed a surprisingly large number of deletion mutants in the spacer providing immunity to the construct (Gudbergsdottir et al., 2011). The authors speculated that there might be some small degree of feedback where the system attacks the host's spacer in addition to that of the MGE. The cellular repair systems may then easily delete the spacer during the repair process. Feedback against self and similar to self DNA, such as targeting closely related housekeeping genes (Gophna and Brodt, 2012) could also impact mating proficiency if the CRISPR system degrades the DNA of exchange partners before it can experience recombination events. It is also important to consider that mechanisms other than CRISPRs have major roles in developing resistance to MGEs (Wilson and Murray, 1991; Bickle and Krüger, 1993; Díez-Villaseñor et al., 2010). For instance, there could be a balance between CRISPRs and restriction/modification systems where one system is lost and another replaces, or complements it such that any one anti-MGE mechanism at any moment in time is in flux.

THE ABSENCE OF INTEINS SUGGESTS BARRIERS TO RECOMBINATION BETWEEN PHYLOGROUPS

Inteins are found pervasively among the archaea (Perler, 2002). They insert into genes and once translated their splicing domains use anauto-catalytic mechanism to self-excise from the protein and re-join the two halves of the polypeptide to generate a functional protein. Inteins associate with homing endonucleases (HEN), found between the splicing domains, to allow their transmission into new hosts. HENs target highly conserved sites in highly conserved genes (Swithers et al., 2009). These HENs appear to be extremely specific in their target sequences as inteins are only found inserted among the most conserved residues of highly conserved protein coding genes (Swithers et al., 2009). Their means of dissemination from host to host is, as yet, unknown although it is clear that it relies on established methods of gene flow within a population (Goddard and Burt, 1999; Gogarten and Hilario, 2006). This suggests that if two hosts have no method of transmitting genes between themselves then the resident inteins will not cross hosts, either. Thus, the patchy distribution of inteins can be interpreted as evidence for a barrier to transfer. This is particularly relevant for the alleles that are not shared between phylogroups A and B. The presence of multiple alleles not seen in the other group argues that the allele has been unable to spread. This is not implying that members of phylogroups A and B do not exchange genes, rather, the sequence divergence and lack of intein spread implies that the recombination process is hindered relative to within group genetic exchange. Indeed, if the mating observed between different *Haloferax* species (see Naor et al., 2012) is possible then almost any sequence divergence between *Halorubrum* phylogroups is akin to a speed bump rather than a mountain in slowing the rate of genetic exchange. Additionally, studies of homologous recombination have found transfers across class-level phylogenetic distance, only at increasingly lower rates as the genetic distance increases (Vulic' et al., 1997; Williams et al., 2012).

AUTHOR CONTRIBUTIONS

Matthew S. Fullmer, J. Peter Gogarten, Antonio Ventosa, and R. Thane Papke participated in the design of this study and helped to draft the manuscript. Shannon M. Soucy generated the intein data and performed the majority of the intein analysis and helped to draft the manuscript. Kristen S. Swithers performed the CRT analysis and helped to draft the manuscript. Andrea M.

Makkay and Ryan Wheeler performed the MLSA PCR. Andrea M. Makkay performed the genome sequencing. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Mohammad A. Amoozegar (University of Tehran, Iran) for allowing us to analyze the Aran-Bidgol strains, and the UConn Bioinformatics Facility for providing computing resources. This research was supported by the National Science Foundation (award numbers, DEB0919290 and DEB0830024) and NASA Astrobiology: Exobiology and Evolutionary Biology Program Element (Grant Number NNX12AD70G).

REFERENCES

Allers, T., Ngo, H.-P., Mevarech, M., and Lloyd, R. G. (2004). Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Appl. Environ. Microbiol.* 70, 943–953. doi: 10.1128/AEM.70.2.943-953.2004

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

AnalystSoft. (2009). *Statistical Analysis Program for Mac OS*. Alexandria, VA: AnalystSoft Inc.

Andam, C. P., Harlow, T. J., Papke, R. T., and Gogarten, J. P. (2012). Ancient origin of the divergent forms of leucyl-tRNA synthetases in the Halobacteriales. *BMC Evolutionary Biology* 12:85. doi: 10.1186/1471-2148-12-85

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bickle, T. A., and Krüger, D. H. (1993). Biology of DNA restriction. *Microbiol. Rev.* 57, 434–450.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., et al. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209. doi: 10.1186/1471-2105-8-209

Bohlin, J., Skjerve, E., and Ussery, D. W. (2008a). Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput. Biol.* 4:e1000057. doi: 10.1371/journal.pcbi.1000057

Bohlin, J., Skjerve, E., and Ussery, D. W. (2008b). Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* 9:104. doi: 10.1186/1471-2164-9-104

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., and Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186, 3980–3990. doi: 10.1128/JB.186.12.3980-3990.2004

Cuadros-Orellana, S., Martin-Cuadrado, A.-B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R. T., et al. (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 1, 235–245. doi: 10.1038/ismej.2007.35

Cui, H.-L., Lin, Z.-Y., Dong, Y., Zhou, P.-J., and Liu, S.-J. (2007). *Halorubrum litoreum* sp. nov., an extremely halophilic archaeon from a solar saltern. *Int. J. Syst. Evol. Microbiol.* 57, 2204–2206. doi: 10.1099/ijs.0.65268-0

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109

DeMaere, M. Z., Williams, T. J., Allen, M. A., Brown, M. V., Gibson, J. A. E., Rich, J., et al. (2013). High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16939–16944. doi: 10.1073/pnas.1307090110

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390–1400. doi: 10.1128/JB.01412-07

Díez-Villaseñor, C., Almendros, C., García-Martínez, J., and Mojica, F. J. M. (2010). Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* 156, 1351–1361. doi: 10.1099/mic.0.036046-0

Dillon, J. G., Carlin, M., Gutierrez, A., Nguyen, V., and McLain, N. (2013). Patterns of microbial diversity along a salinity gradient in the Guerrero Negro solar saltern, Baja CA Sur, Mexico. *Front. Microbiol.* 4:399. doi: 10.3389/fmicb.2013.00399

Dyall-Smith, M. (2009). *The Halohandbook - Protocols for Haloarchaeal Genetics*. Available online at: http://www.haloarchaea.com/resources/halohandbook/index.html

Dyall-Smith, M., Tang, S.-L., and Bath, C. (2003). Haloarchaeal viruses: how diverse are they? *Res. Microbiol.* 154, 309–313. doi: 10.1016/S0923-2508(03)00076-7

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Goddard, M. R., and Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. U.S.A.* 96, 13880–13885. doi: 10.1073/pnas.96.24.13880

Godde, J. S., and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* 62, 718–729. doi: 10.1007/s00239-005-0223-z

Gogarten, J. P., and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evolutionary Biology* 6:94. doi: 10.1186/1471-2148-6-94

Gophna, U., and Brodt, A. (2012). CRISPR/Cas systems in archaea. *Mob. Genet. Elements* 2, 63–64. doi: 10.4161/mge.19907

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J. V. K., Jensen, L. R., She, Q., et al. (2011). Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.* 79, 35–49. doi: 10.1111/j.1365-2958.2010.07452.x

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Held, N. L., Herrera, A., Cadillo-Quiroz, H., and Whitaker, R. J. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS ONE* 5:e12988. doi: 10.1371/journal.pone.0012988

Jansen, R., van Embden, J. D. A., Gaastra, W., and Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* 43, 1565–1575. doi: 10.1046/j.1365-2958.2002.02839.x

Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., and Berg, I. A. (2011). A Methylaspartate cycle in haloarchaea. *Science* 331, 334–337. doi: 10.1126/science.1196544

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 1929–1940. doi: 10.1098/rstb.2006.1920

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Koonin, E. V., and Makarova, K. S. (2009). CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol. Rep.* 1:95. doi: 10.3410/B1-95

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* 8:R61. doi: 10.1186/gb-2007-8-4-r61

Legault, B. A., Lopez-Lopez, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., et al. (2006). Environmental genomics of "*Haloquadratum walsbyi*" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* 7:171. doi: 10.1186/1471-2164-7-171

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., et al. (2012). SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.* 61, 90–106. doi: 10.1093/sysbio/syr095

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., et al. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29, 1718–1725. doi: 10.1093/bioinformatics/btt273

McGenity, T. J., and Grant, W. D. (1995). Transfer of *Halobacterium saccharovorum*, *Halobacterium sodomense*, *Halobacterium trapanicum* NRC 34021 and *Halobacterium lacusprofundi* to the Genus *Halorubrum* gen. nov., as *Halorubrum saccharovorum* comb. nov., *Halorubrum sodomense* comb. nov., *Halorubrum trapanicum* comb. nov., and *Halorubrum lacusprofundi* comb. nov. *Syst. Appl. Microbiol.* 18, 237–243. doi: 10.1016/S0723-2020(11)80394-2

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95, 315–327. doi: 10.1016/j.ygeno.2010.03.001

Naor, A., Lapierre, P., Mevarech, M., Papke, R. T., and Gophna, U. (2012). Low species barriers in halophilic Archaea and the formation of recombinant hybrids. *Curr. Biol.* 22, 1444–1448. doi: 10.1016/j.cub.2012.05.056

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J. O., et al. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20537–20542. doi: 10.1073/pnas.1209119109

Oh, D., Porter, K., Russ, B., Burns, D., and Dyall-Smith, M. (2010). Diversity of *Haloquadratum* and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. *Extremophiles* 14, 161–169. doi: 10.1007/s00792-009-0295-6

Okuda, Y., Sasaki, D., Nogami, S., Kaneko, Y., Ohya, Y., and Anraku, Y. (2003). Occurrence, horizontal transfer and degeneration of VDE intein family in Saccharomycete yeasts. *Yeast* 20, 563–573. doi: 10.1002/yea.984

Oren, A., and Ventosa, A. (1996). A proposal for the transfer of *Halorubrobacterium distributum* and *Halorubrobacterium coriense* to the genus Halorubrum as *Halorubrum distributum* comb. nov. and *Halorubrum coriense* comb. nov., respectively. *Int. J. Syst. Bacteriol.* 46, 1180–1180. doi: 10.1099/00207713-46-4-1180

Oren, A., and Ventosa, A. (2013). Subcommittee on the taxonomy of Halobacteriaceae and Subcommittee on the taxonomy of Halomonadaceae: minutes of the joint open meeting, 24 June 2013, Storrs, Connecticut, USA. *Int. J. Syst. Evol. Microbiol.* 63, 3540–3544. doi: 10.1099/ijs.0.055988-0

Papke, R. T., and Gogarten, J. P. (2012). How bacterial lineages emerge. *Science* 336, 45–46. doi: 10.1126/science.1219241

Papke, R. T., Koenig, J. E., Rodríguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of *Halorubrum*. *Science* 306, 1928–1929. doi: 10.1126/science.1103289

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14092–14097. doi: 10.1073/pnas.0706358104

Perler, F. B. (2002). InBase: the Intein Database. *Nucleic Acids Res.* 30, 383–384. doi: 10.1093/nar/30.1.383

Porter, K., Russ, B. E., and Dyall-Smith, M. L. (2007). Virus–host interactions in salt lakes. *Curr. Opin. Microbiol.* 10, 418–424. doi: 10.1016/j.mib.2007.05.017

Rhodes, M. E., Spear, J. R., Oren, A., and House, C. H. (2011). Differences in lateral gene transfer in hypersaline versus thermal environments. *BMC Evolutionary Biology* 11:199. doi: 10.1186/1471-2148-11-199

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., et al. (2012). Population genomics of early events in the ecological differentiation of Bacteria. *Science* 336, 48–51. doi: 10.1126/science.1218198

Sharma, A. K., Spudich, J. L., and Doolittle, W. F. (2006). Microbial rhodopsins: functional versatility and genetic mobility. *Trends Microbiol.* 14, 463–469. doi: 10.1016/j.tim.2006.09.006

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7. doi: 10.1038/msb.2011.75

Skennerton, C. T., Imelfort, M., and Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* 41, e105. doi: 10.1093/nar/gkt183

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* 6, 181–186. doi: 10.1038/nrmicro1793

Swithers, K. S., Senejani, A. G., Fournier, G. P., and Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evolutionary Biology* 9:303. doi: 10.1186/1471-2148-9-303

Swithers, K. S., Soucy, S. M., Lasek-Nesselquist, E., Lapierre, P., and Gogarten, J. P. (2013). Distribution and evolution of the mobile *vma-1*b intein. *Mol. Biol. Evol.* 30, 2676–2687. doi: 10.1093/molbev/mst164

Touchon, M., and Rocha, E. P. C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE* 5:e11126. doi: 10.1371/journal.pone.0011126

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304

Ventosa, A., Gutiérrez, M. C., Kamekura, M., Zvyagintseva, I. S., and Oren, A. (2004). Taxonomic study of *Halorubrum distributum* and proposal of *Halorubrum terrestre* sp. nov. *Int. J. Syst. Evol. Microbiol.* 54, 389–392. doi: 10.1099/ijs.0.02621-0

Vulic´, M., Dionisio, F., Taddei, F., and Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 94, 9763–9767. doi: 10.1073/pnas.94.18.9763

Whitaker, R. J., Grogan, D. W., and Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* 301, 976–978. doi: 10.1126/science.1086909

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4, 1223–1244. doi: 10.1093/gbe/evs098

Wilson, G. G., and Murray, N. E. (1991). Restriction and modification systems. *Annu. Rev. Genet.* 25, 585–627. doi: 10.1146/annurev.ge.25.120191.003101

Xu, X.-W., Wu, Y.-H., Zhang, H., and Wu, M. (2007). *Halorubrum arcis* sp. nov., an extremely halophilic archaeon isolated from a saline lake on the Qinghai–Tibet Plateau, China. *Int. J. Syst. Evol. Microbiol.* 57, 1069–1072. doi: 10.1099/ijs.0.64921-0

Zhaxybayeva, O., Stepanauskas, R., Mohan, N. R., and Papke, R. T. (2013). Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles* 17, 265–275. doi: 10.1007/s00792-013-0514-z

Zvyagintseva, I. S., and Tarasov, A. L. (1987). Extreme halophilic bacteria from saline soils. *Mikrobiologiya* 56, 839–844.

Conflict of Interest Statement: The authors declare that the research was con- ducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 January 2014; accepted: 18 March 2014; published online: 11 April 2014. Citation: Fullmer MS, Soucy SM, Swithers KS, Makkay AM, Wheeler R, Ventosa A, Gogarten JP and Papke RT (2014) Population and genomic analysis of the genus Halorubrum. Front. Microbiol. 5:140. doi: 10.3389/fmicb.2014.00140

This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology.

4.2 Inteins as Indicators of Gene Flow in the *Halobacteria*

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 June 2014
doi: 10.3389/fmicb.2014.00299

Inteins as indicators of gene flow in the halobacteria

Shannon M. Soucy, Matthew S. Fullmer, R. Thane Papke and Johann Peter Gogarten*

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Edited by:
Jesse Dillon, California State University, Long Beach, USA
Reviewed by:
Julie L. Meyer, University of Florida, USA
Kenneth Mills, College of the Holy Cross, USA
****Correspondence:***
Johann Peter Gogarten, Microbiology Program, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Storrs, CT 06269-3125, USA
e-mail: gogarten@uconn.edu; jpgogarten@gmail.com

This research uses inteins, a type of mobile genetic element, to infer patterns of gene transfer within the Halobacteria. We surveyed 118 genomes representing 26 genera of Halobacteria for intein sequences. We then used the presence-absence profile, sequence similarity and phylogenies from the inteins recovered to explore how intein distribution can provide insight on the dynamics of gene flow between closely related and divergent organisms. We identified 24 proteins in the Halobacteria that have been invaded by inteins at some point in their evolutionary history, including two proteins not previously reported to contain an intein. Furthermore, the size of an intein is used as a heuristic for the phase of the intein's life cycle. Larger size inteins are assumed to be the canonical two domain inteins, consisting of self-splicing and homing endonuclease domains (HEN); smaller sizes are assumed to have lost the HEN domain. For many halobacterial groups the consensus phylogenetic signal derived from intein sequences is compatible with vertical inheritance or with a strong gene transfer bias creating these clusters. Regardless, the coexistence of intein-free and intein-containing alleles reveal ongoing transfer and loss of inteins within these groups. Inteins were frequently shared with other Euryarchaeota and among the Bacteria, with members of the Cyanobacteria (*Cyanothece, Anabaena*), Bacteriodetes (*Salinibacter*), Betaproteobacteria (*Delftia, Acidovorax*), Firmicutes (*Halanaerobium*), Actinobacteria (*Longispora*), and Deinococcus-Thermus-group.

Keywords: gene symbiosis, genome as an ecosystem, inteins, mobile genetic elements, gene flow, horizontal gene transfer, halobacteria

INTRODUCTION

Inteins are self-splicing genetic parasites located in highly conserved sites of slowly evolving genes. They are found in all three domains of life and in viruses (Perler et al., 1997; Pietrokovski, 2001; Gogarten et al., 2002; Swithers et al., 2009). Similar to group I introns, inteins are often associated with a homing endonuclease (HEN). An important difference between inteins and introns is the timing of the splicing activity, which occurs immediately after transcription in introns and after translation in inteins (Hirata et al., 1990; Kane et al., 1990). The association with a HEN domain enables a cyclic invasion pattern, called the homing cycle (Goddard and Burt, 1999; Gogarten and Hilario, 2006). The homing cycle consists of three phases: intein invasion, intein fixation, and eventually loss of the intein enabling invasion to occur again. During invasion and fixation the intein splicing domains are associated with a HEN domain forming a canonical intein (hereafter referred to as a large intein); however, during the loss phase the function of the HEN is often disrupted and begins to degrade, generating a mini-intein. Simulations have shown that intein-containing and intein-free alleles can coexist in well mixed populations under some sets of parameters (Yahara et al., 2009; Barzel et al., 2011). Also, inteins with functioning HEN domains were inferred to have persisted in some eukaryotic lineages for several 100 million years (Butler et al., 2006; Gogarten and Hilario, 2006).

Inteins do not have an apparatus to penetrate the cell envelope. Therefore, they must rely on mechanisms in place within the population for insertion into the cell such as: conjugation, mating, generalized DNA uptake, and viruses or gene transfer agents (Lang et al., 2012). The faster-than-Mendelian inheritance of the large inteins (Gimble and Thorner, 1992), along with a nearly neutral fitness burden, enables these mobile elements to persist in organisms over evolutionary time as long as there are new populations to invade (Goddard and Burt, 1999; Gogarten and Hilario, 2006). Furthermore, the size of the intein (mini or large) provides information about the genomic mobility of the element as mini inteins are rarely integrated into the recipient's genome; whereas large inteins are more frequently integrated due to the activity of the HEN. The conservation of the recognition site provides an invasion target even in distantly related strains and species. Also, inteins have a higher substitution rate relative to their extein hosts (Swithers et al., 2013). This substitution rate gives rise to many evolutionarily informative sites when comparing a large collection of homologous inteins. In this work, we take advantage of these traits and survey the distribution of inteins in the Halobacteria, a highly recombinant class of halophilic Archaea (Williams et al., 2012) known to contain several intein alleles (Perler, 2002). We make use of 118 halobacterial genomes (Supplementary Table 1) and the previously reported and newly discovered intein alleles to survey networks of gene transfer within and outside the

Halobacteria based on the presence-absence profile of the inteins, their sequence similarity, and the phylogenies reconstructed from intein sequences.

MATERIALS AND METHODS

HALOBACTERIAL INTEIN SEQUENCE RETRIEVAL AND ALIGNMENT

Position specific scoring matrices (PSSMs) were created using the collection of all inteins from InBase, the Intein database and registry (Perler, 2002). A custom database was created with all inteins, and each intein was used as a seed to create a PSSM using the custom database. These PSSMs were then used as a seed for PSI-BLAST (Altschul, 1997) searches against each of the halobacterial genomes available from NCBI as of June 2013 as well as a private collection sequenced by our collaborators. To remove false positives, a size exclusion step was then performed on each protein sequence as an intein domain adds 100–700 aa to invaded protein sequences. Inteins were then aligned using Muscle (Edgar, 2004) with default parameters in the SeaView version 4.0 software package (Gouy et al., 2010). Insertions, which passed the size exclusion step, but did not contain splicing domains, were removed and the previous steps were repeated using the resulting dataset on a collection of private genomes from the Papke lab. Prottest 3.2 (Guindon et al., 2010; Darriba et al., 2011) was used to determine an appropriate substitution model for the intein sequences, the WAG model was favored and used for all subsequent trees for consistency. Once the collection of halobacterial inteins was complete, sequences were re-aligned using SATé (Liu et al., 2012) to generate a final alignment using MAFFT (Katoh and Standley, 2013) to align, Muscle (Edgar, 2004) to merge, RAXML (Stamatakis, 2014) for tree estimation, and a WAG model for each allele.

To determine the relationship among all halobacterial inteins, the inteins were aligned using Muscle (Edgar, 2004). Subsequently a tree was built using PhyML v3.0 (Guindon et al., 2010) using a WAG substitution model with a Gamma shape parameter and the proportion of invariant sites estimated from the data.

INTEIN RETRIEVAL OUTSIDE THE HALOBACTERIA

Each halobacterial intein was used as a BLAST (Altschul et al., 1990) query against the non-redundant database on NCBI. Any match with an *e*-value better than 0.000001 was aligned to the dataset to which its query belonged. Sequences were then filtered based on the protein annotation and goodness of fit to the existing alignment. As an additional filtering step each match was used as a query against the non-redundant database and the majority BLAST hit annotations were used to verify the protein identity, as annotations are not always reliable. Remaining sequences were aligned using Clustal Omega 1.1.0 (Sievers et al., 2011) with the profile alignment option in SeaView 4.0 (Gouy et al., 2010). Maximum-likelihood trees were built using PhyML (Guindon et al., 2010) with the WAG model, and rates estimated from the data.

To assess the relative contribution of different genera represented in each intein allele sequence data set, a stacked column graph was created. Sequence density was calculated for each intein allele by dividing the number of intein sequences in each genus by the number of total intein sequences in that allele.

SYMBIOTIC STATE ASSIGNMENT

Intein sequence length was used to determine symbiotic state. For each intein allele the length of the intein sequence was determined. A cutoff length for mini-intein assignment was based on the presence of a gap in intein lengths greater than 100 amino acids within an allele. The third intein state "no-intein" was assigned where the intein was clearly absent from the orthologous protein containing an intein in any of the halobacterial genomes examined. Additionally, once an intein was noted as a mini-intein the alignment was analyzed to ensure the gaps in these sequences correspond to the location of the HEN domain.

RIBOSOMAL PROTEIN REFERENCE TREE

Alignments of 55 ribosomal protein for 21 Halobacteria (Williams et al., 2012) were used to find orthologous proteins in the genomes used in this work. In-house python scripts (data file 1) were used to concatenate the alignments, and PhyML v3.0 (Guindon et al., 2010) was used to build a tree. The tree used the WAG substitution model with the Gamma shape parameter and the proportion of invariant sites and base frequencies estimated from the data.

BAYESIAN CLUSTERING WITH INTEIN SEQUENCES

A concatenation of an intein presence-absence matrix and alignments for each intein allele were generated using in-house python scripts (data file 1). MrBayes version 3.2.1 (Ronquist et al., 2012) was then used to perform a clustering analysis using a partition allowing for character states in the presence-absence matrix and sequence information for each intein allele. The prior for the character portion of the data matrix used a symmetrical Dirichlet distribution with an exponential (1.0), and variable rates so each column was considered independent of the others. The likelihood for the character portion of the alignment used variable coding and 5 beta categories. The prior for the protein sequences in the alignment used a fixed WAG substitution model, with state frequencies estimated from the data, and the likelihood settings used a Gamma shape parameter and the proportion of invariant sites estimated from the data.

RESULTS

HALOBACTERIAL INTEINS

The intein content of a collection of halobacterial genomes was analyzed using an intein-allele-specific PSSM. This survey revealed 13 genes in the Halobacteria invaded by inteins at 24 distinct positions (intein alleles) (**Table 1**). Seven of these intein alleles were not previously reported in the Halobacteria, and two of the seven have not previously been reported to harbor inteins: a DNA ligase gene involved in double strand break repair, and a deaminase gene involved in nucleotide metabolism (**Table 1**). To determine if vertical inheritance was accountable for the distribution of intein alleles, the presence–absence matrix of intein alleles was mapped onto a reference phylogeny (**Figure 1**). Clearly, intein presence-absence is not concordant with the ribosomal protein phylogeny, implicating abundant horizontal genetic transfer (HGT) in creating the observed distribution. The presence of multiple intein alleles in the majority of genomes (70%) might be interpreted

Table 1 | Exteins in the halobacteria.

Intein allele	Extein annotation
cdc21-a	Cell division control protein 21
cdc21-b	
cdc21-c	
polB-d	DNA polymerase B1
polB-a	
polB-b	
polB-c	
*pol-II*a	DNA polymerase II large subunit
*pol-II*b*	
*dtd***	Deoxycytadine triphosphate deaminase
gyrB	DNA gyrase subunit B
helicase-b*	ATP-dependent helicase
*ligase***	ATP-dependent DNA ligase I
rfc-a	Replication factor C small subunit
rfc-d*	
rir1-l*	Ribonucleoside-diphosphate reductase
rir1-k	
rir1-b	
rir1-g	
rir1-m*	
rpolA A	DNA-directed RNA polymerase subunit
udp	UDP-glucose 6-dehydrogenase
topA	DNA topoisomerase I
top6B	DNA topoisomerase VI subunit B

*Denotes intein alleles discovered in this work.

**Denotes extein sequences not previously reported to be invaded by an intein.

to suggest that inteins could spread locally within a single genome.

INTEIN PROPAGATION WITHIN THE HALOBACTERIA

To address the possibility of inteins moving locally within a genome, the phylogenetic relationships among all halobacterial intein sequences were analyzed (**Figure 2**). All of the intein alleles form highly supported clusters with others of the same type, with the exception of two sequences: the *polB*-c inteins of *Haloferax larsenii* and *Haloferax elongans* group inside the *polB*-b intein allele cluster; however, this node is poorly supported (59/100 bootstraps) indicating this relationship could be an artifact produced by poor resolution of the relationships that connect various intein alleles. Furthermore, there is poor support linking all of the intein allele clusters together (less than 70% bootstrap support), indicating sequence conversion (an intein invading an ectopic or atypical locus) between intein alleles, even within the same host protein, is uncommon. Among the inteins analyzed here, at most one invasion of an ectopic site is supported by the data, confirming that this type of event is rare (Perler et al., 1997; Gogarten et al., 2002). These data indicate that HGT is the only plausible explanation for the large number of different intein alleles in this class of organisms. Incongruence between the presence of inteins and ribosomal phylogeny also support this conclusion.

BAYESIAN PHYLOGENETIC ANALYSIS OF INTEINS

In an attempt to resolve the local events (transfers and vertical inheritance within the Halobacteria) that gave rise to the observed intein distribution in the Halobacteria, a Bayesian analysis based on the intein sequences for each allele and on the presence-absence pattern was performed (**Figure 3**). In this analysis two organisms may group together because they both inherited inteins from a common ancestor, or because an intein was recently transferred between them. The paucity of well-supported nodes (nodes with 0.95 or greater posterior probability were considered well-supported) in part reflects the extent to which our sample is biased toward very similar sequences (31% of halobacterial genomes in this study are from *Halorubrum*). Most of the well-supported clusters in the Bayesian tree also occur in the reference tree, suggesting these inteins may be the result of shared vertical inheritance. However, many of these clusters do not have identical intein profiles (clusters 1, 6, 8, and 10), thus HGT between close relatives is a better explanation than vertical inheritance for these clusters. Only three of the clusters, 2, 9, and 12, have branching orders that are different from those observed in the reference tree indicating HGT. Cluster 2 is made up of *Natrinema* spp. *pellirubrum* and *versiforme* which share only the *pol-II*a intein. In the reference tree *Nnm. versiforme* groups with the rest of the *Natrinema*, and *Nnm. pellirubrum* groups with *Haloterrigena thermotolerans*. *Natrinema sp.* J7-2 is the only other member of the *Natrinema* that has an intein in the *pol-II*a position, but the intein in this species is 14 aa shorter than the intein shared by *Nnm. pellirubrum* and *Nnm. versiforme*. *Htg. thermotolerans* shares no inteins with *Nnm. pellirubrum*. Cluster 9 is made up of *Halorubrum* spp. C49 and E3, which share only the *cdc21*-b intein. In the reference tree *Hrr.* E3 groups with *Halorubrum litoreum* and the two share the *pol-II*a intein allele, but no others. *Hrr.* C49 groups with *Halorubrum saccharovorum* and they do not share any inteins. Cluster 12 is made up of *Haloferax* spp. *denitrificans, lucentense, alexandrinus*, and *Haloferax sp.* BAB2207, which all have an intein in the *cdc21*-a position. In the reference tree *Hfx. lucentense, Hfx.* sp. BAB2207, and *Hfx. alexandrinus* all group together, but *Hfx. denitrificans* groups with *Haloferax sulfurifontis,* and they do not share any inteins. The lack of shared inteins between clusters in the reference tree and differences among the inteins shared in these clusters cause these divergences in this tree as compared to the reference tree. This may indicate that the taxa in the Bayesian clusters are exchange partners, or that they share unsampled intermediate exchange partners. Additionally, the majority of clusters share 2 or fewer intein alleles between all members of the cluster (eight out of 12 clusters). The two clusters that share the most intein alleles between all members are Cluster 3, made up of *Haloqudratum walsbyi* strains DSM 16790 and C23 with 13 shared intein alleles, and cluster 7 made up of *Halorubrum* spp. strains SP3 and SP9 sharing 4 intein alleles. Both of these clusters have branching patterns identical to those on the reference tree, indicating that phylogenetic proximity plays a significant role in intein distribution.

Members of the *Halorubrum* genus, not surprisingly, were highly represented in the clusters (four of 12 total). All four of the clusters show a geographic bias. Clusters 6, 8, and 9 were all isolated from the Aran-Bidgol lake in Iran, and cluster 7 was isolated from the Sedom Ponds in Israel (Atanasova et al., 2012). Branch lengths in all of these clusters are very small, suggesting these populations are well mixed with respect to intein sequences. Geography does not seem to play a strong role in linking other

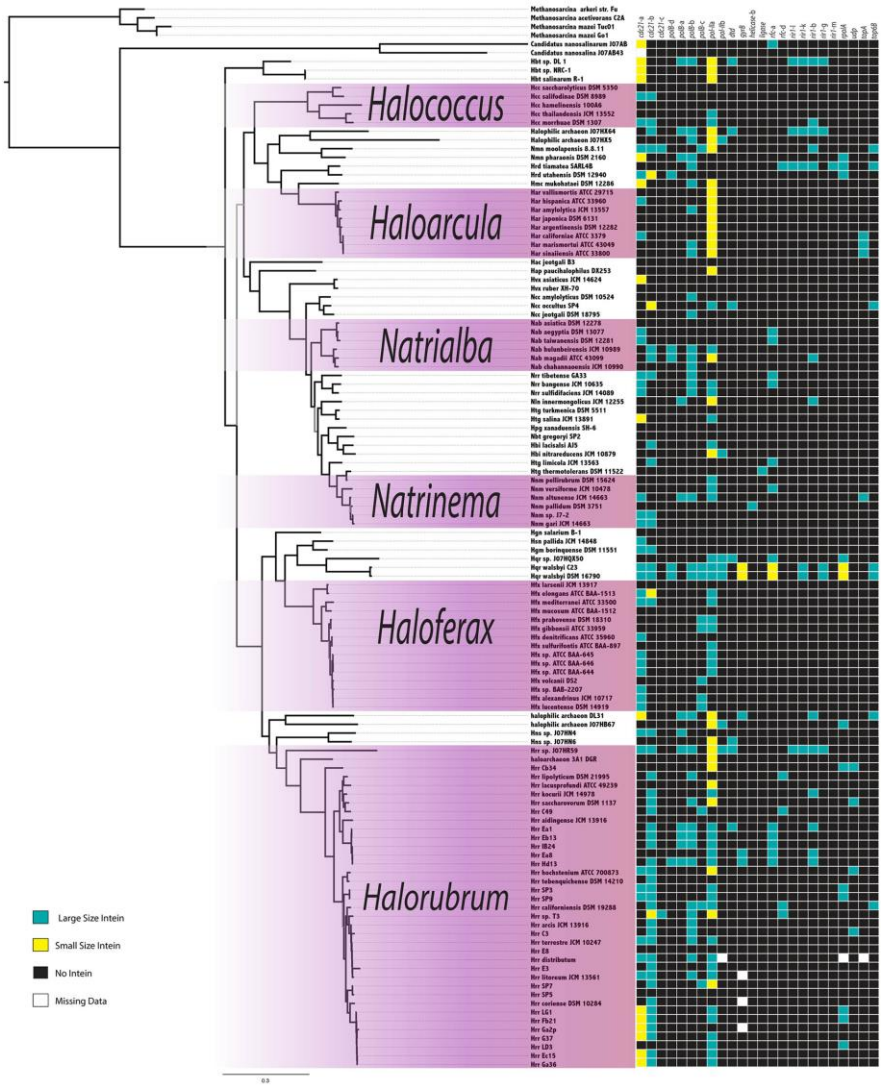


FIGURE 1 | Intein Invasion Pattern in the Halobacteria. Intein pattern of presence-absence is mapped onto the tips of a ribosomal reference tree, teal boxes indicate the presence of a full size intein, yellow boxes indicate the presence of a mini-intein, black boxes indicate the absence of an intein, and white boxes indicate missing data. Purple shaded boxes indicate the genera with more than five species represented on the tree. Nodes with bootstrap support <70 are in gray!"

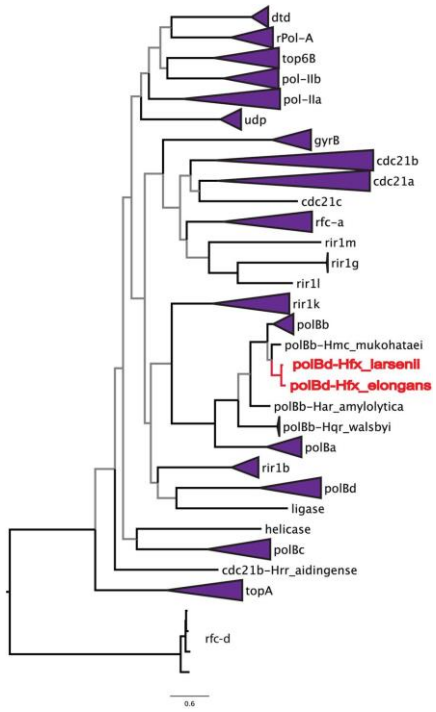


FIGURE2|RelationshipsamonginteinallelesintheHalobacteria. This tree depicts the phylogenetic relationships among intein alleles in the Halobacteria. Inteins that clustered in concordance with the allele were collapsed to a single node, and labeled with the name of the intein allele. Two *polB*-d sequences did not group with the *polB*-d allele, and instead are located amongst the *polB*-c alleles these are indicated in red. Nodes with bootstrap support <70 are colored gray.

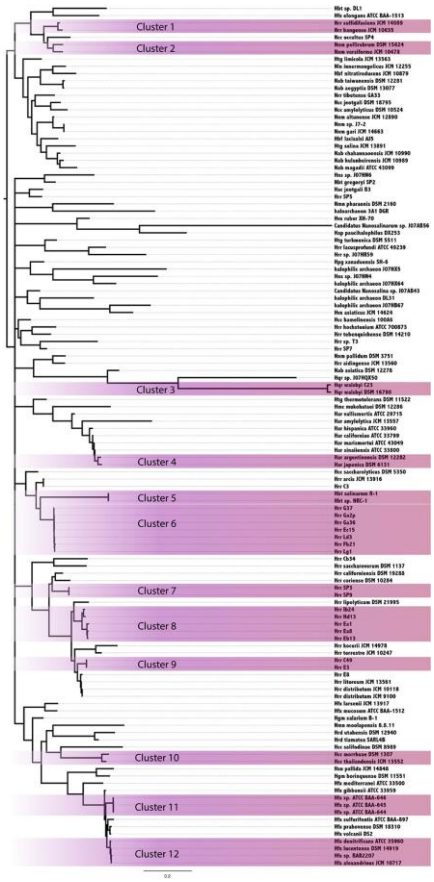


FIGURE 3 | Clustering of Halobacteria based on intein sequences and distribution. Halobacteria were clustered based on intein sequences and the distribution in each genome. Clusters with posterior probability >95% are shaded purple.

well-supported clusters based on intein sequences. Furthermore, evidence of clustering based on geography in the *Halorubrum* is less interesting than the clear separation between groups isolated from the same location (cluster 6, 8, and 9). This separation of species of *Halorubrum* from the same location is echoed in the reference tree, and taken together with the short branch lengths in these clusters indicate that population structure plays a strong role in gene sharing at least for this location (see Fullmer et al., 2014 for in depth discussion). Increased geographical sampling could reveal similar trends in other locations.

INTEIN HOMING IN THE HALOBACTERIA

The existence of a singleton in an intein allele in the genomes analyzed could represent intein invasion from outside the Halobacteria; but could also be due to incomplete sampling. To investigate the phylogenetic distance of invasion events responsible for the observed distribution of inteins, the halobacterial inteins were used as queries to search for homologous sequences in the non-redundant database (Altschul et al., 1990). Intein sequences that matched the alleles in the Halobacteria were found in other Euryarchaeota (but not Crenarchaeota), and Bacteria (**Table 2**). To ascertain whether homing occurred between the Halobacteria and organisms outside the Halobacteria, a maximum likelihood tree was built for each intein allele. The

Table 2 | Taxonomic distribution in each intein allele.

Intein allele	Tree topology	Halobacteria	Bacteria	Other Euryarchaeota
cdc21-a	Monophyletic	55	4	16
cdc21-c	Monophyletic	1	0	0
*dtd***	Monophyletic	6	0	0
gyrB	Monophyletic	6	19	1
*helicase-b**	Monophyletic	1	2	1
*ligase***	Monophyletic	1	0	0
*pol-IIb**	Monophyletic	9	0	1
polB-d	Monophyletic	6	0	1
rfc-a	Monophyletic	16	0	13
*rfc-d**	Monophyletic	5	0	0
rir1-b	Monophyletic	15	55	5
rir1-g	Monophyletic	4	15	0
rir1-k	Monophyletic	5	1	0
*rir1-l**	Monophyletic	3	3	0
rpolA	Monophyletic	10	0	0
top6B	Monophyletic	8	0	0
topA	Monophyletic	4	0	1
udp	Monophyletic	7	2	6
*rir1-m**	Monophyletic	1	4	0
polB-c	Monophyletic	20	1	1
polB-a	Polyphyletic-bacteria	16	2	1
polB-b	Polyphyletic-bacteria	38	3	0
pol-IIa	Polyphyletic-Euryarchaeota	75	0	16
cdc21-b	Polyphyletic-Euryarchaeota	51	1	3

**Denotes intein alleles discovered in this work.*
***Denotes exteins discovered in this work.*

tree topologies were evaluated with respect to the halobacterial inteins. If the halobacterial inteins in the tree were monophyletic it was assumed that except for the initial invasion gene flow for that intein allele occurred within the Halobacteria exclusively. If the halobacterial inteins were polyphyletic, invasion events that generated the observed distribution likely involved organisms outside the Halobacteria either as donors or as recipients. The majority of intein trees, 83%, were monophyletic, reinforcing the idea that recombination is more successful between closely related organisms (Gogarten et al., 2002; Zhaxybayeva et al., 2006; Andam et al., 2010; Papke and Gogarten, 2012; Williams et al., 2012). Interestingly, for trees where the Halobacteria were polyphyletic, the organisms interrupting the clade were Bacteria for two out of the four polyphyletic intein alleles. The sample size restricts building strong claims about HGT between the Halobacteria and the Bacteria. However, this claim is supported by previous evidence of gene exchange between the Bacteria and the Halobacteria (Ng et al., 2000; Khomyakova et al., 2011).

The tight clustering of halobacterial intein sequences and short branches between closely related strains indicate that in the majority cases inteins are inherited vertically or are transferred between closely related strains, and that successful invasion across large genetic distances is rare. Thus, intein alleles that are found in many different genera have been active for many generations, enabling invasion of many lineages, and accumulating examples of rare invasion events such as those that cross domain boundaries. Conversely, a lack of taxonomic diversity cannot be interpreted as a recent invasion as sampling limitations could be responsible for the paucity of samples in that intein allele. While many factors influence the success of intein transfer between divergent organisms, phylogenetic diversity of the organisms invaded by a particular intein allele also is a reflection of the time the intein allele has been present in a linage. Furthermore, a high density of intein sequences in a particular domain or group of genera can be used to determine the most likely reservoir for the circulating intein allele. A stacked column chart was used to quantify the representation of each of the genera in each of the intein alleles (**Figure 4**). Five intein alleles, *cdc21*b, *pol-II*a, *polB*b, *cdc21*a, and *rfc*-d, show polarity in intein density favoring the Halobacteria (specifically *Halorubrum*) as the reservoir for the intein population. This is not surprising as the data indicate that the majority of intein transfer in the Halobacteria is within the class. Additionally, the diversity in five of the intein alleles, *helicase*-b, *cdc21*a, *gyrB*, *rir*1-b, and *udp,* suggests these intein populations may be more ancient than the others in this study as they have had time to accumulate rare, long distance transfers such that the diversity within them spans both class and domain boundaries. Interestingly, the *helicase*-b intein was only recently discovered in this study, though the diversity in the allele gives the impression that this intein has been around for a long time.

TRANSFER OF INTEINS BETWEEN HALOBACTERIAL AND NON-HALOBACTERIAL LINEAGES

Not all inteins are transferred equally; the efficiency of intein invasion is affected largely by the state of the intein. The HEN domain in canonical inteins is required to induce a double strand break and the subsequent homologous repair that results in invasion (Pietrokovski, 2001). Thus, mini-inteins that have lost a functioning HEN domain are mainly transferred vertically (they may be transferred horizontally together with the host gene). If an intein containing allele has been fixed in a population, either a precise deletion of the mini intein encoding DNA could remove the intein from the population or homologous replacement by an intein-free allele transferred from outside the population. Thus, mini-inteins are maintained through strong purifying selection, because any mutation that decreases the self-splicing activity decreases the availability of the host protein (Barzel et al., 2011). The intein states were determined to infer patterns of homing in the Halobacteria. The size of inteins in each allele, along with the position of gaps in the alignment relative to the HEN domain were used as a heuristic for assigning mini-intein status. In most cases there was a clear separation in the distribution of intein lengths (at least 100 amino acids difference in length). The size of more populated intein alleles within the three genera of the Halobacteria with the largest number of available genomes, *Haloarcula*, *Haloferax*, and *Halorubrum*, were recorded in a matrix of intein alleles (**Figure 5**). Many intein alleles show

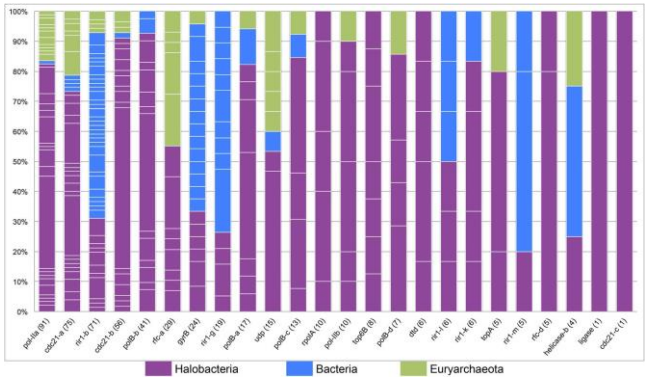


FIGURE 4 | Phylogenetic diversity in halobacterial intein alleles. A stacked column graph depicts the representation of the Halobacteria (in purple), the Bacteria (in blue), and other Euryarchaeota (in green). Intein alleles are ordered by the number of intein sequences recovered for each allele, which is reported in parenthesis after the intein allele name on the *x*-axis. The number of genera for each intein allele is indicated by the number of breaks in the column (white lines) and the height of each of the fragments that make up a column indicate the proportion of sequences in that allele found in a particular genus.

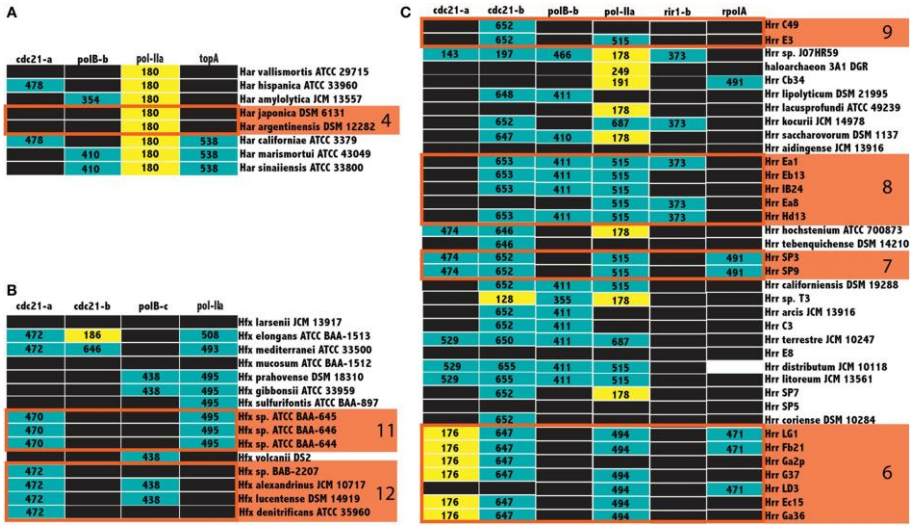


FIGURE 5 | Intein size distributions in the *Haloferax*, *Haloarcula*, and *Halorubrum*. The size of inteins in the *Haloarcula* **(A)**, *Haloferax* **(B)**, and *Halorubrum* **(C)** are indicated in the column corresponding to the intein allele. Mini-inteins are colored yellow, large inteins are colored teal, black boxes indicate no intein, and white boxes indicate missing data, clusters from **Figure 3** are indicated by numbered orange boxes. The *cdc21*-a, and b sequences for *Halorubrum* sp. J07HR59, though smaller than the rest, cannot be considered mini-inteins, as the intein sequences in these positions are complete.

a considerable size variation. This variability can be attributed to the accumulation of insertions and deletions in various lineages over time, which in some lineages leads to loss of the HEN domain. Notably, there is no variability in the size of intein sequences shared by the clusters recovered in the Bayesian analysis (orange boxes **Figure 5)** reinforcing the claim of ongoing gene exchange in these clusters.

Invasion from outside the Halobacteria is one explanation for the polyphyletic topology observed in some halobacterial intein alleles. To determine when these homing events could have occurred, the state of each intein was determined and mapped onto polyphyletic intein allele trees: the results of that analysis are summarized in **Table 3**, with mini-inteins indicated with a star (*), and inteins that group within the Halobacteria indicated by a tilde (~) next to the name of the organism. Many of the intein sequences (5 out of 11) from taxa outside the Halobacteria that interrupt the clade are large-inteins, indicating that interactions between these taxa and the Halobacteria, though rare are ongoing **(Table 3)**. Though the assignment of direction of transfers is extremely preliminary as limited sampling can affect the assignment of direction of transfer, there are some cases with an overwhelming signal where the majority of sequences originate from the Halobacteria, or the Bacteria in the case of *rir1*-m. The mixture of mini and large inteins represented in all of the intein alleles imply most of these inteins are active in the Halobacteria, and notably involve a wide distribution of taxonomic exchange partners.

DISCUSSION

The importance of HGT throughout the tree of life demands the development of a system to monitor gene-flow within and between populations. This research provides fundamental evidence that mobile elements such as inteins can be used to uncover gene flow networks. Inteins have a unique combination of traits that make them ideal tools to study evolution in microbial populations. They have a naturally wide phylogenetic distribution, enabling detection of HGT between distantly related taxa. This is demonstrated in this work by the intein trees where the Halobacteria were polyphyletic (*pol-II*a, *polB*-a, *polB*-b, and *cdc21*b) indicating intein transfer between the Halobacteria and the taxa that interrupt them, as well as by data from other studies where intein transfer has been detected across phyla and domains (Butler et al., 2006; Swithers et al., 2013). Inteins also have a high substitution rate relative to their extein hosts, and a propensity for accumulating insertions and deletions, which makes detection of transfers between close relatives (generally a difficult task) possible; for example, transfer within the *Halorubrum* clusters shown in **Figure 3**. Inteins can be associated with a HEN domain. If they are, they possess the ability to invade intein-free alleles following transfer; if they are not, they rely mainly on vertical inheritance together with the host gene, and the occasional transfer of the host gene. One intein allele, *pol-II*a, is widely distributed in the Halobacteria and there are many examples of mini-intein sequences in this allele. These data suggest that invasion of this allele occurred early in the evolution of the Halobacteria, and that the intein may have been lost in some lineages, but retained as a mini intein in most of the genomes surveyed here. This could also be true for the *cdc21*-a intein; however, the distribution is not as diverse, and considerably fewer mini-inteins were detected. This is more suggestive of an intein that has been active in the Halobacteria for a long period of time, with the different intein states (empty target site, target site invaded by an intein with active HEN, target site occupied by an intein without functioning HEN; Yahara et al., 2009; Barzel et al., 2011) existing and co-existing in different halobacterial lineages.

The genomes analyzed in this work were cultured from salty water and soil samples around the world. The diverse background of the genomes may contribute to the spotty distribution of intein alleles (**Figure 1**). However, genomes isolated from the same location show variation as well **(Figure 3)** (Fullmer et al., 2014), reinforcing the notion that inteins are currently actively propagating in and being eliminated from halobacterial populations. Additionally, previous data have shown recombination occurs at a higher rate than mutation within the Halobacteria, and very little linkage between genes is detected in these genomes (Papke et al., 2004, 2007). These observations indicated gene flow as an important method for niche adaptation in these organisms. In Deep Lake, Antarctica the freezing temperatures limit the rate of replication to approximately 6 times per year and evolution in the halobacterial populations there mainly occurs through gene flow (Demaere et al., 2013). Recent whole genome comparisons revealed frequent gene transfer followed by homologous replacement of the transferred gene within the Halobacteria, hampering attempts to resolve the phylogeny within this group (Williams et al., 2012). Gene flow and recombination between populations and species make it difficult to resolve the species phylogeny among the different genera of Halobacteria (Papke et al., 2004). The use of gene concatenation in building reference trees, as exemplified by the ribosomal protein reference tree used in this work, has been pivotal in determining a branching order for the major clades of organisms, such as the Halobacteria, that participate in a large amount of recombination with close relatives. However, because genetic transfer and homologous recombination occur frequently between close relatives, the resulting phylogeny reflects both, shared ancestry and frequency of gene transfer. Therefore, determining the network of gene flow that overlays the vertical signal is important to the understanding of the evolution of these organisms. Inteins cannot penetrate the cell wall, and thus capitalize on existing gene flow in populations to efficiently invade when the opportunity presents itself. This trait can be exploited to keep track of successful homing events revealed by sequence similarity of inteins in distinct strains.

Halorubrum was the only genus in this study that had a large enough sample size to begin to uncover a signal reflecting population structure. Many of the *Halorubrum* genomes in this study were isolated from the same location, and this collection of genomes showed a clear signal for a structured population. Sixteen genomes from Aran-Bidgol were separated into four well-supported clusters. Three of the four clusters have branching orders identical to those in the reference tree, and the support values for those clusters could be attributed to both transfer within the group and a background phylogenetic signal or ancestral inheritance of similar intein alleles. However, only cluster 7 in the *Halorubrum* shares all intein alleles between all members of

Table 3 | Protein sequence identifiers for intein sequences.

Intein allele	Species name	Accession number	Phylum
cdc21-a	*Archaeoglobus profundus* DSM 5631	YP_004340760.1	Euryarchaeota
	Archaeoglobus veneficus SNP6	YP_003400528.1	Euryarchaeota
	Candidatus Methanomassiliicoccus intestinalis Issoire Mx1	YP_008072558.1	Euryarchaeota
	Croccosphaera watsonii	WP_021836378.1	Cyanobacteria
	Ferroglobus placidus DSM 10642	YP_003435419.1	Euryarchaeota
	Halarchaeum acidiphilum	WP_020220725.1	Halobacteria
	Lamprocystis purpurea	WP_020504136.1	Gammaproteobacteria
	Methanomassiliicoccus luminyensis	WP_019178416.1	Euryarchaeota
	Methanothermococcus okinawensis IH1	YP_004576471.1	Euryarchaeota
	Nocardia asteroides NBRC 15531	GAD83132.1	Actinobacteria
	Nocardiopsis potens	WP_020380316.1	Actinobacteria
	Pyrococcus abyssi GE5	NP_127115.1	Euryarchaeota
	Pyrococcus furiosus DSM 3638	NP_578211.1	Euryarchaeota
	Pyrococcus horikoshii OT3	NP_142122.1	Euryarchaeota
	Pyrococcus sp. NA2	YP_004424138.1	Euryarchaeota
	Thermococcus litoralis DSM 5473	YP_008429717.1	Euryarchaeota
	Thermococcus onnurineus NA1	YP_002306424.1	Euryarchaeota
	Thermococcus sibiricus MM 739	YP_002994932.1	Euryarchaeota
	Thermococcus sp. AM4	YP_002582218.1	Euryarchaeota
	Thermococcus sp. CL1	YP_006424652.1	Euryarchaeota
	Thermococcus zilligii	WP_010479121.1	Euryarchaeota
	Halorubrum sp. SP3	KJ_865687.1	Halobacteria
	Halorubrum sp. SP9	KJ_865689.1	Halobacteria
cdc21-b	*Cyanothece* sp. PCC 7822	YP_003887897.1	Cyanobacteria
	Halarchaeum acidiphilum	WP_020220725.1	Halobacteria
	Candidatus Methanomassiliicoccus intestinalis Issoire-Mx1	YP_008072558.1	Euryarchaeota
	Methanomassiliicoccus luminyensis	WP_019178416.1	Euryarchaeota
	Thermococcus barophilus Halorubrum sp. SP3	YP_004070279.1 KJ_865687.1	Euryarchaeota Halobacteria
	Halorubrum sp. SP7	KJ_865688.1	Halobacteria
	Halorubrum sp. SP9	KJ_865689.1	Halobacteria
polB-d	*Archaeoglobus profundus* DSM 5631	YP_003400528.1	Euryarchaeota
polB-a	*Salinibacter ruber* M8	YP_003572085.1	Bacteroidetes
	Salinibacter ruber DSM 13885	YP_446104.1	Bacteroidetes
	Halarchaeum acidiphilum	WP_020678478.1	Halobacteria
	Methanoculleus bourgensis	YP_006544623.1	Euryarchaeota
polB-b	*Halosimplex carlsbadense*	WP_006885382.1	Halobacteria
	Salinibacter ruber M8	YP_003572085.1	Bacteroidetes
	Salinibacter ruber DSM 13885	YP_446104.1	Bacteroidetes
	Halanaerobium saccharolyticum	WP_005489097.1	Firmicutes
	Halarchaeum acidiphilum	WP_020678478.1	Halobacteria
polB-c	*Thermus scotoductus*	YP_004202875.1	Deinococcus-Thermus
	Methanotorris igneus Kol 5	YP_004483799.1	Euryarchaeota
	Halorubrum sp. SP7	KJ_865686.1	Halobacteria
pol-IIa	*Archaeoglobus veneficus* SNP6	YP_004341738.1	Euryarchaeota
	Halosimplex carlsbadense	WP_006882195.1	Halobacteria
	Methanocaldococcus infernus ME	YP_003616947.1	Euryarchaeota
	Methanococcus aeolicus	ABU41683.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Methanoculleus bourgensis MS2	YP_006544019.1	Euryarchaeota
	Methanoculleus marisnigri JR-1	YP_001048029.1	Euryarchaeota
	Methanofollis liminatans	WP_004037227.1	Euryarchaeota
	Methanolinea tarda	WP_007314808.1	Euryarchaeota
	Methanoplanus limicola	WP_004076782.1	Euryarchaeota
	Methanoplanus petrolearius DSM 11571	YP_003893638.1	Euryarchaeota
	Methanoregula boonei 6A8	YP_001403293.1	Euryarchaeota
	Methanoregula formicica SMSP	YP_007242862.1	Euryarchaeota
	Methanosphaerula palustris E1-9c	YP_002467270.1	Euryarchaeota
	Metahnospirillum hungatei JF-1	YP_503855.1	Euryarchaeota
	Pyrococcus horikoshii OT3	NP_142130.1	Euryarchaeota
	Thermococcus gammatolerans EJ3	YP_002958492.1	Euryarchaeota
	Thermococcus sibiricus MM 739	YP_002994988.1	Euryarchaeota
	uncultured haloarchaeon	ABQ75865.1	Halobacteria
	Halorubrum sp. SP3	KJ_865692.1	Halobacteria
	Halorubrum sp. SP7	KJ_865690.1	Halobacteria
	Halorubrum sp. SP9	KJ_564691.1	Halobacteria
pol-IIb	*Halosimplex carlsbadense*	WP_006882195.1	Halobacteria
	Pyrococcus abyssi GE5	YP_004624494.1	Euryarchaeota
	uncultured haloarchaeon	ABQ75865.1	Halobacteria
gyrB	*Allochromatium vinosum* DSM 180	YP_003443943.1	Gammaproteobacteria
	Anabaena sp. 90	YP_006997726	Cyanobacteria
	Anabaena sp. PCC 7108	WP_016950132.1	Cyanobacteria
	Bacillus subtilis BEST7613	BAM51471.1	Firmicutes
	Calothrix sp. PCC 7103	WP_019489451.1	Cyanobacteria
	Coleofasciculus chthonoplastes	WP_006099284.1	Cyanobacteria
	Cylindrospermopsis reciborskii	WP_006276716.1	Cyanobacteria
	Dactylococcopsis slaina PCC 8305	YP_007173052.1	Cyanobacteria
	Halarchaeum acidiphilum	WP_021780646.1	Halobacteria
	Methanomassiliicoccus luminyensis	WP_019178436.1	Euryarchaeota
	Microcystis aeruginosa	WP_002774451.1	Cyanobacteria
	Moorea producens	WP_008190351.1	Cyanobacteria
	Oscillatoria sp. PCC 10802	WP_017715151.1	Cyanobacteria
	Pleurocapsa sp. PCC 7319	WP_019509077.1	Cyanobacteria
	Prochlorothrix hollandica	WP_017710941.1	Cyanobacteria
	Raphidiopsis brookii	WP_009342634.1	Cyanobacteria
	Rivularia sp. PCC 7116	YP_007054134.1	Cyanobacteria
	Saccharothrix espanaensis DSM 44229	YP_007037469.1	Actinobacteria
	Synechocystis sp. PCC 6803	NP_441040.1	Cyanobacteria
	Trichodesium erythraeum IMS101	YP_723459.1	Cyanobacteria
	uncultured bacterium	EKD46222.1	
helicase-b	*Bacillus amyloliqufaciens* TA208	YP_005540906.1	Firmicutes
	Bacillus subtilis	WP_017696872.1	Firmicutes
	Nanoarchaeota archaeon SCGC AAA011-L22	WP_018204386.1	
rfc-a	*Methanocaldococcus jannaschii* DSM 2661	NP_248426.1	Euryarchaeota
	Methanocaldococcus sp. FS406	YP_003458055.1	Euryarchaeota
	Methanothermococcus okinawensis IH1	YP_004576337.1	Euryarchaeota
	Methanotorris formicicus	WP_007044297.1	Euryarchaeota
	Pyrococcus abyssi GE5	NP_125803.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Pyrococcus furiosus DSM 3638	NP_577822.1	Euryarchaeota
	Pyrococcus horikoshii OT3	NP_142122.1	Euryarchaeota
	Pyrococcus sp. ST04	YP_006353924.1	Euryarchaeota
	Thermococcus kodakorensis KOD1	YP_184631.1	Euryarchaeota
	Thermococcus litoralis DSM 5473	YP_008428897.1	Euryarchaeota
	Thermococcus sp. 4557	YP_004763272.1	Euryarchaeota
	Thermococcus sp. AM4	YP_002582171.1	Euryarchaeota
	Thermococcus sp. CL1	YP_006425306.1	Euryarchaeota
*rpo*lA	*Halorubrum* sp. SP3	KJ_865684.1	Halobacteria
	Halorubrum sp. SP9	KJ_865685.1	Halobacteria
rir1-l	*Chloroherpeton thalassium* ATCC 35110	YP_001995975.1	Chlorobi
	Tepidanaerobacter acetatoxydans Re1	YP_007273179.1	Firmicutes
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
rir1-k	*Deinococcus peraridilitoris* DSM 19664	YP_007181218.1	Deinococcus-Thermus
rir1-b	*Acidovorax avenae* subsp. avenae ATCC 19860	YP_004233126.1	Betaproteobacteria
	Acidovorax sp. CF316	WP_007856012.1	Betaproteobacteria
	Acidovorax sp. NO-1	WP_008903130.1	Betaproteobacteria
	Actinomadura atramentaria	WP_019631066.1	Actinobacteria
	Alicyclobacillus pohliae	WP_018131875.1	Firmicutes
	Aminomonas paucivorans	WP_006300529.1	Synergistetes
	Ammonifex degensii KC4	WP_006300529.1	Firmicutes
	Arhodomonas aquaeolei	WP_018718131.1	Gammaproteobacteria
	Bacillus licheniformis	WP_016885361.1	Firmicutes
	Bacillus subtilis	WP_017697104.1	Firmicutes
	Calothrix sp. PCC 6303	YP_007136749.1	Cyanobacteria
	Candidatus Chloracidobacterium thermophilum B	YP_004863563.1	Acidobacteria
	Candidatus Desulforudis audaxviator MP104C	YP_001717412.1	Firmicutes
	Clostridiaceae bacterium L21-TH-D2	WP_006314960.1	Firmicutes
	Deinococcus radiodurans R1	NP_296095.1	Deinococcus-Thermus
	Delftia acidovorans	WP_016451949.1	Betaproteobacteria
	Delftia sp. Cs1-4	YP_004490724.1	Betaproteobacteria
	Desulfitobacterium hafniense	WP_005810476.1	Firmicutes
	Desulfovibrio magneticus RS-1	YP_002955841.1	Deltaproteobacteria
	Desulfovibrio sp. U5L	WP_009106508.1	Deltaproteobacteria
	Ferroplasma acidarmanus fer1	YP_008141532.1	Euryarchaeota
	Ferroplasma sp. Type II	WP_021787573.1	Euryarchaeota
	Halomonas anticariensis	WP_016418429.1	Gammaproteobacteria
	Halomonas jeotgali	WP_017429019.1	Gammaproteobacteria
	Halomonas smyrnensis	WP_016854101.1	Gammaproteobacteria
	Mahella australiensis 50-1 BON	YP_004462974.1	Firmicutes
	Marinobacter lipolyticus	WP_018405479.1	Gammaproteobacteria
	Methanofollis liminatans	WP_004040239.1	Euryarchaeota
	Methylobacter marinus	WP_020160338.1	Gammaproteobacteria
	Methylococcus capsulatus	WP_017366201.1	Gammaproteobacteria
	Methylomicrobium buryatense	WP_017841702.1	Gammaproteobacteria
	nanoarchaeote Nst1	WP_004578017.1	
	Nocardiopsis halotolerans	WP_017572347.1	Actinobacteria
	Polaromonas sp. JS666	CAJ57177.1	Cyanobacteria
	Pseudanabaena sp. PCC 6802	WP_019499030.1	Cyanobacteria

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Pseudanabaena sp. PCC 7367	YP_007101092.1	Cyanobacteria
	Rhodanobacter fulvus	WP_007082010.1	Gammaproteobacteria
	Rhodanobacter sp. 2APBS1	YP_007588821.1	Gammaproteobacteria
	Rhodanobacter thiooxydans	WP_008437232.1	Gammaproteobacteria
	Rhodothermus marinus SG0.5JP17-172	YP_004824118.1	Bacteroidetes
	Staphylococcus aureus	WP_016187732.1	Firmicutes
	Synechococcus elongatus PCC 6301	CAJ57178.1	Cyanobacteria
	Synechococcus elongatus PCC 7942	YP_400626.1	Cyanobacteria
	Synechococcus sp. PCC 6312	YP_007060778.1	Cyanobacteria
	Thermoanaerobacterium saccharolyticum JW/SL-YS485	YP_006391581.1	Firmicutes
	Thermoanaerobacterium thermosaccharolyticum DSM 571	YP_003851043.1	Firmicutes
	Thermobrachium celere	WP_018663796.1	Firmicutes
	Thermococcus kodakarensis KOD1	YP_184312.1	Euryarchaeota
	Thermodesulfatator indicus DSM 15286	YP_004625205.1	Thermodesulfobacteria
	Thermovirga lienii DSM 17291	YP_004932130.1	Deinococcus-Thermus
	Thermus igniterrae	WP_018110436.1	Deinococcus-Thermus
	Thermus thermophilus HB8	CAJ57170.1	Deinococcus-Thermus
	Thioalkalivibrio sp. ALE11	WP_019570879.1	Gammaproteobacteria
	Thioalkalivibrio sp. ALE30	WP_018881426.1	Gammaproteobacteria
	Thioalkalivibrio sp. HL-Eb18	WP_017926201.1	Gammaproteobacteria
	Thioalkalivibrio sp. K90mix	YP_003459507.1	Gammaproteobacteria
	uncultured bacterium	EKE25755.1	
	Xanthomonas sp. SHU199	WP_017907463.1	Gammaproteobacteria
	Xanthomonas sp. SHU308	WP_017915139.1	Gammaproteobacteria
	zeta proteobacterium SCGC AB-604-B04	WP_018280466.1	Zetaproteobacteria
rir1-g	*Chloroherpeton thalassium* ATCC 35110	YP_001995975.1	Chlorobi
	Deinococcus aquatillis	WP_019011777.1	Deinococcus-Thermus
	Halothece sp. PCC 7418	YP_007166732.1	Cyanobacteria
	Klebsiella pneumoniae	WP_021313783.1	Gammaproteobacteria
	Nocardiopsis dassonvillei subsp. Dassonvillei DSM 43111	YP_003681238.1	Actinobacteria
	Nocardiopsis sp. CNS639	WP_019609645.1	Actinobacteria
	Rhodothermus marinus SG0.5JP17-172	YP_004826277.1	Bacteroidetes
	Tepidanaerobacter acetatoxydans Re1	YP_007273179.1	Firmicutes
	Thermomonospora curvata DSM 43183	YP_003299200.1	Actinobacteria
	Thermus thermophilus HB27	YP_005899.1	Deinococcus-Thermus
	Thermus thermophilus HB8	CAJ57173.1	Deinococcus-Thermus
	Thermus thermophilus JL-18	YP_006059430.1	Deinococcus-Thermus
	Thermus thermophilus SG0.5JP17-16	YP_005639869.1	Deinococcus-Thermus
	Trichodesium erythraeum IMS101	YP_720358.1	Cyanobacteria
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
rir1-m	*Thermus aquaticus*	WP_003044118.1	Deinococcus-Thermus
	Thermus thermophilus HB-8	CAJ57173.1	Deinococcus-Thermus
	Thermus thermophilus SG0.5JP17-16	YP_005639869.1	Deinococcus-Thermus
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
udp	*Fervidibacteria bacterium* JGI 0000001-G10	WP_020250137.1	
	Dictyglomus thermophilum H-6-12	YP_002250310.1	Dictyglomi
	Methanocaldococcus jannaschii DSM 2661	NP_248048.1	Euryarchaeota
	Methanocaldococcus vulcanis M7	YP_003246412.1	Euryarchaeota
	Methanococcus aeolicus Nankai-3	YP_001324612.1	Euryarchaeota
	Methanothermococcus okinawensis IH1	YP_004575831.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Methanotorris igneus Kol 5	WP_007044255.1	Euryarchaeota
	Thermococus gammatolerans EJ3	YP_002960518 .1	Euryarchaeota
topA	*Methanotorris igneus* Kol 5	WP_007044255.1	Euryarchaeota
top6B	*Halarchaeum acidiphilum*	WP_021780130.1	Halobacterium

**Indicates the intein detected is a mini-intein.*
--Indicates taxa that grouped within the halobacterial intein sequences.

the cluster while the other clusters all contain intein alleles that are unique to certain members of the cluster, suggesting ongoing transfer of these inteins within the population. Additionally, three out of the twelve total clusters demonstrate unique branching orders compared to the reference tree, though only five of the clusters reflected in the reference tree have identical intein profiles. The lack of fixation for the intein alleles in the majority of clusters (seven out of twelve) indicates that a signal due to vertical inheritance may aid the formation of the clusters, but that HGT and its bias is the driving force for intein distribution. This analysis demonstrates the utility of intein sequences in distinguishing a population structure amongst genomes isolated from the same location, as demonstrated with the genomes isolated from Aran-Bidgol. These relationships are made evident through analyzing all of the signals from each of the intein alleles represented in the strains, and thus represent a collapsed view of the major gene sharing networks that have shaped the intein profiles of these strains over time. The collapsed networks indicate a higher rate of recombination within compared to between species and groups, a finding similar to the sexual outcrossing in fungal populations where inteins also thrive, as the semi-sexual lifestyle promotes intein homing (Giraldo-Perez and Goddard, 2013).

It is tempting to speculate that strains that harbor an abundance of intein alleles partake in more gene transfer than their counterparts without as many inteins; however, these two phenomena should not be expected to have a strict correlation as HGT between strains that possess only one intein each cannot produce hybrids with more than two inteins each. The number of inteins present in a group of different strains and species may be more reflective of transfers with divergent organisms than within-group transfer frequency.

The presented research demonstrates the utility of intein sequences to follow gene flow within and between populations. Improved reliability to assess the presence and activity of the HEN domain intein will provide a better distinction between vertical and horizontal inheritance of inteins. The overall utility of inteins improves as new intein alleles and new host proteins are reported, increasing the distribution of samples and improving statistical robustness of studies like the one done here. Prior to this work, nine proteins had been reported to contain inteins in the Halobacteria. This work established seven new intein alleles in the Halobacteria, including two proteins not previously reported to contain inteins. The presence of inteins is especially useful in populations where high rates of recombination and widely distributed populations may facilitate the maintenance of intein sequences over long periods of time (Gogarten and Hilario, 2006) and provide a means for distinguishing closely related partners involved in genetic transfers. The phylogenetic distribution of intein alleles, combined with the changing state within intein alleles, and the rapid substitution rate of inteins relative to the extein host sequences (Swithers et al., 2013) will provide a valuable tool to infer gene flow dynamics in and between sampled populations.

AUTHOR CONTRIBUTIONS

Johann Peter Gogarten and Shannon M. Soucy participated in the design of this study and helped to draft the manuscript. Shannon M. Soucy performed the research and all authors contributed to data analysis. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The UConn Bioinformatics Facility provided computing resources for the analyses reported in this manuscript. The *Halorubrum* genomes provided by the Papke lab were sequenced in house by Andrea Makkay and Ryan Wheeler. We would like to thank them for their hard work, as well as acknowledge Dr. Elina Roine and Dennis Bamford (Helsinki University), and Dr. Antonio Ventosa (University of Sevilla) for supplying the sequenced strains. We would also like to recognize labs sequencing genomes and making them available in data repositories such as those hosted by the National Center for Biotechnology Information. This work was supported by the National Science Foundation Grant (DEB 0830024 and DEB0919290) and NASA Astrobiology: Exobiology and Evolutionary Biology Grants (NNX12AD70G and NNX13AI03G).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00299/abstract

REFERENCES

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andam, C. P., Williams, D., and Gogarten, J. P. (2010). Biased gene transfer mimics patterns created through shared ancestry. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10679–10684. doi: 10.1073/pnas.1001418107

Atanasova, N. S., Roine, E., Oren, A., Bamford, D. H., and Oksanen, H. M. (2012). Global network of specific virus-host interactions in hypersaline environments. *Environ. Microbiol.* 14, 426–440. doi: 10.1111/j.1462-2920.2011.02603.x

Barzel, A., Obolski, U., Gogarten, J. P., Kupiec, M., and Hadany, L. (2011). Home and away- the evolutionary dynamics of homing endonucleases. *BMC Evol. Biol.* 11:324. doi: 10.1186/1471-2148-11-324

Butler, M. I., Gray, J., Goodwin, T. J., and Poulter, R. T. (2006). The distribution and evolutionary history of the PRP8 intein. *BMC Evol. Biol.* 6:42. doi: 10.1186/1471-2148-6-42

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Demaere, M. Z., Williams, T. J., Allen, M. A., Brown, M. V., Gibson, J. A. E., Rich, J., et al. (2013). High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16939–16944. doi: 10.1073/pnas.1307090110

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus Halorubrum. *Front. Microbiol.* 5:140. doi: 10.3389/fmicb.2014.00140

Gimble, F. S., and Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. *Nature* 357, 301–306. doi: 10.1038/357301a0

Giraldo-Perez, P., and Goddard, M. R. (2013). A parasitic selfish gene that affects host promiscuity. *Proc. Biol. Sci.* 280:20131875. doi: 10.1098/rspb.2013.1875

Goddard, M. R., and Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci.* 96, 13880–13885. doi: 10.1073/pnas.96.24.13880

Gogarten, J. P., and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol. Biol.* 6:94. doi: 10.1186/1471-2148-6-94

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002). Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* 56, 263–287. doi: 10.1146/annurev.micro.56.012302.160741

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hirata, R., Ohsumk, Y., Nakano, A., Kawasaki, H., Suzuki, K., and Anraku, Y. (1990). Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. *J. Biol. Chem.* 265, 6726–6733.

Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebl, M., and Stevens, T. H. (1990). Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science* 250, 651–657. doi: 10.1126/science.2146742

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., and Berg, I. A. (2011). A methylaspartate cycle in haloarchaea. *Science* 331, 334–337. doi: 10.1126/science.1196544

Lang, A. S., Zhaxybayeva, O., and Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* 10, 472–482. doi: 10.1038/nrmicro2802

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., et al. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.* 61, 90–106. doi: 10.1093/sysbio/syr095

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., et al. (2000). Genome sequence of Halobacterium species NRC-1. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12176–12181. doi: 10.1073/pnas.190337797

Papke, R. T., and Gogarten, J. P. (2012). Ecology. How bacterial lineages emerge. *Science* 336, 45–46. doi: 10.1126/science.1219241

Papke, R. T., Koenig, J. E., Rodríguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of halorubrum. *Science* 306, 1928–1929. doi: 10.1126/science.1103289

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14092–14097. doi: 10.1073/pnas.0706358104

Perler, F. B. (2002). InBase: the Intein Database. *Nucleic Acids Res.* 30, 383–384. doi: 10.1093/nar/30.1.383

Perler, F. B., Olsen, G. J., and Adam, E. (1997). Compilation and analysis of intein sequences. *Nucleic Acids Res.* 25, 1087–1093. doi: 10.1093/nar/25.6.1087

Pietrokovski, S. (2001). Intein spread and extinction in evolution. *Trends Genet.* 17, 465–472. doi: 10.1016/S0168-9525(01)02365-4

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Swithers, K. S., Senejani, A. G., Fournier, G. P., and Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol. Biol.* 9:303. doi: 10.1186/1471-2148-9-303

Swithers, K. S., Soucy, S. M., Lasek-Nesselquist, E., Lapierre, P., and Gogarten, J. P. (2013). Distribution and evolution of the mobile vma-1b intein. *Mol. Biol. Evol.* 30, 2676–2687. doi: 10.1093/molbev/mst164

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4, 1223–1244. doi: 10.1093/gbe/evs098

Yahara, K., Fukuyo, M., Sasaki, A., and Kobayashi, I. (2009). Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18861–18866. doi: 10.1073/pnas.0908404106

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res.* 16, 1099–1108. doi: 10.1101/gr.5322306

Conflict of Interest Statement: The authors declare that the research was con-ducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 January 2014; accepted: 30 May 2014; published online: 26 June 2014. Citation: Soucy SM, Fullmer MS, Papke RT and Gogarten JP (2014) Inteins as indi- cators of gene flow in the halobacteria. Front. Microbiol. 5:299. doi: 10.3389/fmicb. 2014.00299
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology.

Chapter 5. Further Investigation of Gene Transfer Networks using Inteins.

5.1 Introduction

Several lines of evidence from section 4.2 indicate that in order to maximize the information content of inteins, individual intein sequences should be considered. First, the phylogeny of all haloarchaeal intein sequences together has very poor support linking intein alleles together; indicating the life history of each intein allele is unique. Secondly, the concatenated intein sequence phylogeny also has very poor support at most of the bipartitions. This shows that signals of HGT within each intein allele are so different that combining these signals results in noise, that is difficult to interpret beyond the conclusions drawn (inteins are exchanged between close relatives within the same environment). Lastly the phylogenies of intein alleles with many taxa (*pol-II*-a, and *cdc21*-a) show several clusters of *Halorubrum* sp. that group independently from one another. This could represent several different invasion events, or it could symbolize differential evolution of the intein sequence after a single invasion event. To explore these conflicts further, I started working with a student visitor from University of Sao Paulo, Thiberio Rangel. The following work was conducting in collaboration with Thiberio.

We use a a method for analyzing HGT of inteins that is not based on conflict between phylogenetic trees, but on outliers in correlations between pairwise distances estimated using maximum likelihood models. By analyzing the correlation between sets of pairwise distances we can observe sets of sequences that generate a signal conflicting with what we would expect given a reference set of distances. These conflicts represent differential evolutionary trajectories between the reference set and the test dataset, in our case intein alleles. Furthermore, intein alleles that are not in agreement with the reference dataset can be investigated further using

scatterplots and reconciliation of phylogenetic conflict to compare the expected and observed pairwise distances.

5.2 Materials and Methods

5.2.1 Sequence Alignments

Alignments for intein alleles, their corresponding exteins, and ribosomal proteins were generated using the default settings in Muscle v3.8.31 (12), and examined and refined using Seaview v4.4.2(13). Refining the datasets involved removing taxa with partial sequences, or extremely divergent sequences (*Haloquadratum sp.*), these alignments were then used to calculate pairwise distances.

5.2.2 Pairwise Distance Calculation, Correlation, Heat-maps, and Scatterplots

Pairwise Maximum-Likelihood distances matrices were generated for each intein, extein, and ribosomal gene family protein alignment using RAxML v8.1.17 (14), with the –f x –m GTRGAMMA parameters. Correlations between gene families' distance matrices were evaluated through Pearson Correlation Coefficient, since we are only interested in linear relations. The interdependent nature of distance matrix correlations causes it to be more susceptible false-positive values (*i.e.* given the distance between A to B, changing the distance between A to C is also going to affect the distance between B to C), with this in mind we assessed correlation significances via Mantel Significance Test (1000 iterations per pairwise correlation), and p-values were corrected for multiple tests using Benjamini-Hochberg's FDR. Hierarchical clustering in heat maps was generated using single linkage, where the confidence intervals of the linear regressions displayed in scatterplots reflects 95%.

5.2.3 Phylogenetic Trees

Maximum likelihood trees were generated for each intein, extein, and ribosomal protein using RAxML v8.1.17 (14) with the GTR model, optimized substitution rates, and a gamma model of rate heterogeneity.

5.3 Results

5.3.1 Refining the Reference Dataset

We begin by analyzing the agreement between the set of genes that is used to build the reference tree. In section 4.2 we used 56 ribosomal proteins to build the reference phylogeny. We analyzed the pairwise distances within each protein and then looked for correlations in the total set of pairwise distances between all ribosomal proteins. From this analysis we choose a smaller set of 28 ribosomal proteins with a strong correlation of pairwise distance patterns (Figure 1). We then concatenated these proteins to build a new reference tree using the smaller reference dataset (Figure 2). The concatenated dataset was then used as the "reference" for all subsequent comparisons; strong correlation with this dataset indicates vertical evolution of the intein. Whereas conflict with this dataset indicates HGT events have occurred sometime in the life history of the protein.

Figure 1. Correlation between Ribosomal Proteins. Pairwise distances were generated for all members within each protein, and compared between all proteins. The highlighted set is strongly correlated and these proteins were used to represent the reference signal in all subsequent figures.

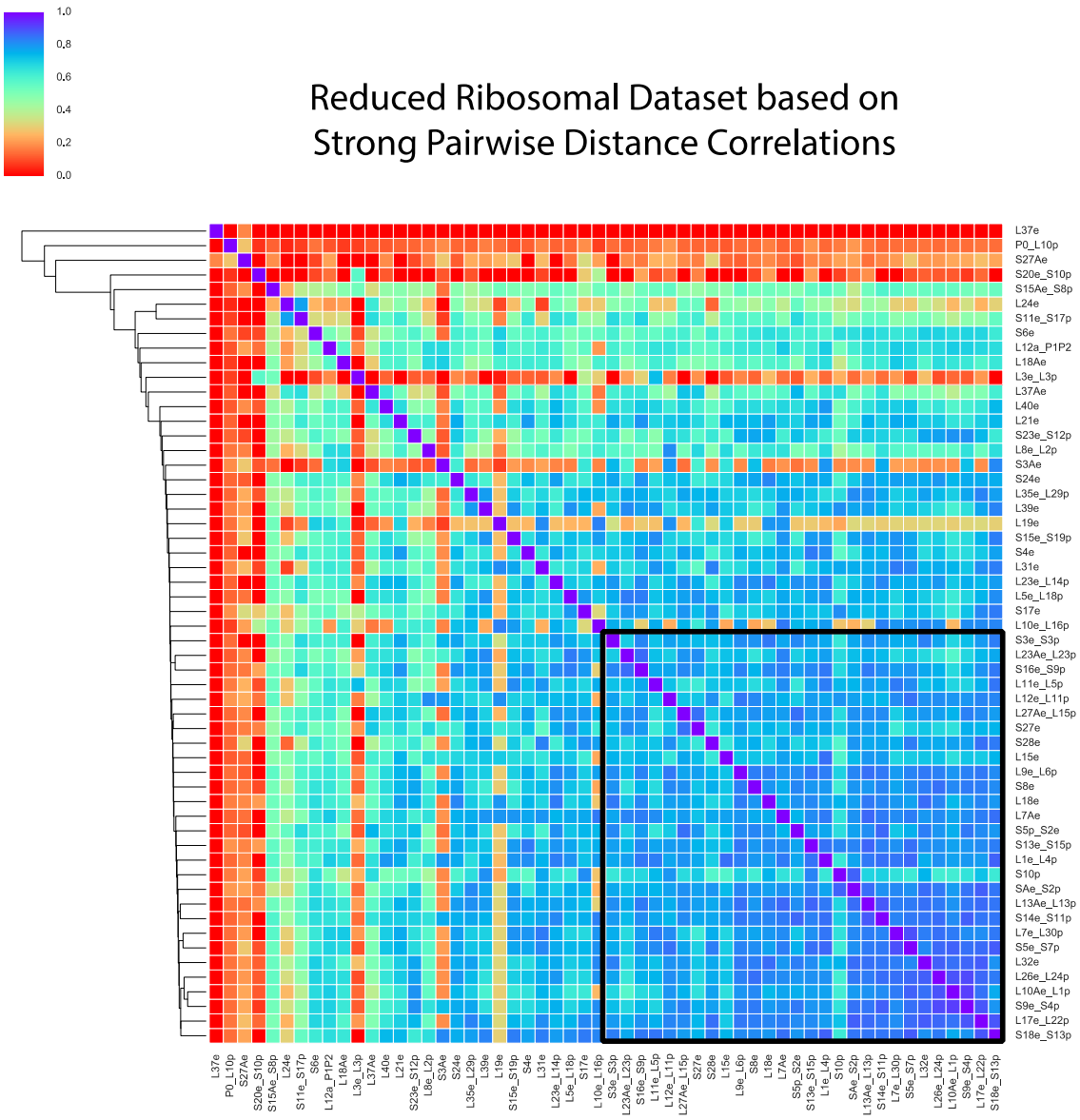
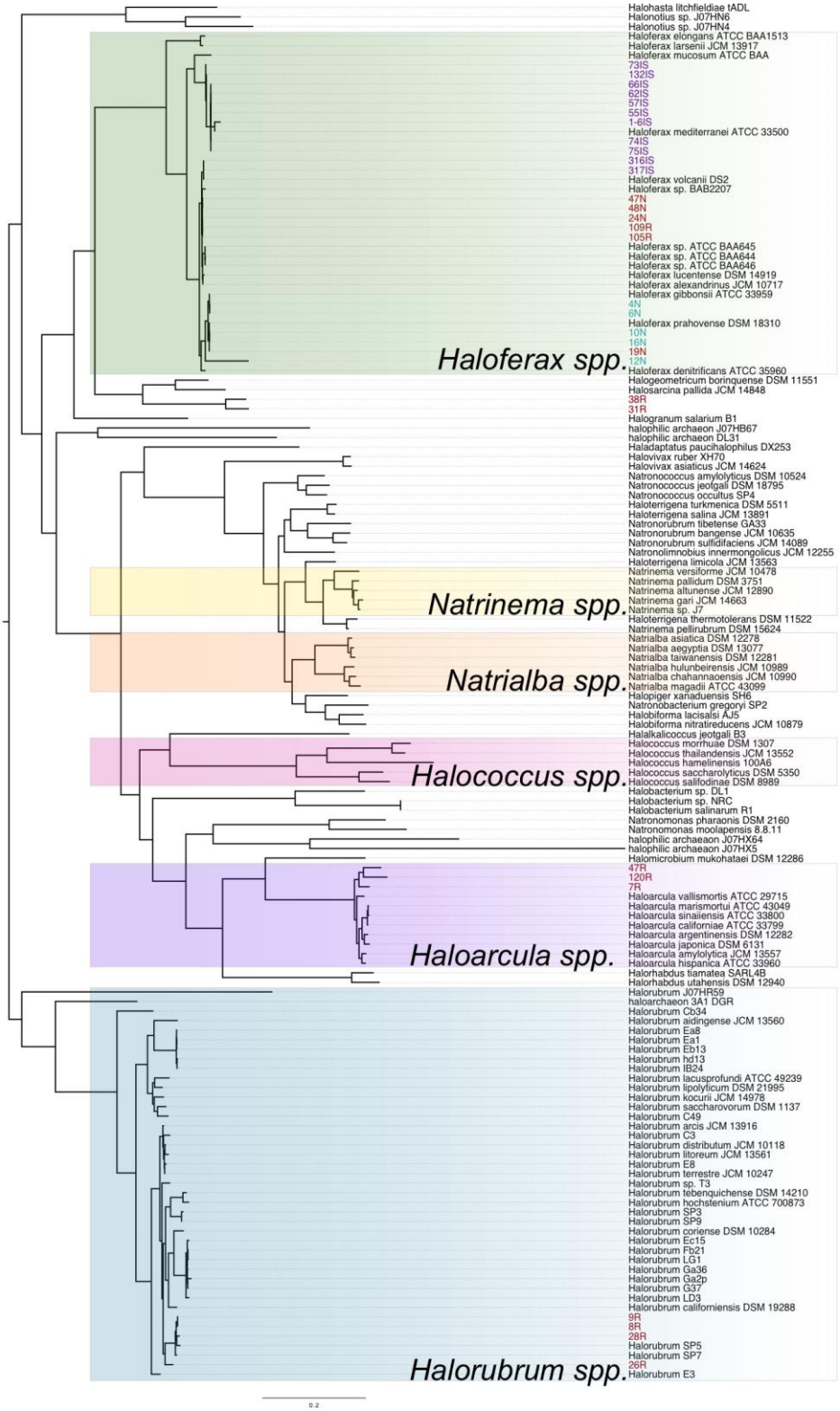
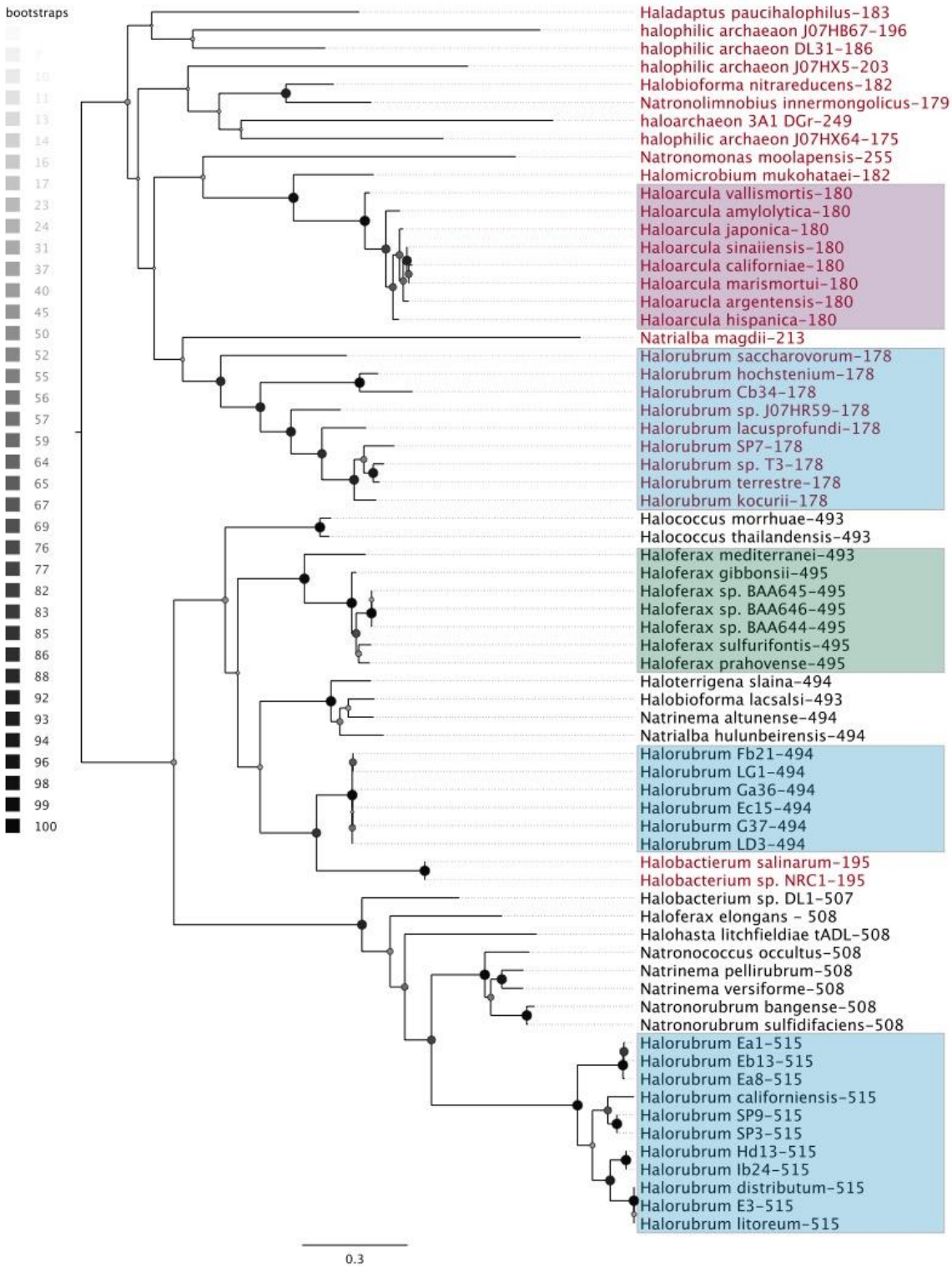


Figure 0-2 Reduced Ribosomal place will have a one with values. Tree-holder better support



5.3.2 Testing the impact of the HEN domain on the correlations

The larger intein phylogenies include mini-inteins (inteins that have lost their homing endonuclease domain) that cluster to the exclusion of full size inteins in the intein phylogenies (Figure 3). Additionally beyond the mini and full size inteins clustering separately the *Halorubrum* sp. cluster in three locations in the Pol II-a phylogeny, and in two locations in the Cdc21a phylogeny (Figure 3). To consider the possibility that the loss of the HEN domain is the cause of this clustering, alignments were generated where the HEN domain was removed from all sequences. This dataset (intein-splicing-only) was then compared to the full-length alignments to look for conflicts that could be caused by differential evolution in the homing endonuclease domain. In all cases the splicing only distances correlated strongly with the full-length alignments (Figure 4), indicating that the different clusters most likely arose through several independent HGT events.


bootstraps
7
10
11
13
14
16
17
23
24
31
37
40
45
50
52
55
56
57
59
64
65
67
69
76
77
82
83
85
86
88
92
93
94
96
98
99
100
Haladaptus paucihalophilus-183
halophilic archaeaon J07HB67-196
halophilic archaeon DL31-186
halophilic archaeon J07HX5-203
Halobioforma nitrareducens-182
Natronolimnobius innermongolicus-179
haloarchaeon 3A1 DGr-249
halophilic archaeon J07HX64-175
Natronomonas moolapensis-255
Halomicrobium mukohataei-182
Haloarcula vallismortis-180
Haloarcula amylolytica-180
Haloarcula japonica-180
Haloarcula sinaiiensis-180
Haloarcula californiae-180
Haloarcula marismortui-180
Haloarucla argentensis-180
Haloarcula hispanica-180
Natrialba magdii-213
Halorubrum saccharovorum-178
Halorubrum hochstenium-178
Halorubrum Cb34-178
Halorubrum sp. J07HR59-178
Halorubrum lacusprofundi-178
Halorubrum SP7-178
Halorubrum sp. T3-178
Halorubrum terrestre-178
Halorubrum kocurii-178
Halococcus morrhuae-493
Halococcus thailandensis-493
Haloferax mediterranei-493
Haloferax gibbonsii-495
Haloferax sp. BAA645-495
Haloferax sp. BAA646-495
Haloferax sp. BAA644-495
Haloferax sulfurifontis-495
Haloferax prahovense-495
Haloterrigena slaina-494
Halobioforma lacsalsi-493
Natrinema altunense-494
Natrialba hulunbeirensis-494
Halorubrum Fb21-494
Halorubrum LG1-494
Halorubrum Ga36-494
Halorubrum Ec15-494
Haloruburm G37-494
Halorubrum LD3-494
Halobactierum salinarum-195
Halobacterium sp. NRC1-195
Halobacterium sp. DL1-507
Haloferax elongans - 508
Halohasta litchfieldiae tADL-508
Natronococcus occultus-508
Natrinema pellirubrum-508
Natrinema versiforme-508
Natronorubrum bangense-508
Natronorubrum sulfidifaciens-508
Halorubrum Ea1-515
Halorubrum Eb13-515
Halorubrum Ea8-515
Halorubrum californiensis-515
Halorubrum SP9-515
Halorubrum SP3-515
Halorubrum Hd13-515
Halorubrum Ib24-515
Halorubrum distributum-515
Halorubrum E3-515
Halorubrum litoreum-515
0.3

A.

B.

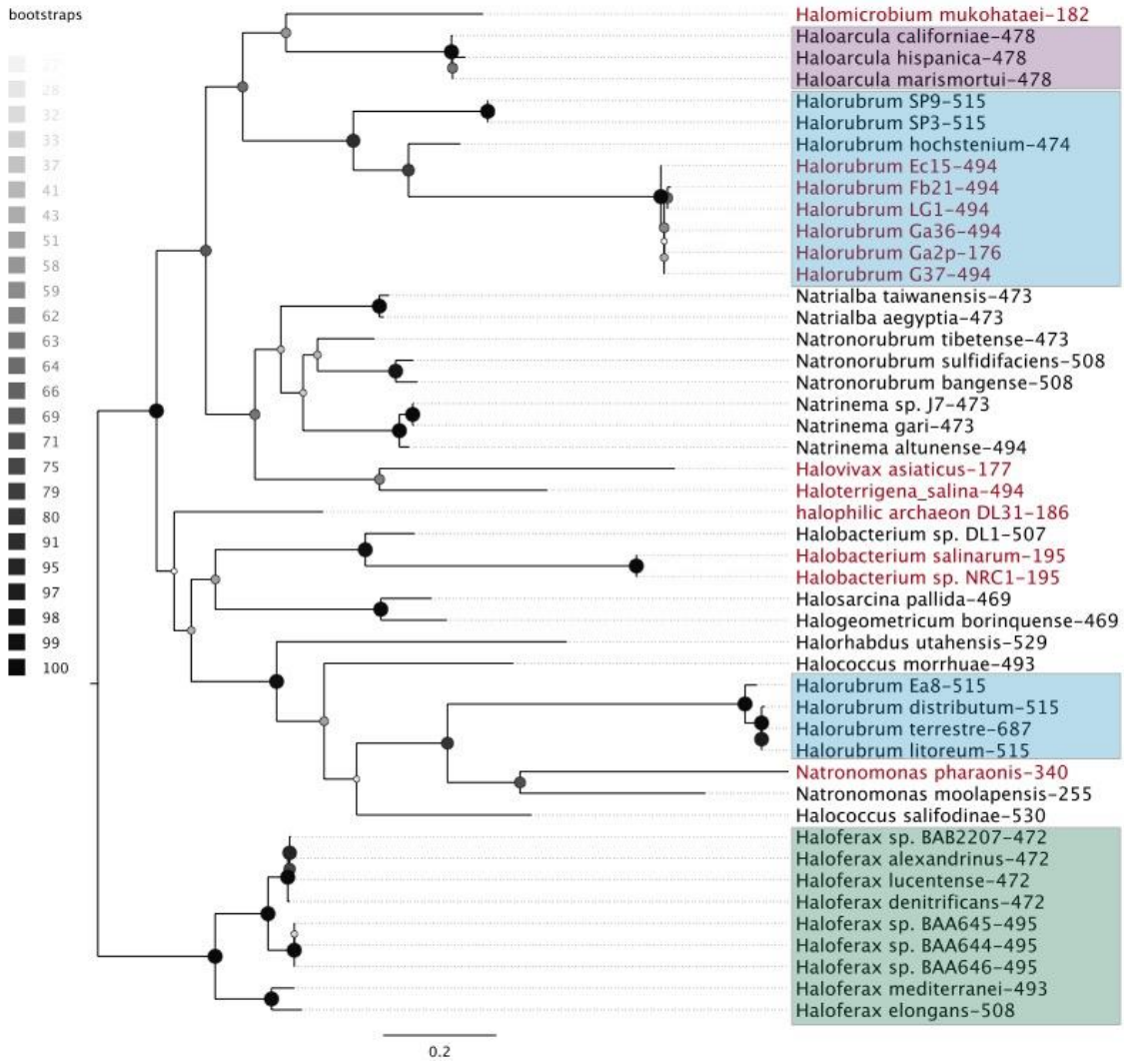


Figure 3. Intein phylogenies for *pol-II*-a and *cdc21*-a. A maximum likelihood phylogeny of *pol-II*a (A) and *cdc21*-a (B) was created from an alignment of all full sequences from (15)**. Bootstrap support values are indicated by the size and color of the circles at each node, larger and darker circles indicate high support. Three major genera are colored, *Halorubrum* sp. in blue, *Haloarcula* sp. in red, and *Haloferax* sp. in Green.**

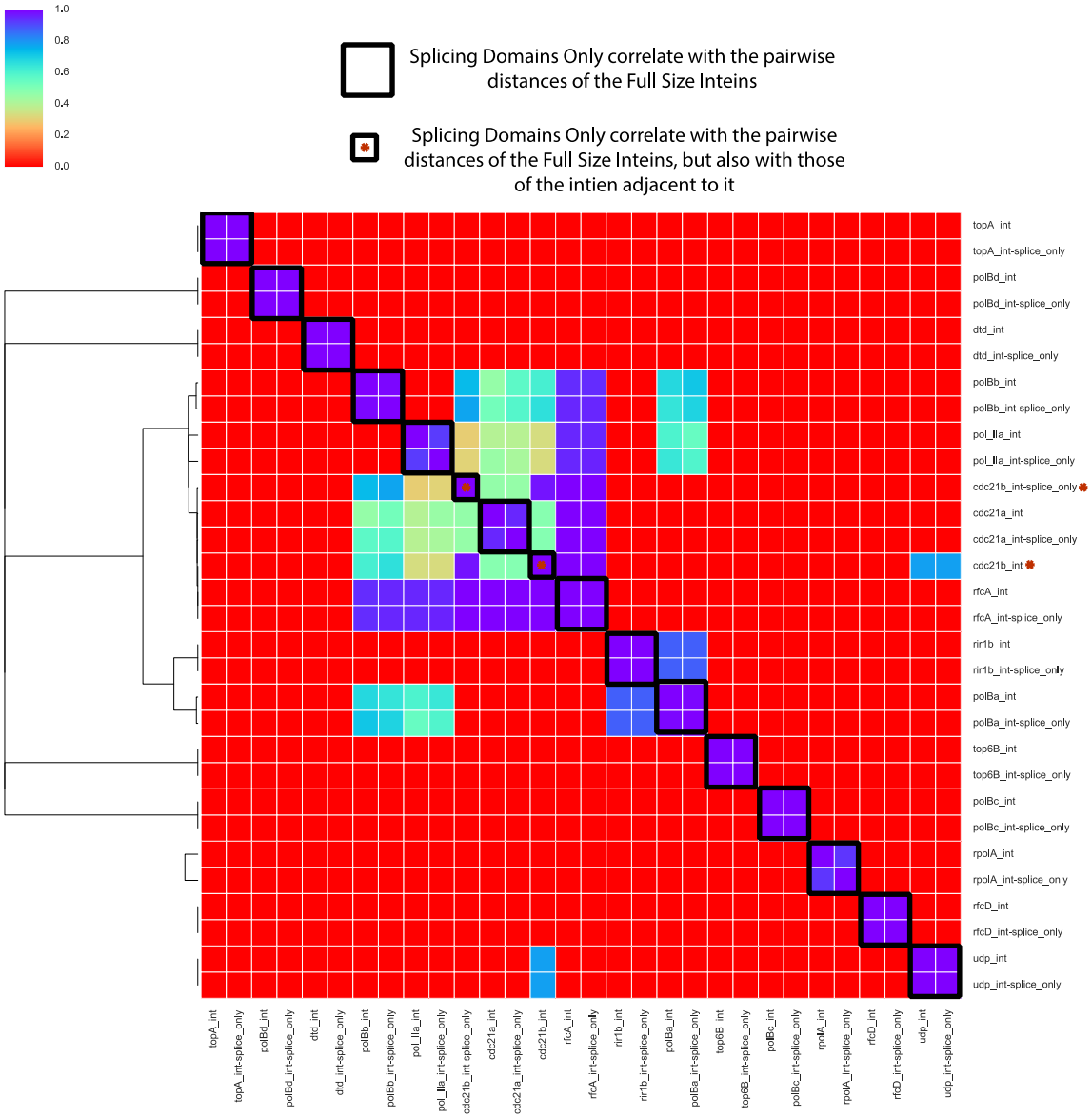


Figure 1. Homing Endonuclease domain is not affect phylogenetic clustering. Highly correlated datasets are outlined, each intein allele correlated strongly with the corresponding splicing only alignment.

5.3.3 Are Inteins transferred with the Extein through Homologous Replacement?

Next we were interested to know if inteins are acquired through homologous replacement with the extein or independent of the extein through homing. There is some evidence that recombination resulting in intein invasion can effect the region of the genome adjacent to the intein insertion site, including the extein. Because inteins reside in proteins that are involved in essential cellular tasks, replacement of these proteins, or even large parts of these proteins could affect the greater network of cellular machinery. To investigate this we looked for correlation between the extein and the corresponding intein allele (Figure 5). Interestingly in most cases there is little agreement between the extein and its corresponding intein, however there are a few exceptions. Four of the twenty-four intein alleles showed agreement between the extein and intein correlations. This could indicate homologous replacement with the extein, or the lineages represented were affected by a single intein invasion event in the history of a common ancestor. In at least two cases the latter scenario is a better explanation of the data: the *udp* intein is only found in *Halorubrum sp.* and for *rfc*-d, all inteins are mini inteins. To further test this we compared the correlation between the extein sequences and the reference signal with the correlations between the intein sequence and reference signal (Figure 6). All four of the intein alleles where the intein and extein datasets are highly correlated are also highly correlated with the reference dataset. This is further support that these patterns are most likely generated through an invasion event in a shared ancestor rather than homologous replacement of the extein and intein sequences.

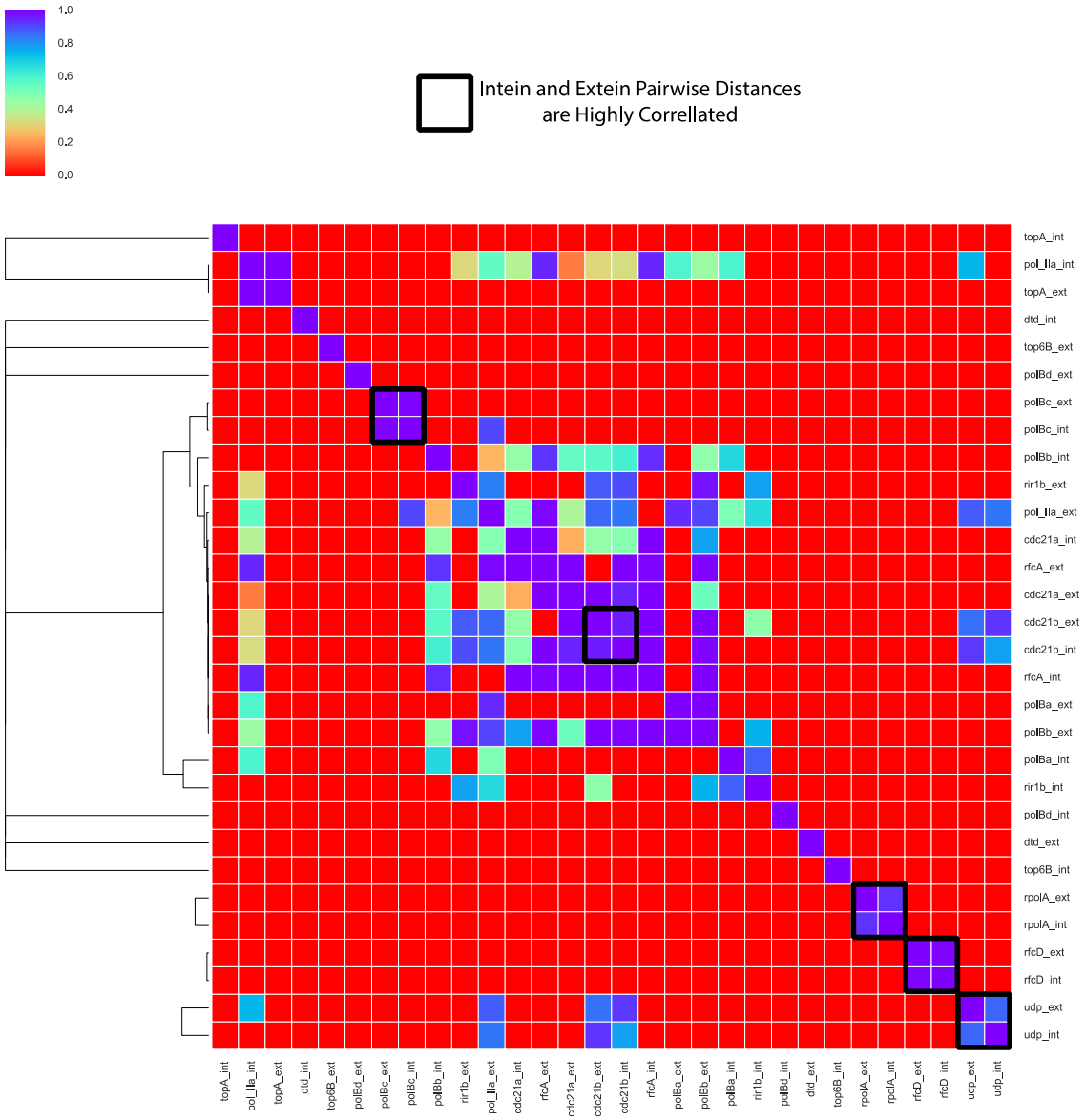


Figure 5. Intein and Extein correlation. Most inteins do not correlate with their corresponding extein, inteins that do show strong correlation with their corresponding extein are outlined in black boxes.

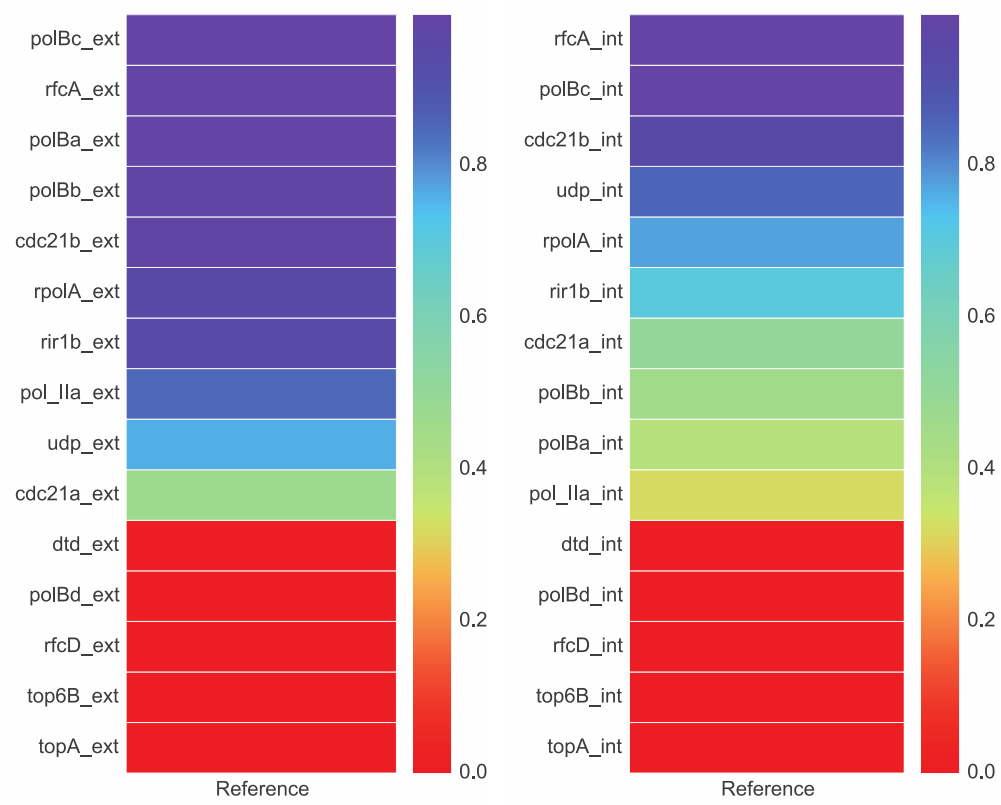


Figure6 Correlation of extein and intein datasets with the reference dataset. Strong correlation is on the violet blue scale, weak correlation is in the green and yellow region, and no correlation is in red. Correlations with $p > 0.07$ were automatically considered no correlation, as the correlation score was not significant.

5.3.4 Exploring Conflicts between Intein and Reference Datasets

To investigate the extent of the HGT signal in each intein dataset we used correlations with the reference set as a measure of the extent of conflict in each intein allele (Figure 6). As expected the agreement with the reference set was much less robust in the intein datasets. We then clustered the intein alleles according to the extent of conflict with the reference dataset. There are 3 main clusters, one that is strongly correlated with the reference set ($R>0.7$), one that is weakly correlated ($0.7 > R > 0.3$), and one with no correlation ($R<0.3$). There doesn't seem to be any characteristic that is shared by inteins in the same cluster, beyond the proportion of conflicts in each dataset.

To further investigate which specific groups were causing the conflicting signal we generated a scatter plot that represents a comparison between the expected distance and the observed distance for each intein allele. A line of best fit is drawn and the p-value reported for each scatter plot indicates the strength of the correlation between the intein allele and the reference set. Points that are a significant distance from the line of best fit represent conflict with the expected signal, and the slope of the line is a representation of the evolutionary rate for the sequences being compared to the reference. Points below the line of best fit represent conflicts that can be resolved through reconciliations with the dataset used in this work, points above the line represent transfer events that involved lineages that are not represented in the dataset used here.

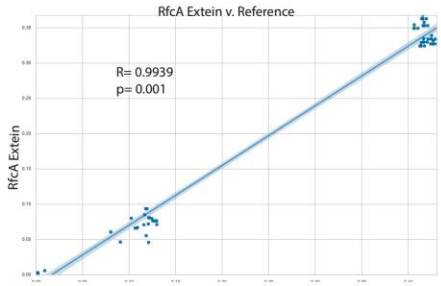
All inteins in cluster 1 (rfcA, polBc, cdc21-b, udp, rpolA, and rir1b) are in good agreement with the reference and most comparisons fall close to the line of best fit (Figure 7). Also the slope of the best-fit line in intein and extein scatter plots for cluster 1 is very similar, indicating similar evolutionary rates relative to the reference dataset. Inteins in cluster 2 had several conflicts and only those below the lines of best-fit can be resolved using the dataset in this work (Figure 7).

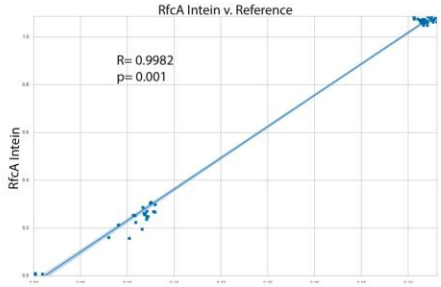
To investigate patterns of intein evolution more closely and try to improve the fit of the line for the larger dataset we constructed a scatter plot where points were colored differently depending on the pairwise comparison. In one set we colored points based on inter or intra-genera comparisons (Figure 8). As expected most intra-genera comparisons fell on the right side of the plot, however in *pol-II*-a there were some intra-genera comparisons that fall away from the cluster on the right. These most likely represent comparisons between species of *Halorubrum* in this group, as the phylogeny shows three distinct clusters. For *cdc21*-a this coloring scheme didn't add more information to scatter plot, all points behaved as expected. To further explore intein evolution in the haloarchaea, we also colored pairwise comparisons based on the presence of the HEN domain. These plots have fewer points as we compared only used data points that compared mini inteins to other mini inteins in the dataset, and the same for full sized inteins. This coloring scheme did not improve the interpretation of the data for *pol-II*-a, as both mini and full size inteins could be found clustering together throughout the plot. Interestingly though, for *cdc21*-a mini inteins and full size inteins formed almost parallel lines of best fit, with the mini-intein line below the full size intein. This could be interpreted to reveal that averaging over the whole intein the substitution rate of mini inteins is smaller than for large inteins containing an HEN. Also the majority of points above the line in the intra-genera comparison are lost in the plot colored by intein size, probably because these conflicts between ribosomal distance and intein distance are due to the lack of the HEN domain.

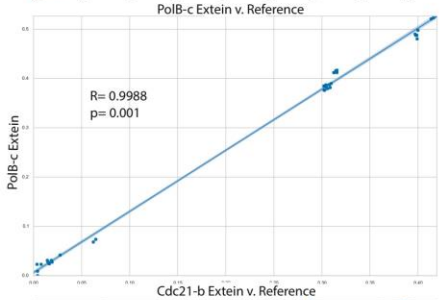
5.4 Future Directions

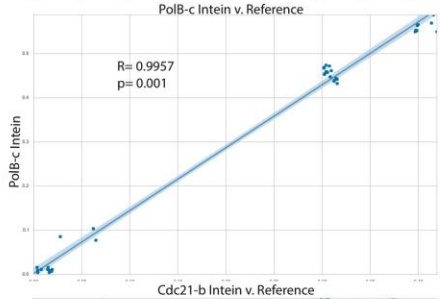
The next step in the process is to identify gene exchange partners in each of the intein alleles. To do this we will use a phylogenetic reconciliation program like ranger-DTL (16) to reconcile gene transfer events indicated by points that fell significantly below the line of best fit in the scatter plot. In this way we are able to exclude reconciliations of gene transfer events that

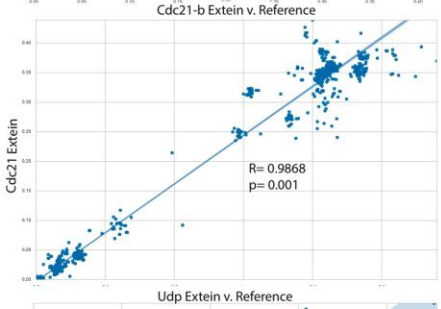
likely involved organisms that are not represented in the current dataset. Once we have a list of reconciliations for each intein allele we can enumerate the pathways of gene exchange represented in total by all intein alleles to build a map of gene exchange in the haloarchaea using inteins.

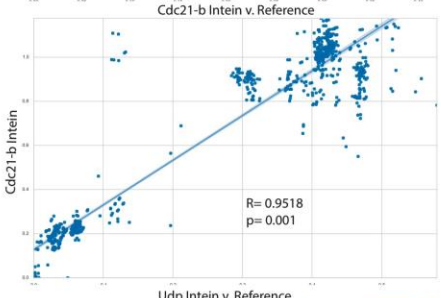

RfcA Extein v. Reference
R= 0.9939
p= 0.001
RfcA Extein

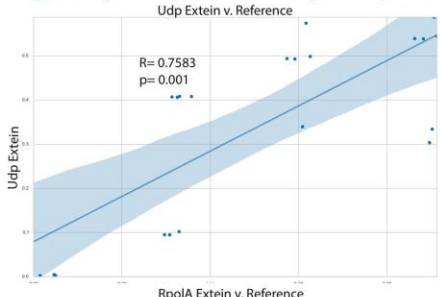

RfcA Intein v. Reference
R= 0.9982
p= 0.001
RfcA Intein

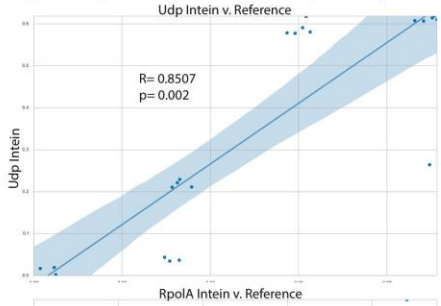

PolB-c Extein v. Reference
R= 0.9988
p= 0.001
PolB-c Extein

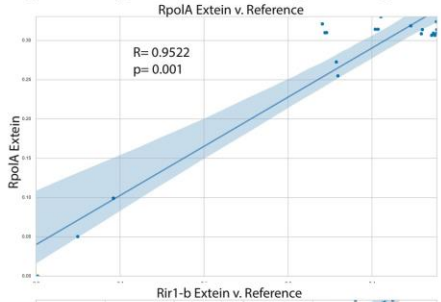

PolB-c Intein v. Reference
R= 0.9957
p= 0.001
PolB-c Intein

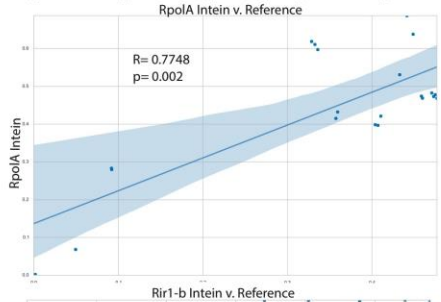

Cdc21-b Extein v. Reference
R= 0.9868
p= 0.001
Cdc21 Extein

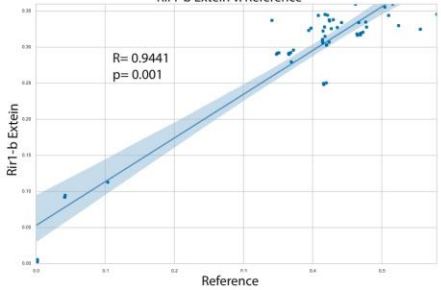

Cdc21-b Intein v. Reference
R= 0.9518
p= 0.001
Cdc21-b Intein


Udp Extein v. Reference
R= 0.7583
p= 0.001
Udp Extein


Udp Intein v. Reference
R= 0.8507
p= 0.002
Udp Intein


RpolA Extein v. Reference
R= 0.9522
p= 0.001
RpolA Extein


RpolA Intein v. Reference
R= 0.7748
p= 0.002
RpolA Intein


Rir1-b Extein v. Reference
R= 0.9441
p= 0.001
Rir1-b Extein
Reference

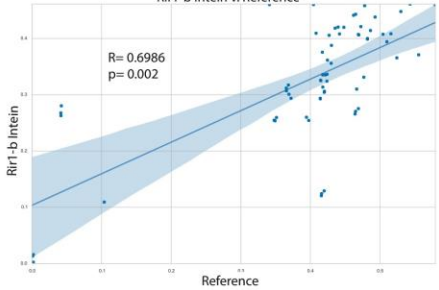

Rir1-b Intein v. Reference
R= 0.6986
p= 0.002
Rir1-b Intein
Reference

Figure 7 Pairwise distance ratios of inteins and exteins relative to the reference distances. Pairwise distances are plotted for all members of each intein allele. Extein comparisons are on the left and inteins are on the right. Intein alleles are listed in order of agreement with the reference dataset as shown in figure 7, intein alleles in cluster 1 are colored blue, and intein alleles in cluster 2 are colored green.

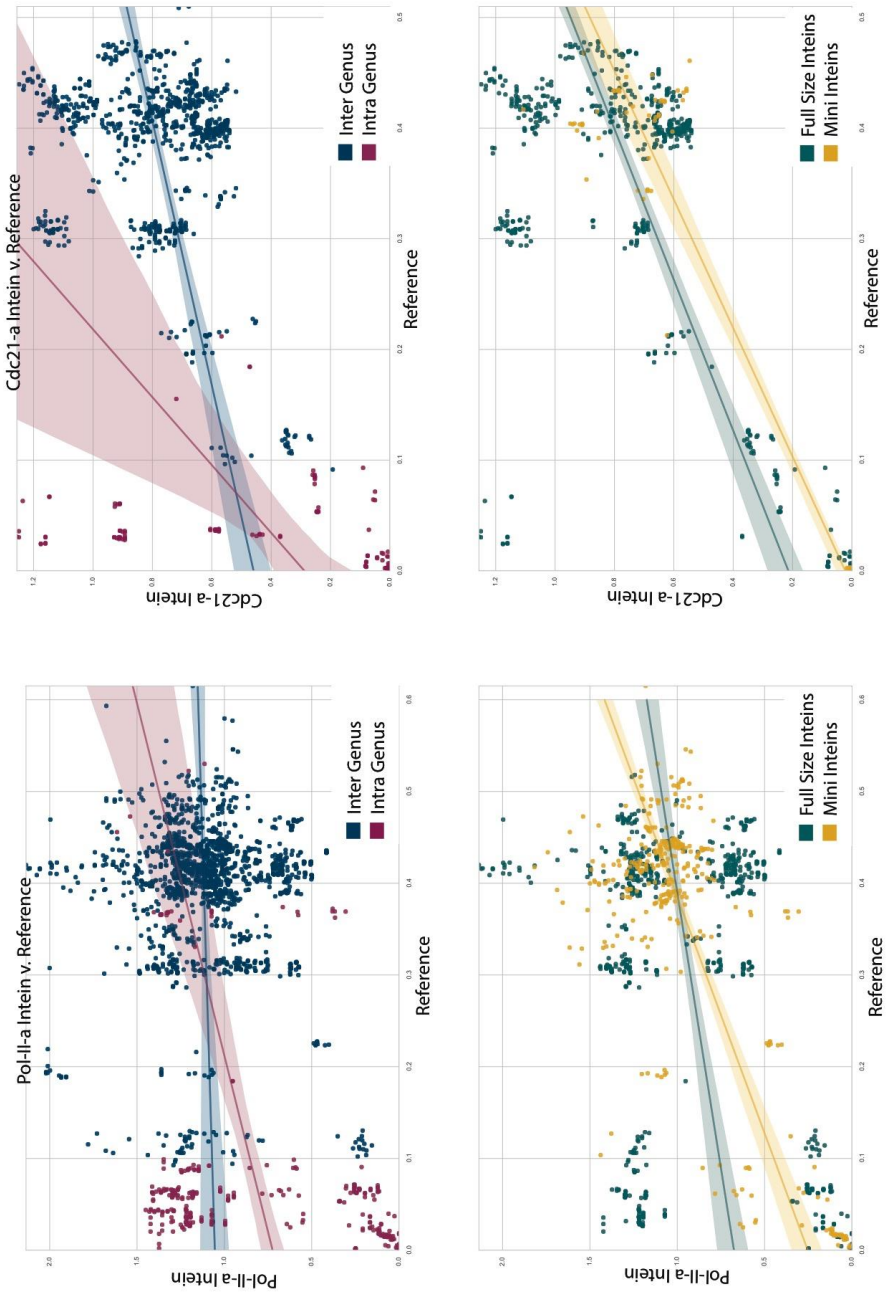


Figure 8 Scatterplot colored by characteristics of genomes compared. The two panels on the left represent *pol-II*-a pairwise distance scatterplots, and on the right are scatterplots of *cdc21*-a. The distance between the genomes being compared colors the two panels on the top. In pink are comparisons within genera, and in blue are comparisons between genera. The bottom panels are colored by the presence of the homing endonuclease domain, in yellow are mini-inteins lacking the domain and in turquoise are full size inteins that contain a homing endonuclease.

Chapter 6. Benefits of Imperfection: Inefficient Homing Endonucleases increase Genetic Variation by Promoting Recombination.

This work was done in collaboration with Dr. Uri Gophna's lab at Tel Aviv University. I was mainly responsible for the simulations and parameter estimation that was based on the bench work conducted in the Gophna lab. This paper was recently accepted for publication by PNAS.

6.1 Abstract

Inteins are parasitic genetic elements that excise themselves at the protein level by self-splicing, allowing the formation of functional, non-disrupted proteins. Many inteins contain a homing endonuclease (HEN) domain, and rely on its activity for horizontal propagation. However, successful invasion of an entire population will make this activity redundant, and the HEN domain is expected to degenerate quickly under these conditions. Several theories have been proposed for the continued existence of the both active HEN and non-invaded alleles within a population. However, to date these models were not directly tested experimentally. Using the natural cell fusion ability of the halophilic archaeon *Haloferax volcanii* we were able to examine this question *in vivo*, by mating *polB* intein positive (insertion site c) and intein negative cells and examining the dispersal efficiency of this intein in a natural, polyploid population. Through competition between otherwise isogenic intein-positive and intein-negative strains we determined a surprisingly high fitness cost of over 7% for the *polB*-c intein. Our laboratory culture experiments and samples taken from Israel's Mediterranean coastline show that the *polB*-c inteins do not efficiently take over an intein-less population through mating, even under ideal conditions. The presence of the HEN/intein promoted recombination when intein-positive and intein-negative cells were mated. Increased recombination due to HEN activity contributes not only to intein dissemination but also to variation at the population level, since recombination tracts during repair extend substantially from the homing site.

6.2 Authors non-technical summary

Parasitic interactions can result in changes to the host's behavior in a way that promotes the distribution or life cycle of the parasite. Inteins are molecular parasites found in all three domains of life. Here we look at the influence of an intein in the DNA polymerase on a population of halophilic archaea in simulations, in experiments, and in the wild. This intein has a fitness cost that is higher than expected for a self-splicing genetic element. In these populations, where mating is independent of host replication, the intein increases the recombination rate between cells with and without inteins. This modification may contribute to the long-term persistence of these genetic parasites, despite the fitness burden they impart on their host.

6.3 Introduction

Homing endonucleases (HENs) are selfish genetic elements found in all domains of cellular life, as well as many viruses. HENs recognize and cleave highly specific target sequences, up to 40 bp long, usually corresponding to a singular site within the genome (17, 18). HENs can reside within conserved genes since they are nearly always located within self-splicing selfish elements: either group I introns [(19), for review see (20)], which excise themselves at the mRNA level, or inteins, which splice out of the protein product [(21), for a recent review see (8)]. HENs contribute to the horizontal transmission of these selfish elements into intron-less or intein-less alleles, by cleaving the vacant allele to induce homologous recombination or reverse transcription, where the allele containing the intron or intein serves as template. Thus the intein or intron effectively invades the vacant site and can later be passed on to daughter cells vertically.

Paradoxically, if a HEN is highly successful in invading cells it will saturate all target cells and then its activity will no longer be under purifying selection. This may result in degeneration of the HEN domain due to accumulation of mutations, and will prevent the future horizontal propagation of the intein in question. This phenomenon has been observed in group I introns of bacteriophages (22, 23), eukaryotes (24) and archaea (25). However, active HENs are often observed, and in most natural populations surveyed to date, there are in fact strains with inteins (or introns) containing degenerate HENs, while other isolates maintain a fully intact HEN (24). The homing cycle model (26, 27) resolves this paradox by hypothesizing that after the intein/intron containing allele has been fixed in the population, and after the HEN has completely decayed due to the lack of selection for function of HEN activity, the splicing element (either intron or intein) can be deleted, thus returning the homing cycle to its beginning. In the case of introns, a precise deletion is somewhat likely and can occur via a processed mRNA intermediate (28, 29). In contrast, a precise deletion of an intein by a random event is expected to be very rare (30). In both introns and inteins, an imprecise deletion in a critical gene would often yield a dysfunctional product that may be highly deleterious or even lethal (7); we suggest that a precise deletion mainly occurs through homologous recombination with an intein-free allele that is transferred into the population, or that survived within the population in case of incomplete invasion.

Recently two models based on incomplete invasion were described. Both are based on the coexistence of intein-free and intein-containing alleles with and without functioning HEN in homogeneous populations. It was demonstrated computationally that if the cost of an inactive HEN is lower than that of the active HEN, the 2 forms can coexist with alleles carrying the empty target site even in a homogeneous population (30, 31). In many instances, the rock-paper-

scissors dynamic between the three allele types (empty target site, intein with HEN, intein without an active HEN) results in oscillations typical for predator-prey system (30). In this scenario both the intein (or intron) and the HEN activity have separate and cumulative fitness costs. Intein-negative cells are slightly more fit than their intein-containing counterparts and therefore out-compete intein-positive cells, while becoming invaded. Cells with an intein but a defective HEN are slightly fitter than those with an intact HEN because they do not suffer the fitness cost from the large size of the HEN or the cost incurred from HEN activity such as off-target cleavage, but are still immune to invasion. Under these assumptions a tri-partite equilibrium between these three cell populations can result that may undergo periodical oscillations (30, 31).

While these models provided fresh insights into the HEN enigma, they assume homogeneous conditions throughout the population, and employ parameter estimates that were never tested, such as fitness costs associated with HEN activity and/or intein presence. Here we use the advanced genetic tools available for halophilic archaea to test the validity of these assumptions and gain a new perspective into HEN/intein dynamics. Halophilic archaea were shown to undergo a unique mechanism of cell fusion (32–34), in which two or more cells can become fused into one cell, containing all the genetic material of the parental cells. This "hetero-diploid" (heterozygous) state, where two different chromosome types are present in the same cell, allows contact between HEN-positive and HEN-negative alleles. Thus, the efficiency of the spread of an active HEN can be assessed in a realistic natural system. We have previously shown that the HEN that resides in the *polB*-c (insertion site c) intein in *Haloferax volcanii* is highly active *in vivo* and converts close to 100% of engineered plasmids that contain its homing site

upon transformation. Interestingly, deleting the entire intein did not significantly alter the growth rate of *H. volcanii* cells (35). An additional aspect of this experimental system is that *H. volcanii* was shown to be polyploid, the number of chromosome copies per cell ranging between 10-20 (36). Here we explore through simulations the consequence of the invading HEN having to convert multiple chromosomes, running the risk of being "back-converted" into a vacant allele through homologous recombination. We demonstrate experimentally that intein invasion is an inefficient process in *H. volcanii* and that having an intein results in a substantial fitness cost. These genetic experiments were followed up by a survey of *H. volcanii* strains isolated from various sampling locations along the eastern Mediterranean shore, facilitating a comparison between laboratory findings and eco-genetic observations.

6.4 Materials and Methods

6.4.1 Culture conditions

H. volcanii cells were routinely grown in rich (YPC) medium, or in CAS medium (see (35)). Nucleic acids and amino acids were added to a final concentrations of 50 µg/ml (Sigma)

6.4.2 Competition assays

H26 and H12 (see Table 1) cultures were grown over night, and when they reached an OD600 ~1, the cells were diluted to OD600 ~0.2, and allowed growth until OD600 ~0.4, where they were all diluted to OD600 ~0.1 and mixed 1:1. The cells were plated to obtain a measurement of ratios at time "0". The cultures were diluted 1:50 every 24 hours for 6 days. The experiment was performed with four biological repeats and each repeat was done in duplicates. In order to distinguish between intein-containing cells and intein-deleted cells, we preformed PCR analysis using primers AP409 and AP410, which amplify the *polB* region surrounding the intein, and that enabled the assessment of the number of colonies on each plate that originated from an intein-positive or intein-negative cells. From the 0 and 3 day time points, about 95 colonies were

screened for each experiment, and for the 6 day time point 48 colonies were screened. We defined 'fitness cost of genotype x vs. genotype y as the relative change in average growth rate calculated as in (37) using the cell numbers at days 0, 3, and 6 with the appropriate dilution factors (see spreadsheet in Supplementary Dataset S1)).

6.4.3 Mating protocol

Each culture was grown to an OD_{600nm} of 1.1-1.3, and 2 ml samples were taken from both strains and applied to a 0.2 micron filter connected to a vacuum to eliminate excess medium. The filter was then placed on a petri dish containing a rich medium (YPC medium + thymidine, see below) for 48 hours at 42°C. The cells were washed then resuspended in casamino (CA) broth, washed twice more in the same medium and plated on selective media.

6.4.4 Examination of intein spread

To estimate the efficiency with which this intein/HEN spreads, we mated strains H729 and HAN17 (Table 1, Supplementary Fig. S1). Following mating, the cells were plated using selection for thymidine and uracil (using CAS medium supplemented with tryptophan), and thus only mated cells could form colonies. From these colonies, which must contain both selection markers, 8 single colonies were picked per mating experiment (4 independent mating experiments were preformed), and were grown in liquid medium in the absence of selection for 4-6 generations, allowing all cell types to grow and hetero-diploid cells to segregate. Subsequently, cultures were re-plated, and from each of these 32 mated cultures, 11-50 single colonies were analyzed to establish their genotype and intein status. To establish whether each colony originated from the Δintein (HAN17) strain or from the wild type *polB* strain (H729, Table 1)), we examined the *trpA* marker, which is 62Kb from *polB* and therefore generally unlinked to it. Each colony was streaked on media with and without tryptophan and the *polB*

allele was tested using PCR (primers AP409 and AP410), and thus trpA+ cells are ones that originated from H729 (with intein) and trpA- cells originate from HAN17.

6.4.5 Determination of recombination frequencies

To estimate recombination frequencies we preformed three mating experiments, all under the same conditions, repeated 3-5 times. The cells after mating were plated using selection for thymidine and uracil, and the medium used contained tryptophan. The mated cells were analyzed as previously described (33), by testing the *trpA* marker by PCR (using primers AP214 and AP215), where colonies containing two copies of the gene, both $trpA^+$ and $trpA^-$ are hetro-diploid colonies, and ones containing either $trpA^+$ or $trpA^-$ are recombinants. We have previously shown that cells that are either $trpA^+$ or $trpA^-$, show a single type of chromosome in other locations along the chromosomes and are true recombinants (33).

6.4.6 Mating experiments to test the linkage between the intein/*polB* and *trpA* loci

To further quantify the success of intein invasion using selection for growth on a selective medium lacking tryptophan, strain HAN17 (Table 1) was transformed with a suicide plasmid containing 500bp each of the upstream and downstream flanking sequences of the gene Hvo0894 cloned into pTA131 (38) thus generating strain UG417 that is $pyrE2^+$ (see Supplementary Fig. S3). Strain UG417 was then mated with strain H729 ($trpA^+int^+$, Table 1, Supplementary Fig. S3) and mated cells were selected on CAS medium lacking thymidine, uracil and tryptophan. This medium only allows growth of heterozygous/heterodipoloids that have both of the parental genotypes or recombinant cells that have wt *pyrE2, hdrB* and *trpA*. Since we selected for three genomic markers two of which are on the UG417 genome we expected that all recombinant cells would be UG417 that have replaced their *trpA* locus with the $trpA^+$ from H729 via homologous recombination. 96 colonies from 3 biological repeats were screened with *hdrB* primers AP121 (5' CCCGCCTCGCCGACGTGCAGT 3') and

AP122 (5' GGAGTTGGTCTGCGAGTGTCG 3') and were all found to be homozygous and hence recombinant. The colonies were further screened with intein primers AP409 and Ap410.

To examine intein invasion rate independently of the *trpA* locus, Strain H729 was transformed with pWL102 (39) a shuttle vector that contains a mevinolin resistance marker. H729- pWL102 was then mated with strain UG417. Mated cells were selected on CAS medium supplemented with tryptophan and mevinolin lacking thymidine and uracil hence selecting for *hdrB* and *pyrE2*, two genomic markers on UG417 (Supplementary Fig. S3). Using this selection only heterozygous/heterodipoloids or cells with the UG417 that obtained pWL102 during mating could grow. This mating assay was repeated 3 times and from each biological repeat 96 colonies were selected and streaked on CAS medium lacking tryptophan as well as CAS medium supplemented with tryptophan to assess recombination at the *trpA* locus. All colonies were also subjected to PCR analysis using intein primers AP409 and Ap410 to assess intein invasion.

6.4.7 Sample collection, 16S rRNA and *polB* gene sequencing

Stagnant seawater or dry salt samples were collected from 14 tidal/spray pools from three rocky shores along Israel's Mediterranean coastline. 5ml of seawater from each pool was spread onto a YPC plate containing ampicillin. Dry salt samples were dissolved in sterile seawater prior to spreading on plates. Colonies with typical *Haloferax* coloring and shape were further examined by 16S rRNA gene PCR using Halobacteriales-specific 16S rRNA gene primers 287F and 958R (40). The ~600 bp amplicon was sequenced by ABI 3730XL sequencers at MCLAB DNA Sequencing Services (San Francisco, CA, USA). Colonies that proved to belong to the *Haloferax* genus (100% identity to a *Haloferax* sp. in NCBI nucleotide BLASTN) were subjected to PCR analysis and sequencing of the *polB* gene. The *polB* gene was amplified using primers AP8 and AP11. Intein presence or absence was determined according to the length of the amplified *polB* product. The *polB* sequence of the isolates was determined using primers AP8, AP11, AP439,

AP409 and AP410. For sequence verification each isolate was sequenced twice using all five primers and only areas covered by at least two overlapping sequences were considered 'error free' and used for genotype determination. Distinct *polB* genotypes showed sequence difference of at least two nucleotides in at least two locations (namely, two adjacent nucleotides did not count as two different nucleotides but only as one).

6.4.8 Phylogenetic reconstruction and sequence logos

polB sequences were aligned in MUSCLE (12) using the default parameters as implemented in Seaview4.3 (13) (Supplementary Dataset S3). The phylogeny of extein sequences was calculated with Phyml (41) in Seaview4.3 (13) under the GTR + Gamma + I model. Numbers give support values calculated using the approximate Likelihood Ratio Test as implemented in Phyml 3.0 (42). Sequences surrounding the insertion site c were extracted and weblogos (43) were calculated from the aligned extein sequences.

6.4.9 Simulations to explore the effects of ploidy

Intein invasion during mating and colony growth was simulated to determine the effects of ploidy, homing endonuclease efficiency, and fitness cost of the intein on the rate of intein invasion. We do not know how the chromosomes segregate following fusion; therefore, we modeled the events following cell fusion in two different ways. In the first, two cells, one intein positive and one intein negative cell, form a fused single cell and each chromosome segregates independently into the daughter cells. In the second simulation chromosomes from each parental type are randomly assorted into daughter cells and one of the daughter cells is chosen at random and followed for a given number of generations. For both simulations during each generation, the probability that an intein negative chromosome is invaded are calculated as follows:

Each un-invaded chromosome may be cut with probability:

$$P_c = E * \frac{X_n}{N}$$

Where E is the efficiency of the homing endonuclease, X_n is the number of homing endonucleases in the nth generation, and N is the total number of chromosomes in the fused cell. This normalizes the total activity so that the HEN activity in the fully invaded cell is the same for different ploidy levels (i.e. the amount of cytoplasm per chromosome is constant).

If a cut occurs on a given chromosome, the repair of the cut with an invaded sequence is dependent on the number of invaded sequences in the cell. The repair of the cut site with an invaded sequence occurs with probability:

$$P_i = \frac{X_n}{N - 1}$$

The minus one in the denominator excludes the cut chromosome from the calculation (i.e., it cannot act as template for its own repair).

Cells with a newly invaded chromosome divide with probability 1-*f*, where *f* is the fitness cost per intein.

6.4.10 Parameter Estimation

The estimated number of generations of growth after mating is 34. This is based on approximately $2*10^8$ cells per colony, or approximately 28 generations of growth. The cells were grown for another 6 generations after the colonies were picked, giving a total of 34 generations.

The fitness cost (*f*) is the decrease in organismal fitness per intein acquired. Based on the co-culture competition experiments (Fig. 2), a completely invaded 20-ploid cell has a fitness decrease of 7.2% compared to an identical un-invaded cell. We use a fitness cost of 0.075% per cell carrying the intein in the population.

The efficiency of the homing endonuclease (*E*) was calculated using deterministic simulations in Excel. The number of invaded chromosomes in a given generation was calculated according to the formula:

$$X_{n+1} = X_n + (N - X_n) * E * \frac{X_n}{N} * \frac{X_n}{N-1} - X_n * f$$

N: number of chromosomes per cell

X_n: number of chromosomes with intein for these mating simulations

n: the generation number.

N-X_n: is the number of alleles that can be invaded;

E: is the efficiency of the homing endonuclease

The factor X_n/N reflects the amount of HE per genome.

The factor $X_n/(N\text{-}1)$ gives the fraction of templates that contain the intein and would lead to insertion of the intein if the template were used to repair the double strand cut.

f : is the fitness cost per intein.

Notice that this formula is equivalent to $X_n + (N - X_n) * P_i * P_C - f * X_n$

In these deterministic calculations, a homing endonuclease efficiency (*E*) of 0.06 leads to invasion of 81% in the case of a 20-ploid cell (spreadsheet in Supplementary Dataset S4). This is consistent with results that 68.5% of un-invaded cells become invaded and 92.5% of invaded cells remain invaded after mating, for a total of 80.5% of cells invaded after mating.

6.4.11 Simulating intein invasion in a population with a limited carrying capacity

We simulated intein invasion in populations that had a specific carrying capacity using a modified discrete logistic equation. We use the following formulas:

$p_{(n+1)} = p_n + p_n * q_n * hm + r * (1-f) * (k - p_n - q_n) * p_n$

$q_{(n+1)} = q_n - p_n * q_n * hm + r * (k - p_n - q_n) * q_n$

Where *p* denotes the proportion of the population that contains the intein and *q* is the proportion of the population that does not contain the intein; *hm* represents the compound probability of an intein + and intein – cell exchanging DNA and the efficiency of the HEN domain making a double strand break and the intein invading; f represents the fitness cost to individual cells for carrying the intein; *r* is the rate of population growth, and k the carrying capacity (we set k=1 to express populations as fractions of the carrying capacity). The growth rate of the intein-free and intein-containing cells is r and (1-f)*r respectively. Following the logistic equation, the growth rate is multiplied by k minus the current population size. For the simulations depicted in Fig. 4 we considered the contribution of intein-containing and intein-free cells to the carrying capacity as equal. To incorporate an impact of the intein on both the growth rate and the carrying capacity, we modified the equation to weigh the contributions of p and q to the carrying capacity (see Supplementary Fig. S2):

$p_{(n+1)} = p_n + p_n * q_n * hm + r*(1-f)*(k-(p_n/(1-f)) - q_n)*p_n$

$q_{(n+1)} = q_n - p_n * q_n * hm + r*(k-(p_n/(1-f)) - q_n)*q_n$

6.5 Results

6.5.1 Intein presence incurs a fitness cost

We had previously observed that the growth rate of the intein deletion strain, is highly similar to its parental strain (33); however, growth rates were compared using growth curve analysis, a method that cannot detect small differences in fitness, or ones associated with the size of the lag phase when growth is resumed after cells from stationary phase are transferred into fresh medium. We therefore performed direct competition assays between a strain containing the intein (H26, see Table 1) and its intein-deletion isogenic strain, (HAN12). Fig. 1 shows the relative abundance of intein-containing and intein-deletion cells in the mixed cultures, at time 0, and after 3 and 6 days of co-growth. It is evident that cells containing the intein grew slower and

were outcompeted by the intein-negative cells. Following the approach described by Lenski et al. (37) we used the change in average growth rate to quantify the intein's effect on host fitness. Calculating the growth rates from three time points in each of eight parallel experiments, we calculated the relative fitness of the intein harboring cells to be 92.8% (Standard Error of the Mean: 0.4%), *i.e.*, the fitness cost of the intein is 7.2% (SEM 0.4). This corresponds to an increase in the average doubling time from 4.17 to 4.49 hours averaged over the repeated culture cycles (see spreadsheet in Supplementary Data 1 for the calculation).

Table 1. Strains used in this study

Strain	**Description**	**Source**
H26	*ΔpyrE2*	(38)
H729	*ΔhdrB*	(33)
HAN12	*ΔpyrE2Δintein*	(35)
HAN17	*ΔpyrE2 ΔtrpA Δintein*	This study
HAN24	*ΔhdrB Δintein*	This study
H53	*ΔpyrE2 ΔtrpA*	(38)
UG417	HAN17 *Hvo00894::pTA131*	This study

6.5.2 Intein spread is not completely efficient during cell fusion events

Examination of the natural invasion capacity of an intein/HEN in archaea requires a system where intein/HEN-positive cells can come into contact and fuse with cells carrying vacant alleles of the same gene that can be invaded. Presumably, invasion requires contact between alleles as well as the expression of the HEN protein domain, and thus in this work only cells that undergo fusion can be invaded. We therefore selected for mated cells so that invasion frequency can be simply assessed, by mating two strains that carry different gene deletions, serving as selectable markers (see Materials and Methods). We used the following strains: H729 - auxotrophic for

thymidine (*ΔhdrB*) and contains an intact intein with a HEN, and HAN17 (see Supplementary Fig. S1): auxotrophic for both uracil and tryptophan (*ΔpyrE*, *ΔtrpA*) and intein-negative – thus containing a vacant HEN cleavage site. Following mating, we selected for mated cells by plating on media lacking thymidine and uracil, such that only cells that underwent mating could grow. The cells that underwent mating were then grown for 4-5 generations in rich media, and subsequently plated, and each colony had its genotype determined. After establishing the original genotype, either H729 or HAN17 using the *trpA* locus, we examined intein presence using PCR for each colony (see Spreadsheet in Supplementary Data S2). As shown in Fig. 2, 68.5% of the cells that had the *trpA* locus from the intein-deleted parent (i.e. a $trpA^-$ allele – HAN17), became intein-positive, and 31.5% of the cells that underwent mating remained intein-negative. We also examined the cells that retained the $trpA^+$ allele from the intein-positive parent H729, and as expected most (92.5%) of the cells remained intein positive; however, over all four biological replicates 7.5% of the cells were now intein-negative (mean fraction of 5.5%, see Fig. 2). This is probably due to a random recombination and gene conversion events, not involving endonuclease activity that resulted in the elimination of the intein-occupied allele. Such events can mechanistically explain how such vacant alleles are formed without resorting to additional molecular mechanisms, such as precise intein deletion. Since even under conditions where all intein-negative cells are forced to make contact with intein containing alleles, homing efficiency was less than 70%, *i.e.*, nowhere near saturation, inteins are unlikely to rapidly invade all cells in a natural population.

6.5.3 Simulation of intein invasion

H. volcanii, like other haloarchaea and methanogens, are polyploid, with the number of genomes per cell varying between about 20 (during exponential growth) and about 10 during stationary

phase (36). To assess the impact of polyploidy on intein invasion of a lineage we performed several simulations. In the simulations depicted in Fig. 3 chromosomes were assumed to segregate randomly after mating. Under this condition a higher ploidy level increases likelihood of complete invasion of the lineage, but the rate of invasion is delayed with higher levels of ploidy.

To illustrate the effect of a high fitness cost on intein invasion of a population we performed simulations of invasion of a population that over time approaches the carrying capacity of the environment (the maximum population size that the environment can support long term) (Fig. 4 and Supplementary Fig. S2). A recent study of gene flow and recombination in a deep lake (9) suggests that gene transfer and homologous recombination continue in the absence of population growth in halophilic archaea. In the simulation in Fig. 4, the intein-free genotype outcompetes the intein-containing genotype during the initial growth phase, but the intein-containing allele spreads in the population, once the overall growth rate declines. Incorporating the fitness effect of the intein into the contribution of intein-containing cells on the carrying capacity of the population does not change the overall outcome, except that the resulting population size decreases as the intein-free cells are invaded by the intein (see Supplementary Fig. S2).

6.5.4 Mating between intein-positive and intein-negative cells results in increased recombination frequencies.

Since the HEN generates double strand breaks, which can be repaired by homologous recombination, we tested the link between recombination frequency and the presence of the intein/HEN. Importantly, recombination frequency can increase when intein/HEN-positive intein/HEN-negative cells come together due to homing, but potentially also due to off-target

cleavage by the HEN at non-specific sites. We tested the overall recombination frequency following natural cell fusion events, among the strains described above. Following mating, the cells first become hetero-diploid (heterozygous), containing chromosomes that originated from two different cells. However, a substantial fraction of the population undergoes recombination and segregation so that the resulting cells contain a single chromosome type. We calculated the frequency of recombination, using PCR [as in (33), see Methods], by measuring the fraction of recombinant colonies from the general mated population (which contains a majority of heterodiploid

cells). The fraction of recombinant colonies that have a single chromosome type but can grow on a medium lacking thymidine, tryptophan and uracil is shown in Figure 5. To estimate the effect of intein on recombination we mated intein-positive cells with intein-negative cells (H729 and HAN17), in this scenario one cell contains an intein with an active HEN while the other chromosome contains an empty *polB* allele which is the recognition site for this HEN. As control experiments we also performed such mating experiments between pairs of strains that were both either intein-positive (H729 and H53) or intein-negative (HAN17 and HAN24). Notably, the recombination frequency was similar for intein-positive and intein-negative pairs of strains, between 33% and 35% recombinants out of the entire mated population. This near identical rate of recombinants implies that off-target DNA cleavage by the PolB HEN is not sufficient to increase recombination rates, and is therefore probably very low. However, when mating an intein-positive strain with an intein-negative partner, the recombination efficiency was markedly higher, with 48.6% of the mated cells being recombinants (Fig. 5). This suggests that the specific cleavage of the "empty" site in the *polB* gene (homing) by the HEN caused an increase in

recombination frequency, probably because of the presence of double-strand breaks due to homing.

6.5.5 Recombination tracts in *H. volcanii* following HEN activity can extend over 50 Kb

These experiments raised the question whether the homologous recombination tract initiated at the intein insertion site can extend well beyond the hundreds of bases within the extein to thousands of bases away, as was previously observed in archaea (44, 45), and bacteria (46). To investigate this, we established a simpler intein invasion assay where the intein donor only has to transfer a small plasmid to the recipient strain and all donors cannot replicate because they cannot synthesize uracil or thymidine. This was achieved by mating a donor intein-positive strain carrying a mevinoline resistance plasmid that is *hdrB*$^-$/*pyrE*$^-$/*trpA*$^+$ with a recipient strain that is mevinoline-sensitive and *hdrB*$^+$/*pyrE*$^+$/*trpA*$^-$. By growing the mated cells on a medium that contains mevinoline and tryptophan, and does not contain uracil or thymidine, donor cells were selected against. This effectively leaves only *hdrB*$^+$/*pyrE*$^+$ cells that received the mevinoline-resistance plasmid, since the chances of two recombination events at very distant loci (over 1Mb) are much smaller than the chances of successful transfer of a small plasmid during mating, see also Supplementary Fig. S3). Following selection colonies were screened by PCR for intein presence or absence as described above, excluding heterozygous colonies (those that contained more than one genotype, less than 25% of the colonies). We observed that using this strain combination and plasmid-transfer based selection, only about 30 % of the colonies were invaded by the intein (Table 2). However, when intein-invaded colonies were streaked on a medium that lacked tryptophan, 23±6% of the intein-positive were *trpA*$^+$ vs. 10±3% of the intein-negative colonies. This demonstrates that the HEN target site, *polB,* and *trpA* gene are linked in recombination despite the fact that these loci are over 60 Kb apart. When we performed an

identical mating experiment selecting on a medium that lacks tryptophan (requiring recombination at the *trpA* locus since only *trpA*$^+$ progeny survive), we observed that 85% of the colonies were intein-positive further supporting the linkage.

These findings are in agreement with previous results showing recombination tracts of hundreds of Kb between *H. volcanii* and *H. mediterranei* (33).

Table 2:
Percent genotypes of colonies obtained in the plasmid-based mating assay for intein invasion (number of colonies are in parentheses)

	trpA$^+$ *Int*$^+$	*trpA*$^+$ *Int*$^-$	*trpA*$^-$ *Int*$^+$	*trpA*$^-$ *Int-*
Mean (3 biological repeats)	23 (18)	10 (8)	7 (5.7)	60 (47.3)
Standard deviation	6	3	3	5

6.5.6 Inteins in natural populations

If indeed not every mating event leads to successful invasion, and given that under natural conditions only a fraction of cells will fuse with one another, we predicted that not all natural isolates will present the same *polB*-c intein genotype. We therefore isolated from tidal pools along Israel's coastline *Haloferax* strains that can be regarded as belonging to *H. volcanii*, or closely related to it (defined as having identical 16S rRNA gene sequences to that of a *Haloferax* sp.).

At the geographic level, two out of the three coastal sampling locations had both, intein-containing and intein-free isolates, based on PCR amplification of the *polB* gene fragment that may contain the *H. volcanii* intein. Of the 14 tidal pools we sampled in the three locations, 12 contained isolates with *polB* sequences of high similarity to *H. volcanii polB*. Of these sites 9 contained either intein-containing isolates or intein-free isolates, but not both, while 3 pools contained a mixed population (Supplementary Fig. S4). Fifty-four different isolates had their

entire *polB* genes amplified and sequenced and two different intein-free genotypes were identified, as well as seven distinct intein-containing genotypes. A phylogenetic reconstruction of extein sequences revealed that inteins have recently invaded intein-less sites in this population (Fig. 6). Further examination of intein sequence revealed that all nucleic acid substitutions within the intein and HEN motif blocks resulted in synonymous ("silent") mutations with the exception of a single proline to serine substitution in a conserved region of the HEN, known as Block B (35), which was present in two genotypes. Thus, most of the isolates that are intein/HEN positive appear to have a genetically intact HEN, showing no sign of degeneration.

6.6 Discussion

None of the current models for the maintenance of inteins with HENs captures the full complexity of these selfish elements in nature. If one assumes efficient invasion, as in the original homing cycle, one would expect most intein-containing cells in nature to have degenerate HENs waiting to be rescued from this dead end state by a precise deletion event. Here we show that in fact even under ideal conditions -- we look only at the cells in which intein-positive and intein-negative cells have to come into contact -- a substantial fraction of cells are not invaded. We also show that a vacant site can be re-created, if rarely, simply by a random (i.e. not targeted) homologous recombination event that converts the intein-positive allele to an intein-negative, as a by-product of polyploidy. In agreement with our genetic experiments, natural populations of *H. volcanii*-related strains can contain both intein-positive and intein-negative cells, and intein-positive isolates show little to no signs of HEN-degeneration.

Our simulations exploring the effect of ploidy on intein invasion suggest that polyploidy facilitates invasion by the HEN (Fig. 3). Higher levels of ploidy slow down the rate of complete invasion (more successful homing events need to occur before invasion of chromosomes is

complete); however, assuming random assortment of chromosomes into the daughter cells, higher ploidy levels lead to a higher frequency of complete invasion in the long run. The finding that in our mating experiments (Fig. 2) the colonies obtained from heterozygous mated cells after five doubling periods no longer reveal any heterozygosity with regards to the intein suggests that the effective ploidy level of the fused cell may be lower than expected from the measurement of chromosomes (36). A high rate of forming homozygotes from heterozygous *H. volcanii* has been described for engineered heterozygotes in the absence of maintaining selection (47) and was attributed to a high rate of gene conversion. Alternatively or in addition, it is possible that chromosomes in the fusion cell do not mix efficiently, and that the first cell division tends to segregate the chromosomes in a way that recreates the genotypes of the two parents. A low level of effective ploidy could also explain why inteins were fixed in only about 80% of the cells following mating (compare Figs. 2 and 3).

Alternatives to a homing cycle model where intein invasion goes to completion posit a fitness cost for the intein (30, 31) and a co-persistence of intein-free and intein-containing alleles (with and without functioning HEN) in the population. Our simulations show that during exponential growth intein-free alleles can outcompete intein-containing ones (see Fig. 4). The fitness cost of the intein is balanced by conversion of intein-free alleles through homing. We experimentally determined a surprisingly high fitness cost of having an intein-positive allele, at least under laboratory conditions and averaged over the growth cycle. Growth retardation was previously shown during expression of group I introns in the 23S rRNA gene of the bacterium *Coxiella burnetii* (48). In accordance with an equilibrium resulting from the fitness cost of the intein, balanced by the spread of the intein, and in accordance with the simulations performed by (30, 31), we find that many natural isolates have survived retaining intein-positive alleles.

Inteins with HEN can be described as parasitic genetic elements (17). Parasites often impact the phenotype of their host to increase the chance of their own propagation (49), and an impact on host behavior and development has also been described for the Saccharomyces intein (*vma1*-a) (50). The observed increase in recombination, when only one of the genotypes of the mating cells is intein-positive (Fig. 5) may be an illustration of this principle. Clearly, the propagation of the intein is enhanced through increased recombination, and therefore, the increased recombination rate is in the best interest of the intein. However, the recombination tract that is generated by the homing activity is by no means limited to the immediate vicinity of the *polB* gene, since the locus we used to test recombination is located over 60 Kb away (Supplementary Fig. S1) from the homing site. Thus, despite the high specificity of the homing site, homing events can exert indirect effects on gene exchange frequency across large regions of the chromosome, and assessing the distance dependence of these effects requires further study. Nevertheless, although recombination will be primarily induced close to the homing site, it could also be more generally increased, because as long as DNA breaks are generated this enhance transcription of homologous recombination genes, such as *radA*, the archaeal homolog of the bacterial *recA* recombinase. Indeed, *recA* transcription is induced in bacteria following DNA breaks (51), and *radA* was the most strongly up-regulated gene in the halophilic archaeon *Halobacterium salinarum* following UV-B exposure (52). Such induction of *radA* is thus expected to increase recombination rates globally, and not just around the homing site. A relationship between intein-presence and increased rate of recombination is also suggested by the mating type switching HO endonuclease in yeast, which also facilitates mating and homologous recombination between cells through cuts made at a single locus (53). This HO endonuclease is a domesticated intein with endonuclease activity, which is most closely related to the *vma1*-a

intein (see above). In addition to the endonuclease domain, the HO endonuclease still contains a self-splicing domain, albeit not shown to be active (27).

The finding of intein-containing alleles with active HENs that coexist in the same population with empty target sites (Fig. 6 and (54)) suggests that the long term survival of inteins does not include the homing cycle going to completion, rather fitness differences between organisms with and without inteins appear to play a role in the long-term coexistence of the HEN with alleles containing empty target sites. In our simplifying simulations, the HEN was found to completely invade a homogeneous population close to its carrying capacity; however, natural populations of *H. volcanii* are neither homogeneous, nor do they exist in a constant environment. In the paper-scissors-rock models describing long term co-existence of HENs with empty target sites (30, 31), inteins without HEN activity provide the crucial link in the intransitive fitness relationships leading to long-term coexistence. Our findings show that under conditions that include occasional growth of the population, the high fitness cost of the intein guarantees the survival of alleles with the empty target site, even in the absence of inteins without HEN activity. Although here we showed a fitness cost for the intein, several reports have indicated a potential positive role for intein presence, claiming a regulatory function under stress conditions (55–59). However, our finding that natural *H. volcanii* populations in the same sampling location tend to be mixed, having both intein-positive and intein-negative strains sharing the same niche, does not support a strong selective pressure for intein retention in this case.

It is tempting to speculate that the increased recombination rate might provide a benefit to the group, and that possibly the intein is maintained in the population through group selection. This reasoning is reminiscent of a "cost of sex" scenario (e.g.,(60, 61)), where under some conditions, such as stress or environmental change, benefits at the population level, resulting

from the increased genetic variation, outweigh the costs (for example the cost of males in parthenogenetic insect species). However, given that the increased recombination rate provides a direct benefit to the molecular parasite, invoking group selection, *i.e.*, the competition between a group that harbors inteins and groups that do not, seems an unnecessary complication; rather this appears as another case (62, 63) where the gene's selfishness drives the survival of the trait at a cost to the individual. While there is a benefit to the larger group, it is likely the benefit to the molecular parasite, not the benefit to the group that guaranties the survival of the parasite.

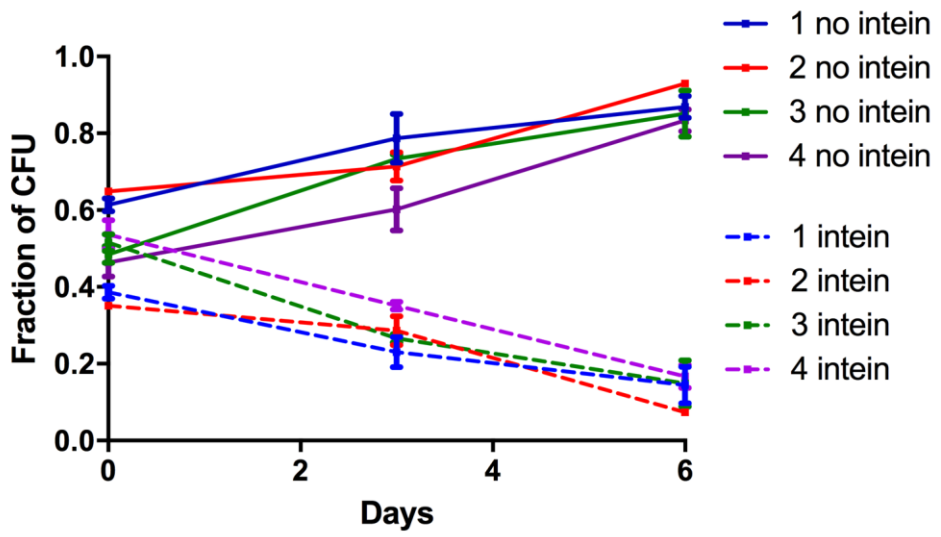


Fig. 1. Relative abundance of intein-containing and intein-free cells in a direct competition assay. *Haloferax volcanii* with and without the *polB*-c intein (otherwise isogenic) were grown in co-culture, aliquots were sampled and the fraction of intein-containing and intein-free colony forming units was determined at different time points. Colors indicate independent parallel experiments.

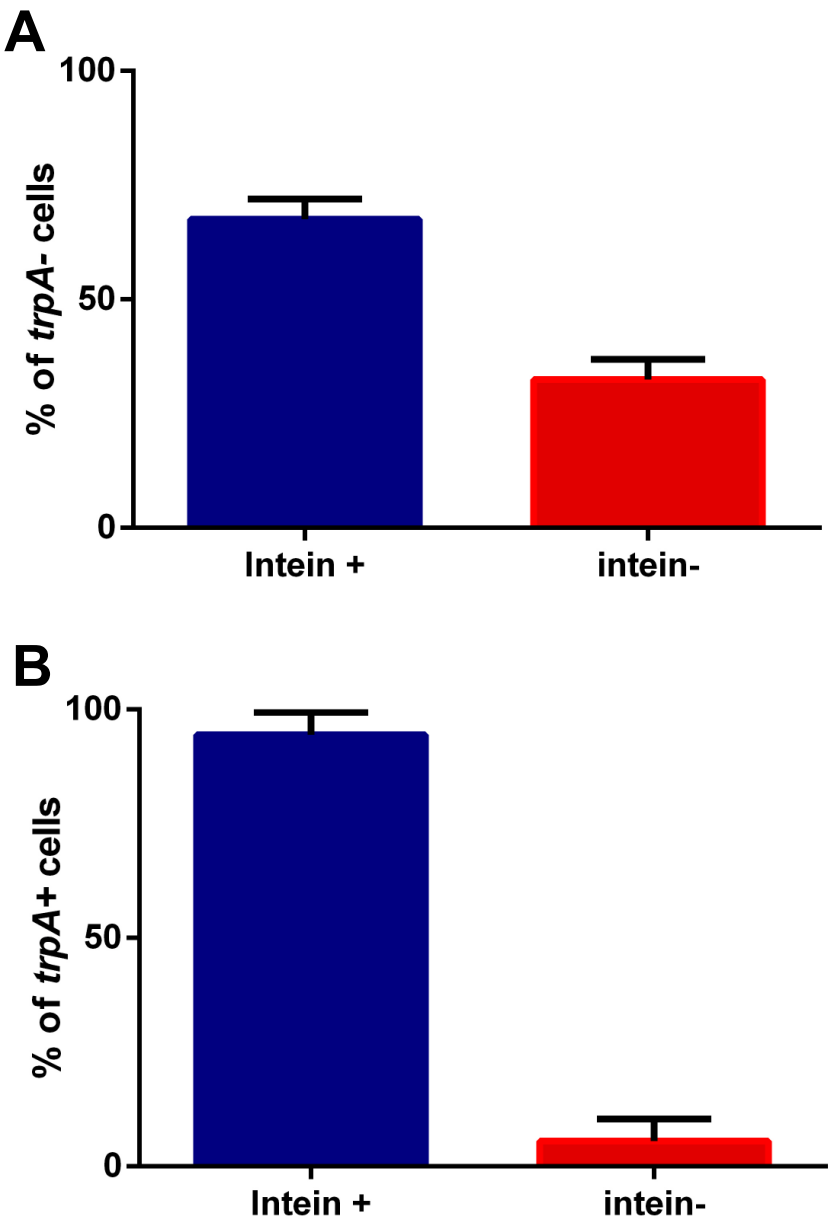


Fig. 2. Percent of intein-containing and intein-free cells following mating between intein-containing and intein-free cells. The percent of cells that had the genotype of the intein-free parent (*trpA*$^{-}$) are depicted in panel A; panel B gives the percent of cells that had the genotype of the intein-containing parent (*trpA*$^{+}$).

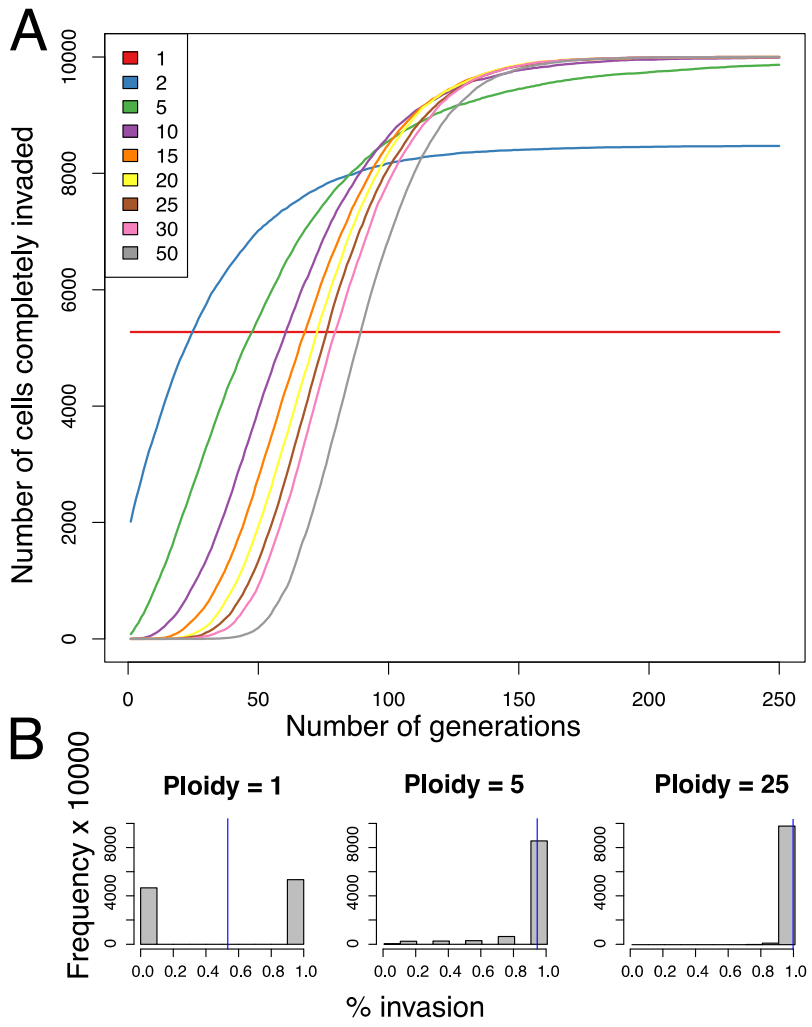


Fig. 3. Simulations to assess the effect of ploidy on intein invasion dynamics after fusion.

Two cells, one completely invaded and one un-invaded, fuse and chromosomes are assumed to assort randomly onto the daughter cells. 10,000 cells are followed for 250 generations. In panel A the number of cells completely invaded after each generation is recorded with the ploidy number as parameter, which is indicated by the colors in the legend. In panel B 10,000 cells are followed over 100 generations and the percent of intein invasion in each cell is recorded, the blue line indicates the mean percent invasion for each ploidy number simulated.

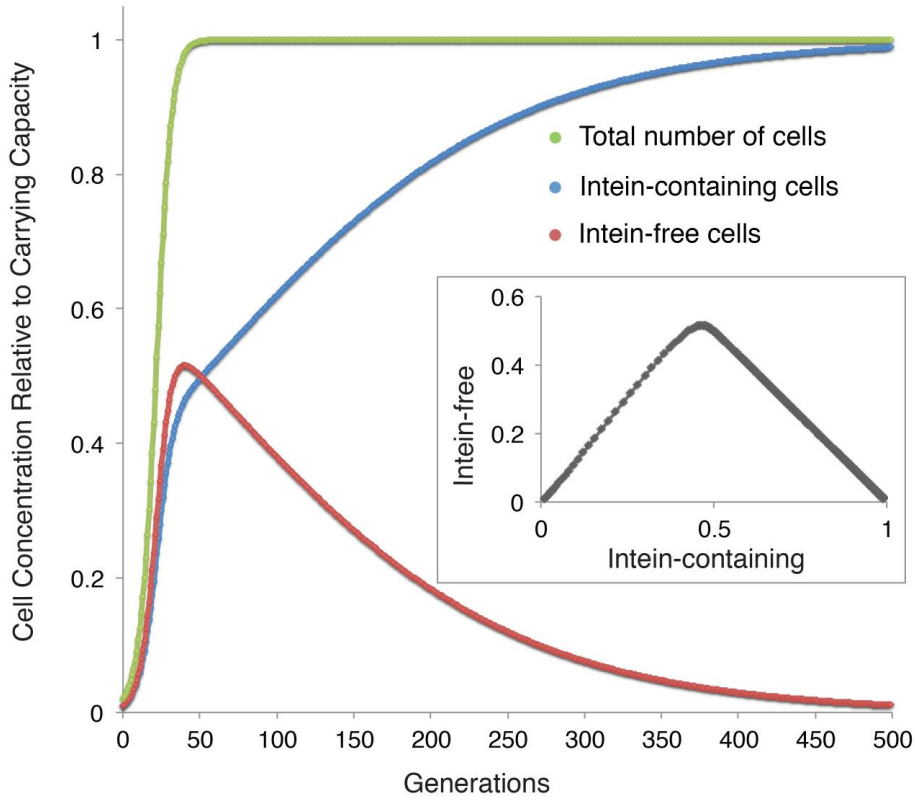


Fig. 4. Simulation of intein invasion dynamics in a population with carrying capacity. Cell number is plotted relative to the carrying capacity over 500 generations. The inset gives concentration of intein-free alleles plotted against the concentration of intein-containing alleles. Note that initially, due to the fitness cost of the intein, the intein-free cells outgrow the intein-containing ones; however, even with the low homing efficiency of 0.01 assumed in this simulation, the intein spreads throughout the population as the growth rate declines closer to carrying capacity. The results remain qualitatively the same even when the carrying capacity is assumed to be equally impacted by the fitness cost (Supplementary Fig. S2). Parameters: homing efficiency=0.01; fitness cost of the intein=0.075; carrying capacity 1; growth rate 0.2 per generation; starting concentration 0.01 each for the intein-containing and intein-free cells.

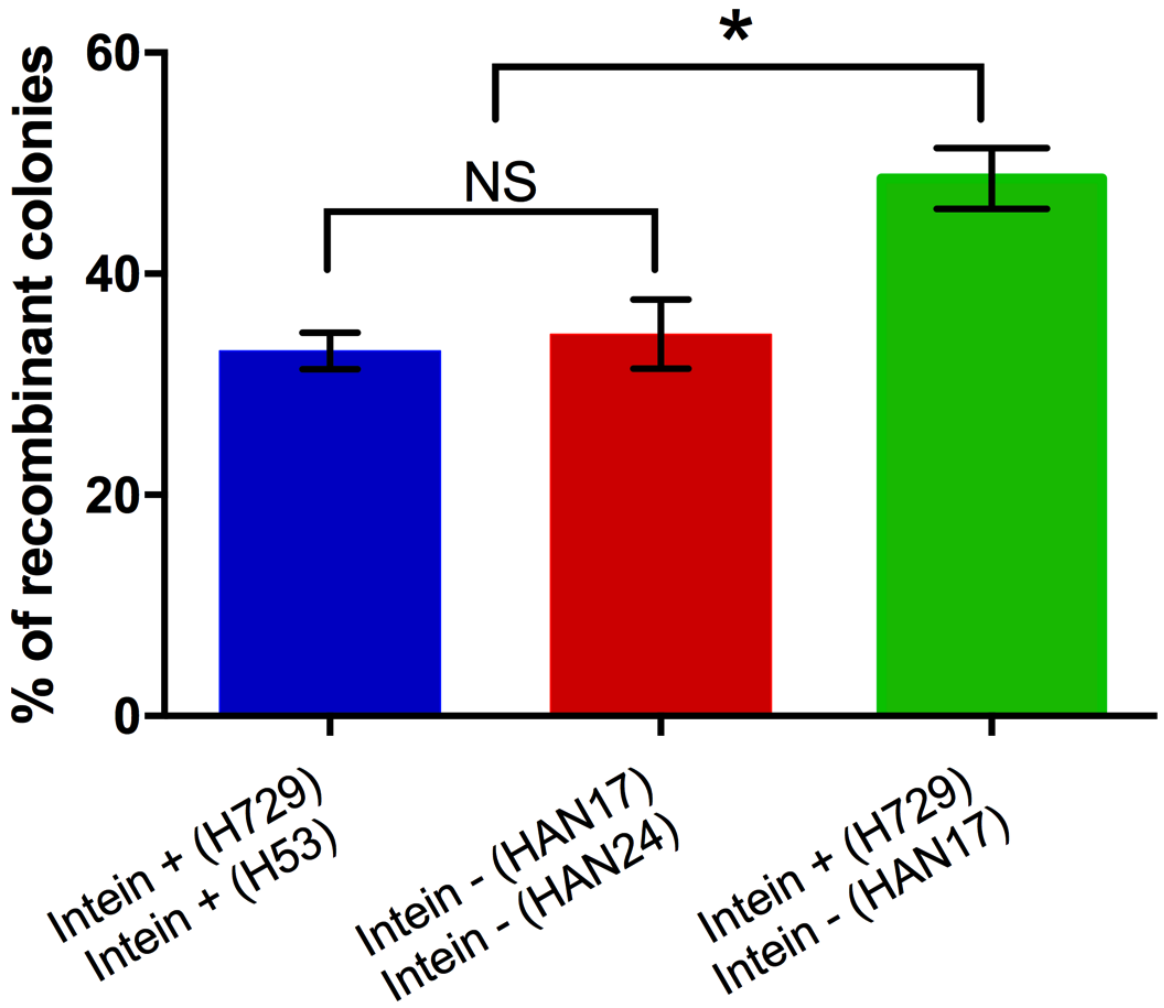


Fig. 5: Recombination frequency in different mating experiments. Intein-free and intein-containing cells were used in mating experiments, and the number of recombinants was determined as described in (33). Recombination rate is higher, when intein-containing cells mate with intein-free cells. Statistical significance by student t test: * P=0.016, "NS" not significant (P=0.86).

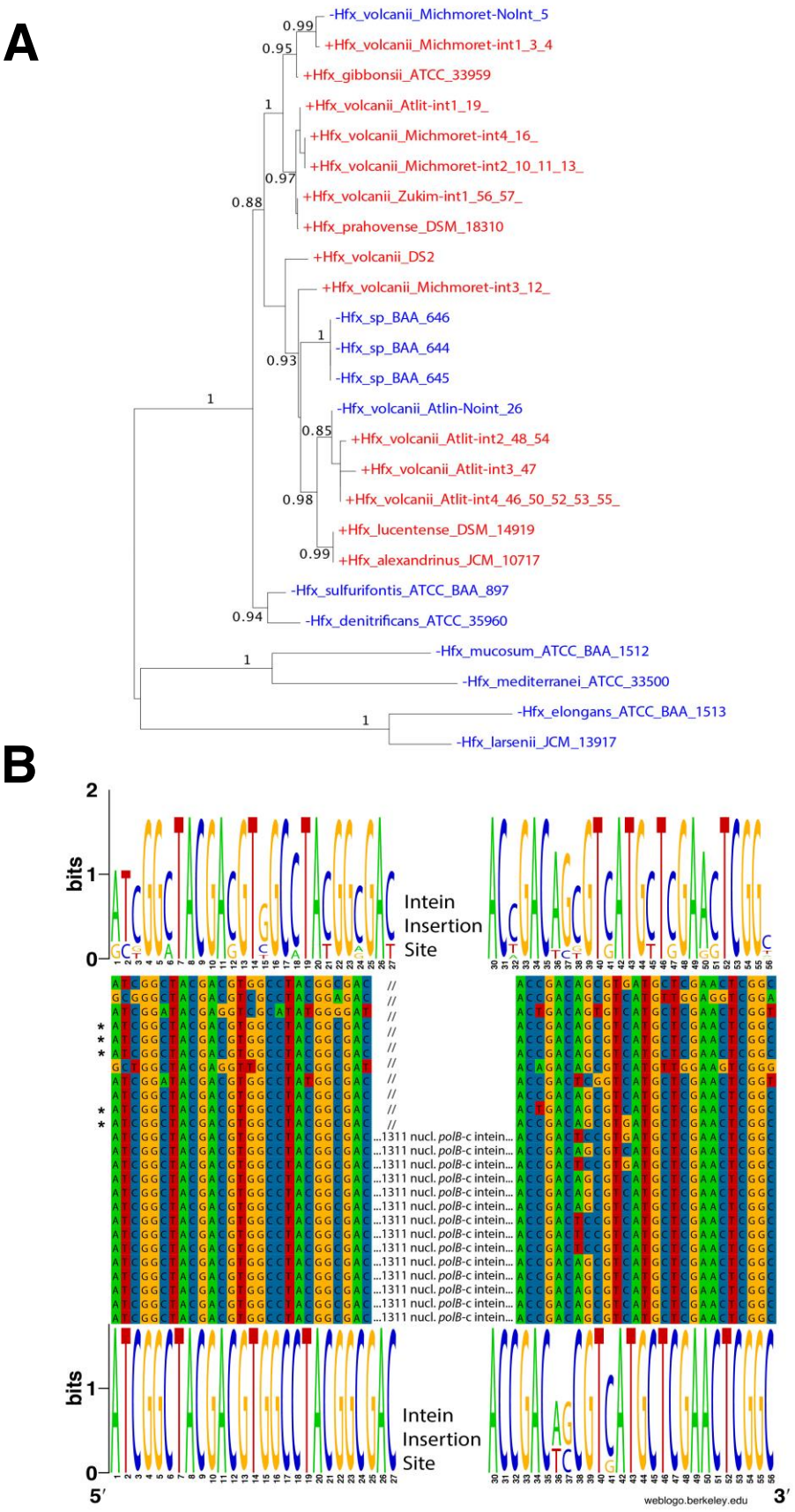

A
-Hfx_volcanii_Michmoret-NoInt_5
+Hfx_volcanii_Michmoret-int1_3_4
+Hfx_gibbonsii_ATCC_33959
+Hfx_volcanii_Atlit-int1_19_
+Hfx_volcanii_Michmoret-int4_16_
+Hfx_volcanii_Michmoret-int2_10_11_13_
+Hfx_volcanii_Zukim-int1_56_57_
+Hfx_prahovense_DSM_18310
+Hfx_volcanii_DS2
+Hfx_volcanii_Michmoret-int3_12_
-Hfx_sp_BAA_646
-Hfx_sp_BAA_644
-Hfx_sp_BAA_645
-Hfx_volcanii_Atlin-Noint_26
+Hfx_volcanii_Atlit-int2_48_54
+Hfx_volcanii_Atlit-int3_47
+Hfx_volcanii_Atlit-int4_46_50_52_53_55_
+Hfx_lucentense_DSM_14919
+Hfx_alexandrinus_JCM_10717
-Hfx_sulfurifontis_ATCC_BAA_897
-Hfx_denitrificans_ATCC_35960
-Hfx_mucosum_ATCC_BAA_1512
-Hfx_mediterranei_ATCC_33500
-Hfx_elongans_ATCC_BAA_1513
-Hfx_larsenii_JCM_13917
0.99
0.95
1
0.97
0.88
1
0.93
1
0.85
0.98
0.99
0.94
1
1
B
bits
Intein
Insertion
Site
...1311 nucl. polB-c intein...
5′
3′
weblogo.berkeley.edu

Fig 6. Maximum likelihood phylogeny for *polB* extein sequences (A) and conservation of *polB*-c intein insertion sites (B). Numbers give support values calculated using the approximate Likelihood Ratio Test as implemented in phyml 3.0 (42). Although drawn as rooted, the tree should be considered unrooted. The finding that sequences without (blue) and with intein (red) do not always form distinct clans (64) reveals that invasion of the *Haloferax* genus with the *polB*-c intein is an ongoing process. Panel B shows a *polB* nucleotide sequence alignment around the intein insertion site c. Web logos (43) give the site conservation for intein minus (top) and intein plus sequences (bottom). The five intein minus sequences that group within the cluster of intein plus sequences are marked with an asterisk. The intein minus sequences show greater nucleotide diversity surrounding the intein insertion site, mainly in synonymous positions -- only two positions at the 5' and close to the 3' end of the alignment represent non-synonymous changes. Homing endonuclease site specificity was shown to tolerate substitutions that result in non-synonymous changes (65), suggesting that none of the depicted *Haloferax* sequences may be immune to intein invasion.

Chapter 7. Intein Epidemiology: Deep Lake, Antarctica.

7.1 Introduction

To date all studies of intein distribution amongst organisms has focused on studies of individual species. The first proposed model of intein propagation and degeneration, the homing cycle, was based on a strain of *Saccharomyces cerevisiae* that contained both intein + and intein – versions of the vacuolar ATPase gene(66). Further analysis revealed that there was a mutation in the homing endonuclease domain of the intein + copy that reduced homing activity by ten-fold (66).

To explain the co-existence of the two alleles the author proposed the homing cycle model of intein evolution (Figure 1). The homing cycle starts with a new un-invaded target acquiring the intein through homing, as long as there are new targets in the population and the intein is exposed to them regularly selection will maintain the function of the homing endonuclease domain. However, over time the homing endonuclease will acquire mutations that affect it's function, as was the case for the VMA-1 intein in *S. cerevisiae*, generating a mini-intein. The mini-intein is not as capable of horizontal transmission and if the homing activity is completely lost the intein can be deleted (66). Based on this model the presence of full-size inteins would indicate that the species had been recently invaded, or was able to frequently invade others in the population. Thus inteins and their association with homing endonuclease domains could be used as an indicator of gene flow.

Recently another model of intein propagation was proposed by Barzel et. al (30), the intransitive fitness model, using *in silico* modeling of a population of inteins (Figure 2). In this model the cycle of transmission and decay is similar, however inteins are maintained long term in populations not through efficient homing as in the homing cycle, but reduced homing efficiency. This model proposes that the long term persistence of inteins, and their homing

endonucleases is a positive correlation between transmission efficiency and host fitness cost. This model also predicts that over long evolutionary periods the intein and homing endonuclease will be maintained at high frequencies in a population through equilibrium, or periodic oscillations through the various states (full-size intein, mini-intein, and no intein).

In order to test these two models we examined the intein distribution, sequence variation, and phylogenies using metagenomic data isolated from Deep Lake, Antarctica (9). Deep Lake is an isolated meromictic (permanently stratified) lake located in a large system of marine derived lakes in the Vestfold Hills of East Antarctica (67). It was formed ~3500 years ago from isostatic rebound, and is the saltiest of the lakes in the region. The biome in Deep Lake is dominated by the Haloarchaea, with 72% of the biomass attributed to just four taxa, listed in order of their abundance: *Halohasta litchfieldiae* tADL, halophilic archaeon DL31, *Halorubrum lacusprofundi,* and *Halobacteria sp.* DL1. Interestingly there is evidence of high rates of inter-genera gene transfer occurring between these community members, and much of the gene transfer seems to be adjacent to or involve mobile genetic elements (transposons, integrases, and plasmids) (9). Additionally, because of the extreme cold temperatures, cell division in this community is limited to a maximum of six times per year (67), implying that most of the evolution in this community is through HGT, rather than high rates of replication as in many prokaryotic species. Previous work by Soucy et. al (15) showed that inteins are highly represented in many different genes, and lineages in the haloarchaea. The availability of completed reference genomes for all four of the most abundant taxa, as well as metagenomic data from several different depths and fractions within the lake make this an ideal system for exploring intein propagation and evolution in the haloarchaea.

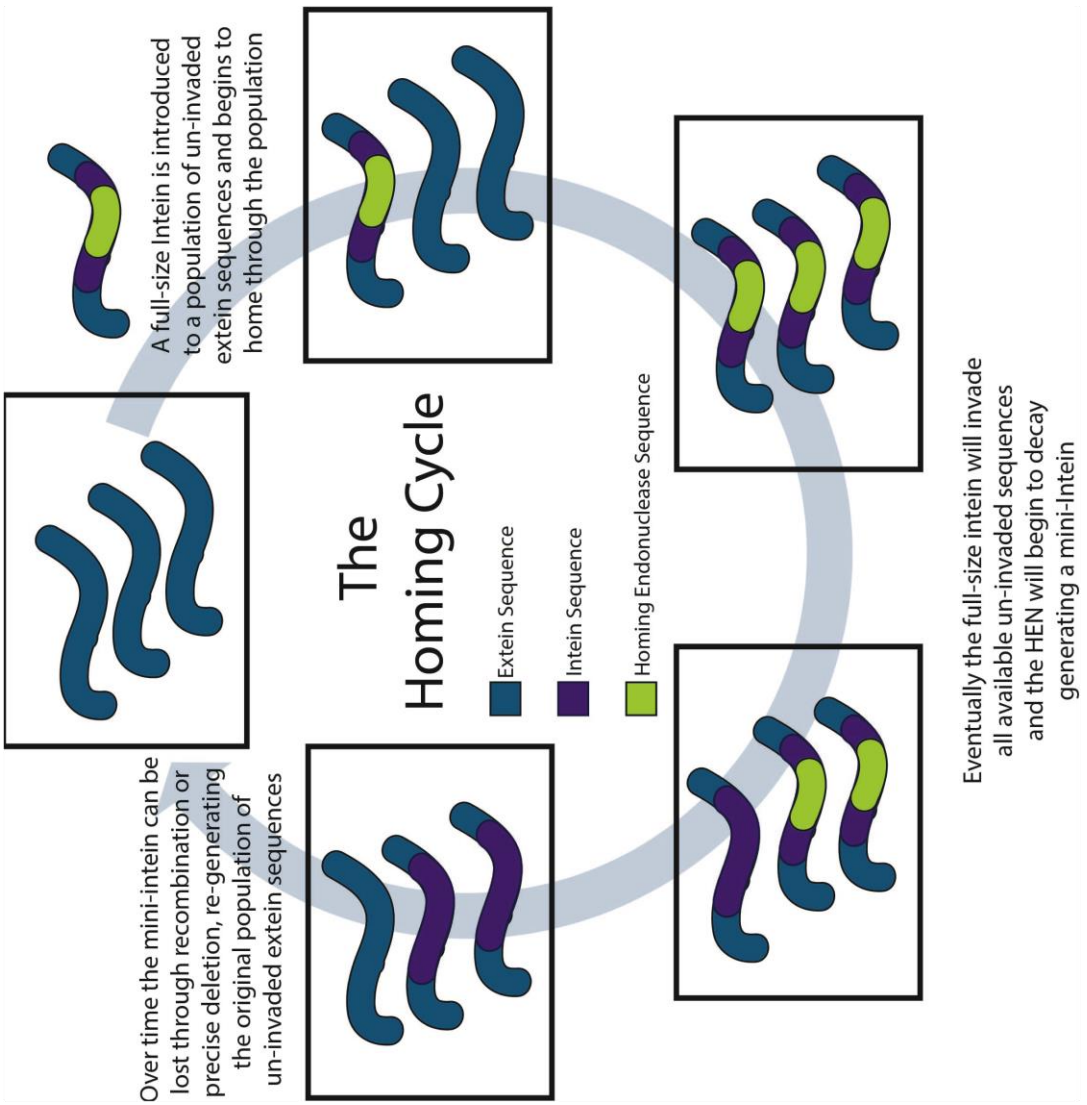
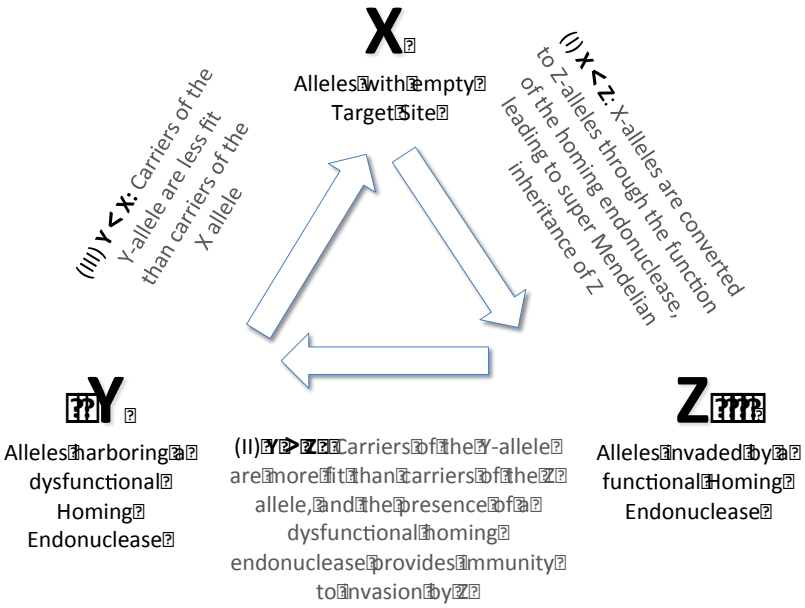


Figure 1 The Homing Cycle. The cycle begins with an un-invaded extein population. A full size intein invades one member of the population and begins exploiting the empty insertion sites in the un-invaded exteins in the population. Eventually all exteins are invaded and selective pressure to maintain the homing endonuclease domain is eventually lost generating a mini-intein. As the mini-intein cannot invade un-invaded inteins eventually the intein can be precisely deleted regenerating the original un-invaded population.



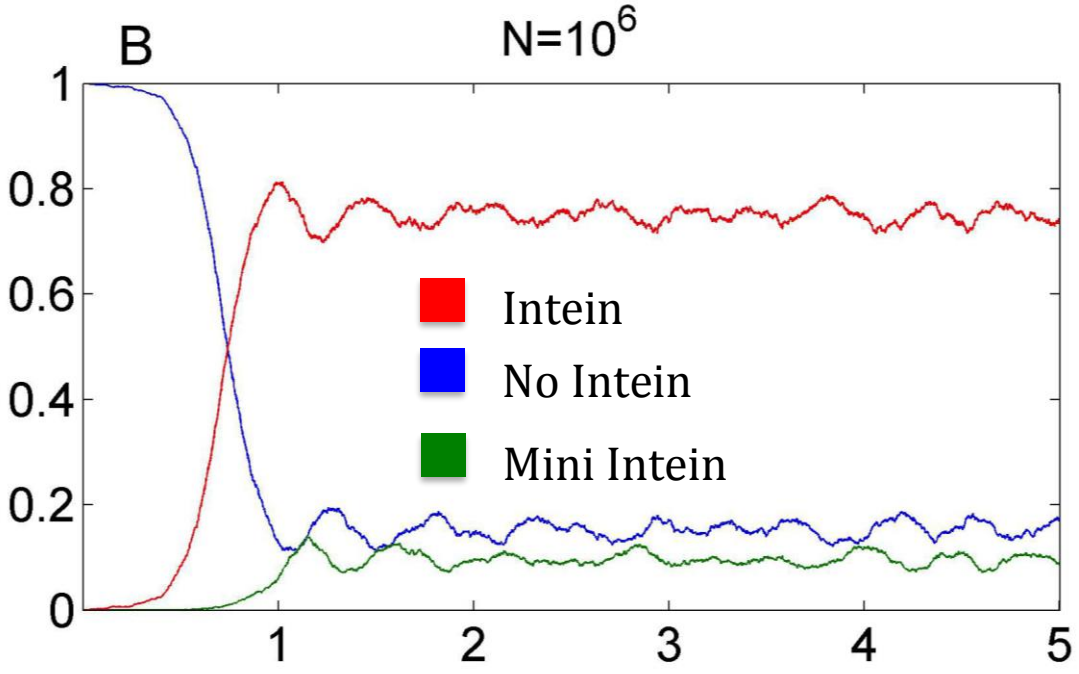


Figure 2 Intransitive Fitness Model. This figure was adapted from Barzel et. al (30) and is based on an *in silico* model of intein invasion. Though the majority of the population has a full size intein, there are small oscillating populations of the other two states (no-intein and mini-inteins) co-existing with the full size intein, each state is more fit than one of the other states, but can also be outcompeted by one of the other states. This model takes into account the fitness cost to the host, unlike the homing cycle.

7.2 Materials and Methods

7.2.1 Reference Genomes

Reference genomes were generated for read mapping later. Intein sequences were retrieved from the reference genomes of *Halorubrum lacusprofundi*, *Halohasta litchfieldiae* tADL, *Halobacteria sp. DL1*, and halophilic archaeon DL31 using the methods outlined in Soucy et. al 2014 (15). All sequences containing inteins from each genome were used for read mapping to reduce the chance of reads mapping to orthologous sequences. Additionally a second reference genome was generated using the intein containing sequences, however the intein was removed from the sequences for the second reference genome, so that reads spanning the insertion site could be recovered. The inteins were removed from the extein sequences using Seaview v4.4.2 alignment viewer(13).

7.2.2 Metagenome Processing and Read Mapping

The metagenomic reads from the 24m 3.0um and 24m 0.8um samples were downloaded from the Joint Genome Institutes Integrated Microbial Genomes and Microbiomes warehouse(68, 69). Reads were trimmed using Sickle v1.33 (70), trimmed reads were mapped onto reference genomes using Bowtie v2.2.5 (71). Bam files were manipulated using samtools v1.2(72) and reads mapped per site was calculated using BEDTools v2.25.0(73). Mapped reads were viewed using Integrated Genome Viewer v2.3.60 (IGV)(74). The number of reads mapped at each site was plotted using RStudio v0.99.489, which was running R v3.0.2(75). Reads mapping across the intein insertion site were aligned and translated using Seaview v4.4.2(13).

7.2.3 Phylogenetic Trees

Phylogenies of inteins, exteins, and a reference dataset using 56 ribosomal proteins as in (45, 15) was generated using phyML v3.0 (76) with an LG model, and a gamma shape parameter. Trees were edited for readability using Figtree v1.4.2 (77).

7.2.4 Sequence Variation Analysis.

In house scripts were used to break the intein sequences of *Hht. litchfieldiae* tADL (*cdc21*-a and *pol-II*-a) and halophilic archaeon DL31 (*cdc21*-a, *polB*-b, *polB*-b, *pol-II*-a, *gyrB*, *rir1*-b, and *top6B*) into 20 base pair segments over a sliding window of 10 base pairs. Also the N-terminal portion of the corresponding extein was subject to the same analysis as a control for each intein sequence. The other two genomes did not have enough coverage to detect variation and thus this work was limited to genomes with sufficient coverage to recover underlying variation in intein sequences. Each 20bp segment that was generated in the earlier step was then used as a seed sequence using GNU grep v2.6.3 program to find reads that contained the seed pattern. Again using in-house perl scripts each match to the mate pair of each seed sequence was also collected. Matching reads were aligned to the parental sequence, and using another in-house script each read was scanned against the reference sequences looking for mismatches. If a mismatch was detected the read was written to a separate file, along with it's mate pair, and the script recorded the sequence position, codon position, and type of mismatch (transition/transversion). Lastly this information was collapsed into a single file for each sample (0.8um and 3.0um) and the number of mismatches and perfect matches per site was calculated. This information was plotted using RStudio v0.99.489 (75).

Structural variants from insertion/deletion events were scanned for using the SVdetect (78) and breakdancer (79) software packages.

7.3 Results

7.3.1 Intein Distribution in Deep Lake Haloarchaea.

There are seven intein alleles in Deep Lake, all seven are carried by halophilic archaeon DL31 (Table 1). *Hht. litchfieldiae* tADL and *Hrr. Lacusprofundi* each had only two inteins, and *Hbt. sp.* DL1 had only one intein. This was surprising, as we had expected the frequent gene exchange to homogenize the community with respect to intein content. Four of twelve intein sequences have no detectable homing endonuclease domain, the *cdc21*-a, *pol-II*-a, and *gyrB* inteins in halophilic archaeon DL31, and the *pol-II*-a intein in *Hrr. lacusprofundi* and are considered mini-inteins in capable of homing into new sites, these inteins are colored yellow in Table 1. This is consistent with previous observations that all three states can be found existing in the same environment (15).

	Cdc21-a	GyrB-b	PolB-b	PolB-c	Pol-II-a	Rir1b-b	Top6B
Halohasta litchfieldiae	**473**				**508**		
Halophilic archaeon DL31	**283**	**433***	**403***	**402***	**186**	**373***	**570***
Halorubrum lacusprofundi					**178**		
***Halobacterium* sp. DL1**	**467**				**507**		

7.3.2 Phylogenetic Comparisons.

To test if any of the inteins had been acquired after Deep Lake was established we created a ribosomal reference tree using 56 ribosomal proteins (Figure 3). The out-group, the closest supported set of sister taxa in the ribosomal phylogeny, was established for each genome from Deep Lake and then incorporated into the phylogenies of the intein and extein sequences. We then use the ribosomal topology as the reference, and compare that with the topology from

each intein, and extein phylogeny. Conflicts with the reference phylogeny could represent HGT events involving inteins. When available genomes from the same genera were used as out-groups (*Halorubrum spp.* and *Halobacteria spp.)*. For *Hht. litchfieldiae* tADL the closest group were the *Halonotius spp.* and for halophilic archaeon DL31 the closest group was another unclassified halophilic archaeon, strain J07HB67 (Figure 3).

All extein phylogenies matched the expected topology based on the reference phylogeny, indicating that any conflicts that arise are because of recombination within the intein sequence (Figure 4). Intein phylogenies could only be built for *pol-II*-a and *cdc21*-a, as they were the only alleles with more than one member. The *pol-II*-a phylogeny was in agreement with the reference topology for most sequences, however *Hbt sp.* DL1 grouped with very high support (b.s. value = 98) outside the *Halohasta/Halonotius* cluster instead of with the other *Halobacteria spp.* (Figure 4). The *cdc21*-a intein of halophilic archaeon DL31 should have grouped adjacent to the *Halohasta/Halonotius* group, but was sister to the *Halobacteria* cluster with moderate support (b.s. value =66) (Figure 4). These conflicts in the topology suggest that there is some gene transfer occurring that involves inteins, however the majority of inteins are in agreement with the expected topology from the ribosomal tree. Inteins that were in a good agreement with the reference topology indicate inteins that most likely were associated with their host genome before this population was established. Inteins that were not in agreement with the reference phylogeny could have been acquired after the population was established, however none of the Deep Lake inteins group adjacent to one another. This indicates that these inteins were probably acquired from a source that is not represented in this sample. This source may be one of the less abundant taxa in Deep Lake, or it could also have been acquired before Deep Lake was

established. These possibilities cannot be distinguished from one another in the scope of this work.

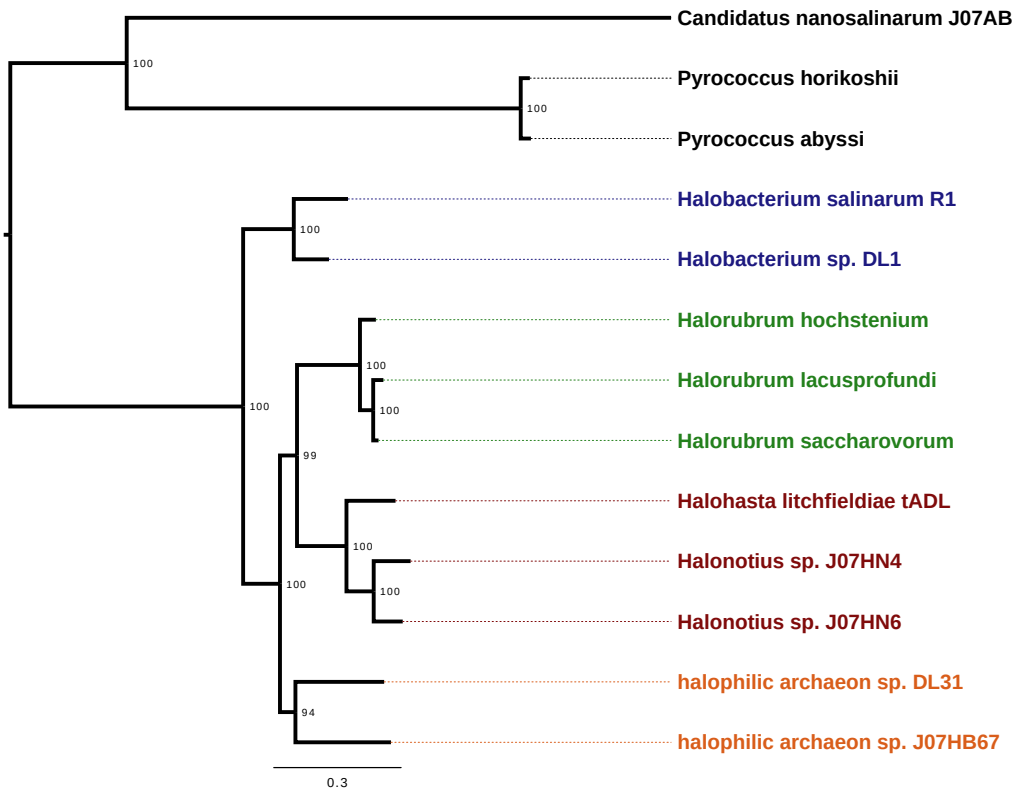


Figure 3 Ribosomal reference phylogeny for Deep Lake taxa with their out-groups. This tree is rooted using the *Pyrococci* as an out-group. Each cluster is indicated by a different color, and each cluster is highly supported.

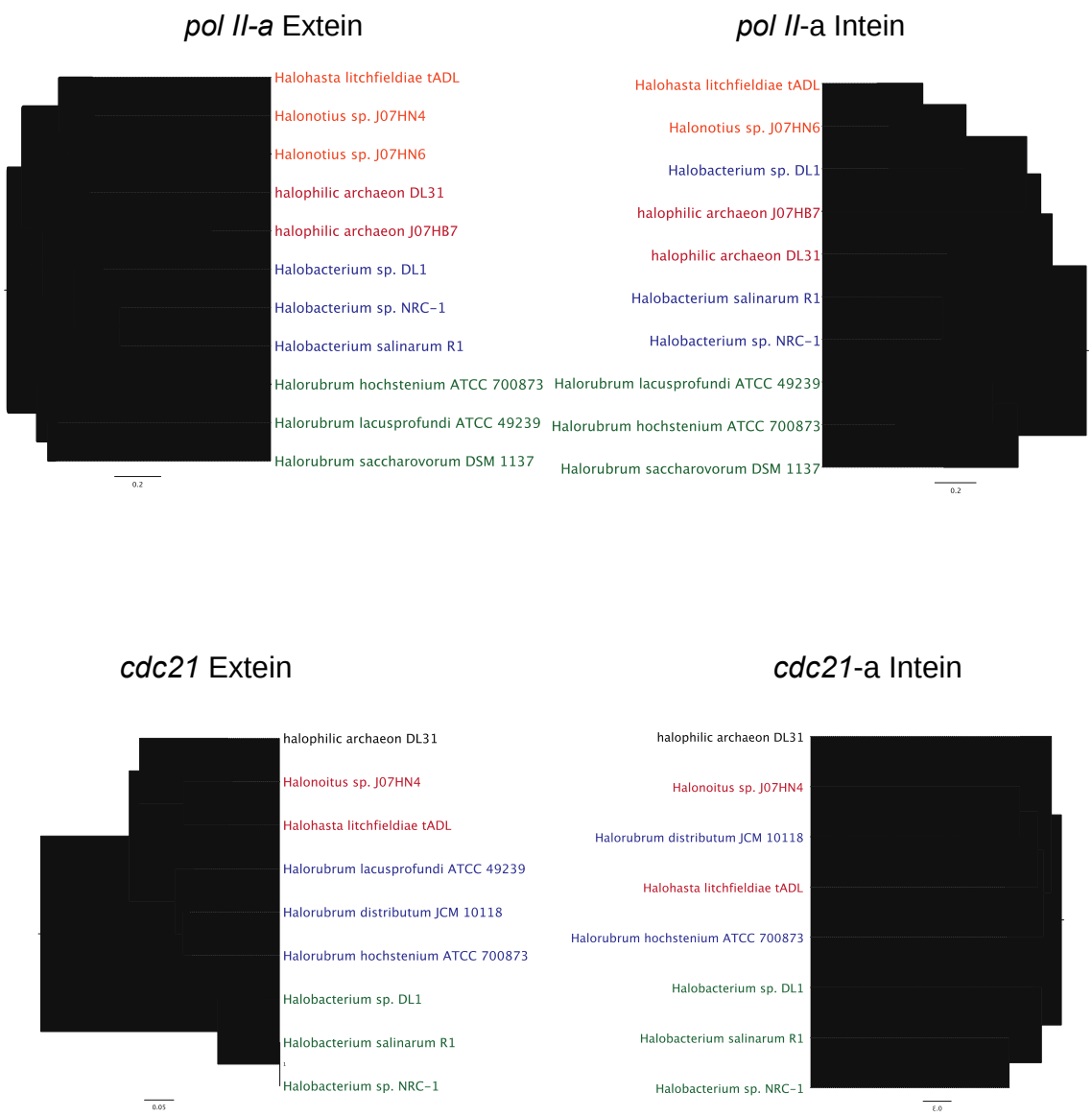


Figure 4 Comparison of Extein and Intein topologies. Clusters of organsims are colored as in the ribosomal reference tree (figure 3). The extein phylogeny is on the left and the intein topology is on thr right. Bootstrap support values are indicated at each node. The *pol-II*-a phylogenies are in the top panel, and the *cdc21*-a phylogenies are in the bottom panel.

7.3.3 Read Mapping.

Metagenome data was used from both the 0.8µm and 3.0µm fractions collected from 24 meters below the surface of Deep Lake (9). These fractions were kept separate during all subsequent analyses to enable comparison of the findings between fractions. Reads were mapped onto two reference genomes, one represented all intein containing genes from the completed genomes, and the other was identical to the first except the inteins were removed, we call this the "Alternate Universe Scenario". The number of reads mapping to each site was compared between the reference genomes, as expected there was a significant drop in the number of reads mapping across the intein insertion for the alternate universe scenario (Figure 5). Reads that mapped across the intein insertion site (IIS) were, collected and aligned to the reference sequence they mapped to, and then translated into a protein sequence. The majority of reads that mapped across the IIS were intein-containing reads that contained more extein than intein, and thus were mapped onto the empty IIS despite the significant sequence differences. For *Hht. litchfieldiae* tADL *pol-II*-a intein fifteen reads that mapped across the IIS were found in the 0.8µm dataset, and twenty reads in the 3.0µm dataset. Also one read was recovered from the 0.8µm dataset mapped across the IIS of the *rir1*-b intein in halophilic archaeon DL31. This data indicates that there are small populations of un-invaded stains co-existing in the population with invaded strains (Figure 6).

To identify the background of the un-invaded reads the mate pairs of reads that mapped across the IIS were collected for each dataset, and aligned to the parental sequence. A sequence fragment of 295 bp for the 0.8µm sample, and 413 bp for the 3.0µm sample was generated using a consensus of the mate pairs aligned to the reference. A phylogeny of these fragments with the corresponding regions of the Deep Lake genome sequences show that the un-invaded organism is most closely related to *Hht. litchfieldiae* tADL (Figure 7). Indicating that there is a sub-

population of *Hht. litchfieldiae* tADL without the intein that co-exists with the intein containing strain. According to the homing cycle the co-existance of intein-free and intein-containing alleles should only be possible if the HEN domain has begun to undergo decay. Except that the majority of inteins in Deep Lake contain a detectable homing endonuclease. The intransitive fitness model supports the co-existence of multiple states of inteins within a single population. The model even predicts that the majority of the populations will carry full-size inteins with smaller sub-populations containing intein-free alleles.

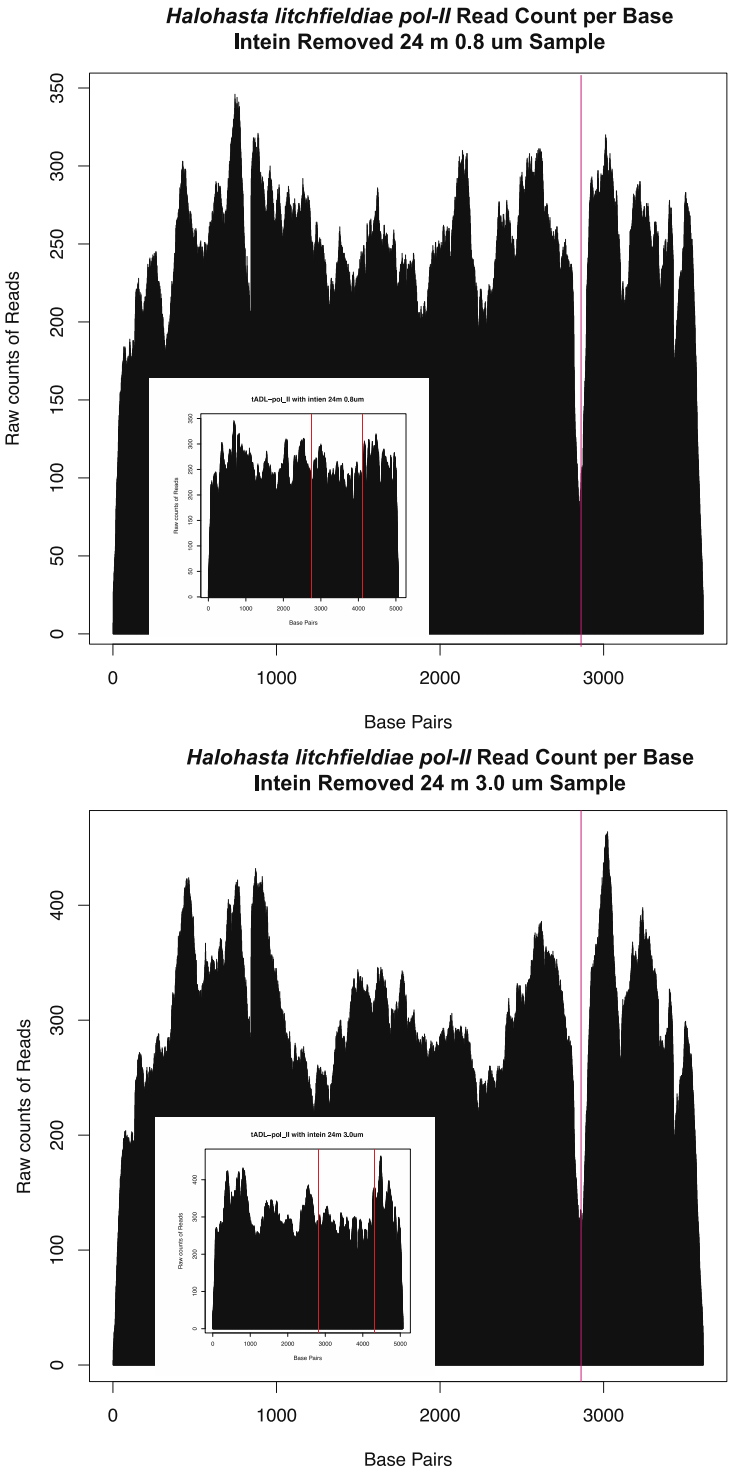


Figure 5 Read counts per base pair for intein-containing and intein-free versions of Deep Lake extein sequences. Read counts for *Halohasta litchfieldiea* tADL *pol-II*-a intein containing sequence is inset inside the read frequency plot for the intein free version. The top panel shows the read counts for the 24m 0.8µm sample, and the bottom panel shows the read counts for the 24 meter 3.0µm sample.

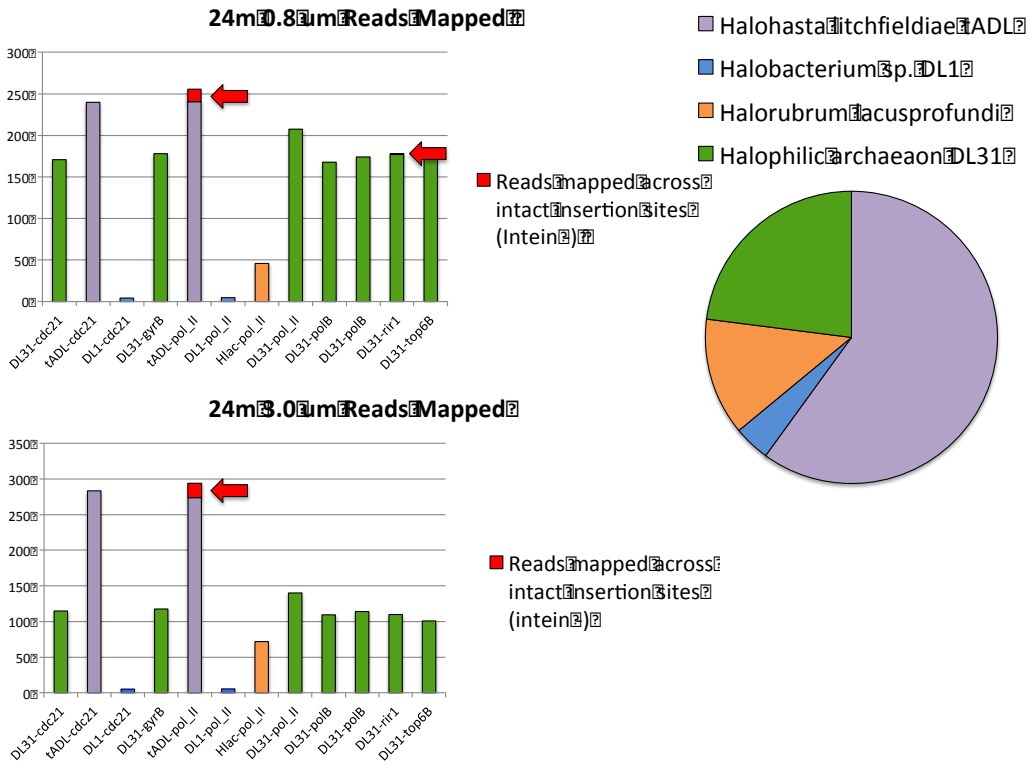


Figure 6 The average number of reads mapped across the intein sequences for each intein in deep lake. In red are the reads that map across the intein insertion site for the corresponding intein allele, and organism. The top bar chart shows read frequencies for the 24-meter 0.8µm sample, and the bottom chart shows the read frequencies for the 24-meter 3.0µm sample. On the right of the figure the abundance of each organism is plotted as a proportion of the reads that map to each of the four genomes, all charts are colored according to the key above the pie chart. The number of reads mapping to each intein sequence correlates with the abundance of each organism in the lake.

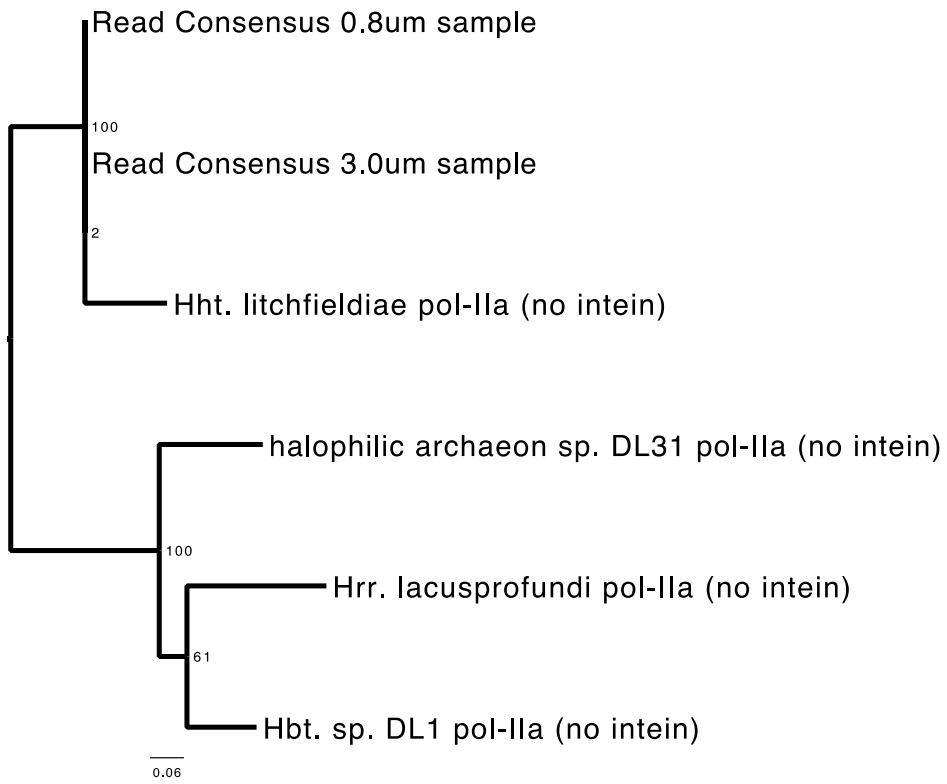


Figure 7 Phylogeny of Deep Lake *pol-II* exteins. This phylogeny was built with only the regions of the Deep Lake exteins that corresponded to the regions that could be built from the mate pairs of reads that map across the intein insertion site. The ininvaded sequences group with high support near *Hht litchfieldiae* tADL indicating these reads probably represent an intein-free subpopulation of *Hht litchfieldiae* tADL.

7.3.4 Sequence Variation Analysis.

In order to look for evidence of HEN decay we created several seed sequences only 20bp long over the entire length of the intein sequence, with a sliding window of ten bases. We used these seed sequences to look for matching patterns in the raw reads, rather than using a read aligner we opted to use a string-matching program to avoid the false positives that were mapped erroneously to the IIS in figure 13. Reads that had a matching pattern were collected, along with their mate pair, and aligned to the parental sequence (where the seeds were generated from). Reads were then analyzed for mismatched bases, if a mismatch was detected the read name, the position of the mismatch, the type of change (transition or transversion), and codon position was noted for each mismatch. The number of mismatches per site was plotted to look for regions that are more prone to decay then others. Surprisingly the majority of mismatches were located between the N-terminal splicing domain and the HEN domain (Figure 8). Also the position in the codon where the mismatch occurs is almost equal at all three sites, and the number of transversions is roughly twice the number of transitions, indicating neutral selection is occurring within the linker regions (Figure 8). This is weak evidence that the HEN is not decaying in these inteins, yet there are un-invaded and invaded species co-existing. More work needs to be done to refine the results of this work, and also compare with other inteins in Deep Lake.

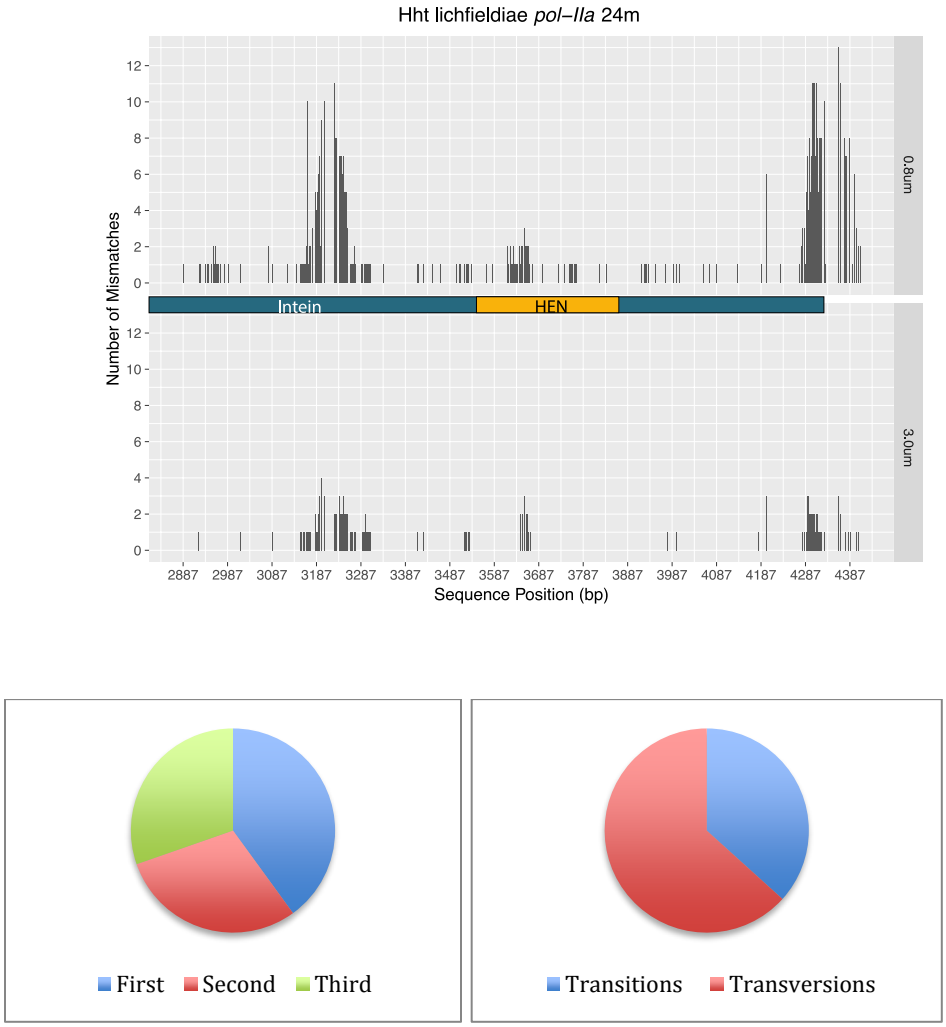


Figure 8 Mismatch frequencies across the *pol-II*-a intein in *Halohasta litchfieliae*. The raw count of mismatched reads at each position is indicated by the height of the columns. The turquoise bar indicates the position of the intein sequence in the alignment, and the gold bar indicates the position of the homing endonuclease domain. The top bar plot represents the 0.8µm sample and the bottom panel represents the 3.0µm sample. The spike at the end of the intein sequence (position 4302) represent the un-invaded reads that were previously recruited in figure 13. The lower pie charts represent the frequency of mismatches in each codon position (left pie chart) and the frequency of transition and transversions (right pie chart) in all mismatches for both samples combined.

An interesting conundrum of intein decay is that any deletions or insertions, which appear to be frequent given the size distribution in most inteins (15), must be in frame. Substitution events that disrupt the codon frame of the intein, will also affect the frame of the c-terminus of the extein as well. We searched for reads that contained gaps when aligned with the reference sequence. We predicted that if we found any gaps they would be in a multiple of three to maintain the frame of downstream sequences. There were four seeds that recruited reads that caused gaps in the reference sequence, represented by only 22 reads (10, 9, 1, and 2 reads per site). Only one of the four gaps occurred in a multiple of three (7, 3, 5, and 11bp). An independent alignment of the reads containing gaps with the reference sequence showed that the seed sequence had forced the gap-causing read onto the alignment, and these reads most likely were erroneously mapped to the reference because of the abundance of perfectly matching reads in the file as well. In order to look for gaps that were larger than could be recovered through seed-recruiting each dataset was analyzed with SVdetect (78) and breakdancer (80), which use the normal distribution of mate-pair insert sizes to identify potential insertion or deletion events, and no events were detected by either software.

7.4 Further Analysis.

More work needs to be done with the seed-recruiting to capture variation. I would like to run this analysis on all seven inteins from halophilic archaeon DL31, and on the *cdc21*-a intein from *Hht. litchfieldiae* tADL. The other two species have coverage too low to hope to find any variation from subspecies existing in the population. I would also like to repeat this analysis on the flanking extein sequence for each intein. We hope to show that substitution events are more frequent in the intein sequence than the extein, and that substitution events are compartmentalized. I would also like to verify that there is a drop in read coverage of similar magnitude for all intein sequences.

Additionally, using the scatterplot analysis in Chapter 6 I would like to identify the genetic background of the intein donors for *Hbt. sp.* DL1 *pol-IIa* intein and the halophilic archaeon DL31 *cdc21*-a intein.

7.5 Outlook.

Though it does look like there is some HGT of inteins in Deep Lake haloarchaea, the majority of inteins (10 of 12) seem to have stably existed within their host since before the lake was established. The donors could not be identified for those inteins that were involved in HGT events, implicating the inteins were donated from organisms that are not represented in this work.

The read mapping data indicates that there are at least two intein states, a full-size intein and no intein, co-existing in a single species within the lake. There could also be a third state and more work with the seed recruiting in other inteins could verify this.

The co-existance of intein states, and the stable co-existance of inteins and their hosts over long periods of time favors the intransitive fitness model of intein evolution. This result is significant because in a population that has a higher rate of recombination compared to replication one would expect that the distribution of inteins would mirror that of a large population of eukaryotic cells, where meiosis is the main method of evolution, but parthenogenesis also occurs at low rates. We did not observe the expected distribution of inteins (complete invasion), instead inteins were mainly concentrated in one organism, halophilic archaeaon DL31. The intein alleles that were present in the other three organisms are also present in most haloarchaea, and most inteins grouped with their respective outgroups indicating long-term association with the host. Furthermore, most inteins had detectable homing endonuclease domains. Though the method of detection cannot verify the ability of the domain to home, the

presence of the domain with no evidence of homing activity between the organisms analyzed here, despite high levels of HGT between these organisms. Taken together this work suggests that the homing endonuclease domain is maintained even with low rates of transmission, supporting the intransitive fitness model.

Chapter 8. Intein Epidemiology: Israeli Rock Pools.

8.1 Introduction.

This work was conducted in collaboration with the Gophna lab. Adit Naor and Neta Altman-Price carried out the bench work and environmental sampling, and I performed the computational analysis. Preliminary data from chapter 6 indicated that the pools on the rocky coast of the Mediterranean Sea contained populations of haloarchaea, more specifically populations of *Haloferax volcanii.* Most of the isolates from these pools contained the *polB*-c intein, as expected based on the reference genome *Haloferax volcanii* DS2, but some strains from each location did not have the *polB*-c intein. This evidence weakly supports the intransitive fitness network where intein-containing and intein-free strains can co-exist. We were interested in learning more about the community structure, the taxonomic diversity in the pools, and the distribution of inteins in the environment. We were looking for additional evidence that there were intein+/- strains co-existing within a single pool. In 2012 Adit Naor collected several environmental isolates from three locations along the Mediterranean coastline. In 2013 Neta Altman-Price visited one of those locations and isolated several more strains. This work clusters the environmental isolates by location, and then determines the closest phylogenetic neighbor (out-group), based on a ribosomal reference tree, and also determines the intein content within each location. The genomes analyzed in this work were isolated from two locations; there are five genomes from Michmoret and fifteen genomes from Atlit. The rock pools that were sampled are located adjacent to the shore; waves crashing over the crevices in these rocks collect seawater, which then evaporates increasing the salinity in the pools. In the heat of the summer accelerated evaporation causes a sheet of salt to form on the top of the pools, and Haloarchaeal blooms beneath the salt sheet give the pool a pinkish orange color. Samples were collected from two pools at Michmoret, and ten pools at Atlit (Figure 1). These communities are of particular

interest as the waves could serve as a distribution mechanism for the haloarchaea between the two sites.

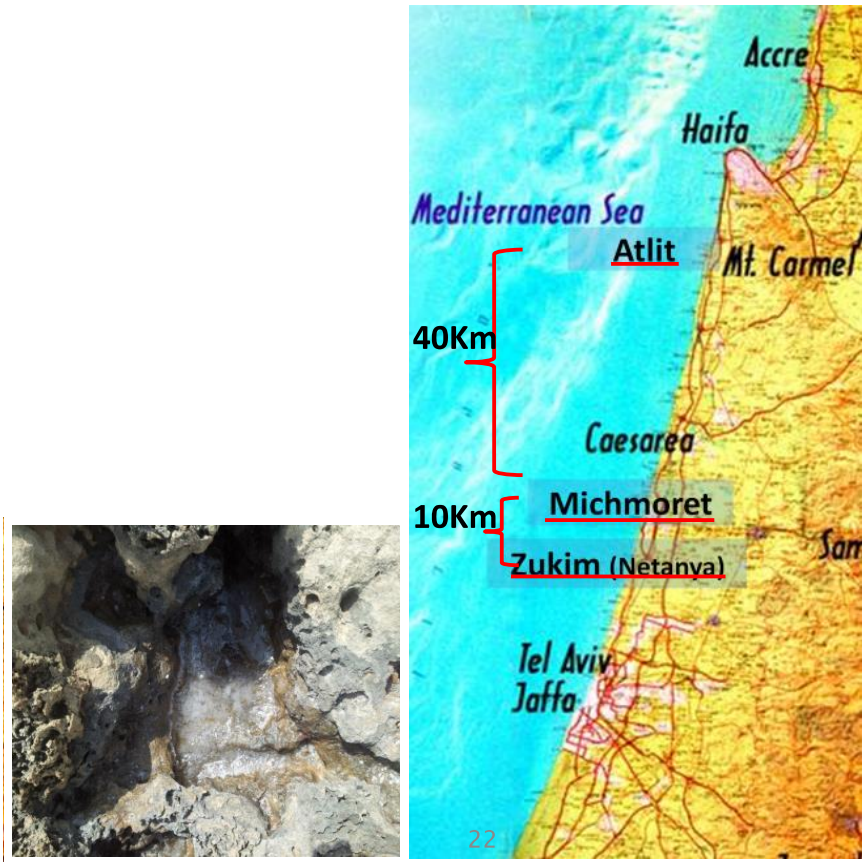


Figure 1 Map and image of sampling sites. On the left is an image of the crystallized sheet that forms on top of the rock pools in August, on the right is a map showing the distance between three sampling sites, the isolates from the third site were not sequenced, and thus were not considered in this work.

8.2 Materials and Methods

8.2.1 Genome Assembly

Reads were processed using cutadapt v1.9.1 to trim low quality bases ($Q_{score} \geq 20$ $p \leq 0.01$) and assembled using SPAdes v3.8 (81). Quality of the assembly was checked with QUAST v4.1 (82) and the assembly was annotated using prokka v1.9 (83).

8.2.2 Intein Retrieval

Each genome was searched for inteins using the same methods as in Soucy et. al 2014 (15). Exteins of each genome were aligned with the data set from Soucy et. al, and aligned using Muscle v3.8.31 (12) to look for the presence of an intein. Intein sequences were extracted and aligned to the intein datasets from Soucy et. al.

8.2.3 Phylogenetic Trees

Maximum likelihood trees were generated for each intein, extein, and ribosomal protein using RAxML v8.1.17 (14) with the GTR model, optimized substitution rates, and a gamma model of rate heterogeneity.

8.3.3 Genome Comparisons

Each genome was compared to all other genomes isolated in the same environment, as well as the closest phylogenetic neighbor from the reference tree using *in silico* DNA-DNA hybridization with the Genome-to –Genome distance calculator (84).

8.3 Results.

8.3.1 Taxonomic distribution in Israeli Rock Pools.

Initially all strains collected in 2012 form both Michmoret and Atlit were reported as strains of *Haloferax volcanii* based on 16s analysis using PCR. Later analysis of these strains

using a concatenation of 56 ribosomal proteins showed that none of these strains were most closely related to *Hfx volcanii* (Figure 2). Strains 10N, 16N, and 19N were most closely related to *Haloferax prahovense*. Strains 24N, 48N, and 47N were most closely related to *Haloferax sp.* BAB2207. Strains 4N and 6N were most closely related to *Haloferax gibbonsii*, and lastly strain 12N was most closely related to *Haloferax denitrificans*.

The 2012 samples from Michmoret were collected from two pools, strain 4N was isolated from one pool and strains 6N, 10N, 12N, and 16N were all isolated from a single rock pool sample. The Atlit strains were isolated in 2012 from three different pools 19N and 24N were each isolated from individual pools, and strains 47N and 48N were isolated from the same pool.

Strains collected in 2013 were more diverse and based on the 16s data the initial assignments showed *Haloferax* species (strains 109R, 105R) isolated as well as species of *Haloarcula* (strains 120R, 7R, and 47R), *Halorubrum* (strains 8R, 9R, 26R, and 28R), and *Halobellus* (strains 31R, and 38R). The ribosomal reference tree showed that strains 109R and 105R are most closely related to *Haloferax sp.* BAB2207, strain 19N was most closely related to *Haloferax prahovense* (Figure 2). The two strains thought to be *Halobellus* were most closely related to *Halogeometricum borinquense* and *Halosarcina pallida*. All of the *Halorubrum* strains were most closely related to *Halorubrum sp.* SP7 and all of the *Haloarcula* strains were most closely related to *Haloarcula vallismortis*.

The 2013 samples were all collected from Atlit, from seven different pools. Strains 109R, 31R, and 28R were all isolated from the same pool, and 8R, 9R, 26R, 38R, 47R, 105R, and 120R were all isolated from separate pools.

With the closest sister taxa from the ribosomal tree, we then performed an *in silico* DNA DNA hybridization (isDDH) between the newly sequenced strains and their closest branching

sister taxa (the out-group). Briefly, isDDH uses the identity between high scoring pairs, and divides the average identity by the length of the high scoring pairs to generate a percent identity between organisms, previous work established <70% isDDH as the species cut off and <80% isDDH as the sub-species limit (54). Several of the taxa showed high identity with their out-group from the ribosomal tree, strains 19N, 10N, and 16N were all sub-species of *Haloferax prahovense* with isDDH values ranging from 86-99.6%. Strains 24N, 109R, 105R, 48N, and 47N were all sub-species of *Haloferax sp.* BAB2207 with isDDH values from 92-95%. Strains 4N and 6N were subspecies of *Haloferax gibbonsii* with isDDH values ranging from 90.8-94.1%. Strains 8R and 9R just barely cleared the species delineation with *Halorubrum sp.* SP7, at 60.8% isDDH. All other strains are considered newly discovered species based on the isDDH standards (Figure 3).

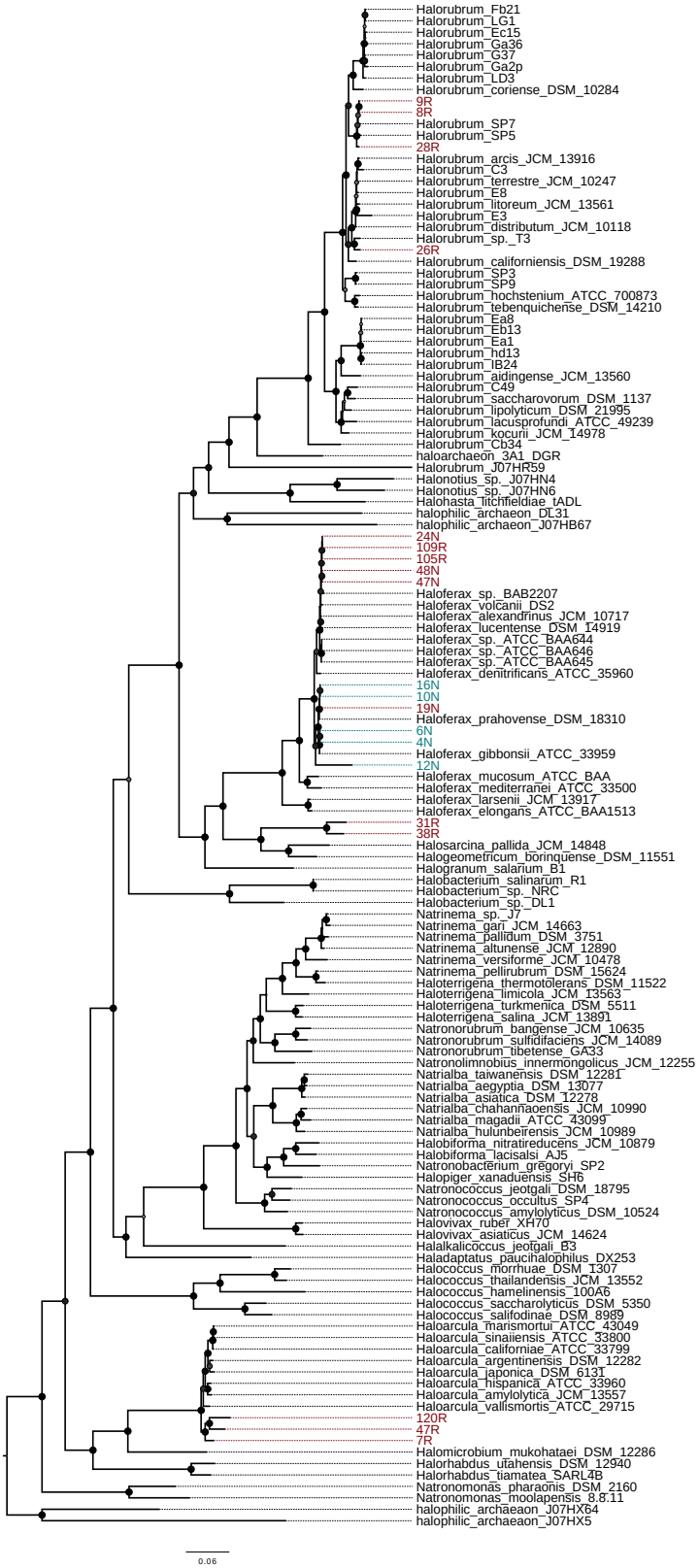


Figure 2 Maximum likelihood phylogeny of 56 ribosomal proteins concatenated. Atlit strains are highlighted in red text, and Michmoret strains are highlighted in blue text. Support values are indicated by the size and color of the node shapes, larger and darker circles indicate higher support.

	10N	16N	Hfx prahovense	4N	6N	Hfx gibbonsii	12N	Hfx denitrificans
10N	100%	99.60%	86.60%	71.00%	70.50%	69.50%	53.60%	48.30%
16N	99.60%	100%	86%	71.10%	70.60%	69.50%	53.70%	48.30%
Hfx prahovense	86.60%	86%	100%	68.40%	68.70%	69.50%	53.90%	30%
4N	71.00%	71.10%	68.40%	100%	90.80%	94.10%	54.20%	49.50%
6N	70.50%	70.60%	68.70%	90.80%	100%	91%	53.80%	49.30%
Hfx gibbonsii	69.50%	69.50%	69.50%	94.10%	91%	100%	55.10%	49.20%
12N	53.60%	53.70%	53.90%	54.20%	53.80%	49.30%	100%	55.10%
Hfx denitrificans	48.30%	48.30%	30%	49.50%	49.30%	49.20%	49.30%	100%

Figure 3 *in silico* DNA DNA hybridization scores between environmental strains and the closest sister taxa from the ribosomal tree (Figure 17). Strain names in green fall below the species threshold and are considered newly discovered organisms. Boxes in yellow indicate the threshold for species delineation, and boxes in red indicate the threshold for sub-species delineation. The above this legend are organisms isolated from Michmoret and on the next page is the data from Atlit

Reference ->	24N	109R	105R	48N	47N	Hfx sp. BAB2207	19N	Hfx prahovense	38R	31R	Hgm borinquense	9R	8R	28R	26R	Hrr sp. SP7	120R	7R	47R	Har vallismortis
24N	100%	100%	100%	94%	94%	92%	46%	46%	23%	22%	21%	22%	21%	22%	15%	21%	21%	21%	21%	19%
109R	100%	100%	100%	94%	94%	95%	46%	46%	23%	22%	21%	22%	21%	22%	22%	21%	21%	21%	21%	19%
105R	100%	100%	100%	94%	94%	95%	46%	46%	23%	22%	21%	22%	21%	22%	22%	21%	21%	21%	21%	19%
48N	94%	94%	94%	100%	96%	92%	46%	46%	23%	22%	20%	21%	21%	22%	22%	21%	21%	20%	20%	19%
47N	94%	94%	94%	96%	100%	90%	46%	45%	23%	22%	21%	21%	21%	22%	22%	21%	22%	21%	21%	19%
Hfx sp. BAB2207	92%	95%	95%	92%	90%	100%	46%	0%	29%	26%	0%	22%	22%	22%	22%	22%	24%	19%	23%	0%
19N	46%	46%	46%	46%	46%	46%	100%	94%	22%	21%	21%	21%	21%	21%	21%	21%	20%	20%	20%	19%
Hfx prahovense	46%	46%	46%	46%	45%	0%	94%	100%	22%	21%	21%	20%	20%	20%	21%	19%	20%	19%	20%	18%
38R	23%	23%	23%	23%	23%	29%	22%	22%	100%	29%	22%	27%	27%	25%	24%	25%	21%	20%	21%	21%
31R	22%	22%	22%	22%	22%	26%	21%	21%	29%	100%	22%	21%	21%	22%	21%	22%	21%	21%	21%	20%
Hgm borinquense	21%	21%	21%	20%	21%	0%	21%	21%	22%	22%	100%	20%	20%	20%	21%	20%	20%	19%	20%	0%
9R	22%	22%	22%	21%	21%	22%	21%	20%	27%	21%	20%	100%	71%	61%	38%	70%	24%	22%	23%	20%
8R	21%	21%	21%	21%	21%	22%	21%	20%	27%	21%	20%	71%	100%	61%	38%	70%	22%	22%	24%	21%
28R	22%	22%	22%	22%	22%	22%	21%	20%	25%	22%	20%	61%	61%	100%	43%	60%	21%	21%	21%	20%
26R	22%	22%	22%	22%	22%	22%	21%	21%	24%	21%	21%	38%	38%	37%	100%	38%	22%	22%	22%	20%
Hrr sp. SP7	21%	21%	21%	21%	21%	22%	21%	19%	25%	22%	20%	70%	70%	60%	38%	100%	20%	20%	20%	21%
120R	21%	21%	21%	21%	22%	24%	20%	20%	21%	21%	20%	24%	22%	21%	22%	20%	100%	83%	94%	37%
7R	21%	21%	21%	20%	21%	19%	20%	19%	21%	21%	19%	23%	24%	21%	22%	20%	83%	100%	83%	37%
47R	21%	21%	21%	20%	21%	23%	20%	20%	20%	21%	20%	22%	22%	21%	22%	20%	94%	83%	100%	37%
Har vallismortis	19%	19%	19%	19%	19%	0%	19%	18%	21%	20%	0%	20%	21%	20%	20%	21%	37%	37%	37%	100%

8.3.2 Intein distribution in the Rock Pools

For all strains that are closely related to their out-group in the isDDH analysis, we compared the intein distribution of the environmental strains to the intein distribution of their out-group (Figure 4). Between Atlit and Michmoret there were seven intein alleles detected, *cdc21*-a, *cdc21*-b, *cdc21*-c, *polB*-b, *polB*-c, *pol-II*-a (both mini and full size), and *rpolA*. Within the same pool strains 10N, 12N, and 16N share the *polB*-c intein, and strains 31R and 28R share the *pol-II*-a intein. These data further support the idea that intein-free and intein-containing alleles can co-exist within a population. We also looked at the distribution of inteins within closely related organisms. Strain 19N had the same intein distribution as it's out-group, but 10N and 16N which share an out-group with 19N were missing the *pol-II*-a intein. Interestingly, strains 24N, 109R, 105R, 48N, and 47N are all highly related to each other and share an out-group- but strains 48N and 47N had a *polB*-c intein and all other strains had no inteins, while their out-group had the *cdc21*-a intein. Strains 6N had no inteins despite being closely related to strain 4N, which has an intein in *polB*-c, as does the out- group for these two strains. Strain 12N is a new species and has an intein in position *cdc21*-a and *polB*-c. Strains 8R and 9R both have the *cdc21*-a intein, but their out-group *Halorubrum sp.* SP7 has the *cdc21*-b intein as well as inteins in *polB*-c and *pol-II*-a. Strain 26R was the only organisms that had the *rpolA* intein, and it also had an intein in *pol-II*-a and *polB*-b. Strain 28R had an intein in the *pol-II*-a position only. All of the *Haloarcula* strains shared inteins in *cdc21*-a and *pol-II*-a.

Strain	year	Region	Conutry	cdc21	dtd	gyrB	helicase	ligase	polB	pol-II	rfc	rir	rpolA	top6B	topA	udp
24N	2012	Atlit Beach	Israel	no	no	no	no	no	no	no	no	no	no	no	no	no
109R	2013	Atlit Beach	Israel	no	no	no	no	no	no	no	no	no	no	no	no	no
105R	2013	Atlit Beach	Israel	no	no	no	no	no	no	no	no	no	no	no	no	no
48N	2012	Atlit Beach	Israel	no	no	no	no	no	**c**	no	no	no	no	no	no	no
47N	2012	Atlit Beach	Israel	no	no	no	no	no	**c**	no	no	no	no	no	no	no
Hfx sp. BAB2207	2012	Gujarat	India	**a**	no	no	no	no	no	no	no	no	no	no	no	no
19N	2012	Atlit Beach	Israel	no	no	no	no	no	**c**	**a**	no	no	no	no	no	no
Hfx prahovense	2007	Telega Lake	Romania	no	no	no	no	no	**c**	**a**	no	no	no	no	no	no
31R	2013	Atlit Beach	Israel	**a,b**	no	no	no	no	no	**a**	no	no	no	no	no	n/a
38R	2013	Atlit Beach	Israel	**c**	no	no	no	no	no	**a**	no	no	no	no	no	n/a
8R	2013	Atlit Beach	Israel	**a**	no	no	no	no	no	no	no	no	no	no	no	no
9R	2013	Atlit Beach	Israel	**a**	no	no	no	no	no	no	no	no	no	no	no	no
Hrr sp. SP7	N/A	Santa Pola	Spain	**b**	no	no	no	no	**c**	**a**	no	no	no	no	no	no
26R	2013	Atlit Beach	Israel	no	no	no	no	no	**b**	**a**	no	no	**a**	no	no	no
28R	2013	Atlit Beach	Israel	no	no	no	no	no	no	**a**	no	no	no	no	no	no
7R	2013	Atlit Beach	Israel	**a**	no	no	no	no	no	**a**	no	no	no	no	no	no
47R	2013	Atlit Beach	Israel	**a**	no	no	no	no	no	**a**	no	no	no	no	no	no
120R	2013	Atlit Beach	Israel	**a**	no	no	no	no	no	**a**	no	no	no	no	no	no

Strain	year	Region	Conutry	cdc21	dtd	gyrB	helicase	ligase	polB	pol-II	rfc	rir	rpolA	top6B	topA	udp
10N	2012	Michmoret	Israel	no	no	no	no	no	**c**	no	no	no	no	no	no	no
16N	2012	Michmoret	Israel	no	no	no	no	no	**c**	no	no	no	no	no	no	no
Hfx prahovense	2007	Telega Lake	Romania	no	no	no	no	no	**c**	**a**	no	no	no	no	no	no
4N	2012	Michmoret	Israel	no	no	no	no	no	**c**	no	no	no	no	no	no	no
6N	2012	Michmoret	Israel	no	no	no	no	no	no	no	no	no	no	no	no	no
Hfx gibbonsii	2012	Alicante	Spain	no	no	no	no	no	**c**	**a**	no	no	no	no	no	no
12N	2012	Michmoret	Israel	**a**	no	no	no	no	**c**	no	no	no	no	no	no	no

Figure 4 Intein presence and absence in environmental strains and their out-groups. A letter in the box under the intein allele name indicates inteins present in each genome, a blue background color indicates full-size inteins, and a yellow box indicates a mini-intein. Strain names in green indicate newly discovered species.

8.3.3 Intein, Extein, and ribosomal comparisons in Israeli Rock Pools.

To determine intein dynamics in the pools we added the intein sequences from these samples to the intein alignments from chapter 5 to determine if the out-group in the extein and intein trees was in agreement with the ribosomal analysis. The extein trees are still running and this section will be updated before the final version of this document has been submitted. For most inteins: polB-c, polB-b, cdc21-c, and cdc21-b all strains grouped as expected, and the extein and intein trees were in good agreement. However, the pol-II-a and cdc21-a phylogenies did not match the topology of the reference or the corresponding extein sequence. In the pol-II-a phylogeny strains 31R and 38R group with *Haloferax elongans* in both the intein and extein trees, rather than it's usual out-group of *Halogeometricum borinquense* (Figure 5 & 6). Also strain 26R groups with *Halorubrum sp.* T3 in the extein tree (Figure 6), and groups outside strain 28R, *Halorubrum sp.* SP3 and *Halorubrum sp.* SP9 in the intein tree (Figure 5). For the cdc21-a phylogeny all strains that have the cdc21-a intein are from Atlit, and they all cluster to the exclusion of all other haloarchaea (Figure 7), this is in contrast to the cdc21 extein tree, where all strains are grouping where expected based on the reference phylogeny (Figure 8). Strain 31R is the deepest branching in the group indicating that the intein may have been passed to the other strains through a relative of this strain. This is strong evidence for intein homing among the rock pools at Atlit beach. Interestingly none of the strains that share the intein were isolated from the same pond.

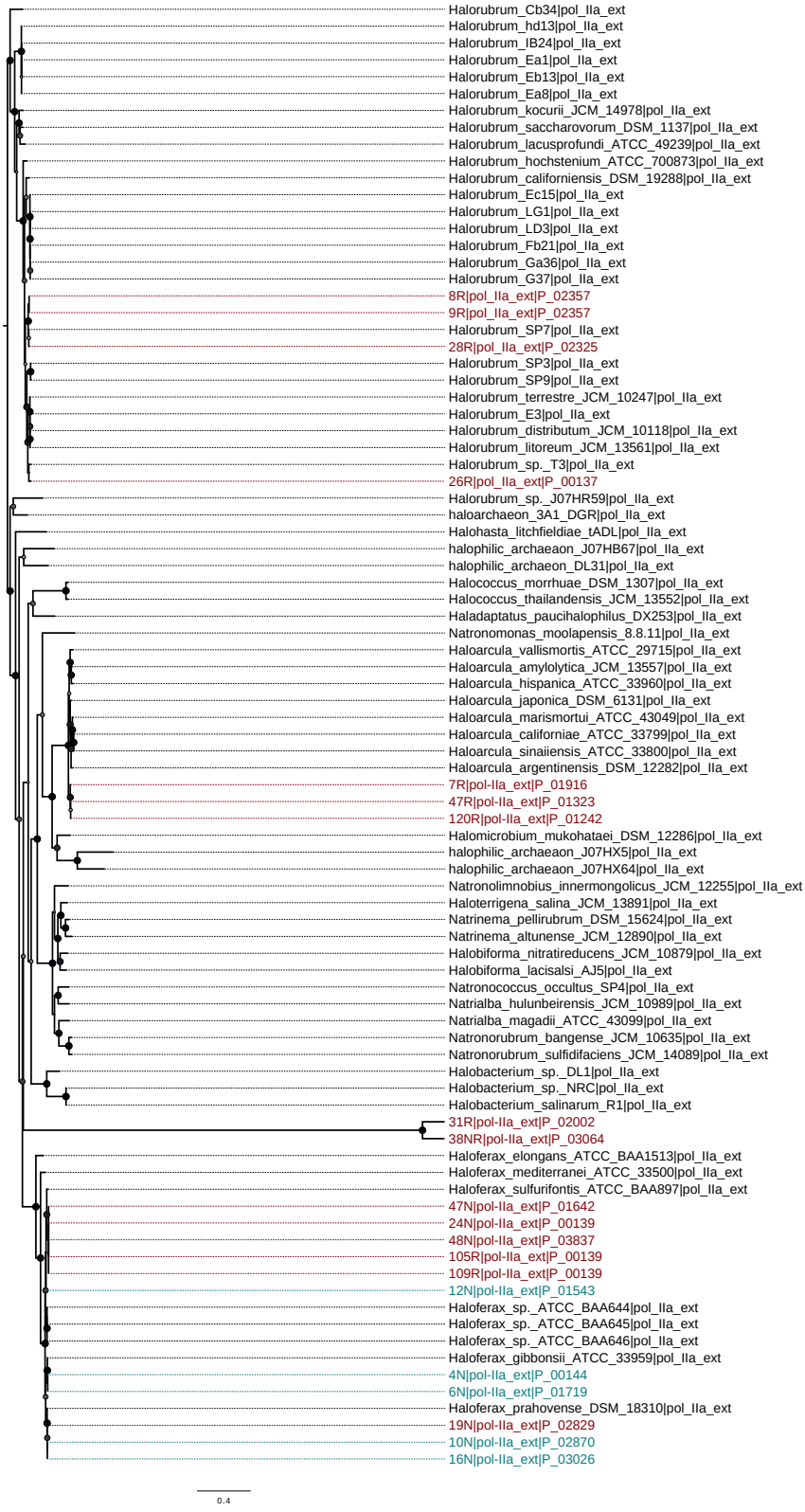


Figure 5 DNA polymerase II extein maximum likelihood phylogeny, strains from Atlit are highlighted in red text, and strains from Michmoret are colored in blue text. Support values are indicated by the size and color of the node shapes, larger and darker circles indicate strong support.

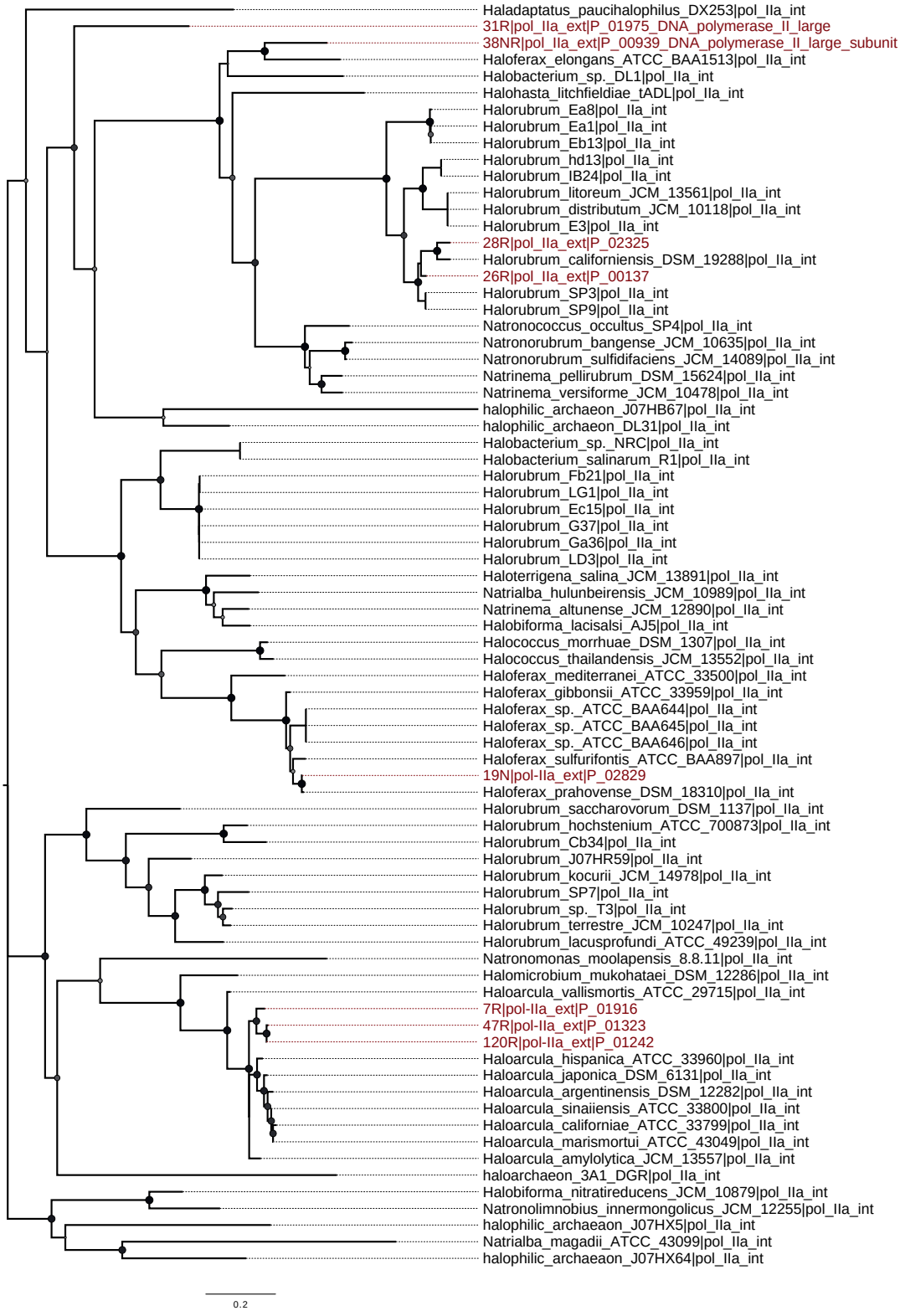


Figure 6 DNA polymerase II intein maximum likelihood phylogeny, strains from Atlit are highlighted in red text, and strains from Michmoret are colored in blue text. Support values are indicated by the size and color of the node shapes, larger and darker circles indicate strong support.

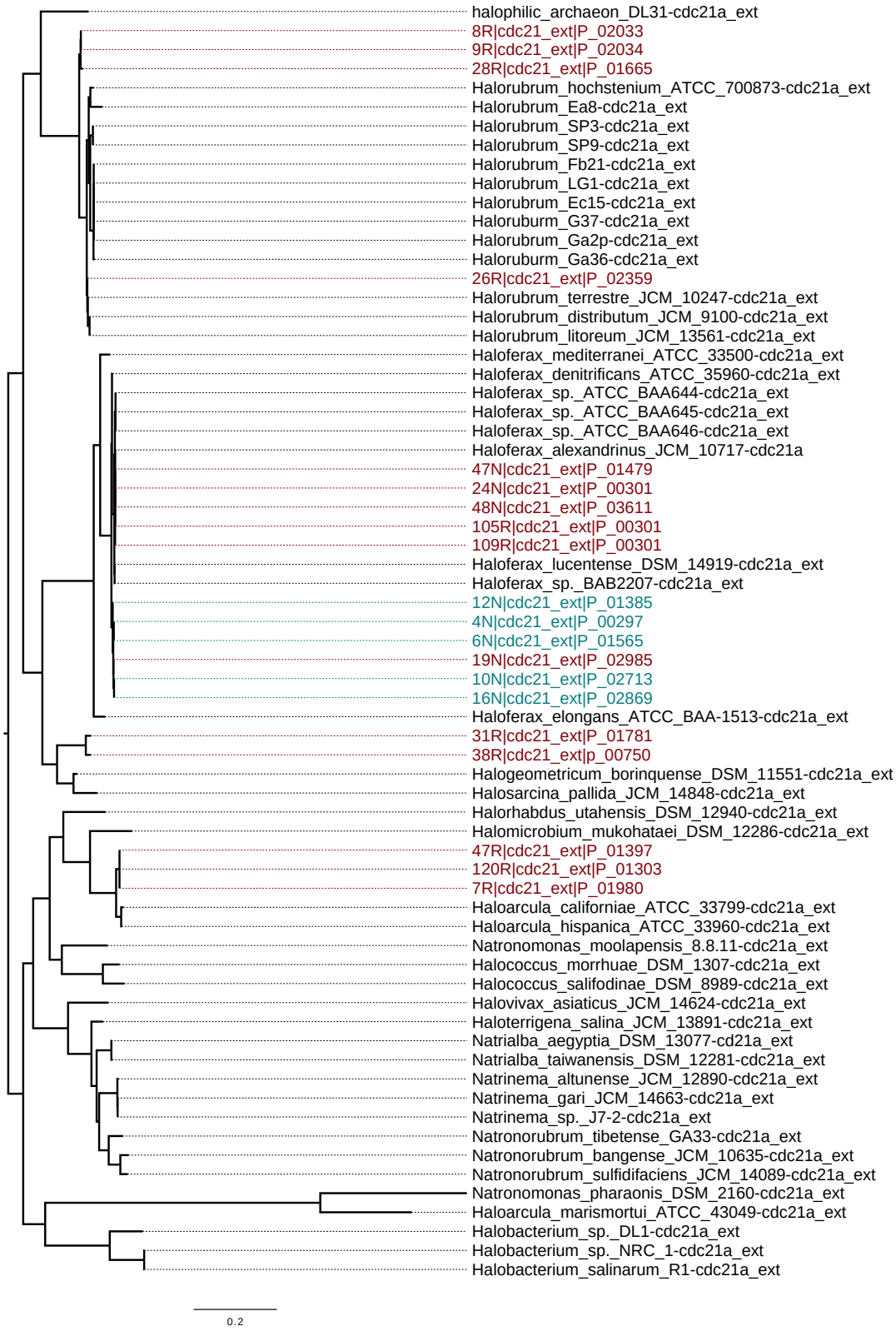


Figure 7 Cell division control protein 21 extein maximum likelihood phylogeny, strains from Atlit are highlighted in red text, and strains from Michmoret are colored in blue text. Support values are indicated by the size and color of the node shapes, larger and darker circles indicate strong support.

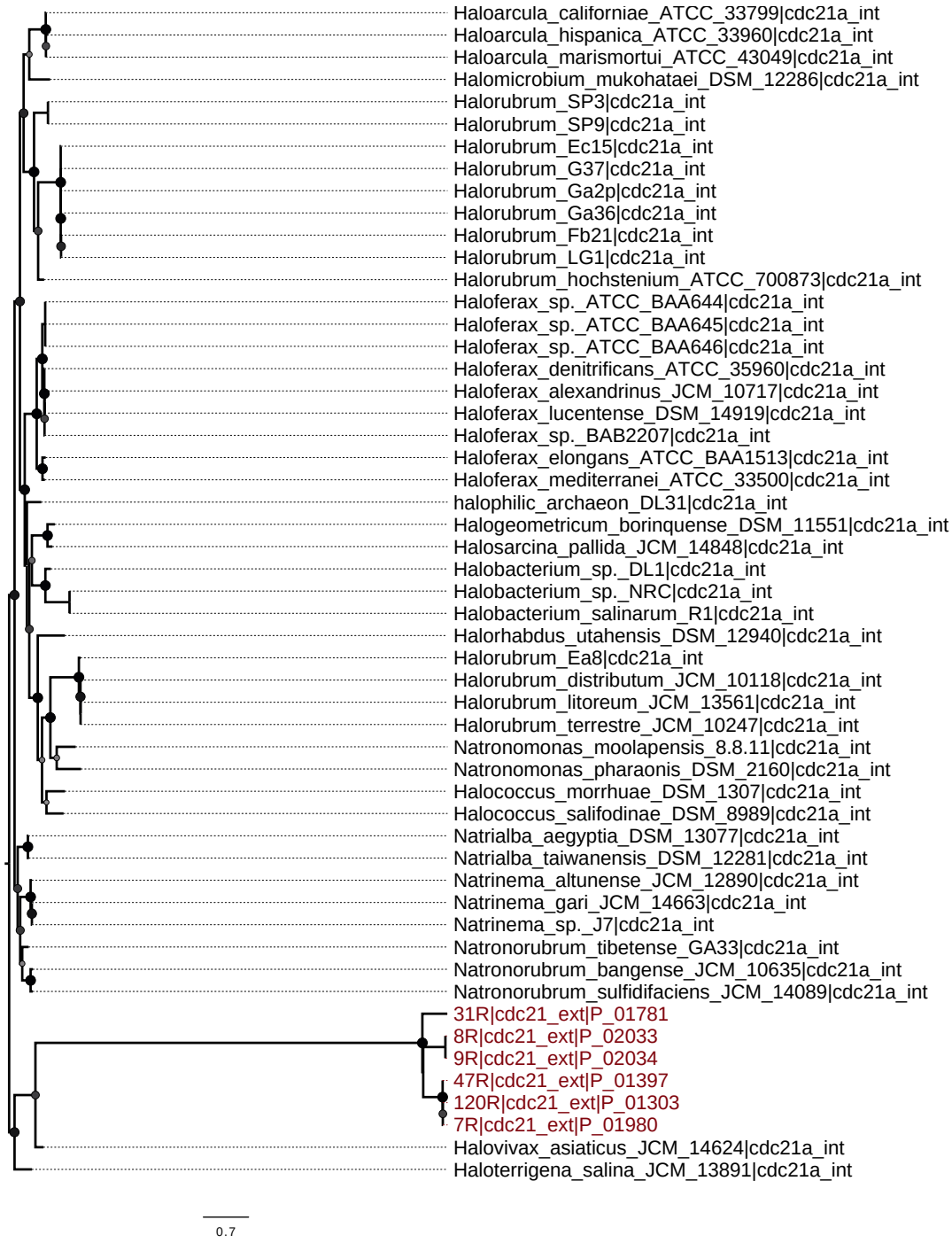


Figure 8 Cell division control protein 21 intein maximum likelihood phylogeny, strains from Atlit are highlighted in red text, and strains from Michmoret are colored in blue text. Support values are indicated by the size and color of the node shapes, larger and darker circles indicate strong support.

8.4 Future Work.

There is some computational analysis in the Gophna lab about interactions between strains based on CRISPR spacers. We will combine the work done here with that analysis to learn more about the evolution and distribution of haloarchaea in the Mediterranean Sea. These data will be compared to the data collected in Deep Lake, and together we will use these datasets with the simulations to compare the lifestyles of haloarchaea between the two environments based on their mobilome content.

Chapter 9. Intein Epidemiology: After Mating lab strains of *Haloferax volcanii* and *Haloferax mediterranei*.

9.1 Introduction.

The work in the following chapter is being done in collaboration with the Gophna lab. A graduate student in the Gophna lab, Israela Turgemen, conducted the bench work for this project and I am carrying out the computational analysis. The genomes in this section were generated after a mating experiment between *Haloferax volcanii* and *Haloferax mediterranei*. After mating the hybrid strains were plated on selective media, requiring mevinolin and novobiocin resistance genes, which are found on the pHv4 and pHm500 plasmids, respectively. After plating on selective media with the plasmid markers, strains were streaked onto media selecting for retention of chromosomal markers from both parental genomes (thymidine-, tryptophan-, and uracil-). Strains that were able to grow after the both selective platings were selected for sequencing. Hybrid strains were subject to *in silico* DNA DNA hybridization with both of the parental genomes. We also analyzed the intein content of these strains to look for intein loss or propagation in these hybrid strains. Lastly we examined the intein, extein, and ribosomal phylogenies to determine if inteins were gained from homologous replacement of the extein sequence, or through homing.

9.2 Materials and Methods

9.2.1 Genome Assembly

Reads were processed using cutadapt v1.9.1 to trim low quality bases ($Q_{score} \geq 20$ $p \leq 0.01$) and assembled using SPAdes v3.8 (81). Quality of the assembly was checked with QUAST v4.1 (82) and the assembly was annotated using prokka v1.9 (83).

9.2.2 Intein Retrieval

Each genome was searched for inteins using the same methods as in Soucy et. al 2014 (15). Exteins of each genome were aligned with the data set from Soucy et. al, and aligned using Muscle v3.8.31 (12) to look for the presence of an intein. Intein sequences were extracted and aligned to the intein datasets from Soucy et. al.

9.2.3 Phylogenetic Trees

Maximum likelihood trees were generated for each intein, extein, and ribosomal protein using RAxML v8.1.17 (14) with the GTR model, optimized substitution rates, and a gamma model of rate heterogeneity.

9.3.3 Genome Comparisons

Each genome was compared to all other hybrid genomes, as well as both parental genomes (*Haloferax volcanii* DS2 and *Haloferax mediterranei* ATCC 33500) using *in silico* DNA-DNA hybridization with the Genome-to –Genome distance calculator (84).

9.3 Results.

9.3.1 Reference Topology.

All of the hybrid strains were in good agreement with the reported background parental genotype (Figure 1). Strains 55IS-132IS have a *Hfx. mediterranei* background, with some *Hfx. volcanii*. Strains 315IS and 317IS are *Hfx. volcanii* background with some *Hfx. mediterranei*. Background parental types were initially determined using PCR, and later verified after genome assembly. The *is*DDH values of each hybrid strain and the two parental backgrounds indicates that there is a range in the amount of recombinant material in each of the hybrid strains, with the majority signal in good agreement with the reported parental background (Table1).

9.3.2 Intein distribution.

The parental strain of *Hfx. mediterranei* has the *cdc21*-a, *cdc21*-b, *polB*-b, and *pol-II*-a intein alleles. The parental strain of *Hfx. volcanii* has only the *polB*-c intein allele. All strains with a parental background of *Hfx. mediterranei* (strains 55IS-132IS) retained the *cdc21*-a and *cdc21*-b inteins, except for strains 132IS and 75IS (Table 1). Also all *Hfx. mediterranei* background strains retained the *pol-II*-a intein, except for strain 75IS (Table 1). Lastly all *Hfx. mediterranei* background strains acquired the *polB*-c intein, and lost the *polB*-b intein. All *Hfx. volcanii* parental background strains maintained the same intein states as their parental strain, except for strain 314IS, which lost the *polB*-c intein (Table 1).

9.3.3 Intein and Extein Phylogenies.

In order to determine if inteins were gained and lost through homologous recombination, or homing, the extein and intein topologies were compared. All intein topologies were as expected based on the distribution of intein alleles (Figure 2). Interestingly most strains with a *Haloferax mediterranei* background maintained their inteins in cdc21-a, cdc21-b, and pol-II-a, however strain 73IS lost all inteins associated with the parental strain and strain 132IS lost the cdc21-a and b inteins. In all cases where an intein was lost the extein grouped with *Haloferax volcanii* rather than *Haloferax mediterranei*, the parental strain (Figures 2, 4, & 6). Indicating that inteins are lost through homologous recombination, rather than precise deletion, at least during mating. Additionally all strains with a *Haloferax mediterranei* background gained the intein in the DNA polymerase B gene, position c from *Haloferax volcanii*. For these strains the extein was also replaced- indicating the intein is acquired not through homing, but through homologous replacement.

9.4 Future Work.

There are a couple of things we are still interested in exploring. I would like to verify that the examples where an intein was gained or lost (section 9.3.3) occurred through homologous replacement and not through recombination. In order to test this I will use GARD(85), and cBROTHER(86) to look for signs of recombination occurring within the extein.

Alternatively the homologous replacement that resulted in the replacement of the alleles above could have been a single event that spans a large genomic distance. In order to determine where recombination occurred in each of the hybrid strains I will employ a method used by Pascal Lapierre in (33) to look for recombination boundaries in each of the hybrid genomes. This method entails collecting identity scores for each open reading frame in each parental genome compared to the hybrids. This work will produce a schematic of each hybrid genome, where the recombinant regions are indicated by a drop in identity to the background parental strain of each hybrid and an increase in the percent identity to the minor parent.

9.5 Outlook.

This work is fundamentally important as the intein dynamics are occurring under controlled circumstances. Thus far we have assumed that intein propagation most likely occurs during mating, when the DNA and protein version of the intein containing sequence can be found in a shared space with the un-invaded copy during the mating process. The change in intein distribution after mating, and also the way new inteins are acquired, through homologous replacement or homing, informs our understanding of how inteins spread in populations, as well as the types of HGT the population experiences. Furthermore the collection of eleven strains that have undergone the same mating process can illustrate the breadth of possibilities that can occur after mating. Some strains like 73IS have retained copies of exteins that do not contain inteins, regardless of the origin of the gene. Strain 132IS has replaced one extein (and lost the associated

inteins), but retained other exteins, generating hybrid replication machinery. Lastly the analysis of recombination boundaries in so many strains, and strains with different parental backgrounds enables us to learn more about the process of recombination during mating.

Strain	Major Parent	Minor Parent	cdc21	dtd	gyrB	helicase	ligase	polB	pol-II	rfc	rir	rpolA	top6B	topA	udp
1_6_IS	Hfx. Mediterranei	Hfx. Volcanii pHv4	a,b	no	no	no	no	b	a	no	no	no	no	no	no
132IS	Hfx. Mediterranei	Hfx. Volcanii	no	no	no	no	no	c	a	no	no	no	no	no	no
55IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
57IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
62IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
66IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
73IS	Hfx. Mediterranei	Hfx. Volcanii	no	no	no	no	no	c	no	no	no	no	no	no	no
74IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
75IS	Hfx. Mediterranei	Hfx. Volcanii	a,b	no	no	no	no	c	a	no	no	no	no	no	no
316IS	Hfx. Volcanii	Hfx. Mediterranei	no	no	no	no	no	c	no	no	no	no	no	no	no
317IS	Hfx. Volcanii	Hfx. Mediterranei	no	no	no	no	no	c	no	no	no	no	no	no	no

Table 1 Intein presence and absence in hybrid strains. Inteins that are from *Haloferax mediterranei* are colored in purple, and inteins from *Haloferax volcanii* are colored in purple.

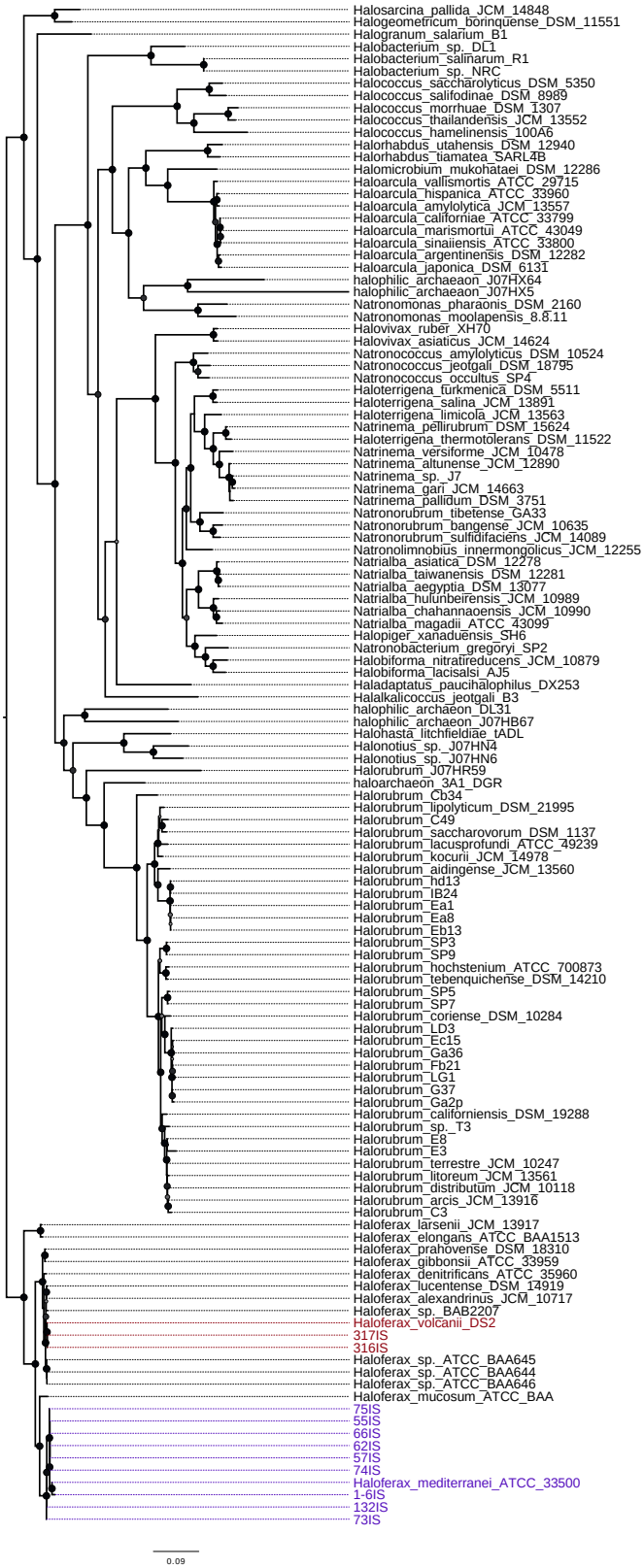


Figure 1. Maximum likelihood ribosomal reference tree, built using a concatenated alignment of 56 ribosomal proteins. Node shapes indicate support values, larger and darker circles indicate strong support. Hybrid strains with a *Haloferax volcanii* parental background are highlighted with red text, and strains with a *Haloferax mediterranei* parental background are highlighted purple.

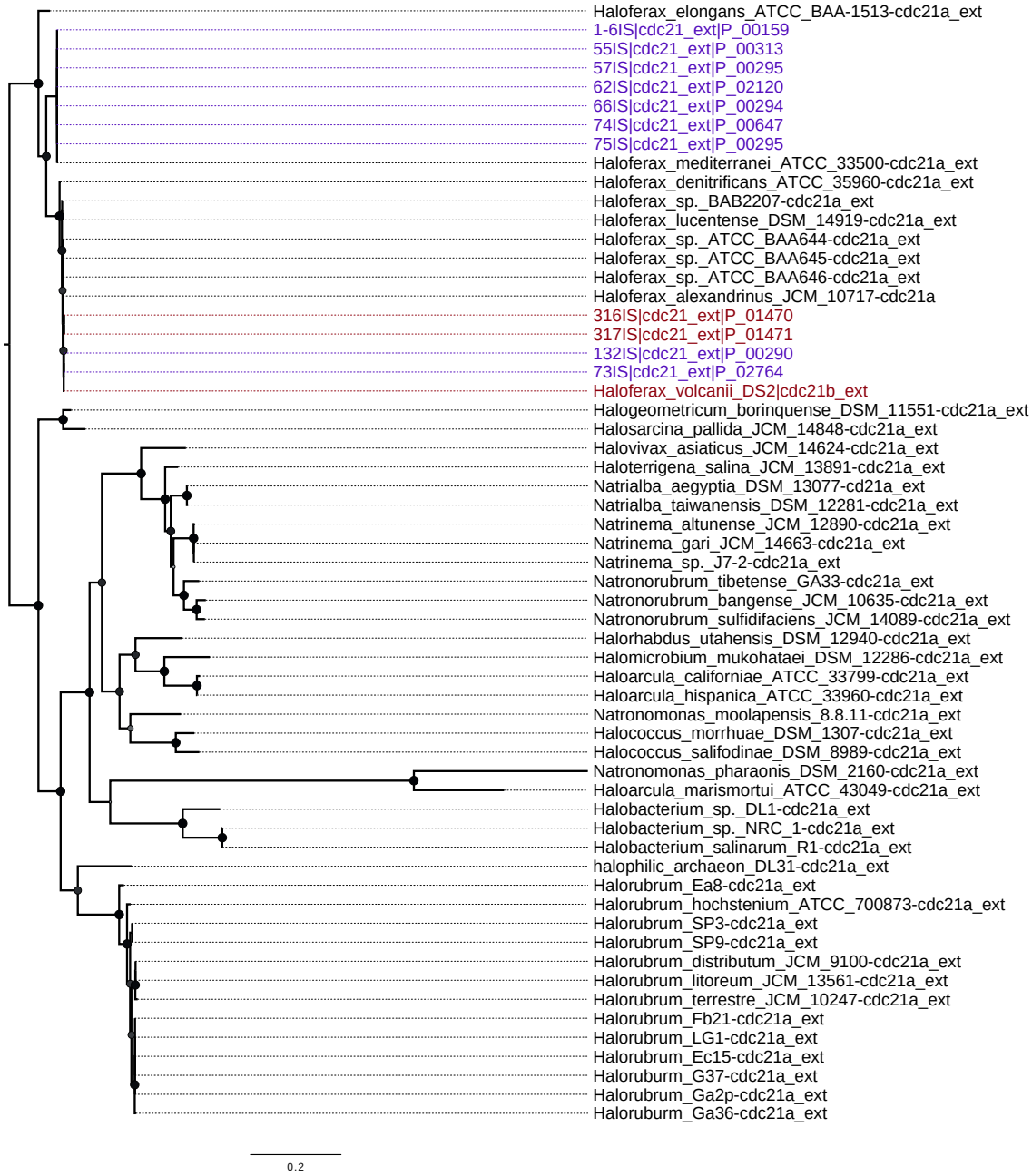


Figure 2 Cell division control protein 21 extein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

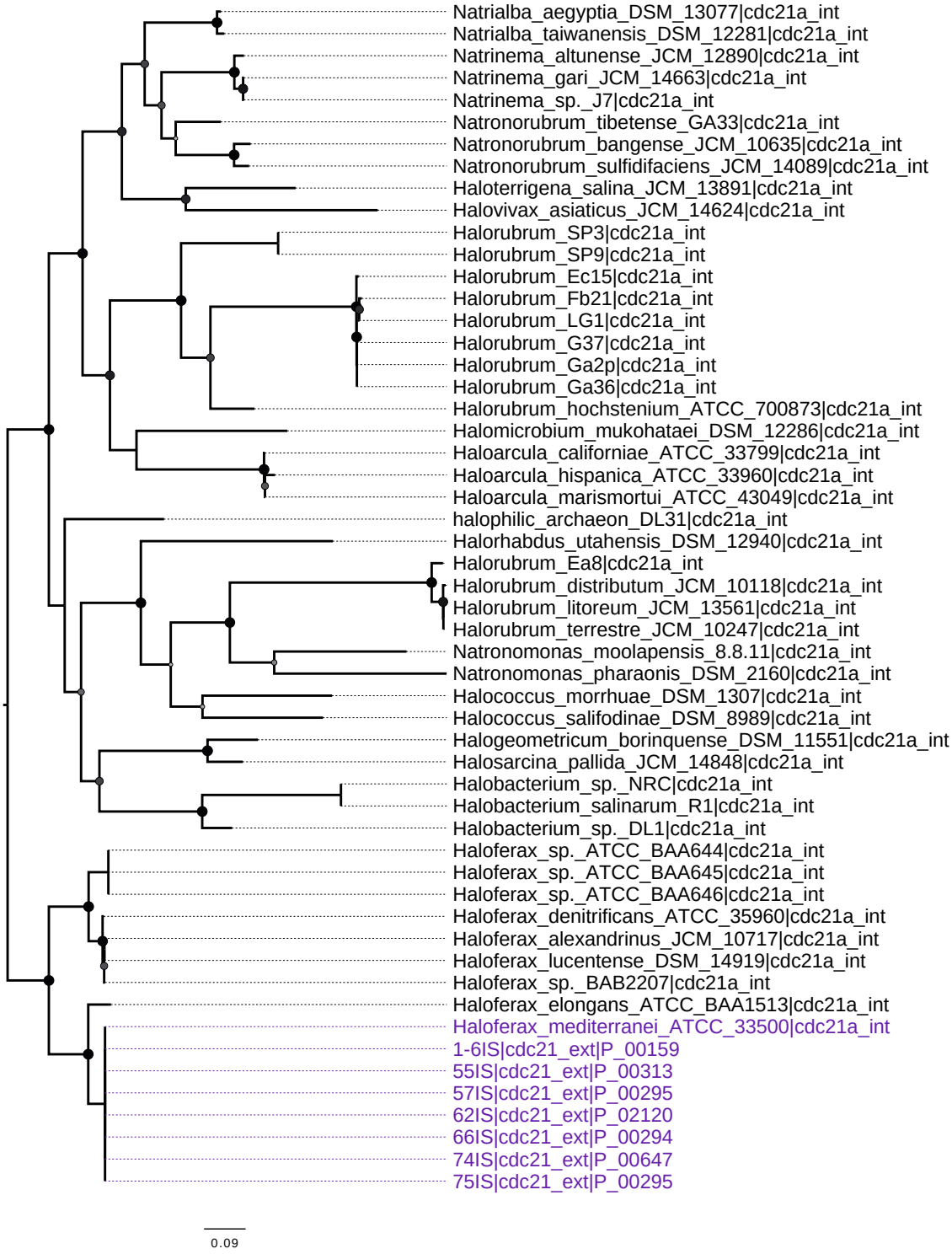


Figure 3 Cell division control protein 21 a intein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

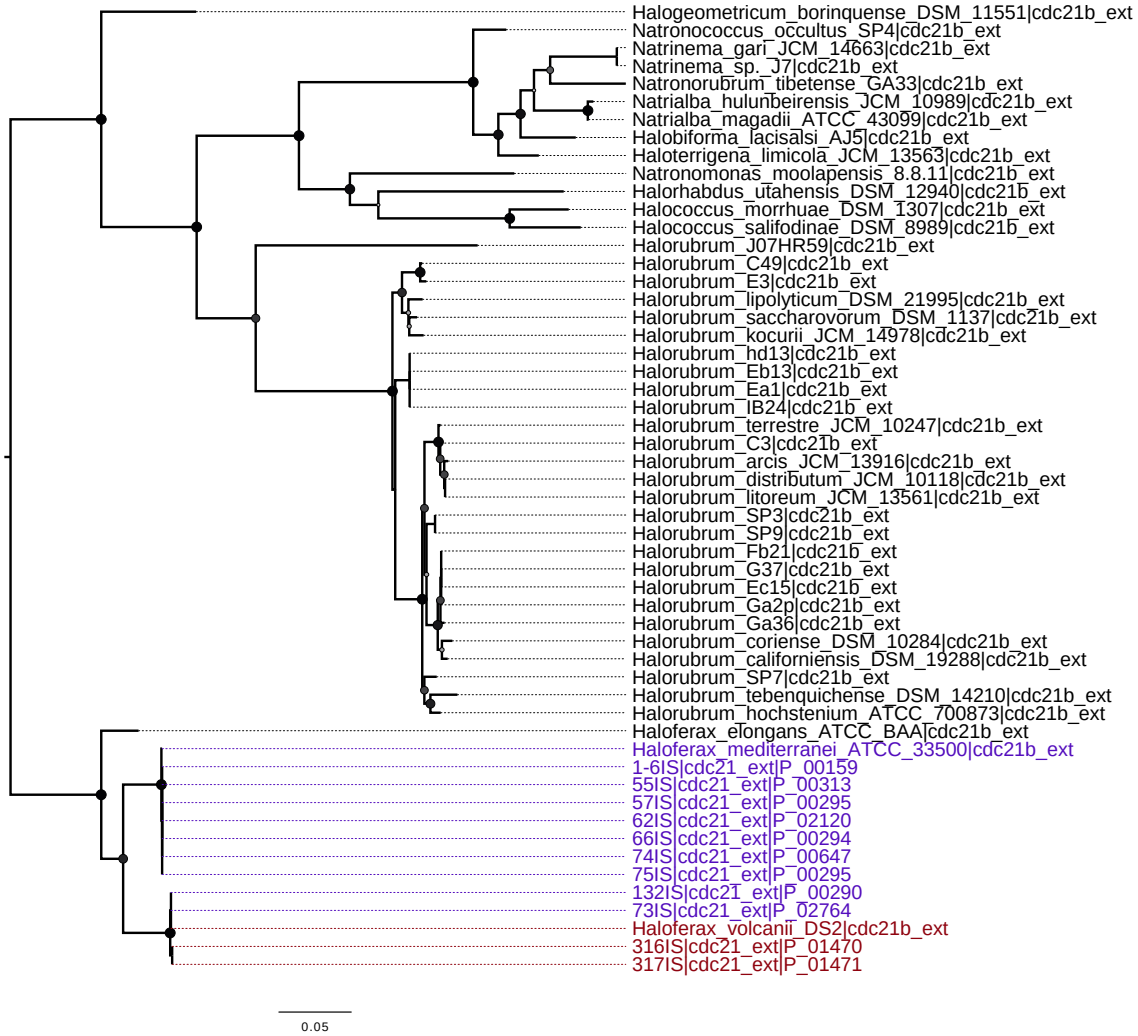


Figure 4 Cell division control protein 21 extein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

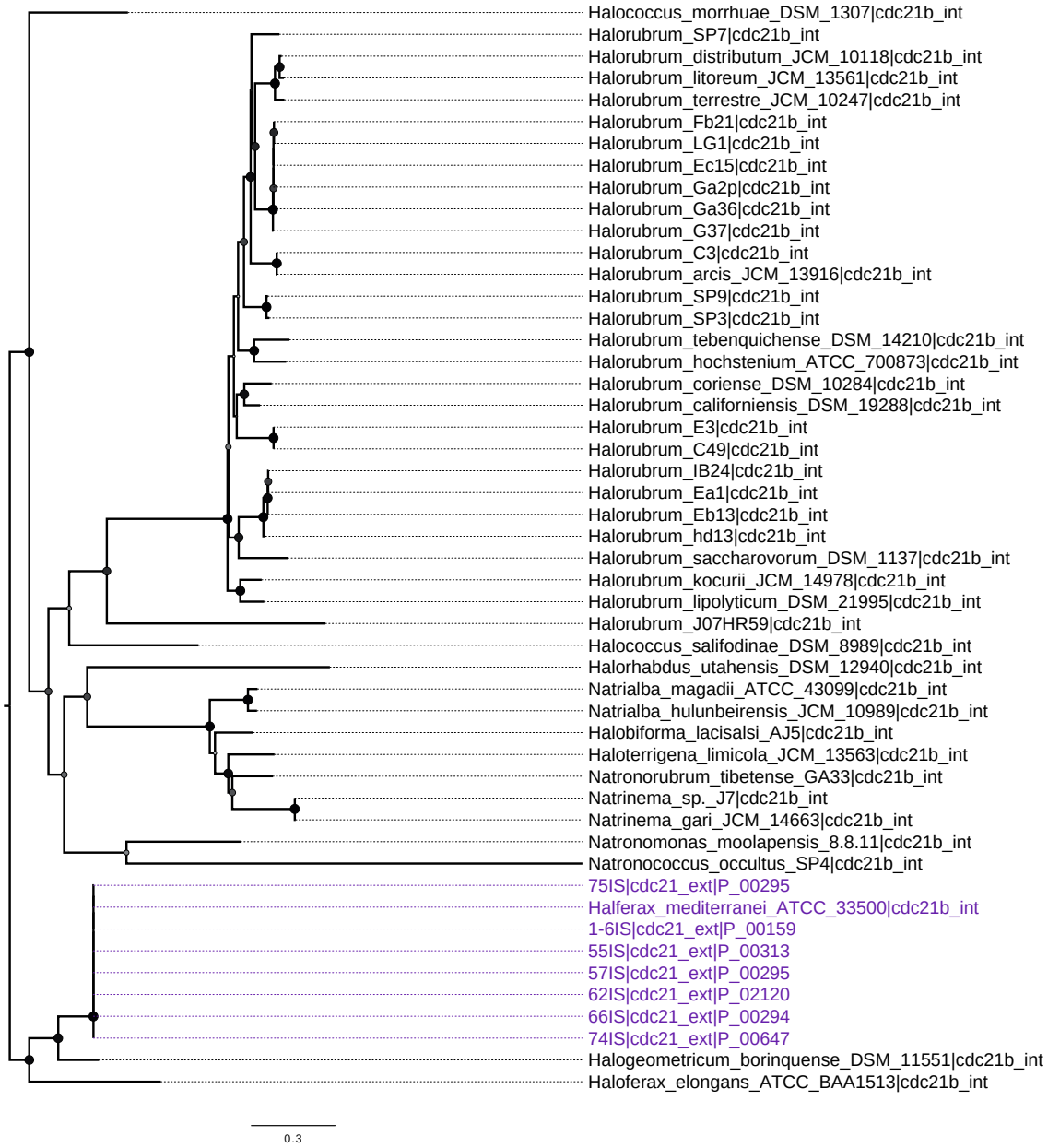


Figure 5 Cell division control protein 21 b intein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

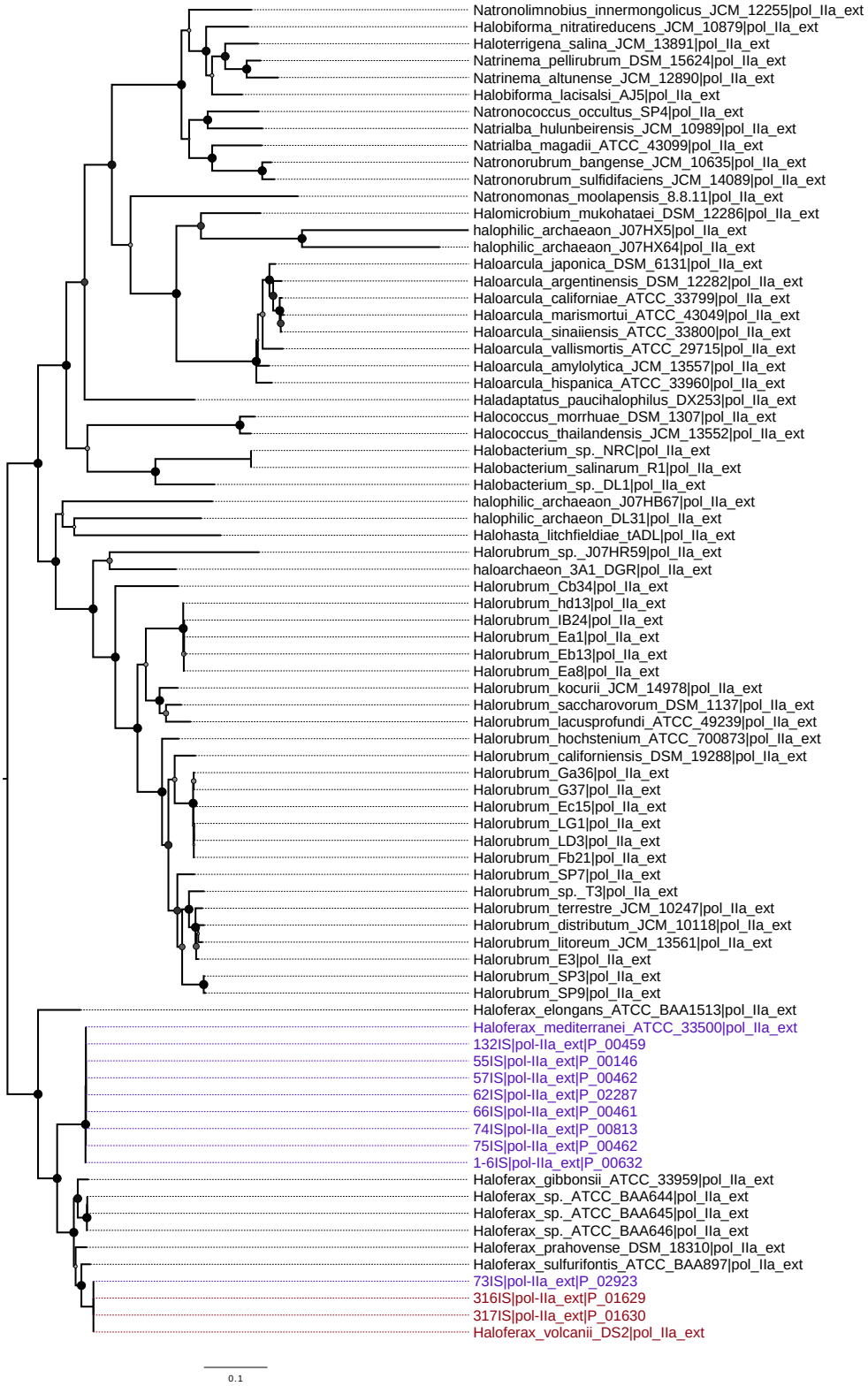


Figure 6 DNA polymerase II extein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

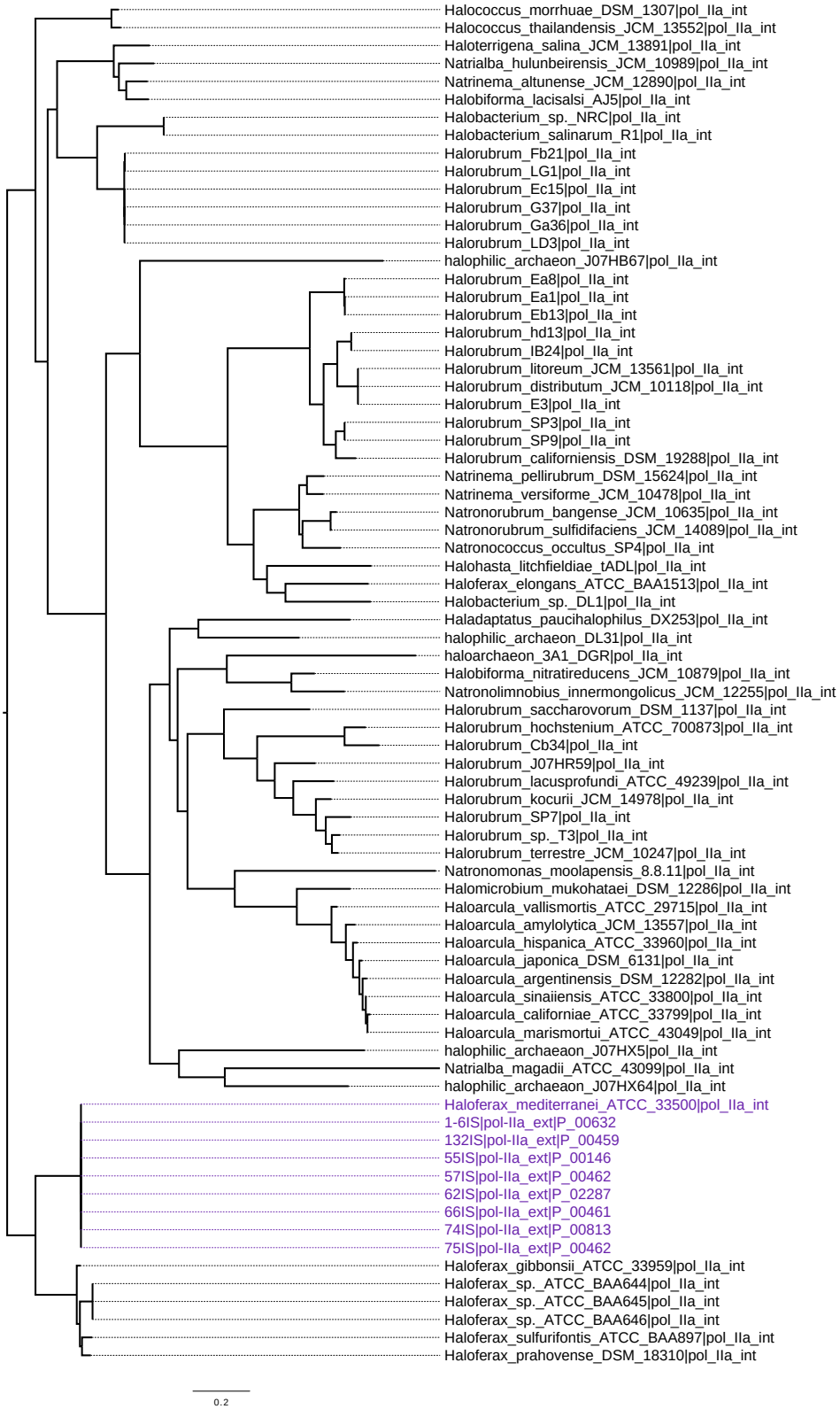


Figure 7 DNA polymerase II a intein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

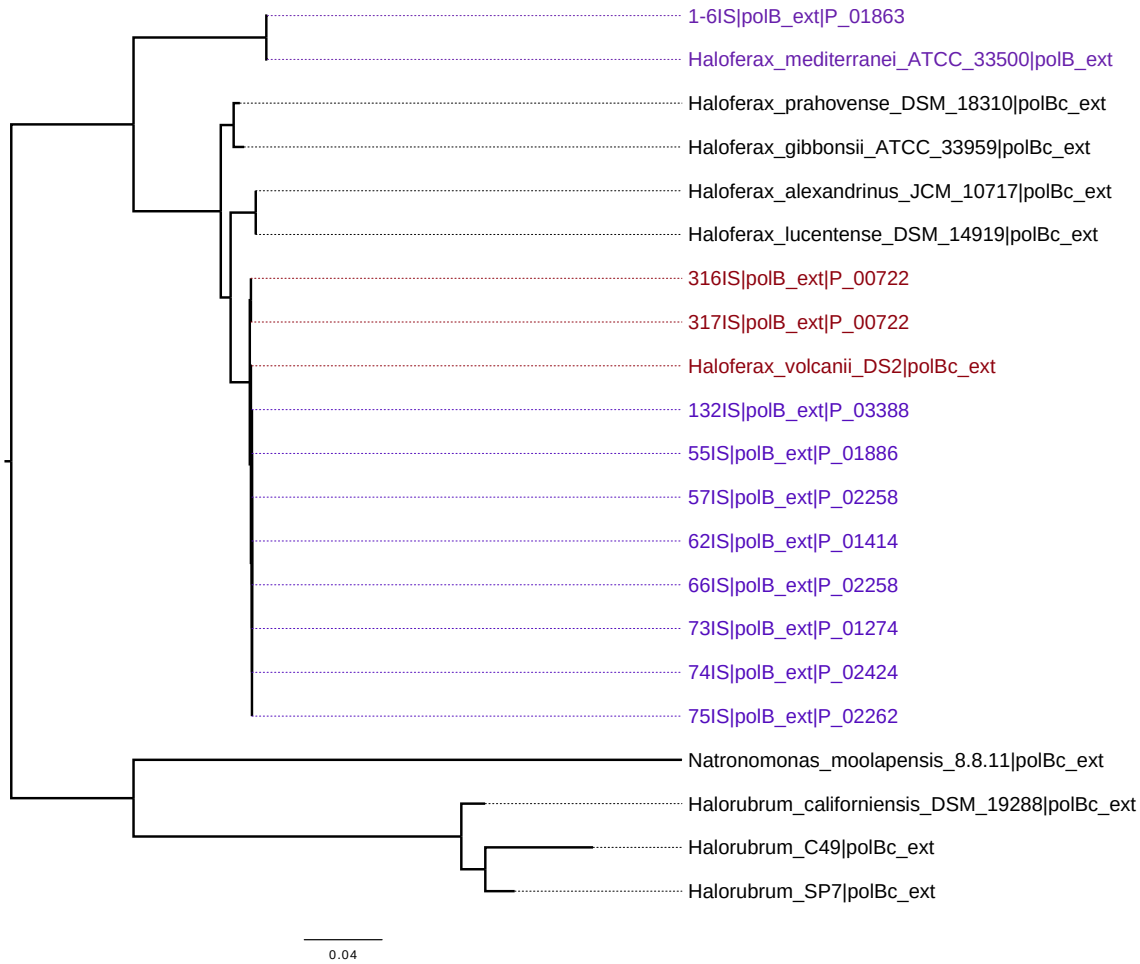


Figure 8 DNA polymerase B extein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

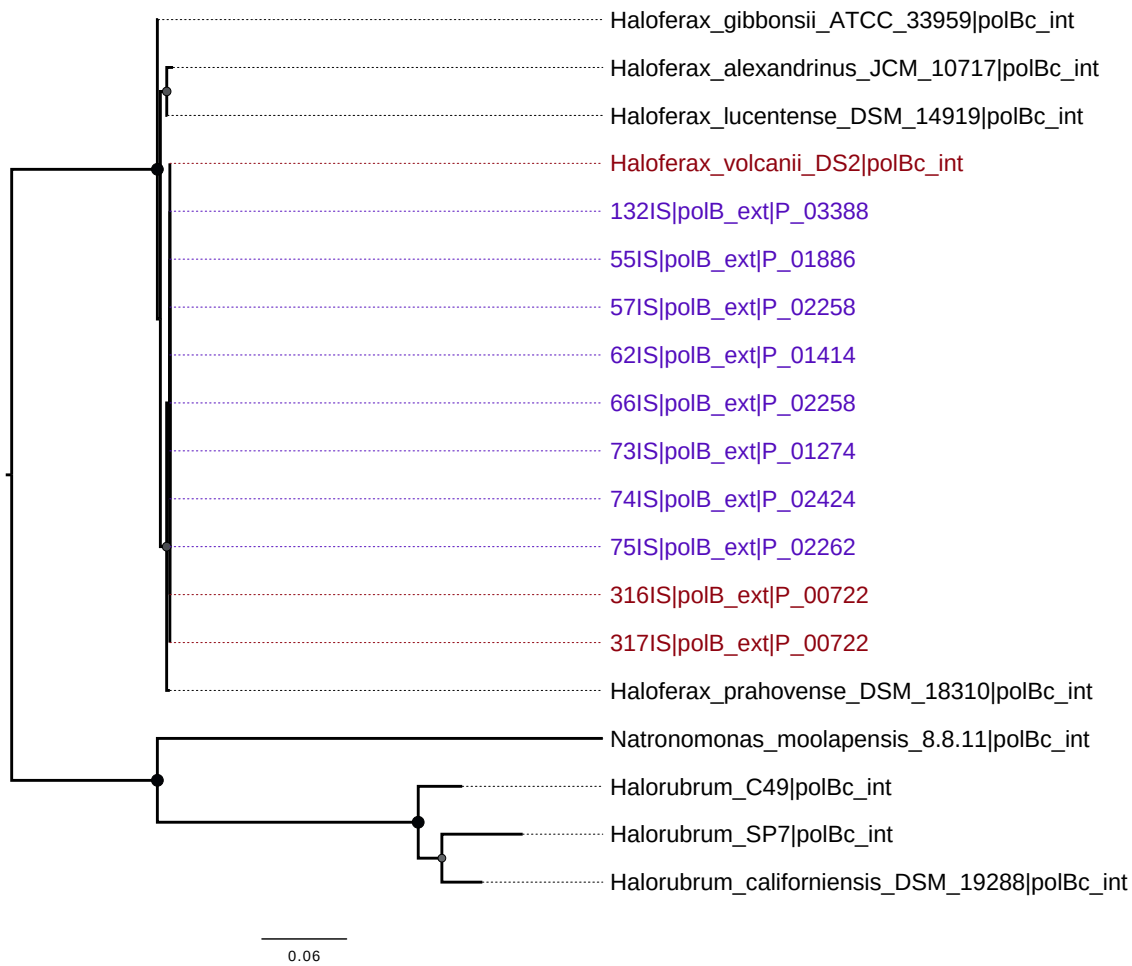


Figure 9 DNA polymerase B c intein maximum likelihood phylogeny. The parental background of each strain is indicated by the color of the text, in red are strains with a *Haloferax volcanii* background, and in purple are strains with *Haloferax mediterranei* background. The node shape indicates support values, larger and darker circles indicate strong support.

Chapter 10. Conclusions and Outlook.

The goal of this research was to examine the boundaries of HGT using the distribution, phylogeny, and symbiotic state of shared inteins in the haloarchaea. This proved to be quite a challenge as each intein allele has a unique phylogenetic history (15) involving both vertical and horizontal gene transfer signals. Through the work done here we have shown that these gene histories can be used to distinguish gene transfer between both distant (87) and closely related organisms, even within a species (54). By combining all intein sequences with a presence absence matrix we consider the strongest signal of gene transfer, and show that most gene transfer occurs within environments between very closely related species (15). Later (chapter 5) we use phylogenetically independent methods to reconcile gene transfer events that can be reconciled within our dataset as opposed to un-represented or un-sampled lineages. Further work will be done to reconcile each gene transfer event that can be explained using our dataset from (15).

The observation that most gene transfer events were occurring within environments is not exactly surprising for the haloarchaea whose distribution punctuated by salinity gradients. In order to build more relevant networks of gene transfer for the haloarchaea, we gathered genomes that were linked by a shared environment and considered them independently. The Deep Lake dataset was highly surprising in that despite evidence of intergenera gene exchange there was no evidence of intein homing or transfer between any of the sequenced genomes, though two of the intein sequences did not agree with the reference, these transfers came from outside of the dataset represented here (chapter 7). Further examination of the metagenomic reads revealed un-invaded exteins from *Halohasta litchfieldiae* tADL, indicating that intein-free alleles of *pol-II-* are able to co-exist with intein-containing versions of closely related organisms (chapter 7).

We had seen preliminary evidence of this in strains of *Haloferax spp.* isolated from the Mediterranean coastline, but we weren't sure to what extent these organisms were exposed to each other between rock pools (chapter 6). The data in Deep Lake showed that inteins were co-existing with intein-free alleles in very closely related strains in a way that resembles the predicted distribution in a large population ($10^{6)}$ by the intransitive fitness model (30) rather than the homing cycle.

In order to explain this co-existence we performed our own modeling of intein invasion using parameters, including the fitness cost of the polB-c intein, measured in lab strains of *Haloferax volcanii*. Our models showed that the intein would eventually go to fixation in a population even with low rates of homing as the population gets close to the carrying capacity (chapter 6). Thus the co-existence of intein-free and intein-containing alleles indicates that homing is rare, and this may be an evolutionary strategy to be maintained in populations over long periods of time. This work also shows that inteins are able to increase the frequency of recombination during mating between intein-containing and intein-free organisms. Taken together this data supports the intransitive fitness model where inteins adopt a strategy such that the homing efficiency is balanced by the fitness cost imposed on the host, and the fitness cost is in part alleviated by the benefit of increased recombination to the population as a whole (chapter 6).

In order to test for evidence of the homing cycle in Deep Lake I looked for evidence of discordant read pairs using two different platforms, and found no evidence of discordant pairs indicating structural variation, associated with loss of the homing endonuclease domain. Next, to look for smaller gaps within the mapped read I used seed-recruiting to look for mismatched base pairs and gaps within reads aligned to the reference. No gaps were found, and mismatched bases

were concentrated in the linker region of the intein rather than the HEN domain as would be expected if the homing endonuclease were decaying (chapter 7). More work needs to be done to confirm that this pattern can be replicated with other inteins, these counts will also be normalized by the number of reads mapped per site before anything definitive can be assumed as the number of mismatched reads is not significant and cannot be distinguished from missed base calls. Thus far this is weak evidence at best that the homing cycle does not describe the evolution of inteins in this population. However, taken together with the lack of discordant reads detected, and the presence of the un-invaded version of the *pol-II*-a intein this strongly supports the intransitive fitness model.

Further analysis of genomes linked by environments (chapter 8) demonstrated many examples where intein-free and intein-containing alleles co-exists in closely related species. The majority of intein phylogenies from these pools were in good agreement with the extein and ribosomal phylogeny. Interestingly in contrast to the data in Deep Lake, all strains from Atlit with the cdc21-a intein group together to the exclusion of the other haloarchaea, indicating the intein was gained through homing within the environment. Furthermore the organisms that share this intein are from several different genera, grouping sister to *Halogeometricum borinquense* (31R), *Halorubrum sp.* SP7 (8R and 9R), and *Haloarcula vallismortis* (7R, 47R, 120R). Also all of the organisms that share the cdc21-a intein were isolated from separate pools, indicating that these pools are probably well mixed. Furthermore, strain 24N was collected in 2012from Atlit and the very closely related 105R and 109R were collected in 2013, indicating there is some stability in the population from year to year.

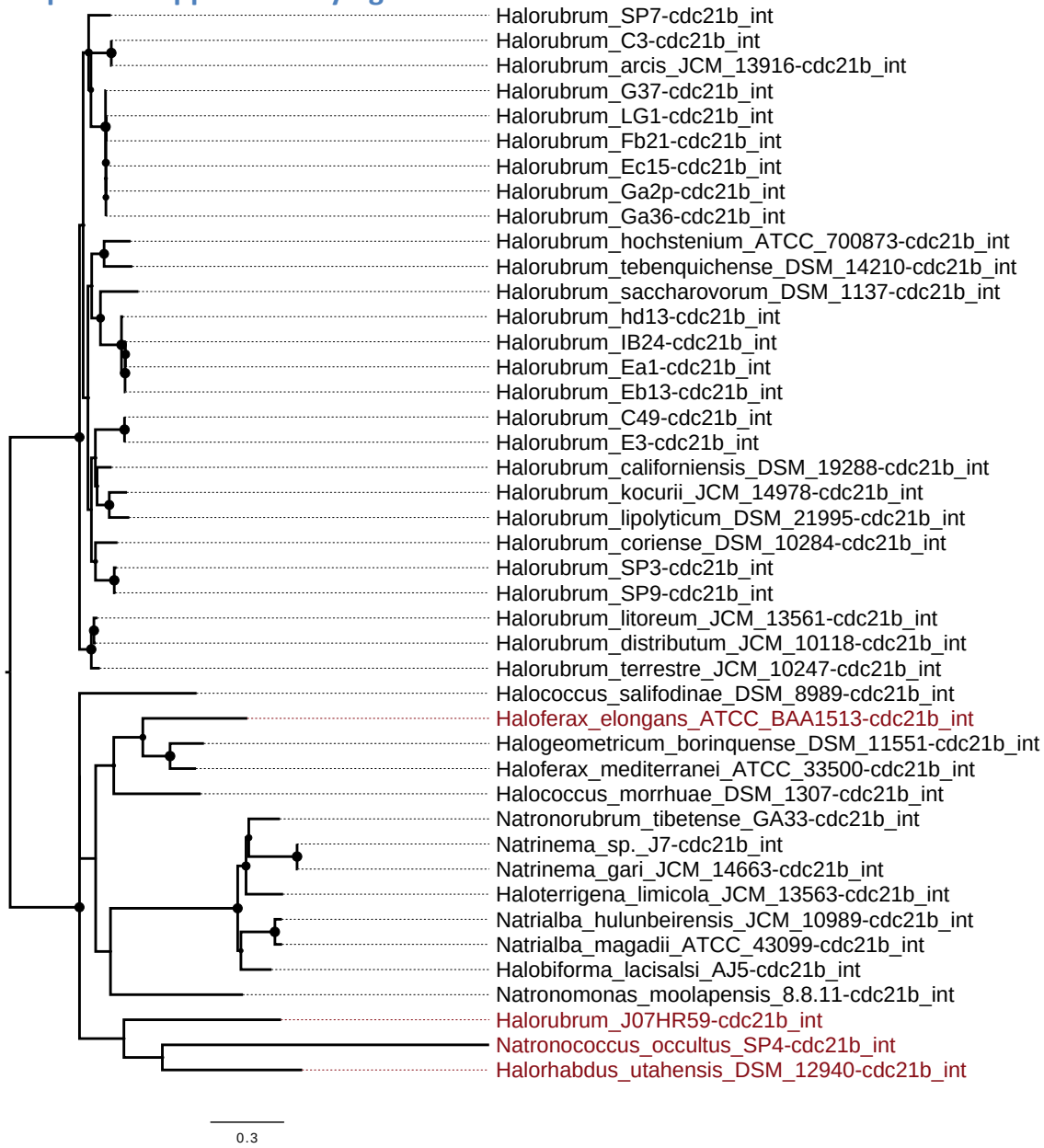
Lastly using data from hybrid genomes generated from mating experiments we showed that all intein dynamics have occurred through homologous replacement with their extein

sequence, and possibly an even larger genomic span as a part of the mating process. Further work will indicate the span of recombination and the breakpoints. This result was surprising given the work in chapter 6 shows that inteins increase the frequency of recombination during mating if one species is intein-free and one contains the intein.

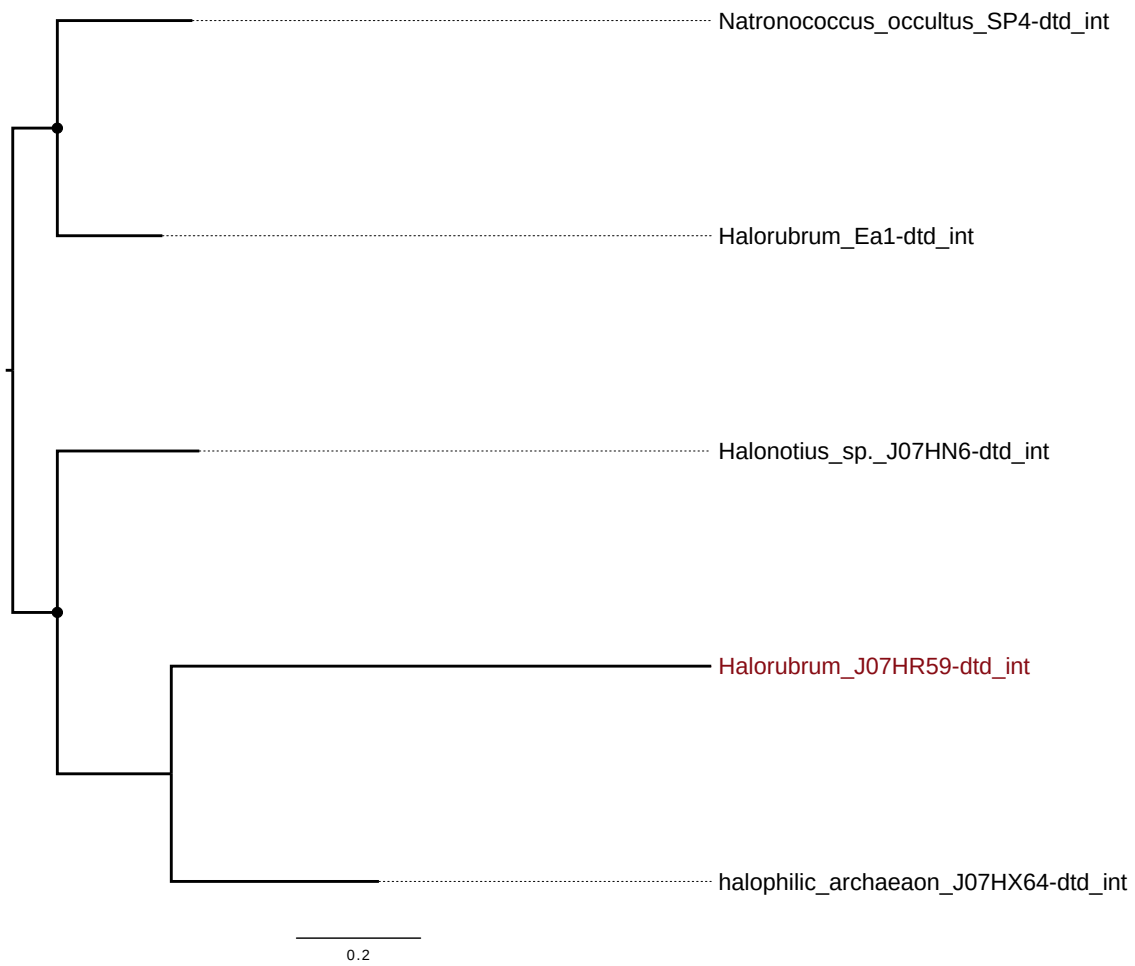
My initial interest in the Microbiology program at the University of Connecticut was the breadth of research pertaining to the various benefits of symbioses. I was drawn to the idea of studying the benefits and functions associated with microbes, where many programs focus on pathogenesis in microbes. I ended up in the Gogarten lab where he was looking for someone to work on inteins. As I stated the initial goal was to use inteins to build networks of gene transfer. The process was not as straightforward as we often speak about in grants, but through several attempts and years of work and collaboration I have worked to develop inteins as a tool to study gene transfer. In the process of doing so I began to think about the symbiotic relationship between a host and a parasitic element, and how the fitness cost of the intein changes depending on the environmental conditions, the diversity in the population, the presence of the homing endonuclease domain, and the presence of empty intein insertion sites. Organically a study that was initially focused on investigating horizontal gene transfer became a study of molecular symbiosis. Instead of studying the benefits of associating with microbes, I've focused on the benefits and consequences of carrying mobile genetic elements not just through my research but also through the reviews I've helped to author. In my next position I plan to continue to explore these concepts using gene transfer agents to examine the domestication of a mobile element.

Appendix A. Supplementary information.

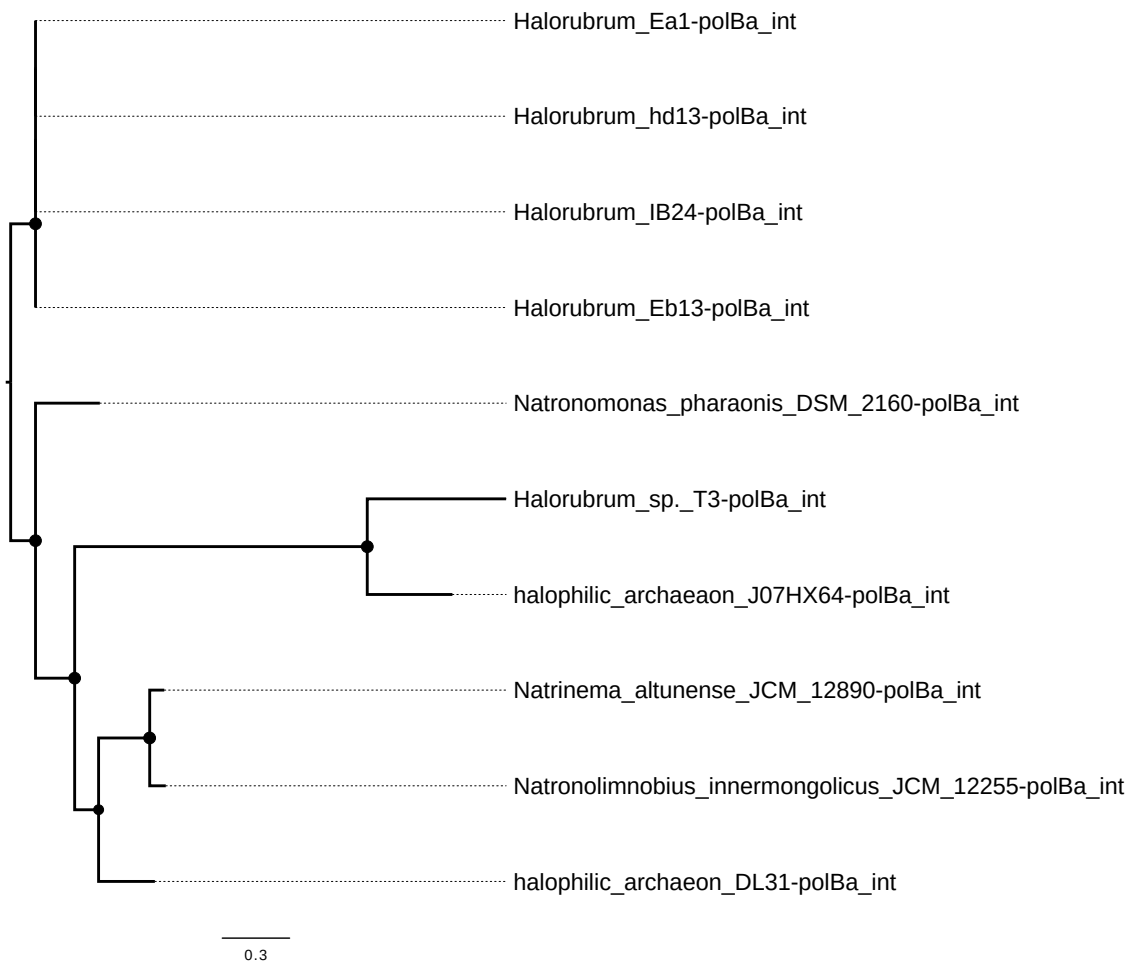
Chapter 5. Supplementary figures.



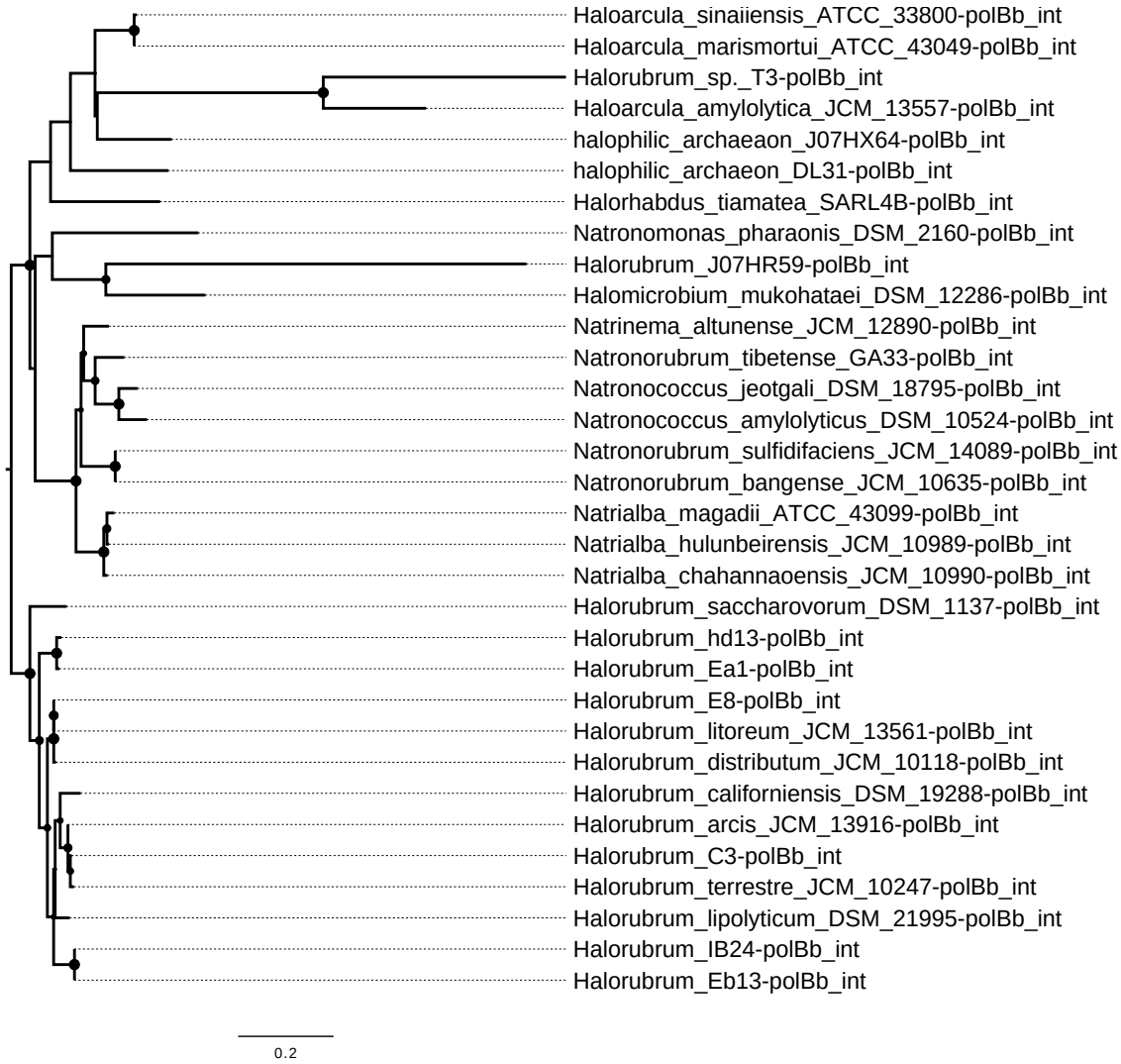
Supplementary Figure 5.1 Maximum likelihood phylogeny of cdc21-b intein. Mini-inteins are highlighted in red text, support for bipartitions is indicated by the node shape, larger circles indicate strong support.



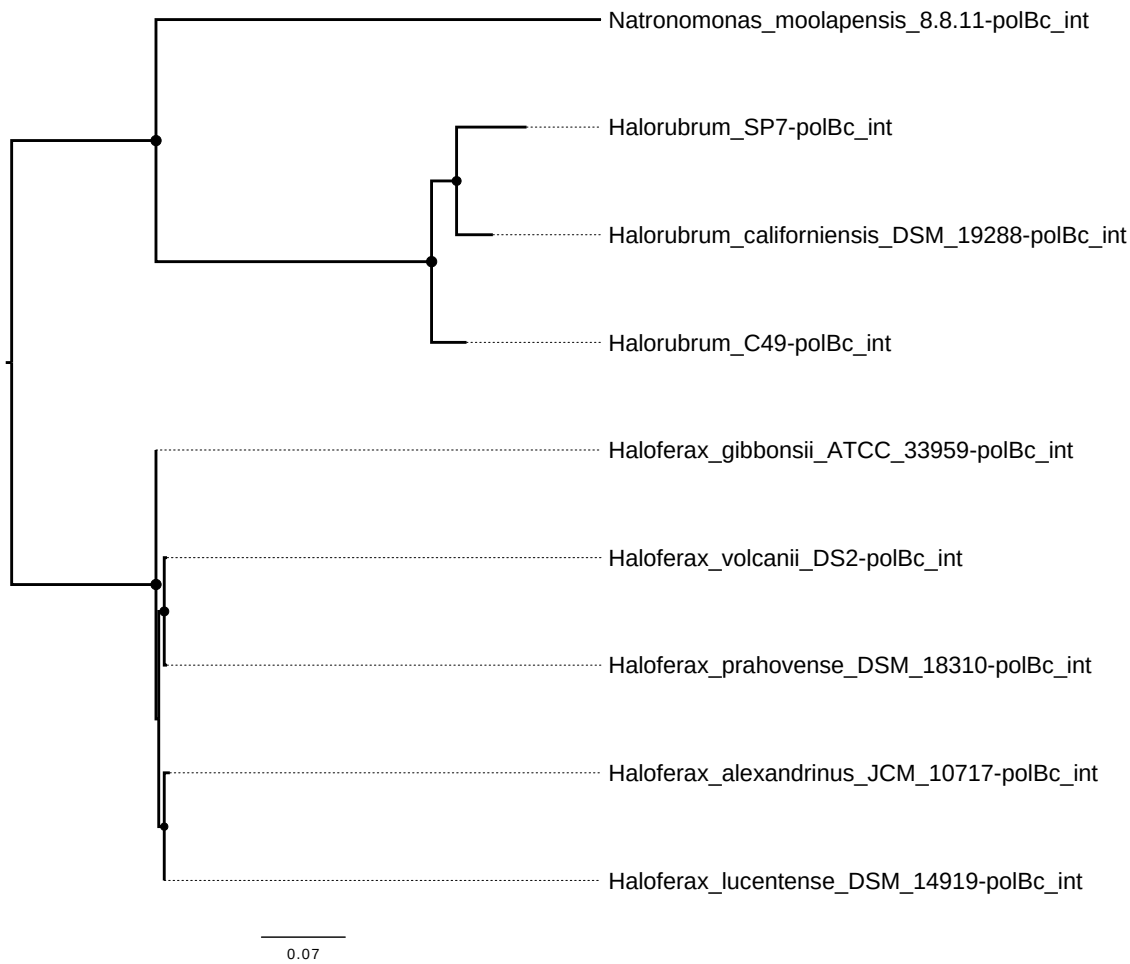
Supplementary Figure 5.2 Maximum likelihood phylogeny of dtd intein. Mini-inteins are highlighted in red text, support for bipartitions is indicated by the node shape, larger circles indicate strong support.



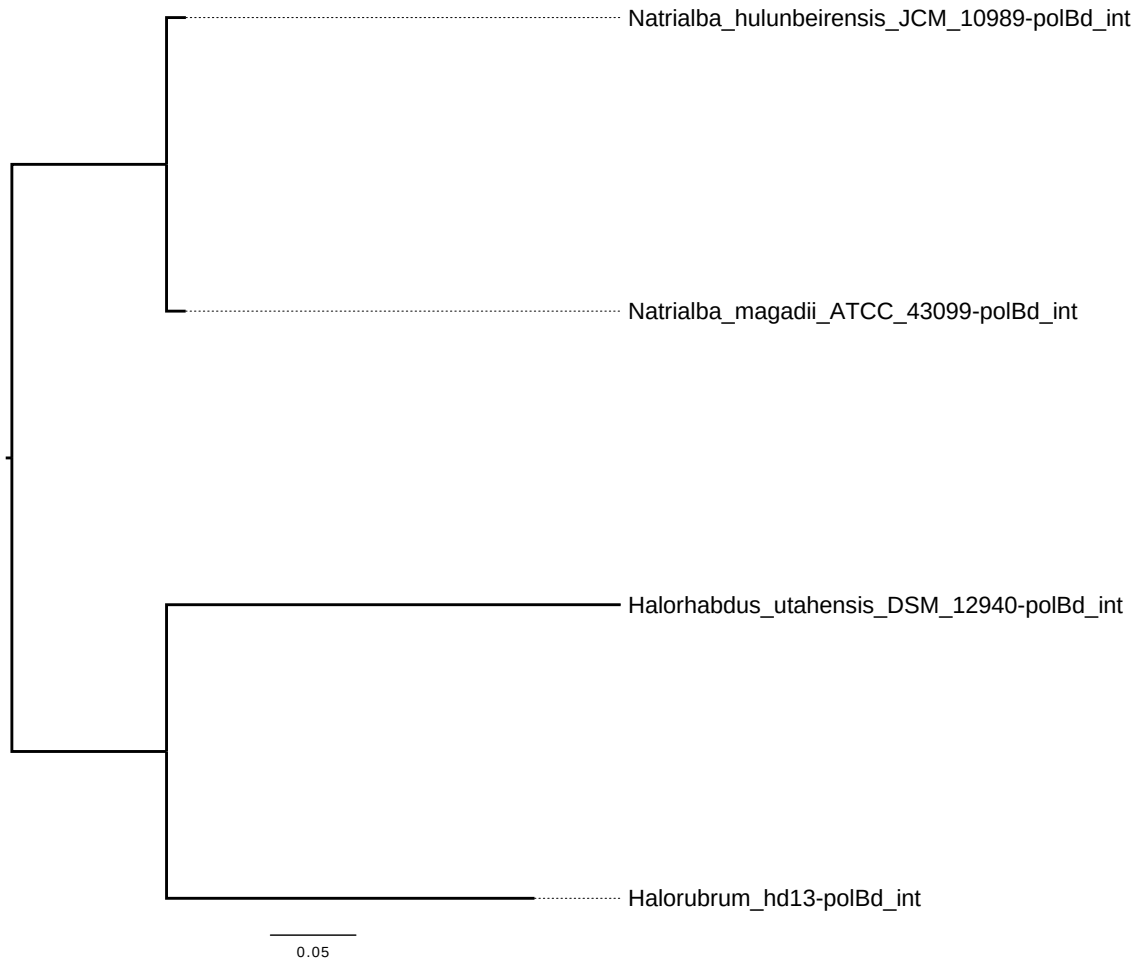
Supplementary Figure 5.3 Maximum likelihood phylogeny of polB-a intein. The node shape indicates support for bipartitions larger circles indicate strong support.



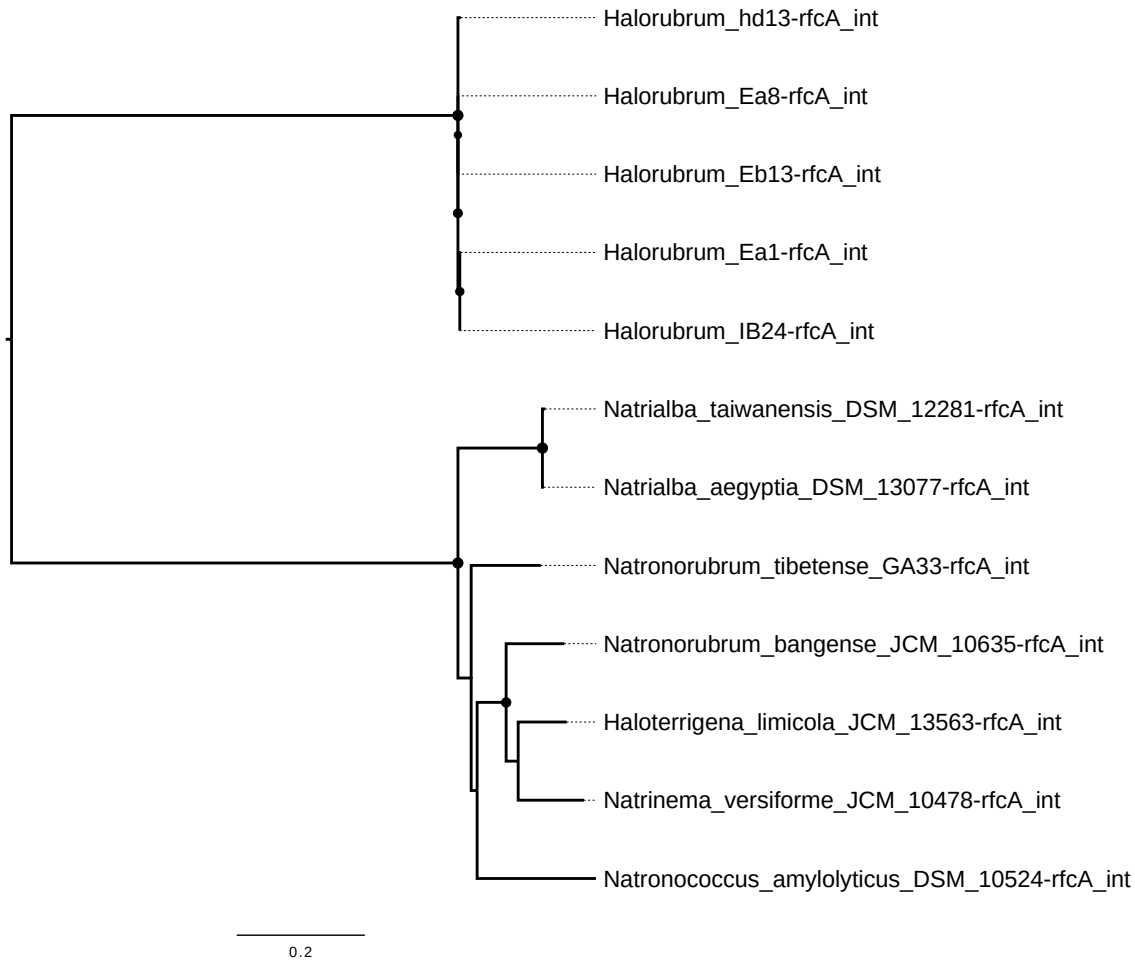
Supplementary Figure 5.4 Maximum likelihood phylogeny of polB-b intein. The node shape indicates support for bipartitions larger circles indicate strong support.



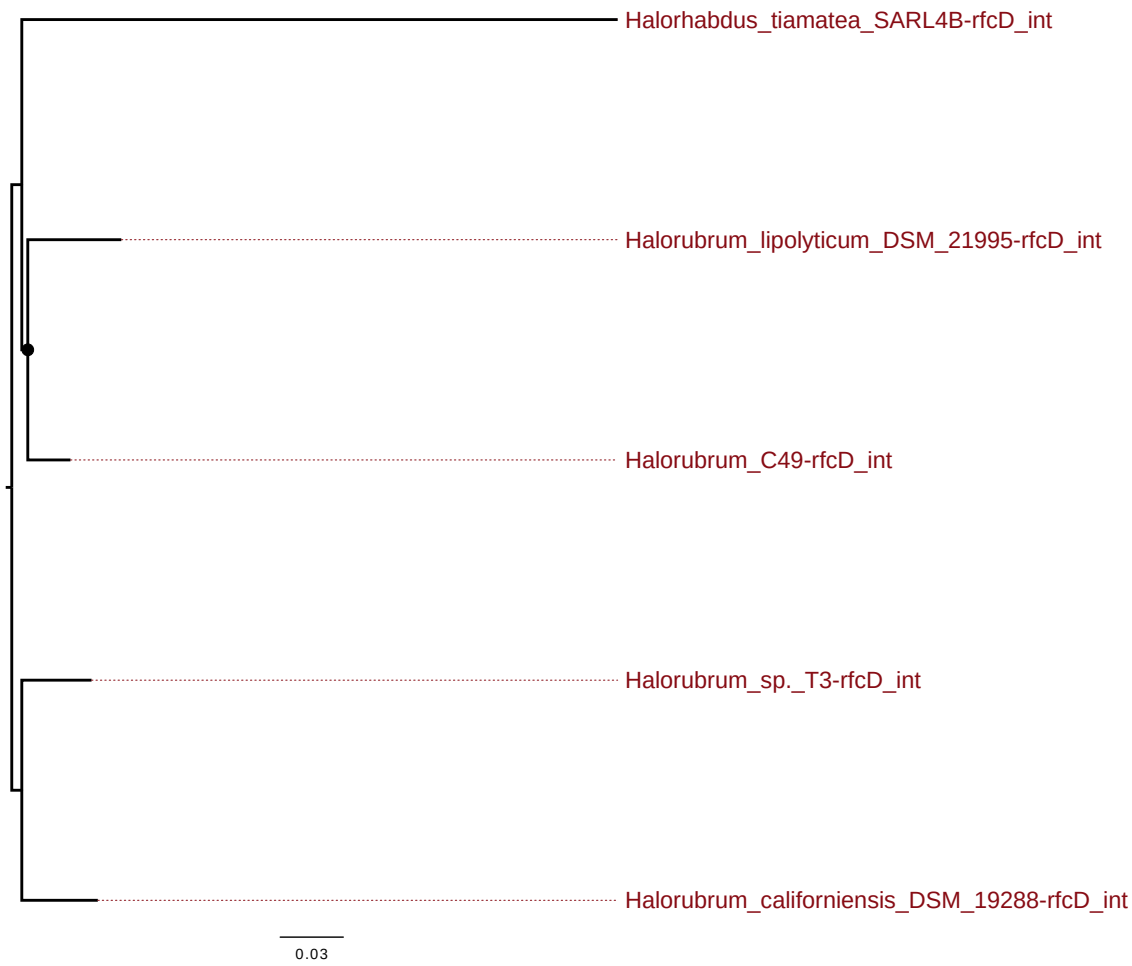
Supplementary Figure 5.5 Maximum likelihood phylogeny of polB-c intein. The node shape indicates support for bipartitions larger circles indicate strong support.



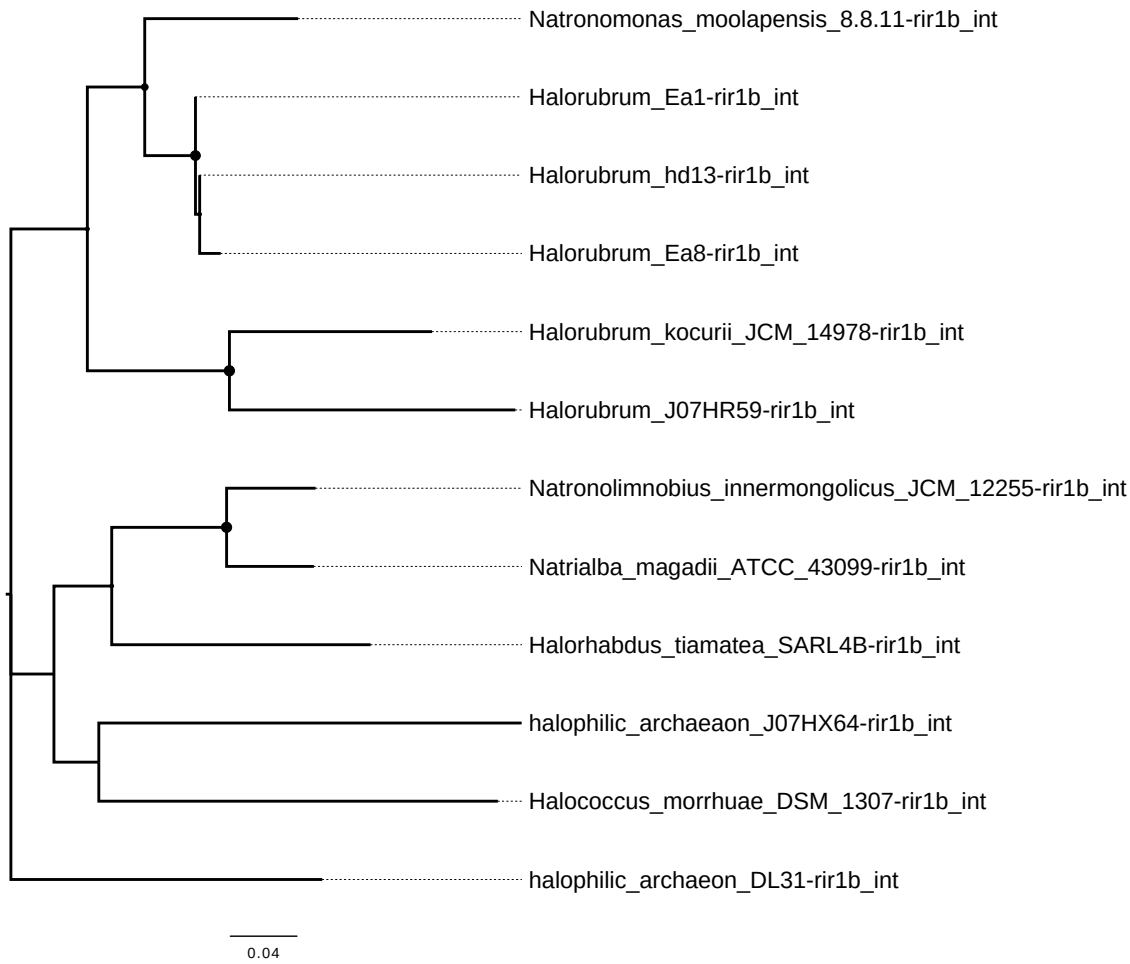
Supplementary Figure 5.5 Maximum likelihood phylogeny of polB-d intein. The node shape indicates support for bipartitions larger circles indicate strong support.



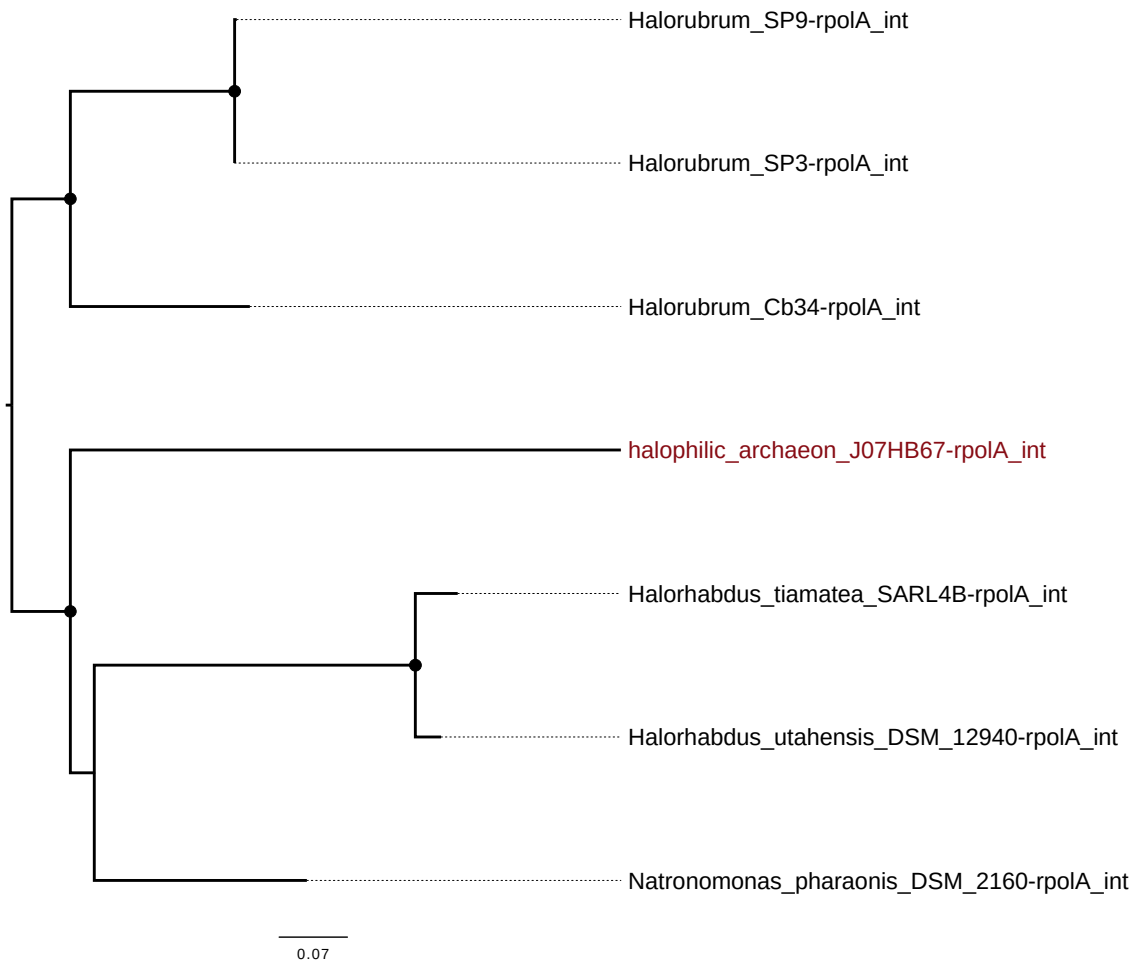
Supplementary Figure 5.6 Maximum likelihood phylogeny of rfcA intein. The node shape indicates support for bipartitions larger circles indicate strong support.



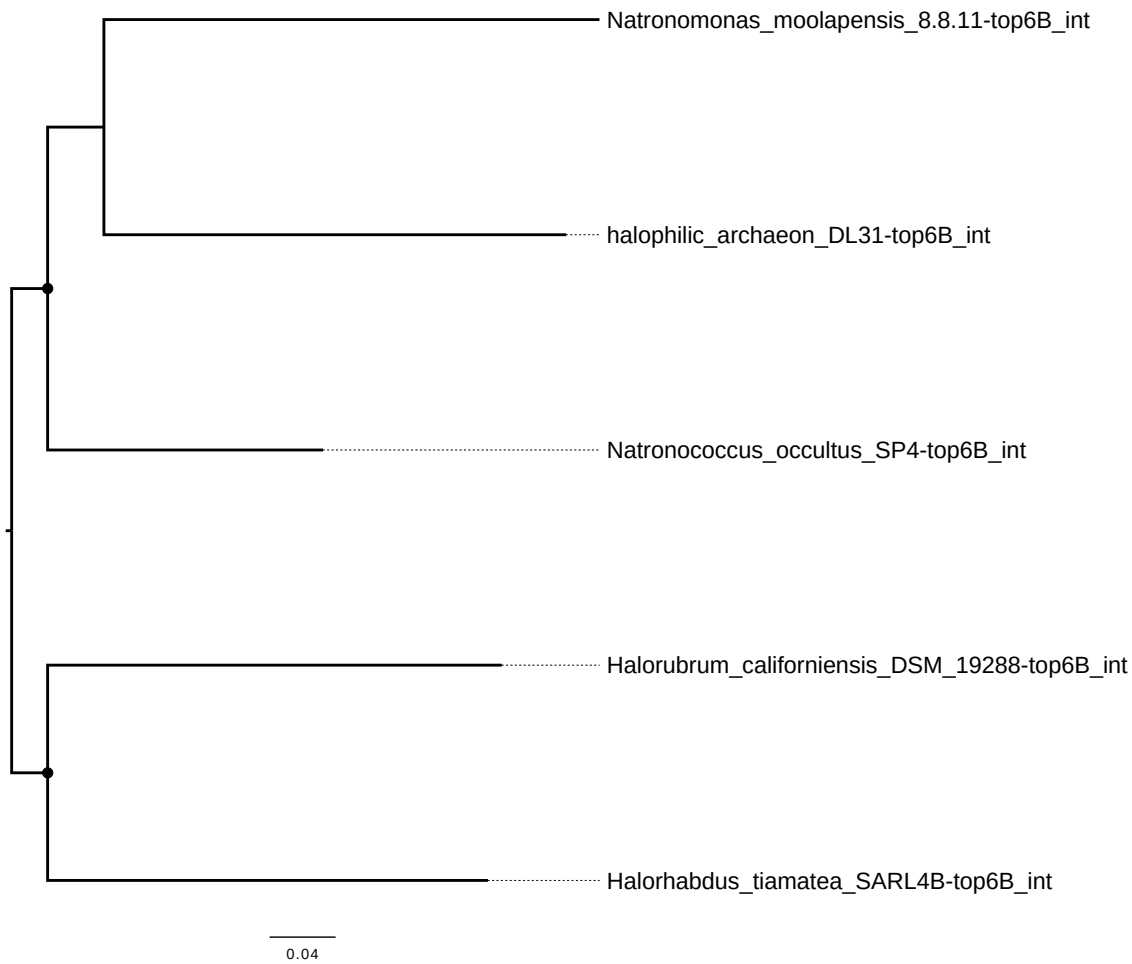
Supplementary Figure 5.7 Maximum likelihood phylogeny of rfcD intein. Mini-inteins are highlighted in red text, support for bipartitions is indicated by the node shape, larger circles indicate strong support.



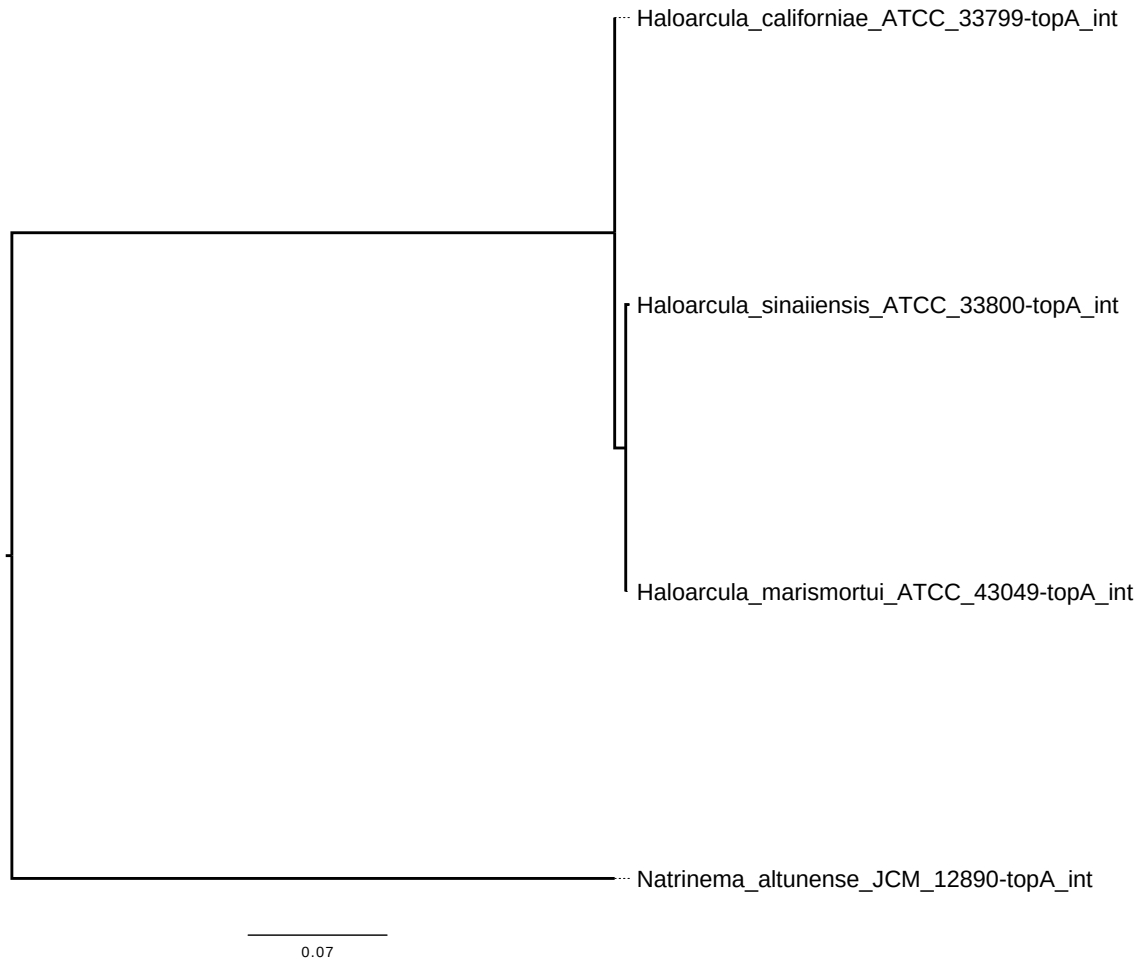
Supplementary Figure 5.8 Maximum likelihood phylogeny of rir1b intein. The node shape indicates support for bipartitions larger circles indicate strong support.



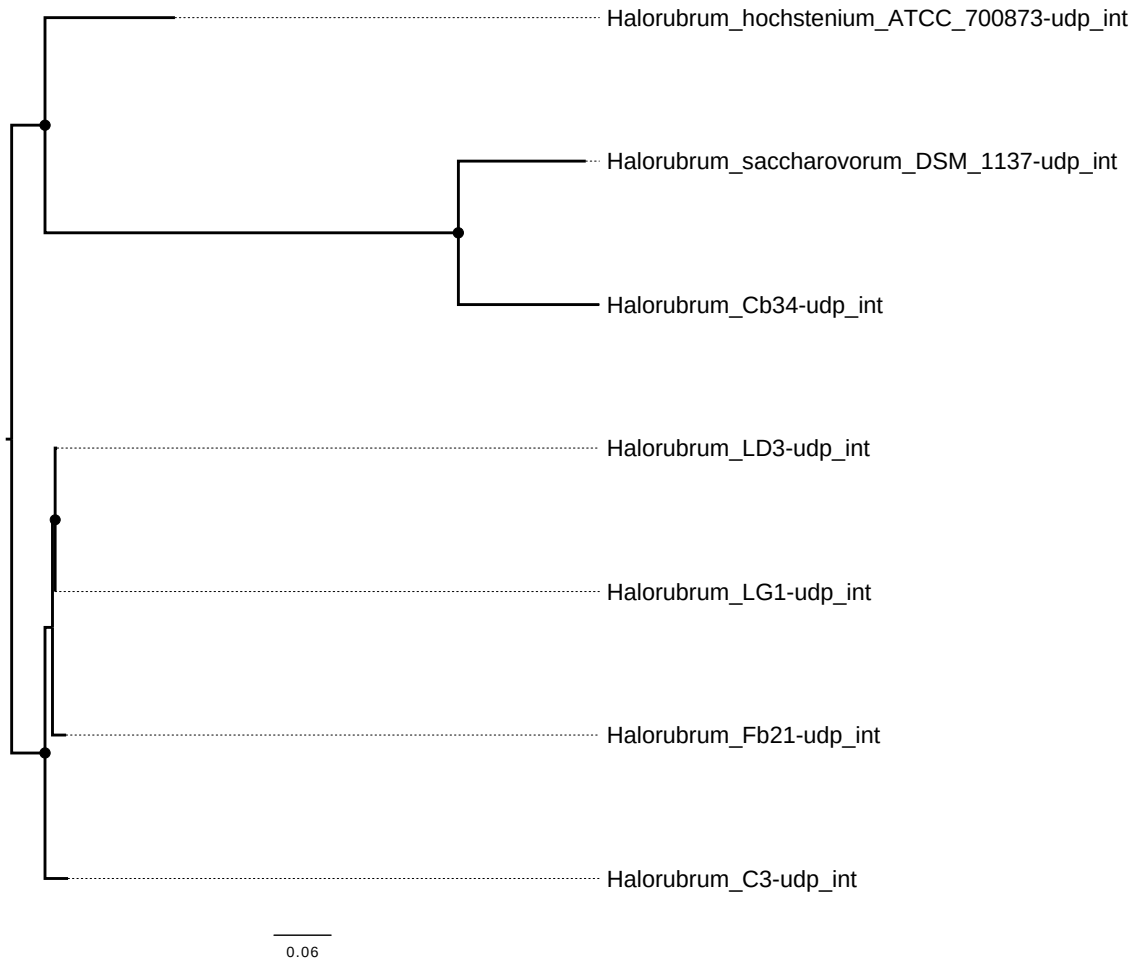
Supplementary Figure 5.9 Maximum likelihood phylogeny of rpolA intein, mini-inteins are highlighted in red text. The node shape indicates support for bipartitions larger circles indicate strong support.



Supplementary Figure 5.10 Maximum likelihood phylogeny of top6B intein. The node shape indicates support for bipartitions larger circles indicate strong support.

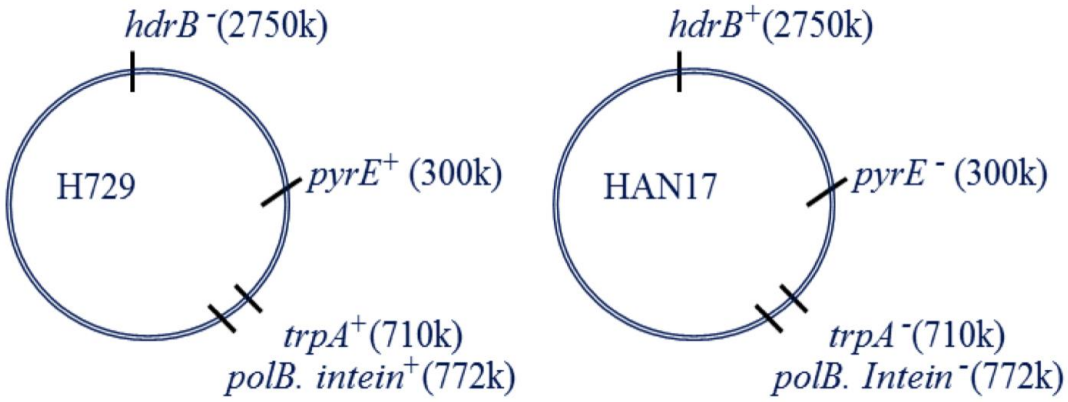


Supplementary Figure 5.11 Maximum likelihood phylogeny of topA intein. The node shape indicates support for bipartitions larger circles indicate strong support.

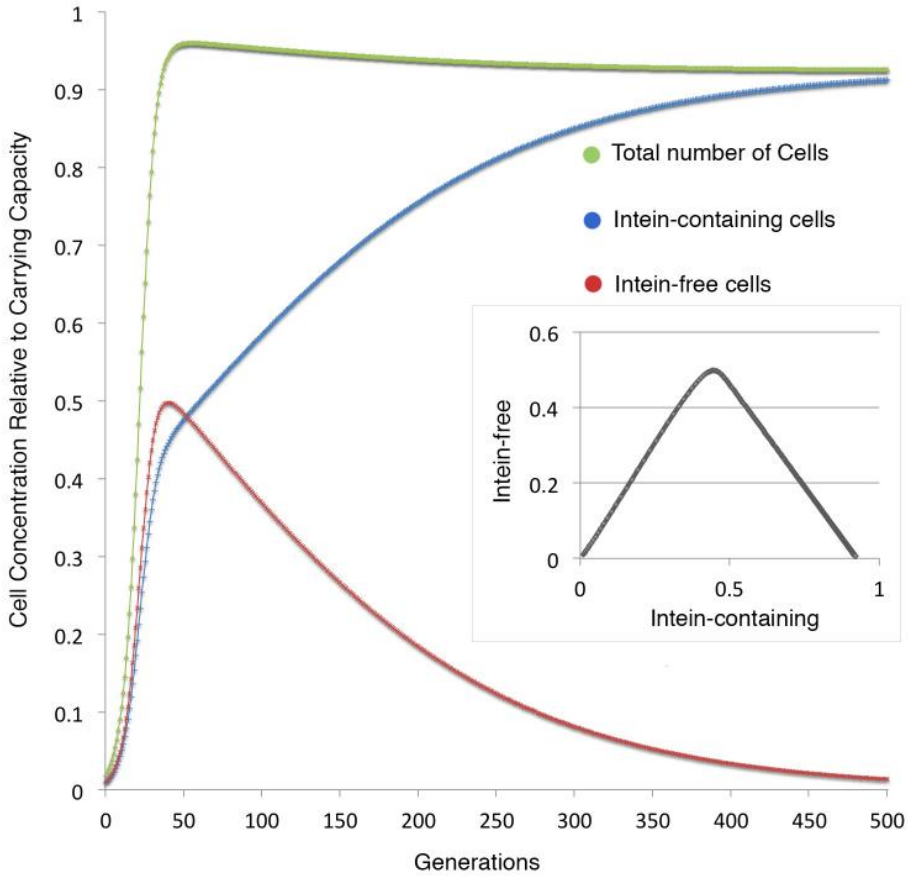


Supplementary Figure 5.12 Maximum likelihood phylogeny of udp intein. The node shape indicates support for bipartitions larger circles indicate strong support.

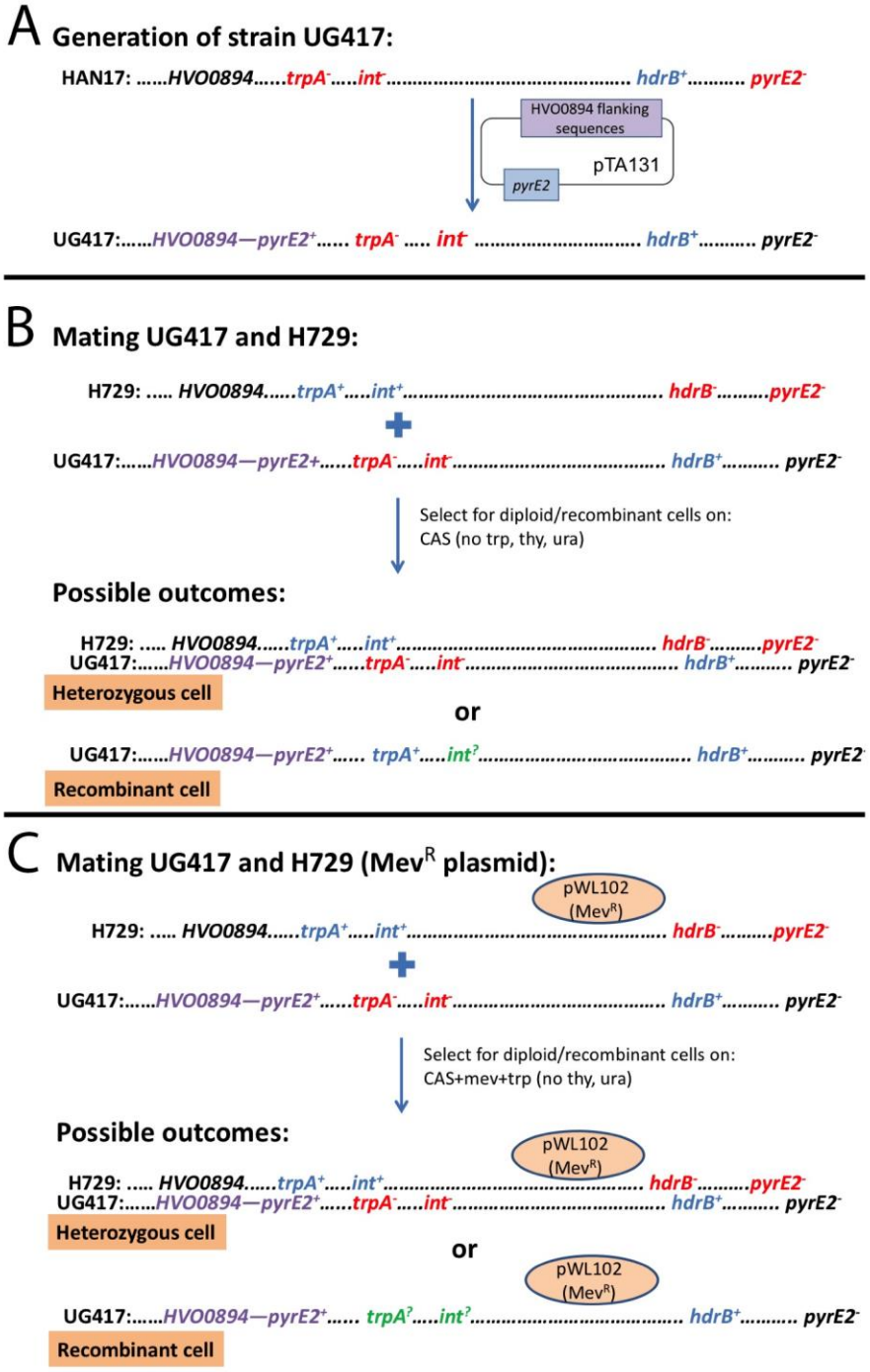
Chapter 6. Supplementary figures.



Supplementary Figure 6.1 Schematic of strains used for mating experiments with the location of selectable markers, as well as the location of the intein, and intein insertion site respectively.



Supplementary Figure 6.2 Schematic of intein dynamics in a population of cells, assuming that the impact of the intein on the host fitness also has an impact on the carrying capacity of the population. The fixation of the intein is slightly slower, but the results are the same as in chapter 6 figure 4.

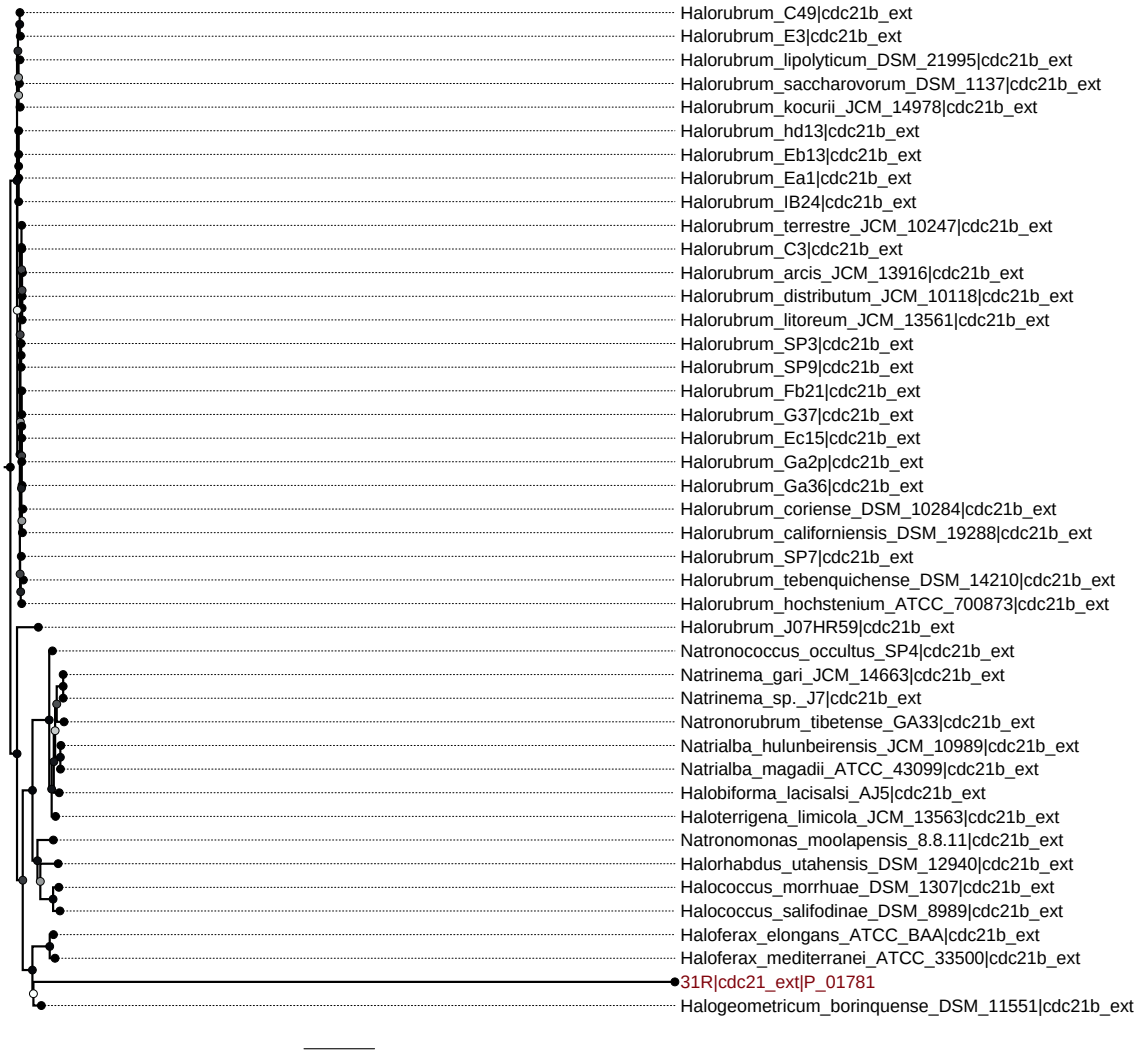


Supplementary Figure 6.3 Schematic of mating experiments and predicted outcomes.

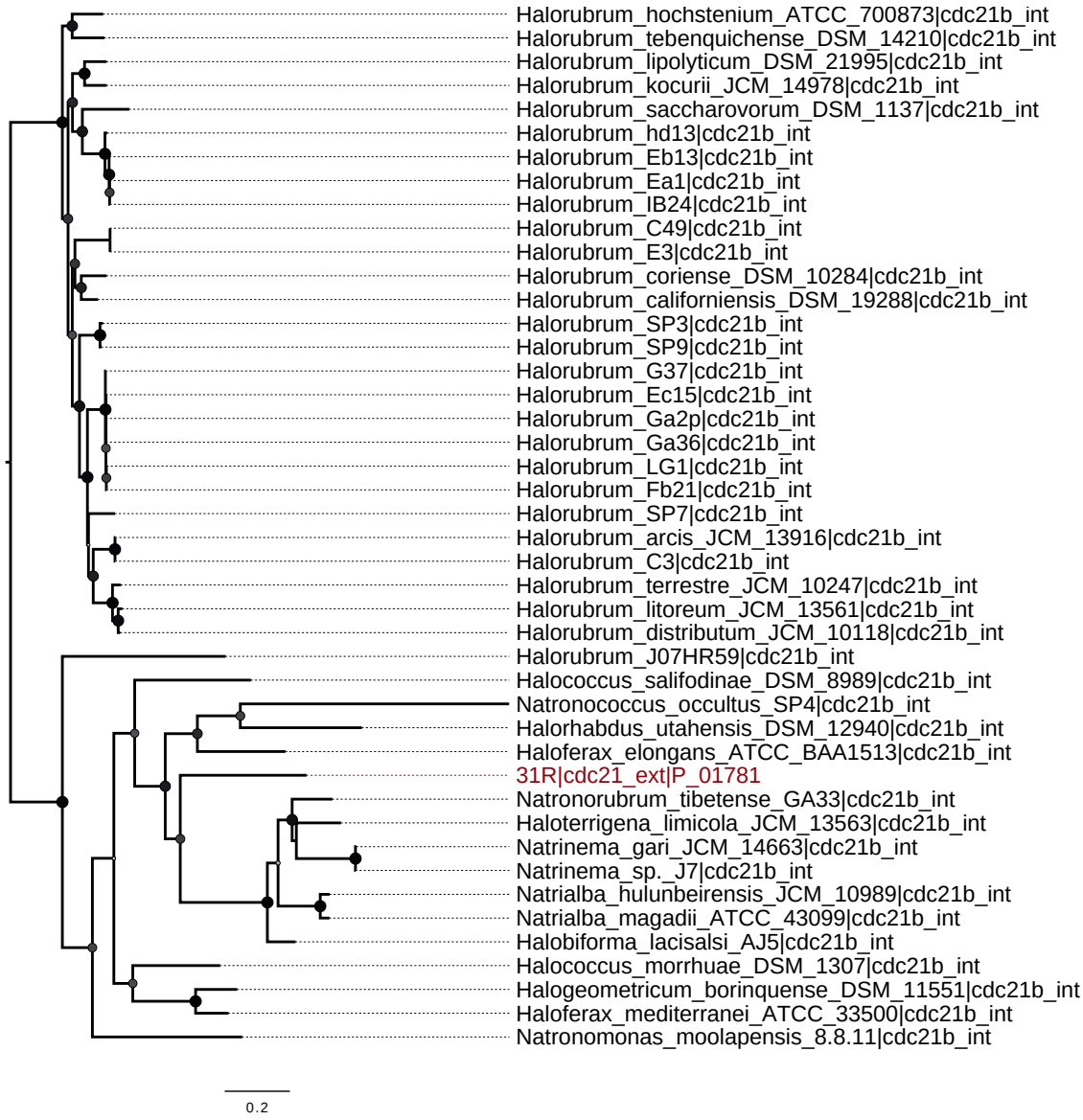
Intein presence	Sampling site/pool	Isolate No.
yes	Michmoret1	1
yes		2
yes	Michmoret2	3
yes		4
yes		5
yes		6
yes		7
no		8
no		9
no		10
no		11
no		12
no	Atlit1	13
yes	Atlit2	14
no		15
no	Atlit3	16
no		17
no	Atlit4	18
no		19
no		20
no	Atlit5	21
no	Atlit6	22
no		23
no		24
no		25
no		26
no		27
no		28
no		29
no		30
no	Atlit7	31
no		32
no		33
no		34
no		35
yes		36
yes	Atlit8	37
yes		38
yes		39
yes		40
yes		41
yes		42
yes		43
yes		44
yes		45
yes		46
no	Atlit9	47
no		48
no		49
no		50
no		51
yes	Zukim	52
yes		53
yes		54

Supplementary Figure 6.4 The distribution of the polB-c intein in twelve pools across three sampling sites.

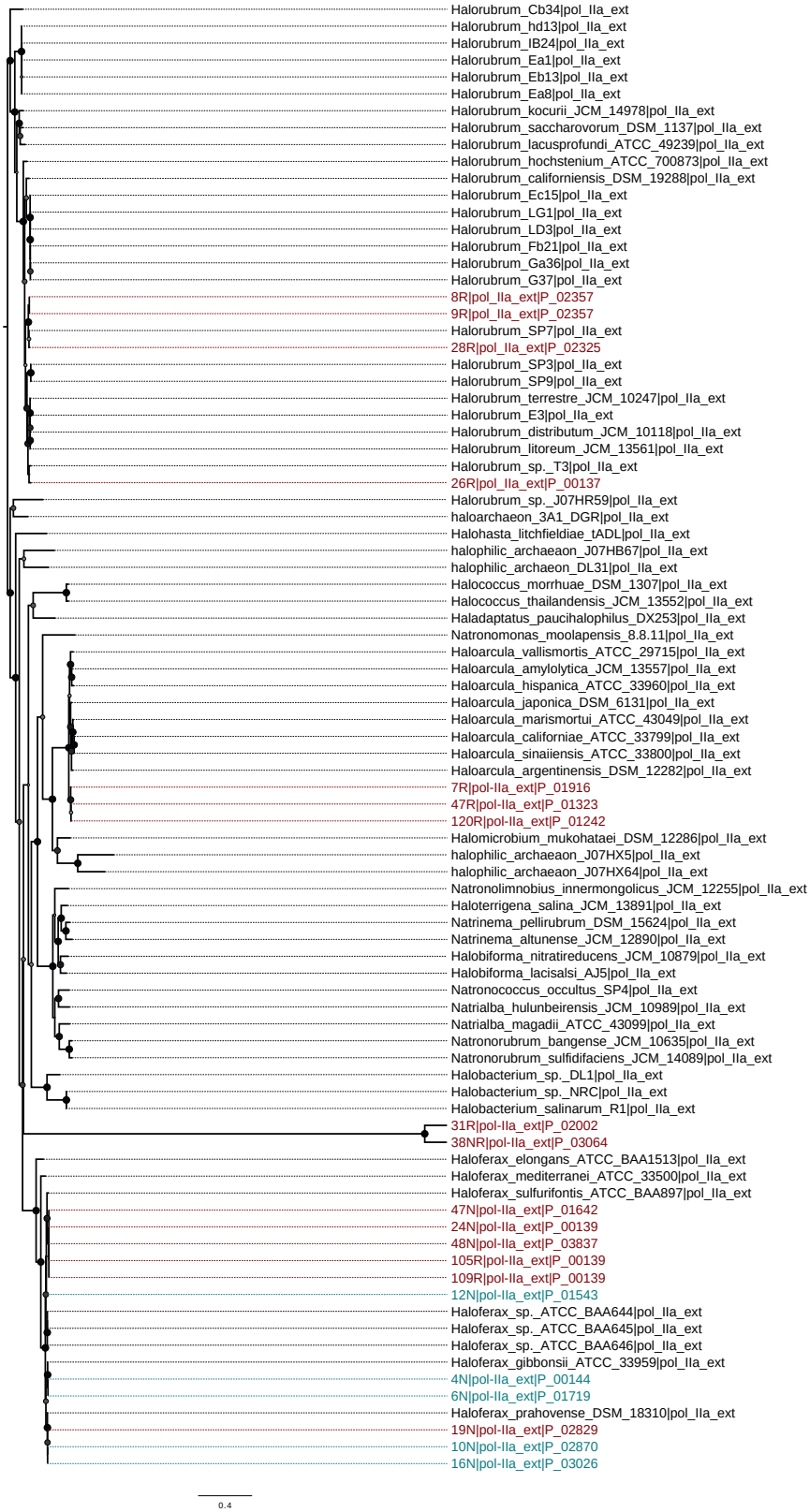
Chapter 8. Supplementary figures.



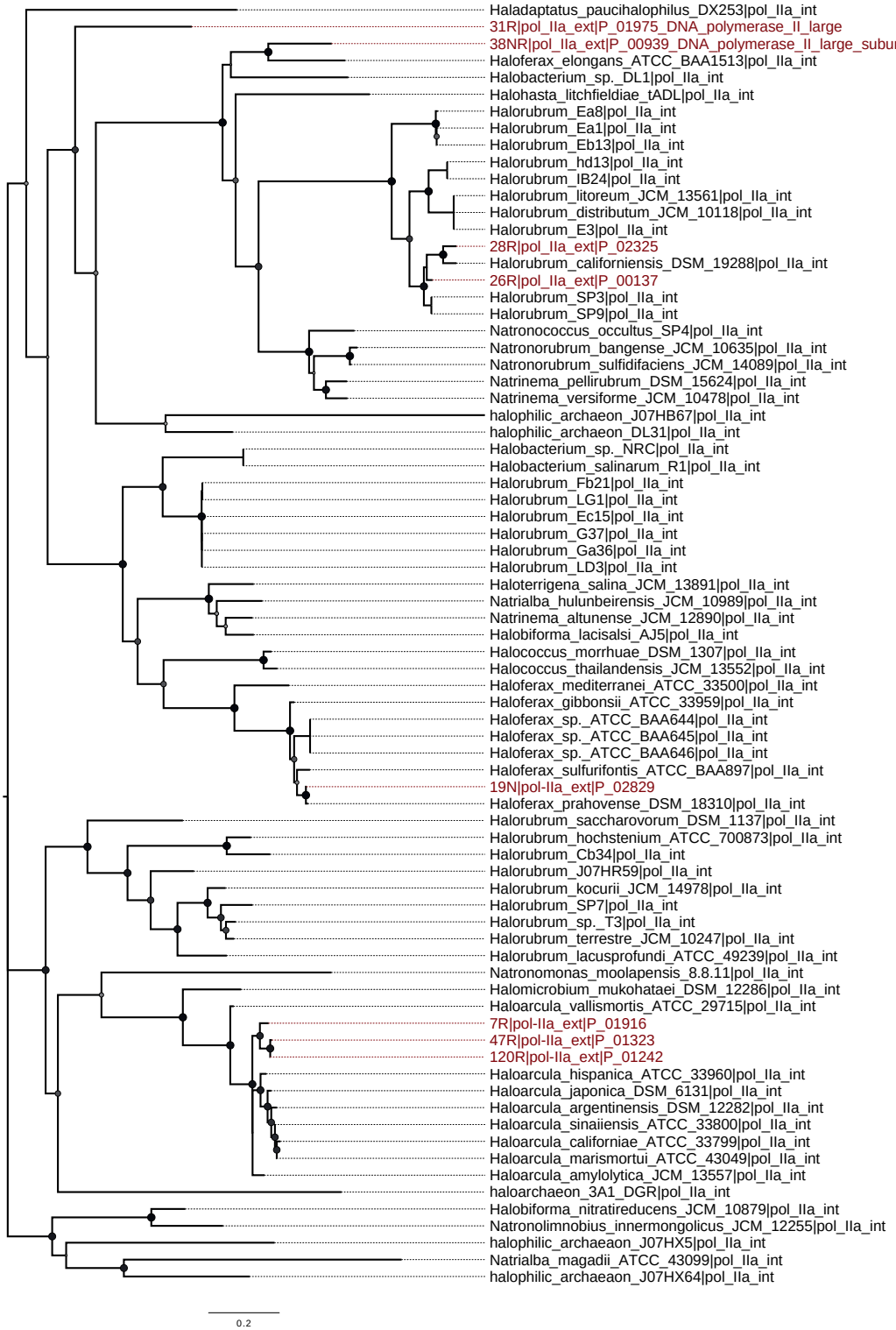
Supplementary Figure 8.1 Cell division control protein 21 extein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



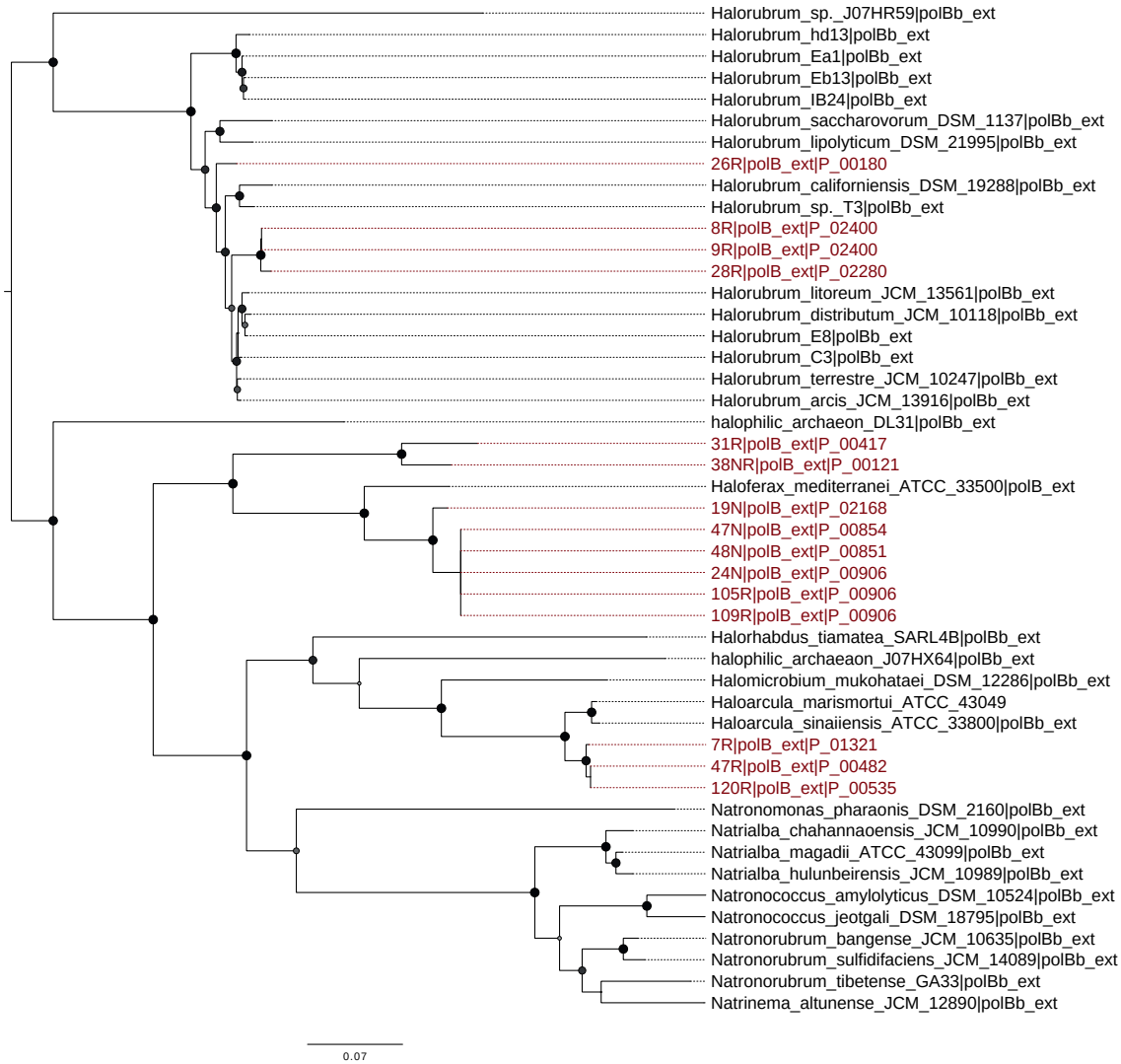
Supplementary Figure 8.2 Cell division control protein 21 intein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



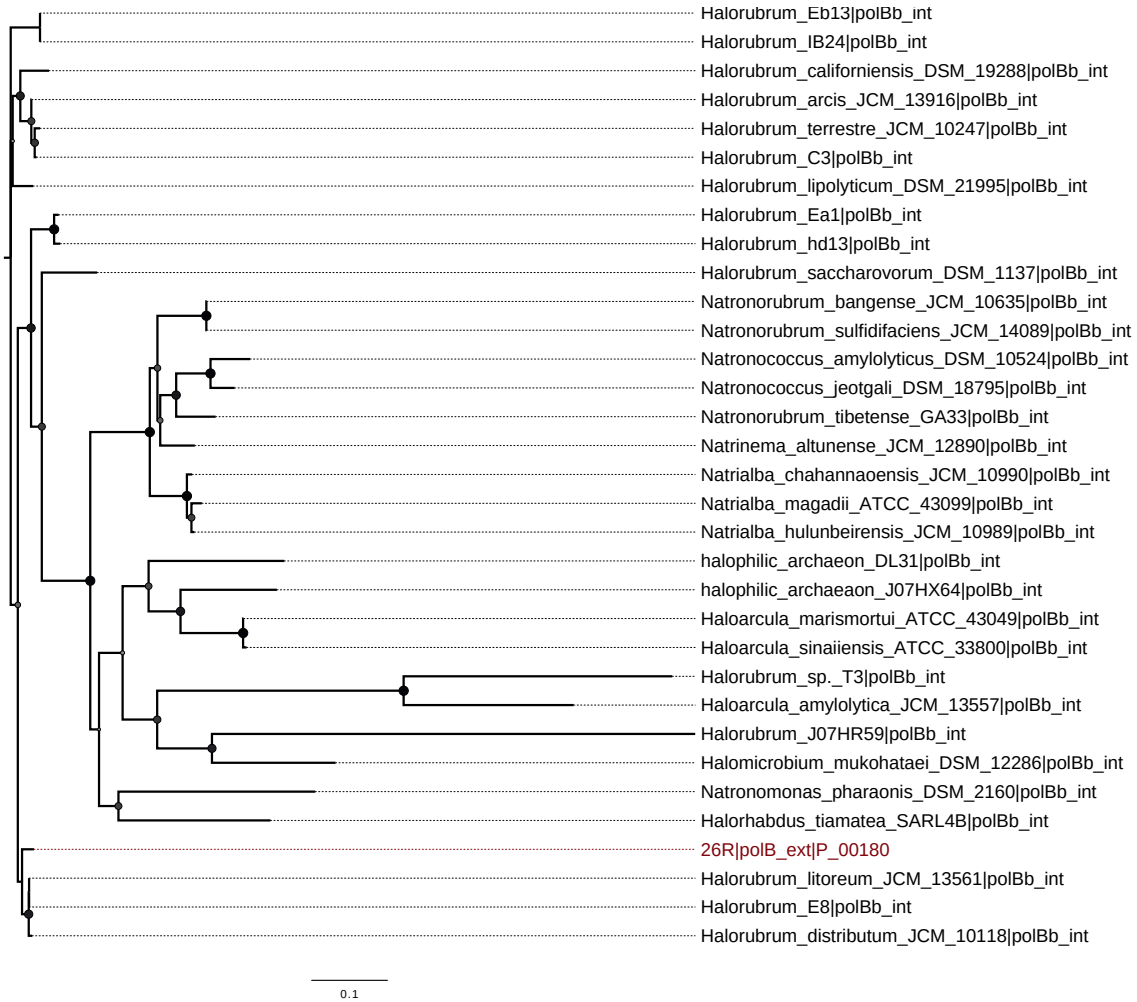
Supplementary Figure 8.3 DNA polymerase II extein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



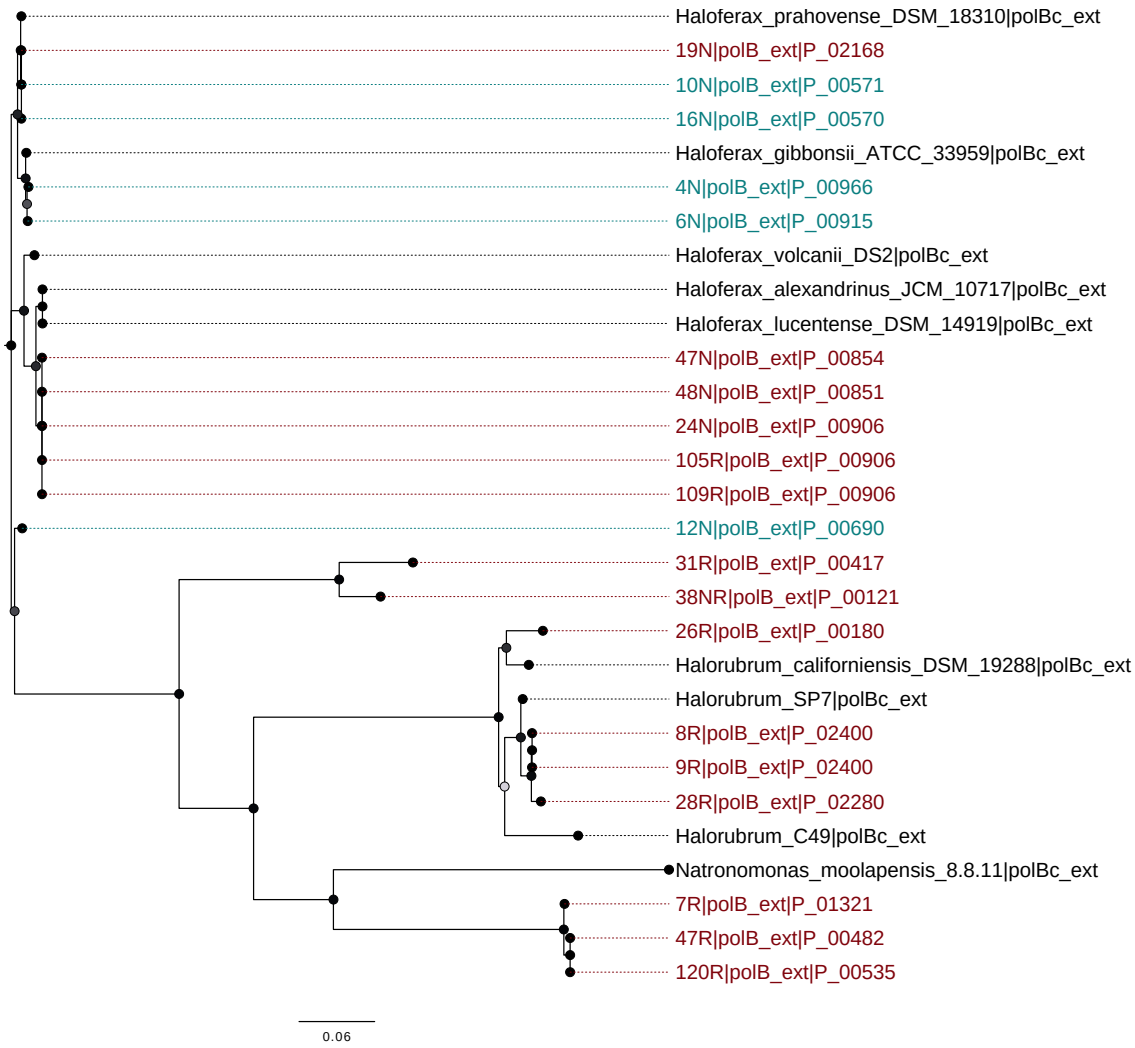
Supplementary Figure 8.4 DNA polymerase II intein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



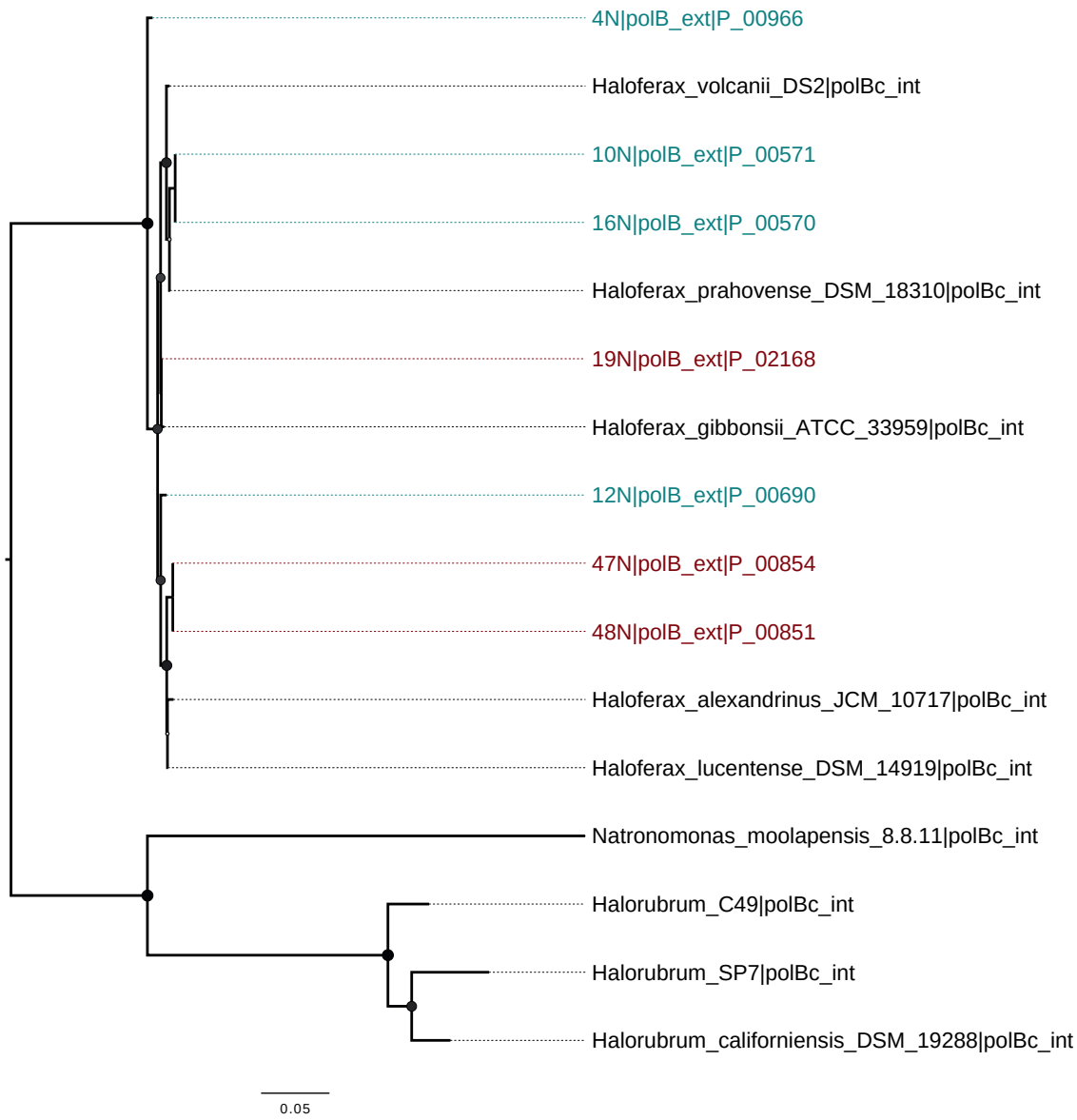
Supplementary Figure 8.5 DNA polymerase B extein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



Supplementary Figure 8.6 DNA polymerase B-b intein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



Supplementary Figure 8.7 DNA polymerase B extein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.



Supplementary Figure 8.8 DNA polymerase B-c intein maximum likelihood phylogeny. The node shapes indicate bipartition support values, larger and darker circles indicate higher support. Environmental isolates are highlighted by location, Atlit in red and Michmoret in blue.

Appendix B. Permission letters from the publishers

UCONN
UNIVERSITY OF CONNECTICUT

Shannon Soucy <shannon.soucy@uconn.edu>

RE: General Query: Other [#6350843]

JOURNALS PERMISSIONS <Journals.Permissions@oup.com> Thu, May 12, 2016 at 11:03 AM
To: Shannon Soucy <shannon.soucy@uconn.edu>

Dear Shannon,

RE. Kristen S. Swithers et al. Distribution and Evolution of the Mobile vma-ib Intein. Molecular Biology and Evolution (2013) 30 (12): 2676-2678

Thank you for your recent email requesting permission to reuse all of your article in a dissertation.

As part of your copyright agreement with Oxford University Press you have retained the right, after publication, to use all or part of the article and abstract, in the preparation of derivative works, extension of the article into a booklength work, in a thesis/dissertation, or in another works collection, provided that a full acknowledgement is made to the original publication in the journal. As a result, you should not require direct permission from Oxford University Press to reuse you article.

However, in line with the journal self-archiving policy, you may only include your **accepted'manuscript'PDF** in your thesis, and public availability must be delayed until **12'months** after first online publication in the journal. You should include the following acknowledgment as well as a link to the version of record. **Please'note inclusion'under'a'Crea7ve'Commons'license'or'any'other'open:access'license'allowing'onward'reuse'is prohibited.**

*This%s%%%re*copyedited,%author*produced%PDF%f%an%ar7cle%accepted%or%publica7on%n%*MBE *following%peer review.%The%defini7ve%publisher*authen7cated%ersion* [insert complete citation information here] *is%available online%at* [insert URL of published article here].

For full details of our publication and rights policy please see the attached link to our website:

http://www.oxfordjournals.org/our_journals/molbev/for_authors/mbe_author_self_archiving_policy.html

UCONN
UNIVERSITY OF CONNECTICUT

Shannon Soucy <shannon.soucy@uconn.edu>

Permission to Reprint Article

Radwa Ibrahim <radwa.ibrahim@hindawi.com> Mon, May 30, 2016 at 7:22 AM
To: Shannon Soucy <shannon.soucy@uconn.edu>

Dear Dr. Soucy,

Open-access authors retain the copyrights of their papers, and all open-access articles are distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided that the original work is properly cited.

Please feel free to contact me if you have further inquiries.

Best regards,

Radwa Ibrahim

--
****
Radwa Ibrahim
Editorial Office
Hindawi Publishing Corporation
http://www.hindawi.com/
****
[Quoted text hidden]

UCONN
UNIVERSITY OF CONNECTICUT

Shannon Soucy <shannon.soucy@uconn.edu>

Permission to Reprint an Article

Permissions@nature.com <Permissions@nature.com> Thu, May 26, 2016 at 4:44 AM
To: Shannon Soucy <shannon.soucy@uconn.edu>

Dear Shannon,

Thank you for contacting Nature Publishing Group. As an author, you have the right to use this manuscript and figures, as per the licence-to-publish you signed:

Ownership of copyright in the article remains with the Authors, and provided that, when reproducing the Contribution or extracts from it, the Authors acknowledge first and reference publication in the Journal, the Authors retain the following non-exclusive rights:

a) To reproduce the Contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

b) They and any academic institution where they work at the time may reproduce the Contribution for the purpose of course teaching.

c) To post a copy of the Contribution as accepted for publication after peer review (in Word or Tex format) on the Authors' own web site or institutional repository, or the Authors' funding body's designated archive, six months after publication of the printed or online edition of the Journal, provided that they also give a hyperlink from the Contribution to the Journals web site.

d) To reuse figures or tables created by them and contained in the Contribution in other works created by them.

The above use of the term 'Contribution' refers to the author's own version, not the final version as published in the Journal.

Kind regards

Melissa Rose

Permissions Assistant

SpringerNature

UCONN
UNIVERSITY OF CONNECTICUT

Shannon Soucy <shannon.soucy@uconn.edu>

Re: (DC): Permissions to Reprint for a Thesis

Frontiers Editorial Office <editorial.office@frontiersin.org> Thu, Jun 2, 2016 at 11:21 AM
To: shannon.soucy@uconn.edu

Dear Ms Soucy,

Thank you for your email.

Under the Frontiers Terms and Conditions, authors retain the copyright to their work. Furthermore, all Frontiers articles are Open Access and distributed under the terms of the Creative Commons Attribution License (CC-BY 3,0), which permits the use, distribution and reproduction of material from published articles, provided the original authors and source are credited, and subject to any copyright notices concerning any third-party content.

You can therefore reuse the article, given that you make sure to properly acknowledge the original source that you intend to reproduce. We recommend to use the format below:

"As originally published in Soucy SM, Fullmer MS, Papke RT and Gogarten JP (2014) Inteins as indicators of gene flow in the halobacteria. Front. Microbiol. 5:299. doi: 10.3389/fmicb.2014.00299"

I hope this answers your question. Please let me know if you have any other questions or concerns.

Kind Regards,

Damaris

Damaris Critchlow
Editorial Operations Specialist

Frontiers | London Office
www.frontiersin.org
WeWork, 1 Fore St
London, UK
Office T +44(0)7934464749

Registered in England (number 9952345)
Registered Office: Munro House, Portsmouth Road, Cobham, Surrey KT11 1PP, United Kingdom
Directors: Kamila Markram, Roger Biggs, Michael Kenyon

Loop | Twitter | Facebook

On Thu, Jun 2, 2016 at 3:55 PM, Frontiers in Microbiology Editorial Office <microbiology.editorial.office@frontiersin.org> wrote:

References

1. Raymond J, Zhaxybayeva O, Gogarten JP, Blankenship RE (2003) Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach. *Philos Trans R Soc Lond B Biol Sci* 358(1429):223–30.
2. Nelson-Sathi S, et al. (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci U S A* 109(50):20537–42.
3. Papke RT, Gogarten JP (2012) Ecology. How bacterial lineages emerge. *Science* 336(6077):45–6.
4. Andam CP, Gogarten JP (2011) Biased gene transfer in microbial evolution. *Nat Rev Microbiol* 9(7):543–55.
5. Rankin DJ, Rocha EPC, Brown SP (2011) What traits are carried on mobile genetic elements, and why? *Heredity (Edinb)* 106(1):1–10.
6. Dziewit L, et al. (2012) Insights into the transposable mobilome of Paracoccus spp. (Alphaproteobacteria). *PLoS One* 7(2):e32277.
7. Swithers KS, Senejani AG, Fournier GP, Gogarten JP (2009) Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol Biol* 9:303.
8. Novikova O, Topilina N, Belfort M (2014) Enigmatic distribution, evolution, and function of inteins. *J Biol Chem* 289(21):14490–7.
9. Demaere MZ, et al. (2013) High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. *Proc Natl Acad Sci U S A* 110(42):16939–44.
10. Swithers KS, Soucy SM, Gogarten JP (2012) The role of reticulate evolution in creating innovation and complexity. *Int J Evol Biol* 2012:418964.
11. Soucy S, Olendzenski L, Gogarten JP (2001) Orthologues, Paralogues and Horizontal Gene Transfer in the Human Holobiont. *eLS* (John Wiley & Sons, Ltd). doi:10.1002/9780470015902.a0005298.pub3.
12. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–7.
13. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27(2):221–4.
14. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–3.
15. Soucy SM, Fullmer MS, Papke RT, Gogarten JP (2014) Inteins as indicators of gene flow in the halobacteria. *Front Microbiol* 5:299.
16. Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 28(12):i283–91.
17. Gogarten JP, Hilario E (2006) Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol Biol* 6(1):94.
18. Burt A, Koufopanou V (2004) Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr Opin Genet Dev* 14(6):609–15.
19. Bonocora RP, Shub DA (2009) A Likely Pathway for Formation of Mobile Group I Introns. *Curr Biol* 19:223–228.

20. Hausner G, Hafez M, Edgell DR (2014) Bacterial group I introns: mobile RNA catalysts. *Mob DNA* 5:8.
21. Pietrokovski S (2001) Intein spread and extinction in evolution. *Trends Genet* 17(8):465–72.
22. Eddy SR, Gold L (1991) The phage T4 nrdB intron: a deletion mutant of a version found in the wild. *Genes Dev* 5(6):1032–41.
23. Foley S, Bruttin A, Brüssow H (2000) Widespread distribution of a group I intron and its three deletion derivatives in the lysin gene of Streptococcus thermophilus bacteriophages. *J Virol* 74(2):611–8.
24. Wikmark O-G, Einvik C, De Jonckheere J, Johansen S (2006) Short-term sequence evolution and vertical inheritance of the Naegleria twin-ribozyme group I intron. *BMC Evol Biol* 6(1):39.
25. Nomura N, et al. (2002) Heterogeneous yet similar introns reside in identical positions of the rRNA genes in natural isolates of the archaeon Aeropyrum pernix. *Gene* 295(1):43–50.
26. Goddard MR, Burt A (1999) Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci* 96(24):13880–13885.
27. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E (2002) Inteins: structure, function, and evolution. *Annu Rev Microbiol* 56:263–87.
28. Jeffares DC, Mourier T, Penny D (2006) The biology of intron gain and loss. *Trends Genet* 22(1):16–22.
29. Derr LK, Strathern JN (1993) A role for reverse transcripts in gene conversion. *Nature* 361(6408):170–173.
30. Barzel A, Obolski U, Gogarten JP, Kupiec M, Hadany L (2011) Home and away- the evolutionary dynamics of homing endonucleases. *BMC Evol Biol* 11:324.
31. Yahara K, Fukuyo M, Sasaki A, Kobayashi I (2009) Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer. *Proc Natl Acad Sci U S A* 106(44):18861–6.
32. Naor A, Gophna U Cell fusion and hybrids in Archaea: prospects for genome shuffling and accelerated strain development for biotechnology. *Bioengineered* 4(3):126–9.
33. Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U (2012) Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr Biol* 22(15):1444–8.
34. Rosenshine I, Tchelet R, Mevarech M (1989) The mechanism of DNA transfer in the mating system of an archaebacterium. *Science (80-)* 245(4924):1387–1389.
35. Naor A, Lazary R, Barzel A, Papke RT, Gophna U (2011) In vivo characterization of the homing endonuclease within the polB gene in the halophilic archaeon Haloferax volcanii. *PLoS One* 6(1):e15833.
36. Breuert S, Allers T, Spohn G, Soppa J (2006) Regulated polyploidy in halophilic archaea. *PLoS One* 1:e92.
37. Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *Am Nat* 138(6):1315–1341.
38. Allers T, Ngo HP, Mevarech M, Lloyd RG (2004) Development of additional selectable markers for the halophilic archaeon Haloferax volcanii based on the leuB and trpA genes. *Appl Env Microbiol* 70(2):943–53.
39. Blaseio U, Pfeifer F (1990) Transformation of Halobacterium halobium: development of

vectors and investigation of gas vesicle synthesis. *Proc Natl Acad Sci U S A* 87(17):6772–6776.
40. Youssef NH, Ashlock-Savage KN, Elshahed MS (2012) Phylogenetic diversities and community structure of members of the extremely halophilic Archaea (order Halobacteriales) in multiple saline sediment habitats. *Appl Environ Microbiol* 78(5):1332–44.
41. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
42. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–21.
43. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14(6):1188–90.
44. Grogan DW, Rockwood J (2010) Discontinuity and limited linkage in the homologous recombination system of a hyperthermophilic archaeon. *J Bacteriol* 192(18):4660–8.
45. Williams D, Gogarten JP, Papke RT (2012) Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol Evol* 4(12):1223–44.
46. Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MCJ (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol Biol Evol* 22(3):562–9.
47. Lange C, Zerulla K, Breuert S, Soppa J (2011) Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon Haloferax volcanii. *Mol Microbiol* 80(3):666–77.
48. Raghavan R, Hicks LD, Minnick MF (2008) Toxic introns and parasitic intein in Coxiella burnetii: legacies of a promiscuous past. *J Bacteriol* 190(17):5934–5943.
49. Hughes DP, Brodeur J, Thomas F (2012) *Host Manipulation by Parasites* (Oxford University Press).
50. Giraldo-Perez P, Goddard MR (2013) A parasitic selfish gene that affects host promiscuity. *Proc Biol Sci* 280(1770):20131875.
51. Michel B (2005) After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biol* 3(7):e255.
52. Boubriak I, et al. (2008) Transcriptional responses to biologically relevant doses of UV-B radiation in the model archaeon, Halobacterium sp. NRC-1. *Saline Systems* 4:13.
53. Strathern JN, et al. (1982) Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the MAT locus. *Cell* 31(1):183–92.
54. Fullmer MS, et al. (2014) Population and genomic analysis of the genus Halorubrum. *Extrem Microbiol* 5. doi:10.3389/fmicb.2014.00140.
55. Miyake T, Hiraishi H, Sammoto H, Ono B (2003) Involvement of the VDE homing endonuclease and rapamycin in regulation of the Saccharomyces cerevisiae GSH11 gene encoding the high affinity glutathione transporter. *J Biol Chem* 278(41):39632–6.
56. Callahan BP, Topilina NI, Stanger MJ, Roey P Van, Belfort M (2011) Structure of catalytically competent intein caught in a redox trap with functional and evolutionary implications. *Nat Struct Mol Biol* 18(5):630–633.
57. Topilina NI, et al. (2015) SufB intein of Mycobacterium tuberculosis as a sensor for oxidative and nitrosative stresses. *Proc Natl Acad Sci*:201512777.
58. Topilina NI, Novikova O, Stanger M, Banavali NK, Belfort M (2015) Post-translational environmental switch of RadA activity by extein–intein interactions in protein splicing.

Nucleic Acids Res 43(13):6631–6648.
59. Novikova O, et al. (2015) Intein Clustering Suggests Functional Importance in Different Domains of Life. *Mol Biol Evol*. doi:10.1093/molbev/msv271.
60. de Vienne DM, Giraud T, Gouyon P-H (2013) Lineage Selection and the Maintenance of Sex. *PLoS One* 8(6):e66906.
61. Nunney L (1989) The Maintenance of Sex by Group Selection. *Evolution (N Y)* 43(2):245–257.
62. Olendzenski L, Gogarten J (2009) Gene transfer: who benefits? *Horiz Gene Transf* 532:3–9.
63. Olendzenski L, Gogarten JP (2009) Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. *Ann N Y Acad Sci* 1178:137–145.
64. Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM (2007) Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol* 22(3):114–115.
65. Scalley-Kim M, McConnell-Smith A, Stoddard BL (2007) Coevolution of a homing endonuclease and its host target sequence. *J Mol Biol* 372(5):1305–19.
66. Gimble FS (2001) Degeneration of a homing endonuclease and its target sequence in a wild yeast strain. *Nucleic Acids Res* 29(20):4215–4223.
67. Cavicchioli R (2015) Microbial ecology of Antarctic aquatic systems. *Nat Rev Microbiol* 13(11):691–706.
68. Markowitz VM, et al. (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* 42(Database issue):D560–7.
69. Markowitz VM, et al. (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* 42(Database issue):D568–73.
70. Joshi NA FJ (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Available at: https://github.com/najoshi/sickle.
71. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–9.
72. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–9.
73. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–2.
74. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178–92.
75. Team Rs (2015) RStudio: Integrated Development for R.Studio, Inc., Boston, MA.
76. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 33(Web Server issue):W557–9.
77. Rambaut A (2015) FigTree:Graphical viewer of Phylogenetic Trees. Available at: http://tree.bio.ed.ac.uk/software/figtree/.
78. Zeitouni B, et al. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 26(15):1895–1896.
79. Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics* 2014. doi:10.1002/0471250953.bi1506s45.
80. Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer: Identification of Genomic

Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics* 45:15.6.1–11.
81. Bankevich A, et al. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 19(5):455–477.
82. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–5.
83. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–9.
84. Auch AF, von Jan M, Klenk H-P, Göker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* 2(1):117–134.
85. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22(24):3096–8.
86. Fang F, Ding J, Minin VN, Suchard MA, Dorman KS (2007) cBrother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics* 23(4):507–8.
87. Swithers KS, Soucy SM, Lasek-Nesselquist E, Lapierre P, Gogarten JP (2013) Distribution and Evolution of the Mobile vma-1b Intein. *Mol Biol Evol*. doi:10.1093/molbev/mst164.