Functional pangenome analysis reveals high virulence plasticity of Aliarcobacter butzleri and affinity to human mucus

Functional pangenome analysis reveals high virulence plasticity of *Aliarcobacter butzleri* and affinity to human mucus

Davide Buzzanca[1,2], Cristian Botta[1], Ilario Ferrocino[1], Valentina Alessandria[1], Kurt Houf[2], Kalliopi Rantsiou[1*]

[1]Department of Agricultural, Forest and Food Sciences (DISAFA), University of Turin, Italy

[2] Department of Veterinary Public Health, Faculty of Veterinary Medicine, Ghent University, Belgium

[*]Corresponding author: kalliopi.rantsiou@unito.it, Largo Paolo Braccini 2, 10095 Grugliasco, Italy, 0039-011-6708870

Keywords: *Aliarcobacter butzleri*, virulence genes, mucus, comparative genomics, pangenome.

ABSTRACT

Aliarcobacter butzleri is an emerging pathogen that may cause enteritis in humans, however, the incidence of disease caused by this member of the *Campylobacteriaceae* family is still underestimated. Furthermore, little is known about the precise virulence mechanism and behavior during infection. Therefore, in the present study, through complementary use of comparative genomics and physiological tests on human gut models, we sought to elucidate the genetic background of a set of 32 *A. butzleri* strains of diverse origin and to explore the correlation with the ability to colonize and invade human intestinal cells *in vitro*.

The simulated infection of human intestinal models showed a higher colonization rate in presence of mucus-producing cells. For some strains, human mucus significantly improved the resistance to physical removal from the *in vitro* mucosa, while short time-frame growth was even observed. Pangenome analysis highlighted a hypervariable accessory genome, not strictly correlated to the isolation source. Likewise, the strain phylogeny was unrelated to their shared origin, despite a certain degree of segregation was observed among strains isolated from different segments of the intestinal tract of pigs. The putative virulence genes detected in all strains were mostly encompassed in the accessory fraction of the pangenome. The LPS biosynthesis and in particular the chain glycosylation of the O-antigen is harbored in a region of high plasticity of the pangenome, which would indicate frequent horizontal gene transfer phenomena, as well as the involvement of this hypervariable structure in the adaptive behavior and sympatric evolution of *A. butzleri*.

Results of the present study deepen the current knowledge on *A. butzleri* pangenome by extending the pool of genes regarded as virulence markers and provide bases to develop new diagnostic approaches for the detection of those strains with a higher virulence potential.

1. INTRODUCTION

Aliarcobacter butzleri (Basonym: *Arcobacter butzleri*) is a Gram negative bacterium belonging to the *Campylobacteriaceae* family often isolated from human stool, animal feces, drinking water, and food [1,2]. It is the most widespread species within the genus *Aliarcobacter* and is considered as an emerging pathogen, transmissible from livestock through the food of animal origin [1,3,4]. In this frame, *A. butzleri* has been isolated from healthy pigs, indicating a possible direct and indirect (cross-contamination dynamics) source of infection mediated by pork [5,6]. Spreading of this pathogen along the food chain is favored by its capability to survive in cold environments [7].

A. butzleri pathogenesis for humans is recognized but the underlying mechanisms are still largely unknown [1,8]. *In vitro* tests, using human cell line models have been employed to simulate adhesion and invasion and infer the virulence potential of strains [9]. Although this approach is a simplification of gut systems, it remains fundamental in the phenotypic investigation of host-pathogen interaction [10]. In this context, the intestinal mucus appears to be relevant and may influence the ability of *A. butzleri* to adhere and invade [11]. The mucus is composed mainly of glycoproteins and is present in different organs such as the stomach and gut. The number of mucus layers is variable along the intestinal tract; in the small intestine and in the colon are present one and two mucus layers, respectively [12]. The presence of mucus on the gut tissue is an important factor that has been shown to influence the development and behavior of intestinal bacteria [13].

Survey studies have been performed to isolate *Aliarcobacter* spp. from different environments, animals and foods [1,8]. Isolates so far have been mainly genetically characterized for their virulence potential, focusing essentially on the presence of putative virulence genes that have been identified based on sequence similarity to other pathogens but without a biological confirmation of their role in pathogenicity [14,15].

The objective of this study was to characterize 32 *A. butzleri* strains, selected based on their source origin, by combining Whole Genome Sequencing (WGS) and physiological data of colonization and invasion of Caco-2 (*Homo sapiens*, Caucasian colon adenocarcinoma), and HT29 MTX (*H. sapiens*, Caucasian colon adenocarcinoma treated with methotrexate), a mucus producer cell line.

2. RESULTS AND DISCUSSION

2.1 SIMULATED INTESTINAL COLONIZATION IS ENHANCED BY HUMAN MUCUS

Thirty-two *A. butzleri* strains previously collected from human and animal feces, pig intestine, animal skin and meat (**Table 1**) were tested on human gut models to define their capability to colonize (cell association) and invade intestinal cells. More specifically, the mixed culture of Caco-2 /HT29-MTX cells and Caco-2 were used as mucus producing (MP) and not-mucus producing (NMP) models, respectively. All *A. butzleri* strains were able to colonize both models after 90 minutes of co-incubation, while strains 7 (isolated from bovine), 26, 28, 31 (isolated from Human feces) presented a not detectable invasion on the MP models; moreover, invasion by strain 28 was not detectable also on NMP models **(Fig. 1)**. However, only a minor part of the bacterial cells colonizing the models (0.64 ± 0.34 % on average) invaded the human cells, corresponding to an average decrease of 3.7 ± 1.5 Log CFU/cm^2 from the initial inoculum **(Supplementary Table 1)**. The multiplicity of infection (MOI) was not the same for all strains. Although this parameter has been shown previously to play a role in the transepithelial resistance of cell lines after 48 hours of contact with *A. butzleri* [16], it appears not to have an effect during adhesion studies, conducted under short term (1-3 hours) of bacterial contact [17,18]. These Overall, our data are in agreement with previous reports and confirm the ability of *A. butzleri* to colonize different cell lines with an invasion efficiency similar to the phylogenetically close species of *Campylobacter jejuni* [19–22].

Comparing colonization data (expressed as ΔLog CFU/cm^2) from MP and NMP models, *A. butzleri* showed an overall greater ($P < 0.001$) colonization capability in presence of human mucus **(Fig. 1A)**. The presence of the human mucus glycoproteins enhanced the colonization capability of all strains,

but significantly ($P < 0.001$) only for three isolates from pig intestine (strains 16, 17, 19). Other than that, no relationship between the two main sources of isolation (human stool and pig intestine in its various sections) and the colonization trend was observed **(Fig. 1B)**. Strains from these two sources highlighted an equal proportion of highly colonizing (positive or close to zero ΔLog) and low colonizing phenotypes (negative ΔLog), regardless of the model used. Positive values observed for some strains suggest bacterial growth during the host-pathogen interaction timeframe, again more evident in presence of mucus. Considering each model, all strains have shown comparable colonization abilities. Finally, considering differences between strains, strain 2 colonized statistically more ($P < 0.05$) than strain 31 in the NMP model.

The effect of mucus in enhancing colonization has already been observed in *Aliarcobacter butzleri* [19]. This is not surprising, since it is a hallmark of intestinal pathogens [13], which must overcome the mucus in order to exert the infection in the host [23]. In this frame, an *in vivo* survey suggested a chemoattractant function of the mucus towards *Aliarcobacter* spp. since it was recovered not only from the inner content but also from the mucus layer of pig intestines [5]. The statistically significant higher mucus-model colonization observed for part of the pig isolates also suggests a rather strain-dependent mucus affinity that may result in exploitation of its protective action against intestinal peristalsis under *in vivo* conditions.

2.2 FUNCTIONAL CHARACTERIZATION OF PUTATIVE ENCODED PROTEOMES

All *A. butzleri* strains were *de novo* sequenced, assembled and subjected to whole-genome comparative analysis. Genomes obtained display a GC content between 26.74 and 27.11 % and a length ranging from 2.04 Mb to 2.50 Mb. We observed several genes/proteins belonging to incomplete (< 60 % of similarity) and questionable (< 90 %) prophage regions, but no intact known prophage region was found in the 32 genomes. Clustered regularly interspaced short palindromic repeat (CRISPR) sequences are present in 23 genomes, of which only ten CRISPR regions are flanked

to CRISPR-associated protein (CAS; general class 1 and 2) sequences. Always concerning mobile genetic elements and signatures of bacteriophages, at least one transposase gene was found in 27 of the 32 genomes (**Supplementary Table 2**). Importantly, the presence of numerous mobile genetic elements are markers of a former evolution and potentially improved fitness, which is important for any pathogen [24].

Classes of COG (Clusters of Orthologous Groups) are homogeneously distributed along the 32 putative encoded proteomes (**Fig. 2A, Supplementary figure 1**). We observed a remarkable abundance (19.4 % on average) of proteins with unknown function or characterized only for general functions, a fact that highlights the limited characterization of *A. butzleri* proteome [8]. Following, signal transduction mechanisms (average of 10.14%) is the most abundant characterized COG class, suggesting the presence of an extended network of control of *A. butzleri* functions [14,15,25]. Predicted proteins involved in the metabolism and transport of amino acids are significantly more abundant (9.16%) than those related to carbohydrate if compared to the COG distribution observed in other bacteria [26]. This is consistent with the limited or null consumption of carbohydrates shown by *A. butzleri* and other *Aliarcobacter* species, which instead utilize organic acids and amino acids as main carbon sources [14,25]. Moreover, the classes of signal transduction mechanism and cell motility represent more than 10 % of the predicted proteins, whereas only 14 genomes (**Supplementary Figure 1B**) harbor genes involved in bacterial cytoskeleton function (COG class Z), which are linked to bacterial motility too [27]. It is noteworthy that proteins involved in motility play a pivotal role in the host/pathogen interaction since the bacterial movement in the gel-like matrix such as the shallow mucus layer can allow a faster pathogen infection [28].

2.3 GENOME-WIDE ANALYSIS SHOWS AN OPEN PANGENOME

The core- and accessory-genomes sizes were estimated (**Table 2**) by clustering the predicted aminoacidic sequences of the 32 annotated genomes through two pangenome computing programs

[29,30]. In both cases the accessory genome resulted to represent from 78% to 75% of the pangenome, comprising most of the hypothetical proteins (up to 90 %) and composed of more than 55 % of singletons (gene family exclusively present in one genome). Similar partitioning of *A. butzleri* pangenome has recently been observed on a set of 49 genomes [15]. This leads us to speculate a wide and open pangenome, which reflects a sympatric evolution with frequent episodes of horizontal gene transfer (HGT), like the exchange of genes involved in pathogenesis and antibiotic resistance, that can confer an adaptive advantage in changing environments [31]. As done for other foodborne pathogens [32,33], with the increasing number of available genomes a large-scale pangenomic survey will be needed soon to confirm these first observations.

The number of genomes here investigated is adequate to infer the structural organization of the pangenome with Markov Random Field networks (**Fig. 3A**), which display the localization of each gene family (nodes) by following a pattern of continuity (edges connect loci that are frequently neighbors) regardless from contigs succession [30]. As previously observed during the validation of this approach on a large set of *Acinetobacter baumannii* genomes [30], the pangenome of our 32 strains shows organized clusters of persistent gene families (present in more than 95 % of the strains) either surrounded or, less frequently, interrupted by islands of dispensable genes (shell and cloud genome). It is noteworthy the presence of a large pangenome plasticity island that represents hotspots of alternative structural organizations along the genomes analyzed, thus possible sites of HGT (**Fig. 3B**). In addition to a predictable presence of hypothetical and not functionally characterized proteins, this island encompasses several accessory gene families generally involved in the COG class of cell wall/membrane and envelope biogenesis, besides proteins more specifically associated with pili/flagella glycosylation, LPS glycosylation/assembling and exopolysaccharide (capsule) secretion (**Fig. 3C**). Additionally, we detect up to 323 regions of genome plasticity (RGP) that can be referred as genomic islands shared by at least two of the genomes [34,35]. As observed for the whole pangenome, these RGP overall encompass gene families involved in the cell wall/membrane and envelope biogenesis, which however represents the second class after genes involved in replication,

recombination, and repair of DNA (**Fig. 3C**). The presence of genes involved in the biosynthesis of capsular polysaccharide, lipopolysaccharide (LPS) glycosylation and flagellin/pilin glycosylation within the island of pangenome plasticity and the RGPs is not surprising, since are accounted as dispensable genetic structures that can be acquired or lost, to face host-to-host transition and colonize new ecological niches [36,37]. For instance, loss of the flagellin glycosylation genes may determine phenotypic changes that decrease recognition of strains by the host–immune system [38], while the polysaccharide chain of LPS (inserted between hydrophobic lipid and hydrophilic O-antigen) possess hypervariable structures that reflect the specific pathogen serological signature [39].

Other genomic regions were entirely constituted by singletons and in some genomes (strains 18, 17, and 3) represent large sections of it (up to 20,000 bp), containing hypothetical proteins, mobile genetic elements (phage proteins, transposases, recombinase) and genes with poorly defined functions. Interestingly, few or no singletons are found in pig duodenum/caecum isolates and 8 of the 9 genomes from pig rectum (**Supplementary Figure 2**). On the contrary, by excluding the singletons we observed that 425 and 140 accessory gene families are significantly (Scoary statistics; *P* [FDR] < 0.05) overrepresented in the genomes of pig rectum and pig duodenum/caecum isolates, respectively. Besides, no overrepresentation of gene families is observed in each of two major ecological sources of isolation, i.e. human stool and all pig intestine. Accordingly, the phylogenetic trees (**Fig. 4**) do not show a clear segregation between these two groups, but regardless of the type of input sequences (i.e. the whole genomes, core genomes, SNPs, and MLST loci) a recurrent clustering pattern that consists of group I (strains 14 and 15, from duodenum and caecum of pig), group II (human strains 1 and 28), group III (strains 12, 19, 20, 21 from pig) and group IV (strains 11, 13, 23 from pig rectum) was observed.

Considering their genomic plasticity (high level of intra-group shared genes, low or absent singletons) and phylogenetic analysis, the isolates from pig duodenum/caecum (group I) and rectum (group III e IV) represent three distinct lineages. This aspect also suggests that distinct genotypes of *Aliarcobacter butzleri* may colonize specific segments of pig intestine, as already observed at the species level for

Aliarcobacter spp. [5]. Moreover, the low genomic plasticity of these three groups and the fact that pig intestine (particularly the rectum section) can be a favorable niche for this pathogen with limited or absent symptoms in the host [40], lead us to speculate host and/or tissue tropism phenomena for these three groups [41]. On the contrary, the remaining strains seem to have undergone more episodes of HGT, likely in reason of frequent host transition events and developed a more host-generalist genotype [42]. In light of our pangenomic observations, *A. butzleri* may represent a pathogen with both host-generalist and host-specialist phenotypes, which can alternately arise in livestock in response to external selective pressures (antibiotics, intensive breeding) and then transmitted to humans, as recently reported at large-scale for *Campylobacter* spp. [38,41].

2.4 REPERTOIRE OF VIRULENCE GENETIC TRAITS

To detect possible genomic signatures linked to *A. butzleri* virulent phenotypes we manually curated the annotated genes by focusing on those sequences putatively associated with host-pathogen interaction in this and other pathogenic bacteria (**Supplementary Table 3**). This produced a list of 100 genes (of which 39 are accessory genes) putatively involved in functions related to human mucosa adhesion/invasion, interaction with host mucosa/mucus, flagellum and motility, as well as proteins more widely correlated to virulence of *A. butzleri* and other pathogens, such as hemolysis, secretion and regulatory systems (**Fig 5**). It is noteworthy that the 32 presence/absence profiles (Pearson's correlation-based dendrogram) cluster as previously observed in the whole-genome phylogenetic dendrogram and almost all these genes are included in accessory gene families of the pangenome. Accordingly, we might speculate that the biodiversity within *A. butzleri* populations are partly shaped from the exchangeable virulome as a genomic tracking of the host-to-host transitions undergone by each strain. Nevertheless, strains origin and other genes not directly involved in virulence mechanisms may play an important role in the phylogenetic segregation of the strains.

2.4.1 GENES COMMONLY RECOGNIZED AS VIRULENCE FACTORS

As first step, we focused on ten genes commonly employed as markers to assay the virulence potential of the *Aliarcobacter* genus [20,43]. Genes correlated to adhesion (*cadF*), invasion (*cj1349*, *ciaB*) and hemolysis (*pldA*, *tlyA*, *mviN*) are present in all the genomes but, except for *pldA* and *tlyA*, were initially annotated as different proteins.

On the other hand, the gene *hecA* (hemagglutinin alternatively annotated as *shlA* or *hpmA*) and *hecB* (hemolysin activation protein here annotated as *shlB*) are present and adjacent to each other in 31 % of the genomes, while *iroE* (encoding a periplasmic enzyme and annotated as *besA*) and the generic virulence factor *irgA* were found in 75 % and 81 % of the genomes, respectively. According to PCR based studies and other genomic surveys [44,45], these latter four genes are less prevalent across the whole *A. butzleri* population. Moreover, the presence of these four genes in our strains do not significantly correlate (Pearson's moment correlation, P [FDR adjusted] > 0.05) with their colonization phenotypes. This is not surprising since they encode for functions useful in following infection phases [46]. Regardless of their impact on the colonization, the initial misannotation observed for most of these genes underlines once more their high polymorphism, which often leads to underestimating virulence potential and diffusion of *A. butzleri* due to false negatives amplifications [14,15].

2.4.2 GENES RELATED TO ADHESION AND INVASION

An important gene involved in the colonization, specifically in the host mucosa adhesion, is *porA* that encodes for a major outer membrane protein responsible for the hypervirulence of *C. jejuni* [47]. Here it was found in all genomes, properly annotated or indicated as putative gene for *Campylobacter* major outer membrane protein. In *A. butzleri* the high polymorphism of this gene and its flanking regions have been recently proposed as a meaningful signature of pathogenicity, not related to the shared ecological origin and whole genome phylogeny [15]. The UPGMA dendrogram of the 32 *porA* sequences (**Supplementary Figure 3**) partially confirms the previous observations, with a grouping

pattern unrelated to the initially shared origin, but instead resembling the phylogenetic segregation previously described (for instance the recurrent groups I, II, III, IV).

Another ubiquitous gene is the *inlJ*, which encodes in *Listeria monocytogenes* for a protein of the LPXTG-internalin family and is involved in host adhesion and invasion [48]. However, other orthologues of the *Listeria monocytogenes* internalin operon are missing in the genomes of all but strain 29 that encompass the internalin A in a different genomic region. The absence of internalin orthologs seems to correlate well with the aforementioned limited invasiveness of *A. butzleri* when compared to *Listeria monocytogenes* [49].

As here phenotypically confirmed, *A. butzleri* can penetrate and likely move throughout the human mucus (**Fig. 1A**). The mucus, having a protective function towards the intestinal epithelium (in our case the cell model layer), must be overcome to allow colonization, capacity observed in the 32 *A. butzleri* strains object of study (**Fig. 1B**) [12]. However, only one gene (encoding an Arylesterase precursor) linked to mucus degradation was detected. Two different Arylesterase forms are present in the genomes, not correlated with greater or lower colonization in presence of mucus (P [FDR] > 0.05) [50].

2.4.3 SECRETION SYSTEMS INVOLVED IN PATHOGENICITY

Several genes of our proposed virulome are part of secretion systems and can play a role in the host-pathogen interaction. Among these, the operon encompassing genes *epsE*/*epsF* and the *xcpO* gene are part of a type II secretion pathway fundamental in the infection mechanism of *Vibrio cholerae* [51], albeit numerous components of the original operon are missing in *A. butzleri* genomes. Similarly, genes (*epsD, epsH epsM, epsN, epsJ*) responsible for exopolysaccharide secretion and biofilm-forming capability in *E. coli* are present, but not organized in a single operon [52]. Some molecules linked to biofilm production can promote bacterial adhesion on human intestinal cells [53].

In this light and considering that *A. butzleri* is proven to form biofilm [54], further investigation to define its *eps* genes role and regulation is now needed.

Moreover, six genes (*virB10, virB8, virB6, virB4, virB3, virB*), encoding a rare T4SS structure (type IV secretion system), recently described in *A. butzleri* [15], were annotated in strain 18. Differently from the previous observation, these genes are not comprised in a single genomic region but are instead spread along with the several islands of singletons found in this genome. The T4SSs are an important virulence mediator in different pathogens, including *C. jejuni*, since are connected to host cell apoptosis, cytotoxicity, bacterial cell survival, adhesion, and invasion to host cell [55–57]. Anyhow, this peculiarity did not result in a greater colonization or invasion activity for this strain.

2.4.4 GENOMIC SIGNATURES RECOGNIZED BY HOST IMMUNE RESPONSE

A consistent fraction of putative virulence genes are involved in the flagellar assembling/motility (flagellins), chemotaxis and urease activity (indirectly responsible for increase of external pH), mostly organized in clusters or anyway located in the same genomic regions [14,15]. In particular, the flagellar proteins are important virulence factors related to human pathogens motility in the proximal mucus layer and recognition by host immune response [13]. Thus, it is intriguing that gene families encoding flagellins are included in the core genome of *A. butzleri*, while we found genes responsible for their glycosylation in the accessory and plastic genomic regions. This suggests the heterogeneous glycans compositions of flagellum may lead to a strain-specific antigenic fraction of this bacterial component [58].

LPS O-antigen plays a pivotal role in the pathogen survival on the human mucosa, modulating host immune response and counteracting its defense mechanisms [11].

All 32 genomes contain at least one copy of the O-antigen ligase gene, of which polymorphism follows the whole-genome phylogeny (formerly groups I, II, III, IV) and goes hand by hand with the structural organization of the surrounding genes (**Supplementary Figure 4, Supplementary Table 4**). Interestingly, in 32 genomes we observed up to 25 different genomic structures flanked to O-

antigen ligase that encompass genes involved in LPS O-antigen assembling [59], such as lipid A biosynthesis protein (*msbB*), LPS transferases (*rfaC*), sugars/glycosides transferases/epimerase/reductase (*rfaF, sunS, pglJ, lacA, epsJ, rfbB, rfbC, rmlD, kfoC*). The lack of a single component of the O-antigen genes cluster (ABC transporters, glycotransferase, glycosyltransferase) can dramatically affect the infectiveness of Gram-negative pathogens [60,61]. Therefore, the role of such variability regarding the genes flanking the O-antigen ligase genes in pathogenicity deserves further investigation. Indeed, the intraspecific complexity of the O-antigen pathway, already noticed in four *A. butzleri* genomes [62], and here confirmed by a large scale genomic comparison, highlights this region of the plastic virulome as one of the most useful to define strain-specific virulence signatures.

2.4.5 GENES INVOLVED IN MULTIPLE VIRULENCE MECHANISMS AND REGULATION

Other meaningful elements of the *A. butzleri* virulome (**Fig. 5**) are represented by membrane components, like TonB transport protein (different protein forms and domains) and the transport complex ExbB/ExbD, which are required for *Shigella dysenteriae* and *E. coli* invasion/spread in human cells [63,64]. Invasion ability shown here and, even more, the capability of *A. butzleri* to cause septicemia by spreading in human cells may suggest similar functions of these genes in *A. butzleri* [1]. Moreover, TonB is involved in the iron uptake as *irgA* [65], by suggesting its possible role in hemolysis [66].

Particularly relevant is the presence of *phoP* and *phoQ* genes (respectively encoding the transcriptional regulatory protein PhoP and sensor protein PhoQ), which constitute a two-component signal transduction system able to regulate intracellular virulence, cell envelope composition, and the within-host lifestyle in Gram negative bacteria [67,68]. Twenty-two genomes contained at least one form of *phoP*, while *phoQ* was only found in eight genomes and not flanking the *phoP* gene. However, several genes encoding for proteins with potential homologous function to *phoP* or *phoQ* were found flanking or nearby the gene encoding for the respective complementary protein. For

instance, in all eight genomes, the *phoQ* gene is located next to the gene *mprA* encoding a transcriptional factor. Interestingly, when this transcriptional factor is joined by the *mprB* gene (regulatory system *mprB*/*mprA*) the *Mycobacterium tuberculosis* infection increase in its persistence [69]. Considering that *mprA* exerts a transcription regulation activity comparable to *phoP*, the genomic continuity with *phoQ* suggests a possible homologous function. On the other hand, several sensor proteins with potential histidine kinase function homologous to *phoQ* are flanking the gene *phoP*, such as the genes *fixL*, *zraS*, *pdtaS* (strain 25), and *ttr* (strain 29). Moreover, several strains harbor genes encoding phosphorelay sensor kinase activity that regulates PhoP-PhoQ in other bacteria, such as the virulence sensor *bvgS* [70] or the couple of genes *dsbA/dsbB*, here annotated as DSBA-like thioredoxin domain protein and thiol-disulfide oxidoreductase *resA*, respectively. Particularly, this latter two-component system activates the *phoP* gene in *E. coli* [71].

In terms of virulence phenotypes, we did not observe a significant correlation (P [FDR] > 0.05) between the colonization/invasion data and the genomic occurrences of *phoP*/*phoQ* or the alternative two-component system above described. Nevertheless, the impact of this signal transduction system on the pathogen phenotype is dependent also on upstream regulators/activators and downstream triggered virulence genes [72], which in *A. butzleri* need to be characterized with future transcriptomic investigations.

3. CONCLUSION

The attention of the scientific community towards *A. butzleri* is significantly rising in the last years, with a parallel increase of concerns about its genomic flexibility, virulence predisposition in humans, adaptability to different hosts. In this frame, we focused our efforts on the first two issues by exploiting the pangenomic approach as an advanced comparative tool, integrating the genomic data with physiological tests on a set of strains tested with human gut models with and without mucus.

In summary, *A. butzleri* strains have shown a similar capability to colonize *in vitro* the human mucosa by adhering and even proliferating within human mucus, without showing marked invasiveness. Notwithstanding, it is not clear if a commensal lifestyle within mucus is conceivable in humans. In

pigs, asymptomatic infections suggest that it may have developed a host specialist lifestyle and such hypothesis is supported by the genomic data of this study. In this context, the open pangenome and the interchangeability of potential virulome have been recently demonstrated and proposed as key genomic features for the host adaptation of this pathogen. Here, also, to confirm these first findings, we link the variable virulome to strains phenotypes, by identifying in the LPS assembling pathway one potential strain-specific signature. Despite the intrinsic limit of pangenomic based comparison that does not necessarily permit to exhaustively explain the multifaceted virulence mechanism of *A. butzleri*, we pointed out and described the presence of putative virulence promoters and antigen recognition markers, such as master regulators.

Therefore, these outcomes will provide concrete guidelines for more comprehensive omics investigation of the *A. butzleri* lifestyle in human mucosa.

4. MATERIALS AND METHODS

4.1 BACTERIAL STRAINS

The *A. butzleri* strains (**Table 1**) were obtained from the Belgian Coordinated Collection of Microorganisms (BCCM; Laboratory for Microbiology, Ghent University, Belgium) isolated from different sources, and stored in Laked Horse Blood (Oxoid, Basingstoke, Hampshire, UK) at -80 °C. Cultivation was performed in microaerophilic conditions at 30°C on agarized Arcobacter broth (CM0965, Oxoid) supplemented with C.A.T supplement (SR0174, Oxoid) [73].

Before each experiment, a single fresh colony was resuspended in Arcobacter broth and incubated at 30 °C for 48 hours. Afterward, 0.5 ml of culture was inoculated on Arcobacter plates supplemented with C.A.T supplement, grown for 48 hours in microaerobic conditions, collected with 1 ml of Ringer's solution (1.15525, Millipore, Burlington, Massachusetts, U.S.A) and thus used as working suspension in the interaction experiments. The bacterial load (Log CFU mL^{-1}) of each working

suspension was determined by measuring OD at 630 nm with ELx880 microtiter plate reader (Savatec, Turin, Italy) and set to the same initial count by an internal standard curve.

4.2 CELL LINES AND HUMAN GUT MODELS

Human colon carcinoma cell lines Caco-2 (86010202, ECACC, European Collection of Authenticated Cell Cultures, Public Health England), HT29 (ATCC® HTB38, ECACC) and HT29 MTX (12040401, ECACC) were cultured in Dulbecco's Modified Eagle's Medium (DMEM 6429; Sigma-Aldrich, St. Louis, Missouri, USA) supplemented with 10% of fetal bovine serum (FBS; F7524 Sigma-Aldrich) and EmbryoMax Penicillin-Streptomycin Solution, 100X (TMS-AB2-C, Sigma-Aldrich). The cell lines were grown in 25 and 75 cm^2 culture flasks (Corning, New York, New York USA) at 37 °C in a humidified atmosphere containing 5% CO_2 and 95% air and sub passaged every 3-4 days (Eppendorf, Galaxy 170 S, Hamburg, Germany) [74].

Two *in vitro* monolayer human epithelial structures were prepared: a mucus-secreting (MP) co-culture of differentiated Caco-2 and HT29-MTX cells in a 9/1 ratio; and two non-mucus-secreting cell models (NMP) represented by a single culture of differentiated Caco-2 cells and a mixed model of Caco-2 and HT29 cells with the same ratio of MP model [75]. Briefly, the cells were seeded at a density of 35,000 cells cm^{-2} and grown in complete culture medium under the same conditions described above, for 14–15 days with regular changes of the media, until functional polarization was reached and models could be considered differentiated and ready for the experiments [76]. Before (3-4 days) the assessment of strains colonization and invasion capability, the MP and NMP models were washed twice with PBS 1X and the complete culture media was replaced with media without antibiotics to allow the pathogen growth.

4.3 ASSESSMENT OF COLONIZATION AND INVASION CAPABILITY

The working suspensions of the strains were inoculated on MP and NMP cell models. Depending on the growth capacity of the individual strains, different inoculum levels could be experimentally achieved; in the majority of cases the density of bacterial suspensions was 7-8 Log_{10} CFU mL^{-1} (**Supplementary Table 1)**. Due to this experimental limitation, the multiplicity of infection (MOI) was not the same for all strains tested. Colonization-invasion assays were performed on two different model wells for each biological replicate. After 90 minutes of co-incubation at 37 °C in a normal atmosphere, the not adherent bacteria were removed by two washing steps with PBS 1X. Colonization and invasion capabilities were evaluated in parallel on MP and NMP models on at least three biological replicates.

To quantify the colonization capability (also defined as cell association), which represents the pathogen ability to adhere and enter the human cells, *A. butzleri* cells were recovered from one duplicate of the cells model by incubating for 30 minutes with 1 mL cm^2 of 0.25 % Triton X-100 (v/v; in PBS 1X). Counts of the resulting suspension were performed employing the CFU method, plating the dilutions on solidified Arcobacter broth supplemented with C.A.T supplement for 48 h at 30 °C in microaerobic conditions.

In parallel, to define the invasion capability (number of bacterial cells that penetrate in the human cells excluding those adherents) the culture media supplemented with 300 μg ml^{-1} of gentamicin sulfate (G1914, Sigma-Aldrich) was added in the cell models for 120' at 37 °C to kill all the extracellular bacteria. After two washing steps with PBS, the internalized viable cells of *A. butzleri* were recovered and enumerated as described for total colonization [20,77,78].

Raw counts data were expressed as Log CFU cm^{-2} of bacteria inoculated (T0), bacteria colonizing the model after washing steps (Tc) and after gentamicin treatment (Ti). Colonization was expressed as Δ Log CFU cm^{-2}, by following the formula: Log CFU $mL^{-1}{}_{Tc}$– Log CFU $mL^{-1}{}_{T0}$. Invasion capability was expressed following the formula: Log CFU $mL^{-1}{}_{Ti}$– Log CFU $mL^{-1}{}_{T0.}$

4.4 GENOME SEQUENCING, ANNOTATION AND BIOINFORMATIC ANALYSIS

Genomic DNA (gDNA) extraction of *A. butzleri* strains was performed by the beads-beating, phenol-chloroform DNA extraction method followed by a RNAse A (5 µg μl^{-1}, MRNA092 Epicenter, Madison, Wisconsin, U.S.A) treatment to digest RNA in the DNA samples, with an incubation of 30 min at 37 °C. The DNA quantification was performed with the employment of Nanodrop (ND 1000, Thermo SCIENTIFIC). The gDNA quality check, to confirm the absence of degradation and impurity, was performed through an electrophoretic run (100 V for 30') on agarose gel 0.8% (w v^{-1}, 0710 VWR) in TAE 1X (Tris – Acetic acid – EDTA, K915 VWR), gelRed (41005, Biotium) was used as DNA intercalating.

Whole genome sequencing (2X150bp, coverage 100X) was performed on Illumina Novaseq 6000 machine by the Novogene company (Cambridge, United Kingdom). Briefly, after a Qubit 2.0 quantification 1 µg of gDNA was used for the library preparation using NEBNext® library prep Kit, randomly fragmented (350 pb) by shearing and then the samples were polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and PCR enriched by P5 and indexed P7 oligos. The PCR product purification was performed with the use of the AMPure XP system, afterwards, the libraries were analyzed by Agilent 2100 Bioanalyzer (size distribution) and quantified using real-time PCR.

Sequencing reads were quality filtered with Solexa QA++ software, and sequences less than 60 bp and dereplicated sequences were removed by Prinseq.

Reads were *de novo* assembled with SPAdes (version 3.11.0) [79] and the quality of the contigs was checked with QUAST software to obtain statistics related to the genomes assembly process and data quality, such as coverage, total genome bp length and the number of contigs [80] (**Supplementary Table 2**).

Genomes were annotated using the Prokka (version 1.11) suite [81] and putative encoded proteins have been manually checked through on UniProt (https://www.uniprot.org/), Pfam

(https://pfam.xfam.org/), and CDD database (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) to understand their functional role [82]. The CRISPR-CAS sequences have been detected with the software CRISPRCasFinder 1.1.2 (https://crisprcas.i2bc.paris-saclay.fr/), while phage sequences were retrieved with Phaster (https://phaster.ca/) [83,84]. Additional analysis on the metabolic pathway was performed on the putative predicted proteome with the software RPS-BLAST 2.2.15 on WebMga (http://weizhong-lab.ucsd.edu/webMGA/), to obtain the related COGs (Clusters of Orthologous Groups) codes and classes [85–87].

Proteins inferred by Prokka were then processed with the parallel use of Roary (version 3.13.0) and PPanGGOlin (version 1.1.85) with default parameters to generate the presence-absence binary matrices of core and accessory genes [29,30]. The structural settlement of the loci (gene families) in the pangenome was inferred through the matrix generated by PPanGGOlin and visualized using the program Gephi 0.9.2-beta (https://gephi.org). The presence of regions of genome plasticity (RGP) has been detected from PPanGGOlin output through the script *ppanggolin rgp -p pangenome.h5* [35]. Associations between binary matrices (presence/absence) of accessory gene families or RGP (singletons excluded) and the main sources of isolation (human stool, pig intestine, and its main sections) were assessed with Scoary scripts [88] and considered significant for *P*-value [FDR adjusted] < 0.05.

Moreover, with the purpose to explore all possible virulence-associated genes present in the genomes, we constructed a repertoire of genes linked to host/pathogen interactions (mucus interaction, adhesion, invasion, modulation of host genes), chemotaxis, motility and general factors related to virulence mechanisms (**Supplementary Table 4**). Genomes, and genes detected using the software described above, were manually curated, and the presence of sequences of interest has been confirmed by BLAST alignment towards reference sequences (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [89].

4.5 PHYLOGENETIC ANALYSES

Phylogenetic UPGMA trees were computed for whole and core genomes of the strains object of the present study using the software ND tree (version 1.2) with the *Campylobacter jejuni* NCTC 11168 (NC_002163.1) genome as outgroup. The MLST sequences were analyzed with the software clustalX (Multiple aligned modes, version 2.0) [90].

An *in silico* MLST analysis has been performed employing the on-line suite MLST 2.0 (https://cge.cbs.dtu.dk/services/MLST/) for all strains [91], by using the sequences used for *Aliarcobacter* spp., specifically *aspA, atpA, glnA, gltA, glyA, pgm, tkt* [92]. After the obtainment of the MLST numeric codes **(Supplementary Table 5)**, the MLST sequences of all strains were stored in FASTA format for phylogenetic analysis.

The Approximately-maximum-likelihood phylogenetic tree of SNPs present in the 32 genomes was produced with the type genome of *A. butzleri* RM4018 synonymous of LMG10828^T as reference (NC_009850.1), using the CSI Phylogeny pipeline (Call SNPs & Infer Phylogeny, CGE, version 1.4) with default options. SNPs detected by the software CSI Phylogeny have been checked with BWA (version 0.7.17) and Samtools software (version 0.1.19) [93].

Phylogenetic trees were visualized with iTOL (version 5.5.1) to obtain the image format choice [94], while the software Morpheus (https://software.broadinstitute.org/morpheus) was used in the heatmap production [95].

4.6 STATISTICAL ANALYSIS

Correlation between presence/absence of virulence-associated genes and colonization/invasion rates was computed by Pearson's product-moment correlation (considered significant for *P*-value [FDR adjusted] < 0.05) in R environment.

Normality and homogeneity of the data from colonization and invasion assays were checked using Shapiro-Wilk's W and Levene's tests, respectively. Kruskal–Wallis (K-W) and ANOVA were used to assess the overall variation and differences between the multiple groups, for nonparametric and parametric data respectively. Pairwise Wilcoxon's test and Duncan's test were used as post hoc

analyses for nonparametric and parametric data respectively. Data were presented in boxplots graph (median, range interquartile, min/max and outliers). Statistics and data plotting were performed with the R program for Statistical Computing 3.6.0 (http://www.r-project.org) unless otherwise stated.

5. AVAILABILITY OF DATA AND MATERIAL

Raw sequence reads were deposited at the Sequence Read Archive of the National Center for Biotechnology Information (Bioproject accession number: PRJNA660594). The genomes assembled sequences, the sequences of the predicted transcripts and amino acidic sequences (.faa, .fna, .gff, .gbf, .sqn, .tbl, .ffn), and the files used to construct the pangenome network (edges.csv, nodes.csv) are available on Zenodo (https://zenodo.org/) at http://doi.org/10.5281/zenodo.4301795.

6. AUTHOR CONTRIBUTIONS

Investigations: DB. Formal analysis: DB, CB, IF. Writing- Original draft preparation: DB, CB. Methodology: CB, VA, KH, KR. Writing – Review & Editing: IF, VA, KH, KR. Conceptualization: VA, KR. Supervision: VA, KH, KR. Resource: KH. Funding acquisition: KR.

REFERENCES

[1] D. Chieffi, F. Fanelli, V. Fusco, *Arcobacter butzleri* : Up-to-date taxonomy, ecology, and pathogenicity of an emerging pathogen, Compr. Rev. Food Sci. Food Saf. (2020) 1541-4337.12577. https://doi.org/10.1111/1541-4337.12577.

[2] A. Oren, G.M. Garrity, List of new names and new combinations previously effectively, but not validly, published, Int. J. Syst. Evol. Microbiol. 70 (2020) 1–5.

https://doi.org/10.1099/ijsem.0.003881.

[3] A.-M. Van den Abeele, D. Vogelaers, J. Van Hende, K. Houf, Prevalence of *Arcobacter* Species among Humans, Belgium, 2008–2013, Emerg. Infect. Dis. 20 (2014) 1746–1749. https://doi.org/10.3201/eid2010.140433.

[4] G. Gölz, T. Alter, S. Bereswill, M.M. Heimesaat, The immunopathogenic potential of *Arcobacter butzleri* - Lessons from a meta-analysis of murine infection studies, PLoS One. 11 (2016) 1–18. https://doi.org/10.1371/journal.pone.0159685.

[5] S. De Smet, L. De Zutter, K. Houf, Spatial Distribution of the Emerging Foodborne Pathogen *Arcobacter* in the Gastrointestinal Tract of Pigs, Foodborne Pathog. Dis. 9 (2012) 1097–1103. https://doi.org/10.1089/fpd.2012.1184.

[6] N. Shange, P. Gouws, L.C. Hoffman, *Campylobacter* and *Arcobacter* species in food-producing animals: prevalence at primary production and during slaughter, World J. Microbiol. Biotechnol. 35 (2019) 146. https://doi.org/10.1007/s11274-019-2722-x.

[7] C.L. Hilton, B.M. Mackey, A.J. Hargreaves, S.J. Forsythe, The recovery of *Arcobacter butzleri* NCTC 12481 from various temperature treatments, J. Appl. Microbiol. 91 (2001) 929–932. https://doi.org/10.1046/j.1365-2672.2001.01457.x.

[8] T.P. Ramees, K. Dhama, K. Karthik, R.S. Rathore, A. Kumar, M. Saminathan, R. Tiwari, Y.S. Malik, R.K. Singh, *Arcobacter* : an emerging food-borne zoonotic pathogen, its public health concerns and advances in diagnosis and control – a comprehensive review, Vet. Q. 37 (2017) 136–161. https://doi.org/10.1080/01652176.2017.1323355.

[9] S.C. Pearce, H.G. Coia, J.P. Karl, I.G. Pantoja-feliciano, N.C. Zachos, K. Racicot, Intestinal in vitro and ex vivo Models to Study Host-Microbiome Interactions and Acute Stressors, Front. Physiol. 9 (2018). https://doi.org/10.3389/fphys.2018.01584.

[10] G.P. Donaldson, S.M. Lee, S.K. Mazmanian, Gut biogeography of the bacterial microbiota, Nat. Rev. Microbiol. 14 (2016) 20–32. https://doi.org/10.1038/nrmicro3552.

[11] A.P. Moran, A. Gupta, L. Joshi, Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract, Gut. 60 (2011) 1412–1425. https://doi.org/10.1136/gut.2010.212704.

[12] G.C. Hansson, Role of mucus layers in gut infection and inflammation, Curr. Opin. Microbiol. 15 (2012) 57–62. https://doi.org/10.1016/j.mib.2011.11.002.

[13] J.-F. Sicard, G. Le Bihan, P. Vogeleer, M. Jacques, J. Harel, Interactions of Intestinal Bacteria with Components of the Intestinal Mucus, Front. Cell. Infect. Microbiol. 7 (2017). https://doi.org/10.3389/fcimb.2017.00387.

[14] F. Fanelli, A. Di Pinto, A. Mottola, G. Mule, D. Chieffi, F. Baruzzi, G. Tantillo, V. Fusco, Genomic Characterization of *Arcobacter butzleri* Isolated From Shellfish: Novel Insight Into Antibiotic Resistance and Virulence Determinants, Front. Microbiol. 10 (2019) 1–17. https://doi.org/10.3389/fmicb.2019.00670.

[15] J. Isidro, S. Ferreira, M. Pinto, F. Domingues, M. Oleastro, J.P. Gomes, V. Borges, Virulence and antibiotic resistance plasticity of *Arcobacter butzleri*: Insights on the genomic diversity of an emerging human pathogen, Infect. Genet. Evol. 80 (2020) 104213. https://doi.org/10.1016/j.meegid.2020.104213.

[16] R. Bücker, H. Troeger, J. Kleer, M. Fromm, J.-D. Schulzke, *Arcobacter butzleri* Induces Barrier Dysfunction in Intestinal HT-29/B6 Cells, J. Infect. Dis. 200 (2009) 756–764. https://doi.org/10.1086/600868.

[17] M.A. Bianchi, D. Del Rio, N. Pellegrini, G. Sansebastiano, E. Neviani, F. Brighenti, A fluorescence-based method for the detection of adhesive properties of lactic acid bacteria to Caco-2 cells, Lett. Appl. Microbiol. 39 (2004) 301–305. https://doi.org/10.1111/j.1472-

765X.2004.01589.x.

[18] J.M. Laparra, Y. Sanz, Comparison of in vitro models to study bacterial adhesion to the intestinal epithelium, Lett. Appl. Microbiol. 49 (2009) 695–701. https://doi.org/10.1111/j.1472-765X.2009.02729.x.

[19] G. Karadas, R. Bücker, S. Sharbati, J.-D. Schulzke, T. Alter, G. Gölz, *Arcobacter butzleri* isolates exhibit pathogenic potential in intestinal epithelial cell models, J. Appl. Microbiol. 120 (2016) 218–225. https://doi.org/10.1111/jam.12979.

[20] G. Karadas, S. Sharbati, I. Hänel, U. Messelhäußer, E. Glocker, T. Alter, G. Gölz, Presence of virulence genes, adhesion and invasion of *Arcobacter butzleri*, J. Appl. Microbiol. 115 (2013) 583–590. https://doi.org/10.1111/jam.12245.

[21] A. Klančnik, I. Gobin, B. Jeršek, S. Smole Možina, D. Vučković, M. Tušek Žnidarič, M. Abram, Adhesion of *Campylobacter jejuni* Is Increased in Association with Foodborne Bacteria, Microorganisms. 8 (2020) 201. https://doi.org/10.3390/microorganisms8020201.

[22] P.H. Everst, H. Goossens, J.P. Butzler, D. Lloyd, S. Knutton, J.M. Ketley, P.H. Williams, Differentiated Caco-2 cells as a model for enteric invasion by *Campylobacter jejuni* and *C. coli*, J. Med. Microbiol. 37 (1992) 319–325. https://doi.org/10.1099/00222615-37-5-319.

[23] A. Alemka, N. Corcionivoschi, B. Bourke, Defense and Adaptation: The Complex Inter-Relationship between *Campylobacter jejuni* and Mucus, Front. Cell. Infect. Microbiol. 2 (2012) 1–6. https://doi.org/10.3389/fcimb.2012.00015.

[24] L. Douidah, L. De Zutter, F. Van Nieuwerburgh, D. Deforce, H. Ingmer, O. Vandenberg, A.M. Van Den Abeele, K. Houf, Presence and analysis of plasmids in human and animal associated *Arcobacter* species, PLoS One. 9 (2014) 1–8. https://doi.org/10.1371/journal.pone.0085487.

603 [25] F. Rovetto, A. Carlier, A.M. Van Den Abeele, K. Illeghems, F. Van Nieuwerburgh, L.
604 Cocolin, K. Houf, Characterization of the emerging zoonotic pathogen *Arcobacter thereius*
605 by whole genome sequencing and comparative genomics, PLoS One. 12 (2017) 1–27.
606 https://doi.org/10.1371/journal. pone.0180493.

607 [26] C. Botta, A. Acquadro, A. Greppi, L. Barchi, M. Bertolino, K. Rantsiou, Genomic
608 assessment in *Lactobacillus plantarum* links the butyrogenic pathway with glutamine
609 metabolism, Sci. Rep. (2017) 1–13. https://doi.org/10.1038/s41598-017-16186-8.

610 [27] M.T. Cabeen, C. Jacobs-wagner, The Bacterial Cytoskeleton, Annu. Rev. Of Genetics.
611 (2010). https://doi.org/10.1146/annurev-genet-102108-134845.

612 [28] M. Furter, M.E. Sellin, G.C. Hansson, W.-D. Hardt, Mucus Architecture and Near-Surface
613 Swimming Affect Distinct *Salmonella* Typhimurium Infection Patterns along the Murine
614 Intestinal Tract, Cell Rep. 27 (2019) 2665-2678.e3.
615 https://doi.org/10.1016/j.celrep.2019.04.106.

616 [29] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T.G.G. Holden, M. Fookes, D.
617 Falush, J.A. Keane, J. Parkhill, Roary: Rapid large-scale prokaryote pan genome analysis,
618 Bioinformatics. 31 (2015) 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

619 [30] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue,
620 A. Calteau, S. Cruveiller, C. Matias, C. Ambroise, E.P.C. Rocha, D. Vallenet,
621 PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph, PLoS
622 Comput. Biol. 16 (2020) 1–27. https://doi.org/10.1371/journal.pcbi.1007732.

623 [31] A.A. Golicz, P.E. Bayer, P.L. Bhalla, J. Batley, D. Edwards, Pangenomics Comes of Age:
624 From Bacteria to Plant and Animal Applications, Trends Genet. 36 (2020) 132–145.
625 https://doi.org/10.1016/j.tig.2019.11.006.

626 [32] T. Gouliouris, K.E. Raven, C. Ludden, B. Blane, J. Corander, C.S. Horner, J. Hernandez-

Garcia, P. Wood, N.F. Hadjirin, M. Radakovic, M.A. Holmes, M. de Goffau, N.M. Brown, J. Parkhill, S.J. Peacock, Genomic surveillance of enterococcus faecium reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom, MBio. 9 (2018) 1–15. https://doi.org/10.1128/mBio.01780-18.

[33] M.R. Davies, L. McIntyre, A. Mutreja, J.A. Lacey, J.A. Lees, R.J. Towers, S. Duchêne, P.R. Smeesters, H.R. Frost, D.J. Price, M.T.G. Holden, S. David, P.M. Giffard, K.A. Worthing, A.C. Seale, J.A. Berkley, S.R. Harris, T. Rivera-Hernandez, O. Berking, A.J. Cork, R.S.L.A. Torres, T. Lithgow, R.A. Strugnell, R. Bergmann, P. Nitsche-Schmitz, G.S. Chhatwal, S.D. Bentley, J.D. Fraser, N.J. Moreland, J.R. Carapetis, A.C. Steer, J. Parkhill, A. Saul, D.A. Williamson, B.J. Currie, S.Y.C. Tong, G. Dougan, M.J. Walker, Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics, Nat. Genet. 51 (2019) 1035–1043. https://doi.org/10.1038/s41588-019-0417-8.

[34] M. Juhas, J.R. Van Der Meer, M. Gaillard, R.M. Harding, D.W. Hood, D.W. Crook, Genomic islands: Tools of bacterial horizontal gene transfer and evolution, FEMS Microbiol. Rev. 33 (2009) 376–393. https://doi.org/10.1111/j.1574-6976.2008.00136.x.

[35] A. Bazin, G. Gautreau, C. Médigue, D. Vallenet, A. Calteau, panRGP: a pangenome-based method to predict genomic islands and explore their diversity, BioRxiv. 1 (2020) 2020.03.26.007484. https://doi.org/10.1101/2020.03.26.007484.

[36] U. Sood, P. Hira, R. Kumar, A. Bajaj, D.L.N. Rao, R. Lal, M. Shakarad, Comparative genomic analyses reveal core-genome-wide genes under positive selection and major regulatory hubs in outlier strains of *Pseudomonas aeruginosa*, Front. Microbiol. 10 (2019). https://doi.org/10.3389/fmicb.2019.00053.

[37] U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, Genomic islands in pathogenic and environmental microorganisms, Nat. Rev. Microbiol. 2 (2004) 414–424.

https://doi.org/10.1038/nrmicro884.

[38] E. Mourkasa, A.J. Taylorb, G. Mérica, S.C. Baylissa, B. Pascoea, L. Mageirosa, J.K. Callanda, M.D. Hitchingse, A. Ridleyf, A. Vidalf, K.J. Forbesg, N.J.C.S. Strachanh, C.T. Parkeri, J. Parkhillj, A.K. Jolley, A.J. Codyk, M.C.J. Maiden, D.J. Kelly, S.K. Sheppard, Agricultural intensification and the evolution of host specialism in the enteric pathogen *Campylobacter jejuni*, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 11018–11028. https://doi.org/10.1073/pnas.1917168117.

[39] B. Liu, Y.A. Knirel, L. Feng, A. V. Perepelov, S.N. Senchenkova, P.R. Reeves, L. Wang, Structural diversity in *Salmonella* O antigens and its genetic basis, FEMS Microbiol. Rev. 38 (2014) 56–89. https://doi.org/10.1111/1574-6976.12034.

[40] I. V. Wesley, A.L. Baetz, D.J. Larson, Infection of cesarean-derived colostrum-deprived 1-day-old piglets with *Arcobacter butzleri*, *Arcobacter cryaerophilus*, and *Arcobacter skirrowii*, Infect. Immun. 64 (1996) 2295–2299.

[41] D.J. Woodcock, P. Krusche, N.J.C. Strachan, K.J. Forbes, F.M. Cohan, G. Méric, S.K. Sheppard, Genomic plasticity and rapid host switching can promote the evolution of generalism: A case study in the zoonotic pathogen *Campylobacter*, Sci. Rep. 7 (2017) 1–13. https://doi.org/10.1038/s41598-017-09483-9.

[42] H.H. De Fine Licht, Does pathogen plasticity facilitate host shifts?, PLoS Pathog. 14 (2018) 1–9. https://doi.org/10.1371/journal.ppat.1006961.

[43] L. Douidah, L. de Zutter, J. Bare, P. De Vos, P. Vandamme, O. Vandenberg, A.-M. Van den Abeele, K. Houf, Occurrence of Putative Virulence Genes in *Arcobacter* Species Isolated from Humans and Animals, J. Clin. Microbiol. 50 (2012) 735–741. https://doi.org/10.1128/JCM.05872-11.

[44] A. Parisi, L. Capozzi, A. Bianco, M. Caruso, L. Latorre, A. Costa, A. Giannico, D. Ridolfi,

675 C. Bulzacchelli, G. Santagada, Identification of virulence and antibiotic resistance factors in
676 *Arcobacter butzleri* isolated from bovine milk by whole genome sequencing, Ital. J. Food
677 Saf. 8 (2019). https://doi.org/10.4081/ijfs.2019.7840.

678 [45] C. Girbau, C. Guerra, I. Martínez-Malaxetxebarria, R. Alonso, A. Fernández-Astorga,
679 Prevalence of ten putative virulence genes in the emerging foodborne pathogen *Arcobacter*
680 isolated from food products, Food Microbiol. 52 (2015) 146–149.
681 https://doi.org/10.1016/j.fm.2015.07.015.

682 [46] S. Ferreira, J.A. Queiroz, M. Oleastro, F.C. Domingues, Genotypic and phenotypic features
683 of *Arcobacter butzleri* pathogenicity, Microb. Pathog. 76 (2014) 19–25.
684 https://doi.org/10.1016/j.micpath.2014.09.004.

685 [47] Z. Wu, B. Periaswamy, O. Sahin, M. Yaeger, P. Plummer, W. Zhai, Z. Shen, L. Dai, S.L.
686 Chen, Q. Zhang, Point mutations in the major outer membrane protein drive hypervirulence
687 of a rapidly expanding clone of *Campylobacter jejuni*, Proc. Natl. Acad. Sci. 113 (2016)
688 10690–10695. https://doi.org/10.1073/pnas.1605869113.

689 [48] C. Sabet, A. Toledo-Arana, N. Personnic, M. Lecuit, S. Dubrac, O. Poupel, E. Gouin, M.-
690 A.M. Nahori, P. Cossart, H.H.H. Bierne, The *Listeria monocytogenes* Virulence Factor InlJ
691 Is Specifically Expressed In Vivo and Behaves as an Adhesin, Infect. Immun. 76 (2008)
692 1368–1378. https://doi.org/10.1128/IAI.01519-07.

693 [49] M.W. Gilmour, M. Graham, G. Van Domselaar, S. Tyler, H. Kent, K.M. Trout-Yakel, O.
694 Larios, V. Allen, B. Lee, C. Nadon, High-throughput genome sequencing of two *Listeria*
695 *monocytogenes* clinical isolates during a large foodborne outbreak, BMC Genomics. 11
696 (2010) 120. https://doi.org/10.1186/1471-2164-11-120.

697 [50] A.P. Corfield, S.A. Wagner, J.R. Clamp, M.S. Kriaris, L.C. Hoskins, Mucin degradation in
698 the human colon: production of sialidase, sialate O-acetylesterase, N-acetylneuraminate

lyase, arylesterase, and glycosulfatase activities by strains of fecal bacteria., Infect. Immun. 60 (1992) 3971–3978. https://doi.org/10.1128/iai.60.10.3971-3978.1992.

[51] M. Sandkvist, Type II Secretion and Pathogenesis, Infect. Immun. 69 (2001) 3523–3535. https://doi.org/10.1128/IAI.69.6.3523.

[52] T. Yoshida, Y. Ayabe, M. Yasunaga, Y. Usami, H. Habe, H. Nojiri, T. Omori, Genes involved in the synthesis of the exopolysaccharide methanolan by the obligate methylotroph *Methylobacillus* sp. strain 12S, Microbiology. 149 (2003) 431–444. https://doi.org/10.1099/mic.0.25913-0.

[53] S. Lebeer, T.L.A. Verhoeven, G. Francius, G. Schoofs, I. Lambrichts, Y. Dufrêne, J. Vanderleyden, S.C.J. De Keersmaecker, Identification of a gene cluster for the biosynthesis of a long, galactose-rich exopolysaccharide in *Lactobacillus rhamnosus* GG and functional analysis of the priming glycosyltransferase, Appl. Environ. Microbiol. 75 (2009) 3554–3563. https://doi.org/10.1128/AEM.02919-08.

[54] S. Ferreira, M.J. Fraqueza, J.A. Queiroz, F.C. Domingues, M. Oleastro, Genetic diversity, antibiotic resistance and biofilm-forming ability of *Arcobacter butzleri* isolated from poultry and environment from a Portuguese slaughterhouse, Int. J. Food Microbiol. 162 (2013) 82–88. https://doi.org/10.1016/j.ijfoodmicro.2013.01.003.

[55] G.G. Sgro, G.U. Oka, D.P. Souza, W. Cenens, E. Bayer-Santos, B.Y. Matsuyama, N.F. Bueno, T.R. dos Santos, C.E. Alvarez-Martinez, R.K. Salinas, C.S. Farah, Bacteria-Killing Type IV Secretion Systems, Front. Microbiol. 10 (2019) 1–20. https://doi.org/10.3389/fmicb.2019.01078.

[56] R.A. Batchelor, Nucleotide sequences and comparison of two large conjugative plasmids from different *Campylobacter* species, Microbiology. 150 (2004) 3507–3517. https://doi.org/10.1099/mic.0.27112-0.

[57] C.E. Alvarez-Martinez, P.J. Christie, Biological Diversity of Prokaryotic Type IV Secretion Systems, Microbiol. Mol. Biol. Rev. 73 (2009) 775–808. https://doi.org/10.1128/mmbr.00023-09.

[58] S. Merino, J.M. Tomás, Gram-negative flagella glycosylation, Int. J. Mol. Sci. 15 (2014) 2840–2857. https://doi.org/10.3390/ijms15022840.

[59] J.M.D. Roberts, L.L. Graham, B. Quinn, D.A. Pink, Modeling the surface of *Campylobacter fetus*: Protein surface layer stability and resistance to cationic antimicrobial peptides, Biochim. Biophys. Acta - Biomembr. 1828 (2013) 1143–1152. https://doi.org/10.1016/j.bbamem.2012.10.025.

[60] S. Kalynych, R. Morona, M. Cygler, Progress in understanding the assembly process of bacterial O-antigen, FEMS Microbiol. Rev. 38 (2014) 1048–1065. https://doi.org/10.1111/1574-6976.12070.

[61] R.F. Maldonado, I. Sá-Correia, M.A. Valvano, Lipopolysaccharide modification in Gram-negative bacteria during chronic infection, FEMS Microbiol. Rev. 40 (2016) 480–493. https://doi.org/10.1093/femsre/fuw007.

[62] F. Fanelli, D. Chieffi, A. Di Pinto, A. Mottola, F. Baruzzi, V. Fusco, Phenotype and genomic background of *Arcobacter butzleri* strains and taxogenomic assessment of the species, Food Microbiol. 89 (2020) 103416. https://doi.org/10.1016/j.fm.2020.103416.

[63] S.A. Reeves, A.G. Torres, S.M. Payne, TonB Is Required for Intracellular Growth and Virulence of *Shigella dysenteriae*, Infect. Immun. 68 (2000) 6329–6336. https://doi.org/10.1128/IAI.68.11.6329-6336.2000.

[64] A.G. Torres, P. Redford, R.A. Welch, S.M. Payne, TonB-Dependent Systems of Uropathogenic *Escherichia coli*: Aerobactin and Heme Transport and TonB Are Required for Virulence in the Mouse, Infect. Immun. 69 (2001) 6179–6185.

https://doi.org/10.1128/IAI.69.10.6179-6185.2001.

[65] W. Simpson, T. Olczak, C.A. Genco, Characterization and Expression of HmuR, a TonB-Dependent Hemoglobin Receptor of *Porphyromonas gingivalis*, J. Bacteriol. 182 (2000) 5737–5748. https://doi.org/10.1128/JB.182.20.5737-5748.2000.

[66] D.J. Morton, R.J. Hempel, T.W. Seale, P.W. Whitby, T.L. Stull, A functional *tonB* gene is required for both virulence and competitive fitness in a chinchilla model of *Haemophilus influenzae* otitis media, BMC Res. Notes. 5 (2012) 327. https://doi.org/10.1186/1756-0500-5-327.

[67] X. Zhuge, Y. Sun, F. Xue, F. Tang, J. Ren, D. Li, J. Wang, M. Jiang, J. Dai, A Novel PhoP/PhoQ Regulation Pathway Modulates the Survival of Extraintestinal Pathogenic Escherichia coli in Macrophages, Front. Immunol. 9 (2018) 1–23. https://doi.org/10.3389/fimmu.2018.00788.

[68] I. Zwir, D. Shin, A. Kato, K. Nishino, T. Latifi, F. Solomon, J.M. Hare, H. Huang, E.A. Groisman, Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 2862–2867. https://doi.org/10.1073/pnas.0408238102.

[69] H. He, T.C. Zahrt, Identification and Characterization of a Regulatory Sequence Recognized by *Mycobacterium tuberculosis* Persistence Regulator MprA, J. Bacteriol. 187 (2004) 202–212. https://doi.org/10.1128/JB.187.1.202.

[70] Y. Eguchi, T. Okada, S. Minagawa, T. Oshima, H. Mori, K. Yamamoto, A. Ishihama, R. Utsumi, Signal Transduction Cascade between EvgA / EvgS and PhoP / PhoQ Two-Component Systems of *Escherichia coli*, J. Bacteriol. 186 (2004) 3006–3014. https://doi.org/10.1128/JB.186.10.3006.

[71] A.M. Lippa, M. Goulian, Perturbation of the Oxidizing Environment of the Periplasm

771 Stimulates the PhoQ/PhoP System in *Escherichia coli*, J. Bacteriol. 194 (2012) 1457–1463.
772 https://doi.org/10.1128/JB.06055-11.

773 [72] L.N. Schulte, M. Schweinlin, A.J. Westermann, H. Janga, S.C. Santos, S. Appenzeller, H.
774 Walles, J. Vogel, M. Metzger, An Advanced Human Intestinal Coculture Model Reveals
775 Compartmentalized Host and Pathogen Strategies during *Salmonella* Infection, MBio. 11
776 (2020). https://doi.org/10.1128/mBio.03348-19.

777 [73] K. Houf, R. Stephan, Isolation and characterization of the emerging foodborn pathogen
778 *Arcobacter* from human stool, J. Microbiol. Methods. 68 (2007) 408–413.
779 https://doi.org/10.1016/j.mimet.2006.09.020.

780 [74] C. Botta, T. Langerholc, A. Cencič, L. Cocolin, *In vitro* selection and characterization of new
781 probiotic candidates from table olive microbiota, PLoS One. 9 (2014).
782 https://doi.org/10.1371/journal.pone.0094457.

783 [75] X.M. Chen, I. Elisia, D.D. Kitts, Defining conditions for the co-culture of Caco-2 and HT29-
784 MTX cells using Taguchi design, J. Pharmacol. Toxicol. Methods. 61 (2010) 334–342.
785 https://doi.org/10.1016/j.vascn.2010.02.004.

786 [76] M. Pinto, S. Robine-Leon, M.-D. Appay, M. Kedinger, N. Triadou, E. Dussaulx, B. Lacroix,
787 P. Simon-Assmann, K. Haffen, J. Fogh, A. Zweibaum, Enterocyte-like differentiation and
788 polarization of the human colon carcinoma cell line Caco-2 in culture, Biol. Cell. 47 (1983)
789 323–330.

790 [77] S. Backert, D. Hofreuter, Molecular methods to investigate adhesion, transmigration,
791 invasion and intracellular survival of the foodborne pathogen *Ccampylobacter jejuni*, J.
792 Microbiol. Methods. 95 (2013) 8–23. https://doi.org/10.1016/j.mimet.2013.06.031.

793 [78] A. Levican, A. Alkeskas, C. Gunter, S.J. Forsythe, M.J. Figueras, Adherence to and Invasion
794 of Human Intestinal Cells by *Arcobacter* Species and Their Virulence Genotypes, Appl.

795 Environ. Microbiol. 79 (2013) 4951–4957. https://doi.org/10.1128/AEM.01073-13.

796 [79] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin,
797 S.I. Nikolenko, S. Pham, A.D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G.
798 Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: A new genome assembly algorithm and its
799 applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.
800 https://doi.org/10.1089/cmb.2012.0021.

801 [80] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome
802 assemblies, Bioinformatics. 29 (2013) 1072–1075.
803 https://doi.org/10.1093/bioinformatics/btt086.

804 [81] T. Seemann, Prokka: Rapid prokaryotic genome annotation, Bioinformatics. 30 (2014) 2068–
805 2069. https://doi.org/10.1093/bioinformatics/btu153.

806 [82] A. Marchler-Bauer, M.K. Derbyshire, N.R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, R.C.
807 Geer, J. He, M. Gwadz, D.I. Hurwitz, C.J. Lanczycki, F. Lu, G.H. Marchler, J.S. Song, N.
808 Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, S.H. Bryant, CDD: NCBI's
809 conserved domain database, Nucleic Acids Res. 43 (2015) D222–D226.
810 https://doi.org/10.1093/nar/gku1221.

811 [83] I. Grissa, G. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly
812 interspaced short palindromic repeats, Nucleic Acids Res. 35 (2007) W52–W57.
813 https://doi.org/10.1093/nar/gkm360.

814 [84] D. Arndt, J.R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, D.S. Wishart, PHASTER: a
815 better, faster version of the PHAST phage search tool, Nucleic Acids Res. 44 (2016) W16–
816 W21. https://doi.org/10.1093/nar/gkw387.

817 [85] A. Marchler-Bauer, CDD: a Conserved Domain Database for protein classification, Nucleic
818 Acids Res. 33 (2004) D192–D196. https://doi.org/10.1093/nar/gki069.

[86] M. Yang, M.K. Derbyshire, R.A. Yamashita, A. Marchler-Bauer, NCBI's Conserved Domain Database and Tools for Protein Domain Analysis, Curr. Protoc. Bioinforma. 69 (2020) 1–25. https://doi.org/10.1002/cpbi.90.

[87] S. Wu, Z. Zhu, L. Fu, B. Niu, W. Li, WebMGA: a customizable web server for fast metagenomic sequence analysis, BMC Genomics. 12 (2011) 444. https://doi.org/10.1186/1471-2164-12-444.

[88] O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary, Genome Biol. 17 (2016) 1–9. https://doi.org/10.1186/s13059-016-1108-8.

[89] S. McGinnis, T.L. Madden, BLAST: At the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res. 32 (2004) 20–25. https://doi.org/10.1093/nar/gkh435.

[90] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics. 23 (2007) 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

[91] M. V Larsen, S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R.L. Marvig, L. Jelsbak, T. Sicheritz-Ponten, D.W. Ussery, F.M. Aarestrup, O. Lund, Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria, J. Clin. Microbiol. 50 (2012) 1355–1361. https://doi.org/10.1128/JCM.06094-11.

[92] W.G. Miller, I. V. Wesley, S.L.W. On, K. Houf, F. Mégraud, G. Wang, E. Yee, A. Srijan, C.J. Mason, First multi-locus sequence typing scheme for *Arcobacter* spp., BMC Microbiol. 9 (2009) 196. https://doi.org/10.1186/1471-2180-9-196.

[93] M.N. Price, P.S. Dehal, A.P. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments, PLoS One. 5 (2010) e9490.

https://doi.org/10.1371/journal.pone.0009490.

[94] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, Nucleic Acids Res. 47 (2019) W256–W259. https://doi.org/10.1093/nar/gkz239.

[95] M.C. Ryan, M. Stucky, C. Wakefield, J.M. Melott, R. Akbani, J.N. Weinstein, B.M. Broom, Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps, F1000Research. 8 (2020) 1750. https://doi.org/10.12688/f1000research.20590.2.

852 **Figures captions**

853

854 **Figure 1.** Colonization and invasion capabilities on mucus producer (MP) and not-mucus producer
855 (NMP) models are expressed as Δ Log CFU/cm^2 (medians ± interquartile range; n=3; dots=outliers)
856 and shown for all 32 strains together **(A)** and individually for each strain **(B)**. The red dotted line
857 marks the ΔLog equal to 0: a condition in which all bacterial cells added colonized/invaded the model.
858 Positive values indicate the potential growth of added bacteria in the model during the co-incubation,
859 while negative values indicate progressively lower colonization/invasion capability. Coding keys of
860 box-plots color are displayed in the caption. Significant differences between models and among the
861 strains are reported in the graph (*P*-value) employing Wilcoxon's test.

862

863 **Figure 2.** Bar-plots displaying the average (± standard deviation) distribution of COG classes in all
864 32 annotated genomes (% of putative proteins assigned to a class compared to the total putative
865 proteins). Coding keys of classes colors are shown in the caption.

866

867

868 **Figure 3.** Partitioned pangenome network **(A)** displaying the genomic diversity of the 32 strains.
869 Nodes represent the gene families and are colored according to the partition (caption), while their size
870 is proportional to the number of genomes in which are present. Edges connect gene families
871 colocalized in the pangenome and their thickness is proportional to the number of genomes sharing
872 that link. Edges are colored as described for nodes, except for edges between partitions (mixed
873 colors). The frame highlights a broad plasticity region of the pangenome (zoomed-in **B**) harboring
874 shell/cloud gene families alternatively present in the 32 genomes (pangenome plasticity region;
875 **Supplementary Table 2)**. Input files (nodes.csv and edges.csv) set up for network visualization in
876 Gephi (https://gephi.org) are provided on Zenodo (http://doi.org/10.5281/zenodo.4301795). Bar-

plots (**C**) showing the functional partitioning of gene families in the pangenome plasticity region (figure B) and all regions of genomic plasticity (RGPs) along the 32 genomes. Asterisks (*) highlight groups of gene families of which function is manually assigned (**Supplementary Table 2**).

Figure 4. Phylogenetic trees of whole genomes (A), core genomes (B), MLST sequences (C) and SNPs (D) of the 32 *A. butzleri* strains. The original source of isolation is indicated and groups of strains that show recurrent clustering patterns are highlighted with colors and named by roman numbers: **I** (strains 14, 15); **II** (strains 1, 8); **III** (strains 12, 19, 20, 21) and **IV (**strains 11, 13, 23**).**

Figure 5. Heatmap representing the absence/presence matrix of putative virulence genes detected in the 32 genomes. Gene names or their annotated product are displayed for each gene considered. Asterix (*) highlight putative virulence genes which annotation was verified by alignment with reference strain LMG 10828^{T}; original annotation in brackets, while caret symbols (^) indicate the presence of non-unique alleles. The groups of strains are indicated from the panes and the group numbers: **I** (strains 14, 15 from pig), **III** (strains 12, 19, 20, 21 from pig) and **IV** (strains 11, 13, 23 from pig), whereas the group **II** (strain 1 and 28 from human) results absent.

Supplementary figures list

Supplementary Figure 1. Bar-plots (**A**) displaying the distribution of COG classes in each of the 32 annotated genomes (% of putative proteins assigned to a class compared to the putative proteins). Coding keys of colors and class codes are shown in the caption. Heatmap (**B**) showing the presence (grey) / absence (white) matrix of genes involved in the accessory pathway and COG classes. Pathways are linked to sequences identified through UniProt codes from the Prokka annotation.

Supplementary figure 2. Singletons distribution along the genomes and composition of the main clusters of singletons (> 10 loci) identified in the accessory genome.

Supplementary figure 3. UPGMA phylogenetic analysis of *porA*. The original source of isolation of the strains is indicated and roman numbers (I–IV) indicate the different groups of strains that coincide with the groups indicated in Figure 4.

Supplementary figure 4. UPGMA phylogenetic analysis of O-antigen ligase. The original source of isolation of the strains is indicated and the roman numbers (I–IV) indicate the different groups of strains. In the case of the O-antigen ligase dendrogram some strains are repeated, this aspect is linked to the presence of several gene copies. The O-antigen ligase sequence of *Klebsiella pneumoniae* ATCC 700721 (used from Prokka for the functional annotation) has been used as outgroup.

Tables captions

Table 1. *A. butzleri* strains object of study. In the table are shown the number of the strains (nr.), C-country of origin (Country), the source of sampling (Source), the specific sampling matrix (Isolation source) and additional information such as official strain codes and information related to the patients from whom the strain was isolated.

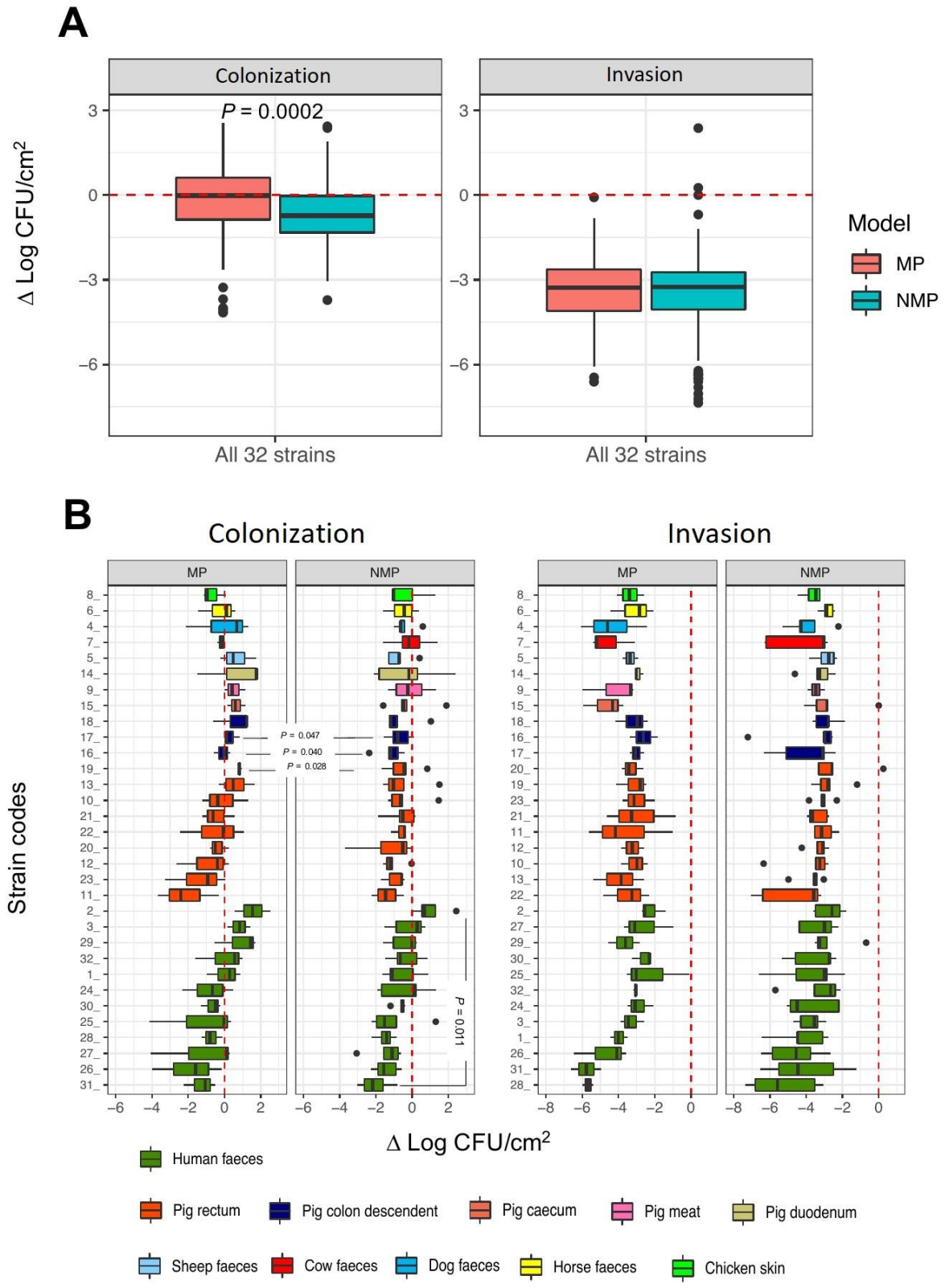
Table 2. Pangenome partitions estimated by two computational methods.

Supplementary tables list

Supplementary Table 1. Bacterial count about colonization/invasion test of the 32 *A. butzleri* strains. The single strain bacterial loads of the initial inoculum (T0), bacteria load detected after the cell layer washing (T1) and after the gentamicin application (T2) are expressed in logarithm$_{10}$ (log) and are relative to Mucus producing models (MP) and Not mucus producing models (NMP). Moreover, the

928 standard deviations (st. dv) and the T0, T1, T2 average of the 32 strains on MP and NMP models with
929 the relative standard deviations are indicated.
930
931
932 **Supplementary Table 2.** Annotation statistics of the 32 *A. butzleri* strains. In the first column are
933 indicated the code of the strains and their source of sampling. In the table are indicated the genome
934 size (Mbp), total genes, number of CDS, number of tRNA, hypothetical proteins, transposase,
935 prophage sequences and CRISPR sequences. The CAS sequences detected in the genomes belong to
936 the general class 1 and general class 2. Sequences putative for the production of protein appertain at
937 the bacteriocins bottromycin, microcin and sactipeptides class are indicated with the number of
938 sequences linked to their translation.
939
940 **Supplementary Table 3.** List of genes putatively involved in *A. butzleri* virulence. In the second
941 column is present the locus tag codes on the type strain LMG10828^T (strain 3), unless otherwise
942 reported in brackets. The protein codes are relative to UniProt code and Pfam databases. Part of the
943 genes (*) involved in antibiotics resistance and general chemotaxis are only reported here and not in
944 **Supplementary Table 4.** List of structures and genes involved in the LPS O-antigen biosynthesis.
945 The presence of O-antigen ligase (**) and genes putatively associated to LPS O-antigen cluster
946 assembling (*) are indicated with asterisks.
947 **Supplementary Table 5.** genes MLST codes of the strains object of study. In the last column is
948 indicated the nearest sequence type code- (nearest ST). Some gene sequences resulted in new alleles,
949 these genes are indicated with an asterisk.
950

951 **Figures**



952

953 **Fig. 1**

954

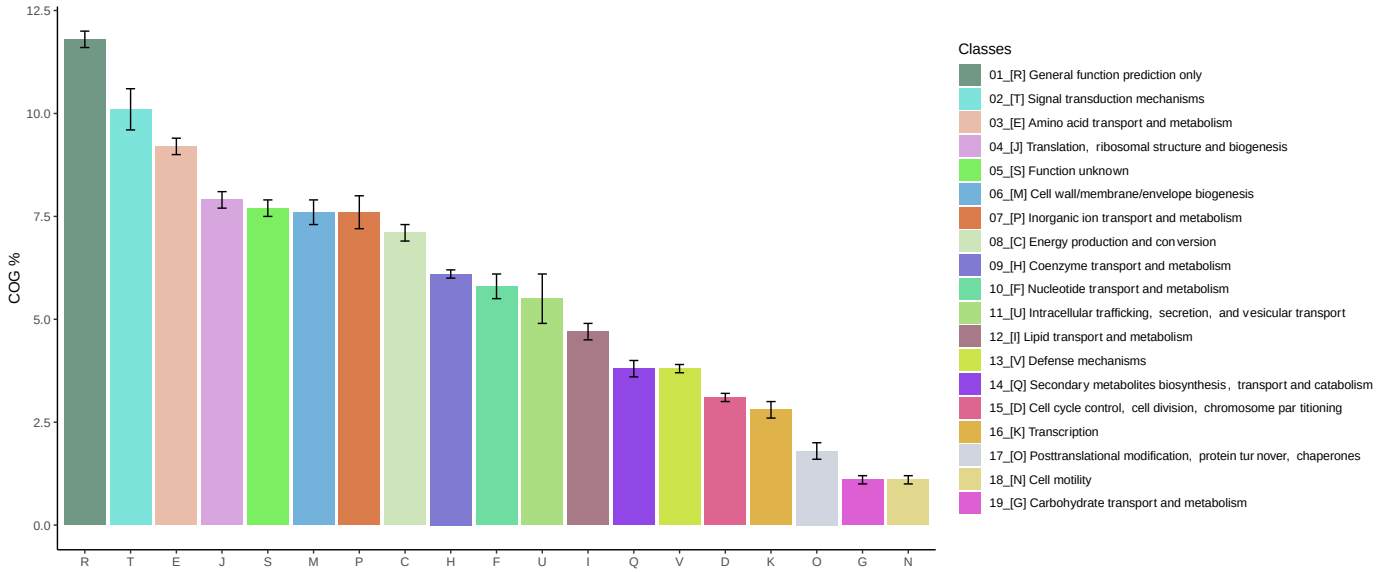

COG %
12.5
10.0
7.5
5.0
2.5
0.0
R
T
E
J
S
M
P
C
H
F
U
I
Q
V
D
K
O
G
N
Classes
01_[R] General function prediction only
02_[T] Signal transduction mechanisms
03_[E] Amino acid transport and metabolism
04_[J] Translation, ribosomal structure and biogenesis
05_[S] Function unknown
06_[M] Cell wall/membrane/envelope biogenesis
07_[P] Inorganic ion transport and metabolism
08_[C] Energy production and conversion
09_[H] Coenzyme transport and metabolism
10_[F] Nucleotide transport and metabolism
11_[U] Intracellular trafficking, secretion, and vesicular transport
12_[I] Lipid transport and metabolism
13_[V] Defense mechanisms
14_[Q] Secondary metabolites biosynthesis, transport and catabolism
15_[D] Cell cycle control, cell division, chromosome par titioning
16_[K] Transcription
17_[O] Posttranslational modification, protein tur nover, chaperones
18_[N] Cell motility
19_[G] Carbohydrate transport and metabolism

Fig. 2

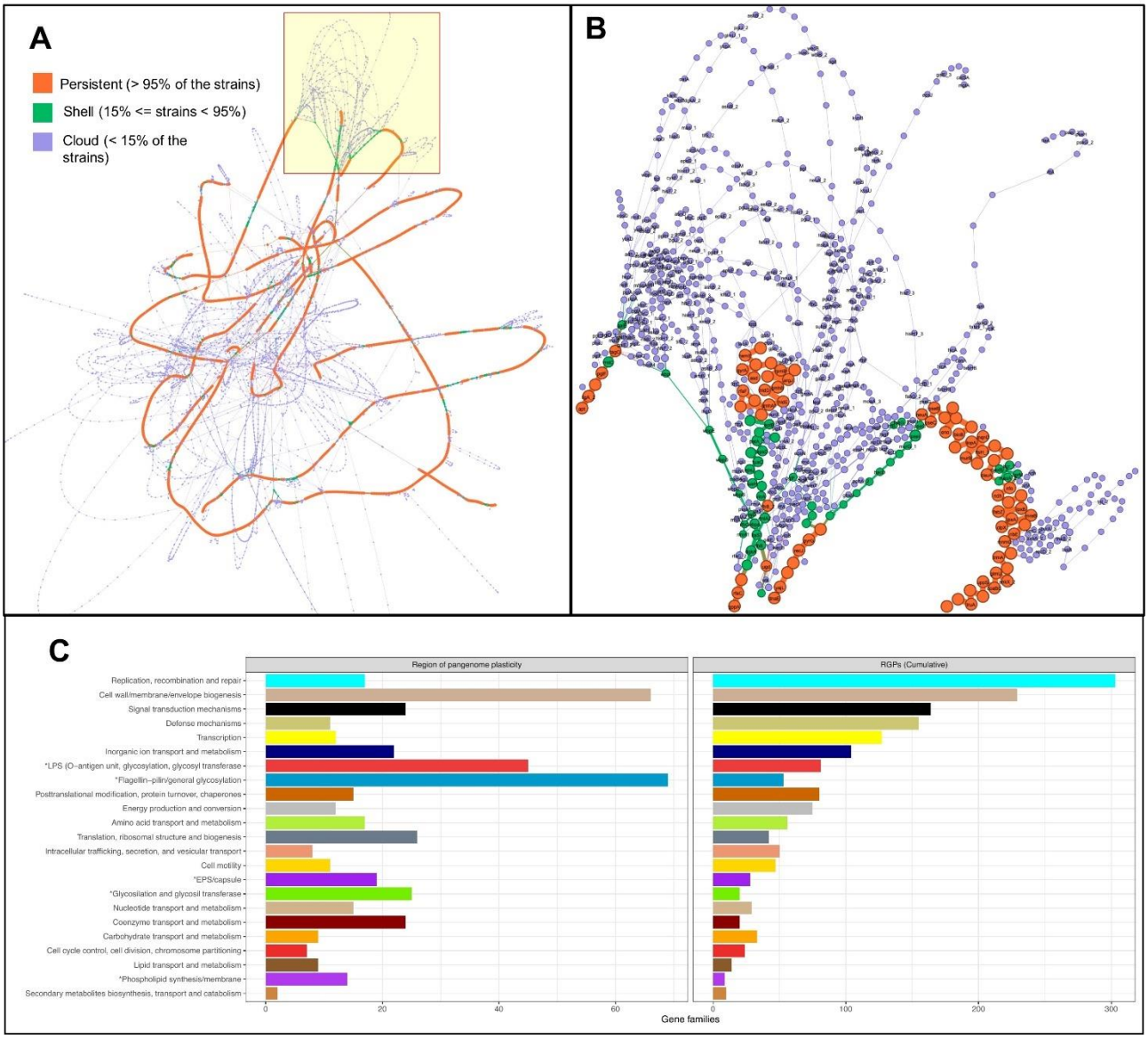

A
Persistent (> 95% of the strains)
Shell (15% <= strains < 95%)
Cloud (< 15% of the strains)
B
C
Region of pangenome plasticity
RGPs (Cumulative)
Replication, recombination and repair
Cell wall/membrane/envelope biogenesis
Signal transduction mechanisms
Defense mechanisms
Transcription
Inorganic ion transport and metabolism
*LPS (O-antigen unit, glycosylation, glycosyl transferase
*Flagellin-pilin/general glycosylation
Posttranslational modification, protein turnover, chaperones
Energy production and conversion
Amino acid transport and metabolism
Translation, ribosomal structure and biogenesis
Intracellular trafficking, secretion, and vesicular transport
Cell motility
*EPS/capsule
*Glycosilation and glycosil transferase
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Carbohydrate transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Lipid transport and metabolism
*Phospholipid synthesis/membrane
Secondary metabolites biosynthesis, transport and catabolism
0
20
40
60
0
100
200
300
Gene families

Fig. 3

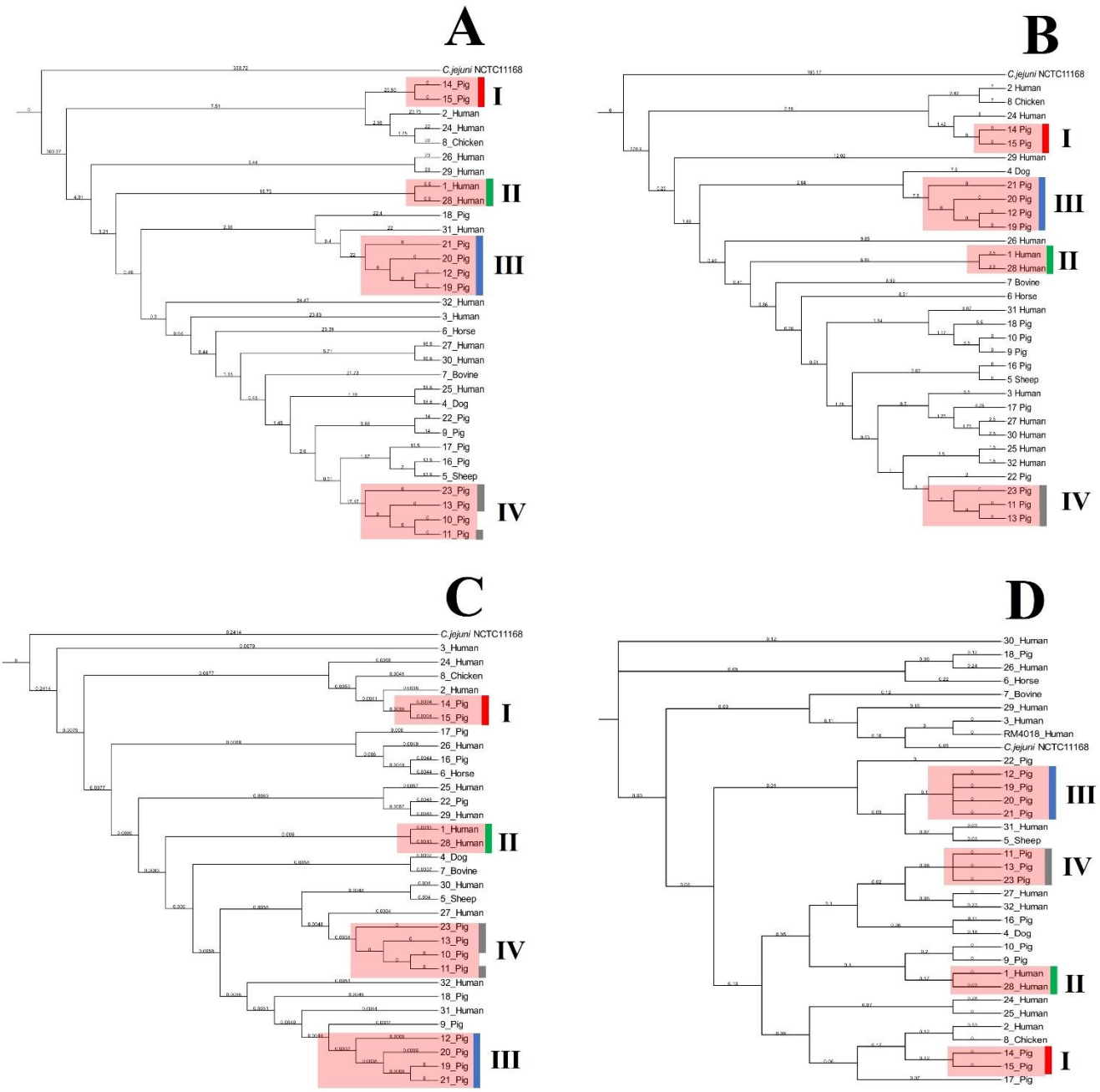

A
C.jejuni NCTC11168
14_Pig
15_Pig
I
2_Human
24_Human
8_Chicken
26_Human
29_Human
1_Human
28_Human
II
18_Pig
31_Human
21_Pig
20_Pig
12_Pig
19_Pig
III
32_Human
3_Human
6_Horse
27_Human
30_Human
7_Bovine
25_Human
4_Dog
22_Pig
9_Pig
17_Pig
16_Pig
5_Sheep
23_Pig
13_Pig
10_Pig
11_Pig
IV
B
C.jejuni NCTC11168
2 Human
8 Chicken
24 Human
14 Pig
15 Pig
I
29 Human
4 Dog
21 Pig
20 Pig
12 Pig
19 Pig
III
26 Human
1 Human
28 Human
II
7 Bovine
6 Horse
31 Human
18 Pig
10 Pig
9 Pig
16 Pig
5 Sheep
3 Human
17 Pig
27 Human
30 Human
25 Human
32 Human
22 Pig
23 Pig
11 Pig
13 Pig
IV
C
C.jejuni NCTC11168
3_Human
24_Human
8_Chicken
2_Human
14_Pig
15_Pig
I
17_Pig
26_Human
16_Pig
6_Horse
25_Human
22_Pig
29_Human
1_Human
28_Human
II
4_Dog
7_Bovine
30_Human
5_Sheep
27_Human
23_Pig
13_Pig
10_Pig
11_Pig
IV
32_Human
18_Pig
31_Human
9_Pig
12_Pig
20_Pig
19_Pig
21_Pig
III
D
30_Human
18_Pig
26_Human
6_Horse
7_Bovine
29_Human
3_Human
RM4018_Human
C.jejuni NCTC11168
22_Pig
12_Pig
19_Pig
20_Pig
21_Pig
III
31_Human
5_Sheep
11_Pig
13_Pig
23_Pig
IV
27_Human
32_Human
16_Pig
4_Dog
10_Pig
9_Pig
1_Human
28_Human
II
24_Human
25_Human
2_Human
8_Chicken
14_Pig
15_Pig
I
17_Pig

Fig. 4

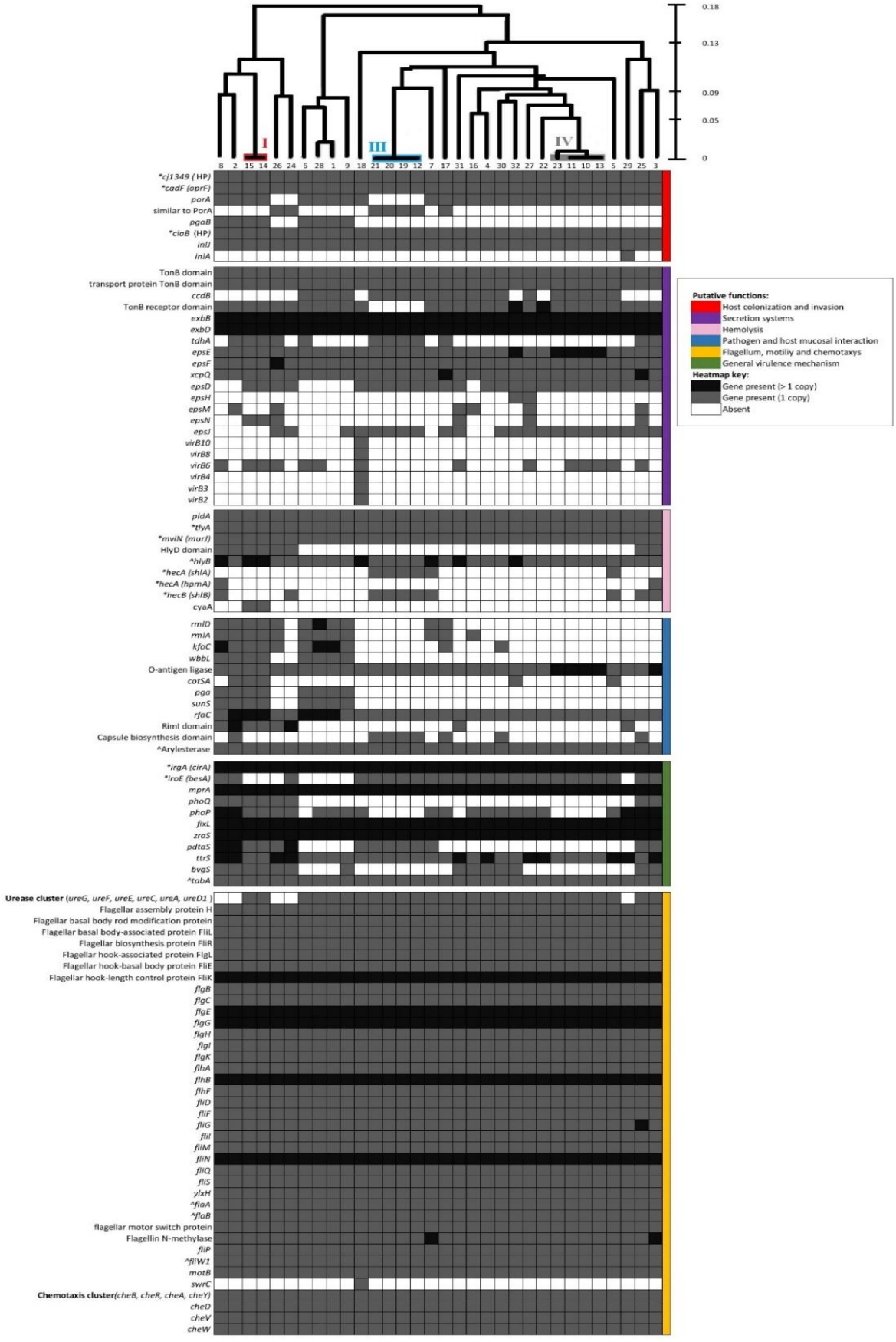

0.18
0.13
0.09
0.05
0
I
III
IV
8 2 15 14 26 24 6 28 1 9 18 21 20 19 12 7 17 31 16 4 30 32 27 22 23 11 10 13 5 29 25 3
*cj1349 (HP)
*cadF (oprF)
porA
similar to PorA
pgaB
*ciaB (HP)
inlJ
inlA
TonB domain
transport protein TonB domain
ccdB
TonB receptor domain
exbB
exbD
tdhA
epsE
epsF
xcpQ
epsD
epsH
epsM
epsN
epsI
virB10
virB8
virB6
virB4
virB3
virB2
pldA
*tlyA
*mviN (murJ)
HlyD domain
^hlyB
*hecA (shlA)
*hecA (hpmA)
*hecB (shlB)
cyaA
rmlD
rmlA
kfoC
wbbL
O-antigen ligase
cotSA
pgα
sunS
rfaC
RimI domain
Capsule biosynthesis domain
^Arylesterase
*irgA (cirA)
*iroE (besA)
mprA
phoQ
phoP
fixL
zraS
pdtaS
ttrS
bvgS
^tabA
Urease cluster (ureG, ureF, ureE, ureC, ureA, ureD1)
Flagellar assembly protein H
Flagellar basal body rod modification protein
Flagellar basal body-associated protein FliL
Flagellar biosynthesis protein FliR
Flagellar hook-associated protein FlgL
Flagellar hook-basal body protein FliE
Flagellar hook-length control protein FliK
flgB
flgC
flgE
flgG
flgH
flgI
flgK
flhA
flhB
flhF
fliD
fliF
fliG
fliI
fliM
fliN
fliQ
fliS
ylxH
^flaA
^flaB
flagellar motor switch protein
Flagellin N-methylase
fliP
^fliW1
motB
swrC
Chemotaxis cluster(cheB, cheR, cheA, cheY)
cheD
cheV
cheW
Putative functions:
Host colonization and invasion
Secretion systems
Hemolysis
Pathogen and host mucosal interaction
Flagellum, motiliy and chemotaxys
General virulence mechanism
Heatmap key:
Gene present (> 1 copy)
Gene present (1 copy)
Absent

Fig. 5

974 **Tables**

975 **Table 1**

Strain code in this study	Source	Additional information	Country
1	Human faeces	Stool sample, (Strain LMG 147149)	Greece
2	Human faeces	Stool sample, (Strain LMG 111199)	Italy
3	Human faeces	Stool sample, (Strain LMG 10828^{T})	U.S.A
4	Dog faeces	/	Belgium
5	Sheep faeces	/	Belgium
6	Horse faeces	/	Belgium
7	Cow faeces	/	Belgium
8	Chicken skin	collected from neck	Belgium
9	Pig meat	/	Belgium
10	Pig rectum	Intestinal content, (rc1-13)	Belgium
11	Pig rectum	Intestinal content, (rc1-14)	Belgium
12	Pig rectum	Intestinal content, (rc2-10)	Belgium
13	Pig rectum	Intestinal content, (rc2-20)	Belgium
14	Pig duodenum	Intestinal content, (dc1-3AAN)	Belgium
15	Pig caecum	Intestinal content, (cm1-2AAN)	Belgium
16	Pig colon descendent	Mucus, (cdm1-1AAN)	Belgium
17	Pig colon descendent	Intestinal content, (cdc2-1AAN)	Belgium
18	Pig colon descendent	Intestinal content, (cdc2-2AAN)	Belgium
19	Pig rectum	Intestinal content , (rc1-2kAAN)	Belgium
20	Pig rectum	Intestinal content, (rc1-3AAN)	Belgium
21	Pig rectum	Mucus, (rm1-2AAN)	Belgium
22	Pig rectum	Intestinal content, (rc2-1AAN)	Belgium
23	Pig rectum	Mucus, (rm2-1AAN)	Belgium
24	Human faeces	Stool sample (male, 90 y/o, diarrhea)	Belgium
25	Human faeces	Stool sample (female, 93 y/o, acute gastroenteritis)	Belgium
26	Human faeces	Stool sample (male, 83 y/o, acute gastroenteritis)	Belgium
27	Human faeces	Stool sample (male, 4 y/o, acute gastroenteritis)	Belgium
28	Human faeces	Stool sample (male 59 y/o)	Belgium
29	Human faeces	Stool sample (male, 51 y/o, diverticulitis)	Belgium
30	Human faeces	Stool sample (male, 55 y/o, traveler's diarrhea)	Belgium
31	Human faeces	Stool sample (female, 80 y/o, flair up colitis ulcerosa)	Belgium
32	Human faeces	Stool sample (female, 79 y/o, recurrent diarrhea episodes)	Belgium

976

977

978

979

980 **Table 2**

	Pangenome partitions						
Methods	Core (> 99%)	Soft core (95% <= strains < 99%)	Persistent (> 95 %)	Shell (15% <= strains < 95%)	Cloud (< 15%)	Accessory /of which singletones (< 99%)	Pangenome
Roary:	1587	76	1663	970	4703	5749 / 3311	7336
PPanGGolin:	1651	155	1806	275	4542	4972 / 2755	6623

981

982