MR. HUGH WHITE (Orcid ID : 0000-0002-3537-0832)

Article type : Original Article

Signatures of selection in core and accessory genomes indicate different ecological drivers of diversification among *Bacillus cereus* clades

Hugh White[1], Michiel Vos[2], Samuel K. Sheppard[3], Ben Pascoe[3] & Ben Raymond[1]

1. Corresponding Author: Centre for Ecology and Conservation, University of Exeter, Penryn campus, Penryn, TR10 9FE, UK. Corresponding author: University of Exeter, Penryn, TR10 9FE, UK. Tel: +44(0)7804109867; E-mail: hw399@exeter.ac.uk
2. European Centre for Environment and Human Health, University of Exeter Medical School, Environment and Sustainability Institute, Penryn Campus, TR10 9FE, United Kingdom
3. Milner Centre for Evolution, Department of Biology & Biotechnology, University of Bath, Claverton Down, Bath, UK

Competing Interests

The authors declare no competing interests

Abstract

Bacterial clades are often ecologically distinct, despite extensive horizontal gene transfer (HGT). How selection works on different parts of bacterial pan-genomes to drive and maintain the emergence of clades is unclear. Focussing on the three largest clades in the diverse and well-studied *Bacillus cereus sensu lato* group, we identified clade-specific core genes (present in all clade members) and then used clade-specific allelic diversity to identify genes under purifying and diversifying selection. Clade-specific accessory genes (present in a subset of strains within a clade) were characterized as being under selection using presence/absence in specific clades. Gene ontology analyses of genes under selection revealed that different core genes and gene functions were enriched in different clades. Furthermore, some gene functions were enriched only amongst clade-specific core or accessory genomes. Genes under purifying selection were often clade-specific, while genes under diversifying selection showed signs of frequent HGT. These patterns are consistent with different selection pressures acting on both the core and accessory genomes of different clades and can lead to ecological divergence in both cases. Examining variation in allelic diversity allows us to uncover genes under clade-specific selection, allowing ready identification of strains and their ecological niche.

Key words: *Bacillus cereus sensu lato*, niche differentiation, psychrotolerance, evolutionary genomics, horizontal gene transfer, comparative genomics

Introduction

Bacterial strains often appear grouped together in distinct phylogenetic clusters, or "clades", despite frequent homogenising horizontal gene transfer (HGT) (Buckee et al., 2008; Fraser, Hanage, & Spratt, 2007; Schloss & Handelsman, 2004). Although uncovered by methods that are blind to ecology (Carroll, Wiedmann, & Kovac, 2020; Guinebretière et al., 2008; Priest, Barker, Baillie, Holmes, & Maiden, 2004), these clades are often ecologically distinct from each other, both in phenotype and genome

content (Cohan, 2016; Hanage, Fraser, & Spratt, 2006). How distinct bacterial phylogenetic clades appear is not fully understood (Doolittle & Papke, 2006). A key question in the debate whether ecological differentiation is determined primarily by selection on the core genome (genes shared by all strains within a clade) or the accessory genome (genes shared only by a subset of those strains) (Maistrenko et al., 2020; McInerney, McNally, & O'connell, 2017; Tettelin, Riley, Cattuto, & Medini, 2008).

Selection in bacteria can be divided into three main categories: purifying selection, which removes deleterious alleles from a population; diversifying selection, which increases allelic diversity when rare alleles confer an advantage (McNally et al., 2019; Molina & Van Nimwegen, 2008) and directional selection, where alleles of genes are replaced by fitter variants (Cohan, 2016). Because recognizing directional selection requires data from a large number of isolates over a substantial time period (Buckee et al., 2008; Chen et al., 2006; Lefébure & Stanhope, 2007), we will focus on purifying and diversifying selection in this study. Purifying selection is prevalent amongst microbes (McNally et al., 2019) and common amongst core genes, either because they are integral to cellular processes or vital for survival in a given habitat (Cohan, 2016; Cohan, 2017). Purifying selection may maintain cohesion within a clade by purging diversity, isolating clades from each other and maintaining their distinctiveness (Cohan, 2011; Cohan, 2016). Diversifying selection also plays a key role in microbial evolution by maintaining multiple allelic variants within a population, a pattern which is common in genes linked to host colonisation, phage resistance and responses to vaccines and antibiotics (Harrow et al., 2021; McNally et al., 2019).

Quantifying the relative impact of selection on bacterial divergence is challenging, as this is dependent upon grouping multiple strains together as a single species, which has proven more difficult for bacteria than for plants and animals (Robinson, Thomas, & Hanage, 2017). In addition to difficulties in recognising directional selection, identifying which regions in the core genome are under selection is challenging due to inconsistent selection across sites within a gene and over time, ubiquity of negative selection across genomes and mutation rate heterogeneity (Chen et al., 2006; Fay & Wu, 2003; Zhang, Nielsen, & Yang, 2005). In contrast, selection on accessory genes is more simply inferred using presence/absence (Méric et al., 2018a; Vasquez-Rifo, Veksler-Lublinsky, Cheng, Ausubel, & Ambros, 2019). In this study, we aimed to identify regions under strong selection while accounting for other factors that influence the rate of molecular evolution (Méric et al., 2018a). Genes with higher or lower allelic diversity compared to the genomic average are likely under diversifying and purifying selection respectively, so we combined our expectations of allelic diversity with a method used to infer differences in allelic diversity between genes (Cohan, 2016; Méric, Mageiros, Pensar, et al., 2018b; Shea et al., 2011).

We applied this method to closely related bacterial clades with distinct ecological niches which we hypothesize have undergone divergent selection pressures. This methodology allows uncovering of the effect of selection within bacterial core genomes and a comparison of selective pressures acting on different bacterial clades.

The *Bacillus cereus* (*Bc*) group contains a number of species with clinical or industrial importance: *Bacillus anthracis* (*Ba*), the causative agent of anthrax (Turnbull, 2002); *Bacillus cereus sensu stricto*, a causative agent of food-poisoning (Messelhäußer & Ehling-Schulz, 2018); *Bacillus thuringiensis* (*Bt*), a group of specialized invertebrate pathogens widely exploited as bio-pesticides (Bravo, Gill, & Soberón, 2007); and *Bacillus mycoides* (*Bm*), a psychrotolerant species which incorporates the former *Bacillus weihenstephanensis* (Lechner et al., 1998; Liu, Lai, & Shao, 2018). The *Bc* group has been well-studied, and there is increasing evidence that different clades have distinct ecological niches (Manktelow, White, Crickmore, & Raymond, 2020; Zheng et al., 2018), making this group ideal for exploring the importance of niche-specific selection to driving clade divergence. For instance, carriage of enterotoxins and insecticidal toxin genes is known to vary strongly between clades (Cardazzo et al., 2008; Méric et al., 2018a); clade also correlates with habitat, thermal niche, and cytotoxicity (Guinebretière et al., 2008; Guinebretière, Velge, Couvert, Carlin, & Debuyser, 2010; Raymond, Wyres, Sheppard, Ellis, & Bonsall, 2010). Thermal niches and clade also predict relative fitness at different temperatures and fitness in a model insect host (Manktelow et al., 2020) and are linked to differences in biogeographical distribution (Drewnowska, Stefanska, Czerniecka, Zambrowski, & Swiecicka, 2020).

The phylogenetic structure of the group is well established and recoverable when alignments of multiple housekeeping genes (MLST) or of the entire core genome are used to create phylogenies (Méric et al, 2018a; Priest et al., 2004). However, the taxonomy of the *Bc* group is much disputed (Carroll et al., 2020; Helgason et al., 2000; Liu et al., 2015). Different authors subdivide the group based on different levels of genetic distinctiveness; consequently, the number of informally recognized clades ranges between five and seven (Guinebretière et al., 2008; Méric et al., 2018a). In this study we will use the five-clade structure initially recovered by MLST (Raymond et al., 2010), as these groups are clearly separated by large phylogenetic distances. This is important to our methodology as selection leading to clade divergence should have occurred far in the past. The gene-by-gene approach used here – which relies on known loci across multiple genomes – means that our methods cannot identify recent directional selection that has occurred only within a subset of a given clade or new imports that do not belong to a recognised locus (Sheppard, Jolley, & Maiden, 2012) and so has not been designed to identify "ecotypes" with recent evolutionary origins (Cohan, 2016). We will focus on clades 1, 2 and 3 in the *Bc* group,

originally named the "*anthracis*", "*kurstaki*" and "*weihenstephanensis*" clades respectively (Priest et al., 2004). While *Bacillus cereus sensu stricto* strains are found in all three clades (Patiño-Navarrete & Sanchis, 2017), Clade 1 contains all *Ba* isolates, Clade 2 contains the majority of insecticidal *Bt* isolates while Clade 3 corresponds to the psychrotolerant *Bacillus mycoides* species (Liu et al., 2018). Bacteria in these clades are readily isolated from both clinical and natural environments and are well represented in genomic databases.

Here, we hypothesised that the three *Bc* clades are ecologically distinct due to selection on their core genomes. We predicted that different genes would be found within the clade-specific core genomes of each clade, and that these genes would have different levels of allelic diversity in each clade, due to differences in selection pressure. We also hypothesised that ecological selection acts on the accessory genome and that HGT would be more frequent amongst diversifying genes, promoting diversification between clades. A large collection of *Bc* isolate genomes were used to reconstruct the five-clade phylogeny identified in previous studies (Méric et al., 2018a; Priest et al., 2004). Based on comparisons of gene-level allelic diversity to the *Bc* strict core genome average (Chattopadhyay et al., 2009; Méric et al., 2018b), we identified genes core to each clade under selection, while presence/absence was used to identify accessory genes under selection (Méric et al., 2018a; Vasquez-Rifo et al., 2019). These genes were subjected to GO analyses to determine functional enrichment, while consistency indices were used to estimate rates of HGT (Méric et al., 2018b).

Materials and Methods

Isolate selection

Bacillus cereus (*Bc*) sequence assemblies were gathered from the *Multispecies BIGSdb* database (Jolley & Maiden, 2010; https://sheppardlab.com/resources/). The isolates belonged to a recognized *Bc sl* species (Bazinet, 2017), were assembled from fewer than 3000 contigs, and had genome sizes in line with previous estimates for the group (Chun et al., 2012; Li et al., 2015; Méric et al., 2018a; Yi et al., 2016). 352 isolate genomes met the selection criteria; of these, 24 isolates could not be assigned to clades with certainty and were removed from the analysis, leaving 328 isolate genomes (Supplementary Table 1).

Creation of a reference pan-genome

The assemblies were aligned using the MAFFT algorithm (Katoh & Standley, 2013) and a gene-by-gene approach. Assembly was conducted in the *BIGSdb* database (Sheppard et al., 2012). Contiguous sequences for each isolate were exported and entered into the Pan-genome Iterative Refinement And

Threshold Evaluation (PIRATE) toolbox (Bayliss, Thorpe, Coyle, Sheppard, & Feil, 2019). In the PIRATE toolbox, genome sequences are passed through multiple cluster thresholds to account for different selection strengths between isolates, avoiding over-clustering and over-splitting of groups (Bayliss et al., 2019). Sequences are filtered from input files and CD-HIT used to create sequence clusters. MCL clustering processes are repeated by PIRATE at default amino acid identity thresholds; the initial clustering at the lowest threshold identified 'gene families' and continued until the highest user-specified threshold. Unique MCL clusters at the highest threshold (95% amino acid identity) were classified as 'unique alleles' (Bayliss et al., 2019). Paralogs were identified and loci were classified, then gene families with multiple loci were checked for over-clustering. Genes were annotated using Prokka (Seemann, 2014). PIRATE produced a gene presence/absence matrix, with each gene possessing its own identifier (Méric et al., 2014). The strict core genome for the entire dataset was identified in Excel by ordering genes based on the percentage of isolates within the group containing this gene. Genes were considered "strict core" if present in all isolates.

Phylogenetic analysis

A maximum likelihood phylogeny was produced using 1004 "strict core" gene sequences. These strict core genes were present in all isolates used in this study. The concatenated sequences were aligned using MAFFT (Katoh & Standley, 2013). A maximum-likelihood phylogeny (Gadagkar, Rosenberg, & Kumar, 2005; Saitou & Imanishi, 1989) was produced using IQ-TREE (Minh et al., 2020) with ModelFinder (Kalyaanamoorthy, Minh, Wong, Von Haeseler, & Jermiin, 2017); the substitution model selected was GTR+F+R10. Inclusion of isolates assigned to clades in a previous study helped with clade recovery (Méric et al., 2018a). The tree was visualised using the *R* package *ggtree* (Yu, Smith, Zhu, Guan, & Lam, 2017).

Identifying core and accessory genes under selection within clades

To derive clade-specific core and accessory genomes, strains within clades 1-3 were extracted in *R* by using *ggtree* (Yu et al., 2017). From these, we reconstructed clade-specific core genomes consisting of genes present in ≥95% of the isolates within each clade. This led to a reduced chance of rejecting "clade-defining" genes that have been lost in very derived isolates. Based on previous observations of allelic diversity and selection (Cohan, 2016; Dugatkin, Perlin, Lucas, & Atlas, 2005; Shea et al., 2011), genes of low allelic diversity were considered to be under purifying selection (i.e. selection leading to a reduced number of different alleles) while genes of high allelic diversity were considered under diversifying selection (i.e. selection leading to a greater number of alleles). All alleles of each gene in the strict core and clade-specific core genomes were found through comparison of the isolates to a representative

FASTA sequence in the Multispecies BIGSdb Genome Comparator under default parameters. Incomplete loci were ignored for pairwise comparison and paralogs were excluded entirely (Jolley & Maiden, 2010). We produced alignments using the MAFFT algorithm (Katoh & Standley, 2013). Diversity per locus was calculated for each gene by dividing the number of distinct alleles by the number of isolates containing that gene (Méric et al., 2018a; Méric et al., 2018b). To distinguish selected regions from neutral ones while accounting for other factors influencing molecular evolution, allelic diversity of each clade-specific core gene was compared to the overall within-clade diversity of the strict core genome (Fay & Wu, 2003; Méric et al., 2018b). Those genes that lay outside two standard deviations of the core genome average (i.e. ~5% of the genes) were considered to have significantly low or high diversity and were therefore considered to be under selection (Cohan, 2016; Dugatkin et al., 2005; Shea et al., 2011). Clade-specific accessory genes were defined as genes present in under 95% of a clade; gene presence and absence was used to identify accessory genes under selection as in previous studies (Méric et al., 2018a; Vasquez-Rifo et al., 2019).

Gene Ontology (GO) analysis

To determine whether selected clade-specific core and accessory genes were enriched for certain functions, each gene was assigned an identification number from the Universal Protein Resource Knowledge Base (UniProtKB) (Boutet, Lieberherr, Tognolli, Schneider, & Bairoch, 2007), based on PIRATE's prediction of their gene name and function (Bayliss et al., 2019). *Bacillus subtilis* identification codes were used because the list of *B. subtilis* UniProtKB codes is more comprehensive than for the list for *B. cereus*, and gene names and functions are equivalent between the species. UniProtKB codes were also assigned for the strict core genes. Where a gene coded for a hypothetical function or had no suitable ortholog amongst *B. subtilis*, the gene was excluded from the analysis. Codes for each set of genes were entered into the Gene Enrichment Analysis tool on the Gene Ontology website, which uses the PANTHER classification system (Mi, Muruganujan, Casagrande, & Thomas, 2013). Over- and under-representation of biological processes compared to the strict core genome was calculated using binomial testing (Rupert Jr, 2012) with replacement, approximating the hypergeometric distribution due to sample size (Rivals, Personnaz, Taing, & Potier, 2007). A Bonferroni correction was used to account for multiple testing (Weisstein, 2004).

Inference of HGT using consistency indices

To examine the impact of HGT on clade formation, consistency indices (CInds) were used to estimate the level of HGT amongst genes under selection (Méric et al., 2018b). CInds were created to

detect homoplasy by comparing the fit of genetic alignment data to a phylogenetic tree. An alignment of allelic sequences from the same gene is compared to a reliable phylogeny produced using multiple conserved genes (Saitou & Imanishi, 1989) to produce a consistency index; lower indices indicate a greater degree of homoplasy. Homoplasy can be caused by independent mutation but is commonly assumed to be caused mainly by homologous recombination (Sanderson & Donoghue, 1989; Schliep, 2011), meaning that consistency indices can be used to infer levels of HGT within a group of bacterial strains.

Only genes that were present in all strains in the phylogeny and were considered under either purifying or diversifying selection in at least one clade were included in the consistency index analysis. Consistency indices were calculated for each gene using the *R Phangorn* Package (Schliep, 2011) and the maximum-likelihood group phylogeny was used for comparisons. The process was repeated for all genes in the strict core genome (n=1004). The average CInd of each gene set was compared using a Wilcoxon-Mann-Whitney test. The frequency distribution of CInds for both gene sets was also examined. Both analyses have previously been conducted to test for significant differences in consistency indices between sets of genes (Méric et al., 2018b).

Results

The *Bacillus cereus* group phylogeny has a distinct clade structure

The strict core genome phylogeny divided *Bacillus cereus* isolates into genetically distinct clades. 328 genomes from the Multispecies BIGSdb database (Jolley & Maiden, 2010) met criteria for the study, with an average size of ~5.6mb ± 0.3mb (Supplementary table 1) and an average contig number of 285. Variation in assembly sequence size and contig number was consistent with other published estimates of *Bacillus* group genome sizes (Chun et al., 2012; Li et al., 2015; Méric et al., 2018a; Takeno et al., 2012; Yi et al., 2016). The group pan-genome produced by PIRATE contained 36,687 genes, consisting of 1004 strict core genes excluding homologs and 35,679 accessory genes. A maximum-likelihood tree was produced using the concatenated strict core genome sequences and was consistent with the five-clade phylogeny proposed by previous studies (Méric et al., 2018a; Sorokin et al., 2006) (Fig. 1a). The three largest clades, clades 1-3, contained 94, 95 and 78 isolates respectively (Fig. 1a).

Functional enrichment is dependent on clade and whether the genes are core or accessory

Analysis of clade-specific core genes under selection suggests different selective pressures acting on each *Bc* clade. Allelic diversity was calculated for each gene that was present in all strains within a

specific clade – the clade-specific core genes – and compared to the strict core genome average to identify genes under purifying or diversifying selection (Fig. 1b). Out of 4383 clade-specific core genes across 3 clades, 261 had allelic diversity significantly lower than the within-clade strict core genome average (two standard deviations below the mean), while 161 had significantly higher allelic diversity than the within-clade strict core genome average (two standard deviations above the mean) (Supplementary table 2). Despite some genes appearing in multiple clade-specific core genomes, most genes were conserved or diverse only within one clade (Fig. 2). Genes found to be conserved or diverse in previous studies were also found to be conserved or diverse respectively in this study. These included the *cspA* gene, coding for a highly conserved cold-shock protein used to classify the psychrotolerant *Bacillus mycoides* species (Lechner et al., 1998), and the *hag* gene which encodes a diverse bacterial flagella protein (Xu & Côté, 2006). Genes linked to functions such as protein export were conserved in all clades (Bost & Belin, 1997; Fröderberg, Houben, Baars, Luirink, & De Gier, 2004) (Supplementary table 2) and as expected, Clade 3 contained many highly conserved cold-shock proteins (Ermolenko & Makhatadze, 2002). Genes under diversifying selection in all clades included genes coding for flagellin (Xu & Côté, 2006) and the bacteriophage membrane receptor *yueB* (São-José, Baptista, & Santos, 2004). A notable gene under diversifying selection in Clade 2 was *emrB*, a multi-drug export protein (Lomovskaya & Lewis, 1992).

Clade-specific accessory genes under selection were identified through presence/absence to a specific clade. 5239, 7559 and 5605 genes were found only in Clade 1, Clade 2 and Clade 3 respectively and present in less than 95% of the clade. Accessory genes under positive selection in each clade showed functions that are distinct to each clade. Of these, several are worthy of note; the Clade 1-specific accessory genome included the gene *InlA*, which codes for internalin-A and allows the invasion of mammal cells (Dhar, Faull, & Schneewind, 2000), the Clade 2-specific accessory genome included Cry toxins – key *Bt* insecticidal toxins – such as *cry2Ab* (Zheng et al., 2018) and the Clade 3-specific accessory genome contained the *binA* gene, which produces a homolog to an insecticidal binary toxin component (Palma, Muñoz, Berry, Murillo, & Caballero, 2014) (Supplementary Table 2).

GO analyses suggest clade-specific selection acting on the core and accessory genomes of each *Bc* clade

Binomial testing was used to measure the functional enrichment of biological processes (Ashburner et al., 2000; Gene Ontology Consortium, 2019) within clade-specific core and accessory genomes (Mi et al., 2013) by comparison to the strict core genome. This methodology allowed ecological characterisation of the clades and avoided *a priori* assumptions of relevance. Additionally, it avoids characterising a clade by the possession of any one gene, as has often been the case in the *Bc sl* group (Bravo et al., 2007; Lechner et al., 1998). There was significant functional enrichment of biological

processes amongst conserved and diverse clade-specific core genes of all clades; conserved clade-specific genes were often linked to translation (Fig. 3a). However, some enrichment was clade-specific: Clade 3 contained a greater number of conserved genes linked to negative regulation of transcription and fewer conserved genes linked to biosynthesis and stimulus response than would be expected based on the strict core genome (Fig. 3a). The same was found to be the case for diverse clade-specific genes; genes with uncharacterised functions were more common than expected within Clade 1 and less common than expected in Clades 2 and 3, but only Clade 2 showed unique functional enrichment, with more genes linked to antibiotic and antimicrobial resistance than expected. Functional enrichment of biological processes was robust when the criteria for considering genes under selection within a clade were relaxed to include ~10% of the clade-specific core genomes as opposed to ~5% as described above.

Like the clade-specific core genomes, there were disparities in functional enrichment between the clade-specific accessory genomes. While some processes – such as antibiotic biosynthesis – were enriched in all clade-specific accessory genomes, there were differences between the clades regarding the enrichment of other biological processes (Fig. 3b). Interestingly, biological processes enriched within clade-specific accessory genomes were not the same as those enriched within that clade's specific core genome. For instance, Clade 3 accessory genes were more likely to be linked to motility and secondary metabolism, while its clade-specific core genome was not. Clade 3 was also not significantly enriched for accessory genes linked to negative regulation of transcription, while its core genome was (Fig. 3b).

Genes under diversifying selection undergo more frequent horizontal gene transfer

Two sets of clade-specific core genes were suitable for consistency index analysis; 42 genes of low allelic diversity and 24 genes with high allelic diversity were present in all 328 strains and therefore their gene phylogeny could be compared to the strict core genome phylogeny to check for inconsistencies that suggest HGT. Consistency indices were calculated for clade-specific core genes of high and low diversity, as well as for all genes in the strict core genome. The consistency indices of each gene set suggest that genes under diversifying selection undergo frequent HGT, while HGT is uncommon amongst conserved genes (Fig. 4a and 4b); the mean consistency index of conserved clade-specific core genes (0.46 ± 0.02) was significantly higher than the mean of the strict core genome (0.34 ± 0.003) (Wilcoxon-Mann-Whitney test; $U = 33722$, $p = 4.435e^{-11}$). In contrast, mean consistency index of diverse clade-specific core genes (0.28 ± 0.018) was significantly lower than for the strict core genome (Wilcoxon-Mann-Whitney test; $U = 7893$, $p = 0.00385$) (Fig. 5).

Discussion

This study aimed to explore ecological differentiation between closely related bacterial clades and the role of selection in driving and maintaining this distinctiveness. To accomplish this, we tested bacterial genomes from an economically important and well-studied model group for signatures of selection. The *Bacillus cereus* (*Bc*) group contains many different strains, all thought to be well-adapted to exploit protein-rich food such as cadavers (Manktelow et al., 2020; Rasigade, Hollandt, & Wirth, 2018). Despite high levels of genetic similarity, the clade structure of the group is distinct and robust to multiple phylogenetic methods. Clades have been associated with differences in fitness and virulence gene complement, as well as with distinct biogeographic and thermal niches (Cardazzo et al., 2008; Drewnowska et al., 2020; Guinebretière et al., 2008; Guinebretière et al., 2010; Manktelow et al., 2020; Méric et al., 2018a; Zheng et al., 2018); here, we show that clade-specific core and accessory genomes bear signatures consistent with niche-specific selection.

We identified genes under putative purifying and diversifying selection within clade-specific core genomes by comparison to diversity in the strict core genome. As mentioned, identifying genes undergoing selection presents computational and data sampling challenges (Buckee et al., 2008; Zhang et al., 2005); additionally, selection must be distinguished from other factors affecting allelic diversity (Chen et al., 2006; Fay & Wu, 2003; Zhang et al., 2005). This was achieved by using allelic diversity and comparison to the average genomic diversity to identify outliers under strong selection (Méric et al., 2018b). Genes with very low or very high allelic diversity compared to the average are likely to be under strong purifying or diversifying selection (Cohan, 2016; Dugatkin et al., 2005; Shea et al., 2011). Amongst gene sets with non-normally distributed allelic diversity values, using percentile values to encapsulate the most extreme 5% of the data would be suitable; however, due to normal distribution of the data in this study, mean and SD filtering of allelic diversity provided a way to quickly identify genes under strong selection. It should be noted that low allelic diversity may occur due to purifying selection or due to directional selection combined with HGT (i.e. gene-specific sweeps) (Cohan, 2016); this may explain the low numbers of conserved genes within Clade 2. However, because the majority of conserved genes also showed low levels of HGT (Fig. 5), we feel confident that the majority of conserved genes are the result of purifying selection; an in-depth examination could identify genes from among these sets that are more likely to have undergone gene-specific sweeps.

Analysis of clade-specific conserved core genes suggested that core genes under purifying selection differed significantly between clades and supported previous hypotheses about the ecological distinctiveness of major *Bc* clades (Fig. 3a). For example, consider our analysis of Clade 3, now recognised as *B. mycoides* (Carroll et al., 2020). Here, the analysis of clade-specific core genes identified the cold-

shock protein gene *cspA*, a unique sequence signature of which was used to originally classify the psychrotolerant *Bacillus mycoides* species (Lechner et al., 1998). Furthermore, Clade 3 possessed many conserved genes linked to ribosome assembly and negative regulation of transcription, and few linked to metabolism, biosynthetic processes and external stimuli responses (Ermolenko & Makhatadze, 2002). These features are characteristic of adaptation to low temperatures, where metabolic functions are downregulated in response to cold (Barria, Malecki, & Arraiano, 2013; López-Maury, Marguerat, & Bähler, 2008; Tribelli & López, 2018). This supports other studies indicating that strains within Clade 3 are psychrotolerant specialists (Lechner et al., 1998; Liu et al., 2018; Manktelow et al., 2020) and demonstrates how the methodology used here can identify important genes with specific variants within ecologically distinct groups. Different patterns of enrichment amongst conserved clade-specific core genes also suggest that the clades are ecologically distinct, and purifying selection may maintain new species by purging novel variation caused by mutation and horizontal gene transfer (Cohan, 2016; Cohan, 2017).

We found evidence that diversifying selection within clade-specific core genomes acts on different genes depending on the clade. While the *hag* flagellin gene was extremely diverse across all three clades, only genes of high allelic diversity *within Clade 2* were enriched for functions linked to flagellum-dependent motility. Flagellin is a common receptor for bacteriophages, and since variations in flagellin structure may prevent phage infection, this is a trait likely to be under diversifying selection (Nobrega et al., 2018). Clade 2 also has the largest proportion of isolates encoding insecticidal toxins and carries a greater number of insecticidal toxins than other clades (Méric et al., 2018a; Zheng et al., 2018). This supports the hypothesis that this clade is dominated by specialist insect pathogens (Raymond & Bonsall, 2013; Raymond & Federici, 2017; Raymond et al., 2010) and provides further evidence for the ecological distinctiveness of the clades.

Flagellar motility may also be important during the early stages of insect infection (Mazzantini et al., 2016); *Bt* mutants with reduced flagellar motility have reduced virulence when infecting larvae (Zhang, Lövgren, Low, & Landén, 1993). Diverse *Bt* genes were also more likely to be linked to antimicrobial resistance (Supplementary table 2). Antimicrobial resistance mechanisms are common in *Bc* strains (Abriouel, Franz, Omar, & Gálvez, 2011; Bernhard, Schrempf, & Goebel, 1978) and are often under diversifying selection, which can result in the emergence and maintenance of allelic diversity for that trait (Levin, 1988; McNally et al., 2019). Diversifying selection on antibiotic resistance may be prevalent amongst Clade 2 strains because competition to enter insect cadavers first is intense (Garbutt, Bonsall, Wright, & Raymond, 2011; Van Leeuwen, O'neill, Matthews, & Raymond, 2015). Therefore, overcoming

host defences and securing the first infection of a host may provide an advantage in pathogenic bacteria that is not seen in necrotrophic bacteria.

One of the aims of this study was to assess the importance of selection in maintaining bacterial species. Alternative drift-based models of bacterial speciation assume that genetic differences between taxa are self-reinforcing (Fraser et al., 2007). HGT can erode differences between neutrally diverging lineages and greater genetic distance leads to reduced HGT via a range of mechanisms (Fraser et al., 2007) There is evidence for these kinds of forces operating in the *Bc sl* group; for instance, HGT predominantly occurs within clades (Didelot, Barker, Falush and Priest, 2009). Nevertheless, one notable result of this study was the variation in inferred levels of HGT between loci under different forms of selection. Here, we used consistency indices to infer the prevalence of HGT. High consistency indices amongst conserved genes – such as the *cspA* gene – indicate low levels of HGT (Méric et al., 2018b); in contrast, low consistency indices in diverse genes such as the *hag* gene imply high levels of HGT (Fig. 4 and Fig. 5). At a fundamental level, all chromosomal genes undergo horizontal gene transfer at similar rates (Gogarten et al., 2002). However, the subsequent fate of horizontally transferred alleles differs depending on gene and gene function; this may be due to variation in selection strength and type between genes (Nakamura et al., 2004; Kivisaar, 2019). Our results indicate that the effects of HGT are strongly modulated by selection in the *Bc sl* group. When novel allelic diversity is favoured under diversifying selection, HGT can supply that diversity. However, purifying selection can also purge clade-specific allelic variants that incur strong selective disadvantages in the 'wrong' genetic background (Vos et al, 2015). Moderate levels of HGT therefore do not impeded speciation, as seen in other species (Melendrez et al., 2016). Background levels of HGT are important, but selection can clearly act to promote clade identity and genetic coherence in the face of HGT.

While unlikely to be an issue in *Bacillus cereus* due to intermediate levels of homologous recombination (Patiño-Navarrete & Sanchis, 2017), consistency indices are likely most effective at identifying patterns of HGT when levels are low or intermediate; at high rates of HGT genes may be spread sufficiently widely so that genes received via HGT cannot be distinguished from genes received via linear descent (Andam & Gogarten, 2011; Sanderson & Donoghue, 1989). Spotting inconsistencies may also be difficult in conserved genes due to the small number of differences between genes. However, given levels of HGT are roughly intermediate for all gene sets (~0.5) and that conserved genes with small differences are sufficiently different to be used for reconstructing phylogenies (Saitou & Imanishi, 1989), these would seem to be minor concerns.

The role of accessory genomes in ecological specialisation is widely accepted (Brockhurst, Harrison, Hall, & Richards, 2019; Cobo-Simón & Tamames, 2017); *Bacillus thuringiensis*, which carries key virulence factors primarily on large plasmids, is a well-known example (Zheng et al., 2018). As with the core genome analysis, accessory genes unique to each clade were significantly enriched for specific biological processes. Furthermore, the processes enriched within a clade-specific core genome often differed from the processes enriched within the specific accessory genome of the same clade. For instance, the Clade 3 accessory genome was enriched for genes linked to motility and secondary metabolic processes, while its core genome was not. The utility of presence/absence for identifying accessory genes under selection is still debated, as strains accumulate a mix of deleterious, beneficial and neutral genes and the frequency of beneficial accessory genes is unclear (Vos & Eyre-Walker, 2017). Despite, presence/absence of specific accessory genes has been found to be biological meaningful in other studies (Cohen, Ashkenazy, Levy Karin, Burstein, & Pupko, 2013; Méric et al., 2018a; Vasquez-Rifo et al., 2019). With this considered, our results would suggest that both the core and accessory genome determine a strain's ecology.

While these results indicate the importance of chromosomal core and accessory genes to strain ecology, they should be taken with caution for three reasons. First, enrichment within the accessory genome may not be representative of all strains within a clade; the majority of genes in bacterial pan-genomes are either common ("core" or nearly core) or extremely rare (accessory) (Haegeman & Weitz, 2012). Because the *Bc sl* clades consist of isolates assigned to different species or ecotypes – for instance, both Clades 1 and 2 contain strains identified as *Bt* (Méric et al., 2018a)– the enrichment of certain biological processes within a clade's accessory genome may be due to high numbers of rare genes that are possessed by a minority of the clade in question. Second, the different functional enrichment in clade-specific core and accessory genomes may reflect differences in selection over time as opposed to differences in function; within one species of bacterium, accessory gene content change occurs at faster rates but is retained less readily than amino-acid substitution in the core genome (Wielgoss et al., 2016), implying that accessory genomes reflect current selection and core genomes reflect past selection. Third, this study did not attempt to incorporate plasmid sequences into the analysis. While it was not possible to differentiate between chromosomal and plasmid DNA in all isolates, we did not explicitly analyse plasmid sequences in this study. While some plasmids are stably associated with *Bc* lineages and therefore considered part of the "core genome" (Méric et al., 2018a; Zheng et al., 2018), many plasmids are highly mobile and carry genes encoding several key virulence traits (Patiño-Navarrete & Sanchis, 2017; Schnepf et al., 1998). While analysis of the selection pressures that formed the *Bc* clades will benefit by excluding

plasmid sequences (by reducing the confounding effect highly mobile plasmids may have on analysis), future researchers may wish to incorporate these important parts of *Bc sl* pan-genome. Therefore, future iterations of this methodology may benefit from two modifications: splitting analysis of the accessory genome into genes of intermediate and low frequency within a clade (Inglin, Meile, & Stevens, 2018) and the incorporation of plasmid sequence data.

It is interesting that Clade 1 does not appear to possess any significant clade-specific enrichments, aside from deficiencies in certain biosynthetic processes (Fig. 3b) and the possession of the internalin-A protein gene *inlA* (allowing for epithelial cell invasion) in its clade-specific core genome (Dhar et al., 2000). We hypothesised that the *anthracis* clade would consist of necromenic (cadaver-associated) bacteria that may specialise on vertebrates (Manktelow et al., 2020) although *B. anthracis* itself is a clonal expansion and represents only a small part of the diversity in this group. Clade 1 includes at least six currently recognised species, though one proposed revision suggests lumping all these groups into a single taxon based on a 92.5 % average nucleotide identify (ANI) (Carroll et al., 2020). Regardless of current taxonomic disputes, the clade splits into two groups separated by a 94% ANI. These two branches of Clade 1 were previously described as *PanC* Groups II and III, corresponding to *Bacillus paranthracis* and allies and *Bacillus albus/wiedmannii* and allies respectively (Guinebretière et al., 2008; Guinebretière et al., 2010). There is evidence for differences in phenotype and biogeography between these groups (Drewnowska et al., 2020; Guinebretière et al., 2008; Guinebretière et al., 2010). "Lumping" these groups into a single clade may be obscuring ecological distinctiveness in Clade 1. While useful for identifying the selection pressures that formed the *Bc* clades and that are currently creating diversity within each clade, our results should not be taken to mean that the clades are ecologically monolithic. Repeating this analysis using the seven-clade phylogeny of Guinebretière *et al* (2008) and with greater representation in these sub-groups may reveal ecological distinctions that were not seen in this study.

This possibility suggests how this selection-informed analysis may be used for refining taxonomic decision making. Methods based on raw genetic differences, such as ANI, appear highly objective; however, decisions still need to be made on how to apply rules and what level of differentiation is appropriate for describing species in a particular group (Carroll et al., 2020; Vos, 2011). There are advantages in describing species as units with real ecological and phenotypic distinctiveness; if groups recognised by ANI-based decisions also show coherent patterns of selection, it provides another means of assessing whether a species definition is of practical value. Another pragmatic application of genome-wide analysis of conserved genes is its value in identifying key ecological traits and single loci that can be

used for species-level identification; one example from this study is the wealth of psychrotolerance traits found in clade 3, exemplified by the conserved cold-shock gene *cspA*.

In conclusion, this study showed that functional enrichment in both core and accessory genes is heavily dependent on clade in the *Bc* bacterial group. Key ecological traits associated with *Bacillus* species – such as antimicrobial and insecticidal activity in *Bt* strains and psychrotolerance in *Bm* strains – were among those enriched in specific clades, supporting the hypothesis that clades within the group formed due to different selection pressures and have distinct ecologies. The core and accessory genomes of each clade appear to experience selection on different traits, highlighting the importance of considering both when determining clade ecology. High levels of HGT amongst diversifying core genes suggests that HGT plays a key role in promoting diversification within the *Bc sl* group. Lastly, this analysis identified genes, such as the *cspA* gene in Clade 3, that can be used to identify strains to the clade level and to infer their ecological niche, allowing easier determination of strains' potential to harm humans and to act as biopesticides, with the commensurate benefits to agricultural and medical practices.

Acknowledgements

We wish to thank Dr Sion Bayliss for running the PIRATE pipeline and Dr Manmohan Sharma for access to the University of Exeter remote servers and advice on effectively constructing maximum-likelihood trees. This work was funded by the BBSRC South West Biosciences Doctoral Training Partnership (Grant number BB/M009122/1), who also provided training to the primary researcher.

References

Abriouel, H., Franz, C. M., Omar, N. B., & Gálvez, A. (2011). Diversity and applications of *Bacillus* bacteriocins. *FEMS microbiology reviews*, *35*(1), 201-232.

Andam, C. P., & Gogarten, J. P. (2011). Biased gene transfer and its implications for the concept of lineage. *Biology direct*, *6*(1), 1-16.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25-29.

Barria, C., Malecki, M., & Arraiano, C. M. (2013). Bacterial adaptation to cold. *Microbiology*, *159*(Pt_12), 2437-2443.

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., & Feil, E. J. (2019). PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience*, *8*(10), giz119.

Bazinet, A. L. (2017). Pan-genome and phylogeny of *Bacillus cereus sensu lato*. *BMC evolutionary biology*, *17*(1), 1-16.

Bernhard, K., Schrempf, H., & Goebel, W. (1978). Bacteriocin and antibiotic resistance plasmids in *Bacillus cereus* and *Bacillus subtilis*. *Journal of bacteriology*, *133*(2), 897-903.

Bost, S., & Belin, D. (1997). *prl* mutations in the *Escherichia coli* secG gene. *Journal of Biological Chemistry*, *272*(7), 4087-4093.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). Uniprotkb/swiss-prot. In *Plant bioinformatics* (pp. 89-112). Humana Press.

Bravo, A., Gill, S. S., & Soberon, M. (2007). Mode of action of *Bacillus thuringiensis* Cry and Cyt toxins and their potential for insect control. *Toxicon*, *49*(4), 423-435.

Brockhurst, M. A., Harrison, E., Hall, J. P., Richards, T., McNally, A., & MacLean, C. (2019). The ecology and evolution of pangenomes. *Current Biology*, *29*(20), R1094-R1103.

Buckee, C. O., Jolley, K. A., Recker, M., Penman, B., Kriz, P., Gupta, S., & Maiden, M. C. (2008). Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. *Proceedings of the National Academy of Sciences*, *105*(39), 15082-15087.

Cardazzo, B., Negrisolo, E., Carraro, L., Alberghini, L., Patarnello, T., & Giaccone, V. (2008). Multiple-locus sequence typing and analysis of toxin genes in *Bacillus cereus* food-borne isolates. *Applied and environmental microbiology*, *74*(3), 850-860.

Carroll, L. M., Wiedmann, M., & Kovac, J. (2020). Proposal of a taxonomic nomenclature for the *Bacillus cereus* group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. *MBio*, *11*(1), e00034-20.

Chattopadhyay, S., Weissman, S. J., Minin, V. N., Russo, T. A., Dykhuizen, D. E., & Sokurenko, E. V. (2009). High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proceedings of the National Academy of Sciences*, *106*(30), 12412-12417.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., ... & Gordon, J. I. (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proceedings of the National Academy of Sciences*, *103*(15), 5977-5982.

Chun, J. H., Hong, K. J., Cha, S. H., Cho, M. H., Lee, K. J., Jeong, D. H., ... & Rhie, G. E. (2012). Complete genome sequence of *Bacillus anthracis* H9401, an isolate from a Korean patient with anthrax.

Cobo-Simón, M., & Tamames, J. (2017). Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC genomics*, *18*(1), 1-11.

Cohan, F. M. (2011). Are species cohesive?—a view from bacteriology. *Population genetics of bacteria: a tribute to Thomas S. Whittam*, 43-65.

Cohan, F. M. (2016). Bacterial speciation: genetic sweeps in bacterial species. *Current Biology*, *26*(3), R112-R115.

Cohan, F. M. (2017). Transmission in the origins of bacterial diversity, from ecotypes to phyla. *Microbiology spectrum*, *5*(5), 5-5.

Cohen, O., Ashkenazy, H., Levy Karin, E., Burstein, D., & Pupko, T. (2013). CoPAP: coevolution of presence–absence patterns. *Nucleic acids research*, *41*(W1), W232-W237.

Dhar, G., Faull, K. F., & Schneewind, O. (2000). Anchor structure of cell wall surface proteins in *Listeria monocytogenes*. *Biochemistry*, *39*(13), 3725-3733.

Didelot X, Barker M, Falush D, Priest FG. (2009). Evolution of pathogenicity in the *Bacillus cereus* group. *Systematic and Applied Microbiology*, *32*(2), 81-90.

Doolittle, W. F., & Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome biology*, *7*(9), 1-7.

Drewnowska, J. M., Stefanska, N., Czerniecka, M., Zambrowski, G., & Swiecicka, I. (2020). Potential enterotoxicity of phylogenetically diverse *Bacillus cereus sensu lato* soil isolates from different geographical locations. *Applied and environmental microbiology*, *86*(11), e03032-19.

Dugatkin, L. A., Perlin, M., Lucas, J. S., & Atlas, R. (2005). Group-beneficial traits, frequency-dependent selection and genotypic diversity: an antibiotic resistance paradigm. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1558), 79-83.

Ermolenko, D. N., & Makhatadze, G. I. (2002). Bacterial cold-shock proteins. *Cellular and Molecular Life Sciences CMLS*, *59*(11), 1902-1913.

Fay, J. C., & Wu, C. I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annual review of genomics and human genetics*, *4*(1), 213-235.

Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, *315*(5811), 476-480.

Fröderberg, L., Houben, E. N., Baars, L., Luirink, J., & De Gier, J. W. (2004). Targeting and translocation of two lipoproteins in *Escherichia coli* via the SRP/Sec/YidC pathway. *Journal of Biological Chemistry*, *279*(30), 31026-31032.

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *304*(1), 64-74.

Garbutt, J., Bonsall, M. B., Wright, D. J., & Raymond, B. (2011). Antagonistic competition moderates virulence in *Bacillus thuringiensis*. *Ecology letters*, *14*(8), 765-772.

Gene Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, *47*(D1), D330-D338.

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, *19*(12), 2226-2238.

Guinebretière, M. H., Thompson, F. L., Sorokin, A., Normand, P., Dawyndt, P., Ehling-Schulz, M., ... & De Vos, P. (2008). Ecological diversification in the *Bacillus cereus* group. *Environmental Microbiology*, *10*(4), 851-865.

Guinebretière, M. H., Velge, P., Couvert, O., Carlin, F., Debuyser, M. L., & Nguyen-The, C. (2010). Ability of *Bacillus cereus* group strains to cause food poisoning varies according to phylogenetic affiliation (groups I to VII) rather than species affiliation. *Journal of clinical microbiology*, *48*(9), 3388-3391.

Haegeman, B., & Weitz, J. S. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC genomics*, *13*(1), 1-15.

Hanage, W. P., Fraser, C., & Spratt, B. G. (2006). Sequences, sequence clusters and bacterial species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1475), 1917-1927.

Harrow, G. L., Lees, J. A., Hanage, W. P., Lipsitch, M., Corander, J., Colijn, C., & Croucher, N. J. (2021). Negative frequency-dependent selection and asymmetrical transformation stabilise multi-strain bacterial population structures. *The ISME journal*, *15*(5), 1523-1538.

Helgason, E., Økstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., ... & Kolstø, A. B. (2000). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. *Applied and environmental microbiology*, *66*(6), 2627-2630.

Inglin, R. C., Meile, L., & Stevens, M. J. (2018). Clustering of pan-and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. *BMC genomics*, *19*(1), 1-15.

Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, *11*(1), 1-11. Available through https://sheppardlab.com/resources/.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587-589.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kivisaar, M. (2019). Mutation and recombination rates vary across bacterial chromosome. *Microorganisms*, *8*(1), 25.

Lechner, S., Mayr, R., Francis, K. P., Prü, B. M., Kaplan, T., Wießner-GunkeL, E. L. K. E., ... & Scherer, S. (1998). *Bacillus weihenstephanensis* sp. nov. is a new psychrotolerant species of the *Bacillus*

cereus group. *International Journal of Systematic and Evolutionary Microbiology*, *48*(4), 1373-1382.

Lefébure, T., & Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome biology*, *8*(5), 1-17.

Levin, B. R. (1988). Frequency-dependent selection in bacterial populations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *319*(1196), 459-472.

Li, Q., Xu, L. Z., Zou, T., Ai, P., Huang, G. H., Li, P., & Zheng, A. P. (2015). Complete genome sequence of *Bacillus thuringiensis* strain HD521. *Standards in genomic sciences*, *10*(1), 1-8.

Liu, Y., Lai, Q., Göker, M., Meier-Kolthoff, J. P., Wang, M., Sun, Y., ... & Shao, Z. (2015). Genomic insights into the taxonomic status of the *Bacillus cereus* group. *Scientific reports*, *5*(1), 1-11.

Liu, Y., Lai, Q., & Shao, Z. (2018). Genome analysis-based reclassification of *Bacillus weihenstephanensis* as a later heterotypic synonym of *Bacillus mycoides*. *International journal of systematic and evolutionary microbiology*, *68*(1), 106-112.

Lomovskaya, O. L. G. A., & Lewis, K. I. M. (1992). *emr*, an *Escherichia coli* locus for multidrug resistance. *Proceedings of the National Academy of Sciences*, *89*(19), 8938-8942.

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics, 9*(8), 583-593. doi:10.1038/nrg2398

Maistrenko, O. M., Mende, D. R., Luetge, M., Hildebrand, F., Schmidt, T. S., Li, S. S., ... & Bork, P. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME journal*, *14*(5), 1247-1259.

Manktelow, C. J., White, H., Crickmore, N., & Raymond, B. (2021). Divergence in environmental adaptation between terrestrial clades of the *Bacillus cereus* group. *FEMS Microbiology Ecology*, *97*(1), fiaa228.

Mazzantini, D., Celandroni, F., Salvetti, S., Gueye, S. A., Lupetti, A., Senesi, S., & Ghelardi, E. (2016). FlhF is required for swarming motility and full pathogenicity of *Bacillus cereus*. *Frontiers in microbiology*, *7*, 1644.

McInerney, J. O., McNally, A., & O'connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature microbiology*, *2*(4), 1-5.

McNally, A., Kallonen, T., Connor, C., Abudahab, K., Aanensen, D. M., Horner, C., ... & Corander, J. (2019). Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *MBio*, *10*(2), e00644-19.

Melendrez, M. C., Becraft, E. D., Wood, J. M., Olsen, M. T., Bryant, D. A., Heidelberg, J. F., Rusch, D. B., Cohan, F. M., Ward, D. M. (2016). Recombination does not hinder formation or detection of ecological species of *Synechococcus* inhabiting a hot spring cyanobacterial mat. *Frontiers in microbiology*, *6*, 1540.

Meric, G., Mageiros, L., Pascoe, B., Woodcock, D. J., Mourkas, E., Lamble, S., ... & Sheppard, S. K. (2018). Lineage-specific plasmid acquisition and the evolution of specialized pathogens in *Bacillus thuringiensis* and the *Bacillus cereus* group. *Molecular ecology*, *27*(7), 1524-1540.

Méric, G., Mageiros, L., Pensar, J., Laabei, M., Yahara, K., Pascoe, B., ... & Sheppard, S. K. (2018). Disease-associated genotypes of the commensal skin bacterium *Staphylococcus epidermidis*. *Nature communications*, *9*(1), 1-11.

Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C., Jolley, K. A., & Sheppard, S. K. (2014). A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. *PloS one*, *9*(3), e92798.

Messelhäußer, U., & Ehling-Schulz, M. (2018). *Bacillus cereus*—a multifaceted opportunistic pathogen. *Current Clinical Microbiology Reports*, *5*(2), 120-125.

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551-1566.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530-1534.

Molina, N., & Van Nimwegen, E. (2008). Universal patterns of purifying selection at noncoding positions in bacteria. *Genome research*, *18*(1), 148-160.

Nakamura, Y., Itoh, T., Matsuda, H., & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics*, *36*(7), 760-766.

Nobrega, F. L., Vlot, M., de Jonge, P. A., Dreesens, L. L., Beaumont, H. J., Lavigne, R., ... & Brouns, S. J. (2018). Targeting mechanisms of tailed bacteriophages. *Nature Reviews Microbiology*, *16*(12), 760-773.

Palma, L., Muñoz, D., Berry, C., Murillo, J., & Caballero, P. (2014). Draft genome sequences of two *Bacillus thuringiensis* strains and characterization of a putative 41.9-kDa insecticidal toxin. *Toxins*, *6*(5), 1490-1504.

Patiño-Navarrete, R., & Sanchis, V. (2017). Evolutionary processes and environmental factors underlying the genetic diversity and lifestyles of *Bacillus cereus* group bacteria. *Research in Microbiology*, *168*(4), 309-318.

Priest, F. G., Barker, M., Baillie, L. W., Holmes, E. C., & Maiden, M. C. (2004). Population structure and evolution of the *Bacillus cereus* group. *Journal of bacteriology*, *186*(23), 7959-7970.

Rasigade, J. P., Hollandt, F., & Wirth, T. (2018). Genes under positive selection in the core genome of pathogenic *Bacillus cereus* group members. *Infection, Genetics and Evolution*, *65*, 55-64.

Raymond, B., & Bonsall, M. B. (2013). Cooperation and the evolutionary ecology of bacterial virulence: the *Bacillus cereus* group as a novel study system. *Bioessays*, *35*(8), 706-716.

Raymond, B., & Federici, B. A. (2017). In defence of *Bacillus thuringiensis*, the safest and most successful microbial insecticide available to humanity—a response to EFSA. *FEMS microbiology ecology*, *93*(7), fix084.

Raymond, B., Wyres, K. L., Sheppard, S. K., Ellis, R. J., & Bonsall, M. B. (2010). Environmental factors determining the epidemiology and population genetic structure of the *Bacillus cereus* group in the field. *PLoS Pathogens*, *6*(5), e1000905.

Rivals, I., Personnaz, L., Taing, L., & Potier, M. C. (2007). Enrichment or depletion of a GO category within a class of genes: which test?. *Bioinformatics*, *23*(4), 401-407.

Robinson, D. A., Thomas, J. C., & Hanage, W. P. (2017). Population Structure of Pathogenic Bacteria. In M. Tibayrenc (Ed.), *Genetics and evolution of infectious diseases* (pp. 43-57): Elsevier.

Rupert Jr, G. (2012). *Simultaneous statistical inference*. Springer Science & Business Media.

Saitou, N., & Imanishi, T. (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree.

Sanderson, M. J., & Donoghue, M. J. (1989). Patterns of variation in levels of homoplasy. *Evolution*, *43*(8), 1781-1795.

São-José, C., Baptista, C., & Santos, M. A. (2004). *Bacillus subtilis* operon encoding a membrane receptor for bacteriophage SPP1. *Journal of bacteriology*, *186*(24), 8337-8346.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592-593.

Schloss, P. D., & Handelsman, J. (2004). Status of the microbial census. *Microbiology and molecular biology reviews*, *68*(4), 686-691.

Schnepf, E., Crickmore, N. V., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., ... & Dean, D. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiology and molecular biology reviews*, *62*(3), 775-806.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Shea, P. R., Beres, S. B., Flores, A. R., Ewbank, A. L., Gonzalez-Lugo, J. H., Martagon-Rosado, A. J., ... & Musser, J. M. (2011). Distinct signatures of diversifying selection revealed by genome analysis of

respiratory tract and invasive bacterial populations. *Proceedings of the National Academy of Sciences*, *108*(12), 5039-5044.

Sheppard, S. K., McCarthy, N. D., Falush, D. & Maiden, M. C (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* , *320*(5873), 237-239.

Sheppard, S. K., Jolley, K. A., & Maiden, M. C. (2012). A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes*, *3*(2), 261-277.

Sorokin, A., Candelon, B., Guilloux, K., Galleron, N., Wackerow-Kouzova, N., Ehrlich, S. D., ... & Sanchis, V. (2006). Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains. *Applied and Environmental Microbiology*, *72*(2), 1569-1578.

Takeno, A., Okamoto, A., Tori, K., Oshima, K., Hirakawa, H., Toh, H., ... & Ohta, M. (2012). Complete genome sequence of *Bacillus cereus* NC7401, which produces high levels of the emetic toxin cereulide.

Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current opinion in microbiology*, *11*(5), 472-477.

Tribelli, P. M., & López, N. I. (2018). Reporting key features in cold-adapted bacteria. *Life*, *8*(1), 8.

Turnbull, P. C. B. (2002). Introduction: anthrax history, disease and ecology. *Anthrax*, 1-19.

Van Leeuwen, E., O'neill, S., Matthews, A., & Raymond, B. (2015). Making pathogens sociable: the emergence of high relatedness through limited host invasibility. *The ISME journal*, *9*(10), 2315-2323.

Vasquez-Rifo, A., Veksler-Lublinsky, I., Cheng, Z., Ausubel, F. M., & Ambros, V. (2019). The *Pseudomonas aeruginosa* accessory genome elements influence virulence towards *Caenorhabditis elegans*. *Genome biology*, *20*(1), 1-22.

Vos, M. (2011). A species concept for bacteria based on adaptive divergence. *Trends in microbiology*, *19*(1), 1-7.

Vos M, Hesselman MC, Te Beek TA, van Passel MW, Eyre-Walker A. (2015). Rates of lateral gene transfer in prokaryotes: high but why?. *Trends in microbiology, 23*(10), 598-605.

Vos, M., & Eyre-Walker, A. (2017). Are pangenomes adaptive or not?. *Nature microbiology*, *2*(12), 1576-1576.

Weisstein, E. W. (2004). Bonferroni correction. MathWorld. *A Wolfram Web,[Online], Available: http://mathworld. wolfram. com/BonferroniCorrection. html*.

Wielgoss, S., Didelot, X., Chaudhuri, R. R., Liu, X., Weedall, G. D., Velicer, G. J., & Vos, M. (2016). A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. *The ISME journal*, *10*(10), 2468-2477.

Xu, D., & Côté, J. C. (2006). Sequence diversity of the *Bacillus thuringiensis* and *B. cereus sensu lato* flagellin (H antigen) protein: comparison with H serotype diversity. *Applied and environmental microbiology*, *72*(7), 4653-4662.

Yi, Y., de Jong, A., Spoelder, J., Elzenga, J. T. M., van Elsas, J. D., & Kuipers, O. P. (2016). Draft genome sequence of *Bacillus mycoides* M2E15, a strain isolated from the endosphere of potato. *Genome announcements*, *4*(1), e00031-16.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28-36.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, *22*(12), 2472-2479.

Zhang, M. Y., Lövgren, A., Low, M. G., & Landén, R. (1993). Characterization of an avirulent pleiotropic mutant of the insect pathogen *Bacillus thuringiensis*: reduced expression of flagellin and phospholipases. *Infection and immunity*, *61*(12), 4947-4954.

Zheng, J., Gao, Q., Liu, L., Liu, H., Wang, Y., Peng, D., ... & Sun, M. (2017). Comparative genomics of *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple invertebrate hosts. *MBio*, *8*(4), e00822-17.

Data Accessibility

- Genetic data can be accessed from public databases by referring to the strain accession numbers in Supplementary Table 1.
- Sample metadata is available from the Multispecies BIGSdb (Jolley & Maiden, 2010; https://sheppardlab.com/resources/) and are available in Supplementary Table 1. Metadata includes Multispecies BIGSdb ID, the clade the strain was assigned to in this study, isolate identifier, aliases, pathotype, species source, lineage, serovar, clinical isolate, sequence length (bp) and accession number.
- The PIRATE Pipeline is available through GitHub (https://github.com/SionBayliss/PIRATE).

- Details of the clade-specific core genes that showed extremely high and low allelic diversity can be found in Supplementary Tables 2a and 2b. Details of clade-specific accessory genes can also be found in Supplementary Table 2c.
- UniprotKB codes are available for each gene from UniProt Knowledgebase (UniProtKB; https://www.uniprot.org/) and are listed next to their respective gene in Supplementary Table 2. Metadata includes PIRATE ID number, the clades in which a gene was conserved/diverse/accessory, consensus gene name, consensus gene product, UniProtKB code.
- Raw output from the PIRATE pipeline (both the Excel summary and the identified "gene family" .FASTA files), the maximum likelihood tree file and IQ tree command lines, output from the Gene Ontology analysis tool, and the raw output from analysis of consistency indices will be made available publicly through Open Research Exeter (ORE; https://ore.exeter.ac.uk/repository/handle/10036/10890) upon acceptance and publication. The IQ tree and Rscripts used to generate relevant output (Maximum likelihood phylogeny/Fig. 1A, Gene Ontology graph Fig. 3, and consistency index graph Fig. 5) will also be stored here.

Author contributions

Hugh White designed the study, performed research, analysed data and was the primary writer for the manuscript. Samuel K. Sheppard and Ben Raymond helped design the study, and Samuel K. Sheppard contributed the use of PIRATE and the BIGSdb database. Samuel K. Sheppard, Ben Raymond and Michiel Vos all contributed to the writing of the manuscript.

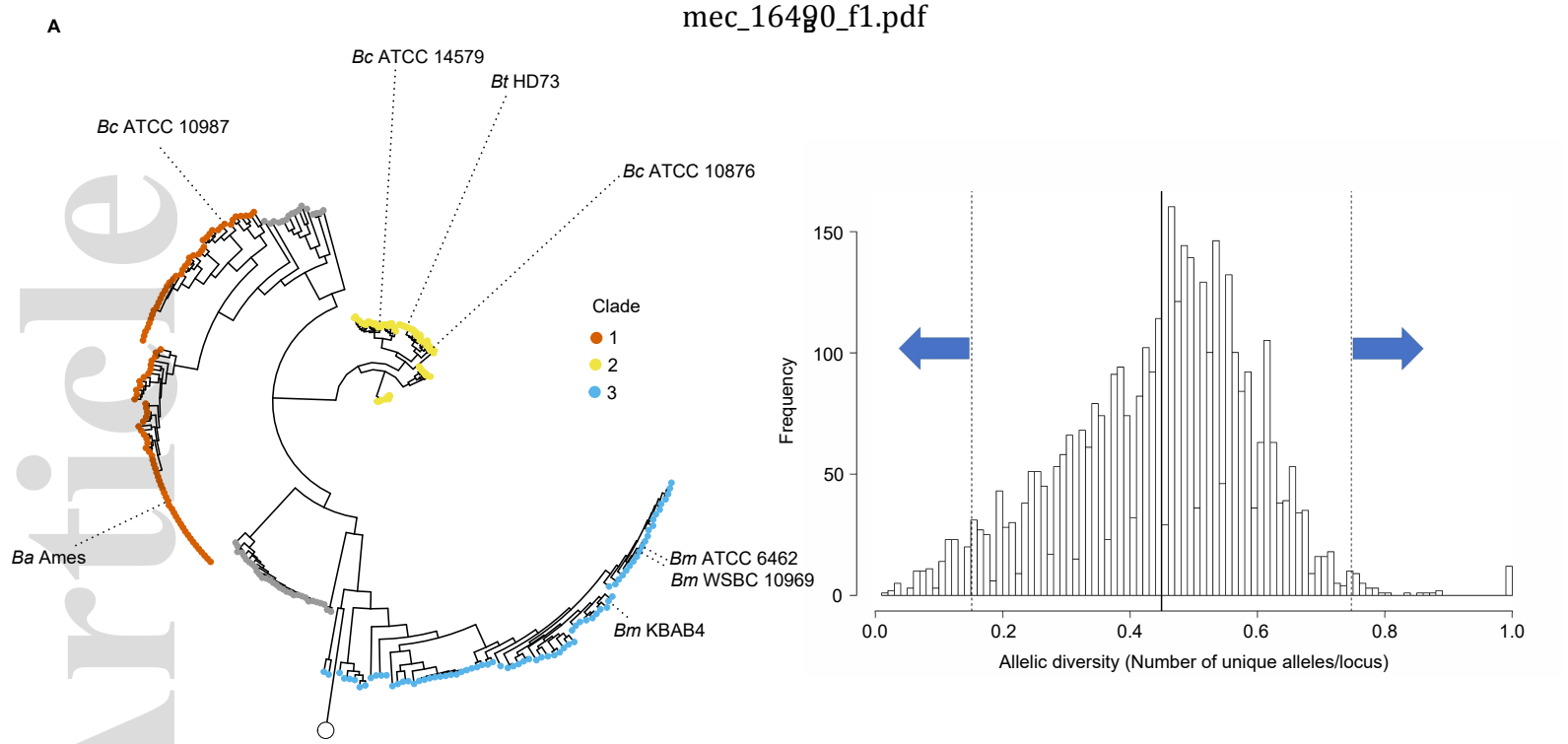

Figure 1: Identifying core and accessory genes under selection within the *Bacillus cereus sensu lato* (*Bc sl)* phylogeny
A Maximum likelihood phylogeny of the *Bc sl* group strains used in this study. The concatenated core genome sequences were aligned using MAFFT and fed into IQ-TREE. Clade identity was determined through reference to type strains and consultation of existing clade metadata. **B** The process by which genes of low and high diversity were identified. Graph shows the frequency distribution of allelic diversity values across genes within a clade's flexible core genome. Solid line shows the mean diversity of the strict core genome within the clade and the dashed lines show the second standard deviation interval of the strict core genome.

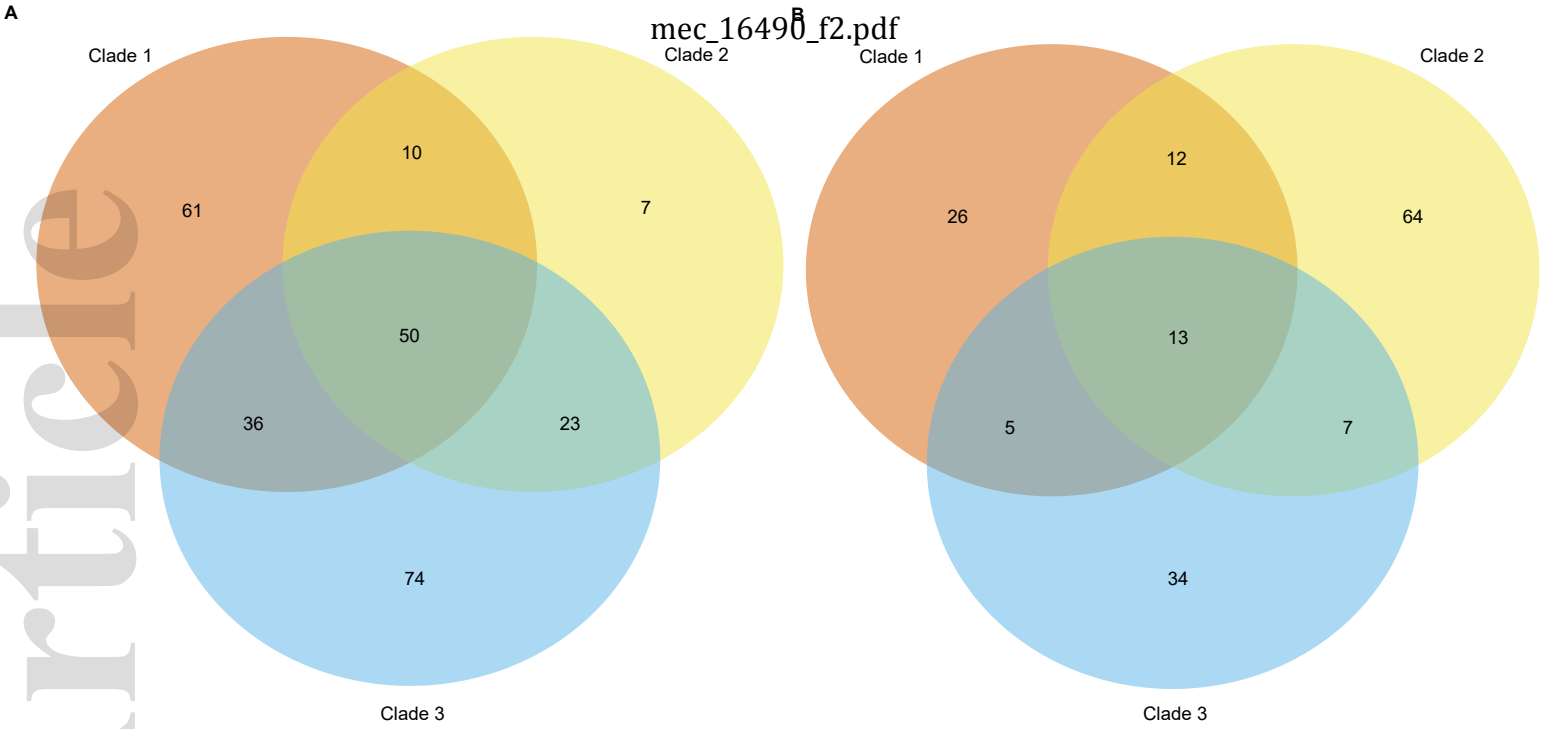

Figure 2: Venn diagram showing patterns of purifying and diversifying selection in the clade-specific core genomes of the three largest *Bc sl* clades. Numbers indicate the total number of genes that are experiencing selection (either purifying or diversifying), and location of numbers indicates whether the genes are experiencing selection in one clade or in multiple clades. **A** Genes under purifying selection (n = 261). **B** Genes under diversifying selection (n = 161).

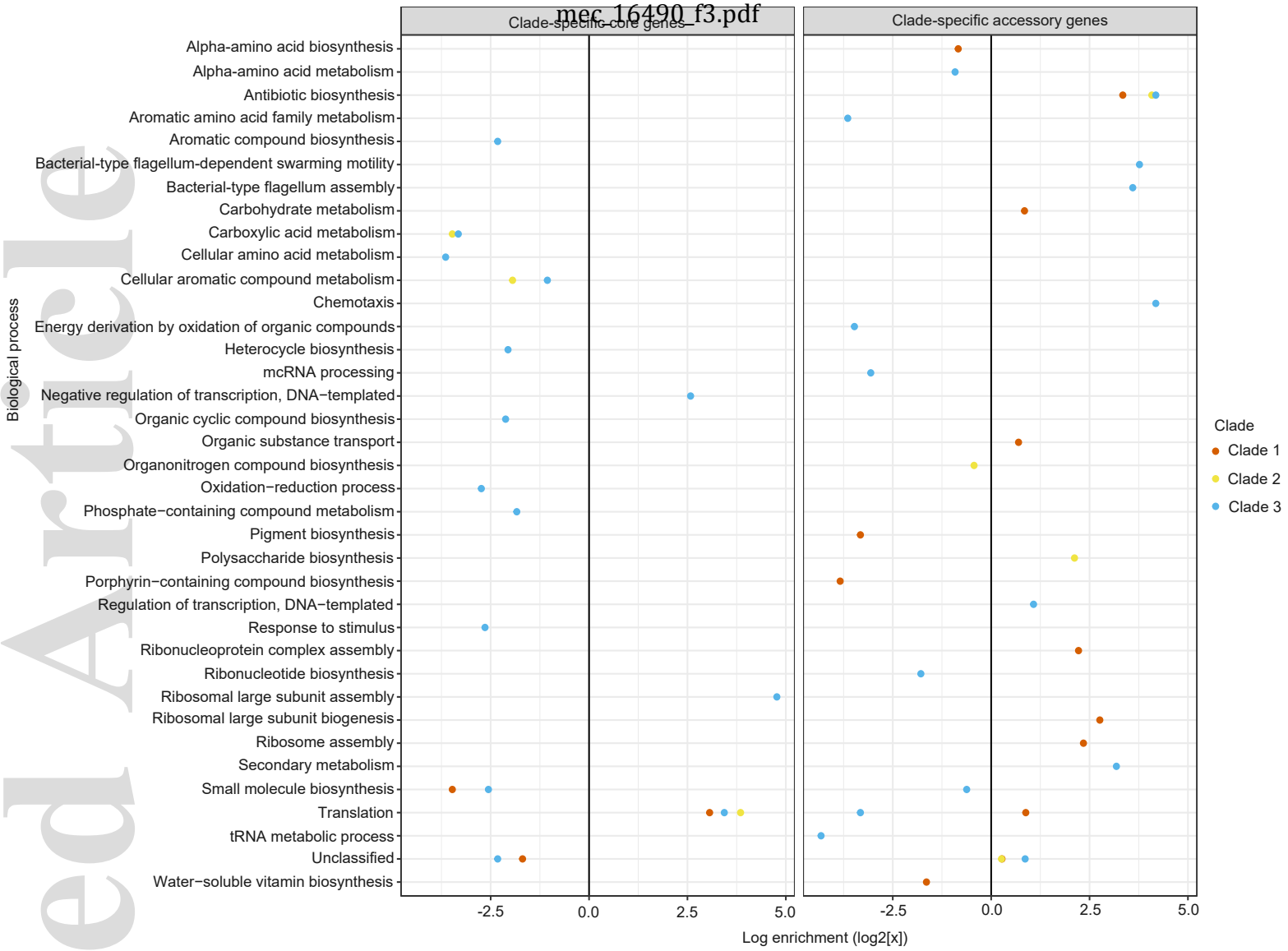

Figure 3: Significant enrichment of biological processes within clade-specific conserved core genes and clade-specific accessory genes across the major *Bc sl* clades. Enrichment values were calculated using the Gene Ontology Enrichment analysis software available online, using a Binomial test with Bonferroni correction. Only significant enrichments are shown.

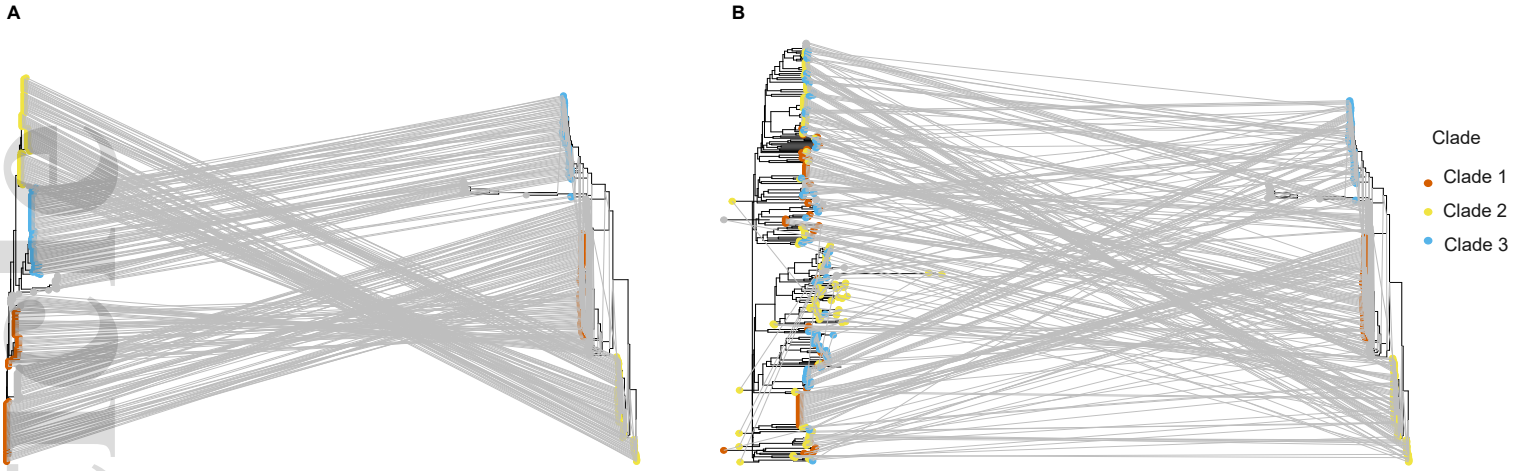

Figure 4: Consistency between phylogenies based on a single gene and a core genome. **A** Consistency between the conserved cold shock protein (*CspA*) gene phylogeny and the strict core genome (n=1004) of the *Bc sl* group. A given isolate in each tree is connected to the same isolate in the other tree by a line. **B** Consistency between the diverse flagellin (*hag*) gene phylogeny and the strict core genome (n=1004) of the *Bc sl* group. A given isolate in each tree is connected to the same isolate in the other tree by a line.

mec_16490_f5.pdf

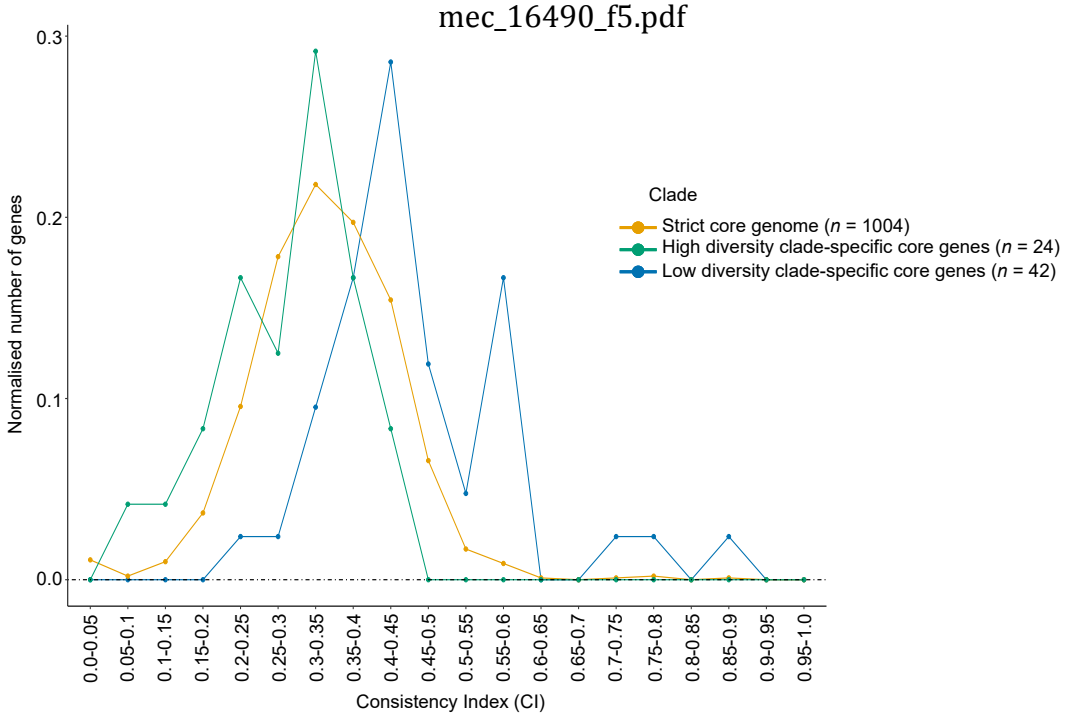

Figure 5: Consistency index distribution amongst genes under purifying selection (n = 42) and diversifying selection (n = 24). Consistency indices were calculated for each gene using the phangorn package in R and a maximum-likelihood phylogeny was created using alignments of the concatenated core genome. Average consistency index of each gene set was compared to that of the strict core genome (n = 1004). Normalised numbers of genes represent the number of genes with a given consistency index while controlling for the size of the dataset.