RESEARCH ARTICLE

Co-invading symbiotic mutualists of *Medicago polymorpha* retain high ancestral diversity and contain diverse accessory genomes

Stephanie S. Porter\(^1,\ast,†\), Joshua J. Faber-Hammond\(^1\) and Maren L. Friesen\(^2,3,4\)

\(^1\)School of Biological Sciences, Washington State University, 14204 NE Salmon Creek Ave, Vancouver, WA 98686, USA, \(^2\)Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI, 48824, USA, \(^3\)Department of Plant Pathology, Washington State University, P.O. Box 646430 Pullman, WA 99164, USA and \(^4\)Department of Crop and Soil Sciences, Washington State University, P.O. Box 646420 Pullman, WA 99164, USA

\(\ast\)Corresponding author: Washington State University, School of Biological Sciences, 14204 NE Salmon Creek Ave, Vancouver, WA 98686, USA. Tel: (360) 546-9478; E-mail: stephanie.porter@wsu.edu

One sentence summary: The beneficial nitrogen-fixing symbionts that facilitate a plant’s biological invasion appear to have spread from Europe with their host, retain high levels of genomic diversity and harbor novel genes.

Editor: Angela Sessitsch

Stephanie S. Porter, http://orcid.org/0000-0002-1193-9705

ABSTRACT

Exotic, invasive plants and animals can wreak havoc on ecosystems by displacing natives and altering environmental conditions. However, much less is known about the identities or evolutionary dynamics of the symbiotic microbes that accompany invasive species. Most leguminous plants rely upon symbiotic rhizobium bacteria to fix nitrogen and are incapable of colonizing areas devoid of compatible rhizobia. We compare the genomes of symbiotic rhizobia in a portion of the legume’s invaded range with those of the rhizobium symbionts from across the legume’s native range. We show that in an area of California the legume *Medicago polymorpha* has invaded, its *Ensifer medicae* symbionts: (i) exhibit genome-wide patterns of relatedness that together with historical evidence support host-symbiont co-invasion from Europe into California, (ii) exhibit population genomic patterns consistent with the introduction of the majority of deep diversity from the native range, rather than a genetic bottleneck during colonization of California and (iii) harbor a large set of accessory genes uniquely enriched in binding functions, which could play a role in habitat invasion. Examining microbial symbiont genome dynamics during biological invasions is critical for assessing host-symbiont co-invasions whereby microbial symbiont range expansion underlies plant and animal invasions.

Keywords: co-invasion; microbial invasion; invasion genomics; rhizobia; legume; mutualism

INTRODUCTION

Biological invasions are a major driver of species diversity loss, leading to extinctions and substantial economic losses (Vitousek, D’Antonio and Loope 1997; Mack et al. 2000; Pimentel 2011; Lowry et al. 2013). Invasive plants and animals associate with diverse microbiota as they spread (van der Putten, Klironomos and Wardle 2007; Traveset and Richardson 2014). While pathogens amplified by invasive species have been well documented (Vitousek, D’Antonio and Loope 1997; Mack et al.
Genotypes in the native and invaded range, as occurs for acrnovelsymbiontsintheinvadedrange. In this case, hosts (Sullivan et al. 2014) can acquirenovelsymbiontsintheinvadedrange.Mesorhizobium braver (Lesins and Lesins 1979; Silva, Kan and Martinez-Romero 2001; Reeve, Chain and O’Hara 2010; Sugawara et al. 2013; Haskett et al. 2016). Medicago polymorpha originates from Mediterranean Europe and North Africa and is a globally distributed invader (Lesins and Lesins 1979; Silva, Kan and Martinez-Romero 2007). It is specialized on the rhizobium Ensifer medicae, and also rarely nodulates with Ensifer mellioti in its native range (Rome et al. 1996). This study provides a genome-wide perspective on rhizobium nodulating M. polymorpha at one reserve in its invaded range. We compare rhizobium genomes from this site to published genomes from the same species of rhizobium across the native range of its host to ask: (i) Do genome-wide patterns of relatedness suggest M. polymorpha symbionts in the sampled portion of the invaded range are novel chimeras, co-invading species, novel partner species or cosmopolitan generalist symbionts? (ii) Do population genomics support reductions in symbiont diversity in the invaded range, as would be predicted by a colonization bottleneck? And (iii) Do invaded range symbiont genomes harbor unique accessory genes that could have been acquired during adaptation to their new habitat?

**METHODS**

**Sample collection**

Seventeen Ensifer medicae strains were isolated from one root nodule from each of 17 field-collected Medicago polymorpha plants from the McLaughlin Natural Reserve in California, USA (38.8709975, -122.419128) (Table S1, Supporting Information), as described in Porter and Rice (2013).

**Genomic sequencing**

Californian isolates were grown in 10 mL tryptone yeast broth cultures at 30°C shaking at 200 rpm until dense, and then cells pelleted and frozen at ~80°C. DNA extraction and genomic library construction followed Dunham & Friesen (2013) with modifications as in Porter et al. (2016). Samples were quantified using a Qubit fluorometer (Thermo Fisher, Waltham, MA) and ~500 ng of DNA was used for genomic library construction. Libraries were sequenced in paired 76 bp format on an Illumina GAIIx at the University of Southern California.

Population genetic comparisons can distinguish among four invasion scenarios. First, familiar symbionts from the host’s native range could co-invade a new area with their host, as has occurred for a number of mycorrhizae (Hayashi et al. 2001; Dickie et al. 2010; Nuñez and Dickie 2014; Bogar, Dickie and Kennedy 2015; Hayward et al. 2015). Here, the same symbiont genotypes are present both ranges, but historical records indicate that symbionts were absent in the invaded range prior to host introduction (Pringle et al. 2009). Second, major genomic rearrangements (chimerism) and horizontal gene transfer (HGT) of genomic islands can give rise to evolutionary innovations that help microbes colonize novel environments (Ochman, Lawrence and Groisman 2000; Kuo and Ochman 2009; Litchman 2010; Darmon and Leach 2014). For example, European Lotus corniculatus colonized New Zealand in concert with the transfer of the symbiosis island from its European rhizobial partner, Mesorhizobium loti, into a non-nodulating native New Zealand Mesorhizobium (Sullivan et al. 1995, 1996). Third, hosts can switch partners to acquire novel symbionts and Mimosa pigra and their nitrogen-fixing symbionts (Lafay and Burdon 2006; Parker, Wurtz and Paynter 2006). Finally, if symbionts are cosmopolitan, the same symbiont genotypes are available to the host in novel and native ranges, as occurs in plants that associate with cosmopolitan arbuscular mycorrhizal fungi (Moora et al. 2011). Here, there would be no bottleneck as expected in co-invasions. In some cases, symbionts could invade first, spread on native alternative hosts (Wolfe et al. 2010; Bonito et al. 2011), and then facilitate the spread of invasive hosts. Understanding the prevalence of these mechanisms of symbiont invasion is critical for integrating microbes into invasion biology.

Leguminous plants comprise some of the most noxious invasive species (Daehler 1998; Richardson et al. 2000), but we are only beginning to understand the role of their rhizobium symbionts in invasion (Traveset and Richardson 2014; Simonsen et al. 2017; LaPierre et al. 2017). Rhizobia are horizontally transmitted bacteria that persist between hosts as soil saprobes (Narożna et al. 2015) and rhizobial limitation can reduce the spread of invasive plants (Parker 2001; Parker, Malek and Parker 2006; Simonsen et al. 2017). Plants house rhizobia intracellularly in root nodules where they provide reduced nitrogen in exchange for photosyntheses (Oldroyd 2013); the genes underlying these processes cluster on transmissible plasmids or islands (Calibet et al. 2001; Reeve, Chain and O’Hara 2010; Sugawara et al. 2013; Haskett et al. 2016). Medicago polymorpha is a globally distributed invasive species (Daehler 1998; Richardson 2000), but we are only beginning to understand the role of their rhizobium symbionts in invasion biology.
In total, 17 E. medicae genome assemblies from across the native range of M. polymorpha were downloaded from NCBI and MicroScope assembly databases (Epstein et al. 2012; Sugawara et al. 2013; Reeve et al. 2015). These strains originated from France, Italy, Greece, Jordan, Iraq and Syria, which span the Mediterranean region of Europe and the Middle East where M. polymorpha is native (Lesins and Lesins 1979). Ensifer mелиоти strain Rm1021 (Galibert et al. 2001), the sister species of E. medicae (Bailly et al. 2006, 2011; Sugawara et al. 2013), was used as an outgroup for phylogenetic analyses.

Assembly and annotation

Californian draft genomes were assembled using the A5 pipeline (Tritt et al. 2012), using the May 18th 2012 Linux x64 A5 version and run with default parameters. Assembly statistics, including genome size, number of contigs and average contig length were calculated using Prinseq v0.20.4 (Schmieder and Edwards 2011; Table S1, Supporting Information). All 35 genome assemblies (including E. mелиоти) were submitted to RAST v2.0 for functional annotation by RASTtk (Aziz et al. 2008).

To obtain gene ontology (GO) terms for consensus sequences for each orthogroup, we created consensus sequences for each orthogroup using the EMBASS v6.5.7 cons tool running on protein sequences with default parameters (Rice, Longden and Bleasby 2000). Consensus amino acids were called at a locus if it was present in at least 50% of orthogroup sequences. Consensus sequences were mapped and annotated with BLAST2GO v3.2 using the top five BLASTP hits from NCBI’s nr database (Ana Conesa 2008) and BLAST2GO’s embedded InterProScan search. This allowed us to extract and analyze top BLAST hits and GO information for 6858 out of the 12 138 genes in the pangenome.

Core and accessory genomic variation

Orthologs were identified using OrthoMCL (Wang et al. 2003) with the inflation value for MCL clustering at 1.5 and other parameters at default values. Copy numbers of each orthogroup by strain indicate the number of core vs. accessory and single copy vs. duplicate genes. Core genes were defined to have at least one ortholog copy in all E. medicae strains examined, while genes with variable presence/absence rates were considered accessory genes, excluding singletons. Singletons were defined as genes found in only one E. medicae strain. Additionally, we used the following definitions in enrichment tests: (i) private core genes are present in all strains in one geographic range, but accessory in the other range, while (ii) private accessory genes were accessory in one range and completely absent in the other. There were no orthogroups that were core in one range and completely absent the other.

To identify SNPs in the core genome, ClustalOmega v1.1.0 was used to align orthogroup amino acid sequences from all single-copy core genes using default parameters. Orthogroups with more than one copy in a strain were excluded from this analysis to maintain a balance in copy number for downstream statistical tests. Nucleotides were substituted for their respective translated protein sequences from each multiple sequence alignment using pal2nal.pl (http://www.bork.embl.de/pal2nal/). Nucleotide MSAs were sorted by replicon, and SNPs were then called for each set using parseSNPs.py (https://gist.github.com/peterk87/8409706). The SNP fasta output file was reformatted through custom scripts and PGDSpider v2.0.9.1 for use in STRUCTURE v2.3.0, Genepop v4.1.2 and SplitsTree v4.14.2 (Raymond and Rousset 1995; Huson 1998; Fritchard, Stephens and Donnelly 2000; Hubisz et al. 2009; Lischer and Excoffier 2012).

To determine the genomic organization of core and accessory variation in E. medicae, we used the finished WSM419 genome that consists of a chromosome (3.78 Mbp), two mega-plasmids: psmed01 (1.57 Mbp), psmed02 (1.25 Mbp) and an additional plasmid, psmed03 (0.22 Mbp) (Reeve, Chain and O’Hara 2010). Representative gene sequences from each orthogroup were mapped to WSM419 using default parameters in Bowtie2 v2.2.6 (Langmead and Salzberg 2012) and filtered to include only those with primarily unique loci in the genome (Bowtie2 mapping scores ≥ 30). Unmapped sequences are reported as unknown loci.

To determine whether Californian E. medicae have lost major portions of the genome relative to native range strains, we visualized the genome content using BlastRing (BRIC) v0.95 (Alikhan et al. 2011). The WSM419 E. medicae reference genome was divided into individual replicons and the 34 genome assemblies were BLASTed to this reference. Rings were organized by native versus invaded range for California for visualization.

Patterns of relatedness

To determine evolutionary relationships between E. medicae strains from the invaded and native host range, we built restricted maximum likelihood trees and neighbor nets for (i) SNPs in single-copy core genes in each replicon, and (ii) accessory gene content. Analyses used all 17 native and 17 invaded range strains and the sister species, E. mелиоти, as the outgroup. Trees were constructed in RAxML v8.0.6 (Stamatakis 2014) using concatenated single copy core gene sequences. For the chromosome, we concatenated three sets of 500 randomly selected genes of the 1831 total genes to check for possible tree disparities. For psmed01 and psmed02, we concatenated all 603 and 89 genes, respectively, and for psmed03, no genes were core. Nucleotide alignments converted from protein MSAs with pal2nal.pl (as described above) were concatenated using FASCONCAT v1.0 (Kück and Meusemann 2010). Accessory gene input data were in binary presence/absence format for 7878 orthogroups. The RAxML model used for nucleotide sequences was GTR+GAMMA and for accessory genes BINGAMMA, both with 1000 iterations and 100 bootstraps. We also built neighbor nets using SplitsTree v4.14.2 (Huson 1998) based on (i) SNPs segregating in the core genome, and (ii) accessory gene presence/absence.

Population genetic diversity and population structure

To assess whether there is evidence of a population bottleneck during invasion, we calculated expected heterozygosity (He) in each range. This analysis focused on all non-singleton SNPs in the core genome and used Genepop Software v4.1.2 (Raymond and Rouset 1995). Heterozygosity, He, was calculated for each polymorphic site using the allele frequency for each geographic range separately and for the total population. Changes in population-level He rates in California compared to the native range were compared via two-tailed T-tests using a Bonferroni corrected P-value of 0.0125 for four tests (each replicon and across genome). Private SNP alleles were tallied for each replicon.

Fst was calculated at each site in the core genome using the following formula: Fst = (Ht − Hs)/Ht. Ht represents He calculated for the total sampled population of 34 Ensifer strains and Hs is the average of individual He values from each of the two ranges. Fst values were averaged for each replicon and across the genome.
In order to define patterns of genetic diversity in the invaded range site, we ran STRUCTURE on the core genome SNPs and on accessory genome presence/absence on the 17 Californian strains, with the burnin period and MCMC repetitions set at 10 000 and 15 iterations, respectively, and K ranging from 1 to 12. We determined the best K by identifying peak values of ΔK when plotted against K (Evanno, Regnaut and Goudet 2005).

Unique gene content and functional implications

To determine whether Californian or native range E. medicae harbor unique genes, we examined private core and accessory pangenomic compartments by range. We also identified genes with variable frequencies between ranges by conducting two-tailed T-tests on the copy number for each orthogroup. We calculated the proportion of orthogroups with duplicates in at least one strain, number of accessory vs. core genes per genome for range comparisons and ran functional enrichment tests on pangenome subsets.

To examine the functional implications of private accessory genomic compartments, we conducted enrichment tests. We primarily conducted two analyses using the private accessory orthogroups from each range as the test sets and the remaining orthogroups in the pangenome as the reference set for each comparison. Singletons were excluded from this analysis since they may be partial sequences resulting from incomplete genome assembly and thus skew functional annotation results. Enrichment analysis on GO terms was implemented via Fisher’s Exact Test through BLAST2GO with a False Discovery Rate of 0.05. We also tallied top species BLAST hits for private accessory genes in each range to address whether range-exclusive genes have their origin as rare Ensifer genes or from HGT.

RESULTS

Assembly and annotation statistics

Draft genomes from the McLaughlin Reserve in the invaded range (excluding Str5) assembled into an average of 276 contigs with an average length of 48.5 kb for a total mean genome size of 6.9 Mb. Despite these similar metrics, it is possible that differences in sequencing depth (~20X vs ~10X coverage (Epstein et al. 2012)) resulted in rare sequencing errors being retained more often in invaded range strains. The Str5 assembly resulted in an unusually large number of contigs (2373), low average contig size (2.4 kb), and slightly smaller genome size (5.8 Mb), yet has a comparable number of RAST annotated features with other strains. Annotated genomes (all strains in both ranges) have an average of 7285 transcribed features distributed amongst 6878 orthogroups/genes. The average genome contains 6574 single-copy genes (Fig. 1A) and 304 genes that were duplicated at least once in the genome based on OrthoMCL assignments (Fig. S1, Supporting Information).

There are more single-copy orthogroups per strain in the invaded range than in the native range (a mean of 6845 vs. 6302; $\text{T}_{16} = 6.22, P = 1.22 \times 10^{-5}$), while the number of multi-copy orthogroups did not differ between ranges (a mean of 291 vs. 316; $\text{T}_{16} = -0.89, P = 0.39$) (Fig. 1A). In total, there are 12 138 unique orthogroups in pangenome of the selected strains of E. medicae (Table S2, Supporting Information). The mode value of gene copies per orthogroup is 34 (for 2905 orthogroups), primarily representing single copy genes found in all 34 strains (Fig. 1B). Excluding singletons, there is an abundance of rare accessory orthogroups (3780) found in five or fewer strains. Each strain harbors an average of 269 singleton (i.e. unique) orthogroups, with Str5 yielding the highest number of singletons (1534), many of which are partial gene fragments likely derived from assembly artifacts. Without singletons, Str5 showed a comparable number of genes with closely related strains; therefore, singletons from all strains were excluded from downstream analyses to retain only conservative gene calls.

Of the 12 138 orthogroups in the E. medicae pangenome, 73% of sequences mapped to a replicon in the WSM419 reference genome (Fig. 1C), but a portion of these yielded low mapping scores (<30) indicating they either had weak alignments or were repetitive sequences that mapped to multiple locations. A majority of sequences (67.3%) aligned to the reference genome with strong Bowtie2 mapping scores (>30), while 4098 genes (33.8%) were not confidently mapped to any molecule. A majority of private accessory genes (2752/3804; 72.3%) from both ranges remained unmapped. All 4260 orthogroups core to all 34 strains
were found in the mapped portion of the pangenome since they include an ortholog in WSM419.

Patterns of relatedness

Ensifer medicae in the sampled portion of the invaded range in California are primarily a derived subset of native range diversity based on patterns of relatedness at SNPs in core genes. Californian E. medicae appears to be comprised of three major clades (Clade 1, 2 and 3) based on phylogenies estimated with RAxML (Fig. 2), population groupings from STRUCTURE (Fig. S2, Supporting Information), neighbor nets estimated in SplitsTree (Figs S3 and S5, Supporting Information) and clustering of strains by accessory gene presence/absence data (Fig. S6, Supporting Information). Among the 32 genomes in our study with > 4 contigs, contig size (a measure of assembly quality) is weakly negatively correlated ($R = -0.53$, $T_{30} = -3.5$, $P < 0.01$) with branch length (Fig. S7, Supporting Information). Str5 has a low number of recovered reads and numerous small contigs, which can drive the overestimation of sequence divergence from other strains and likely underlie its long branch lengths. Due to the random nature of sequencing errors, these are unlikely to affect tree or network topology, which we confirmed by building neighbor networks and conducting STRUCTURE analysis both with and without Str5.

Within a clade, Californian strains tend to exhibit consistent patterns of close relatedness across replicons, with the exception of four strains (Str7, Str8, Str9 and Str18) that group in different clades for different replicons (Fig. 2). The relationship among clades, however, is different on different replicons (Fig. 2). The psmed01 phylogenetic tree and neighbor net both show Clade 1 as basal and closest to the E. melloti outgroup (Fig. S4, Supporting Information), which may point to a unique origin of the Clade 1 psmed01 plasmid. Neighbor nets with and without Str5 yielded similar patterns of relatedness among the other 33 strains, confirming that the inclusion of this low-coverage assembly has little impact on overall strain clustering (Figs S3 and S8, Supporting Information).

STRUCTURE indicates there are either two or three subpopulation clusters in the invaded range depending on the molecule examined (Fig. S2, Supporting Information). Among Californian strains, peaks of $K$ indicate four subpopulation clusters ($K = 4$) for the chromosome and accessory gene presence/absence, and two ($K = 2$) for psmed01 and psmed02. However, for the chromosome and accessory gene presence/absence, the four optimum clusters consist of the three expected clusters based on phylogenetic data, and the outlier Str5. STRUCTURE analysis repeated with Str5 removed best supports three subpopulations ($K = 3$) for the chromosome and accessory data, which correspond to the clades observed in the RAxML phylogenies and SplitsTree networks.

Major genomic loss events

California E. medicae do not appear to have lost major portions of the genome relative to native range strains. Visual comparisons of invaded and native range genomes using BlastRing alignments to the WSM419 genome revealed that all genomes from both ranges contain the majority of genes contained within the WSM419 chromosome, psmed01 and psmed02 replicons (Fig. 3). However, there are substantial differences in the genomic architecture of psmed03 between all strains and the WSM419 reference, which contains a large region of tandem duplications. Invaded range genomes contain 89% of the annotated genes in the WSM419 reference genome, while native range genomes contain an average of 88.4%. In comparison, psmed03 was highly variable with an average of 63.7% of the 157 WSM419 genes found in invaded range strains and 43.2% found in native range strains.

Population genetic diversity and population structure

Our results support a scenario in which E. medicae did not undergo a major population bottleneck during co-invasion with their plant host. The average expected $H_e$ from each locus differed between ranges (Bonferroni correction for four tests; $P < 0.0125$; Table 1). He is greater in the invaded than in the native range for loci located on the WSM419 chromosome (+30.9%; $T_{6,280} = 21.18$, $P = 3.05E−96$), WSM419 psmed01 (+140.8%; $T_{6,396} = 62.38$, $P < 0.0001$), and the genome overall (+71.6%; $T_{13,339} = 55.99$, $P < 0.0001$), although not for psmed02. The Fst between ranges is 0.0983 (Table 1). Psmed01 is the most divergent replicon ($Fst = 0.1301$) and the chromosome is least divergent with $Fst = 0.0671$. Invaded range genomes harbor more private alleles overall ($5511$), than do native range strains ($2445$). Compared to native range strains, invaded range strains have 100% and 211% more private alleles on the chromosome and psmed01 replicon, respectively, but 30% fewer private alleles on the psmed02 replicon. Divergence between native and invaded range strains appears to be largely driven by private alleles on psmed01 from Clade 1. He on this plasmid is twice that of native range strains, and Clade 1 harbors over half of invaded range private alleles.

To exclude SNPs resulting from sequencing errors, the metrics above are calculated after excluding singleton variable sites. This does not appear to overestimate $H_e$ and $Fst$ because differences in $H_e$ and private allele count are even greater when singleton SNPs, many of which are likely real mutations and rare alleles, are included (Table S3, Supporting Information).

Unique gene content and functional implications

A large set of core genes is shared between California and the native range. 4260 (35%) orthogroups are core, while 7878 (65%) are accessory. California strains contained more accessory genes (mean of 2877) than did native range strains (mean of 2359) ($T_{13} = 5.51$, $P = 4.75E−05$). In particular, all strains in Californian Clade 1 contain more accessory genes than any other strains (mean of 3276) (Fig. S9, Supporting Information).

Invaded range strains in California harbor more private accessory genes than do native range strains (Fig. 4). There are no private core genes in either range that were completely absent from the other range, suggesting that the core genomes are relatively stable. Ignoring the four mosaic genomes (Str7, Str8, Str9 and Str18), 857 of the 2319 total private accessory genes in the native range strains. 4260 (35%) orthogroups are core, while 7878 (65%) are accessory. California strains contained more accessory genes (mean of 2877) than did native range strains (mean of 2359) ($T_{13} = 5.51$, $P = 4.75E−05$). In particular, all strains in Californian Clade 1 contain more accessory genes than any other strains (mean of 3276) (Fig. S9, Supporting Information).

Invaded range strains in California harbor more private accessory genes than do native range strains (Fig. 4). There are no private core genes in either range that were completely absent from the other range, suggesting that the core genomes are relatively stable. Ignoring the four mosaic genomes (Str7, Str8, Str9 and Str18), 857 of the 2319 total private accessory genes in the native range strains. 4260 (35%) orthogroups are core, while 7878 (65%) are accessory. California strains contained more accessory genes (mean of 2877) than did native range strains (mean of 2359) ($T_{13} = 5.51$, $P = 4.75E−05$). In particular, all strains in Californian Clade 1 contain more accessory genes than any other strains (mean of 3276) (Fig. S9, Supporting Information).

Invaded range strains in California harbor more private accessory genes than do native range strains (Fig. 4). There are no private core genes in either range that were completely absent from the other range, suggesting that the core genomes are relatively stable. Ignoring the four mosaic genomes (Str7, Str8, Str9 and Str18), 857 of the 2319 total private accessory genes in the native range strains. 4260 (35%) orthogroups are core, while 7878 (65%) are accessory. California strains contained more accessory genes (mean of 2877) than did native range strains (mean of 2359) ($T_{13} = 5.51$, $P = 4.75E−05$). In particular, all strains in Californian Clade 1 contain more accessory genes than any other strains (mean of 3276) (Fig. S9, Supporting Information).

Invaded range strains in California harbor more private accessory genes than do native range strains (Fig. 4). There are no private core genes in either range that were completely absent from the other range, suggesting that the core genomes are relatively stable. Ignoring the four mosaic genomes (Str7, Str8, Str9 and Str18), 857 of the 2319 total private accessory genes in the native range strains. 4260 (35%) orthogroups are core, while 7878 (65%) are accessory. California strains contained more accessory genes (mean of 2877) than did native range strains (mean of 2359) ($T_{13} = 5.51$, $P = 4.75E−05$). In particular, all strains in Californian Clade 1 contain more accessory genes than any other strains (mean of 3276) (Fig. S9, Supporting Information).
Figure 2. Major clades of diversity present across replicons in the native range of *E. medicae* have representatives in the invaded range in California. Phylogeny estimated for *E. medicae* strains using restricted maximum likelihood in RAxML based 500 randomly selected single copy core genes from: (A) the chromosome, (B) the psmed01 replicon and (C) the psmed02 replicon. Strains from the invaded range in California are color coded by cluster as determined by STRUCTURE: blue, Clade 1; red, Clade 2; green, Clade 3; yellow, mosaic genomes. Relationship to the sister species, *E. meliloti*, is shown in inset figures. All nodes had bootstrap values >85 except for those shown in the figure.
Expected heterozygosity, private alleles, and FST calculated using alleles for single copy core gene SNPs within and between subpopulations.

<table>
<thead>
<tr>
<th># SNPs</th>
<th>Expected heterozygosity</th>
<th>Private alleles</th>
<th>FST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CA</td>
<td>NAT</td>
<td>Overall</td>
</tr>
<tr>
<td>Full genome</td>
<td>13 340</td>
<td>0.2844</td>
<td>0.1657</td>
</tr>
<tr>
<td>Chromosome</td>
<td>6281</td>
<td>0.2520</td>
<td>0.1924</td>
</tr>
<tr>
<td>psm001</td>
<td>6397</td>
<td>0.3240</td>
<td>0.1346</td>
</tr>
<tr>
<td>psm002</td>
<td>662</td>
<td>0.2099</td>
<td>0.2130</td>
</tr>
</tbody>
</table>

*Singleton alleles were excluded from these calculations to eliminate bias from potential sequencing errors. Calculations were made using all SNPs combined and those mapped to individual replicons in the WSM419 reference genome. No genes on psm003 were core to the full set and thus we could not identify SNPs from this molecule. CA: California, Nat: native range. *Statistically significant differences between native and invaded range at P < 0.0125 (Bonferroni corrected cutoff) as determined by two-tailed T-test.

DISCUSSION

Beneficial symbionts play a critical role in the invasion biology of exotic plants and animals (Richardson et al. 2000; Pringle et al. 2008; Himler et al. 2011; Traveset and Richardson 2014). However, the invasion genomics of most beneficial microbes remains a black box (van der Putten, Klironomos and Wardle 2007; Litchman 2010; Le Roux et al. 2017). Using population genomics, we show that in an area the legume M. polymorpha has invaded, its E. medicae symbionts: (i) likely co-invaded with their host from Europe into California, (ii) did not experience a genetic bottleneck during colonization of California, but rather show the introduction of the majority of native range genetic diversity and (iii) harbor a large set of accessory genes uniquely enriched in binding functions, which could play an adaptive role in colonization.

Symbiotic invasion scenarios

Our data support a scenario of co-invasion of E. medicae symbionts with their host, M. polymorpha. Approximately 89% of the gene content from the E. medicae WSM419 reference genome (Galibert et al. 2001; Sugawara et al. 2013) is present in symbiont genomes in both the native and invaded host range. This high similarity supports the conclusion, based on 16S or MLST sequences, that M. polymorpha associates with the native range E. medicae species in their invaded ranges in California (Silva, Kan and Martinez-Romero 2007; Porter and Rice 2013). In addition, Californian strains from the three major clades are largely derived from E. medicae clades present across M. polymorpha’s native range. This pattern varies across the genome: patterns of relatedness among SNPs in the chromosome and psm002 replicon suggest a scenario of a dispersal of E. medicae from M. polymorpha’s native range to California, while those in the psm001 replicon suggest Californian strains from Clade 1 could have a unique origin or derive from an unsampled native range E. medicae strain. In psm003, a highly variable small plasmid (213 kb in WSM419), patterns of gene presence and absence in Californian strains suggest that this replicon is derived relative to native range strains, though psm003 contains only accessory genes, precluding an SNP-based analysis of relatedness.

Our data do not support a scenario of genomic chimerism facilitating E. medicae invasion, which would require symbiosis genes transferring into a novel chromosomal background. This is in contrast to cases where genomic chimerism events within Mesorhizobium and Bradyrhizobium have facilitated range expansion of L. corniculatus (Sullivan et al. 1995; Sullivan and Ronson 1998), Biserula pleleus (Nandasena et al. 2007) and C. scoparius (Horn et al. 2014). In the M. polymorpha invasion, Californian E. medicae are derived from native range E. medicae across most of their genome. This suggests that Ensifer genomes have more stable symbiotic replicons than other genera (Soberón-Chávez and Nájera 1989; Segovia, Pinero and Palacios 1991; Laguerre and Bardin 1993; Bromfield, Barran and Wheatcroft 1995; Sullivan 2014). However, invasion genomics of most beneficial microbes remains a black box (van der Putten, Klironomos and Wardle 2007; Litchman 2010; Le Roux et al. 2017). Using population genomics, we show that in an area the legume M. polymorpha has invaded, its E. medicae symbionts: (i) likely co-invaded with their host from Europe into California, (ii) did not experience a genetic bottleneck during colonization of California, but rather show the introduction of the majority of native range genetic diversity and (iii) harbor a large set of accessory genes uniquely enriched in binding functions, which could play an adaptive role in colonization.

ENSIFER species (Fig. S11, Supporting Information), which could indicate many of these genes are rare orthologs or paralogs of genes that have been sequenced in strains of E. mediceae or E. meliloti. A small portion of genes have stronger homology with genes from non-Ensifer species and may have been acquired through HGT. Californian private accessory genes exclusive to invaded Clade 1 are enriched for DNA recombination, transposase activity, DNA mediated transposition, transposition and DNA metabolic process (Table S4, Supporting Information). Invasive range Clade 2 and 3 private accessory genes did not yield enriched GO terms.
congeneric alfalfa (*Medicago sativa*), which nodulates with *E. medicae*, show North American soils initially lacked compatible rhizobia (*Coburn 1907; Westgate 1908; Wing 1912*). Thus, we conclude that *E. medicae* has co-invaded with *M. polymorpha*, possibly along with other compatible medic hosts. Co-invasions are a common pattern of legume-rhizobium invasion: *Mimosa* species co-invade with their beta-proteobacterial symbionts (*Chen et al. 2005; Andrus, Andam and Parker 2012; Klonowska et al. 2012; Gehlot et al. 2013*) and *Dipogon lignosus* co-invades with *Burkholderia* symbionts (*Liu et al. 2014*). Host-symbiont co-invasions may be common among diverse hosts as exemplified by the stinkbug, *Mecapota cibaria*, which has co-invaded with the nutritional symbiont, *Ishikawella capulata* (*Brown et al. 2014*).

**Population genomics of host-microbial symbiont co-invasion**

Invaded range populations often exhibit low genetic diversity due to population bottlenecks at introduction (*Druligosch and Parker 2008; Barrett 2015; Bock et al. 2015*). For example, an invasive mycorrhiza, *Amanita phalloides*, exhibits lower diversity and fewer private alleles in the portion of its range where it is exotic (*Pringle et al. 2009*). In contrast, we detected higher levels of diversity in invaded range *E. medicae* than in native range *E. medicae*, despite comparing a continent-wide native range *E. medicae* pangenome, to a pangenome sampled from a single reserve in the invaded range. Compared to *E. medicae* from *M. polymorpha*s native range, the chromosome and *psmed01* replicon from invaded range *E. medicae* exhibit higher heterozygosity at SNP loci and more private alleles. Much of this high diversity and high number of private alleles is driven by the *psmed01* replicon, whereas the *psmed02* replicon shows no significant difference in heterozygosity between ranges. This is surprising because *psmed02*, the replicon with the most symbiosis-related genes (*Galardini et al. 2011; Sugawara et al. 2013*), is considered a hotbed of recombination, gene gain and loss, and rapid evolution (*Bailly et al. 2011*), as is the *psymA* homolog in the sister species, *E. meliloti* (*Giuntini et al. 2005*). Our finding of increased nucleotide and private allele diversity in *psmed01* and the chromosome, but not *psmed02*, in the Californian invaded range could result from selective forces acting on introduced variation, multiple introductions of highly diverse lineages from the native range, or de novo mutation. Purifying selection in native range *E. medicae* is thought to be weaker on the megaplasmids than on the chromosome (*Epstein et al. 2012*) possibly due to lower rates of plasmid gene expression (*Morrow and Cooper 2012*). Our findings suggest that different patterns of recombination and/or selection could underlie evolution in the native and invaded range.

While *E. medicae* harbors an order of magnitude less genotypic diversity than rhizobia such as *Rhizobium leguminosarum* (*Epstein et al. 2012; Velázquez et al. 2015*), using whole-genome analysis we were able to identify three clades of deep diversity that were present in our invaded range sample. We also found evidence for the lack of co-transmission across replicons, reflecting recombination or HGT, similar to that previously observed for *Ensifer* (*Bailly et al. 2011, Epstein et al. 2012*). *Ensifer medicae* nodulating another invasive medic, *Medicago lupulina*, in the Eastern USA contained even lower genetic diversity than we found evidence for the lack of co-transmission across replicons, reflecting recombination or HGT, similar to that previously observed for *Ensifer* (*Bailly et al. 2011, Epstein et al. 2012*). *Ensifer medicae* nodulating another invasive medic, *Medicago lupulina*, in the Eastern USA contained even lower genetic diversity than we identified (*Harrison et al. 2017*). A longer and more mixed history of Spanish and English colonization (*Spira and Wagner 1983; De Haan and Barnes 1998*), combined with a Mediterranean climate could result in more diverse *E. medicae* in California than in

---

**Figure 4.** Comparison of *E. medicae* orthogroup attributes from pangenomes in the invaded vs. native range of the symbiosis. Accessory genes are those not found in all strains, while core genes are found to have at least one copy per strain. ‘Private Accessory’ is defined as orthogroups with at least one gene copy in one range but no copies in the other. ‘Private Core’ is defined as orthogroups with 100% allele frequency in one range but not the other.

**Table 2.** Orthologs at significantly higher frequencies in strains from either geographic range based on two-tailed T-test per orthogroup.

<table>
<thead>
<tr>
<th>Significance threshold</th>
<th>Higher frequency in California invaded range</th>
<th>Higher frequency in native range</th>
</tr>
</thead>
<tbody>
<tr>
<td>P &lt; 0.05</td>
<td>1480</td>
<td>498</td>
</tr>
<tr>
<td>P &lt; 0.01</td>
<td>761</td>
<td>162</td>
</tr>
<tr>
<td>Bonferroni corrected</td>
<td>18</td>
<td>10</td>
</tr>
</tbody>
</table>
the Eastern USA. Alternatively, *M. lupulina* may associate with a smaller subset of *E. medicae* genotypes than does *M. polymorpha* (Brockwell and Hely 1966).

**Novel microbial genomic content in the invaded range**

The set of *E. medicae* strains from a single site in California harbor more accessory genes than those from sites spanning Europe and North Africa. Gene gain and loss play a major role in *E. medicae* population genomics (Bailly et al. 2011; Epstein et al. 2012; Sugawara et al. 2013; Epstein, Sadowsky and Tiffin 2014) and 65% of *E. medicae* orthologs were accessory genes in our study. California strains harbor more private accessory genes and have more accessory genes at higher frequency than the native range strains. While the majority of *E. medicae* accessory genes appear to be deleterious and subject to purifying selection in the native range (Epstein, Sadowsky and Tiffin 2014), novel conditions in California could drive selection for a greater number of adaptive accessory genes, or make *E. medicae* more vulnerable to the acquisition of deleterious accessory genes (Lee and Marx 2012; Epstein, Sadowsky and Tiffin 2014).

Compared to genes shared with native range strains, the California private accessory genome is enriched in loci with binding functions, which suggests that this molecular function could play a role in invasion. While most accessory genes share sequence identity with previously sequenced *Ensifer* strains, a subset do not and may have been acquired horizontally from other genera. Across the native range, private accessory genes are primarily enriched in transposon-related genes and other repetitive elements. The set of genes exclusive to Clade 1 strains in California are also enriched for transposon-related GO terms, including recombination, transposition and DNA metabolic processes. Clade 1 strains have the largest accessory genomes, indicating that gene gain and loss may play an especially large role in this lineage.

**CONCLUSIONS**

Our data support a host-symbiont co-invasion scenario for *M. polymorpha* and *E. medicae*, as opposed to invasion via novel microbial chimeras, novel microbial partner species or cosmopolitan generalist symbionts. Co-invading symbionts at a single reserve in California have greater genome-wide diversity than do symbionts from across the host's native range, suggesting no symbiont genetic bottleneck. In fact, compared to native range symbionts, invaded range symbionts harbor more accessory genes, which are uniquely enriched in binding molecular functions and could underlie novel adaptations. There is great potential for microbial genomics to shed light on symbiont invasion biology, which is a critical facet of the invasion biology of host plants and animals.

**SUPPLEMENTARY DATA**

Supplementary data are available at FEMSEC online.

**ACKNOWLEDGEMENT**

The authors thank J.P. Dunham for assistance with preparing genomic libraries, A. Garoutte for bioinformatics processing of raw sequence reads, B. Epstein for providing genomic information for some native range *E. medicae* strains, the authors of the publications on native range *E. medicae* genomes, and the editor and two anonymous reviewers for suggestions that improved the manuscript. The authors have no conflicts of interest to declare.

**Data Accessibility.** All sequence data are available under the NCBI BioProject PRJNA352431. The Whole Genome Shotgun sequences have been deposited at DDBJ/EMBL/GenBank under the accessions NBUA00000000-NBUQ00000000.

**FUNDING**

This work was funded by NSF DEB-1355216 to S.S. Porter and NSF DEB-1354878 to M.L. Friesen.

**Conflict of interest.** None declared.

**REFERENCES**


Coburn FD. *The Book of Alfalfa*. Orange Judd Company, 1907.

Conflict of interest. None declared.

**References**


Coburn FD. *The Book of Alfalfa*. Orange Judd Company, 1907.


Oldroyd GED. Speak, friend, and enter: signalling systems that establish of legumes outside their native range at a global scale. Nat Commun 2017;8, DOI: 10.1038/ncomms14790.


Simonsen AK, Dinnage R, Barrett LG et al. Symbiosis limits establishment of legumes outside their native range at a global scale. Nat Commun 2017;8. DOI: 10.1038/ncomms14790.


