TRANSPARENT PROCESS
OPEN ACCESS
THE EMBO JOURNAL

A versatile genetic toolbox for *Prevotella copri* enables studying polysaccharide utilization systems

Jing Li[1], Eric J C Gálvez[1,2,†], Lena Amend[1,†], Éva Almási[1], Aida Iljazovic[1], Till R Lesker[1], Agata A Bielecka[1], Eva-Magdalena Schorr[1] & Till Strowig[1,2,3,*]

Abstract

Prevotella copri **is a prevalent inhabitant of the human gut and has been associated with plant-rich diet consumption and diverse health states. The underlying genetic basis of these associations remains enigmatic due to the lack of genetic tools. Here, we developed a novel versatile genetic toolbox for rapid and efficient genetic insertion and allelic exchange applicable to** ***P. copri*** **strains from multiple clades. Enabled by the genetic platform, we systematically investigated the specificity of polysaccharide utilization loci (PULs) and identified four highly conserved PULs for utilizing arabinan, pectic galactan, arabinoxylan, and inulin, respectively. Further genetic and functional analysis of arabinan utilization systems illustrate that** ***P. copri*** **has evolved two distinct types of arabinan-processing PULs (PUL^{Ara}) and that the type-II PUL^{Ara} is significantly enriched in individuals consuming a vegan diet compared to other diets. In summary, this genetic toolbox will enable functional genetic studies for** ***P. copri*** **in future.**

Keywords genetic manipulation; human diet; hybrid two-component system; polysaccharide utilization locus; *Prevotella copri*

Subject Categories Metabolism; Methods & Resources; Microbiology, Virology & Host Pathogen Interaction

DOI 10.15252/embj.2021108287 | Received 18 March 2021 | Revised 8 September 2021 | Accepted 24 September 2021 | Published online 21 October 2021

The EMBO Journal (2021) 40: e108287

Introduction

The complex microbial communities residing in the intestine affect the physiology of the host influencing the balance between health and disease (Bäckhed *et al*, 2005; Hooper, 2009). Yet, extensive interpersonal variability in the human gut microbiota composition and function complicates the establishment of links between the presence of specific bacterial gene content to human phenotypes. Functional genetic study of these bacteria is essential to dissect the genetic basis underlying the microbe-driven host phenotypes. However, many commensal bacterial species have so far eluded efforts for genetic engineering.

For instance, no genetic tools have been established for *Prevotella copri*, a common human gut microbe, whose prevalence and relative abundance have been linked to various beneficial and detrimental effects on human health (Ley, 2016; Claus, 2019; Maeda & Takeda, 2019). Specifically, *P. copri* has been found to be enriched in individuals at risk for rheumatoid arthritis (Scher *et al*, 2013; Alpizar-Rodriguez *et al*, 2019; Wells *et al*, 2020) and in patients with enhanced insulin resistance and glucose intolerance (Pedersen *et al*, 2016). Conversely, others found *P. copri* to be positively correlated with improved glucose and insulin tolerance during intake of fiber-rich prebiotic diets (Kovatcheva-Datchary *et al*, 2015; De Vadder *et al*, 2016). Besides the lack of tools for genetic engineering, the establishment of functional links between *P. copri* and disease outcomes has been additionally complicated by its fastidious nature *in vitro*, tremendous strain-level diversity, resulting in the recent recognition of multiple genetically distinct clades and the lack of corresponding diverse isolates (Tett *et al*, 2019).

In contrast to *P. copri*, members of the genus *Bacteroides*, as the best example, have been extensively studied via a variety of genetic tools (Koropatkin *et al*, 2008; Goodman *et al*, 2011; Mimee *et al*, 2015; Lim *et al*, 2017; García-Bayona & Comstock, 2019; Bencivenga-Barry *et al*, 2020). These studies have, for instance, identified genes required for various bacterial physiological functions and provide approaches to investigate bacteria–host interactions. Of those, the genes for degrading plant- and animal-derived polysaccharides that are resistant to human digestion have been highlighted due to their important role in affecting bacterial fitness in the microbiome (Kaoutari *et al*, 2013; Porter & Martens, 2017; Wexler & Goodman, 2017). These genes are typically organized in so called polysaccharide utilization loci (PULs) that differ in polysaccharide specificity (Koropatkin *et al*, 2012). PULs are defined by the presence of one or more genes homologous to *Bacteroides thetaiotaomicron susD* and *susC* encoding outer membrane proteins that bind and import starch oligosaccharides (Shipman *et al*, 2000; Martens *et al*, 2009). The SusC/D protein complex (Glenwright *et al*, 2017) cooperates with diverse carbohydrate-degrading enzymes (CAZyme), e.g., glycosyl hydrolases (GHs) and polysaccharide

1 Department of Microbial Immune Regulation, Helmholtz Centre for Infection Research, Braunschweig, Germany
2 Hannover Medical School, Hannover, Germany
3 Centre for Individualized Infection Medicine, Hannover, Germany
*Corresponding author (lead contact). Tel: +49 531 6181 4700; E-mail: till.strowig@helmholtz-hzi.de
†These authors contributed equally to this work

lyases (PLs), which are typically encoded in close proximity to the *susC/D* homologs in the genome. Many PULs in *B. thetaiotaomicron* contain genes encoding sensor-regulator systems, such as hybrid two-component systems (HTCSs) (Xu *et al*, 2003; Sonnenburg *et al*, 2006). HTCS proteins are chimeric proteins harboring the functional domains of a periplasmic sensor, a histidine kinase, and a DNA-binding response regulator enabling HTCSs to recognize distinct signal components degraded from complex carbohydrates and to initiate the upregulation of CAZyme-encoding genes in a positive feedback loop (Sonnenburg *et al*, 2006; Martens *et al*, 2011).

Notably, higher prevalence of intestinal *Prevotella* spp. was found in populations consuming a plant-rich diet, e.g., vegetarians in Western populations (De Filippo *et al*, 2010; Wu *et al*, 2011; Ruengsomwong *et al*, 2016; Fragiadakis *et al*, 2019), suggesting that they encode efficient machineries for degradation of plant-derived polysaccharides. Yet, due to the lack of genetic tools, the characterization of carbohydrate utilization has been limited to bioinformatic and phenotypic studies (De Filippis *et al*, 2019; Fehlner-Peach *et al*, 2019). Specifically, two recent studies described extensive variability among clades and also strains within clades in the ability to directly utilize diverse complex plant carbohydrates (Fehlner-Peach *et al*, 2019; Tett *et al*, 2019). While combinations of comparative genomics and phenotypic assays can be used to predict the ability to utilize specific polysaccharides, such approaches rely on well-characterized genetic elements as a reference, which makes it difficult to identify genes with unknown functions and thereby hinders the establishment of casual relationship between the genetic content and phenotypes. Moreover, substrate predictions based on gene annotations from genetically distinct bacteria might be incomplete or inaccurate. This observation is supported by the presented data below that some *P. copri* strains harboring PULs lacking *susC/D* genes can still grow on the substrates of those PULs, suggesting the functional redundancy between various PUL components.

Here, we described a newly established genetic toolbox for approaching gene insertion, deletion, and complementation in *P. copri*. Using the genetic tools as well as high-throughput sequencing and bioinformatic analysis, we identified four highly conserved *P. copri* PULs responsible for utilization of specific plant polysaccharides via HTCS activation, and demonstrated that *P. copri* strains have evolved two types of arabinan processing PULs. These studies not only build up a universal genetic manipulation system for an abundant bacterial species in the microbiome, but also present its applications on future efforts of understanding *P. copri* biology, e.g., nutrient acquisition. Because the workflow of establishing the genetic manipulation system for *P. copri* can be potentially modified and applied to other underexplored gut bacteria, our studies shed light on the future microbiome research on intricate interactions between bacteria–bacteria and host–bacteria during human health and disease.

Results

Development of conjugation-based gene insertion system for *P. copri*

As targeted gene inactivation approaches enable gene function studies, they are frequently carried out in *Bacteroides* spp. by transferring a suicide plasmid from a donor strain into the recipient followed by selection of bacterial clones, which underwent homologous recombination (Koropatkin *et al*, 2008; García-Bayona & Comstock, 2019; Bencivenga-Barry *et al*, 2020). To adapt the system for *P. copri*, we considered several key differences between *Bacteroides* spp. and *P. copri* including oxygen and antibiotic sensitivity as well as promoter sequences driving expression of the selectable marker gene.

Because oxygen exposure has been reported to promote mating between *Escherichia coli* (donor) and *Bacteroides* spp. (recipient) (Salyers *et al*, 1999), conjugation for *Bacteroides* spp. is routinely performed for at least 15 h under aerobic conditions followed by transferring the cultures to anaerobic conditions permitting growth (García-Bayona & Comstock, 2019; Bencivenga-Barry *et al*, 2020). We initially tested aerotolerance of three *P. copri* strains, the type strain (DSM18205) and two strains (HDD04 and HDB01) from our laboratory collection containing recent isolates from healthy and diseased individuals (Fig 1A and Dataset EV1). Exposure to air decreased viability three to four orders of magnitude for the *P. copri* strains within only 4 h, which is in sharp contrast to *B. thetaiotaomicron* that displayed only a 28.6% drop in viability (Fig EV1A). Hence, all genetic manipulations for *P. copri* were subsequently carried out in anaerobic conditions.

Escherichia coli S17-1 λpir is commonly used as a donor for *Bacteroides* spp. in aerobic conditions, but it may show impaired growth under anaerobic conditions. Hence, we compared anaerobic growth of *E. coli* S17-1 λpir and another donor strain, *E. coli* β2155 (Dehio & Meyer, 1997; Demarre *et al*, 2005). Notably, while *E. coli* β2155 is auxotrophic for diaminopimelic acid (DAP), it displayed in the presence of DAP faster and more robust anaerobic growth compared to *E. coli* S17-1 λpir both in liquid culture and on agar plates (Fig EV1B and C). Thus, we further tested the possibility of using *E. coli* β2155 as the donor for delivering vectors into recipient *P. copri*.

Figure 1. Development of a conjugation-based gene insertion platform for *Prevotella copri* strains from multiple clades. ▶

A Phylogenetic tree of *P. copri* species complex using reference strains (n = 17) from four *P. copri* clades (Tett *et al*, 2019) and novel isolates (n = 11) used in this study. *Prevotella copri* clades are indicated by different colors.

B Schematic illustration for targeted gene insertion system in *P. copri*. Primer binding sites (P1–P6) are indicated.

C Detection of plasmid integration in *P. copri* DSM 18205 and HDD04 by PCR.

D–I Optimization of conjugation efficacy: Influence on yield of transconjugants of promoter sequences of selection marker (D), length of homology arm (E), donor *Escherichia coli* strain (F), conjugation ratio of donor to recipient strains (G), recipient *P. copri* strains that are erythromycin-sensitive (H), or tetracycline-sensitive (I). Comparisons in (D–G) were performed using *P. copri* HDD04. ND: not detectable.

Data information: Values and error bars represent the mean of at least three biological replicates and their standard deviations (SDs), respectively. Statistical significance between groups was calculated by Student's t test (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; and ****$P < 0.0001$; NS, $P > 0.05$, not statically significant).

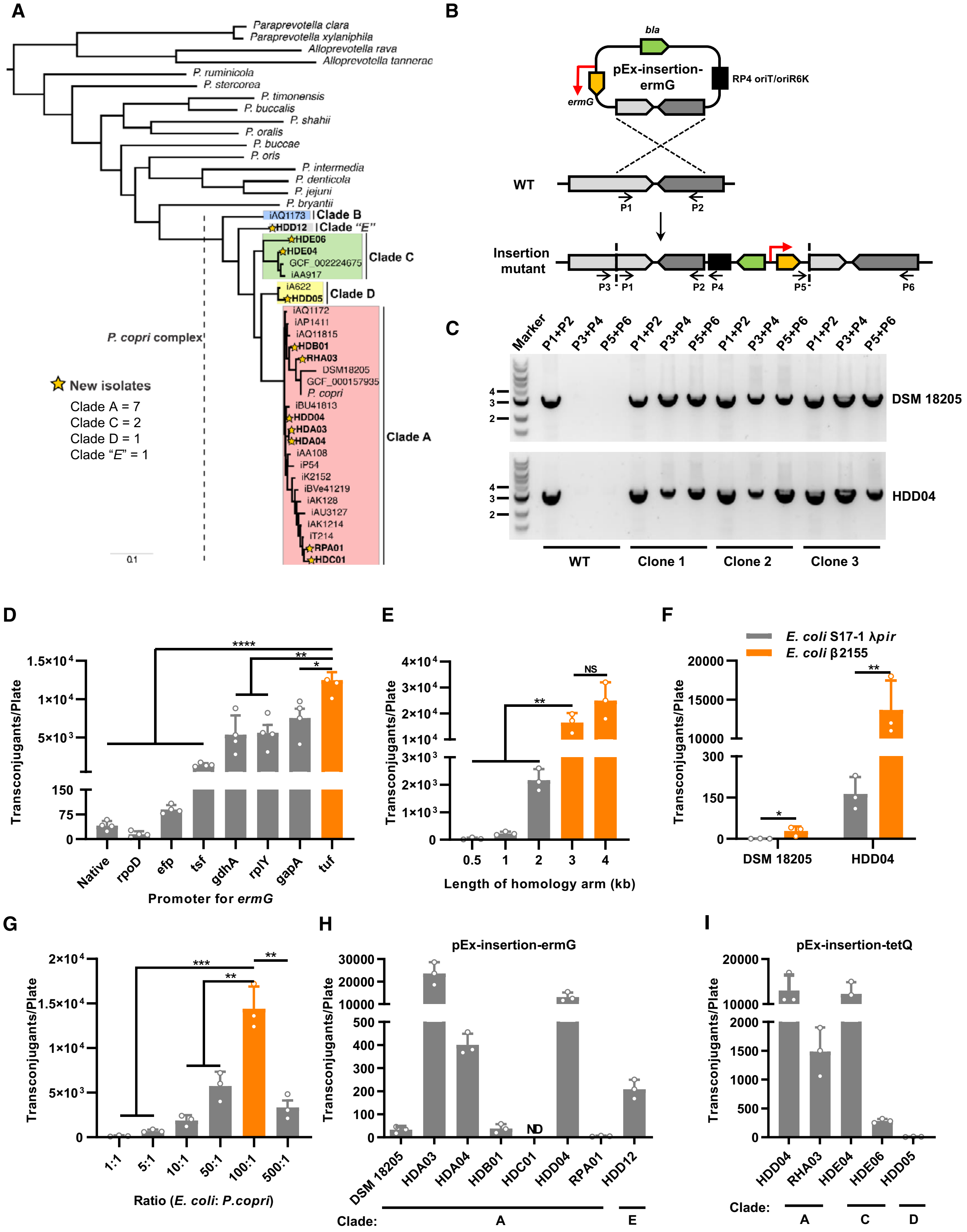
A
Paraprevotella clara
Paraprevotella xylaniphila
Alloprevotella rava
Alloprevotella tannerae
P. ruminicola
P. stercorea
P. timonensis
P. buccalis
P. shahii
P. oralis
P. buccae
P. oris
P. intermedia
P. denticola
P. jejuni
P. bryantii
iAQ1173 Clade B
HDD12 Clade "E"
HDE06
HDE04
GCF_002224675
iAA917
Clade C
iA622
HDD05
Clade D
iAQ1172
iAP1411
iAQ11815
HDB01
RHA03
DSM18205
GCF_000157935
P. copri
iBU41813
HDD04
HDA03
HDA04
iAA108
iP54
iK2152
iBVe41219
iAK128
iAU3127
iAK1214
iT214
RPA01
HDC01
Clade A
P. copri complex
New isolates
Clade A = 7
Clade C = 2
Clade D = 1
Clade "E" = 1
0.1
B
bla
pEx-insertion-ermG
RP4 oriT/oriR6K
ermG
WT
P1
P2
Insertion mutant
P3
P1
P2
P4
P5
P6
C
Marker
P1+P2
P3+P4
P5+P6
P1+P2
P3+P4
P5+P6
P1+P2
P3+P4
P5+P6
P1+P2
P3+P4
P5+P6
4
3
2
DSM 18205
4
3
2
HDD04
WT
Clone 1
Clone 2
Clone 3
D
Transconjugants/Plate
1.5×10⁴
1×10⁴
5×10³
150
75
0
****
**
*
Native
rpoD
efp
tsf
gdhA
rplY
gapA
tuf
Promoter for ermG
E
Transconjugants/Plate
4×10⁴
3×10⁴
2×10⁴
1×10⁴
3×10³
2×10³
1×10³
0
**
NS
0.5
1
2
3
4
Length of homology arm (kb)
F
E. coli S17-1 λpir
E. coli β2155
Transconjugants/Plate
20000
15000
10000
300
150
0
*
**
DSM 18205
HDD04
G
Transconjugants/Plate
2×10⁴
1.5×10⁴
1×10⁴
5×10³
0
***
**
**
1:1
5:1
10:1
50:1
100:1
500:1
Ratio (E. coli: P.copri)
H
pEx-insertion-ermG
Transconjugants/Plate
30000
20000
10000
500
400
300
200
100
0
ND
DSM 18205
HDA03
HDA04
HDB01
HDC01
HDD04
RPA01
HDD12
Clade:
A
E
I
pEx-insertion-tetQ
Transconjugants/Plate
20000
10000
2000
1500
1000
500
0
HDD04
RHA03
HDE04
HDE06
HDD05
Clade:
A
C
D

Figure 1.

The suicide plasmid, pExchange-tdk, is extensively used for gene deletion in *Bacteroides* spp. (Koropatkin *et al*, 2008). This plasmid possesses (i) a R6K origin limiting it to replicate only in host–bacteria carrying the *pir* gene, and (ii) an erythromycin-resistant gene (*ermG*) for selecting *Bacteroides* transconjugants. Firstly, no *pir* homologs were identified in *P. copri* strains, suggesting the feasibility of using pExchange-tdk-based vectors for plasmid integration in *P. copri*. Secondly, the susceptibility of *P. copri* strains to erythromycin was tested. From one type strain and 11 distinct strains of our collection representing four *P. copri* clades, eight strains from two clades (Fig 1A) were sensitive to erythromycin driving us to initially utilize an erythromycin-based selection system (Dataset EV2). To achieve a stable expression of *ermG* in *P. copri*, we inserted a strong promoter of a *P. copri* housekeeping gene, i.e., elongation factor Tu gene (*tuf*) (Fig EV1D), in front of the *ermG* coding sequence into pExchange-tdk and removed the counterselection marker (thymidine kinase gene). These modifications resulted in a new shuttle vector referred to as pEx-insertion-ermG (Fig 1B and Appendix Fig S1A). To consider potential negative positional effects for *P. copri* growth due to plasmid insertion, we individually cloned three different 3-kb regions (DSM18205_00642-43, 00941-42, and 02334-35) containing the 3′-end coding sequences of genes from *P. copri* DSM 18205, as homology arms for approaching plasmid integration without disrupting any functional genes. In addition, these DNA regions are relatively conserved in genomes of our *P. copri* isolates, allowing us to rapidly expand the testing into different strains.

The initial conjugation was performed between *E. coli* β2155 carrying the respective plasmids and *P. copri* DSM 18205. Of note, *E. coli* β2155 can grow weakly on BHI blood agar plates even without DAP based on our observations. Hence, after the co-incubation of *E. coli* and *P. copri* on BHI blood agar with DAP, besides erythromycin, we additionally added gentamicin to inhibit the *E. coli* donor and positively select for *P. copri* transconjugants. This selection yielded eight erythromycin resistant (Erm^R) colonies only when pEx-insertion-ermG-DSM18205_02334-35 was used, which indeed suggests the presence of positional effects (Fig 1B). Since DSM18205_02334 encodes a putative β-glycosidase gene (*bgl*), we refer to the plasmid as pEx-insertion-ermG-DSM-bgl. Integration of the transferred plasmid was determined in three individual colonies and colony PCR amplified fragments expected for successful integration (Fig 1C and Dataset EV1). DNA sequences of PCR products were further confirmed by Sanger sequencing. We next attempted conjugation using pEx-insertion-ermG-HDD04-bgl and HDD04 as the conjugation recipient, which strikingly resulted in a more than 10,000 Erm^R colonies suggesting a large strain variability in conjugation efficiency and prompting us to further optimize the approach.

Optimization of conjugation-based gene insertion for *P. copri*

Building on these proof-of-concept data, the genetic elements of the plasmid and the conjugation procedures were systematically adjusted by varying factors that likely affect the conjugation efficiency. This included the promoter of *ermG*, the length of homology arm for plasmid integration, the donor *E. coli* strains and ratio of donor to recipient for conjugation, as well as the recipient *P. copri* strains.

First, an efficient expression of the selection marker is the prerequisite to obtain Erm^R transconjugants. Besides the promoter of the *tuf* gene, we chose another six different promoters of housekeeping genes showing diverse gene expression in *P. copri* HDD04 in BHI liquid media supplemented with fetal bovine serum (BHI+S; Figs 1D and EV1D; Dataset EV5). Notably, the numbers of transconjugants varied approximately 300-fold depending on the promoter, but none of the other promoters yielded higher numbers than *tuf*, suggesting that high *ermG* expression levels are required for transconjugant survival under erythromycin selection (Figs 1D and EV1D). Second, comparison of homology arms between 0.5 and 4 kb demonstrated the highest yield of transconjugants with homology arms of 3 and 4 kb (Fig 1E). Third, we assessed the above-mentioned *E. coli* strains as conjugation donors. In line with the ability for anaerobic growth, conjugation with donor *E. coli* β2155 increased the yield of transconjugants for both DSM 18205 and HDD04 approximately 85-fold compared to *E. coli* S17-1 λpir, indicating the advantage of donor strain for advancing anaerobic conjugation (Fig 1F). More transconjugants were also obtained as ratio of donor to recipient increased until 100:1, after which it decreased again (Fig 1G). Last, we evaluated a larger panel of Erm^S *P. copri* strains using strain-specific homology arms as extensive sequence variations among strains are present. Except for one strain with undetectable production of Erm^R colonies all other seven strains exhibited extensive diversity in the number of transconjugants varying by approximately 10^4-fold between the strains with the lowest (RPA01, mean = 5 CFUs) and highest (HDA03, mean = 2.4×10^4 CFUs) yield (Fig 1H).

While these iterative improvements allowed the targeted insertion in seven strains from the clade A and "E" (a newly observed clade, unpublished observation), the other four strains from the clade A, C, and D could not be assessed due to their Erm^R phenotype (Dataset EV2). Hence, we screened the antibiotic susceptibility of HDD04 and DSM 18205 identifying tetracycline, chloramphenicol, and spectinomycin as additional selective antibiotics (Dataset EV2). Next, *P. copri* HDD04 was utilized to test the feasibility of tetracycline, chloramphenicol, and spectinomycin resistance genes, i.e., *tetQ*, *catA*, *aadA*, for selection. Conjugation using pEx-insertion carrying *tetQ* (pEx-insertion-tetQ-bgl, Appendix Fig S1B), but not *catA* or *aadA*, successfully resulted in HDD04 transconjugants after selection with the respective antibiotics. We therefore performed the conjugation for four Erm^R but tetracyclin-sensitive (Tet^S) strains, i.e., RHA03, HDE04, HDE06, and HDD05, followed by tetracycline selection, resulting in tetracycline resistant (Tet^R) transconjugants (mean = 8.3 to 1.2×10^4 CFUs; Fig 1I).

In summary, the comprehensive and stepwise adaptation of a gene insertion system to a genetically inaccessible bacterium, i.e., *P. copri,* illustrates the significant influence of multiple variables for a successful production of genetic mutants, thereby providing a valuable template for initiating the construction of genetic tools for other commensals in future. These experiments together demonstrate the feasibility of gene insertion in *P. copri* strains from distinct clades enabling functional studies.

Genetic inactivation of PUL regulators to identify their polysaccharide substrates

The prevalence and relative abundance of *P. copri* has been linked to plant-rich diets in humans and mouse models (De Filippo *et al*, 2010; Wu *et al*, 2011; Ruengsomwong *et al*, 2016; Fragiadakis *et al*, 2019; Kovatcheva-Datchary *et al*, 2019; Gálvez *et al*, 2020), yet the underlying molecular mechanism is still poorly understood. To demonstrate the utility of our gene insertion system in identifying gene functions, we decided to investigate the genetic basis for utilization of distinct carbohydrates in *P. copri*. Since no available chemically defined cultivation system for *P. copri* exist thus far, we modified the minimal medium (MM) (Martens *et al*, 2008) originally used for cultivation of *B. thetaiotaomicron* by supplementing additional defined nutrients (see Materials and Methods), enabling the growth of all *P. copri* strains tested from our strain collection (n = 12) with glucose as a sole carbon source (Dataset EV3). Specifically, 10 out of 12 strains reached a maximal optical density (OD$_{600}$ max) of 0.6–1.0 in MM + Glucose overnight, while two strains (HDA03 and RHA03) showed only moderate growth (approximately OD$_{600}$ max 0.3). This minimal medium was then used to extensively characterize polysaccharide utilization in HDD04, as it showed robust growth in MM and high number of transconjugants. HDD04 grew on various plant cell wall glycans, such as arabinan, arabinogalactan, and arabinoxylan (Dataset EV3). Beyond the utilization of plant cell wall glycans, HDD04 also showed growth on plant and animal cell storage carbohydrates such as inulin (0.651 ± 0.016) and glycogen (0.818 ± 0.026), but grew poorly on levan (0.122 ± 0.009), and could not grow on starch. In parallel, PULs were identified in *P. copri* using a bioinformatic approach PULpy (preprint: Stewart *et al*, 2018) followed by manual curation. Specifically, the PUL repertoire of HDD04 was predicted based on *susC/D*-like pairs resulting in 29 PULs in comparison with 19 PULs in the *P. copri* reference strain (DSM 18205; Dataset EV3), suggesting a much broader carbohydrate utilization capability of HDD04 compared to DSM 18205. CAZyme content in the PULs and their substrate were predicted using bioinformatic approaches (dbCAN2 tool and dbCAN-PUL) (Zhang *et al*, 2018; Ausland *et al*, 2021) (Fig EV2).

To directly link distinct PULs and growth phenotypes on polysaccharides, we focused on HTCS genes, the typical activator associated with PULs (Sonnenburg *et al*, 2006; Martens *et al*, 2011). Genome-wide screening for HTCS genes in HDD04 by homology search using the known domains of HTCS (Terrapon *et al*, 2015) combined with a protein BLAST on National Center for Biotechnology Information (NCBI) identified 10 gene candidates as our targets. We associated nine out of 10 HTCS gene candidates with their closest predicted PULs (e.g. HDD04_00018 is named as *htcs*-PUL3) with only HDD04_0019 being a solitary HTCS gene.

Ten HTCS insertion mutants were generated by integrating a modified pEx-insertion-ermG plasmid (Appendix Fig S1C) into the coding sequences of their periplasmic sensor domains followed by a screening for growth defects in MM plus polysaccharides to identify their respective substrates (Fig 2A). Of note, to block potential effects of transcriptional and translational readthrough for the HTCS genes after plasmid integration, pEx-insertion-ermG was modified to include T1–T2 terminators and TAA encoding stop codon in front of the cloned homology arm, respectively (pEx-insertion-ermG-T1T2; Fig 2A and Appendix Fig S1C). A strain with plasmid integrated into an intergenic region (between HDD04_00165 and 00166) was utilized as a positive control. The polysaccharides (n = 14) and glucose that can support the growth of HDD04 to an OD$_{600}$ max of > 0.2 within 120 h were investigated (Fig 2B). Compared to the control strain, 6/10 HTCS gene mutants displayed similar growth patterns, e.g., *htcs*-PUL3, as shown in Fig 2B, demonstrating that they are not essential for growth on the tested polysaccharides. Strikingly, the other four HTCS mutants each showed dramatic growth defects (OD$_{600}$ max < 0.1) on only one specific polysaccharide (Figs 2C and D). Specifically, gene disruptions of *htcs*-PUL14, *htcs*-PUL21, *htcs*-PUL24, and *htcs*-PUL26 abolished the capacities of HDD04 grown on arabinan, pectic galactan, arabinoxylan, and inulin, respectively (Fig 2C and D). To link the HTCS gene and the nearest PUL for easy identification, we have tentatively designated these genes as *htcs*$^{\mathrm{D\_Ara}}$ (*htcs*-PUL14, HDD04_02372), *htcs*$^{\mathrm{D\_PecGal}}$ (*htcs*-PUL21, HDD04_02939), *htcs*$^{\mathrm{D\_AraXyl}}$ (*htcs*-PUL24, HDD04_03129), and *htcs*$^{\mathrm{D\_Inu}}$ (*htcs*-PUL26, HDD04_03217).

Together, these experiments not only demonstrate the utility of the gene inactivation strategies to perform functional studies in *P. copri*, but also uncovered the link between PUL-associated regulatory genes and metabolic phenotypes on utilizing specific polysaccharides for *P. copri*.

Construction of an allelic exchange system for validating the function of PUL regulators on polysaccharide degradation

Although our genetic insertion system is efficient in generating mutants for rapid determinations of phenotypes, it some limitations: (i) The selective pressure provided by antibiotics is constantly required in the medium for plasmid integrants with corresponding antibiotic resistance markers; (ii) it would be challenging to target relatively smaller genes because the yield of conjugants is negatively correlated with the homology arm cloned in the conjugative plasmid as shown in Fig 1E; (iii) the integration of the plasmid may cause a polar effect on the expression of down-stream genes, especially complicating the characterization of each gene's role in an operon. Therefore, we aimed to establish an allelic exchange system for unmarked gene deletion and complementation in *P. copri*. One of the most widely used system for allelic exchange in bacteria is based on the levansucrase gene (*sacB*), which catalyzes the hydrolysis of sucrose and synthesizes the toxic compound levan (Gay *et al*, 1985; Recorbet *et al*, 1993). Two key criteria have been identified to limit its application: (i) The inability of bacteria to grow properly on agar plates in the presence of relatively high concentrations of sucrose and (ii) whether the expression of *sacB* can effectively select gene deletion mutants in the presence of sucrose.

Hence, we first determined the growth of *P. copri* DSM18205 and HDD04 on agar plates with increasing concentrations of sucrose. As media base, we employed yeast extract and tryptone (YT) supplemented with horse blood instead of BHI to reduce salt concentrations, which have been demonstrated to decrease the sucrose sensitivity of *E. coli* (Blomfield *et al*, 1991) and have been observed by us to interfere with *P. copri* growth on high sucrose concentration. Both strains showed the normal colony numbers and morphology until a sucrose concentration of 6%, while *E. coli* tolerated up to 10% sucrose (Fig EV2A). The inability to grow under these conditions was likely caused by osmotic pressure, as similar results were obtained for *P. copri* strains with increasing the concentration of glucose in YT + blood agar (Fig EV2B). In order to ensure the selectivity of sucrose without affecting growth of *P. copri*, a working concentration of 5% sucrose was chosen. Notably, 5/10 strains were not able to grow well in the YT + blood media in the presence of 5% sucrose reflecting their distinct growth properties compared to other strains (Fig EV2C).

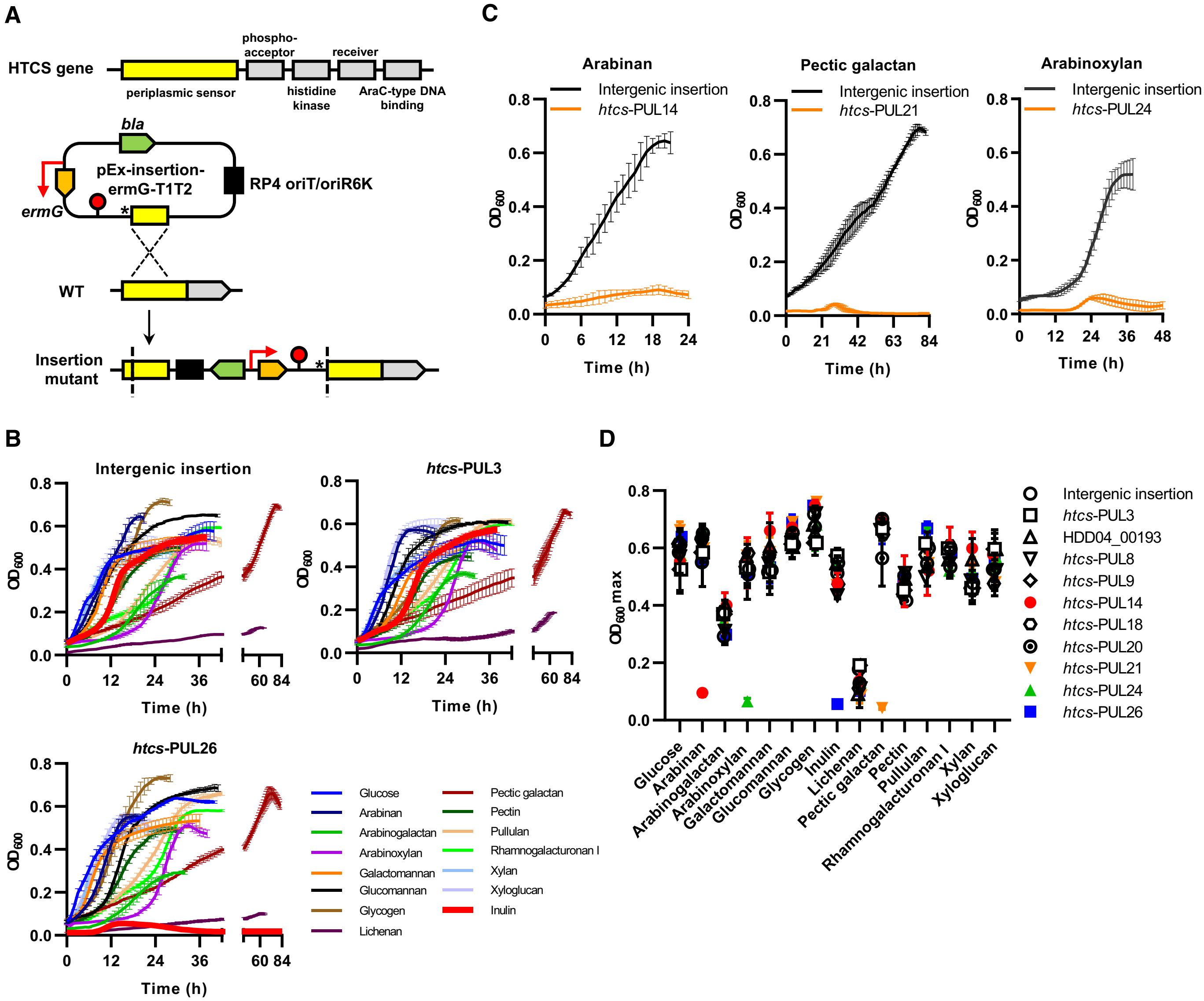

Figure 2. Identification of HTCS and associated PULs essential for utilization of polysaccharides using targeted gene inactivation.

A Schematic illustration for gene inactivation strategy targeting HTCS gene candidates in *Prevotella copri* HDD04. A classical Reg_prop HTCS (Martens *et al*, 2011) harboring core domains is shown as an example. Partial coding sequence of the periplasmic sensor domain was cloned into the pEx-insertion-ermG-T1T2 vector as the homolog arm for genetic insertion of the HTCS gene. This vector contains T1–T2 terminators (red ball-and-stick symbol) and TAA encoding stop codon (asterisk) for avoiding possible transcriptional and translational readthrough for the targeted gene after plasmid integration.

B Growth of *P. copri* HDD04 strains with plasmid insertions at an intergenic region (control) and two representative putative HTCS genes, respectively, in minimal media (MM) supplemented with glucose or indicated polysaccharides.

C Growth of *P. copri* HDD04 strains with respective *htcs*-PUL14, *htcs*-PUL24, and *htcs*-PUL24 insertion compared to an intergenic insertion mutant on arabinan, pectic galactan, and arabinoxylan, respectively.

D Maximum growth (OD_{600} max) of mutant strains (n = 11) in MM supplemented with glucose or indicated polysaccharides.

Data information: Error bars represent the standard error of the means (SEMs) in (B) and SDs in (C) and (D) of the biological replicates from three carbohydrate arrays with each carbohydrate tested in duplicate, respectively.

Next, a derivative vector of pEx-insertion-ermG named pEx-deletion-ermG was created by joining the promoter of *gdhA* gene to a promoterless copy of *sacB* and inserting it down-stream of *ermG* (Fig 3A and Appendix Fig S1D). The homology arm for targeting the *bgl* gene (Fig 2A) was cloned into the pEx-deletion-ermG, resulting in pEx-deletion-ermG-HDD04-bgl. We individually integrated the pEx-deletion-ermG-bgl and pEx-insertion-ermG-bgl into HDD04 (Fig 3B). Erm^R colonies were readily obtained for both plasmids and displayed normal colony morphology as wild type, indicating that the expression of *sacB* did not affect the growth of *P. copri* in the absence of sucrose (Fig 3B). Plating these Erm^R colonies containing pEx-deletion-ermG-bgl in the presence of sucrose significantly reduced CFU by 10^4-fold (Fig 3B). In contrast, the same strain carrying pEx-insertion-ermG-*bgl* exhibited equivalent viability in the presence and absence of sucrose. These results illustrated that expression of *sacB* effectively functioned as sucrose-based selection.

We subsequently assessed the false positive rate of this counterselection system by plating Erm^R on YT+Erm+Suc plates and found that approximately 1 out of 6.7 × 10^5 cells in the bacterial population were Suc^R but still Erm^R, i.e. carried the plasmid. Of note, 50 from 500 colonies were randomly picked and restreaked on

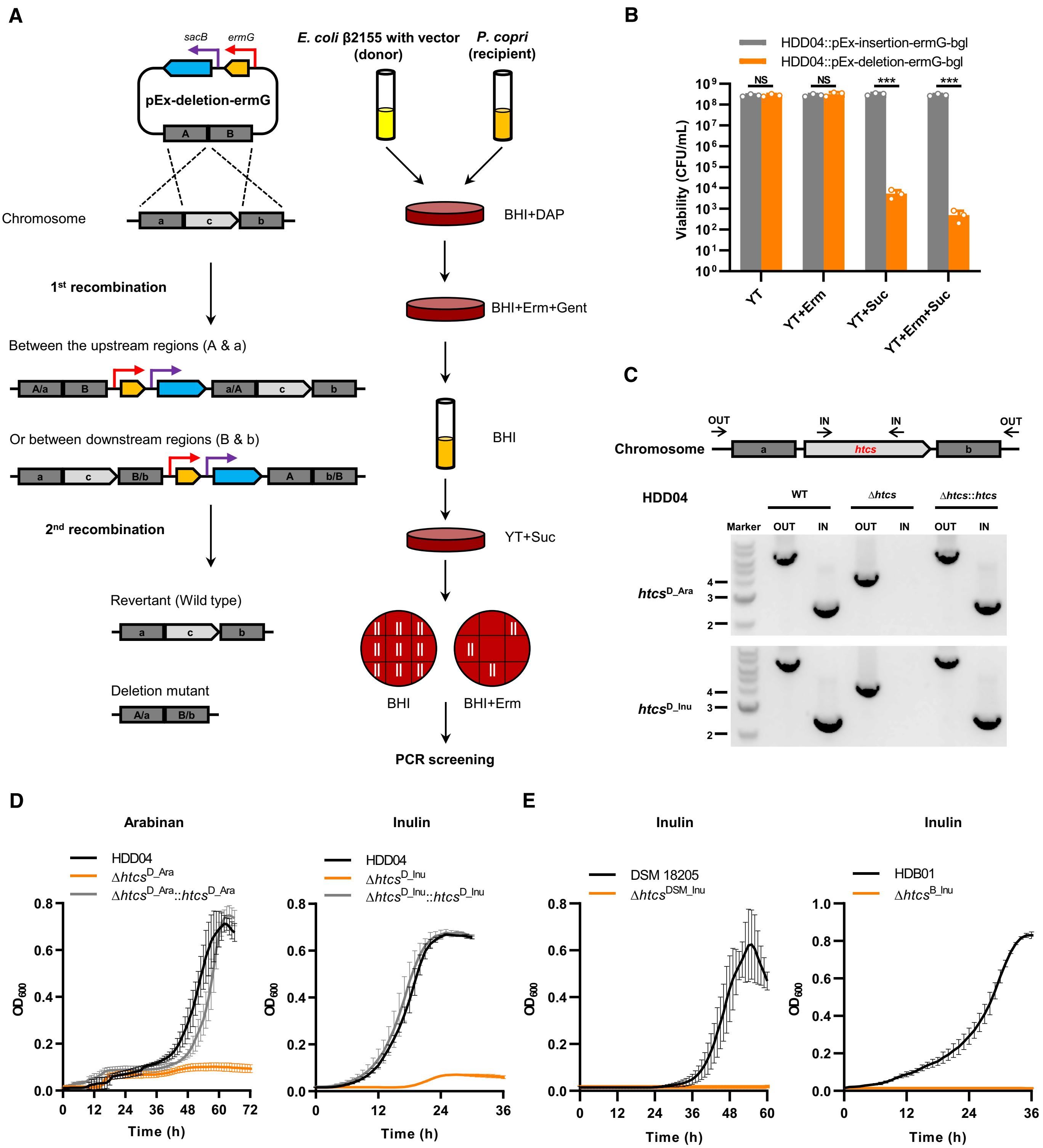

Figure 3.

Figure 3. Development of a conjugation-based gene deletion and complementation platform for *Prevotella copri* strains.

A Schematic illustration for allelic exchange using pEx-deletion-ermG.

B Viability of *P. copri* HDD04 with integration of pEx-insertion-ermG-bgl or pEx-deletion-ermG-bgl into the chromosome on YT agar plates with indicated supplements. Error bars represent the mean of three biological replicates ± SDs (***$P < 0.001$).

C, D Generation of *P. copri* HDD04 Δ*htcs*^{Ara} and Δ*htcs*^{Inu} and quantification of growth. Detection of deletion and complementation of two HTCS genes in *P. copri* HDD04 by PCR using "OUT" and "IN" primer pairs (C) and growth of indicated strains in arabinan or inulin (D), respectively.

E Growth of the wild-type and *htcs*^{Inu} mutant strains of DSM 18205 and HDB01, respectively.

Data information: In (B, D, and E), the data represent the means of three biological replicates ± SDs.

YT + Erm and YT + Suc plates to evaluate whether the Suc^{R} phenotype of these "escapers" was due to genetic mutations. Unexpectedly, they all showed Erm^{R} but Suc^{S} phenotypes. We further sequenced the *sacB* gene and its promoter sequences in 10 random-selected clones, yet none of them had mutations. This suggested that the Suc^{R} phenotype of these escapers were attributable to phenotypic but not genetic causes. Similar level of selectivity in the *sacB*-sucrose-based system was recapitulated in other five *P. copri* strains (Appendix Fig S1D and E, and Fig EV2D). Taken together, these results demonstrate the utility of the *sacB*-based counterselection system for targeted allelic exchange in *P. copri*.

Genetic deletion and complementation demonstrate essential function of HTCS genes on degrading plant polysaccharides

As a proof of concept, we chose two HTCS genes, *htcs*$^{D\_Ara}$ and *htcs*$^{D\_Inu}$ as our targets for genetic deletion and complementation. A schematic for describing the allelic exchange methodology during gene-editing process including plasmid construction, allelic exchange, and mutant selection is presented in Fig 3A. In brief, we (i) cloned up- and down-stream regions of the target gene into the pEx-deletion-ermG plasmid and transferred the plasmid into *E. coli* β2155; (ii) performed conjugation between *P. copri* and *E. coli* carrying the plasmids followed by selection of plasmid integrants (1st recombination); (iii) passaged the Erm^{R} clones without any selection, permitting the spontaneous allelic exchange (2nd recombination); (iv) carried out the selection of bacteria that had lost the plasmid (revertant and deletion mutant); (v) validated the Erm^{S} phenotype of selected clones and screened for the clones with gene deletion mutations. Of note, PCR screening and Sanger sequencing of Erm^{S} clones obtained after the counterselection step revealed that 37.5–56.3% of clones are confirmed deletion mutants (Dataset EV4) with all remaining clones being revertants, showing the precise performance of our targeting system. Following these procedures, *htcs*$^{D\_Ara}$ and *htcs*$^{D\_Inu}$ were successfully deleted in HDD04, generating Δ*htcs*$^{D\_Ara}$, and Δ*htcs*$^{D\_Inu}$ gene deletion strains accordingly (Fig 3C). We subsequently complemented the deletion mutants with the corresponding HTCS genes, respectively, through a similar genetic procedure except that we cloned the target gene and its flanking regions into pEx-deletion-ermG (Fig 3C). In line with our previous findings in Fig 2D, the HTCS-deficient mutants failed to grow on their previously identified substrates, while complementation of HTCS genes in the mutant strains restored the growth to the levels of the wild-type HDD04 strain (Fig 3D).

To demonstrate that the allelic exchange system can be also applied to *P. copri* strains with relatively lower yields of transconjugants, individual deletion of homologous *htcs*$^{D\_Inu}$ genes were performed in the DSM 18205 (*htcs*$^{DSM\_Inu}$, DSM18205_02724) and HDB01 (*htcs*$^{B\_Inu}$, HDB01_02906) strains (Fig EV2E). Similarly, deletion of *htcs*$^{D\_Inu}$ displayed dramatic grow defects in both DSM 18205 and HDB01, respectively (Fig 3E).

In conclusion, these results demonstrate the utility of our novel allelic exchange system in the type strain and other *P. copri* isolates for establishing a causal relationship between genotypes and phenotypes, as best exemplified by specific HTCS genes and the growth phenotypes on arabinan and inulin.

Plant-derived polysaccharides induced the transcription of distinct PUL-associated genes *in vitro* and *in vivo*

To determine whether polysaccharides induce the expression of specific PULs associated with the identified HTCS genes or rather broader changes in multiple PULs, we performed transcriptome profiling of *P. copri* HDD04 cultures grown in MM supplemented with either glucose or one of four plant polysaccharides as the sole carbohydrate (Dataset EV5). As we expected, the *susC/D*-like elements in PUL14, 21, 24, and 26 exhibited the largest upregulation upon their respective polysaccharide substrates compared to MM + Glucose (Fig 4A), further confirming our genetic characterization (Figs 2 and 3). In line with the previous reports, these PULs also encode the CAZymes that likely primarily targeted the substrates, and were presumably activated via the cognate HTCS regulators in a feedback loop. We therefore defined these four PULs as PUL$^{D\_Ara}$, PUL$^{D\_PecGal}$, PUL$^{D\_AraXyl}$, and PUL$^{D\_Inu}$ (Fig 4B). Using average fold change of *susC/D*-like genes in each PUL as reference, PUL$^{D\_Inu}$ showed a relatively low (3.3-fold) upregulation in response to inulin, whereas PUL$^{D\_Ara}$, PUL$^{D\_PecGal}$, and PUL$^{D\_AraXyl}$ displayed much higher induction levels (PUL$^{D\_Ara}$: 266.1-fold; PUL$^{D\_PecGal}$: 983.7-fold; PUL$^{D\_AraXyl}$: 159.3-fold; Fig 4A). It is worth noting that the disruption of HTCS gene in PUL24 resulted in the growth deficiency on a simple xylan, i.e., wheat arabinoxylan, where the xylan backbone is the main component with arabinose decorations, but did not affect the growth pattern on the corn xylan, which has extensive side residues forming heterogeneous and polydispersed glycans (Figs 2C and 4A). These data more precisely defined the specificity of PUL24 for arabinoxylan utilization compared to the previous prediction using bioinformatic analysis and growth assays (Fig EV3A) (Fehlner-Peach *et al*, 2019). Interestingly, genome-wide only *susC/D*-like genes from PUL$^{D\_Inu}$ were significantly upregulated, indicating an extremely specific system for processing inulin by *P. copri*. Yet, multiple *susC/D*-like elements were induced (> 10-fold change) by the other three plant-derived polysaccharides (Fig 4A). For instance, *susC/D*-like pairs in PUL$^{D\_PecGal}$ (PUL24, 23.9-fold) were also greatly expressed when cells were exposed to arabinan. The transcriptional response of *susC* homologs in four identified PULs to each tested polysaccharide were further validated

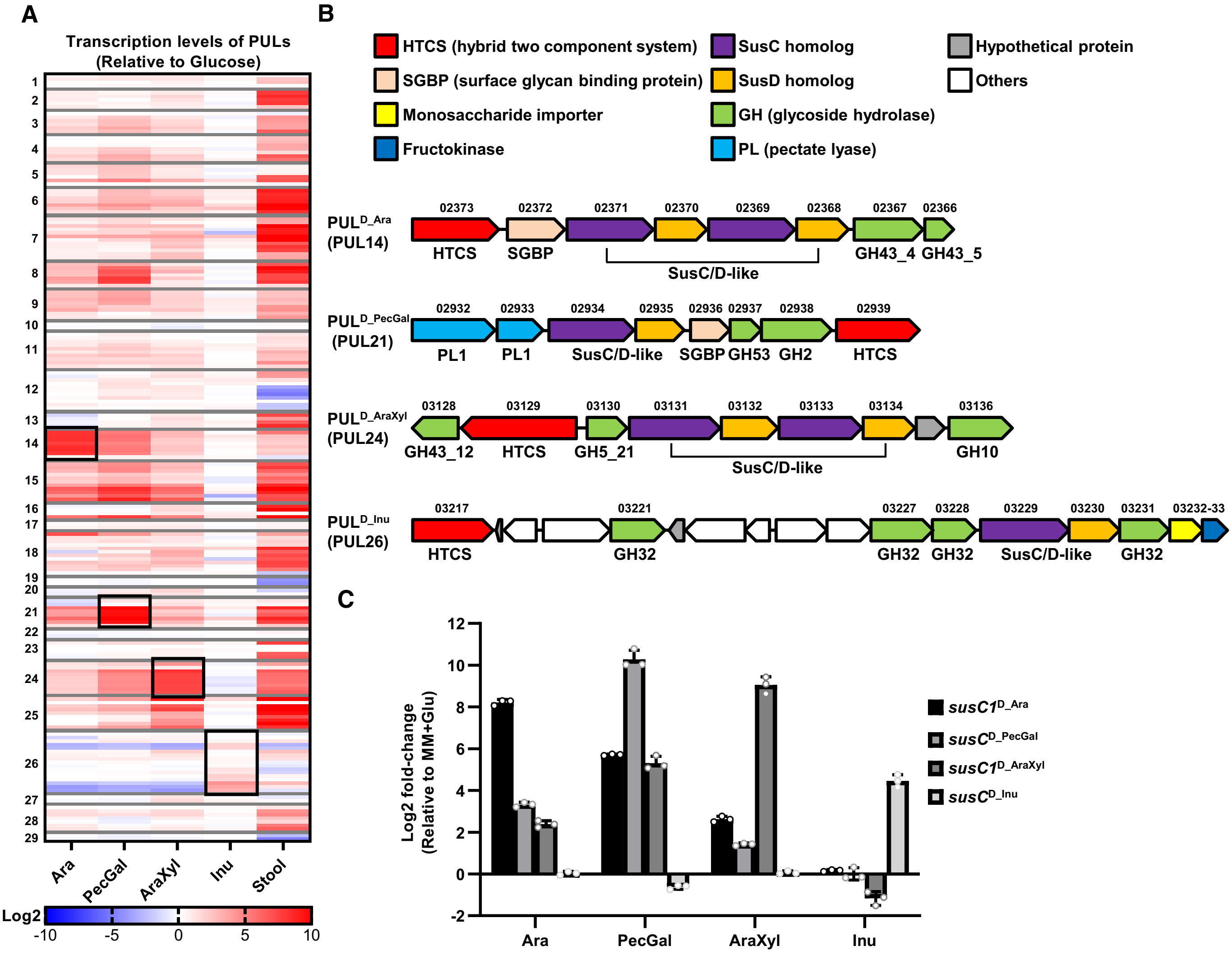

Figure 4. Transcriptional adaptation of *Prevotella copri* HDD04 to distinct plant polysaccharides and in the human gut.

A Heatmap showing the induction of *susC* and *susD* homologs in each predicted PUL from *P. copri* HDD04 in MM supplemented with indicated polysaccharides ($n = 3$) or in a fecal sample ($n = 1$). Average gene expression of every *susC/D* element (two or four genes) was calculated and normalized to expression in glucose. The heatmap shows the average log2 fold change of *susC/D* pairs within the predicted PULs. PUL14 (PUL$^{D\text{–}Ara}$), PUL21 (PUL$^{D\text{–}PecGal}$), PUL24 (PUL$^{D\text{–}AraXyl}$), and PUL26 (PUL$^{D\text{–}Inu}$) are highlighted by black frames.

B Genetic architectures of PUL$^{D\text{–}Ara}$, PUL$^{D\text{–}PecGal}$, PUL $^{D\text{–}AraXyl}$, and PUL$^{D\text{–}Inu}$, in *P. copri* HDD04. Genes encoding proteins with predicted functions are displayed by arrows with different colors. Numbers above the gene represent the locus tags based on the sequenced genome of *P. copri* HDD04. GH families that genes encode are shown based on the bioformatic analysis as described in Dataset EV3 and Materials and Methods.

C *In vitro* transcriptional response of targeted *susC* homologs in MM+indicated polysaccharide in comparison with MM+Glucose reference.

by real-time quantitative PCR (RT–qPCR; Fig 4C). The upregulation of multiple PULs during growth on pectic glycans, e.g., pectic galactan, have been also observed in *Bacteroides* spp. due to the complexity of these polysaccharides (Martens *et al*, 2011; Luis *et al*, 2018). For instance, PUL$^{D\text{–}PecGal}$ was likely induced by the degraded components of pectin fragment attached to the main-chain arabinofuranosyl residues of sugar beet arabinan (Fig EV3B). Surprisingly, PUL15 (HDD04_02377-87) possesses genes encoding GH13 and GH97, which were previously demonstrated to degrade various glucans (Koropatkin *et al*, 2008; Cerqueira *et al*, 2020), was commonly upregulated by arabinan, pectic galactan, and arabinoxylan. Besides the regulation of PUL-associated genes, there were also genes encoding polysaccharide catabolism enzymes that displayed >10-fold upregulation by the polysaccharides, e.g. one putative extracellular exo-alpha-L-arabinofuranosidase precursor gene in MM+Arabinan (HDD004_02362, Dataset EV5). Similarly, a putative gene operon was strongly upregulated in response to arabinan and arabinoxylan, which has a high similarity to the arabinose utilization system in *B. thetaiotaomicron* (Schwalm *et al*, 2016) (Fig EV3C).

To identify whether *P. copri* actively utilizes these PULs *in vivo*, a metatranscriptome analysis was performed from a stool sample collected from the donor, from which HDD04 was isolated. Strikingly, except PUL$^{D\text{–}Inu}$ that displayed a slightly increased transcription compared to MM+Glucose (average of fold-change *susC/D*-like genes: 1.2-fold), the *susC/D* homologs in the other three identified PULs (PUL$^{D\text{–}Ara}$: 4.2-fold; PUL$^{D\text{–}PecGal}$: 215-fold; PUL$^{D\text{–}AraXyl}$: 44.6-fold) were actively expressed. Moreover, 16 other PULs displayed upregulation from 2.36-fold (PUL1) to more than 8,000-fold change (PUL25), which suggests additional substrates from the human diet can be targeted by various PULs in *P. copri* (Fig 4A). Collectively, these analyses illustrate that *P. copri* carries out an efficient and diverse polysaccharide processing by orchestrating its associated gene expression profile *in vitro* and *in vivo*. Further functional gene studies will be required to understand which polysaccharides can be utilized *in vivo* by *P. copri*.

PUL$^{D\text{–}Ara}$, PUL$^{D\text{–}PecGal}$, PUL$^{D\text{–}AraXyl}$, and PUL$^{D\text{–}Inu}$ are conserved among genetically diverse *P. copri* strains

Recent studies reported that *P. copri* isolates exhibited extensive genomic and phenotypic variations (De Filippis *et al*, 2019; Fehlner-Peach *et al*, 2019; Tett *et al*, 2019). To examine whether utilization of arabinan, pectic galactan, arabinoxylan, and inulin is a common capacity based on the genetic content of *P. copri* species, we performed a comparative genomic analysis to identify corresponding PULs in strains from our *P. copri* strain collection. Notably, PULs carrying homologous HTCS/SusC genes compared to HDD04 were found in each of the *P. copri* strains (Fig 5 and Dataset EV3; Fig EV4). The gene organization and content of these PULs varied from conserved, e.g., PUL^{AraXyl}, to variable, e.g., PUL^{Inu} (Fig EV4). Hence, we next determined the growth in MM supplemented with arabinan, pectic galactan, arabinoxylan, or inulin. Most *P. copri* strains grew on the tested polysaccharides with the exception of HDA03 that could not grow on pectic galactan and arabinoxylan (Fig EV4). Strikingly, genetic evidence potentially explaining the inability to use specific polysaccharides could be easily identified (Fig EV4). Specifically, natural mutations in the HTCS genes of PUL^{AraXyl}, i.e., a truncation of HTCS^{AraXyl}, and of PUL^{PecGal}, i.e., deletion, are likely responsible for the “no growth” phenotypes of HDA03 on these polysaccharides. Additionally, the cognate first *susC* gene of PUL^{AraXyl} as well as the SGBP gene of PUL^{PecGal} were truncated into two segments in HDA03, which could further contribute to the inability to utilize these polysaccharides.

Notably, the PUL^{PecGal} in HDE04 and HDD12 and PUL^{Inu} in HDD05 and HDD12 do not contain *susC/D*-like elements, suggesting potential functional complementation between various PULs. These cases of “incomplete” PULs highlight the limitations of using *susC/D* homologs as genetic markers to predict the growth phenotypes on specific polysaccharides substrates.

Two types of arabinan processing PULs in *P. copri* display clade-specific distribution and diet-dependent expansion in the human gut microbiome

While functionally all tested strains were able to use arabinan, we found that the arabinan processing PULs genetically displayed two distinct structures among the 12 strains (Fig 5). Specifically, two

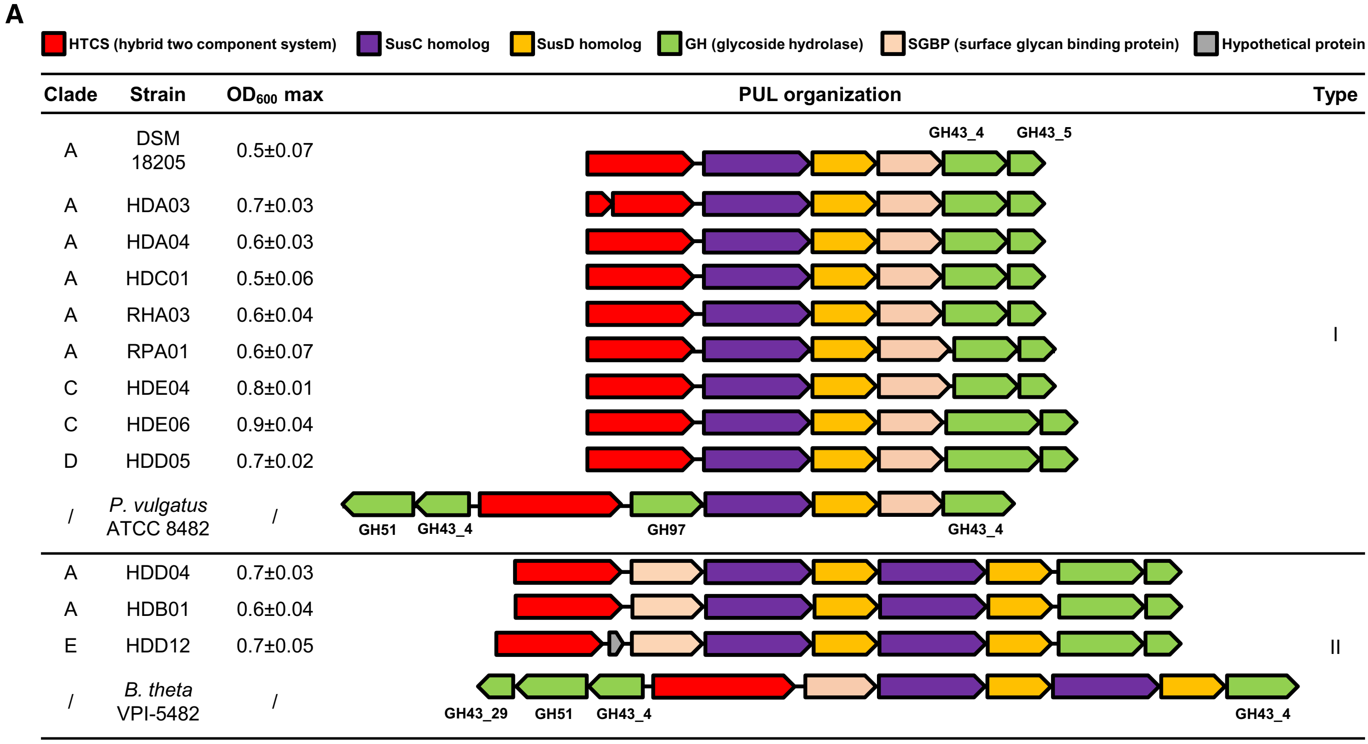

Figure 5. Gene organizations of two types of arabinan processing PULs in multiple *Prevotella copri* and two *Bacteroides* type strains with the capacities of growing on arabinan

strains from clade A and the single strain from clade E feature an almost identical gene organization with a SGBP-like gene in front of two pairs of *susC/D*-like genes, whereas the remaining strains from clade A, as well as the strains from clade C and D encode a single pair of *susC/D*-like genes followed by a SGBP-like gene. In the following, we refer to two types of PULs as type-I (single copy) and type-II (tandem repeat), and the two *susC/D*-like pairs in type-II PUL^{Ara} as *susC1*, *susD1*, *susC2*, and *susD2*, respectively. Of note, similar PUL^{Ara} types have been noticed in *Phocaeicola vulgatus* (formerly *Bacteroides vulgatus*) with type I and *B. thetaiotaomicron* with type II (Fig 5) (Martens *et al*, 2011; Lynch & Sonnenburg, 2012; Patnode *et al*, 2019).

Phylogenetic analysis of protein sequences encoded by HTCS genes from the 12 *P. copri* strains as well as *Ph. vulgatus* ATCC 8482 and *B. thetaiotaomicron* VPI-5482 shows a clade-driven evolutionary pattern, which closely resembled that of genome-based tree of the *P. copri* complex (Figs 1A and 6A). The inability of the mutant strain DSM18205 *htcs*$^{DSM\_Ara}$ to grow on arabinan validates that the HTCSs of two arabinan utilizing systems are functionally conserved in the two types of PUL^{Ara} (Fig 6B). In contrast, the *P. copri*-derived SusC and SusD homologs form three distinct evolutionary branches in the tree that differ from the proteins in *B. thetaiotaomicron* and *Ph. vulgatus,* but still show a relative similarity of these proteins corresponding to the PUL^{Ara} types, respectively (Figs 6A and EV5A). Similar to the *susC/D*-like genes, the SGBP-like proteins are clustered by PUL^{Ara} type rather than *P. copri* clade (Fig EV5B). We next performed functional studies complementing the phylogenetic analysis. Because the *susC1* gene but not the *susC2* gene is required in *B. thetaiotaomicron* for growth on arabinan as described previously (Luis *et al*, 2018), the genes encoding *susC* in type-I, and *susC1* and *susC2* in type-II system of *P. copri* were individually in-frame deleted to explore their necessity in *P. copri* (Fig 6B). In agreement with previous observations in *B. thetaiotaomicron*, only *susC1* but not *susC2* is essential for type-II PUL^{Ara} carrier HDD04 (Fig 6B). Moreover, deletion of *susC* in type-I PUL^{Ara} abolished the growth capacity of *P. copri* DSM 18205 on arabinan (Fig 6B). These results indicated that *P. copri* strains encode highly similar sensory/regulatory systems for sensing arabinan-derived ligands and transcriptional activation of PUL^{Ara}, but that PUL^{Ara} encodes distinct modules, i.e., *susC-susD-SGBP* in type I and *SGBP-susC1-susD1-susC2-susD2* in type II, for carbohydrate binding and importing.

To gain a broader understanding of the prevalence of PUL^{Ara} types in *P. copri* as well as related *Prevotella* spp., *Bacteroides* spp., and *Phocaeicola* spp., PUL^{Ara} were predicted from 1,602 non-redundant genomes retrieved from the NCBI genome database (n = 1,504) and from a recent comprehensive metagenomic *P. copri* survey (n = 98) (Tett *et al*, 2019). Together with our strains (n = 12), we identified that 499 out of 1,614 genomes encode either type-I or type-II PUL^{Ara}, suggesting that the arabinan utilization potential is frequently found in these genera but not ubiquitous (Fig 6C and Dataset EV6). In agreement with the results from the analysis of our limited strain collection, type-I PUL^{Ara} is present in *P. copri* clades A (55.3%), C (72.7%) and D (75%), while the type-II is encoded by *P. copri* strains from the clade A (36.8%) and the single strain from the clade E (HDD12; Fig 6C). Notably, none of the two types as well as the arabinose utilization operon we identified was found in any of the 53 screened genomes of clade B strains (Fig 6C), which is consistent with previous reports that this clade lacks the genes and capacity for arabinose and arabinan utilization (Tett *et al*, 2019). The two types of PUL^{Ara} are also widespread in other members of the *Prevotella* spp., *Bacteroides* spp., and *Phocaeicola* spp. (Fig 6C). Notably, our extended analysis shows that while some species displayed either a type I- or type II-dominated distribution, e.g., *Ph. vulgatus* (n = 90 genomes, type I: 90%, type II: 0%), other species such as *Phocaeicola plebeius* (n = 20 genomes, type I: 20%, type II: 50%) and *B. thetaiotaomicron* (n = 42 genomes, type I: 7.14%, type II: 88.1%) displayed both types of arabinan utilization systems similar to *P. copri* species (Fig 6C). As the evolutionary analysis of *Bacteroides* spp. and *Phocaeicola* spp. is limited, these dominations could also result from the clade-specific distribution as *P. copri*. Interestingly, only 1 out of 55 *Bacteroides ovatus* genomes carry type-I PUL^{Ara}, which not only suggests the occurrence of horizontal gene transfer of type-I PUL^{Ara} between *B. ovatus* and other type-I PUL^{Ara} carriers (Fig 6C), but also provides an explanation for the apparent lack of arabinan utilization in most *B. ovatus* strains (Martens *et al*, 2011).

The increased relative abundance of *P. copri* in the gut microbiota has been associated with fiber-rich diets (De Filippo *et al*, 2010; Wu *et al*, 2011; Ruengsomwong *et al*, 2016; Fragiadakis *et al*, 2019). As the *in vitro* results described above suggest a potential contribution of the PUL^{Ara} for fitness advantage within the ecosystem, we investigate whether *P. copri* encoding different types of PUL^{Ara} displays a diet-modulated abundance in the human gut. Therefore, we performed a specialized analysis of a publicly available dataset from one recent study comparing the differences in the gut microbiota of individuals consuming omnivore, vegetarian, and vegan diet (De Filippis *et al*, 2019). In a previous study, we identified five distinct *P. copri* metagenome-assembled genomes (MAGs) from four clades in individuals of that cohort (Fig 6D). MAG610 (clade C) encodes a type-I PUL^{Ara}, whereas two MAGs 609 and 611 (clade A and C) carry the type-II counterpart. Two MAGs 612 and

Figure 6. Genetic and phylogenetic analyses of PUL^{Ara} types in members of the genera *Prevotella*, *Phocaicola* and *Bacteroides*.

▶

A Phylogenetic trees of HTCS and SusC-like proteins encoded by two types of arabinan processing PULs in diverse *Prevotella copri* strains, *Bacteroides thetaiotaomicron* VPI-5482, and *Phocaeicola vulgatus* ATCC 8482. The isolates from distinct *P. copri* clades are indicated by dots in different colors. The proteins from type-II PUL^{Ara} are highlighted in bold.

B Growth of *P. copri* DSM 18205 and HDD04 wild-type strains and indicated mutants in MM+Arabinan.

C Distribution of two types of arabinan processing PULs in the members of genera *Prevotella*, *Bacteroides* and *Phocaeicola*. Total number of genomes for each clade or species group that were analyzed is indicated above the bars.

D The association between the two types of PUL^{Ara} in *P. copri* and host dietary preference. The relative abundance of identified *P. copri* MAGs for each individual was grouped based on the presence or absence of two types of PUL^{Ara} in each dietary habit. Asterisks indicate Wilcoxon U-test significant differences (*$P < 0.05$; NS, $P > 0.05$, not statically significant).

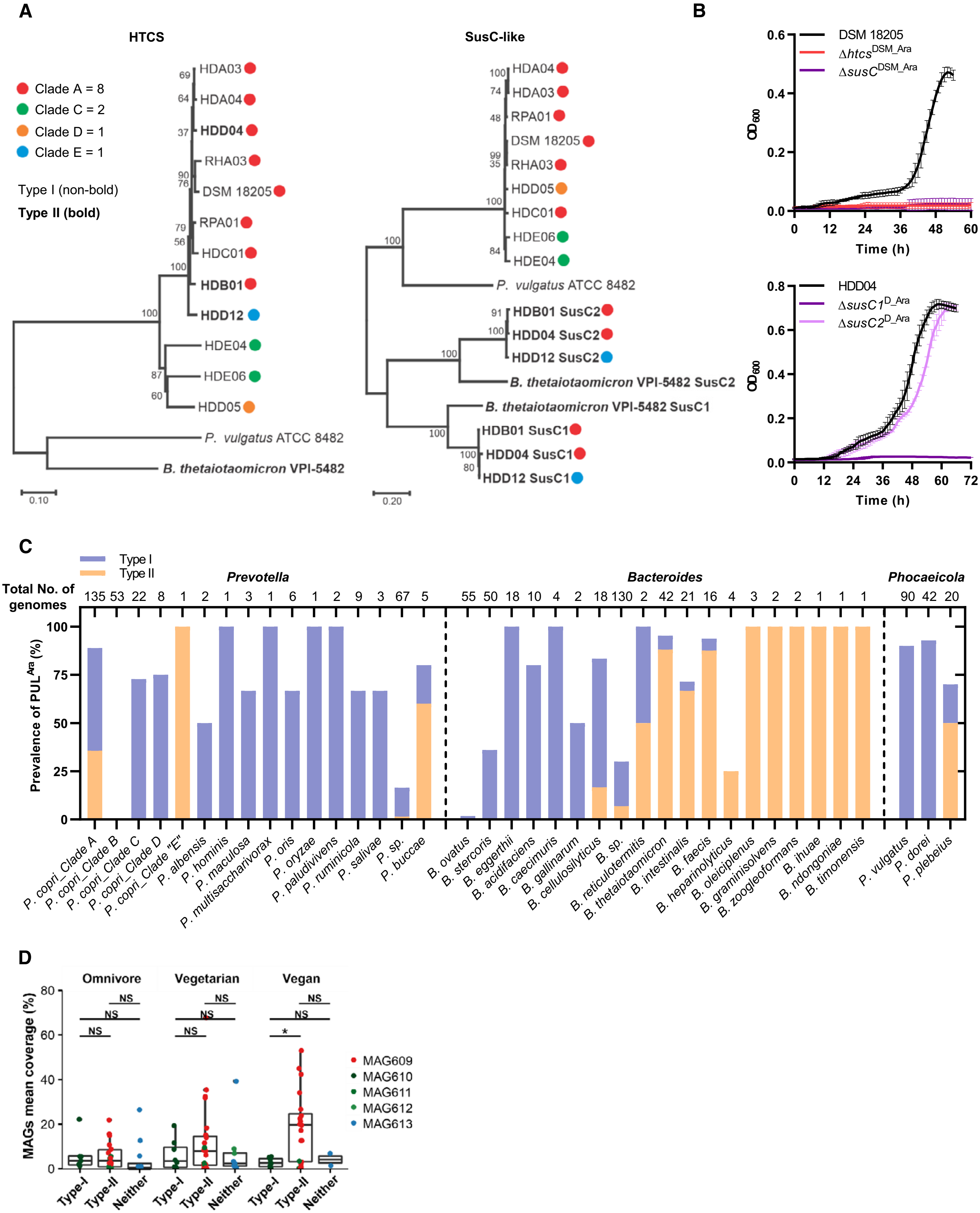
A
HTCS
SusC-like
● Clade A = 8
● Clade C = 2
● Clade D = 1
● Clade E = 1
Type I (non-bold)
Type II (bold)
HDA03
HDA04
HDD04
RHA03
DSM 18205
RPA01
HDC01
HDB01
HDD12
HDE04
HDE06
HDD05
P. vulgatus ATCC 8482
B. thetaiotaomicron VPI-5482
69
64
37
90
76
79
56
100
100
87
60
0.10
HDA04
HDA03
RPA01
DSM 18205
RHA03
HDD05
HDC01
HDE06
HDE04
P. vulgatus ATCC 8482
HDB01 SusC2
HDD04 SusC2
HDD12 SusC2
B. thetaiotaomicron VPI-5482 SusC2
B. thetaiotaomicron VPI-5482 SusC1
HDB01 SusC1
HDD04 SusC1
HDD12 SusC1
100
74
48
99
35
100
84
100
91
100
100
100
100
80
0.20
B
DSM 18205
Δhtcs$^{DSM\_Ara}$
ΔsusC$^{DSM\_Ara}$
OD$_{600}$
0.6
0.4
0.2
0.0
0
12
24
36
48
60
Time (h)
HDD04
ΔsusC1$^{D\_Ara}$
ΔsusC2$^{D\_Ara}$
OD$_{600}$
0.8
0.6
0.4
0.2
0.0
0
12
24
36
48
60
72
Time (h)
C
Type I
Type II
Prevotella
Bacteroides
Phocaeicola
Total No. of genomes
135 53 22 8 1 2 1 3 1 6 1 2 9 3 67 5
55 50 18 10 4 2 18 130 2 42 21 16 4 3 2 2 1 1 1
90 42 20
Prevalence of PUL^{Ara} (%)
100
75
50
25
0
P. copri_Clade A
P. copri_Clade B
P. copri_Clade C
P. copri_Clade D
P. copri_Clade "E"
P. albensis
P. hominis
P. maculosa
P. multisaccharivorax
P. oris
P. oryzae
P. paludivivens
P. ruminicola
P. salivae
P. sp.
P. buccae
B. ovatus
B. stercoris
B. eggerthii
B. acidifaciens
B. caecimuris
B. gallinarum
B. cellulosilyticus
B. sp.
B. reticulotermitis
B. thetaiotaomicron
B. intestinalis
B. faecis
B. heparinolyticus
B. oleiciplenus
B. graminisolvens
B. zoogleoformans
B. ihuae
B. ndongoniae
B. timonensis
P. vulgatus
P. dorei
P. plebeius
D
Omnivore
Vegetarian
Vegan
MAGs mean coverage (%)
80
60
40
20
0
NS
NS
NS
NS
NS
NS
NS
NS
*
Type-I
Type-II
Neither
Type-I
Type-II
Neither
Type-I
Type-II
Neither
● MAG609
● MAG610
● MAG611
● MAG612
● MAG613

Figure 6.

613 (clade C and B) contain neither of the PULs (Fig 6D). The relative abundance of each MAG was then calculated per individual and grouped based on the presence of type-I or type-II PUL^{Ara} in the individuals with three distinct dietary preferences (Fig 6D). While there was no significant difference based on PUL^{Ara} presence and type in omnivore and vegetarian diets, remarkably, in vegans, PUL^{Ara} type-II positive MAGs showed higher abundance than those with type-I PUL^{Ara} (Fig 6D). A possible explanation for the expansion of *P. copri* strains encoding type-II PUL^{Ara} in vegans could be distinct capabilities of the two PUL^{Ara} to degrade a so far unknown subtype of polysaccharide. Moreover, in the case of type-II PUL^{Ara}, it is possible that the two distinct SusC/D pairs not only form separate homodimer but also form heterodimers, which may enable type-II PUL^{Ara} to access more types of polysaccharides degradation intermediates. Further genetic and functional characterizations will be required to understand the differential precise nature of two PUL^{Ara} systems. Yet, taken together our comprehensive analyses illustrate the importance of arabinan utilization systems for *P. copri* fitness *in vitro* and *in vivo*.

Discussion

Studying the biology of many human commensals is hindered by diverse obstacles, e.g., challenging cultivation, extensive strain-level variation, low number of publicly available strains, and the lack of genetic tools (Ley, 2016; De Filippis *et al*, 2019; Fehlner-Peach *et al*, 2019; Tett *et al*, 2019). In this study, we developed a genetic toolkit that allows versatile genetic manipulations for a wide range of *P. copri* strains, and applied our genetic platform coupled with genomic, transcriptomic, and phenotypic approaches to provide insights into the genetic basis of polysaccharide utilization of this prevalent gut bacterium. Particularly, the recognition that *Prevotella* spp. including *P. copri* are a dominant part of the "non-westernized" microbiome as well as their unexplained antagonism with *Bacteroides* spp. has elevated the interest in the members of this genus (Arumugam *et al*, 2011; Ley, 2016; Johnson *et al*, 2017; Costea *et al*, 2018). While a series of genetic tools have been created for the genus *Bacteroides* (Koropatkin *et al*, 2008; Goodman *et al*, 2011; Mimee *et al*, 2015; Lim *et al*, 2017; García-Bayona & Comstock, 2019; Bencivenga-Barry *et al*, 2020), none have been reported for *P. copri* despite its diverse associations to human diseases. Moreover, few functional genetic studies reported on targeting genes in animal-derived *Prevotella*, i.e., *Prevotella ruminicola* (Shoemaker *et al*, 1991; Gardner *et al*, 1996; Ogata *et al*, 1999) and *Prevotella bryantii* (Accetto *et al*, 2005; Accetto & Avguštin, 2007). Together, this suggests the existence of limitations preventing the simple transfer of established genetic tools to *P. copri*.

Here, by redesigning and optimizing the key genetic elements in the conjugative plasmids and experimental procedures, we developed an anaerobic conjugation-based system, which overcomes the genetic intractability of diverse *P. copri* strains. For instance, the optimization of promoter strength for the selection marker resulted in a 306.6-fold increase in conjugation outcomes. Another key factor was the evaluation of different donor and recipient strains improving conjugation outcome by 83.7-fold and up to 14,160-fold, respectively. Of note, naturally occurring antibiotic resistances can complicate gene targeting, but the combination of plasmids carrying either *ermG* or *tetQ* antibiotic markers allowed the generation of insertion mutants for 11 out of 12 *P. copri* strains in our strain collection representing multiple distinct clades. At least one of the strains may even represent a distinct species, i.e., the HDD12 strain was annotated as *Prevotella hominis*, further supporting the utility of our approach to diverse *Prevotella* strains.

In a proof of concept of our approach, we focused on one strain, i.e., HDD04, with high conjugation capacity and robust growth in chemically defined medium, targeting 10 HTCS regulators for controlling PUL expression, and identified four HTCS genes with essential functions in utilizing plant polysaccharides. Next, we adapted the *sacB*-sucrose system to *P. copri* that enabled the selection of mutants after allelic exchange with a 3.1×10^{-7} to 4.9×10^{-6} false positive rate. This efficacy is similar to the that obtained for a recent system used for allelic exchange in *Bacteroides* spp. (Bencivenga-Barry *et al*, 2020). Of note, compared to other counterselection systems, such as the genes encoding 30S ribosomal protein S12 (*rpsL*) for Proteobacteria and thymidine kinase (*tdk*) for *Bacteroides* spp. (Dean, 1981; Reyrat *et al*, 1998; Koropatkin *et al*, 2008), the *sacB*-sucrose system does not require any genetic modification of the recipient strain in advance. The utility of the *sacB*-sucrose system was demonstrated by deletion of HTCS genes in three strains, including ones with relatively lower conjugation capacity. Despite these advantages, the allelic exchange system for *P. copri* does have limitations; e.g., it requires multiple selection steps and the ability to tolerate the osmotic pressure of high sucrose concentrations. Other counterselection markers, such as inducible antibacterial effectors (Bte1 and Bfe1) utilized in *Bacteroides* species (García-Bayona & Comstock, 2019; Bencivenga-Barry *et al*, 2020), can be envisioned, yet it has been noticed that *P. copri* DSM 18205 was not affected by the Bfe1 effector (Chatzidaki-Livanis *et al*, 2016). Another limitation relates to complementation of gene deletions, e.g., in *B. thetaiotaomicron*, genetic complementation is accomplished by integrating a plasmid carrying the complemented gene into the chromosome (Wang *et al*, 2000; Koropatkin *et al*, 2008). Specifically, the integration vector pNBU2 encodes a tyrosine integrase, which mediates sequence-specific recombination between the attN site of pNBU2 and one of two attBT sites located in the 3′ ends of the two tRNASer genes on the *B. thetaiotaomicron* chromosome. Because no identical or similar attachment DNA sequences were found in *P. copri*, we so far carried out the complementation by the same allelic exchange approach as used for gene deletion. Thus, development of a similar integration vector for *P. copri* or a *P. copri* parent strain carrying attachment sites for pNBU2 has the potential to simplify the process of genetic complementation.

The high prevalence and increased abundance of *P. copri* in the intestinal microbiota is frequently associated with consumption of fiber-rich diets, which has inspired the research for underlying genetic basis of polysaccharide utilization in *P. copri*. Combinations of comparative genomics and phenotypic assays have predicted the substrates for the PULs harboring well-defined CAZymes-coding genes (Fehlner-Peach *et al*, 2019). However, the direct contribution of specific PULs for polysaccharide substrates has not been documented for *P. copri*. Our genetic studies confirmed the previous bioinformatic prediction of arabinan as the substrate for the PUL14 homologs (Fehlner-Peach *et al*, 2019), Additionally, three new PUL/polysaccharide combinations were identified in *P. copri*. It is intriguing that PUL24, whose homologous gene cluster was predicted based on bioinformatic analysis as xylan-processing PUL, was identified in this study to be specifically essential for wheat arabinoxylan utilization, but non-essential for complex xylan from corn, suggesting the contribution of other CAZymes out of PUL24 to deconstruction of xylans with complicated structures. Two distinct PULs, PUL-XylL and PUL-XylS, in *B. ovatus* have been previously demonstrated to be induced by different types of xylans, but confer to growth on the simple and complex xylans, respectively. In *P. copri* HDD04, the CAZymes present in PUL25 (e.g., GH43_7, GH31, and GH3) as well as the ones that were not predicted as PUL components (e.g., HDD04_02869-70 around PUL20) were also co-upregulated with PUL24 by arabinoxylan to a lesser degree. These CAZymes are likely responsible for degrading other types of xylans besides wheat arabinoxylan. Collectively, genetic approaches coupled with phenotypic and transcriptome analyses present a framework for a more accurate characterization of PULs. Notably, the gene organization and content of PULs varied between *P. copri* strains from conserved to variable. For instance, the strictly conserved synteny of the arabinan and arabinoxylan processing PULs suggests that they have been under the positive selection pressure, as discussed below. In contrast, the gene content of the PULs for pectic galactan and inulin are relatively variable, suggesting the non-essentiality of certain genes for supporting growth of *P. copri* on these two carbon sources, including even *susC/D*-like genes. These observations support the model that *susC/D*-like element in some but not all PULs are essential for the uptake of polysaccharides and are therefore suboptimal markers for carbohydrate utilization potential. Finally, *P. copri* strain carry either one of two types of PUL$^{\mathrm{Ara}}$, i.e., type-I with a single *susC/D* pair or type-II PUL$^{\mathrm{Ara}}$ with tandem-repeat *susC/D*, which appears to be shared phenomenon in the *Bacteroides*, *Phocaeicola*, and *Prevotella* genera. In *P. copri*, the distribution of the two distinct PUL types showed largely clade-specific features, i.e., clade A encoded both types of PUL$^{\mathrm{Ara}}$, clades C and D only encoded type I, and none of identified PUL$^{\mathrm{Ara}}$ was found in clade B (Fehlner-Peach *et al*, 2019; Tett *et al*, 2019). While the HTCS regulators independent of the PUL$^{\mathrm{Ara}}$ type shared high homology between members of the same clade, the *susC/D*-like and SGBP-like genes clustered by PUL type. Notably, in a limited explorative analysis PUL$^{\mathrm{Ara}}$ type-specific domination was observed in individuals consuming a vegan diet, suggesting the advantage of type-II over type-I in utilizing arabinan or potential other arabinose-based polysaccharides from dietary fibers in the human gut.

In summary, we have demonstrated the versatile capacities of the genetic toolbox for, firstly, generating a series of individual gene insertion mutants for phenotypic screening in parallel; secondly, enabling targeted gene deletion and complementation to establish causal relationship between genotypes and phenotypes; thirdly, determining the impact of homologous genes in distinct *P. copri* strains on specific polysaccharide utilization. The toolbox will enable the dissection of more sophisticated biological interactions of *P. copri* with the human hosts during health and disease, such as investigate associations of *P. copri* to host metabolism *in vivo* (Kovatcheva-Datchary *et al*, 2015; De Vadder *et al*, 2016; Pedersen *et al*, 2016). Importantly, the platform was designed using general principles highlighting key technical details that can be modified and applied to other *Prevotella* species and even prominent bacterial genera from humans and other habitats. Moreover, these principles can be utilized for further development of high-throughput genetic screening, such as transposon mutagenesis (Goodman *et al*, 2011) and CRISPRi (Peters *et al*, 2016), thereby advancing studies into systematically understanding the ecological and metabolic processes of microbiota and their impacts on host health and disease.

Materials and Methods

Preparation of culture media for *P. copri*

BHI + S liquid medium

The fetal bovine serum was heated at 56°C for 30 min to inactivate complement. 9.25 g of brain heart infusion (BHI) powder was dissolved in 225 ml double-distilled water (ddH$_2$O) in 500-ml glass bottle. The medium was supplemented with 10% fetal bovine serum (FBS), and placed on the hotplate stirrer with 250°C for 20 min. The heated medium was then cooled down to room temperature, supplemented with 1 μg/ml vitamin K3, and filter-sterilized using a filter unit (0.22 mm pore diameter).

Minimal medium

Minimal medium (MM) was prepared on the basis of a minimal media supporting the growth of various *Bacteroides* spp. strains (Martens *et al*, 2008). It contained 100 mM KH$_2$PO$_4$ (pH 7.2), 15 mM NaCl, 8.5 mM (NH4)$_2$SO$_4$, 100 μM MgCl$_2$, 1.4 μM FeSO$_4$·7H$_2$O, 50 μM CaCl$_2$, 1.9 μM hematin, 1 mg/l vitamin K$_3$, 5 μg/l vitamin B$_{12}$, 0.5 g/l cysteine. We modified this media to contain additionally 10 ml/l amino acid mix solution (250 mg each of L-alanine, L-arginine, L-asparagine, L-aspartic acid, L-cysteine, L-glutamic acid, L-glutamine, glycine, L-histidine, L-isoleucine, L-leucine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, l-tryptophan, L-tyrosine, L-valine, and 312 mg of L-lysine monohydrochloride into 1 l ddH$_2$O), 10 ml/l purine/pyrimidine solution (200 mg each of adenine, guanine, thymine, cytosine, and uracil into 1 l ddH$_2$O, pH 7.0), 10 ml/l ATCC Vitamin Mix, 10 ml/l ATCC Trace Mineral Mix, and 1 μl/l vitamin K1 solution were added. The commercial carbohydrates were prepared as described previously (Martens *et al*, 2011). Briefly, 10 g/l carbohydrate stock solutions (2× concentration) were sterilized by filtering (for glucose) or autoclaving at 121°C for 15 min (for complex polysaccharides, Dataset EV7). When needed, 10 g/l carbohydrate solutions were added into 2× MM at a volume ratio of 1: 1.

YT + S agar

Sucrose was dissolved in ddH$_2$O at 0.5 g/ml (50%) as a stock solution. 5 g yeast extract and 10 g tryptone were dissolved in 900 ml ddH$_2$O. The resulting medium (YT) was autoclaved, cooled down to 50°C, and supplemented with 5% sucrose (100 ml 50% sucrose stock solution) for counter-selection for gene deletion and complementation of *P. copri* strains. When necessary, the different volumes of sucrose solution were added in media.

Bacterial culture conditions

All strains, plasmids, and primers used are listed in Dataset EV1. *Escherichia coli* strains were grown aerobically at 37°C on Luria-Bertani (LB) media. *Escherichia coli* β2155 were specially cultured on LB supplemented with 0.3 mM 2,6-Diaminopimelic acid (DAP)

(Demarre *et al*, 2005). *Prevotella copri* was cultured in BHI+S liquid media, minimal media plus a carbon source, on BHI agar supplemented with 5% defibrinated horse blood or on YT agar supplemented with 5% defibrinated horse blood and 5% sucrose unless otherwise specified. Cultures were routinely grown and manipulated in an anaerobic chamber (Coy Laboratory Products) with an atmosphere of 20% CO_2, 10% H_2, and 70% N_2 at 37°C. When necessary, antibiotics were added to the medium as follows: 7 μg/ml vancomycin, 100 μg/ml ampicillin, 200 μg/ml gentamicin, 20 μg/ml erythromycin for selecting *P. copri* DSM 18205 derived plasmid integrants; 5 μg/ml for other *P. copri* strains, and 20 μg/ml tetracycline for RHA03 and HDE06; 2.5 μg/ml for HDD04, HDE04 and HDD05.

Isolation of P. copri from humans

Prevotella copri was isolated from fecal samples of *P. copri*-positive donors previously determined by 16S rRNA sequencing. Briefly, the fresh fecal samples were collected and further processed in an anaerobic chamber. A pea-sized fecal pellet was resuspended in 5 ml BHI + S and filtered through 70 μm cell strainer. We performed a serial 10-fold dilution of the flow-through, and streaked out the diluted samples with the dilution factors of 10^{-3} to 10^{-6} on BHI blood agar plates supplemented with vancomycin. The plates were then incubated anaerobically at 37°C for 48–72 h. Individual colonies were picked into BHI+S broth and the resulting cultures were screened by PCR for *P. copri*-positive cultures using *P. copri*-specific primers (P_copri_69F/P_copri_853R). The pure *P. copri* isolates were obtained by steaking out the *P. copri*-positive cultures above, and confirmed by Sanger sequencing using the 16S rRNA gene-specific primers as described previously (16S_27F/16S_1492R) (Miller *et al*, 2013). The fresh culture of *P. copri* was mixed with an equal volume of 50% glycerol in BHI medium in sealed glass vials as bacterial glycerol stocks, and cryopreserved at −80°C immediately.

DNA extraction from human feces and P. copri cultures

The DNA extraction from fecal samples of *P. copri*-positive donors or *P. copri* strains cultured in BHI+S broth (OD_{600} = 0.6) was performed using ZymoBIOMICS DNA Miniprep Kit based on the instruction manual. We measured the concentration of purified DNA samples by Qubit Fluorometer (Thermo Scientific), and analyzed by agarose gel electrophoresis, NanoDrop^{TM} 2000 (Thermo Scientific), and Bioanalyzer (Agilent Technologies).

Whole-genome sequencing, assembly, and annotation

The DNA library for genome sequencing of *P. copri* strains was performed using NEBNext® Ultra™ II FS DNA Library Prep Kit (New England Biolabs) for Illumina with parameters as followed: 500 ng input DNA and 5 min at 37°C for fragmentation; > 550-bp DNA fragments for size selection; primers from NEBNext Multiplex Oligos for Illumina Kit (New England Biolabs) for barcoding. The library was sequenced on the Illumina Miseq 2 × 250 bp The obtained reads were thus assembled with SPAdes version v3.10.0 using "careful" mode (Bankevich *et al*, 2012). Short contigs were then filtered by length and coverage (contigs > 500 bp and coverage > 5×). Gene prediction and annotation was performed using PROKKA version v1.13.3 (Seemann, 2014) with default parameters. The locus IDs of annotated genes in this study and the NCBI database are shown in Dataset EV8.

Phylogenetic analyses

Placement of *P. copri* complex. The phylogenomic analyses were conducted as previously described on the characterization of the *P. copri* Complex (Tett *et al*, 2019) using PhyloPhlAn3 (Asnicar *et al*, 2020) with reference set of *P. copri* strains (Tett *et al*, 2019) and the newly *P. copri* strains isolated in this study. The phylogenetic analysis in Fig 1A was built using the 400 universal marker genes of the PhyloPhlAn database using the parameters "--diversity low", and "--accurate" option. The configuration file (config_file.cfg) was set with the following tools and parameters:

Diamond version v0.9.9.110 (Buchfink *et al*, 2015) with "Blastx" for the nucleotide-based mapping, "Blastp" for the amino acid-based mapping, and "--more-sensitive --id 50 --max-hsps 35 -k 0" in both cases. MAFFT version v7.310 (Katoh & Standley, 2013), with "--localpair --maxiterate 1,000 --anysymbol --auto" options. trimAl version 1.2rev59 (Capella-Gutiérrez *et al*, 2009), with "-gappyout" option. IQ-TREE multicore version v1.6.9 (Nguyen *et al*, 2015), with "-nt AUTO -m LG" options. RAxML version 8.1.15 (Stamatakis, 2014), with "-p 1989 -m GTRCAT -t" options.

For Figs 6 and EV5, the amino acid sequences of HTCS, SusC-like, SusD-like, and SGBP-like proteins encoded by arabinan processing PULs (PUL^{Ara}) from *B. thetaiotaomicron* VPI-5482, *Ph. vulgatus* ATCC 8482 and 12 *P. copri* strains were used for the phylogenetic trees via the MEGA-X software, respectively. The evolutionary history was inferred using the Neighbor-Joining method (Saitou & Nei, 1987) with 1,000 bootstrap replicates. The two types of PUL^{Ara} extracted from all recovered assemblies from the genus *Prevotella* (n = 8), *Bacteroides* (n = 13) and *Phocaeicola* (n = 3) was calculated for the analysis of phylogenetic distributions, respectively.

Determination of P. copri sensitivity to oxygen

Bacteroides thetaiotaomicron VPI-5482, *P. copri* HDD04, HDB01, and DSM 18205 were grown in BHI+S broth anaerobically. The fresh bacterial cultures were divided into 1 ml aliquots into 2 ml tubes with the caps being open, respectively. These aliquots were aerobically incubated at 37°C. At four time points (0, 1, 2, 4 h), three aliquots from respective cultures were placed back to the anaerobic chamber and performed serial dilutions for counting CFUs.

Determination of P. copri sensitivity to antibiotics

The wells of a non-tissue culture flat bottom 96-well were loaded with 198 μl BHI+S media in the presence of 2-fold serial dilutions of erythromycin, tetracycline, chloramphenicol, spectinomycin, apramycin, and hygromycin ranging from 0.04 to 400 μg/ml. BHI+S media without antibiotics was loaded as a positive control. *Prevotella copri* strains were grown in BHI+S broth to an optical density (OD_{600}) of 0.5–0.7. 2 μl bacterial culture was then inoculated into each well (inoculation ratio of 1:100). Absorbance at OD_{600} of each well was measured at an interval of 1 h for 5 days using the microplate reader (BioTek). Assays were performed in triplicate. To ensure that the concentrations of erythromycin and tetracycline used for selecting transconjugants of different *P. copri* strains were sufficient for killing all the wild-type *P. copri* cells, 1 ml of fresh *P. copri* culture (10^8–10^9 bacterial cells) was plated on the BHI blood agar with respective concentration of antibiotics in triplicate.

Prediction of HTCS genes in *P. copri*

The genes of *P. copri* HDD04 that encodes proteins containing all the domains of HTCS (PF07494-PF07495-PF00512-PF02518-PF00072-PF12833) according to the Pfam classification were identified as described previously (Terrapon *et al*, 2015). The hmmsearch was carried out using default parameters (Eddy, 1998).

Molecular cloning

The relevant primers and plasmids are described in Dataset EV1. PCR amplification for cloning was carried out using Q5 High Fidelity DNA Polymerase (New England Biolabs). The PCR products were purified, and followed by DNA assembling with PCR amplified plasmid using Gibson reaction (HiFi DNA Assembly Master Mix, New England Biolabs). The assembled products were transformed into *E. coli* β2155 by chemical transformation. The resulting colonies were randomly picked to detect the inserts and their sizes by colony PCR using OneTaq DNA Polymerase (New England Biolabs). Genetic modifications generated on plasmids were verified by sequencing at Microsynth Seqlab (Microsynth AG, Germany).

Specifically, the pEx-insertion vector was constructed as follows: Firstly, the thymidine kinase gene (*tdk*) and its promoter was deleted from the multiple cloning site using DNA assembling with PCR amplified plasmid, resulting in pExchange. Secondly, 300 bp of the *tuf* (elongation factor Tu, DSM18205_02600) promoter was inserted into pExchange exactly before the coding sequence of *ermG*, generating the pEx-insertion-ermG vector. For first trial of genetic insertion in *P. copri*, 3-kb DNA sequences from DSM18205_00642-43, 00941-42, and 02334-35 were cloned into the multiple cloning site of pEx-insertion-ermG as the homology arm for plasmid integration. The pEx-insertion-ermG with the DNA region from DSM_02334-35 (DSM_02334: putative β-glycoside hydrolase, *bgl*) was designated as pEx-insertion-ermG-DSM-bgl.

Based on pEx-insertion-ermG as the vector backbone, similar cloning procedures were performed for constructing various plasmids carrying: (i) different promoter sequences for driving selective marker; (ii) different sizes of homology arms varying from 0.5 kb to 4 kb for integration in *P. copri* HDD04; (iii) 3-kb cloned homologous regions from different *P. copri* strains in our collection; (iv) a pLGB30-derived *tetQ* selective marker (García-Bayona & Comstock, 2019) instead of *ermG*; (v) the T1-T2 terminators copied from pSAM (Goodman *et al*, 2011) for blocking the transcriptional readthrough for the HTCS genes after plasmid integration.

The pEx-deletion-ermG and pEx-deletion-tetQ vectors were created by inserting the counter-selection marker following *ermG* and *tetQ*, respectively. The counter-selection marker is a DNA fragment generated by splicing 300-bp of the *gdhA* (HDD04_01507) promoter and the *sacB* gene from the pEX18Ap plasmid (Hoang *et al*, 1998). The pEx-deletion-ermG-bgl was similarly generated as described above. For in-frame deletion of genes in *P. copri*, the approximately 2-kb regions flanking the target gene were amplified, and assembled with PCR amplified pEx-deletion-ermG. For gene complementation, the target gene flanking with approximately 1-kb up- and down-stream regions were entirely amplified, and cloned into pEx-deletion-ermG.

Genetic manipulations of *P. copri*

Overnight culture of the *E. coli* donor strain was subcultured into LB medium containing ampicillin and DAP and *P. copri* subcultured into BHI+S medium. When they were grown to exponential phase (OD_{600} = 0.5–0.7), *E. coli* culture was transferred into the anaerobic chamber. The following procedures of genetic manipulations for *P. copri* including plasmid insertion, in-frame deletion, and complementation were performed in the anaerobic chamber. For conjugation, 1 ml *E. coli* culture (~ 5 × 10^9 CFUs) was centrifuged at 8,000 *g* for 3 min to pellet the bacterial cells, followed by resuspension in 100 μl fresh *P. copri* culture (~ 5 × 10^7 CFUs) to get a ratio of donor: recipient of 100:1. Specially, if *P. copri* HDA04 or HDD12 culture was used as the recipient strain for conjugation, to obtain the same donor/recipient ratio above, 20 μl HDA04 culture plus 80 μl BHI + S medium or 10 μl HDD12 culture plus 90 μl BHI + S medium was used to resuspend the *E. coli* pellet, respectively. The resuspension was then plated on a BHI blood agar with DAP for 18 h at 37°C for bacterial conjugation unless otherwise stated. Bacterial cells were washed off from the plate using 1 ml BHI + S medium, mixed well, and plated serial dilutions or the whole bacterial pellet after centrifugation on BHI blood agar plates containing gentamicin in addition of erythromycin or tetracycline. Colonies generated from transconjugants were visible after incubation of plates for 2–4 days according to properties of the *P. copri* derivatives. If necessary, the CFUs were counted for quantification of transconjugant yields. Insertion of the plasmid was verified by amplifying two joints between the bacterial chromosome and vector via colony PCR using P3/P4 and P5/P6 primer pairs, with P1/P2 amplified DNA as a control.

For in-frame deletion and complementation, the insertion mutants were grown in liquid BHI+S without selection and then subcultured every 12 h for allelic exchange. The final culture was plated onto YT agar plates supplemented with 5% sucrose to select the revertants (wild type) and gene deletion mutants with loss of the vector. After incubation of plates for 2–4 days, individual colonies were restreaked onto BHI blood plates in the presence and absence of erythromycin using the same inoculating loop, respectively, to further confirm erythromycin sensitivity of the clones. Erythromycin-sensitive clones were subsequently screened for the genetic modifications (gene deletion or complementation) by PCR and verified by sequencing at Microsynth Seqlab (Germany).

Prediction of PULs in *P. copri* genomes

The prediction of PULs in *P. copri* genomes and MAGs was described previously (Gálvez *et al*, 2020). Briefly, PULs and *susC/D*-like gene annotations were carried out using PULpy (preprint: Stewart *et al*, 2018) (commit 8955cdb, https://github.com/WatsonLab/PULpy). Annotation of carbohydrate-active enzymes (CAZymes) surrounding the *susC*/*susD*-like pairs was performed using dbCAN2 tool (Zhang *et al*, 2018) version v2.0.6 (CAZyDB=07312019, https://github.com/linnabrown/run_dbcan). Putative substrates of these CAZymes were predicted by dbCAN-PUL as described previously (Ausland *et al*, 2021).

Measurement of *P. copri* growth on a carbohydrate array

The growth curves of *P. copri* strains cultured in minimal medium (MM) supplemented with a sole carbohydrate were measured as previously described with the following modifications (Martens *et al*, 2011). The wells of a non-tissue culture flat bottom 96-well were loaded with 100 μl sterilized carbohydrate stocks (2× concentration). Each carbohydrate was added into at least three wells. *Prevotella copri* was grown in MM + Glucose to an OD_{600} value of approximately 0.6. 400 μl culture was then centrifuged to pellet the bacterial cells. The pellet was washed by 1 ml 2× MM without any carbohydrates and resuspended in 10 ml 2× MM as a seed culture. Each well of the plate was loaded with 100 μl seed culture. Absorbance at OD_{600} of each well was measured for 5 days by the microplate reader (BioTek) at 1-h intervals with 15-s pre-shaking.

In Figs 2D and EV4, and Dataset EV3, the maximal OD_{600} values subtracting the background reads (OD_{600} max) in the curves were identified for calculating means and standard deviations (SDs). Because we observed that the presence of erythromycin in MM significantly affect the duration of lag phase, but not the growth pattern of *P. copri*. The growth curves of HTCS gene insertion mutants and relevant intergenic insertion control were therefore shown starting from the OD_{600} values increased by 10% of the OD_{600} max to the OD_{600} max in Fig 2B and C.

RNA extraction from human feces and metatranscriptome sequencing

The fecal sample from the human donor carrying *P. copri* HDD04 was immediately collected into DNA/RNA Shield Fecal Collections Tubes and stored at 4°C for stabilizing RNA. An aliquot of 400 μl content from the tube was used for isolating RNA using ZymoBIOMICS RNA Miniprep Kit following the instruction manual.

RNA extraction and RNA-seq library preparation

Prevotella copri HDD04 was grown in BHI+S broth to the exponential phase (OD_{600} = 0.6). 5 ml fresh cultures were treated by RNAprotect (New England Biolabs) based on the manufacturer's instructions, pelleted by centrifugation, and stored at −80°C until further processing. The bacterial RNA was isolated using ZymoBIOMICS RNA Miniprep Kit following the instruction manual. RNA quality was evaluated by agarose gel electrophoresis, NanoDrop^{TM} 2000 (Thermo Scientific), and Bioanalyzer (Agilent Technologies) according to RNA integrity score (RIN > 8.0). Bacterial ribosomal RNA (rRNA) was then depleted by Ribo-Zero Gold rRNA Removal Kit (Epidemiology) as described in the commercial protocol. Libraries for Illumina sequencing were prepared using the NEBNext$^{®}$ Ultra$^{™}$ Directional RNA Library Prep Kit for Illumina$^{®}$ (New England Biolabs) following manufacturer's protocol. For each sample, 100 ng of fragmented mRNA was used as an input for cDNA synthesis and Illumina sequencing adaptor ligation.

For other treatment in Fig 4, *P. copri* HDD04 was initially grown in minimal media plus glucose (MM + Glucose) to an OD_{600} value of 0.5. 2 ml culture was then centrifuged to pellet the bacterial cells and resuspended using an equal volume of minimal media without carbohydrates (MM), followed by another centrifugation and resuspending in the same volume of MM. 40 μl suspension was inoculated into 4 ml MM plus glucose, arabinan, arabinoxylan, pectic galactan, and inulin, respectively. Three replicates were performed for each carbon source. Once *P. copri* grown to OD_{600} = 0.5, 750 μl culture was taken and treated by RNAprotect (New England Biolabs). Bacterial ribosomal RNA (rRNA) was thus depleted using Pan-prokaryote riboPOOL^{TM} Kit (siTOOLs Biotech) as described in the manual. The cDNA library preparation and sequencing was carried out as described above.

RNA-seq analysis

Reads were quality filtered using Trimmomatic (Bolger *et al*, 2014) version v0.33 with as follow parameters (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:35 HEADCROP:3). After quality control reads were aligned to each *P. copri* reference genome using STAR (Dobin *et al*, 2013) version v2.5.2a. Reads count was performed using HTSeq (Anders *et al*, 2015) version v0.11.2. With the aim to control for interspecies multi-mapping, reads were split by mapping to multiple references using (BBsplit). References genomes were selected from the reconstructed MAGs and one representative strain for each of the *P. copri* clades in combination with each donor's isolate from the clade A.

For *in vivo* and *in vitro* differential gene expression, gene read counts were transformed using TPMs normalization and differential gene expression was quantified in R using the DEseq analysis with a single replicate (iDEG) package (Li *et al*, 2019).

For the transcriptome *in vitro* with supplemented polysaccharides, samples were proceeded as described above. Normalization and differential expression were quantified in R using the DEseq2 package (Love *et al*, 2014) version v1.26.0 using the samples grown in MM + glucose as a control.

Measurements of gene transcription by RT–qPCR

The preparation of *P. copri* cultures and extraction of total RNA were performed as described above. Reverse transcription was carried out with ProtoScript$^{®}$ II First Strand cDNA Synthesis Kit (New England Biolabs) using Random Primer Mix and 800 ng purified RNA as template for 20 μl reaction. The abundance of transcript for target *susC*-like genes and reference *tuf* gene was quantified with KAPA SYBR$^{®}$ FAST qPCR mix (KAPA Biosystems) using 0.5 ng/μl template cDNA, 25 nM of each target gene-specific primer. The reaction was performed in a 96-well plate on the Roche Lightcycler 480. Using the ddCT method, raw values were normalized to values for the *tuf* gene and then the fold change was calculated by dividing MM+specific polysaccharide values by values obtained from MM+glucose.

Reconstruction of *P. copri* MAGs

The reconstruction of *P. copri* MAGs from a recent dataset (De Filippis *et al*, 2019) was performed as described previously (Gálvez *et al*, 2020). In brief, the sequencing data of the gut microbiome from 101 healthy Italian individuals with distinct diets (Omnivore, n = 25; Vegetarian, n = 39; Vegan, n = 37; NCBI SRA: SRP126540 and SRP083099) were analyzed as follows: (i) Sample-wise assembly, annotation, and integrative genomic binning was carried out with ATLAS metagenomic workflow (Kieser *et al*, 2020) (commit a007857, https://github.com/metagenome-atlas/atlas); (ii) genome abundance estimates were calculated for each sample by mapping the reads to the non-redundant MAGs using BBmap and determining the median coverage across each of the MAGs.

Quantification and statistical analysis

Statistical analysis

Statistical analyses were carried out in R (R Core Team, 2019) and figures were produced using 690 the package ggplot2 (Wickham, 2016). Datasets were analyzed using the GraphPad Prism 8. Pairwise comparisons were performed using Student's t test with a paired, two-tailed distribution. More statistical details are indicated in the associated figure legends when required.

Data and software availability

The accession numbers for all whole-genome sequencing and (meta)transcriptome data reported in this manuscript are available under NCBI BioProject ID: PRJNA684333 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA684333).

Expanded View for this article is available online.

Acknowledgements

We thank members of the Strowig laboratory for valuable discussions. We thank Andrew Goodman for helpful discussions. We thank the genome analytics core facility of the Helmholtz Institute for Infection Research. J. L. was funded by Singh-Chhatwal Stipend from Helmholtz Centre for Infection Research and Humboldt Research Fellowship from Alexander von Humboldt Foundation. Open Access funding enabled and organized by Projekt DEAL.

Author contributions

JL and TS designed the experiments and wrote the paper. JL conducted the most of experiments and data analysis. LA and AI performed the strain isolation. JL, EJCG, and TRL performed the bioinformatic analysis. LA and EA assisted in growth assays. LA and AAB assisted in RNA preparation for RNA-seq. EA and E-MS assisted in molecular cloning and genetic manipulation.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Accetto T, Avguštin G (2007) Studies on Prevotella nuclease using a system for the controlled expression of clones genes in *P. bryantii* TC1-1. *Microbiology* 153: 2281 – 2288

Accetto T, Peterka M, Avguštin G (2005) Type II restriction modification systems of *Prevotella bryantii* TC1-1 and *Prevotella ruminicola* 23 strains and their effect on the efficiency of DNA introduction via electroporation. *FEMS Microbiol Lett* 247: 177 – 183

Alpizar-Rodriguez D, Lesker TR, Gronow A, Gilbert B, Raemy E, Lamacchia C, Gabay C, Finckh A, Strowig T (2019) *Prevotella copri* in individuals at risk for rheumatoid arthritis. *Ann Rheum Dis* 78: 590 – 593

Anders S, Pyl PT, Huber W (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166 – 169

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M *et al* (2011) Enterotypes of the human gut microbiome. *Nature* 473: 174 – 180

Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U *et al* (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* 11: 2500

Ausland C, Zheng J, Yi H, Yang B, Li T, Feng X, Zheng B, Yin Y (2021) dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Res* 49: D523 – D528

Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307: 1915 – 1920

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19: 455

Bencivenga-Barry NA, Lim B, Herrera CM, Trent MS, Goodman AL (2020) Genetic manipulation of wild human gut bacteroides. *J Bacteriol* 202: e00544-19

Blomfield IC, Vaughn V, Rest RF, Eisenstein BI (1991) Allelic exchange in *Escherichia coli* using the *Bacillus subtilis* sacB gene and a temperature-sensitive pSC101 replicon. *Mol Microbiol* 5: 1447 – 1457

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114 – 2120

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12: 59 – 60

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)* 25: 1972 – 1973

Cerqueira FM, Photenhauer AL, Pollet RM, Brown HA, Koropatkin NM (2020) Starch digestion by gut bacteria: crowdsourcing for carbs. *Trends Microbiol* 28: 95 – 108

Chatzidaki-Livanis M, Geva-Zatorsky N, Comstock LE (2016) *Bacteroides fragilis* type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species. *Proc Natl Acad Sci USA* 113: 3627 – 3632

Claus SP (2019) The strange case of *Prevotella copri*: Dr. Jekyll or Mr, Hyde? *Cell Host Microbe* 26: 577 – 578

Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, de Vos WM, Ehrlich SD, Fraser CM, Hattori M *et al* (2018) Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol* 3: 8 – 16

De Filippis F, Pasolli E, Tett A, Tarallo S, Naccarati A, De Angelis M, Neviani E, Cocolin L, Gobbetti M, Segata N *et al* (2019) Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. *Cell Host Microbe* 25: 444 – 453.e3

De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci USA* 107: 14691 – 14696

De Vadder F, Kovatcheva-Datchary P, Zitoun C, Duchampt A, Bäckhed F, Mithieux G (2016) Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metab* 24: 151 – 157

Dean D (1981) A plasmid cloning vector for the direct selection of strains carrying recombinant plasmids. *Gene* 15: 99 – 102

Dehio C, Meyer M (1997) Maintenance of broad-host-range incompatibility group P and group Q plasmids and transposition of Tn5 in *Bartonella henselae* following conjugal plasmid transfer from *Escherichia coli*. *J Bacteriol* 179: 538 – 540

Demarre G, Guérout AM, Matsumoto-Mashimo C, Rowe-Magnus DA, Marlière P, Mazel D (2005) A new family of mobilizable suicide plasmids based on broad host range R388 plasmid (IncW) and RP4 plasmid (IncPα) conjugative machineries and their cognate *Escherichia coli* host strains. *Res Microbiol* 156: 245 – 255

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15 – 21

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14: 755 – 763

Fehlner-Peach H, Magnabosco C, Raghavan V, Scher JU, Tett A, Cox LM, Gottsegen C, Watters A, Wiltshire-Gordon JD, Segata N *et al* (2019) Distinct polysaccharide utilization profiles of human intestinal *Prevotella copri* isolates. *Cell Host Microbe* 26: 680 – 690.e5

Fragiadakis GK, Smits SA, Sonnenburg ED, Van Treuren W, Reid G, Knight R, Manjurano A, Changalucha J, Dominguez-Bello MG, Leach J *et al* (2019)

Links between environment, diet, and the hunter-gatherer microbiome. *Gut Microbes* 10: 216 – 227

Gálvez EJC, Iljazovic A, Amend L, Lesker TR, Renault T, Thiemann S, Hao L, Roy U, Gronow A, Charpentier E *et al* (2020) Distinct polysaccharide utilization determines interspecies competition between intestinal *Prevotella* spp. *Cell Host Microbe* 28: 838 – 852.e6

García-Bayona L, Comstock LE (2019) Streamlined genetic manipulation of diverse bacteroides and parabacteroides isolates from the human gut microbiota. *MBio* 10: e01762-19

Gardner RG, Russele JB, Wilson DB, Wang GR, Shoemaker NB (1996) Use of a modified Bacteroides-Prevotella shuttle vector to transfer a reconstructed beta-1,4-D-endoglucanase gene into Bacteroides uniformis and *Prevotella ruminicola* B(1)4. *Appl Environ Microbiol* 62: 196 – 202

Gay P, Le Coq D, Steinmetz M, Berkelman T, Kado CI (1985) Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. *J Bacteriol* 164: 918 – 921

Glenwright AJ, Pothula KR, Bhamidimarri SP, Chorev DS, Baslé A, Firbank SJ, Zheng H, Robinson CV, Winterhalter M, Kleinekathöfer U *et al* (2017) Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* 541: 407 – 411

Goodman AL, Wu M, Gordon JI (2011) Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. *Nat Protoc* 6: 1969 – 1980

Hoang TT, Karkhoff-Schweizer RR, Kutchma AJ, Schweizer HP (1998) A broad-host-range F1p-FRT recombination system for site-specific excision of chromosomally-located DNA sequences: application for isolation of unmarked *Pseudomonas aeruginosa* mutants. *Gene* 212: 77 – 86

Hooper LV (2009) Do symbiotic bacteria subvert host immunity? *Nat Rev Microbiol* 7: 367 – 374

Johnson EL, Heaver SL, Walters WA, Ley RE (2017) Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *J Mol Med* 95: 1 – 8

Kaoutari AE, Armougom F, Gordon JI, Raoult D, Henrissat B (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* 11: 497 – 504

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772 – 780

Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA (2020) ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* 21: 257

Koropatkin NM, Martens EC, Gordon JI, Smith TJ (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* 16: 1105 – 1115

Koropatkin NM, Cameron EA, Martens EC (2012) How glycan metabolism shapes the human gut microbiota. *Nat Rev Microbiol* 10: 323 – 335

Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, Hallen A, Martens E, Björck I, Bäckhed F (2015) Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of prevotella. *Cell Metab* 22: 971 – 982

Kovatcheva-Datchary P, Shoaie S, Lee S, Wahlström A, Nookaew I, Hallen A, Perkins R, Nielsen J, Bäckhed F (2019) Simplified intestinal microbiota to study microbe-diet-host interactions in a mouse model. *Cell Rep* 26: 3772 – 3783.e6

Ley RE (2016) Prevotella in the gut: choose carefully. *Nat Rev Gastroenterol Hepatol* 13: 69 – 70

Li Q, Zaim SR, Aberasturi D, Berghout J, Li H, Vitali F, Kenost C, Zhang HH, Lussier YA (2019) Interpretation of 'Omics dynamics in a single subject using local estimates of dispersion between two transcriptomes. *AMIA Annu Symp Proc* 2019: 582 – 591

Lim B, Zimmermann M, Barry NA, Goodman AL (2017) Engineered regulatory systems modulate gene expression of human commensals in the gut. *Cell* 169: 547 – 558.e15

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550

Luis AS, Briggs J, Zhang X, Farnell B, Ndeh D, Labourel A, Baslé A, Cartmell A, Terrapon N, Stott K *et al* (2018) Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nat Microbiol* 3: 210 – 219

Lynch JB, Sonnenburg JL (2012) Prioritization of a plant polysaccharide over a mucus carbohydrate is enforced by a Bacteroides hybrid two-component system. *Mol Microbiol* 85: 478 – 491

Maeda Y, Takeda K (2019) Host–microbiota interactions in rheumatoid arthritis. *Exp Mol Med* 51: 1 – 6

Martens EC, Chiang HC, Gordon JI (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* 4: 447 – 457

Martens EC, Koropatkin NM, Smith TJ, Gordon JI (2009) Complex glycan catabolism by the human gut microbiota: the bacteroidetes sus-like paradigm. *J Biol Chem* 284: 24673 – 24677

Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN *et al* (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* 9: e100122

Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF (2013) Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. *PLoS One* 8: e56018

Mimee M, Tucker AC, Voigt CA, Lu TK (2015) Programming a human commensal bacterium, *Bacteroides thetaiotaomicron*, to sense and respond to stimuli in the murine gut microbiota. *Cell Syst* 1: 62 – 71

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32: 268 – 274

Ogata K, Aminov RI, Tajima K, Nakamura M, Matsui H, Nagamine T, Benno Y (1999) Construction of *Prevoltella ruminicola-Escherichia coli* shuttle vector pRAM45 and transformation of *P. ruminicola* strains by electroporation. *J Biosci Bioeng* 88: 316 – 318

Patnode ML, Beller ZW, Han ND, Cheng J, Peters SL, Terrapon N, Henrissat B, Le Gall S, Saulnier L, Hayashi DK *et al* (2019) Interspecies competition impacts targeted manipulation of human gut bacteria by fiber-derived glycans. *Cell* 179: 59 – 73.e13

Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BAH, Forslund K, Hildebrand F, Prifti E, Falony G *et al* (2016) Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535: 376 – 381

Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, Hawkins JS, Lu CHS, Koo B-M, Marta E *et al* (2016) A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* 165: 1493 – 1506

R Core Team (2019) *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Porter NT, Martens EC (2017) The critical roles of polysaccharides in gut microbial ecology and physiology. *Annu Rev Microbiol* 71: 349 – 369

Recorbet G, Robert C, Givaudan A, Kudla B, Normand P, Faurie G (1993) Conditional suicide system of *Escherichia coli* released into soil that uses the *Bacillus subtilis* sacB gene. *Appl Environ Microbiol* 59: 1361 – 1366

Reyrat JM, Pelicic V, Gicquel B, Rappuoli R (1998) Counterselectable markers: untapped tools for bacterial genetics and pathogenesis. *Infect Immun* 66: 4011–4017

Ruengsomwong S, La-Ongkham O, Jiang J, Wannissorn B, Nakayama J, Nitisinprasert S (2016) Microbial community of healthy Thai vegetarians and non-vegetarians, their core gut microbiota, and pathogen risk. *J Microbiol Biotechnol* 26: 1723–1735

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406–425

Salyers AA, Shoemaker N, Cooper A, D'Elia J, Shipman JA (1999) 8 genetic methods for bacteroides species. *Methods Microbiol* 29: 229–249

Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, Rostron T, Cerundolo V, Pamer EG, Abramson SB *et al* (2013) Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *Elife* 2: e01202

Schwalm ND, Townsend GE, Groisman EA (2016) Multiple signals govern utilization of a polysaccharide in the gut bacterium *Bacteroides thetaiotaomicron*. *MBio* 7: e01342-16

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

Shipman JA, Berleman JE, Salyers AA (2000) Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. *J Bacteriol* 182: 5365–5372

Shoemaker NB, Anderson KL, Smithson SL, Wang GR, Salyers AA (1991) Conjugal transfer of a shuttle vector from the human colonic anaerobe *Bacteroides uniformis* to the ruminal anaerobe *Prevotella* (*Bacteroides*) *ruminicola* B14. *Appl Environ Microbiol* 57: 2114–2120

Sonnenburg ED, Sonnenburg JL, Manchester JK, Hansen EE, Chiang HC, Gordon JI (2006) A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing *in vivo* to carbohydrate metabolism. *Proc Natl Acad Sci USA* 103: 8834–8839

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

Stewart RD, Auffret MD, Roehe R, Watson M (2018) Open prediction of polysaccharide utilisation loci (PUL) in 5414 public Bacteroidetes genomes using PULpy. *bioRxiv* https://doi.org/10.1101/421024 [PREPRINT]

Terrapon N, Lombard V, Gilbert HJ, Henrissat B (2015) Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* 31: 647–655

Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M *et al* (2019) The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* 26: 666–679.e7

Wang J, Shoemaker NB, Wang G-R, Salyers AA (2000) Characterization of a Bacteroides mobilizable transposon, NBU2, which carries a functional lincomycin resistance gene. *J Bacteriol* 182: 3559–3571

Wells PM, Adebayo AS, Bowyer RCE, Freidin MB, Finckh A, Strowig T, Lesker TR, Alpizar-Rodriguez D, Gilbert B, Kirkham B *et al* (2020) Associations between gut microbiota and genetic risk for rheumatoid arthritis in the absence of disease: a cross-sectional study. *Lancet Rheumatol* 2: e418–e427

Wexler AG, Goodman AL (2017) An insider's perspective: bacteroides as a window into the microbiome. *Nat Microbiol* 2: 17026

Wickham H (2016) *ggplot2: elegant graphics for data analysis*. New York, NY: Springer-Verlag. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R *et al* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334: 105–108

Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI (2003) A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science* 299: 2074–2076

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 46: 95–101