Original Article

CAMBRIDGE
UNIVERSITY PRESS

Rapid PCR identification of *Prevotella copri* in an Australian cohort of pregnant women

Lawrence Gray[1,2] , Kyoko Hasebe[1] , Martin O'Hely[1,3], Anne-Louise Ponsonby[3,4], Peter Vuillermin[1,2,3], Fiona Collier[1,2,3] and the BIS Investigator Group[1,2,3,4]

[1]Deakin University, Geelong, Australia; [2]Barwon Health, Geelong, Australia; [3]Murdoch Children's Research Institute, Melbourne, Australia and [4]The Florey Institute of Neuroscience and Mental Health, Melbourne, Australia

Abstract

Gut bacteria from the genus *Prevotella* are found in high abundance in faeces of non-industrialised communities but low abundance in industrialised, Westernised communities. *Prevotella copri* is one of the principal *Prevotella* species within the human gut. As it has been associated with developmental health and disease states, we sought to (i) develop a real-time polymerase chain reaction (PCR) to rapidly determine *P. copri* abundance and (ii) investigate its abundance in a large group of Australian pregnant mothers.

The Barwon Infant Study is a pre-birth cohort study ($n = 1074$). Faecal samples were collected from mothers at 36 weeks gestation. Primers with a probe specific to the V3 region of *P. copri* 16S rRNA gene were designed and optimised for real-time PCR. Universal 16S rRNA gene primers amplified pan-bacterial DNA in parallel. Relative abundance of *P. copri* was calculated using a $2^{-\Delta Ct}$ method.

Relative abundance of *P. copri* by PCR was observed in 165/605 (27.3%) women. The distribution was distinctly bimodal, defining women with substantial ($n = 115/165$, 69.7%) versus very low *P. copri* expression ($n = 50/165$, 30.3%). In addition, abundance of *P. copri* by PCR correlated with 16S rRNA gene MiSeq sequencing data ($r^2 = 0.67$, $P < 0.0001$, $n = 61$).

We have developed a rapid and cost-effective technique for identifying the relative abundance of *P. copri* using real-time PCR. The expression of *P. copri* was evident in only a quarter of the mothers, and either at substantial or very low levels. PCR detection of *P. copri* may facilitate assessment of this species in large, longitudinal studies across multiple populations and in various clinical settings.

Introduction

The increased incidence of immune-mediated non-communicable diseases (NCDs), such as allergic disease and diabetes, in developed nations has paralleled changes in our environment. The highly urbanised industrialised, "Western" environments reduce environmental bacterial diversity[1] and opportunities for human-microbial interaction throughout the life course.[2] The effects of environmental changes may be mediated by our commensal microbiota. The human gut microbiome contains the largest number of bacteria in the body, acting in a symbiotic relationship with the host to mediate barrier function of the gut, modulate immune responses, synthesise cofactors and ferment dietary fibre to create accessible carbohydrates for host gut enterocytes.[3] Increasing evidence exists that the maternal gut microbiome influences offspring immune development, establishing the immune profile of infants at birth.[4]

Studies of human gut microbiota by 16S rRNA gene amplicon sequencing have identified enterotypes associated with a predominance of three key bacterial genera: *Prevotella*, *Bacteroides* and *Ruminococcus*.[5,6] These enterotypes have been associated with specific dietary patterns[7] and geographically distinct populations.[8] From infancy, the *Prevotella* enterotype is most abundant in the faeces of non-industrialised, high-fibre eating societies[9–11] and in lower abundance in the faeces of people living in industrialised, developed Western populations.[5,7] Within these Western populations, the *Prevotella* enterotype exhibits a bimodal distribution, with either high or low abundance, but rarely intermediate abundance.[12] High abundance of the *Prevotella* enterotype is associated with high-faecal microbial diversity.[12]

Prevotella copri (*P. copri*) is the predominant *Prevotella* species within the human gut microbiome[13,14] and is associated with both host benefits[15,16] and a number of NCDs that have their origins in early life.[17–19] Beneficially, lesser abundance of faecal *P. copri* has been observed in childhood atopic dermatitis.[16] While in contrast, greater abundance has been associated with early stage rheumatoid arthritis.[17] Conflictingly, greater abundance of faecal *P. copri* has been associated not only with both increased insulin sensitivity and glycaemic control[15] but also with increased insulin resistance and impaired glucose tolerance.[20] In this context, *P. copri* is a potentially useful biomarker and the ability to rapidly and accurately detect the presence

Table 1. Details of primers and probe

Primer/probe details		Sequence	Position	Amplicon size (bp)
P. copri, 16S rRNA (NR_113411)	For-	5′-CGCGAACTGGTTTCCTTGA-3′	632–687	56
	Rev-	5′-ACCGCTACACCACGAATTCC-3′		
	Taqman probe	FAM-5′-ACGCACAAAGTGGG-3′-MGBNFQ		
Universal bacterial 16S rRNA (V4 region)	For-	5′-CCTAYGGGRBGCASCAG-3′	341–533	193
	Rev-	5′-TTACCGCGGCTGCTGGCA-3′		

of faecal *P. copri* and level of carriage without perfoming 16S rRNA gene or metagenomic sequencing would improve the cost implications and feasibility of large-scale studies of the relationships between gut bacteria and NCDs.

The aim of this study was to develop a rapid method for identification and determination of relative abundance of *P. copri* and investigate the distribution of abundance. This was performed in a cohort of pregnant women from the Barwon Infant Study (BIS), a large Australian birth cohort.[21] Primers and probes for a real-time Taqman polymerase chain reaction (PCR) were designed, and the abundance of *P. copri* was determined relative to amplification of pan bacterial 16S rRNA genes.

Methods

Faecal collection and processing

Samples were selected from women in the BIS, a pre-birth cohort study in the southeast of Australia ($n = 1064$ mothers and 1074 infants) assembled using an unselected antenatal sampling frame. The eligibility criteria and cohort profile are described elsewhere.[21] Faecal samples were collected from pregnant women in the BIS at 36 weeks gestation and stored at $-80\,°C$ prior to DNA extraction. DNA was extracted using the Qiagen PowerSoil® DNA Isolation Kit (Cat#12888-100) (QIAGEN Pty Ltd., Victoria, Australia). The quality of the DNA was assessed by NanoDrop and samples diluted to concentrations of 10 ng/µl and 20 ng/µl, for 16S rRNA gene diversity profiling ($n = 410$) and PCR ($n = 605$), respectively. Within these, the total number with both 16S rRNA gene MiSeq diversity profiling and *P. copri* PCR was 351 samples.

Relative abundance of P. copri by 16S rRNA gene MiSeq diversity profiling

A sample of the BIS cohort ($n = 410$) was selected for 16S rRNA gene bacterial diversity profiling sequencing, which was performed as described in Appendix *Sample collection and processing*. Briefly, the V4 region of the 16S rRNA gene was sequenced on the Illumina MiSeq platform. Operational taxonomic units (OTUs) were formed based on 97% identity, and taxonomic identification was based on the SILVA v123Nr99 database. Candidate OTUs representing *P. copri* were selected on the basis of (i) assignment to the SILVA genus-level taxon "Prevotella 9" consistent with *P. copri* and (ii) examining the alignment of their representative sequences with the Microbial Nucleotide BLAST tool.[22] The OTU best-matching *P. copri* JCM 13464 16S rRNA gene (Accession No: AB649279) was taken to represent *P. copri*. For each sample, the relative abundance of *P. copri* was determined by dividing the number of sequences assigned to this OTU by the total number of sequences assigned to any OTU.

Relative abundance of P. copri by PCR

P. copri PCR

Primers specific to the V4 region of *Prevotella* 16S rRNA gene and a 20-mer Taqman 6-carboxyfluorescein (FAM)-labelled probe were designed using Primer Express (Thermo Fisher Scientific). The primer sequences and the PCR conditions are described in Table 1. In brief, the PCR reaction comprised 10 µl of Taqman Universal Master Mix II (Cat Nos 4440040, Thermo Fisher Scientific Australia), 1 µl of forward primer (1 µM), 1 µl of reverse primer (1 µM), 0.2 µl of Taqman *P. copri* probe, 2 µl of template DNA (20 ng/µl) and 5.8 µl of PCR grade sterile water. The PCR cycling conditions were $50\,°C$ hold for 2 min, $95\,°C$ hold for 10 min to activate the hot start DNA polymerase, followed by 42 cycles of $95\,°C$ for 15 s and $60\,°C$ for 1 min.

#### Universal 16S rRNA gene PCR

16S rRNA gene universal primers (341F-533R) were utilised to amplify the V4 region of pan bacterial species (Table 1). Cycle threshold (Ct) values for each sample were used for the calculation of *P. copri* expression (details below). The 16S rRNA gene PCR reaction comprised of 10 µl of PowerUp SyBR Green Master Mix (catalogue #A25742, Thermo Fisher Scientific Australia), 1 µl of forward primer (1 µM), 1 µl of reverse primer (1 µM), 2 µl of template DNA (20 ng/µl) and 6 µl of PCR grade sterile water. The PCR cycling condition was set as $50\,°C$ hold for 2 min, $95\,°C$ hold for 10 min, $95\,°C$ hold for 15 s, $60\,°C$ hold for 30 s, $72\,°C$ hold for 45 s, $95\,°C$ hold for 15 s, $59\,°C$ hold for 30 s, $72\,°C$ hold for 45 s, $95\,°C$ hold for 15 s, $58\,°C$ hold for 30 s, $72\,°C$ hold for 45 s, followed by 42 cycles of $95\,°C$ hold for 15 s, $57\,°C$ hold for 30 s and $72\,°C$ hold for 45 s. A melting curve was performed after 42 cycles at $95\,°C$ hold for 15 s, $60\,°C$ hold for 1 min, $95\,°C$ hold for 30 s and $60\,°C$ hold for 15 s.

PCR controls

DNA isolated from (i) a pure colony of *P. copri* (DSMZ 18205, www.dsmz.de) and (ii) a BIS maternal faecal sample that had a very high abundance of *P. copri* by 16S rRNA gene MiSeq diversity profiling was used as positive controls (Positive Control-1 and Positive Control-2, respectively). Two samples that were negative for *P. copri* but had exhibited 16S rRNA gene sequence for a related species, *Prevotella jejuni*, were selected as sensitivity controls. The positive controls were run on each PCR plate.

Linearity of P. copri and universal 16S rRNA gene PCR reactions

DNA from both the *P. copri* positive controls was serially diluted (four-fold dilutions from 20 ng/µl to 4.8×10^{-6} ng/µl), and samples were amplified for *P. copri* and universal 16S rRNA gene. The curves generated were assessed for linearity and
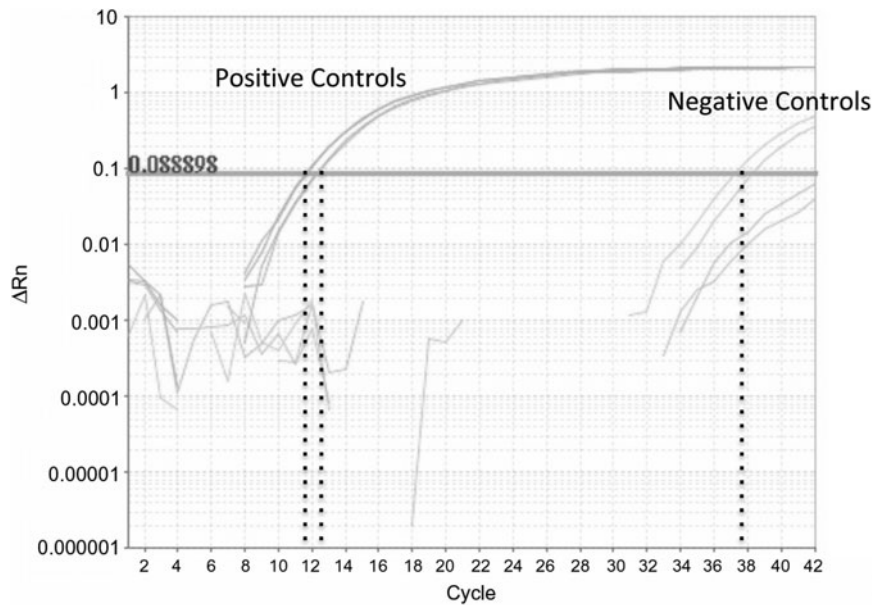
Fig. 1. Real-time PCR amplification of *P. copri in faecal DNA*. A Taqman reaction with primers and FAM conjugated probe was used to detect the presence of *P. copri*. The samples selected as positive controls were amplified (pure bacterial *P. copri* DNA and high abundance *P. copri* faecal DNA with starting DNA concentration of 20 ng/µl) and had Ct values of 11.7 and 12.4, respectively. Samples selected as negative controls (high abundance *P. jejuni* faecal DNA with starting DNA concentration of 20 ng/µl) had a Ct value >37.0.

provided an estimated threshold of reliable detection for both *P. copri* and universal 16S rRNA gene.

Calculation of relative abundance of P. copri *by PCR*

The relative abundance of *P. copri* by PCR was calculated (when there was reliable detection for both *P. copri* and universal 16S rRNA gene) by the $2^{-\Delta Ct}$ method,[23] where ΔCt was the differences in Ct values for *P. copri* and universal 16S rRNA gene, and relative abundance of *P. copri* was described in arbitrary units (a.u.). Comparison of the relative abundance of *P. copri* by PCR with the relative abundance by 16S rRNA gene sequencing was performed by linear regression after log transforming both measures.

Abundance of P. copri *by whole genome NGS*

Bacterial DNA from seven women's samples was sheared, adapter ligation and nick repair performed using the Ion Plus Fragment Library kit, cleaned and quantitated, and metagenomic sequencing performed on the Ion Torrent S5xl System (Thermo Fisher Scientific). The total reads for the samples were between 3,824,400 and 9,384,897. The assembled contigs were mapped to the NCBI Reference Sequence: NZ_ACBX00000000.2, *Prevotella copri* DSM 18205, whole genome shotgun sequencing project, and the proportion of mapped reads/total reads assessed for each of the seven samples.

Results

16S rRNA gene MiSeq profile

Six OTUs were assigned to the "Prevotella 9" taxon. The most frequent, which we refer to by the non-canonical name *OTU41*, represented 97.0% of all "Prevotella 9" OTU counts and its representative sequence was an exact match with *P. copri* JCM 13464 16S rRNA gene (Accession No: AB649279). Thus, we used the relative abundance of OTU41 as a marker for the relative abundance of *P. copri*, concluding that it was present in 79/410 samples at relative abundances between 0.0037% and 66% (mean 7.5% when present, 1.4% overall; median 0.37% when present).

P. copri *PCR*

The Taqman primer/probe PCR reaction worked efficiently, with positive controls giving Ct values <15 and sensitivity controls undetermined up to 34 cycles (Fig. 1). This indicates high specificity for the PCR. Stocks of the positive control DNA were used in all PCR assays that followed, with Ct values set to the same level throughout all semi-quantitative PCR assays.
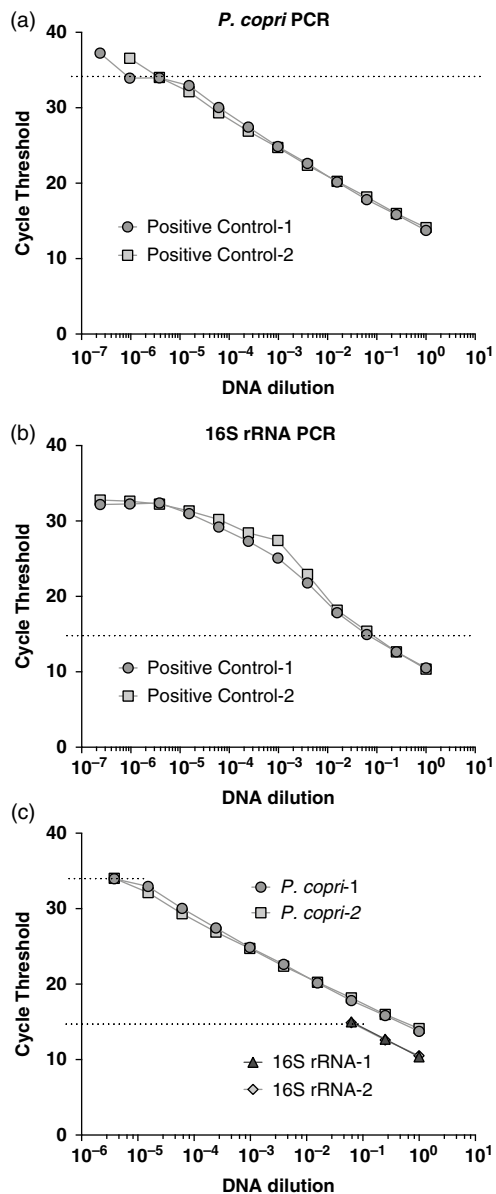
Universal 16S rRNA gene PCR

The same set of samples was run using the universal 16S rRNA gene primers. A Sybr Green reaction was used with a number of touchdown steps for the primer annealing. The positive control DNA was also used in all the PCR assays, with Ct values set to the same level throughout all PCR assays.

Assessment of linearity of PCR reactions

The *P. copri* PCR was linear for serially diluted samples (one in four dilutions, 20 ng/µl to 3.8×10^{-6} ng/µl). This indicated high sensitivity for the detection of *P. copri*. The universal 16S rRNA gene PCR reaction, however, was only linear from 20 ng/µl to 1.25 ng/µl (1/16 dilution). An estimated threshold for reliable detection of *P. copri* was ≤34 Ct corresponding to the detection of negative controls (Fig. 1) and the linearity threshold (Fig. 2a). For reliable calculation of relative expression, the threshold for universal 16S rRNA gene was ≤15 Ct for each sample (Fig. 2b, 2c).

Relative abundance of P. copri *by PCR*

In this group of pregnant Australia women, detectable levels of *P. copri* were observed in 165/605 (27.3%) women. The relative abundance of *P. copri* by PCR ranged from 1.5×10^{-5} to 141.6 arbitrary units (a.u.) and after log-transformation showed a distinct bimodal pattern (Fig. 3). Based on this distribution, we assigned those mothers with relative abundance by PCR less than 10^{-2} a.u. as having "very low" carriage of *P. copri* ($n = 50$), equivalent to a Ct >32.9 ± 2.0 (mean ± SD), while mothers with relative abundance by PCR greater than 10^{-2} a.u. were defined as having "substantial" carriage of *P. copri* ($n = 115$).

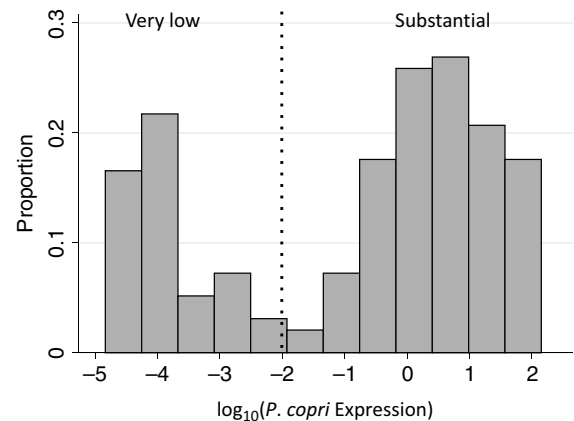Journal of Developmental Origins of Health and Disease
231

Fig. 2. Determination of linearity of the real-time PCR reactions for *P. copri* and 16SrRNA gene sequencing. The linearity of *P. copri* and universal 16S rRNA gene PCR reactions was determined by amplification of serially dilutions (¼ sequentially) of the positive control DNA (from 20 ng/μl). The Ct thresholds (·········) for calculation of *P. copri* expression was estimated to be 34 Ct for the (a) *P. copri* reaction and 15 Ct for universal 16S rRNA PCR reaction in both (b) positive controls and (c) pure *P. copri*.
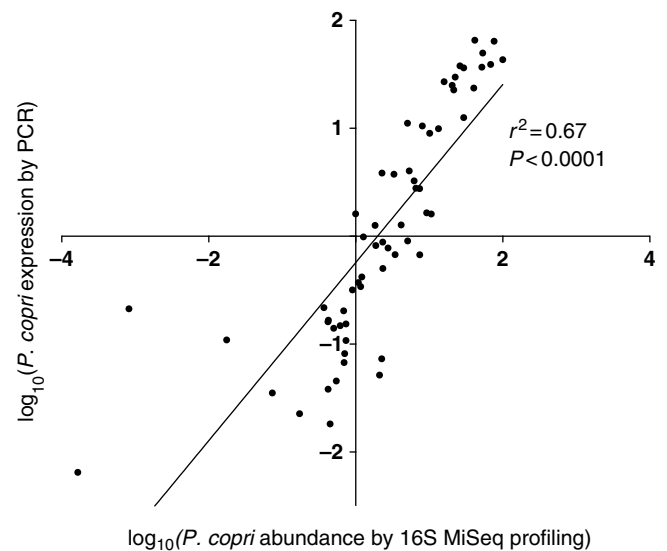

Fig. 3. Bimodal distribution of *P. copri* expression. *P. copri* expression (arbitrary units, a.u.) in faecal samples from pregnant mothers was determined by real-time PCR. 165/605 samples were classified as *P. copri* positive, and the distribution of expression was bimodal, with 50 samples $<10^{-2}$ a.u. classified as "very low", and 115 samples $>10^{-2}$ a.u. classified as "substantial".

Fig. 4. Scatter plot of relative abundance of *P. copri* measured by real-time PCR and 16S rRNA gene sequencing. The relative abundance of *P. copri* by PCR was strongly correlated with the relative abundance determined by 16S rRNA gene sequencing using Illumina MiSeq ($r^2 = 0.67$, $P < 0.0001$, $n = 61$).

Correlation between the relative abundance of P. copri *by PCR with other methods*

16S rRNA gene MiSeq sequencing
The relative abundance of *P. copri* by PCR was strongly correlated with the relative abundance determined by Illumina MiSeq 16S rRNA gene sequencing ($r^2 = 0.67$, $P < 0.0001$, $n = 61$) (Fig. 4), and the correlation was strongest when the relative abundance by both methods was greater than 10^{-1}. There were 39 samples deemed positive for *P. copri* by PCR (relative abundance of *P. copri* by PCR, 1.5×10^{-5} to 0.8 a.u.) but not detected by 16S rRNA gene sequencing. Of these 39, 30 expressed *P. copri* in the very low carriage range and 29 had <5000 total sequence read counts in the MiSeq 16S rRNA gene profiling.

Whole genome NGS
There was absolute agreement between the *P. copri* PCR and the unbiased next generation sequencing (NGS) with 4/7 of the samples not detecting *P. copri* (Supplementary Fig. 1). Of the three samples with detectable *P. copri*, the methods correlated almost perfectly ($r^2 = 0.99$, $P = 0.001$) (Supplementary Fig. 1).

Discussion

Our study utilised a semi-quantitative real-time PCR method to determine the relative abundance of *P. copri* and demonstrate a bimodal distribution of *P. copri* in faecal microbiota of pregnant Australian women. To our knowledge, this is the first time that the carriage pattern of this key bacterial species, which is implicated in the development of NCDs, has been reported by direct identification. The probe developed was specific for the detection

of *P. copri*, and the derived relative abundance showed a high level of agreement with the relative abundance determined by 16S rRNA gene sequencing data. The assay will assist in rapid and inexpensive semi-quantification of this important bacterium in future studies.

Targeted PCR has been used extensively in clinical microbiology laboratories for routine diagnosis of infectious diseases,[24] and primers that are able to amplify the pan bacterial 16S rRNA genes have been employed to enable semi-quantitation of the target.[25] Here, we used specific primers and probe to identify a particular gut bacterial species. Although the amplicon is not large, the amplified sequence is 100% matched in coverage and identity to *P. copri* only. A related *Prevotella* species, *P. jejuni*, also found in faeces, had 98% coverage and 91% identity for the amplicon, but reassuringly was not detected in the accepted Ct range for the assay. The approach used here for identification and quantification of a bacterial species against bacterial content could be applied to the determination of specific strains of bacteria. Strain differences may be key to understanding the apparently contradictory association of bacteria with both health and disease.[26] As an example of this, a recent study showed that non-Western individuals harboured *P. copri* strains with a higher number of genes encoding carbohydrate catabolism and starch binding mechanisms, while Western individuals had strains with increased genetic capacity for protein and vitamin utilisation.[27] PCR potentially provides an efficient and cost-effective method of strain detection and quantification in large studies.

Reported associations between *P. copri* and human health outcomes may be contingent on abundance state. In previous studies, based on 16S rRNA gene sequencing, a bimodal abundance distribution of the genus *Prevotella* has been reported;[12] however, this is the first time to our knowledge that a bimodal distribution of carriage of *P. copri* has been observed in a large population cohort. We defined the relative abundance of *P. copri* by PCR as either substantial ($>10^{-2}$a.u.) or very low ($<10^{-2}$a.u.). It is important to recognise that $<10^{-2}$a.u means that carriage was only detected after 33 PCR cycles, and the meaningfulness of detection at such exceptionally low levels is uncertain. It is however possible that low abundance carriage represents a reservoir to maintain microbial diversity. In murine models, Sonnenburg et al. demonstrated that long-term dietary change may alter faecal bacterial composition irreversibly.[28] Importantly, bacterial diversity loss was only partly reversible with a return to a broader diet. These findings parallel changes in faecal bacterial composition identified in modern industrialised societies. Highly Westernised gut microbiomes lack diversity,[8,10,11] thought to be related to accumulated species loss as a result of narrow diet[28] and exposure to antibacterial agents.[29] A lower abundance state may be important in maintaining gut microbiome diversity by allowing repopulation with bacterial species in response to environmental perturbations.

Greater abundance of *P. copri*, defined by 16S rRNA gene sequencing, has been associated with both benefits to the host,[15,16] as well as inflammatory disease states.[17,18] Although less evidence exists to date, lower abundance carriage of *P. copri* may also influence health outcomes. Recently, it has been reported that a lower relative abundance of *P. copri*, defined by 16S rRNA gene sequencing, was associated with atopic dermatitis.[16] Further research in larger cohorts using PCR is required to examine whether *P. copri* expression state, and in particular very low carriage, is associated with health outcomes.

Strategies for reducing the complexity of microbiome data are crucial to assessing associations within an epidemiological framework, as they enable meaningful approaches to important issues such as the selection of covariates to include in adjusted analyses[30] and assessment of causal inference features.[31] In general, data dimension reduction steps are applied during the analysis of data derived via DNA sequencing platforms (16S rRNA gene or metagenomic) and include the calculation of summary measurements such as alpha diversity, phylum level abundance or principal co-ordinates, and classification into enterotypes or according to presence/absence of selected taxa. A limitation of this approach is that the sequencing techniques and subsequent bioinformatic analyses remain time-consuming and expensive. The microbiome is highly dynamic and identifying associations that are likely to be causal represents a crucial challenge for the field.[32] In addition to the improved biologic insights provided by advances in sequencing technology and bioinformatic techniques, we need studies involving large numbers of participants, with frequent longitudinal measures, and adequate consideration of covariates; and of course replication. In this context, simple, inexpensive measures of microbiome biomarkers, such as *P. copri* carriage, may compliment in-depth, resource-intensive, sequencing studies.

The strengths of this study include the use of a large population cohort, as well the use of negative controls, a bacterial DNA with a high expression of an alternate *Prevotella* species. The PCR for *P. copri* appears to be both specific and sensitive to very low levels of the bacteria. The expression quantitation also shows good correlation with 16S rRNA gene sequencing *P. copri* abundance but may be even more sensitive for the very low carriage state. The main limitation of this study is that participants were all pregnant women from a defined geographical area. Pregnancy, sex and environment may all have an effect on the composition of the gut microbiome, and this may limit the generalisability of the findings. However, the cohort demographics reflect the Australian population more broadly,[21] and thus these findings may be relevant to other Westernised populations.

Conclusion

We developed a rapid and cost-effective technique for identifying the faecal bacteria *P. copri* using quantitative real-time PCR. The expression of *P. copri* in our cohort was distributed bimodally, with both substantial and very low faecal expression. Further investigation of the very low carriage state may be important in understanding the role of *P. copri* in health and disease or maintaining a microbiome community structure in Westernised environments. The use of real-time PCR measures may facilitate large, longitudinal studies across multiple populations and settings, which will compliment smaller, resource-intensive sequencing-based studies.

Conflict of Interest. None.

Ethical Standards. Ethics approval (10/24) for this study was obtained from the Barwon Health Human Research Ethics committee.

References

1. Flandroy L, Poutahidis T, Berg G, *et al.* The impact of human activities and lifestyles on the interlinked microbiota and health of humans and of ecosystems. *Sci Total Environ.* 2018; 627, 1018–1038.
2. Parajuli A, Gronroos M, Siter N, *et al.* Urbanization reduces transfer of diverse environmental microbiota indoors. *Front Microbiol.* 2018; 9, 84.
3. Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. *Expert Rev Anti Infect Ther.* 2010; 8(4), 435–454.
4. Gray LE, O'Hely M, Ranganathan S, Sly PD, Vuillermin P. The maternal diet, gut bacteria, and bacterial metabolites during pregnancy influence offspring asthma. *Front Immunol.* 2017; 8, 365.
5. Arumugam M, Raes J, Pelletier E, *et al.* Enterotypes of the human gut microbiome. *Nature.* 2011; 473(7346), 174–180.
6. Costea PI, Hildebrand F, Arumugam M, *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol.* 2018; 3(1), 8–16.
7. Wu GD, Chen J, Hoffmann C, *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* 2011; 334(6052), 105–108.
8. Yatsunenko T, Rey FE, Manary MJ, *et al.* Human gut microbiome viewed across age and geography. *Nature.* 2012; 486(7402), 222–227.
9. De Filippo C, Cavalieri D, Di Paola M, *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci USA.* 2010; 107(33), 14691–14696.
10. Schnorr SL, Candela M, Rampelli S, *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat Commun.* 2014; 5, 3654.
11. Martinez I, Stegen JC, Maldonado-Gomez MX, *et al.* The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Rep.* 2015; 11(4), 527–538.
12. Lahti L, Salojarvi J, Salonen A, Scheffer M, de Vos WM. Tipping elements in the human intestinal ecosystem. *Nat Commun.* 2014; 5, 4344.
13. Ley RE. Gut microbiota in 2015: prevotella in the gut: choose carefully. *Nat Rev Gastroenterol Hepatol.* 2016; 13(2), 69–70.
14. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* 2015; 16, 82.
15. De Vadder F, Kovatcheva-Datchary P, Zitoun C, Duchampt A, Backhed F, Mithieux G. Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metab.* 2016; 24(1), 151–157.
16. Mahdavinia M, Rasmussen HE, Botha M, *et al.* Effects of diet on the childhood gut microbiome and its implications for atopic dermatitis. *J Allergy Clin Immunol.* 2018; 143(4), 1636–1637e.5.
17. Scher JU, Sczesnak A, Longman RS, *et al.* Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife.* 2013; 2, e01202.
18. Dillon SM, Lee EJ, Kotter CV, *et al.* Gut dendritic cell activation links an altered colonic microbiome to mucosal and systemic T-cell activation in untreated HIV-1 infection. *Mucosal Immunol.* 2016; 9(1), 24–37.
19. Kristensen NB, Bryrup T, Allin KH, Nielsen T, Hansen TH, Pedersen O. Alterations in fecal microbiota composition by probiotic supplementation in healthy adults: a systematic review of randomized controlled trials. *Genome Med.* 2016; 8(1), 52.
20. Pedersen HK, Gudmundsdottir V, Nielsen HB, *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature.* 2016; 535(7612), 376–381.
21. Vuillermin P, Saffery R, Allen KJ, *et al.* Cohort profile: the Barwon infant study. *Int J Epidemiol.* 2015; 44(4), 1148–1160.
22. (NCBI) NCfBI. Microbial Nucleotide BLAST Tool [Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes.
23. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001; 25(4), 402–408.
24. Maurin M. Real-time PCR as a diagnostic tool for bacterial diseases. *Expert Rev Mol Diagn.* 2012; 12(7), 731–754.
25. Sontakke S, Cadenas MB, Maggi RG, Diniz PP, Breitschwerdt EB. Use of broad range16S rDNA PCR in clinical microbiology. *J Microbiol Methods.* 2009; 76(3), 217–225.
26. Nicol MP, Wilkinson RJ. The clinical consequences of strain diversity in *Mycobacterium tuberculosis. Trans R Soc Trop Med Hyg.* 2008; 102(10), 955–965.
27. De Filippis F, Pasolli E, Tett A, *et al.* Distinct genetic and functional traits of human intestinal *Prevotella copri* Strains are associated with different habitual diets. *Cell Host Microbe.* 2019; 25(3), 444–453 e3.
28. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. *Nature* 2016; 529(7585), 212–215.
29. Yassour M, Vatanen T, Siljander H, *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med.* 2016; 8(343), 343ra81.
30. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol.* 2019; 34(3), 211–219.
31. Ponsonby A-L, Dwyer T. Biomedicine must look beyond P values. *Nature* 2014; 507, 169.
32. Schmidt TSB, Raes J, Bork P. The human gut microbiome: from association to modulation. *Cell.* 2018; 172(6), 1198–1215.
33. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol.* 2013; 79(17), 5112–5120.
34. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods.* 2013; 10(10), 996–998.
35. Schloss PD, Westcott SL, Ryabin T, *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009; 75(23), 7537–7541.
36. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 2013; 8(4), e61217.
37. R Core Team. *R: A Language and Environment for Statistical Computing*, 2018. R Foundation for Statistical Computing, Vienna. Available from: https://www.r-project.org/.

Appendix

Sample collection and sequencing

Faecal samples were collected from pregnant women at 36 weeks gestation. Prior to arriving at the laboratory, samples were either stored temporarily at 4 °C or frozen at −20 °C. Samples were thawed and aliquoted to approximately 200 mg and stored for 1–3 years at −80 °C. In a subset of 410 mothers, DNA was extracted using the Qiagen PowerSoil® DNA Isolation Kit, Cat#12888-100 and transported to the J. Craig Venter Institute, Rockville, MD, USA. The V4 variable region of the 16S rRNA gene was amplified using adaptor and barcode-ligated V4 primers (forward: 5′-GTGCCAGCMGCCGCGGTAA-3′, reverse 5′-GGACTACH VGGGTWTCTAAT-3′). PCR amplicons were generated from approximately 100 ng of extracted DNA using Platinum Taq polymerase (ThermoFisher Inc.) and cycling conditions: 94 °C followed by 35 cycles of 94 °C for 30 s, 57 °C for 30 s and 72 °C for 30 s, followed by a final extension cycle of 72 °C for 5 min. Amplicons were purified using the Qiaquick PCR purification kit (QIAGEN Inc.) following manufacturer's specifications.

Purified amplicons were quantified using SybrGold, normalised and pooled in preparation for Illumina MiSeq sequencing. Libraries were sequenced using the dual index V2 chemistry kit 2x250bp on an Illumina MiSeq according to manufacturer's specifications.[33] USEARCH software[34] merged corresponding paired-end reads, filtered (to remove merged reads with mismatches, too many or too few base pairs), clustered into 97% identity OTUs, identified OTU representative sequences and removed chimeras. The mothur software suite[35] assigned representative sequences to taxa described in the SILVA v123Nr99 taxonomic database. The final descriptions of OTUs present in each sample were composed in USEARCH. Samples with fewer than 2500 read pairs were excluded from further analysis. Microbiome data were managed using the phyloseq package in R.[36,37]