

Genome - microbiome dynamics in human evolution and in emergence of diseases

Nicole Sónia Neto Pedro

D
2023



NICOLE SÓNIA NETO PEDRO

GENOME-MICROBIOME DYNAMICS IN HUMAN EVOLUTION AND IN EMERGENCE OF DISEASES

Tese de Candidatura ao grau de Doutor em Ciências
Biomédicas;
Programa Doutoral da Universidade do Porto
(Instituto de Ciências Biomédicas Abel Salazar)

Orientadora – Doutora Luísa Pereira

Categoria – Investigador

Afiliação – Instituto de Investigação e Inovação em
Saúde, Universidade do Porto; Instituto de Patologia
e Imunologia Molecular da Universidade do Porto,
Porto, Portugal.

Coorientador – Doutor François-Xavier Ricaut

Categoria – Investigador

Afiliação – Laboratoire Évolution and Diversité
Biologique (EDB UMR5174), Université de Toulouse
Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France

Coorientadora – Doutora Joana Pereira

Categoria – Investigador

Afiliação – Instituto de Investigação e Inovação em
Saúde, Universidade do Porto; Instituto de Patologia
e Imunologia Molecular da Universidade do Porto,
Porto, Portugal

Cotutora – Professora Doutora Ana Tomás

Categoria – Professor Associado

Afiliação – Instituto de Ciências Biomédicas Abel
Salazar da Universidade do Porto, Porto, Portugal

O trabalho apresentado nesta tese foi realizado no Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), no Instituto de Investigação e Inovação em Saúde da Universidade do Porto (i3S) e no Laboratoire Évolution and Diversité Biologique (EDB UMR5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France.



Este trabalho foi financiado através de uma bolsa de doutoramento (SFRH/BD/136299/2018) pela Fundação para a Ciência e a Tecnologia (FCT) e o Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) com o apoio da União Europeia (UE) e do Fundo Social Europeu (FSE), e pelo projecto PAPUAEVOL n° ANR-20-CE12-0003-01 financiado pela Agence Nationale de la Recherche, Ministério dos Negócios Estrangeiros e Europeus francês (Missão Pré-histórica Francesa na Papua Nova Guiné), e a Embaixada Francesa na Papua Nova Guiné.



DECLARAÇÃO DE HONRA

Declaro que a presente tese é de minha autoria e não foi utilizada previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referência. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Nicolle Pedro

PRECEITOS LEGAIS

According to the paragraph d) of the article 3 of the general regulations of the third cycle of studies of the University of Porto (alínea “d)” do artigo 3.º do regulamento Geral dos Terceiros Ciclos de estudos da Universidade do Porto), the author declares that this thesis used the results of the published articles and the in preparation articles indicated below. In compliance with the legislation, the author of this thesis declares that she actively participated in the conception, execution, analysis, and discussion of the results, as well as in the elaboration of the publications, under the name Pedro N. Paper I and Paper II do not require authorization by the publisher according to the permission rights displayed in their websites. Paper I, published by Springer Nature: <https://www.nature.com/nature-portfolio/reprints-and-permissions/permissions-requests>. Paper II, published by John Wiley & Sons: <https://www.wiley.com/en-us/network/publishing/research-publishing/trending-stories/how-to-clear-permissions-for-a-thesis-or-dissertation>.

Paper I. **Pedro N**, Brucato N, Fernandes V, André M, Saag L, Pomat W, Besse C, Boland A, Deleuze JF, Clarkson C, Sudoyo H, Metspalu M, Stoneking M, Cox MP, Leavesley M, Pereira L, Ricaut FX. 2020. Papuan mitochondrial genomes and the settlement of Sahul. *Journal of Human Genetics* 65(10):875 – 887.

Paper II. **Pedro N**, Brucato N, Cavadas B, Lisant V, Camacho R, Kinipi C, Leavesley M, Pereira L, Ricaut FX. 2022. First insight into oral microbiome diversity in Papua New Guineans reveals a specific regional signature. *Molecular Ecology*. 00:1 – 14.

Paper III. **Pedro N et al.** Host diversity versus environmental influence on the Papua New Guinean oral microbiome. (In preparation).

“Always and never are two words
you should always remember
never to use.”

- Wendell Johnson

~ XI ~

ACKNOWLEDGEMENTS

Esta tese é o resultado de vários anos de trabalho, e não teria sido possível sem a ajuda e apoio essencial de inúmeras pessoas. Agradeço a todos os que contribuíram das mais variadas formas para a sua conclusão.

Em primeiro lugar gostaria de agradecer à minha orientadora, Doutora Luísa Pereira por me ter ajudado a enveredar nesta etapa de crescimento profissional. A sua disponibilidade constante, confiança, ensinamentos e orientação foram muito importantes para me ajudar a desenvolver esta tese. Estou também grata pelos momentos de distração diários, pelas diversas histórias e o companheirismo destes últimos anos. Agradeço verdadeiramente esta oportunidade!

I want to thank my co-supervisor, Dr. François-Xavier Ricaut for the very warm welcome to Toulouse. I appreciate all the material and tools you provided that were indispensable for the realization and conclusion of this work. I truly appreciate all the guidance and dedication you showed, and how you always encouraged me to become more independent and pro-active.

Agradeço à minha coorientadora Doutora Joana Barbosa Pereira pela disponibilidade para ajudar e a amizade sempre demonstradas. Obrigada pela companhia nas pausas para café e pelos variados momentos de descontração.

Quero também agradecer ao Professor Doutor Eduardo Rocha por me ter aceite no Doutoramento em Ciências Biomédicas, à Professora Doutora Ana Tomás pela participação neste projecto como a minha tutora, ao Professor Doutor Sobrinho Simões por me acolher no IPATIMUP, e ao Professor Doutor Mário Barbosa e Professor Doutor Claudio Sunkel pela inclusão no i3S.

I would like to thank Dr. Nicolas Brucato, for all the great brainstorming, for your encouraging attitude and all the support and availability to help. I would also like to thank the EDB lab for hosting me in all my visits.

Ao grupo da Diversidade Genética, Ana, Carla, Fernanda, Susana, Ricardo, Rita e Verónica, agradeço toda a companhia e ajuda ao longo destes anos. Obrigada também às ex-GD, Joana, Marisa e Patrícia, por toda a amizade e companheirismo.

Finalmente, gostaria de agradecer às pessoas que me são mais próximas. Ao Bruno, por todo o apoio pessoal e profissional, e por todo o carinho e dedicação. Às minhas irmãs, por todos os momentos de convívio, de diversão e jogatinas; as “Manas chatas” que estão sempre lá para os bons e os maus momentos. Aos meus sobrinhos, que conseguem sempre animar a “Ticole” e fazem dela uma tia babada. E aos meus pais, que sempre me encorajaram a seguir os meus sonhos, e permitiram que eu chegasse a esta fase. A todos, o meu muito obrigada!

ABSTRACT

New Guineans and Aboriginal Australians are the descendants of the earliest modern human group(s) to leave Africa and are the oldest locally continuous populations found outside Africa. Papua New Guinea (PNG) has varied geography and topography that can be divided in the lowlands and the highlands in the main Island, and dozens of smaller islands. It also holds one of the highest genetic and linguistic diversity on the planet. Their high genetic and cultural diversity, different environments and high pathogenic pressure makes PNG a unique case study to look at biological evolution and adaptation in the human population.

Genetic and archeology evidence date the settlement of PNG around 55 thousand years ago (kya), and at that time, PNG and Australia were connected in a prehistoric continent, named Sahul. Further knowledge of the dispersal and population structuring following its settlement is scarce, so we analyzed a dataset of 915 complete mtDNA genomes from across PNG, which includes 379 mtDNA sequences from newly collected samples, to reconstruct the phylogenetic and phylogeographic history of northern Sahul. Our results support the hypothesis that two groups of settlers arrived in Sahul around the same time frame of 50-65 kya, and settled the Northern Sahul and the Southern Sahul separately. In spite of major climatic, cultural, and historical changes, limited dispersal was observed in Northern Sahul over time. After a 20 ky period of isolation, mtDNA lineages diversification showed greater exchanges within Northern Sahul and to Southern Sahul, Wallacea and Sunda. In the Holocene, the arrival of lineages from Island Southeast Asia stimulated the interactions of New Guinea populations with remote Oceania.

The 50 ky span of PNG populations in their unique environment created a valuable resource for the study of adaptive evolution in human population. The oral microbiota, as a boundary between the human body and the surrounding environment, can help address this subject. The oral microbiota is highly complex and can be colonized by around 700 different types of microorganisms including bacteria, fungi, viruses, archaea, and protozoa. Healthy individuals have an oral microbiota composed mainly by Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria and Proteobacteria, but

different populations can present different abundances of these taxa. We used a qualitative (microbiome array) and quantitative (non-human aligned reads from human whose genome sequences) approach to perform the first known characterization of the oral microbiome in 85 PNG individuals from different geographic locations. Within PNG, the oral microbiome showed significant geographical structure, especially in the less urbanized locations. In a global comparison, *Rothia* and *Neisseria* were found to be enriched in Papua New Guineans and Aboriginal Australians, suggesting a common oral microbiome signature.

The uncovered Sahul oral microbiome signature could be a result of a genetic and/or environmental adaptation. To assess the possible genetic impact in the oral microbiome, we performed a microbiome quantitative trait analysis (mbQTL) on 252 Papuans living in different environments. For the analysis of a potential environmental influence, we inferred the metabolic impact of the underlying microbial community through the abundance of bacterial pathways in the oral microbiome between a subset of PNG (n=30), Philippines (n=20) and Western (n=7) population groups. Six interactions between host genomic variants and four bacterial genera were identified, namely *Gemella* with the *BACE2* gene, and *Schaalia* with *ISG20* gene. The *Neisseria* and *Rothia* higher abundance could not be explained by association with host genetic factors, but the metabolic pathways allowed to confirm the statistically significant higher contribution of both *Neisseria* and *Rothia* in iron-related pathways in PNG versus *Veillonella* in Philippines and Western population groups. These results confirm that this specific oral microbiome profile in PNG might have been driven by iron deficiency, as both bacterial genera are associated to iron-related functions. More studies in isolated populations are important to fully uncover fitness advantages conferred to the human hosts in selective environments.

RESUMO

Os Nova Guineenses e os Australianos Aborígenas são descendentes dos primeiros grupos de humanos anatomicamente modernos a sair de África e são das mais antigas populações locais contínuas existentes fora de África. A Papua Nova Guiné (PNG) apresenta uma geografia e topografia complexas e podem ser dividida nas terras baixas e nas terras altas, onde albergam uma das maiores diversidades linguísticas e genéticas do planeta. A sua elevada diversidade genética e cultural, ambientes diferentes e alta pressão patogénica fazem da PNG um caso de estudo único para analisar a evolução e adaptação biológica da população humana.

As evidências genéticas e arqueológicas indicam que a colonização da PNG ocorreu há cerca de 55 mil anos e, nessa altura, a PNG e a Austrália estavam ligadas num só continente pré-histórico, chamado de Sahul. Pouco se sabe sobre a dispersão e estruturação da população após a sua colonização, pelo que analisámos um conjunto de 915 genomas completos de DNA mitocondrial (mtDNA) de toda a PNG, que inclui 379 novas sequências de mtDNA, para reconstruir a história filogenética e filogeográfica do norte do Sahul. Os nossos resultados apoiam a hipótese de que dois grupos de colonizadores chegaram ao Sahul no mesmo período de tempo (50-65 mil anos), e colonizaram o Norte e o Sul do Sahul separadamente. Apesar das grandes mudanças climáticas, culturais e históricas, observou-se uma dispersão muito limitada no Norte do Sahul ao longo do tempo. Após um período de isolamento de 20 mil anos, verificou-se uma diversificação das linhagens de mtDNA, com interacções dentro e fora do Norte do Sahul, para o Sul do Sahul, para a Wallacea e mais além. No Holoceno, a chegada de populações asiáticas vindas do sudeste asiático estimulou as interacções entre as populações da Nova Guiné com a as da Oceânia.

O isolamento das populações da PNG durante 50 mil anos neste ambiente único, levou à criação de um caso de estudo único para analisar a evolução adaptativa da população humana. Como interface entre o corpo humano e o ambiente exterior, a microbiota oral é uma excelente ferramenta para abordar estas questões. A microbiota oral é altamente complexa e diversificada, podendo ser colonizado por cerca de 700 tipos diferentes de microrganismos, incluindo bactérias, fungos, vírus, archaea, e

protozoários. Os indivíduos saudáveis têm uma microbiota oral composta principalmente por Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria e Proteobacteria, mas diferentes populações podem apresentar diferentes abundâncias destes taxa. Utilizamos uma abordagem qualitativa e quantitativa para realizar a primeira caracterização do microbioma oral em 85 indivíduos da PNG de diferentes localizações geográficas. Dentro da PNG, o microbioma oral apresentou elevada estruturação geográfica, especialmente nos locais menos urbanizados. Numa comparação global, os géneros *Rothia* e a *Neisseria* eram significativamente mais abundantes na PNG e nos Australianos Aborígenas, o que sugere um perfil do microbioma oral partilhado entre as populações do Sahul.

O perfil de microbioma oral descoberto no Sahul pode ser o resultado de uma adaptação genética e/ou ambiental. Para avaliar o possível impacto genético no microbioma oral, realizámos uma análise de *microbiome quantitative trait loci* (mbQTL) em 252 Nova Guineenses de diferentes locais geográficos. Para analisar uma potencial influência ambiental, inferimos a actividade metabólica da comunidade microbiana através da abundância das bactérias no microbioma oral, num subconjunto de indivíduos da PNG (n=30), Filipinas (n=20) e populações ocidentais (n=7). Foram identificadas seis interacções entre variantes genómicas do hospedeiro e quatro géneros bacterianos, nomeadamente *Gemella* com o gene *BACE2*, e *Schaalia* com o gene *ISG20*. A maior abundância de *Neisseria* e a *Rothia* detetada nas populações do Sahul não puderam ser explicadas por factores genéticos, mas as vias metabólicas permitiram confirmar que tanto a *Neisseria* como a *Rothia* apresentavam contribuições estatisticamente significativas em vias relacionadas com o ferro na PNG, enquanto que nas Filipinas e nas populações ocidentais a maior contribuição nessas vias é da *Veillonella*. Estes resultados confirmam que este perfil de microbioma oral específico da PNG pode ter sido motivado por deficiência de ferro, uma vez que ambos os géneros bacterianos apresentam funções relacionadas com o ferro. Estes resultados destacam a importância destes estudos em populações isoladas, importantes para revelar as adaptações conferidas aos humanos em ambientes selectivos.

TABLE OF CONTENTS

DECLARAÇÃO DE HONRA VII

PRECEITOS LEGAIS..... IX

ACKNOWLEDGEMENTS.....XIII

ABSTRACT XV

RESUMO XVII

TABLE OF CONTENTSXIX

LIST OF FIGURESXXI

LIST OF TABLES..... XXVIII

LIST OF ABBREVIATIONS XXX

I. INTRODUCTION..... 1

 1. *Human evolution* 3

 1.1. Human Evolutionary Genetics 3

 1.2. The Neutral theory of molecular evolution 7

 1.3. Natural selection 8

 1.3.1. Pathogen-driven natural selection 9

 1.3.2. Environment-driven natural selection 10

 1.3.3. Nutrient-driven natural selection 11

 2. *Human Microbiome* 12

 2.1. Microbiome technologies 14

 2.1.1. Microbiome microarray 15

 2.1.2. 16S rRNA gene sequencing 16

 2.1.3. Shotgun metagenomics 16

 2.2. Microbiome diversity and metabolism 17

 2.3. Microbiome and host interaction 19

 2.4. Microbiome of the oral cavity 21

 2.4.1. Factors that affect the oral microbiome 23

 2.4.1.1. Life stage and age 23

 2.4.1.2. Food consumption 24

 2.4.1.3. Lifestyle and personal habits 25

 2.4.1.4. Dental care 26

 2.4.1.5. Antibiotics 27

 2.4.1.6. Genetics 27

 3. *Case study: Papua New Guinea* 28

 3.1. The settlement of Sahul: Archeology studies 30

 3.2. The settlement of Sahul: Genetic studies 32

 3.3. Food habits 35

 3.4. Disease burden 36

3.5. Microbiome studies in PNG 38

II. AIMS..... 41

III. RESEARCH WORK..... 45

PAPER I. PAPUAN MITOCHONDRIAL GENOMES AND THE SETTLEMENT OF SAHUL 47

PAPER II. FIRST INSIGHT INTO ORAL MICROBIOME DIVERSITY IN PAPUA NEW GUINEANS REVEALS A SPECIFIC REGIONAL SIGNATURE..... 63

PAPER III. HOST DIVERSITY VERSUS ENVIRONMENTAL INFLUENCE ON THE PAPUA NEW GUINEAN ORAL MICROBIOME 79

IV. CONCLUSION..... 113

REFERENCES 123

APPENDICES 165

Appendix A - Supplementary information from Paper I 167

Appendix B - Supplementary information from Paper II 189

Appendix C - Supplementary information from Paper III 225

LIST OF FIGURES

Introduction

Figure 1. Map of Human mitochondrial DNA haplogroups, from their origin in Africa and colonization of Eurasia and the Americas. Figure adapted from Wallace (2015)..6

Figure 2. Schematic representation of the different types of natural selection. A) Purifying or negative selection, involving the removal of deleterious alleles from the population, B) Positive Selection, increasing the frequency of an advantageous allele that can lead to adaptation and C) Balancing Selection or heterozygote advantage where the presence of heterozygotes is favored in the population. Figure adapted from Quintana-Murci (2016)..... 9

Figure 3. The relative abundances of the six dominant bacterial phyla in microbiomes from: A) diverse human body parts in healthy state and (B) gut, comparing between health and disease for inflammatory bowel disease, type 2 diabetes, necrotizing enterocolitis and H. pylori infection. Figure adapted from Spor et al. (2011). 13

Figure 4. The timeline for microbial ecology research, from the first reports by Leeuwenhoek in 1676 to the development of NGS techniques. Figure adapted from Escobar-Zepeda et al. (2015). 15

Figure 5. Standard heritability versus Community heritability analysis. Figure adapted from van Opstal and Bordenstein (2015)..... 20

Figure 6. Representation of the diverse ecological niches in the oral microbiome and the major microbial components of the oral cavity. Figure adapted from Lee et al. (2021). 22

Figure 7. The dynamics of the oral microbiome from infancy to adult life, with emphasis on the factors that can modulate the oral microbiota. Figure adapted from Sedghi et al. (2021). 24

Figure 8. Geographic representation of the topography in Papua New Guinea with a color code for the elevation. Figure adapted from <https://www.worldatlas.com/maps/papua-new-guinea>. 29

Figure 9. Map of the Sunda, Wallacea and Sahul regions, with dark green representing the present-day geography and the light shading the extent of the continents at the LGM. The potential northern (pink line) and southern (blue line) routes for the settlement of Sahul (Bird et al. 2019). Estimated times of arrival based on archeological sites are also indicated for eastern New Guinea (Summerhayes et al. 2010), New Ireland (Leavesley et al. 2002), North Solomon islands (Wickler 1990) and the Madjedbebe (Clarkson et al. 2017). 32

Figure 10. Leading Causes of Death in Papua New Guinea (top 15) for the year of 2019, indicated in deaths per 100,000 population. Causes of death are grouped in non-communicable diseases (in blue), communicable, maternal, perinatal and nutritional conditions (in red) and injuries (in green). Data obtained from WHO (2020). 36

Paper I

Fig. 1 Genetic relationships and increasing population sizes across Sunda and Sahul. **a** Seven geographical regions: coastal PNG (light blue), highland PNG (dark blue), Near Oceania (pink), Remote Oceania (red), Australia (Orange), Wallacea (light green), and Sunda (dark green). **b** Tree of haplogroup P. Subclades are represented by triangles, while single lineages are represented by lines. The tree is scaled to kya (thousands of years ago) using the maximum likelihood molecular clock for the whole mtDNA genome with the mutation rate of Fu et al. (details of age estimates are reported in Table S1). **c** Tree of haplogroups M and Q. **d** Bayesian skyline plot representing median

estimates of effective population size for each the seven geographic regions based on P, M, and Q mtDNA lineages.52

Fig. 2 Spatial distribution of main Sahul mtDNA haplogroups. Inverse distance weighted interpolation shows areas with higher frequencies in darker shading, taking into account only the effects of distance (color scale in regions with few or no samples should not be considered as having high statistical confidence). Data details are provided in Table S5 and the triangles represent the central point for each region used in the interpolation. **a** Distribution of haplogroup M27. **b** Haplogroup M28. **c** Haplogroup M29. **d** Haplogroup P. **e** Haplogroup Q.53

Fig. 3 Proposed movements of maternal lineages P, M, and Q. Dark shading represents modern coastlines; light shading illustrates the extent of the Sunda continent at the Last Glacial Maximum. Red octagrams represent the probable approximate origins of haplogroups. Arrows represent probable migration paths during: **a** the initial settlement of Sahul (~50 kya; green); **b** the Last Glacial Maximum (~28 kya; blue); **c** the postglacial warming period through to the Holocene (~18 kya; pink); and **d** the Late Holocene (~3.5 kya; orange).58

Paper II

Figure 1 Schematic representation of the experimental strategy (sample preparation, sample analysis, data analysis and output) for the two microbial detection techniques.67

Figure 2 (a) Locations of the five sampled groups in PNG. (b) Box-plot representing the concordance in percentage of genera inferred in both methods by applying a threshold of abundance to the WGS results of 1%, 0.5%, 0.1% and 0.01% (Table S3). (c) Correlation of alpha-diversity values (Chao1 index) obtained for each individual oral microbiome characterized by both microbiome array and WGS methodologies. A regression line for the linear model (in blue) and the confidence interval (grey shadow) was added as well as the R and p-value for the model.68

Figure 3 Diversity of the oral microbiome samples using qualitative results (microbiome array). (a) Venn diagram representing the percentage of genus in common between the five groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and east highlanders living in Port Moresby (in pink); (b) alpha-diversity measure of the oral microbiome with Chao1 index between the different groups; statistical analysis was performed using pairwise t tests and correction with Holm's method (*p < .05; **p < .01); and (c) phylogenetic tree of genera found in the oral microbiome of the distinct groups, with emphasis in the genera significantly different. The values are the percentage of samples in each group in which that genus was detected with the MDA, and the size is proportional to that value. Statistical analysis was performed using chi-squared tests with multiple testing correction (*p < .05; **p < .01; ***p < .001).71

Figure 4 Profile of microbial composition of oral microbiome based on quantitative data (WGS results) with the taxa discriminating between the five distinct groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and east highlanders living in Port Moresby (in pink). We applied the LEfSe algorithm and LDA calculation in a “one-against-all” multi-class analysis, requiring only one pairwise Wilcoxon tests to be significant. Only discriminating features with p < .05 and LDA score over 3 are presented.72

Figure 5 Comparison of oral microbiome profiles between PNG and other populations based on quantitative for review only data (WGS results). (a) Unrooted neighbour-joining tree of the mean weighted UniFrac distances between the five PNG groups (Port Moresby, Simbu province, East Sepik province, Western province and east highlanders living in Port Moresby), the Philippines hunter-gatherers (HG) and traditional farmers (TF) and the Western controls (HMP, 2012; Lassalle et al., 2018). (b) Heatmap for abundances of the top 11 genera detected with quantitative results in the eight groups also present in the unrooted neighbour-joining tree (>1% of abundance in all samples) and other published populations of aboriginal Australian children and Batwa pygmies (Handsley-Davis et al., 2021; Nasidze et al., 2011).74

Paper III

Figure 1. The 20 significant miQTLs, corresponding to six interactions between host genomic regions/variants and four bacterial genera of the oral microbiome.91

Figure 2. Relative abundance values per genotype and worldwide map with allele frequencies. A) *Gemella* and rs7283882 SNP located in the chromosome 21; B) *Schaalia* and rs7182861 SNP located in chromosome 15; C) *Klebsiella* and rs11908491 SNP located in chromosome 20; D) *Salmonella* and rs4315878 SNP located in chromosome 5.94

Figure 3. The impact of *Gemella* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Philippines group. B) Krona plot with the significantly different pathways between PNG and the Western controls. C) Bar-plot with abundance per individual samples of the several bacterial genera that contribute to the Glycolysis IV pathway, grouped by PNG (blue), Philippines (orange) and Western (green) and D) Boxplot for the abundance of *Gemella* in the three groups, PNG (blue), Philippines (orange) and Western (green).96

Figure 4. The impact of *Neisseria* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Philippines group. B) Krona plot with the significantly different pathways between PNG and the Western controls. C) Bar-plot with abundance of per individual samples of the several bacterial genera that contribute to the superpathway of heme b biosynthesis from glutamate, grouped by PNG (blue), Philippines (orange) and Western (green).98

Figure 5. The impact of *Rothia* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Western controls. B) Bar-plot with abundance of per individual samples of the several bacterial genera that contribute to the heme b biosynthesis pathway, grouped by PNG (blue), Philippines (orange) and Western (green). C) Schematic representation of the eight enzymatic steps (arrows) of heme biosynthesis (enumerated next to the reaction) and the subsequent erythropoiesis and hemolysis in the vascular system. Figure adapted from Heckl et al. (2021).99

Appendix A

Figure S1. Maximum-Parsimony Phylogenetic Tree of M27 Mitogenomes.
..... <https://cloud.i3s.up.pt/index.php/s/xEHBWERJDJpmFF4>

Figure S2. Maximum-Parsimony Phylogenetic Tree of M28 Mitogenomes.
..... <https://cloud.i3s.up.pt/index.php/s/RKwZeZoZJSDcDQ3>

Figure S3. Maximum-Parsimony Phylogenetic Tree of M29'Q Mitogenomes.
..... <https://cloud.i3s.up.pt/index.php/s/kH4dPWgriJJxMG7>

Figure S4. Maximum-Parsimony Phylogenetic Tree of P Mitogenomes.
..... <https://cloud.i3s.up.pt/index.php/s/YDXmTtMSjaNeWGB>

Appendix B

Figure S1. Alpha-diversity plot for all samples with Chao1 index. Sample PNG54 from Port Moresby (in red) was removed as an outlier for Alpha diversity (Chao1 = 903). Sample is considered an outlier if value falls out of the mean +/- 3SD (637) for all samples.191

Figure S2. Correlation between alpha-diversity and (A) Age and (B) Gender of the individuals (F=Female and M=Male). Pearson correlation was performed for the Age correlation, obtaining a p-value of 0.36, and a t-test was performed in the alpha-diversity between Female and Male (p-value = 0.4).192

Figure S3. Profile of microbial composition of samples with phylogenetic tree of all genera found in the oral microbiome of the different groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and East highlanders living in Port Moresby (in pink).193

Figure S4. Beta diversity and microbiome profiling of the different groups. A nMDS plot and PERMANOVA analysis for the (A) Jaccard distance and the (C) unweighted UniFrac distance between the five tested groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and East highlanders living in Port Moresby (in pink). Unrooted neighbor-joining trees of the mean (B) Jaccard distance and (D) unweighted UniFrac distances between the 5 groups. Stress value for the NMDS in Jaccard is 0.2182 and for the unweighted UniFrac is 0.2187.194

Appendix C

Figure S1. Correlation between bacterial genera abundances inferred through Bracken 2 and MetaPhlAn algorithms, in a) the total 57 samples from PNG, Philippines and Western (in red); B) the samples from PNG (in green); C) the samples from the Philippines (in blue); and D) the samples from the Western population (in brown).....227

LIST OF TABLES

Appendix A

Table S1. List of 379 new Papua New Guinean complete mitogenome sequences considered in this Study. <https://cloud.i3s.up.pt/index.php/s/nkokm5Lmf38xWLj>

Table S2. List of all 539 published complete mitogenome sequences considered in this Study. <https://cloud.i3s.up.pt/index.php/s/gtQ6b6SPkHYmQKw>

Table S3. Total number of newly (n=379) and published (n=539) mitogenomes used in this study.169

Table S4. TMRCA age estimates using ML for all M, P and Q Papuan subclades. ...169

Table S5. Region frequencies of haplogroups M27, M28, M29, Q and P.185

Appendix B

Table S1. Demographic information (Age (A), Gender (G), Origin for both parents, Sampling place) for all the individuals in this study, indication of which samples we obtained Array microbiome results and WGS microbiome results. For these last, we also added the number of unmapped reads used for microbiome inference and the sequencing depth of the WGS.195

Table S2. Frequency of all the genera detected in all samples with the WGS microbiome results and the Array microbiome results.199

Table S3. Comparison of the genera detected in both methods per sample by applying a threshold of abundance to the WGS results of 0.01%, 0.1%, 0.5% and 1%.219

Table S4. Qualitative results indicating the number of samples for each group in which the different genus were detected.221

Table S5. Relative abundance of each of the top Phyla and Genera for the five groups in this study (Port Moresby, Simbu province, East Sepik province, Western province, East highlanders living in Port Moresby) and in comparison with other published populations (Australian Aboriginal Children, Batwa Pygmies, Philipines HG, Philippine TF, Western controls).223

Appendix C

Table S1. Demographic information (Age, Gender, Province and Region of origin of the father, Sampling place) for all the individuals in this study, and the number of unmapped reads used for microbiome inference from the WGS.227

Table S2. Statistically significant ($p\text{-value} < 5 \times 10^{-08}$) mbQTLs in the PNG groups. For each SNP, information is provided for the reference (R) and alternative (A) alleles, the associated genus, genomic region (gene or intergenic region), the p -value and the genotype frequencies for European, East Asian and African population groups obtained from the 1000 genomes data and for PNG calculated from the genetic data of this study.236

Table S3. Table S3. Statistically significant ($q\text{value} < 0.05$) results of the bacterial pathways analysis between PNG and the Philippines and Western population groups. Each significant association is related to the contribution of one taxon to one bacterial pathway. For each pathway is provided information for the Superclass, Class and Feature. <https://cloud.i3s.up.pt/index.php/s/GPwsmz43PzZYkET>

LIST OF ABBREVIATIONS

ACE	Abundance-based coverage estimators
AIDS	Acquired immunodeficiency syndrome
CRS	Cambridge reference sequence
DNA	Deoxyribonucleic acid
Gb	Gigabase
GWAS	Genome wide association studies
HBV	Hepatitis B virus
HIF	Hypoxia inducible factor
HIV	Human immunodeficiency virus
HPV	Human papilloma virus
HWE	Hardy-Weinberg equilibrium
Kya	Thousand years ago
LGM	Last Glacial Maximum
LP	Lactase persistence
mbQTL	Microbiome Quantitative trait locus
mtDNA	Mitochondrial DNA
NCD	Non-communicable disease
nDNA	Nuclear DNA
Ne	Effective population size
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
PNG	Papua New Guinea
rCRS	Revised Cambridge reference sequence
rRNA	Ribosomal RNA
SNP	Single nucleotide polymorphisms
TNG	Trans New Guinea
UTR	Untranslated region
WES	Whole exome sequencing
WGS	Whole genome sequencing

I. INTRODUCTION

1. Human evolution

The study of human evolution is the result of cooperation between several research fields, such as linguistics, archaeology and genetics (Cavalli-Sforza et al. 1988). In the last two decades, this last field in particular has made significant contributions in this area due to the development of new and affordable technologies and algorithms. Genetics focuses on the study of genes and heredity in organisms, which is encoded in an acid nucleic macromolecule, named deoxyribonucleic acid (DNA). The set of all the genetic information of an organism is called the genome. Human evolutionary genetics looks into the genetic variation that leads to changes in the human genomes along generations and in the genomes of related species, allowing to make genetic inferences on human evolution (Jobling et al. 2014). This area has recently achieved worldwide attention when the 2022 Nobel Prize in Physiology or Medicine was attributed to Svante Pääbo, a geneticist specialized in the field of evolutionary genetics, for his discoveries concerning the genomes of extinct hominins and human evolution (NobelPrize.org 2022). Svante Pääbo's more notorious research was the sequencing of the Neanderthal genome (Prüfer et al. 2014), and the discovery of a previously unknown hominin, Denisova (Reich et al. 2010).

1.1. Human Evolutionary Genetics

Human population genetics aims to compare the genetic profile of different populations, analyze selected variants or alleles, and infer the level of admixture from other populations (Hellenthal et al. 2014). The human genome is mainly composed of the nuclear DNA (nDNA), with around 3.2 gigabases (Gb) duplicated, organized as 23 pairs of chromosomes (22 autosomal and one sex-specific). About 99% of the human genome is composed of non-coding regions, that are beginning to gain attention as evidence shows that some may play functional roles, being highly conserved and under strong negative selection (Hardison 2000). The human nuclear genome contains around 20,000 genes, a number much lower than initially expected given the derived position the human species occupies in the tree of life. However, the human genes are highly complex and undergo more alternative splicing than other mammals (Kim et al. 2007), which translates into over 250,000 peptides (Ezkurdia et al. 2014).

The field of human population genetics improved greatly with the increased level of resolution of next-generation sequencing (NGS) technologies that enable large-scale sequencing and genotyping of the genome. Although expensive in the beginning, the continuous improvement of the high-throughput sequencing technologies has led to lower costs, with a range of solutions for the genotyping of single nucleotide polymorphisms (SNPs) in the human genome (Maroti et al. 2018). Whole genome sequencing (WGS), as its name indicates, is the technique that determines all the nucleotides within a genome. It is the most expensive sequencing technique available and the one that provides complete information. The sequencing of the first complete human genome in 2003 by the Human Genome Project (HGP 2004) with a cost of around 1 billion US dollars, covered about 92% of the human genome. Only in 2022 the remaining 8% of the human genome were sequenced by the Telomere-to-Telomere Consortium (Nurk et al. 2022). Nowadays, WGS with high coverage (30X) can get a price as low as 500€, and low coverage (1X) can cost 80€. By combining these two approaches, it is possible to create a reference panel at a 30X coverage and use it to impute the missing information in the 1X coverage samples, greatly reducing the costs of WGS (Lou et al. 2021; Pasaniuc et al. 2012). For the study of adaptive history of populations, WGS is the best approach, and its successive improvements, namely long-read sequencing, will soon allow to uncover the full spectrum of human genetic variation, resolving some of the missing heritability (Logsdon et al. 2020). An alternative to WGS is whole exome sequencing (WES), which reports all exons of protein-coding genes and covers approximately 1% of the human genome. It is constituted of roughly 180,000 exons and about 30,000,000 base pairs, allowing for significant insights into human genetic variation in the region with direct functional impact in protein structures (Ng et al. 2009). Sometimes, the exon information is added up with the adjacent untranslated regions (UTR), as these are also functionally important, containing many gene expression regulatory elements (Ferreira et al. 2022). Albeit being limited to the variation found in the coding regions of the genomes, the lower cost of WES allows for an increase in the sample size of the study compared to WGS. In the context of medical studies, sample size and the interpretability of functional impact may be critical to achieving meaningful success in the detection of relevant rare variants (Bomba et al. 2017). The availability of several human exomes and complete genomes allowed to identify millions of variable

positions along the chromosomes, information that has been used to design SNP microarrays (Maroti et al. 2018). These are one of the fastest and cheapest high-throughput techniques that focuses on genotyping specific bases (which are polymorphic in the human species) throughout the entire genome. It is usually composed of relevant and common variants that are population-specific or disease-associated and have been widely used in genome-wide association studies (GWAS) comparing case and control cohorts (International HapMap 2003; Page et al. 2015).

The human genome also comprises the mitochondrial DNA (mtDNA), a double stranded circular molecule with around 16,569bp that can be found in the mitochondria. With few exceptions (Luo et al. 2018), mtDNA has the particularity of being maternally transmitted, which means that it does not undergo recombination and is inherited as a haplotype or lineage. In 1981, Anderson et al. (1981) reported the first complete human mtDNA sequence, also known as the Cambridge Reference Sequence (CRS), that was later revised in 1999 (and known as rCRS (Andrews et al. 1999)). The mtDNA mutates 5 to 10 times faster than the nDNA, with more mutations occurring in the control region, especially the hypervariable regions I and II, and less in the coding region (Melvin and Ballard 2017). This higher mtDNA mutation rate can be explained by (Birky 2001): a mtDNA repair system more prone to errors; a stronger mutagenic environment, due to high concentrations of reactive oxygen species within mitochondria; and it undergoes more replications per unit of time which can generate more mtDNA errors. The lack of recombination, high mutation rate and high turnover makes mtDNA a marker of choice in the study of human evolutionary history (Ingman et al. 2000). Phylogenetic methods allow to infer the way the mtDNA haplotypes derive from a common ancestor, which will create a haplogroup, and with constant diversity accumulation over time, sub-haplogroups will start to appear. The human mtDNA tree (Torroni et al. 2006) includes all haplogroups and sub-haplogroups of the human population, reflecting the evolution history of humans (Figure 1).

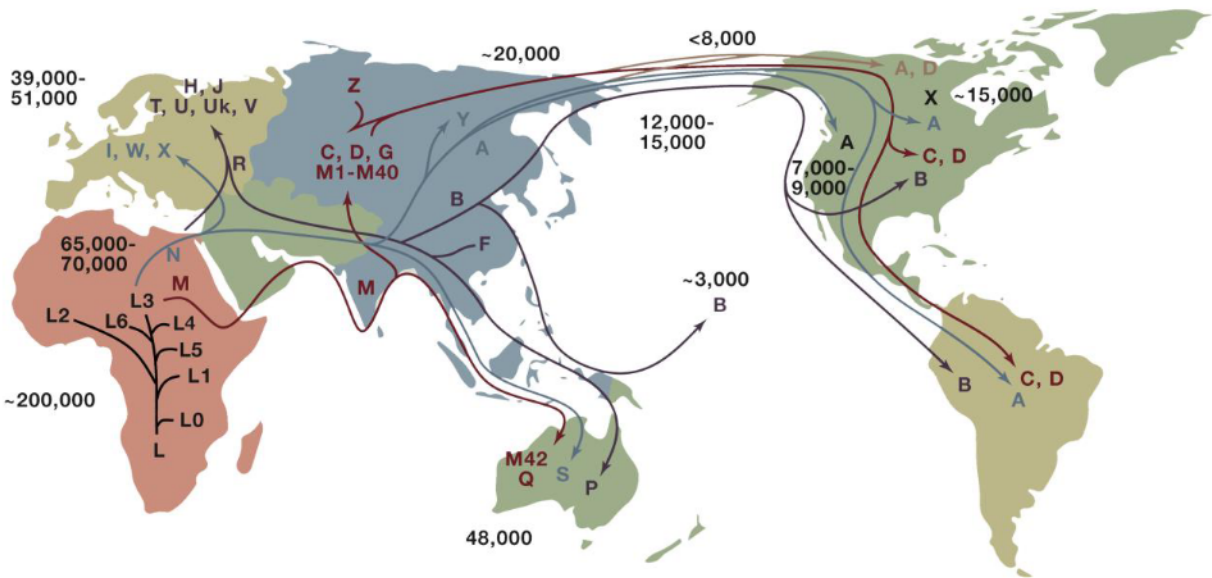


Figure 1. Map of Human mitochondrial DNA haplogroups, from their origin in Africa and colonization of Eurasia and the Americas. Figure adapted from Wallace (2015).

After emerging 200 thousand years ago (kya) in Africa, the first anatomically modern humans populated the rest of the world in subsequent migrations. First, the sub-Saharan spread-out (Rito et al. 2013) that generated a grand group of African haplogroups, namely L0, L1, L2, L3, L4, L5 and L6, referred to as macrohaplogroup L. Then, the out of Africa migration occurred approximately 70-65 kya (Soares et al. 2012), when few individuals affiliated in the L3 haplogroup successfully left Africa and originated two descendent macrohaplogroups, M and N, that then colonized the rest of the world. The East Asian and European settlement began around 50 and 45 kya, respectively, with the macrohaplogroup N dispersing throughout the globe and evolving into multiple European, Asian, and Native American mtDNA lineages, while macrohaplogroup M dispersal was mostly confined to Asia and descendent Native American haplogroups (Wallace 2015) the exception being M1 haplogroup, which migrated back-to-Africa, being thus present in East and North Africa (Pennarun et al. 2012). Island Southeast Asia and Melanesia were colonized by modern humans at least ~50,000 years ago (Barker et al. 2007; Soares et al. 2009), by descendants of the M and N macrohaplogroups, with M29'Q, S and P lineages exclusively found in Australia and New Guinea Islands (Hudjashov et al. 2007). Ultimately, the migration to North

Asia by the haplogroups A, C, and D took place 20 kya and then to the American continent by 15 kya (deMenocal and Stringer 2016).

1.2. The Neutral theory of molecular evolution

The study of human evolution and the reconstruction of past demographic events relies on the analysis of genetic diversity in populations. Several factors can shape population genetic diversity such as genomic forces, demographic forces, and natural selection (Jobling et al. 2014). All these forces together have an impact on the levels of genetic diversity which, in turn, can have an impact on phenotypic variation, both within the population or between different human populations.

Mutations are crucial evolutionary forces that allow species biological evolution. They can arise by two processes, either because of physical or chemical damage (for example, ultra-violet radiation) to the DNA molecule or because of errors occurring during DNA replication (Jobling et al. 2014). If the error occurs in a somatic tissue, it will not be transmitted to the next generation, but if the error occurs in a germinal tissue (a cell developing into an egg or a sperm), it may be transmitted to the offspring and contribute to species evolution. Once these mutations occur, they can undergo genetic recombination, the reshuffling of the paternal and maternal genetic material, allowing new combinations of alleles to emerge (Armour 2009).

The Neutral theory of molecular evolution states that most of the mutations do not change our fitness, so their evolution in time is just driven by randomness, also known as genetic drift. Mutations will drift away with time and be lost in the population, or they can increase in frequency and eventually be fixed. It occurs in all populations of non-infinite size, but the magnitude of the genetic drift relates to the size of the population being sampled (Kimura 1971). The Hardy-Weinberg equilibrium (HWE) assumes that populations are constant in size, but that is not usually the case. In fact, when the effective population size (N_e ; the number of individuals in the population involved in producing the next generation) is sharply reduced in size, genetic drift can have major effects on that population (Wang, Santiago, et al. 2016). The drastic reduction can occur due to the bottleneck effect, where a major event leads to the

reduction in population size, or due to the founder effect, when a small group splits off from the main population to form a colony. In both cases the surviving/founding individuals may not represent the full genetic diversity of the original population. During these reductions, drift acts in an intense way, with loss in genetic diversity, until the population regains a sufficiently large size for drift to be buffered (Jobling et al. 2014).

1.3. Natural selection

Some of these mutations can have a functional effect, by for instance changing a protein or the levels of gene expression, and have an impact on the phenotype of the individual. In this case, these mutations can be targeted by natural selection because they have conferred a selective disadvantage or advantage to human populations (Quintana-Murci 2016). Purifying or negative selection (Figure 2A) involves the removal of deleterious alleles from the population. It is the most widespread type of selection throughout the human genome, easily inferred from the higher proportion of non-synonymous variants when compared to synonymous variants in the younger than older branches of the mtDNA mammalian trees (Soares et al. 2013) and in the nDNA (Albers and McVean 2020). Positive selection (Figure 2B) increases the frequency of an advantageous allele and can lead to adaptation. In this case, new mutations that occurred in the past will rapidly increase in frequency until they reach fixation. Balancing selection (Figure 2C) occurs when the allele is only advantageous at the heterozygous state. The advantageous allele will not be fixed or lost; it will stay at intermediate frequency in the population (Quintana-Murci 2016).

Genome-wide evaluations of selection in humans identified two main motors: pathogens and metabolites (Karlsson et al. 2014), which in many cases are interconnected (e.g., viruses that use lipids to enter cells; (Sierra et al. 2017)). A better knowledge on how populations adapt to nutrients/pathogen/environment pressures over time will reveal genetic variations coding advantageous phenotypic traits, which may have been crucial for human past and present survival (Barreiro and Quintana-Murci 2010; Ko et al. 2013).

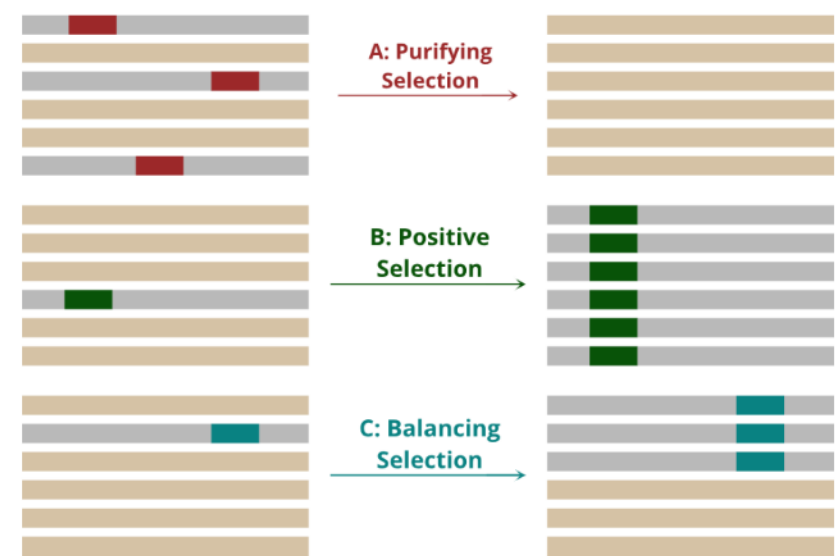


Figure 2. Schematic representation of the different types of natural selection. A) Purifying or negative selection, involving the removal of deleterious alleles from the population, B) Positive Selection, increasing the frequency of an advantageous allele that can lead to adaptation and C) Balancing Selection or heterozygote advantage where the presence of heterozygotes is favored in the population. Figure adapted from Quintana-Murci (2016).

1.3.1. Pathogen-driven natural selection

Examples of pathogen-driven selection are abundant (Karlsson et al. 2014), with a classic example of balancing selection consisting in deleterious mutations at *β-globin* (*HBB*), *glucose-6-phosphate dehydrogenase* and *glycophorin A, B and C* genes. These mutations trigger the development of sickle cell anemia, glucose-6-phosphate dehydrogenase deficiency and thalassemia when in homozygosity, respectively, however in the heterozygous state, they confer protection against malaria infection (Barreiro and Quintana-Murci 2010). *Plasmodium falciparum* invades the red blood cells to breed and release new parasites capable of infecting other red blood cells (Luzzatto 2012). When the parasite invades the abnormal red blood cells it triggers them to sickle, and they are removed by the immune system before the parasite can complete its life cycle and infect other red blood cells, leading to less parasites and milder illness. Most commonly, this protection is caused by three different altering

variants of β -globin (*HBB*), each conferring different levels of resistance against malaria in different regions: the HbS allele displays the highest frequencies in equatorial Africa, Cameroon, Guinea, Uganda, and Kenya, as well as in parts of the Middle East, Saudi Arabia, and India (Flint et al. 1998); HbC is primarily observed in Africa, but more restricted to West African groups (Tishkoff et al. 2001); and the HbE allele is mainly found in Asia, from northern India to China and South East Asia (Chotivanich et al. 2002).

The *Apolipoprotein L1* locus is another example of selection due to pathogen adaptation in Africans. Two variants (G1 and G2) modify the protein, enabling it to lyse the *African trypanosomes* parasite, and conferring protection to the African sleeping sickness (Ko et al. 2013). However, these variants are also associated with high levels of hypertension and with chronic kidney disease. As such, these deleterious alleles do not disappear in African populations given the positive selection of the heterozygous individuals due to the local pathogens.

Recently, and in light of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic, a study involving East Asian populations uncovered several selection events in human genes, more specifically genes encoding for virus-interacting proteins, including SARS-CoV-2, that likely started more than 20,000 years ago (Souilmi et al. 2021). This signal was also detected for genetic variants of the *Zinc Finger DHHC-Type Palmitoyltransferase 20 (ZDHHC20)* gene in New Guineans and Bismarck Islanders (Brucato et al. 2022). These adaptive events appear to be limited to East Asian populations, overlapping with the fact that East Asia is the geographical region where several modern coronavirus epidemics have emerged (Souilmi et al. 2021).

1.3.2. Environment-driven natural selection

Throughout early migrations, modern humans encountered a variety of habitats, from tropical to arctic, from high altitudes to lowlands, and even toxic environments, leading to adaptation to these new environmental changes (Fan et al. 2016). A variant in fatty acid desaturase enzymes allowed the Inuit populations from arctic

environments (Alaska, Canada, Greenland) to adjust to a new and necessary marine-based diet, enriched in omega-3 polyunsaturated fatty acids (Fumagalli et al. 2015; Tishkoff 2015). Signatures of positive selection in genes involved in the hypoxia-inducible factor (HIF) pathway were uncovered in the genomes of populations living in hypoxic high-altitude regions in Asia and South America, such as the Tibetan Plateau and the Andean Altiplano (Fan et al. 2016). In the highlanders from Papua New Guinea (PNG), phenotypic differences were found when compared with lowlanders, such as larger lung volume, chest depth and higher hemoglobin concentration (André et al. 2021). These may hint to altitude adaptation in PNG highlanders, although genetic analyses are needed to confirm the existence and uncover which genes are under positive selection. Hypoxia tolerance was not only investigated in terms of high-altitude adaptation, but also in breath-hold diving (Ilardo et al. 2018). In Southeast Asia, the Bajau people, also known as “Sea Nomads”, are renowned for their extraordinary breath-holding abilities due to their marine hunter-gatherer form of subsistence. Ilardo et al. (2018) showed that these breath-holding abilities are due to natural selection on genetic variants in the *phosphodiesterase 10A* gene, which increases spleen size and provide them with a larger reservoir of oxygenated red blood cells, and on the *Bradykinin Receptor B2* gene that affects the human diving reflex.

1.3.3. Nutrient-driven natural selection

In the Neolithic revolution, around 10 kya, there was a radical shift in the human form of subsistence: a meat-based diet from the hunter-gatherer lifestyle was substituted by an agriculture-based diet, with the incorporation of cereals/fibers and dairy products (Bocquet-Appel 2011). This transition to primary food production was marked by the appearance of social stratification, with increasing population numbers and density, sedentary lifestyle, the construction of cities and creation of communities. A very common food adaptation due to selection is the adult lactase persistence (LP) in Europeans and East Africans (Gerbault et al. 2011), occurring in the *Lactase-Minichromosome Maintenance Complex Component 6* (*LCT-MCM6*) locus. Despite the same functional effect, the protective variant for European populations is distinct from the variants present in the East African, which indicates independent and convergent selection for the adaptations to milk culture in the two continents (Enattah et al. 2008).

In Asia, small percentages of LP are found even in known herder populations. Heyer et al. (2011) analyzed the Kazakh population (traditionally herders) and the Tajiko-Uzbek population (agriculturalists) from Central Asia and found that LP in less than 30% and Gallego Romero et al. (2012) found a prevalence of LP around 45% in herders from South Asia. In both cases, LP was associated with the European variant, most likely related to a recurrent gene flow from Europe (Segurel et al. 2020). These results challenge the understanding of this adaptive story because milk consumption does not always correlate with LP frequency: traditional herders from Mongolia and Central Asia present a low frequency of LP even though they consume milk, and Hadza and Yaaku hunter-gatherers from East Africa do not consume milk but have a high frequency of LP (Segurel et al. 2020). It is not entirely clear what could cause these differences but it is more likely due to a combination of cultural, nutritional, and environmental factors, such as milk fermentation prior to consumption that reduce the amount of lactose, the type of livestock and the actual proportion of milk consumed in their diet (Evershed et al. 2022; Segurel et al. 2020).

2. Human Microbiome

The human body is colonized by trillions of microbial cells, namely bacteria, fungi, and viruses (Sender et al. 2016), which collectively are referred to as the microbiota, while the genetic makeup of these microbial cells is termed the microbiome. In contrast to the human genome, which is not as easily influenced by external factors, the human microbiome is a highly dynamic ecosystem with body-site-specific diversity, including but not limited to the skin, the oral cavity, the esophagus tract, the colon, the vagina, and semen (Blum 2017) (Figure 3). In addition to its intra-individuality, the microbiota also fluctuates throughout life and coexists with its host, being an active part of many biological processes in the human body (Blum 2017). The human microbiome provides a fundamental internal ecosystem for physiological processes, such as protection against pathogen infection, nutrient processing, stimulation of angiogenesis, development and maintenance of the intestinal epithelial barrier, and maintenance of the immune system, among others, thus evolving together with the host to form a superorganism (Cho and Blaser 2012). This involvement of the microbiome in many biological functions renders essential to maintain a homeostatic

balance, as a compromised state can bring problems to the host, such as the appearance of several diseases (Gilbert et al. 2018), as inflammatory bowel disease, type 2 diabetes mellitus and necrotizing enterocolitis (Spor et al. 2011).

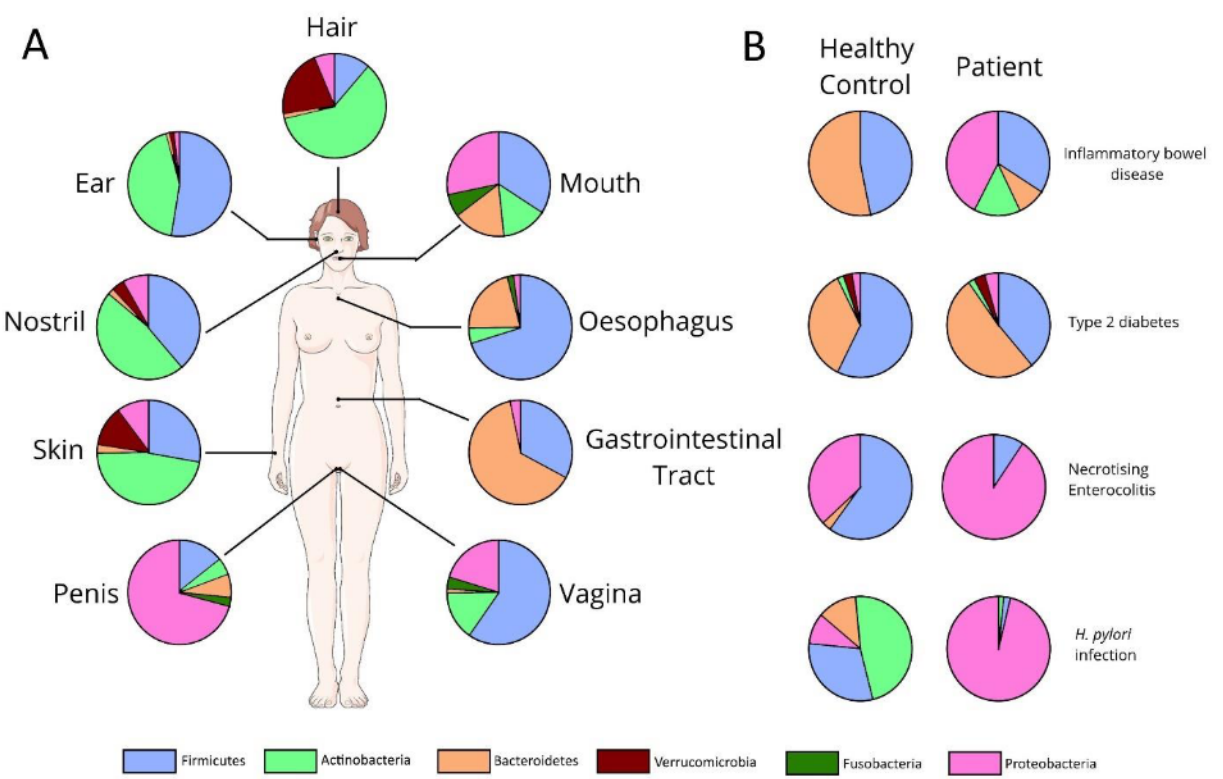


Figure 3. The relative abundances of the six dominant bacterial phyla in microbiomes from: A) diverse human body parts in healthy state and (B) gut, comparing between health and disease for inflammatory bowel disease, type 2 diabetes, necrotizing enterocolitis and *H. pylori* infection. Figure adapted from Spor et al. (2011).

Innumerable external factors can shape the microbiome, namely the use of antibiotics or probiotics, changes in diet, and states of infection in the host organism (Gilbert et al. 2018). For example, diet is a profound conditioner of the gut microbial community structure (David et al. 2014), while vigorous scrubbing can briefly alter the skin microbiome (Grice and Segre 2011). Similarly, the oral microbiome is influenced by prolonged smoking habits, which promote an anaerobic environment in the respiratory tract (Wu et al. 2016). Regarding the effects of the microbiome on

health, Deshpande et al. (2018) associated the presence of gram-positive bacteria in the esophagus with good health, while the dominance of gram-negative bacteria was associated with early stages of esophageal adenocarcinoma. Additionally, a well-known indicator of gut dysbiosis (disruption or imbalance of the microbiota) is the change in the Firmicutes/Bacteroidetes ratio (Stojanov et al. 2020). The understanding of how the microbiome varies within a person or between people over time will elucidate the mechanisms by which the microbiome influences the preservation of health or the onset and progression of disease (Gilbert et al. 2018).

2.1. Microbiome technologies

The study of microbial communities has evolved dramatically since Leeuwenhoek first reported the presence of oral microorganisms in 1676 (Davis 2020), to the full characterization of the microbiome using NGS techniques (Figure 4). Early studies of host-associated bacteria were dependent on the isolation and growth of bacterial colonies with culture medium methods. These methods enabled scientists to understand the microorganisms' physiologies (Blevins and Bronze 2010) but were limited in the proportion of microorganisms that could be cultivated and, as such, a large portion of the microbiota remained unidentified. The study of microorganisms progressed with the help of microscopy and the development of staining techniques, increasing its resolution (Beveridge 2001; Blevins and Bronze 2010).

It soon became evident that the number of observed microorganisms was much higher than the number of microorganisms obtained with culture methods. Based on this, Winogradsky concluded that the microorganisms needed specific conditions to grow, and revolutionized microbial ecology studies by creating several environments for culture media production (Escobar-Zepeda et al. 2015). The study of microorganisms based solely on their morphology features, growth, and selection of some biochemical profiles (Oliver et al. 1991; Roszak et al. 1984) was limited but remained the only method of microbial research for 300 years. Only in 1977 did Carl Woese proposed the use of small subunit ribosomal RNA genes (rRNA) as markers for taxonomy (Woese and Fox 1977) and together with the Sanger automated sequencing (Sanger et al. 1977), a powerful tool for assessing microbial diversity in the human microbiota was

created. In 1990, Giovannoni et al. (1990), performed the first microbial diversity study using 16s rRNA genes, and led the way to the beginning of microbiome studies, which were later revolutionized by the metagenomics era. Microbiomic studies have greatly improved with the creation of The Human Microbiome Project (Human Microbiome Project 2012), the first worldwide and body-wide metagenomic microbiome survey of 300 healthy adult donors, covering the microbiome of 18 body sites. Nowadays, different techniques can be used for microbial diversity studies, but the microbiome microarray, the 16s rRNA gene and metagenomics will be further explained.

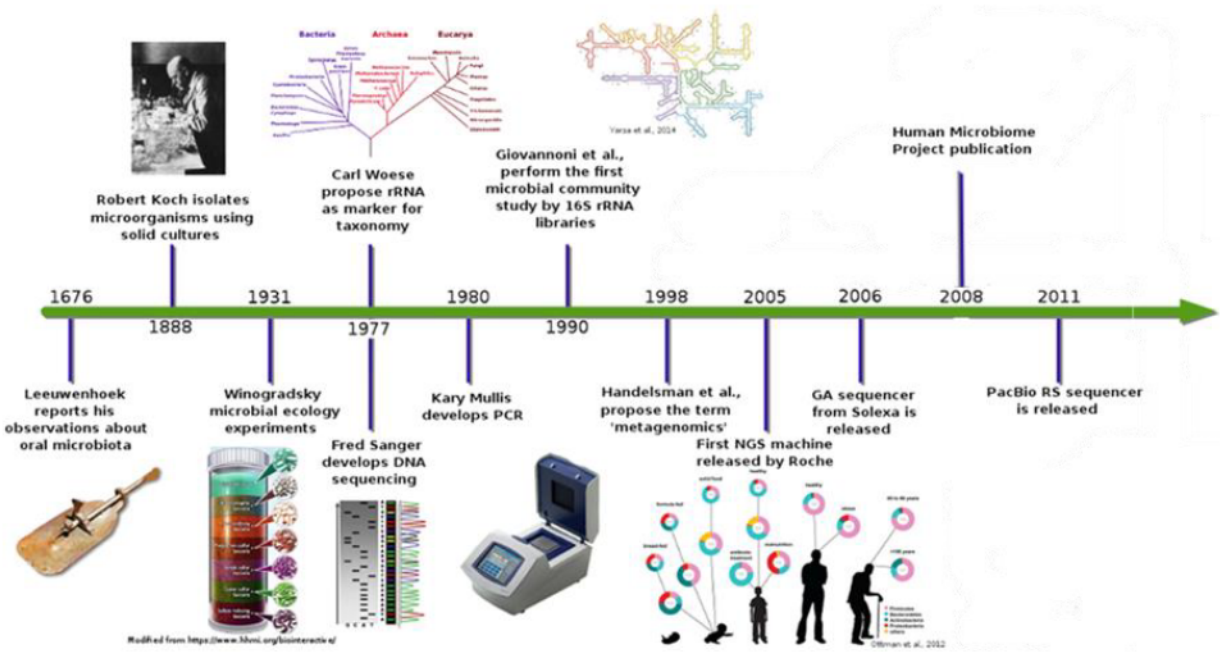


Figure 4. The timeline for microbial ecology research, from the first reports by Leeuwenhoek in 1676 to the development of NGS techniques. Figure adapted from Escobar-Zepeda et al. (2015).

2.1.1. Microbiome microarray

Microbiome diversity characterization can be performed with microbiome microarrays, which can rapidly identify the microbial community at the species or strain level, giving a profile of which microorganisms are present in the analyzed samples. One example is the Thermo Fisher Scientific Axiom™ Microbiome Array (MDA; from Thermo Fisher

Scientific, Waltham, MA, USA) that harbors a total of 1.38 million probes to detect over 12,513 microbial species including archaea, bacteria, fungi, protozoa, and viruses (Thissen et al. 2019). Probes are designed against conserved regions (for the detection of higher taxonomic ranks) and unique regions (species/strain specific) of the microbial genomes, with a sensitivity of 100-1 000 genome copies, and achieving 100% accuracy in species identification (Thissen et al. 2014). It can run 96 samples simultaneously in a 96-well microarray plate and give us the results in less than one week. While less costly than the other technologies, this technology is only able to detect presence or absence of the different microorganisms in the sample (non-quantitative) and only for the species included in the microarray. This might pose a problem for the study of oral microbiome in remote populations, due to the microbial sequences included in the microarray being biased towards more studied populations, a common problem encountered with the use of microarray methodologies.

2.1.2. 16S rRNA gene sequencing

As a cost-effective method for profiling bacterial communities, the 16S rRNA gene sequence became one of the most widely used methods. The 16S rRNA gene contains both highly conserved regions, ideal for primer design, and hypervariable regions, which allow high phylogenetic resolution. It has nine hypervariable regions that are separated by nine highly conserved regions, but most studies do not use the whole rRNA gene sequence, only a few of the hypervariable regions (Rajendhran and Gunasekaran 2011). The choice of which variable regions to amplify and sequence depends on the needs of the project, as some regions offer more resolution at the genus/species levels of the microbial community than others (Bukin et al. 2019). One of the limitations of these studies is the lack of comparability between studies that analyze different hypervariable regions (Yang et al. 2016).

2.1.3. Shotgun metagenomics

Shotgun metagenomics sequences all the DNA material in a sample, yielding information on many genes rather than only one. This increased resolution (at the

species level) comes with the need for increased sequencing depth, and the associated higher cost (Ranjan et al. 2016). Additionally, as it is not only dependent on the 16S rRNA gene that is uniquely present in bacteria, it can also be used to investigate non-bacterial components of the microbiome (e.g., fungi, viruses, and archaea). As it has information of all the genes from the microbial community, it can also be used to direct evidence of profiling the abundance of microbial metabolic pathways (Beghini et al. 2021). However, the analysis of the results for shotgun sequencing relies on computationally advanced methods that require expertise (Peterson et al. 2021).

2.2. Microbiome diversity and metabolism

Two metrics are largely used to describe and compare communities (Escobar-Zepeda et al. 2015): Alpha (α) is a metric for local diversity of a community; and Beta (β) measures how different the community samples are in an area.

Alpha diversity measures the diversity in the samples and can be divided in two aspects: the richness in species, which quantifies the number of species in a sample; and the evenness of the sample that takes into account the relative abundance of each species in the sample (Whittaker 1972). There are a great number of alpha diversity metrics, some that only consider one of these aspects, and others that consider the whole sample diversity. The Simpson's dominance index expresses dominance or relative concentration based on the first or first few species; and the Shannon-Wiener index expresses the relative evenness or equitability of the species through the whole sequence. These two metrics consider both evenness and richness in their analysis (Thukral 2017). The Chao1 and abundance-based coverage estimators (ACE) metrics only measure the richness considering observed species in samples given all the identified bacterial (Hughes et al. 2001). The most appropriate metric to use depends on many factors, but mostly on sample size. While the Simpson's metric takes into account the most abundant species, the Shannon index is also based in the rare species, which can only be successfully measured with a large dataset. As such, by assigning greater weight on the most abundant species, they are much less sensitive to sample size. In contrast, with increasing numbers of individuals or samples

collected, rare species are continuously added to the sample, making the Shannon index a more appropriate indicator of species diversity (Gotelli and Chao 2013).

Beta diversity is a measure of quantification of the number of different communities in a specific region, informing about the degree of differentiation among biological communities. As with alpha diversity, there are beta-diversity indices that only take into account richness and just compare the number of shared and unique taxa between the samples, while there are others that also consider evenness in their calculations (Chao and Chiu 2016). The Jaccard index is one example of the former, as it analyzes the similarity between two sets of data only by looking into what is common and different between them, treating the common and rare species equally. When also considering the evenness of the samples, the Bray-Curtis dissimilarity indices analyzes abundance similarity, thereby species are not treated as equal (Chao et al. 2006). Alternatively, the UniFrac method compares microbial communities not only by their similarities but also through phylogenetic distance information. This method comes with two options that can produce different but complementary results, the weighted UniFrac that accounts for differences in relative abundances in the community, and the unweighted UniFrac that only considers the richness of species (Lozupone et al. 2006).

Other analysis applied to microbiome data go beyond taxonomy and explore the abundance of gene families, pathways, and systems, to identify the functional capabilities most valuable to organisms in the environment under study (Sharon et al. 2011). These analyses can be done at the single gene level, by focusing on the abundance of gene families, or at the pathway level where all the genes involved in the pathway are considered. These processes start by the identification of the genes in the microbiome data, and then aligning them against databases to predict their function (Beghini et al. 2021). Several databases are available, and they must be selected based on the type of analysis to be performed. For gene level analysis there is the COG (Tatusov et al. 2003), Pfam (Finn et al. 2006), and TIGRFAM (Haft et al. 2003) databases, while for systems or pathway level analysis, available databases are KEGG (Kanehisa and Goto 2000) or MetaCyc (Caspi et al. 2016). In addition to the function prediction, the contribution of each bacteria is also calculated, from the

number of reads in the microbiome data that are associated with the specific function. Ultimately, from the functional and abundance information, it is possible to compare gene families or pathways' abundances between the different samples and identify potential enriched pathways (Beghini et al. 2021; Sharon et al. 2011).

2.3. Microbiome and host interaction

The variability of the oral microbiome due to host genetics versus external factors has been largely discussed (Sedghi et al. 2021). Twin-based studies provide one method for quantifying heritability of microbial taxa (Figure 5). In such analyses, heritability is measured by comparing variation in microbial taxon abundances that is attributable to human genetics, more specifically, each microbe is a quantitative "trait", and the additive genetic factors affecting that microbe's abundance are estimated (van Opstal and Bordenstein 2015). A study characterizing the microbiome in supragingival plaque of monozygotic and dizygotic twin pairs, described that the similarity between the oral microbiome was higher for monozygotic twins than dizygotic twins, even when sharing the same environment (Gomez et al. 2017). Younger twins living together and exposed to similar environmental versus older twins living apart for many years, showed that host genetics affect the composition of the oral microbiota at a species level (Esberg et al. 2020).

However, some authors suggest that the heritability of the microbiome should be viewed in a more comprehensive manner. The standard heritability is based on the premise of a unidirectional interaction between host and microbiome, where colonization is controlled by the host. A more community-based interaction should be considered, as the microbiota interacts with each other as well as with the host (van Opstal and Bordenstein 2015). Thus, the community heritability takes into account the host-microbiome ecosystem and supposes that host genetic variation will have effects on microbial community assembly (Figure 5).

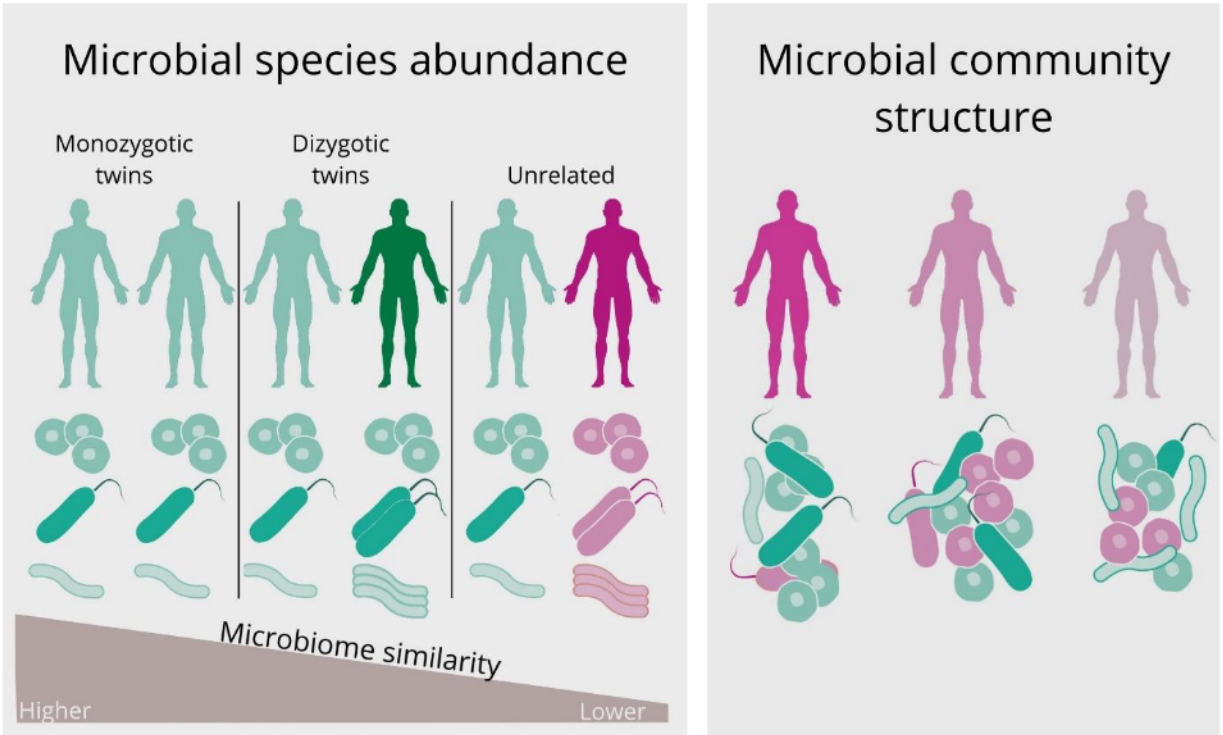


Figure 5. Standard heritability versus Community heritability analysis. Figure adapted from van Opstal and Bordenstein (2015).

The incorporation of these two approaches will allow a more comprehensive view in the study of heritability of the microbiome (van Opstal and Bordenstein 2015), already seen in some microbiome studies, where authors use beta-diversity (Hua et al. 2016) or microbial pathways (Goodrich et al. 2017) as the “trait” to be correlated with the host genetics.

Microbiome-human association studies (microbiome GWAS or microbiome quantitative trait loci abbreviated as mbQTL) aims to determine association between genotypic diversity in the host and microbiome patterns. Although studies in twins have suggested that the microbiome is influenced by the genetic profile of the host, studies aiming at evaluating direct associations are sparse and they mostly relate to the most studied microbiome, the gut microbiome. A rare example of a replicated association in several European populations is between variants near the lactase gene, responsible for lactose tolerance, and the abundance of *Bifidobacterium* (Blekhman et al. 2015).

Other results have been reported, but none was replicated and validated in different cohorts. One of the main limiting factors is the cohort size to achieve genome-wide significance, and the fact that the available data may have been obtained from different methods (different regions of the 16S rRNA, shotgun sequencing) makes the compilation of big databases for the analysis more challenging (Awany et al. 2018). Additionally, there is still a high disparity in the sampling population of microbiome-GWAS, with continuous exclusion of non-European populations (Popejoy and Fullerton 2016). The need to study non-European populations, in particular the African population, may unveil important host-genetic interactions in traits/diseases (Pedro et al. 2021), as it is well known that variants associated with diseases found in populations of European descent do not always replicate in non-Europeans (Peprah et al. 2015).

2.4. Microbiome of the oral cavity

The oral cavity is considered one of the most dynamic ecosystems in the human body, with each of its various structures such as the palate, tongue, saliva, and dental plaque, among others, having its own ecological niche and environment (Figure 6) (Xu et al. 2015). While the different oral niches have their own microbial composition and structure, in response to salivary flow, mastication, chewing, and other factors that select suitable microorganisms for each oral niche, they can be usually divided into three main parts, mainly the saliva, the tooth surface (plaques in the teeth surface), and soft tissues (surfaces of the tongue, buccal mucosa, palate, and gingiva) (Zhang et al. 2018). The salivary microbiome can be considered representative of the entire oral microbiota, as it includes the microorganisms that are shed from the various oral niches (Li et al. 2022).

The oral cavity harbors about 50 to 100 billion microorganisms, among bacteria, fungi, viruses, and archaea. Most of the microorganisms found in the oral cavity belong to the bacteria kingdom (Dewhirst et al. 2010). The core oral bacterial microbiome in healthy individuals (Caselli et al. 2020) includes Actinobacteria (genera *Corynebacterium*, *Rothia*, and *Actinomyces*), Bacteroidetes (genera *Prevotella*, *Capnocytophaga*, and *Porphyromonas*), Firmicutes (genera *Streptococcus* and

Granulicatella), *Fusobacteria* (genus *Fusobacterium*), and *Proteobacteria* (genera *Neisseria* and *Haemophilus*). Though present in smaller amounts, the remaining groups of microorganisms have important functions in maintaining the microbial balance of the oral cavity. The oral virome consists of eukaryotic viruses such as *Anelloviridae*, *Herpesviridae*, and *Papillomaviridae* (Bottalico et al. 2011; Wylie et al. 2014), and phages. There are approximately 100 species of fungi in the oral cavity, with the more common genera being *Candida*, *Cladosporium*, *Aureobasidium*, *Saccharomycetales*, *Aspergillus*, *Fusarium*, and *Cryptococcus* (Ghannoum et al. 2010; Hong et al. 2020). Oral archaea are far less numerous and diverse than oral bacteria, and are mostly methanogens (Wade 2013). Most of the microbial species present in the oral cavity are commensal species but they can become pathogenic in response to changes in the environment (Lamont et al. 2018). As the oral mucosa is the first barrier in the digestive tract, it plays an essential immune role (Fábián et al. 2012).

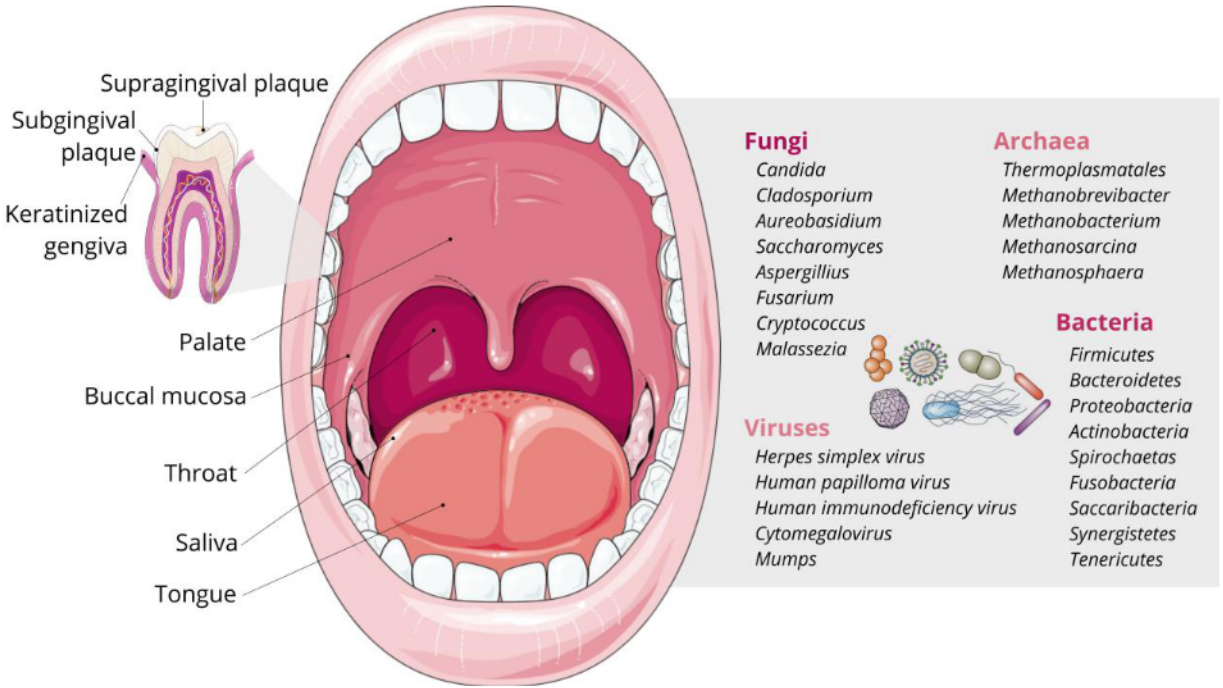


Figure 6. Representation of the diverse ecological niches in the oral microbiome and the major microbial components of the oral cavity. Figure adapted from Lee et al. (2021).

2.4.1. Factors that affect the oral microbiome

In the oral cavity, several factors can shape the abundance and composition of the oral microbial communities (Figure 7), as well as their structure and metabolic functions (Li et al. 2022). This can affect the normal balance of the oral microbiome, which can ultimately lead to the occurrence of oral caries, periodontitis, dysbiosis or even systemic diseases (Sampaio-Maia et al. 2016; Valm 2019).

2.4.1.1. Life stage and age

The oral cavity in the prenatal state is thought to be sterile and starts to be colonized soon after delivery. The oral microbiota of infants begins similar to that of their mothers, but early on there will be influences from diet, human interactions and genetics to modulate its oral microbiome (Sedghi et al. 2021). During the first months of a baby's life, the absence of teeth leads to microbial colonization of only mucosal surfaces, and later, with the eruption of teeth, microbes begin to colonize hard tissues (Sampaio-Maia and Monteiro-Silva 2014). Then, as children age and with the acquisition of permanent dentition, their oral microbiotas begin to stabilize right into adult life most likely due to more defined lifestyle patterns, namely oral-hygiene habits and more balanced dietary habits (Gomez and Nelson 2017). Nevertheless, some events in the host's life will create fluctuations in the oral microbiome (Li et al. 2022). For example, the oral microbiome can be affected by the hormones produced during the female menstrual cycle, with changes in *Campylobacter*, *Haemophilus*, *Oribacterium*, and *Prevotella* (Bostanci et al. 2021). Pregnancy also affects the oral microbiome, with higher abundances of *Neisseria*, *Porphyromonas*, and *Treponema* and lower abundances of *Streptococcus* and *Veillonella* (Saadaoui et al. 2021). Older age has been associated with a decline in the bacterial diversity of oral microbiotas (Burcham et al. 2020), a correlation also observed in the gut microbiome (Figure 7) (Claesson et al. 2012).

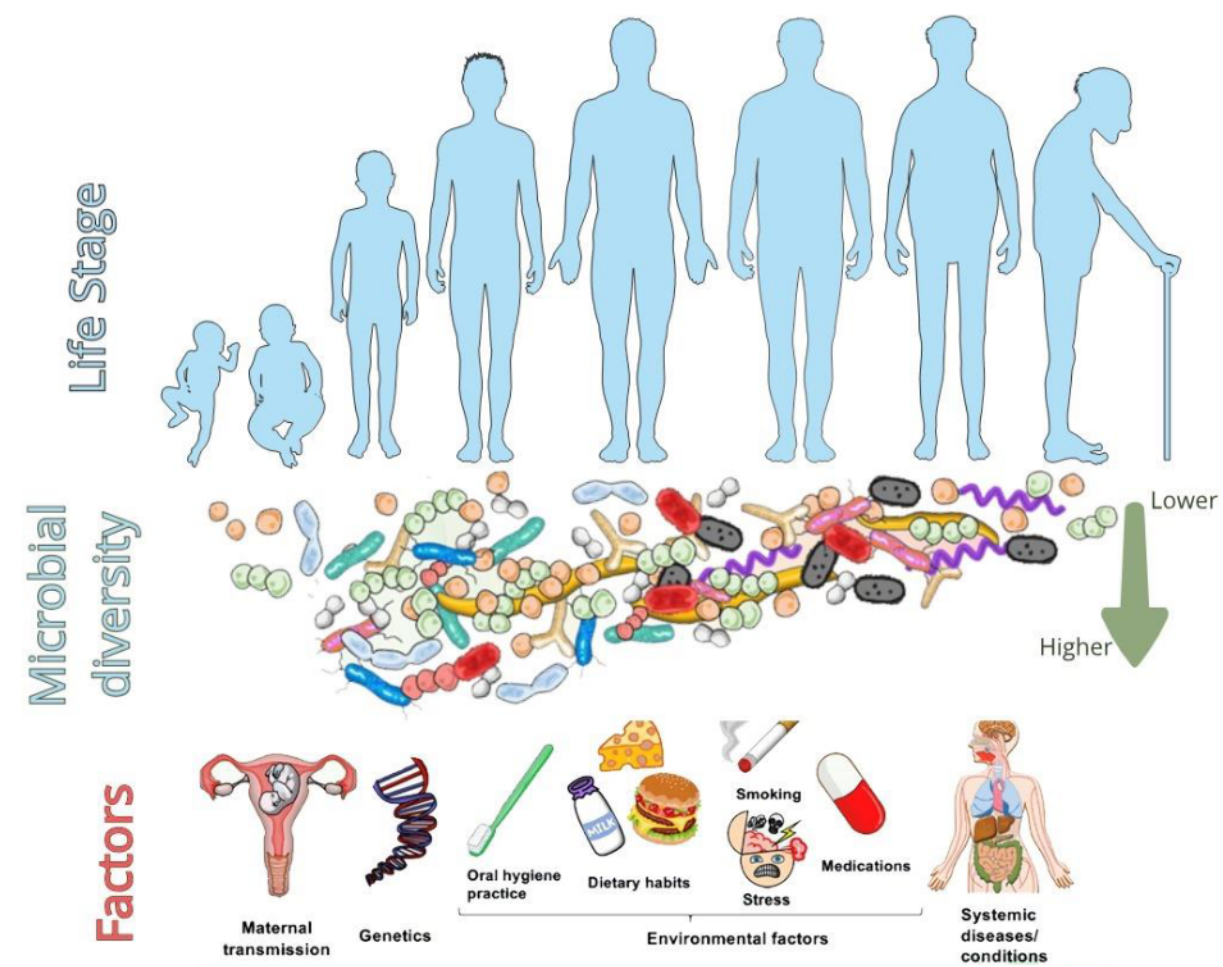


Figure 7. The dynamics of the oral microbiome from infancy to adult life, with emphasis on the factors that can modulate the oral microbiota. Figure adapted from Sedghi et al. (2021).

2.4.1.2. Food consumption

As the first step in food processing of the human digestive system, the relationship between diet and the oral microbiota has been a main point of focus. The changes in the diet caused by the initial domestication of plants and animals in the Neolithic, and later due to the advances in technology after the Industrial Revolution, were the two key periods that significantly altered the central constituents of the human diet (Santonocito et al. 2022). In the pre-agriculture period, diet was limited to minimally processed, wild plant and animal foods, characteristic to a hunter-gatherers lifestyle

(Cordain et al. 2000). With the introduction of agriculture, cereals and dairy products started to be included into the human diet. More recently, the increase in food-processing procedures following the Industrial Revolution led to the introduction of fatty domesticated and processed meats, saturated fats, refined grains, sugar, alcohol and salt, greatly affecting the oral microbiota. This new diet rich in fat and ultra-processed foods and poor in fruit and vegetable consumption defines the commonly known “Westernized diet”. Sugar and refined carbohydrates have been vastly associated with the abundance of oral bacteria, more specifically caries-associated bacteria, as it creates a favorable environment for the development of dental and periodontal diseases (Adler et al. 2013; Li et al. 2022).

In modern populations, studies have already shown that even in the first months of life, breast-fed infants had a higher proportion of *Streptococcus*, whereas formula-fed infants had higher proportions of *Actinomyces* and *Prevotella* (Oba et al. 2020). Moreover, the development of oral candidiasis was diminished in breast- and mixed-fed infants when compared to solid food-fed infants (Oba et al. 2020). In adults, a comparison between vegan and omnivorous individuals showed differences at all taxonomic levels below phyla, with an increased abundance of *Neisseria subflava*, *Haemophilus parainfluenzae*, *Rothia mucilaginosa* and *Capnocytophaga* spp. in vegans, and an increase abundance of *Prevotella melaninogenica* and *Streptococcus* spp. in omnivorous. Vegans also showed a higher number of species associated with periodontal disease (e.g. *Campylobacter rectus* and *Porphyromonas endodontalis*). Consumption of a variety of fatty acids and fiber was associated with bacterial richness and community structure. Furthermore, several metabolic pathways were found differentially abundant in vegans versus omnivores, indicating possible downstream effects of the different nutrient intake (Hansen et al. 2018).

2.4.1.3. Lifestyle and personal habits

Smoking can affect the composition of the oral microbiota, as it creates an anaerobic environment in the oral cavity, changes the pH of the oral saliva and promotes antibacterial effects due to its toxic substances. These changes in the environment may affect the adhesion of the normal oral bacteria and promote the colonization of

other types of microorganisms (Jia et al. 2021; Wu et al. 2016). Wu et al. (2016) described an increase in Firmicutes and Actinobacteria, and a decrease in Proteobacteria in smokers. However, due to the dynamic state of the microbiota, it has been shown that a former smoker can restore its oral microbiota to a healthy state after a certain period of non-smoking habits (Jia et al. 2021; Wu et al. 2016; Yang et al. 2019).

Excessive alcohol consumption can also cause an imbalance of the oral microbiome (Yussof et al. 2020). However, moderate and regular consumption of red wine has some antibacterial effects and does not affect the diversity of the human salivary microbiota (Barroso et al. 2015; Sánchez et al. 2019).

Betel nut (from the *Areca catechu* palm tree) chewing is a lifestyle habit common to populations in South and Southeastern Asia and the Pacific (Gupta and Warnakulasuriya 2002). It is a leading cause of oral cancer (Chattopadhyay et al. 2019), but it also affects the oral microbiome by reducing its alpha diversity (both evenness and richness) and changing the abundance of common bacteria. As with smoking, microbiome alternations induced by betel nut chewing may be reversible with cessation of this habit (Hernandez et al. 2017).

2.4.1.4. Dental care

Oral hygiene habits can greatly influence the oral microbiome, and may help in the prevention of caries and periodontal disease (Sedghi et al. 2021). Accordingly, the absence of oral hygiene can significantly alter the proportions of microbes in the dental plaque, including the ones responsible for periodontal disease (*Tannerella forsythia*, *Treponema denticola*, and *Porphyromonas gingivalis*). Interestingly, the absence of oral hygiene does not affect gingival inflammation, even with the increased amount of plaque on teeth (Baumgartner et al. 2009).

Toothbrushing and flossing are essential to clean and remove remnants of food that may harbor pathogenic bacteria, while toothpaste and mouthwashes have general

antimicrobial properties to reduce microbial load (Okuda et al. 1998; Terezhalmay et al. 2005). However, toothbrushes can also serve as reservoirs for pathogenic bacteria that can colonize the oral cavity, which highlights the importance of correct cleaning and storage of personal dental hygiene material (Shang et al. 2020). Furthermore, novel toothpastes are now designed to shape the oral microbiome by fostering species associated with healthy oral bacterial communities (Adams et al. 2017).

2.4.1.5. Antibiotics

Antibiotics have a much higher effect in the gut microbiome (Li et al. 2022), but it can also affect the composition and functions of the oral microbiome (Moraes et al. 2020). It has been shown that antibiotic use reduces the alpha-diversity of the amoxicillin versus the untreated group, with a continuous decrease even after 6 months (Menon et al. 2019; Monroy-Pérez et al. 2020). More specifically, the use of amoxicillin was shown to reduce the abundances of *Neisseria*, *Streptococcus*, and *Veillonella* in the oral microbiota (Larsson Wexell et al. 2016).

2.4.1.6. Genetics

For oral microbiome in particular, few studies have been performed. In European populations, Demmitt et al. (2017) identified variants in the *IMMPL2* and *INHBA-AS1* genes on chromosome 7 associated with variation in the abundance of *Granulicatella* and Poole et al. (2019) reported that gene copy number of the *AMY1* locus correlated with an increase in salivary *Porphyromonas*. In a Chinese cohort, Liu et al. (2021) found several associations, between *APPL2* gene with *Prevotella jejuni*, *Oribacterium uSGB 3339* and *Solobacterium uSGB 315*; the *SLC2A9* gene with *Oribacterium uSGB 1215*, *Oribacterium uSGB 489* and *Lachnoanaerobaculum umeaense*; the *OR11H1* gene with species F0422 uSGB 392; and the *LOC105371703* uncharacterized RNA gene with *Eggerthia*. In addition to associations with relative abundance, Liu et al. (2021) reported four loci, namely *NFIB*, *LINC02578*, *LOC105373105*, and *EIF3E*, associated with α -diversity (Chao1 index) of tongue dorsum samples and two loci, *SLC25A42* and *LINC02225*, associated with the α -diversity of salivary samples. The analyses with β -

diversity (Bray-Curtis Similarity) found association with one locus, *CTNNA3-DNAJC12*, in the tongue dorsum and one locus, *G2E3-AS1*, in the salivary microbiome.

3. Case study: Papua New Guinea

PNG lies in the southwest Pacific, in the north of the Oceania continent, right above Australia. It is separated from Southeast Asia by the Wallace Line, a strong ecological barrier for the spreading of fauna since the Pleistocene, which created two distinct biogeographic areas: the biota of mainland Asia and the biota of New Guinea and Australia (Cámara-Leret et al. 2020). It comprises the eastern half of the island of New Guinea (the western half belongs to Indonesia), and around 60 smaller islands to the east, the main ones being New Britain, New Ireland, Bougainville, and Manus. Mainland PNG accounts for 85% of PNG's total land area, about 462,840 km², and comprises complex geography and topography, dividing it into two regions: lowlands, including by open beaches, coastal swamps and dry savannah (Bourke and Harwood 2009), and the highlands, with characteristic dense rainforests, large wetland river basins and high mountain ranges (the highest point is Mount Wilhelm with 4509 m) (Figure 8) (Cleary et al. 2022). The lowlands have a much lower population density in contrast to a high population density in the highlands (Bourke and Harwood 2009), which is largely a result of high malaria endemicity in the coastal regions, in contrast to rare/absent malaria throughout the highlands (Bergström et al. 2017).

As PNG is entirely in the tropics, its climate is monsoonal, with the "wet" monsoon season extending from December to March, and the "dry" monsoon season from May to October. Average annual rainfall is high (2000-2500 mm), with even the so-called dry season having 50-100 mm of rain per month, in some areas. There is a high discrepancy in rainfall throughout the island, with some areas receiving more than 5000 mm annually, while others such as Port Moresby (the capital) receiving 1000 mm or less. Lowlands and coastal areas have higher temperatures and humidity than the highlands, and they do not vary much year-round (CountryReports 2022). The rugged landscapes and difficult terrain that is characteristic to PNG is responsible for highly isolated communities.

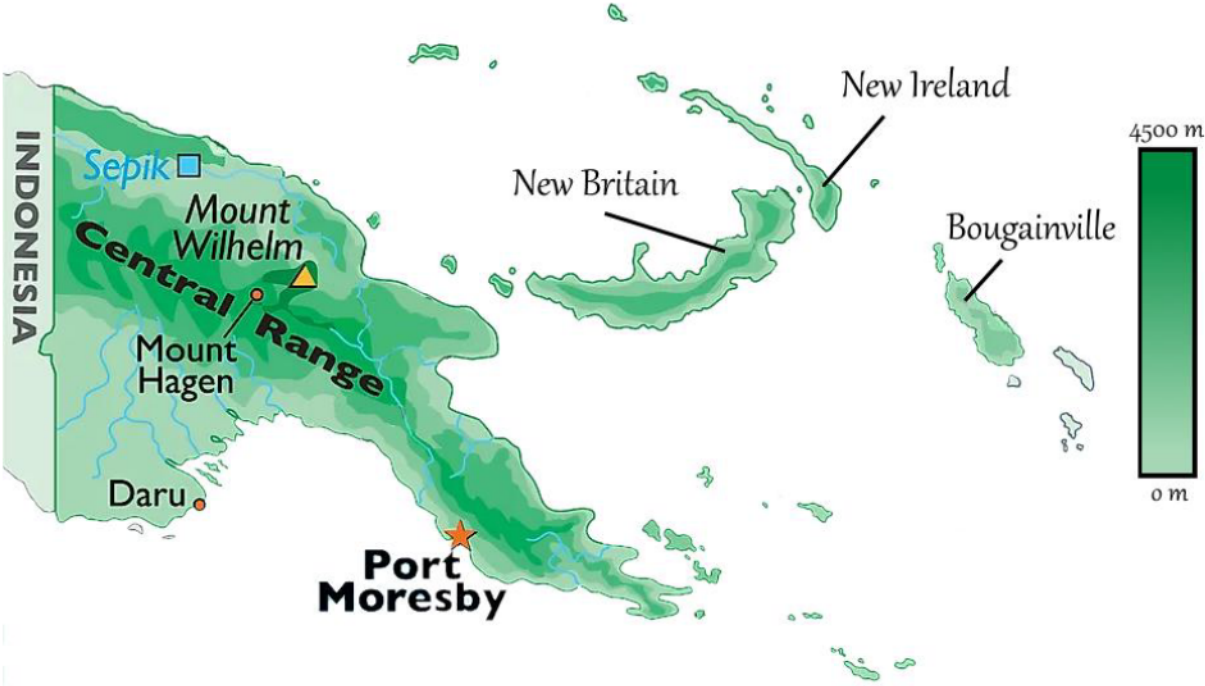


Figure 8. Geographic representation of the topography in Papua New Guinea with a color code for the elevation. Figure adapted from <https://www.worldatlas.com/maps/papua-new-guinea>.

Over 22,000 villages and a few urban centers harbor the 8 million people that live in PNG, who speak more than 800 “Papuan” languages (Pawley et al. 2005), in one of the most culturally, ethnically, and linguistically diverse countries in the world (Cleary et al. 2022). The Trans-New Guinea (TNG) family language includes an estimated 480 languages and is spoken across all the PNG highlands and large parts of the lowlands. Most of the central range, 2000 kilometers along the center of New Guinea, is occupied exclusively by TNG languages (Pawley et al. 2005). The next more common family language in PNG is the Austronesian (around 200 languages), introduced during the migration of Austronesians from Island Southeast Asia to the Pacific islands in the mid-Holocene, and is mostly represented in the coastal regions, the areas more affected by the migration (Foley 2000). The languages in the PNG region are spoken by an average of only 3000 speakers spread over 15-20 villages, increasing linguistic diversity. No language has ever emerged as more powerful, either numerically or economically, to replace the innumerable languages in PNG (Foley 2000).

Socially and politically, PNG is organized in clans, typically distributed across multiple villages or even several language groups (Foley 2000). The family unit in PNG is usually patriarchal and patrilocal, comprising a husband, with or without his parents, a wife, and children, with extended families often gathering for meals or to celebrate work, cultural or religion ceremonies (Bray and Smith 1985). Papuan daily life is strongly governed by cultural traditions with basic social patterns following traditional gender specialization. Men are responsible for deforestation and planting all the crops, control over the family income, repairing and building of houses and fences, and the butchering of animals. Women have the typical household tasks, including daily cooking, smaller gardening activities, taking care of the animals, and knitting the traditional grass skirts and bags called bilum (Bray and Smith 1985).

3.1. The settlement of Sahul: Archeology studies

The archeological sites available in the prehistoric continent of Sahul, a single landmass connecting the nowadays New Guinea, Australia, and Tasmania are sparse, which complicates the study of migration history in this region, in general and through ancient DNA (O'Connell et al. 2018). Limited success on ancient DNA studies were obtained for Pacific Islands, namely Vanuatu and Tonga, dating as far as 3.1–2.7 kya and 2.7–2.3, respectively, and Wallacea sites in North Moluccas, Sulawesi and East Nusa Tenggara dating as far as 2.6 kya and in South Sulawesi dating 7.3–7.2 kya (Carlhoff et al. 2021; Lipson et al. 2018; Oliveira et al. 2022; Skoglund et al. 2016).

The few archeological sites in Sahul allowed to infer an initial occupation at around 50 kya (Allen and O'Connell 2014). One site within northern Australia (Madjedbebe site) pointed to signs of an earlier arrival, around 65 kya (Figure 9) (Clarkson et al. 2017) but this date has not been universally accepted by the archaeological community. A more conservative estimate of the colonization of Sahul between 50–60 kya has been proposed (O'Connell et al. 2018). This is in accordance with dates obtained for the few archeological sites available in New Guinea, which indicate that the eastern part of New Guinea was reached 49 kya (Summerhayes et al. 2010), followed by New Ireland around 45 kya (Leavesley et al. 2002) and North Solomon islands by 33 kya (Wickler 1990). The journey between Sunda (the continental landmass of mainland and western

islands of Southeast Asia) and Sahul, involved the crossing of stretches of water between the Wallacean islands, a group of more than 1000 islands in the eastern section of modern Indonesia. The path taken to reach Sahul is still a matter of debate (Figure 9). Some studies support a northern route through Sulawesi and the Moluccas, favored on the basis of fewer crossings and overall shorter crossing distances (Kealy et al. 2018). In opposition, other studies indicate a southern path along the island chain near Timor, more parsimonious due to inter-visibility based on island shape and size and smaller water crossings for the initial part of the journey (Bird et al. 2018).

The study of Wallacea is still a work in progress, as the archeological evidence available in Wallacea is more recent than the archeological record from Sahul (Carlhoff et al. 2021). To date, the oldest sites associated with early modern humans found in Wallacea do not go further than 46 kya (Brumm et al. 2021; Hawkins et al. 2017; Shipton et al. 2020). Additionally, important archeological sites with evidence for the migration of modern humans to Sahul might be concealed by the submersion of islands within the Wallacean archipelago (Kealy et al. 2015) following the Last Glacial Maximum (LGM) around 9 kya (Lewis et al. 2013).

A few archeological sites have been uncovered in Sunda, namely the Niah Cave in Borneo (40-42 kya) (Barker et al. 2007), Tabon Cave on southwest Philippines (around 47 kya) (Détroit et al. 2004) and Callao Cave in northern Luzon (around 67 kya) (Mijares et al. 2010). The oldest one, in Sumatra, dates early modern human occupation as far as 73-63 kya (Westaway et al. 2017).

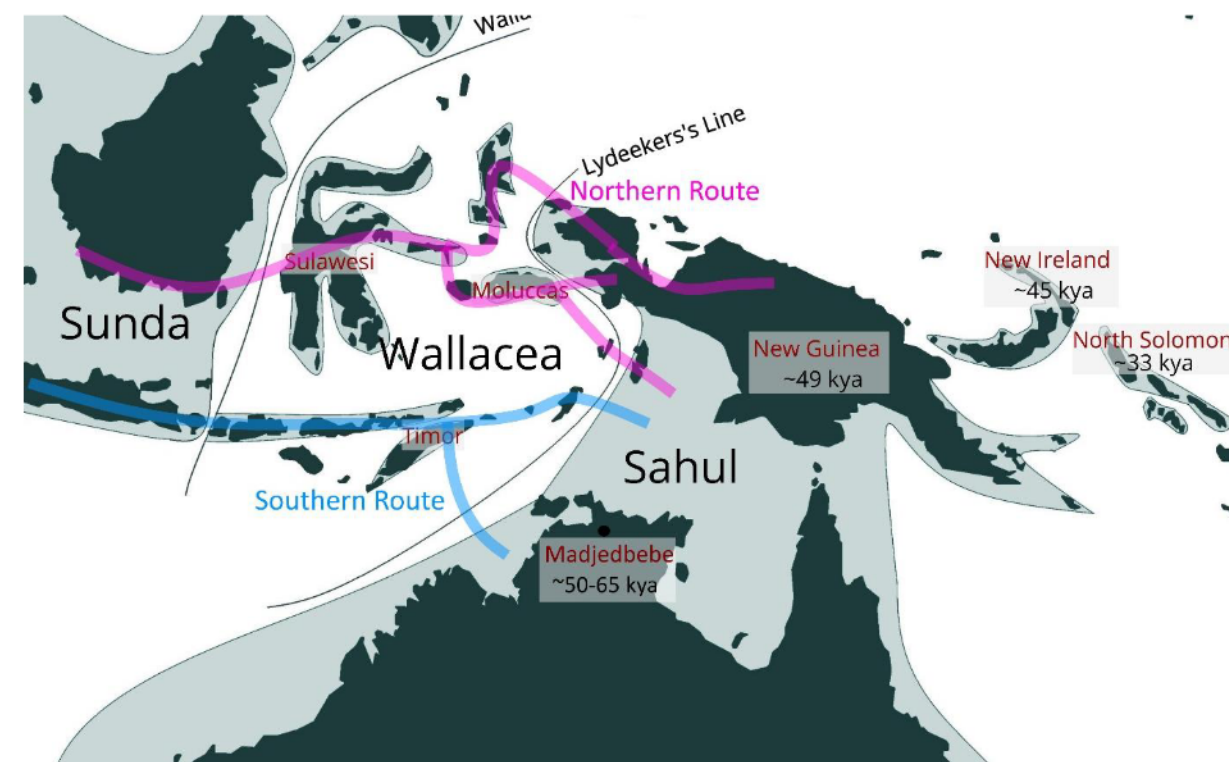


Figure 9. Map of the Sunda, Wallacea and Sahul regions, with dark green representing the present-day geography and the light shading the extent of the continents at the LGM. The potential northern (pink line) and southern (blue line) routes for the settlement of Sahul (Bird et al. 2019). Estimated times of arrival based on archeological sites are also indicated for eastern New Guinea (Summerhayes et al. 2010), New Ireland (Leavesley et al. 2002), North Solomon islands (Wickler 1990) and the Madjedbebe (Clarkson et al. 2017).

3.2. The settlement of Sahul: Genetic studies

The genetic characterization in PNG, although still incipient, has been already instrumental in supplementing essential information on the settlement of Sahul. With the technological advances, genome-wide studies have been performed, either by the genotyping of millions of SNPs through the use of SNP microarrays (Bergström et al. 2017) or through the sequencing of complete genomes of PNG populations (Brucato, André, et al. 2021; Malaspinas et al. 2016). All these studies have uncovered a strong

and ancient genetic differentiation within PNG that is still observed in the present populations.

In accordance with archeology data, genetic studies indicate that Sahul was settled by modern humans around 50 kya (Bergström et al. 2017; Malaspinas et al. 2016). Initial studies suggested that less than 150 people first colonized Sahul (Williams 2013), however, some argue that the initial colonization of the continent was deliberate and would have required organized sea travel, involving hundreds of people (Bird et al. 2019; O'Connell and Allen 2012). Genomic data corroborates this last hypothesis, as only a large population (greater than 1000) would have created the genetic diversity observed today (Merriwether et al. 2005). Demographic models also estimate that an initial population size between 1300 and 1550 people was required to avoid extinction in Sahul (Bradshaw et al. 2019; Brucato, André, et al. 2021).

This group of Sahul settlers are among the earliest modern human groups to leave Africa and are believed to be the ancestors of the current New Guineans and Aboriginal Australians that diverged from Eurasians 51–72 kya (Malaspinas et al. 2016). They are the oldest non-African locally continuous populations, and with the highest traces of archaic introgression from Denisovans (~4%), indicating a possible Denisovan presence in the Sahul region when the modern humans arrived (Jacobs et al. 2019). Additionally, Pagani et al. (2016) also found genetic signatures in present-day Papuans that may indicate that at least 2% of their genome come from introgression with modern human settlers from a largely extinct out of Africa migration earlier than 75 kya, but it is still highly controversial. Based on estimations from WGS, Aboriginal Australians and Papuans diverged from each other around 37 kya (Malaspinas et al. 2016). This suggests an ancient population structure in Sahul related to paleo-environmental changes leading up to the LGM, when the ice sheets grew to their maximum positions between 33 and 27 kya (Clark et al. 2009).

Genetic data provide evidence that population growth and divergence occurred within New Guinea, separating the lowland and highland groups, only in the postglacial warming period around 10–20 kya, when climatic conditions were more favorable (Malaspinas et al. 2016). Following this period, a spread of plant cultivation around 9

kya was observed in the highlands group (Denham et al. 2003) which may be correlated to the dispersal and expansion of the TNG languages (Pawley et al. 2005). During the mid-Holocene, around 3 kya, a migration of Austronesian groups closely related to mainland East Asians impacted the genetic structure of New Guinea, affecting mostly the coast of New Guinea and offshore islands (Kayser et al. 2008; Skoglund et al. 2016), bringing new languages, cultural practices and animals (pig, chicken). This Austronesian migration led to the settlement of the Pacific Islands (Matisoo-Smith 2015).

In relation to the uniparental markers, mt-DNA studies revealed five haplogroups present in PNG, namely Q, P, M27, M28, and M29 (Friedlaender et al. 2005; Merriwether et al. 2005). Earlier studies indicated that haplogroups P and Q constituted approximately 90% of the observed mtDNA lineages in New Guinea (Tommaseo-Ponzetta et al. 2002). Haplogroup Q is the most frequent and diverse haplogroup in Northern Sahul, found in the Highlands, Coastal New Guinea, and Near Oceania. It has three subhaplogroups Q1, Q2, and Q3, occurring more frequently in most, but not all, Papuan-speaking areas (Friedlaender et al. 2005). Haplogroup P is very unevenly distributed through PNG, with higher frequencies in the highlands and in certain islands of the Louisiade Archipelago (Friedlaender et al. 2005). The haplogroup P3, in particular, is the only mtDNA haplogroup shared between Australia and New Guinea (Huoponen et al. 2001). The haplogroups M27, M28, and M29 are almost exclusively found in Near Oceania (e.g., the Bismarck and Solomon archipelagos), and may have originated in this region as they are most diverse and frequent here (Merriwether et al. 2005). PNG genetic structure is even more evident in the paternal lineages (Y-chromosome) than in the maternal lineages, due to cultural practices (Bergström et al. 2017; Kayser et al. 2006; Mona et al. 2009). The majority of populations still abide to patrilocality, where upon marriage it is the female that moves to the village of the husband, leading to a female-based small-scale migration that reflects in the less differentiated mtDNA between neighboring populations than Y-chromosome lineages (Kayser et al. 2000; Kayser et al. 2006; Langergraber et al. 2007; Mona et al. 2009).

These almost 50 kya of Sahul isolation, high genetic diversity, heterogeneity of environmental conditions and high pathogenic pressure may have led to independent genetic evolution and diversification, making PNG a good and unique case study to look at biological evolution and adaptation.

3.3. Food habits

Most populations in PNG live in a semi-traditional way and depending upon local farming, largely based in starchy staples, such as taro, yam, unripe bananas, sago and sweet potato (Schapper 2017). The inclusion of tree crops, such as canarium nuts, okari, pandanus, candlenuts and breadfruit, add much energy and oil to the diet. Other edible vegetables part of their diets are the leaves of various fig species, of *Rungia klossii*, also known as “Mushroom plant”, and of *Dicliptera papuana*, sugarcane, vegetable cane, and bamboo (Bourke and Harwood 2009). Protein came primarily from hunting, fishing, and gathering of fungi and grubs, and later (in the last 3ky) was enriched by the domestication of pigs and other animals. There is some variation on available food throughput PNG, with a higher availability of starch-rich plant in the highlands, while the products of hunting and gathering are more common in the populations nearby forest (Schapper 2017). The use of local plants for almost 50 ky in PNG (Summerhayes et al. 2010), for a longer time than European used wheat and milk, might have driven unique human genetic and microbial adaptations and interactions.

More recently, the importation of foods largely altered the dietary profile in PNG, with access to greater dietary diversity, reflecting a shift towards increased consumption of protein-dense foods (Schmidt and Fang 2021). Importation of rice and wheat comprised 24% of total agricultural products imported in the last years, with also a growth in the importation of animal-sourced food, such as poultry and preserved and prepared fish. However, importation data also suggest a rapid growth in the total import value of processed (49% of total agricultural products imports) and ultra-processed (almost 70%) foods (Schmidt and Fang 2021). This last growth in particular is concerning, given the continuing growing evidence linking these foods with poor

health, including greater risk of cardiovascular disease, hypertension, and cancer (Baker et al. 2020; Elizabeth et al. 2020).

3.4. Disease burden

PNG, as a culturally diverse and low resource country, suffers with high mortality countrywide, attributed to conditions of poverty, endemic infections, emerging infections, non-communicable diseases (NCDs) and injuries (Figure 10) (Kitur et al. 2019). However, in a country where most deaths occur outside of health facilities, the available health reports are limited and they do not represent the picture of the country, so there is inadequate information regarding the burden of disease and cause of death in PNG .

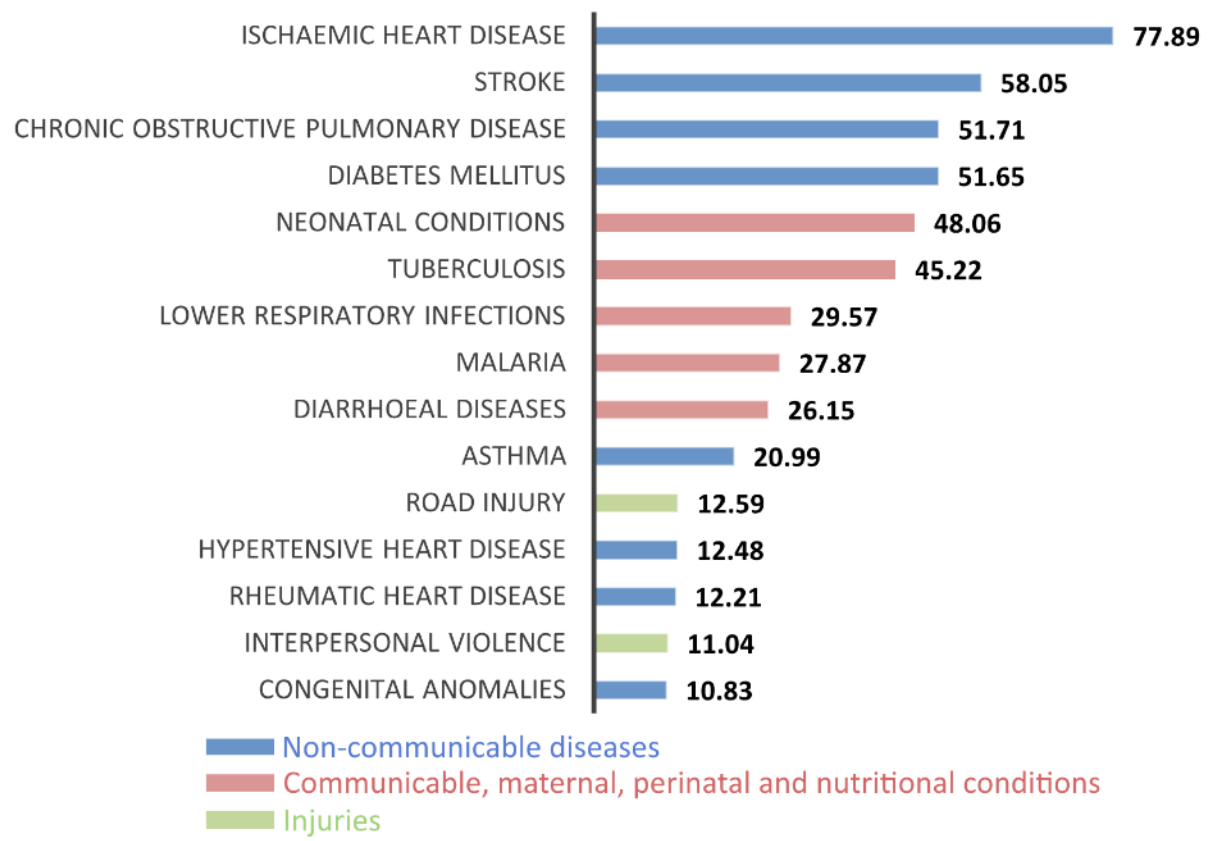


Figure 10. Leading Causes of Death in Papua New Guinea (top 15) for the year of 2019, indicated in deaths per 100,000 population. Causes of death are grouped in non-

communicable diseases (in blue), communicable, maternal, perinatal and nutritional conditions (in red) and injuries (in green). Data obtained from WHO (2020).

More than half of deaths in PNG are caused by endemic infectious diseases, such as malaria, pneumonia and diarrhea, in addition to conditions of poverty that lead to maternal, child and nutritional problems (Gouda et al. 2013). These mortality causes are more evident in provinces that are relatively remote, report high malaria cases and have the highest all-cause mortality in PNG (Kitur et al. 2019). Malaria infection, in particular, is shaped by the diverse landscape of PNG and human adaptation. For example, in the population living above 1,300 m, the temperatures are too low for sustained malaria transmission, and below 600 m partial immunity to infection was acquired by the local population (Cleary et al. 2022). Gerbich blood group-negative individuals showed resistance to infection by both *Plasmodium falciparum* and *Plasmodium vivax*, as their red blood cells express a truncated form of the glycophorin C receptor, which prevents these parasites to bind to the red blood cells (Zimmerman et al. 1999). Hepatitis B virus (HBV) is also endemic in PNG, affecting around 8% of the population (Kitau et al. 2015). Interestingly, the HBV strain found in PNG is unique to this region (HBV/C3), as HBV codiverged with modern humans during early global migration and unique strains can still be found in Sahul Indigenous populations (Yuen et al. 2019). As the main HBV transmission route in endemic populations is via mother-to-child for genotypes B and C infections (such as the one found in PNG), the HBV genome analysis can be used to infer ancient human population movements, corroborating the dates obtained from genetic and archeology studies (~50 kya) (Yuen et al. 2019). Other devastating disease under positive evolutionary selection in PNG is Kuru, an epidemic prion disease transmitted by endocannibalism and restricted to the Fore population (Asante et al. 2015; Mead et al. 2009). Variants in the prion protein gene were found exclusively in people who lived in the region in which kuru was prevalent, which were found to provide strong protection against the disease in the heterozygous state.

High mortality from emerging infections are linked to the Human Immunodeficiency Virus (HIV), Tuberculosis, and Human Papilloma Virus (HPV) (Kitur et al. 2019). These are highest in provinces that are accessible by the Highlands highway and National

Capital District, as they facilitate travel between the provinces that report the highest rates of sexually transmitted infections, such as HIV and HPV, responsible for deaths from acquired immunodeficiency syndrome and cervical cancer in PNG (Coghlan et al. 2011; NDOH 2019).

Endemic (such as chronic respiratory diseases, cancers, renal failure, liver cirrhosis) and emerging NCDs (such as diabetes, stroke, ischemic heart diseases, lung cancer) are becoming more prominent than infectious diseases and conditions of poverty in the provinces with higher socioeconomic status, such as the National Capital District and the Highlands (Gouda et al. 2013). Modernization, including high rates of cigarette smoking, inactivity and unhealthy diets are responsible for high rates of diabetes and coronary heart diseases in these provinces (Chang et al. 2014; Rarau et al. 2017).

Injuries, including homicide, falls, road traffic accidents, bites of venomous animals, fires, poisoning, and suicide affects males twice as much as females (Carter et al. 2016). The highest burden of injuries is reported in the provinces with the highest rates of violence and road traffic accidents. More remote provinces in the highlands have the highest rates of violence-related deaths, including homicide, while cities and coastal areas record more deaths resulting from road traffic accidents and falls from buildings (Gouda et al. 2013; Kitur et al. 2019).

3.5. Microbiome studies in PNG

Microbiome studies in PNG populations are few and are limited to the gut microbiota of healthy individuals from different locations. Greenhill et al. (2015) used reverse transcriptase real-time polymerase chain reaction to determine the prevalence of the established important groups of bacteria in the gut microbiota of individuals from the highlands and lowlands of PNG, living a predominantly traditional lifestyle. Highlanders presented a higher diversity of bacteria detected, but both rural groups showed similarities to those observed in other low-income settings where traditional diets are consumed. Martinez et al. (2015) analyzed the 16S rRNA of the fecal microbiota for two groups of highlanders from different locations of PNG, with a

traditional way of subsistence. The gut microbiota of these groups was compared to a group of individuals living in an industrialized country, USA. They concluded that the fecal microbiota in PNG is more diverse (several bacterial lineages found in PNG were undetectable in the USA cohort) but less individualized than in the USA, with the bacterial species shared between the groups presenting vastly different abundance profiles.

II. AIMS

The aim of this thesis is to characterize the ethnic diversity, population structure and oral microbiome in PNG, an understudied population with one of the most unique human biological diversities worldwide. This project will significantly impact the field of human evolution by creating a more ethnically diverse microbiome and genomic panels for Melanesian populations. Additionally, by integrating oral microbiome and genetic methodologies, we aim to uncover the biological adaptation of the oral microbiome in PNG populations living in a very heterogenous environment (physical/pathogenic). To achieve this, the research work of this thesis was divided in the following specific goals:

1. Characterization of the settlement of PNG

As a necessary introduction to the ethnic diversity observed in the region, we analyzed complete Papuan mitochondrial genomes from across PNG to reconstruct the phylogenetic and phylogeographic history of northern Sahul. We sought to shed some light into four key points about the settlement of Sahul that remain debatable: the number of groups of settlers involved in the settlement, the date of first human arrival, the route(s) taken by the first settlers, and the population substructure following the peopling of Sahul.

2. Evaluate the oral microbial diversity between the different regions of PNG

We performed the first known characterization of the oral microbiome in PNG individuals living in different environments, comparing these results with microbiome data obtained from other worldwide populations. We generated microbiome data by both qualitative (Microbiome Array) and quantitative methods (non-human reads from WGS), to characterize the oral microbiome patterns across the diverse PNG groups living in different environments.

3. Associations between microbiome diversity and human genomic data

We performed mbQTL analysis in a large a diverse PNG cohort to identify associations between human variants (complete genomes) and microbial genera (non-human reads of the WGS). Additionally, microbial functional profiling was inferred in a smaller dataset of the oral microbiome data from PNG and other worldwide populations, to detect the presence/absence and abundance of microbial pathways. This analysis allowed to identify enriched pathways and the respective contributing species in the Papuan populations versus the other groups.

III. RESEARCH WORK

PAPER I. PAPUAN MITOCHONDRIAL GENOMES AND THE SETTLEMENT
OF SAHUL

Journal of Human Genetics (2020) 65:875–887
https://doi.org/10.1038/s10038-020-0781-3



ARTICLE



Papuan mitochondrial genomes and the settlement of Sahul

Nicole Pedro^{1,2} · Nicolas Brucato³ · Veronica Fernandes^{1,2} · Mathilde André^{3,4} · Lauri Saag⁴ · William Pomat⁵ · Céline Besse⁶ · Anne Boland ⁶ · Jean-François Deleuze⁶ · Chris Clarkson⁷ · Herawati Sudoyo⁸ · Mait Metspalu⁴ · Mark Stoneking⁹ · Murray P. Cox¹⁰ · Matthew Leavesley^{11,12,13} · Luisa Pereira^{1,2} · François-Xavier Ricaut ³

Received: 5 February 2020 / Revised: 29 April 2020 / Accepted: 13 May 2020 / Published online: 1 June 2020
© The Author(s) 2020. This article is published with open access

Abstract

New Guineans represent one of the oldest locally continuous populations outside Africa, harboring among the greatest linguistic and genetic diversity on the planet. Archeological and genetic evidence suggest that their ancestors reached Sahul (present day New Guinea and Australia) by at least 55,000 years ago (kya). However, little is known about this early settlement phase or subsequent dispersal and population structuring over the subsequent period of time. Here we report 379 complete Papuan mitochondrial genomes from across Papua New Guinea, which allow us to reconstruct the phylogenetic and phylogeographic history of northern Sahul. Our results support the arrival of two groups of settlers in Sahul within the same broad time window (50–65 kya), each carrying a different set of maternal lineages and settling Northern and Southern Sahul separately. Strong geographic structure in northern Sahul remains visible today, indicating limited dispersal over time despite major climatic, cultural, and historical changes. However, following a period of isolation lasting nearly 20 ky after initial settlement, environmental changes postdating the Last Glacial Maximum stimulated diversification of mtDNA lineages and greater interactions within and beyond Northern Sahul, to Southern Sahul, Wallacea and beyond. Later, in the Holocene, populations from New Guinea, in contrast to those of Australia, participated in early interactions with incoming Asian populations from Island Southeast Asia and continuing into Oceania.

Introduction

Supplementary information The online version of this article (<https://doi.org/10.1038/s10038-020-0781-3>) contains supplementary material, which is available to authorized users.

The island of New Guinea comprises an area of 785,000 km² and hosts around 12 million people (8 million

✉ François-Xavier Ricaut
francois-xavier.ricaut@univ-tlse3.fr

¹ Instituto de Investigação e Inovação em Saúde, Universidade do Porto (i3S), 4200-135 Porto, Portugal
² Instituto de Patologia e Imunologia Molecular, Universidade do Porto (Ipatimup), 4200-465 Porto, Portugal
³ Laboratoire Évolution and Diversité Biologique (EDB UMR 5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, 118 route de Narbonne, Bat 4R1, 31062 Toulouse, France
⁴ Institute of Genomics, University of Tartu, 51010 Tartu, Tartumaa, Estonia
⁵ Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea
⁶ Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France
⁷ School of Social Science, University of Queensland, Australia, St Lucia, QLD 4072, Australia

⁸ Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta 10430, Indonesia
⁹ Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
¹⁰ Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand
¹¹ Strand of Anthropology, Sociology and Archaeology, School of Humanities and Social Sciences, University of Papua New Guinea, PO Box 320, University 134, National Capital District, Papua New Guinea
¹² College of Arts, Society and Education, James Cook University, P. O. Box 6811, Cairns, QLD 4870, Australia
¹³ ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Wollongong, Wollongong, NSW 2522, Australia

in Papua New Guinea and 4 million in the western Indonesian half of the island), with the highest density in the intermountain valleys 1400–1850 m above sea level (masl). This is one of the most bio-culturally diverse regions on Earth [1] with more than 900 languages spoken, mostly Papuan, but with some Austronesian languages arriving in the last 3 ky [2, 3].

At the time of the initial arrival of modern humans at least 50 kya [4] and substantially earlier [5], New Guinea, Australia, and Tasmania were connected into a single landmass called Sahul, until rising sea levels during the Holocene 9 kya flooded the Torres Strait [6]. In present-day geography, the Pleistocene Sahul continent can be divided into Northern Sahul, representing New Guinea and Near Oceania, while Southern Sahul corresponds to Australia. New Guinea represents approximately a third of the Sahul landmass and the most mountainous part of it, with peaks reaching 4900 masl.

To reach Sahul, the initial settlers from Sunda had to cross up to 90 km stretches of water between the Wallacean islands using still debated southern and/or northern routes via Timor and/or Sulawesi [7–10]. Bradshaw et al. [11] modeled an initial population size of between 1300 and 1550 individuals for the peopling of Sahul, indicating this was a planned crossing involving substantial numbers of people. The eastern part of New Guinea was reached by 49 kya (Ivane valley [12]), the islands of New Ireland by 45 kya, involving a sea crossing of 50 km (Buang Merabak [13]) and the North Solomons by 33 kya with an open sea crossing of 80–180 km depending on the route taken (Kilu cave on Buka island [14]).

New Guineans derive from a biological and cultural mixture of these first Papuan settlers, who arrived around 50 kya [4], and mid-Holocene Austronesian groups closely related to mainland East Asians [2], the latter having the strongest impact on the coast of New Guinea and offshore islands [2, 15–17]. However, early plant domestication and an independent development of agriculture in Highland New Guinea by 9 kya [18] suggest a complex multi-directional exchange of artifacts, plants, animals, and technologies between New Guinea (e.g., banana and sugar cane) and Island Southeast Asia (e.g., pigs and chickens) starting from the mid-Holocene and possibly earlier [19, 20].

Recent genomic data suggest that Indigenous Australians and Papuans diverged from Eurasians 51–72 kya and from each other around 10–32 kya [21]. A later divergence between lowland and highland groups in New Guinea is also attested from the postglacial warming period (10–20 kya) with highland population growth following the spread of plant cultivation around 9 kya [21, 22] and leading to the expansion of Trans-New Guinea languages [23]. Strong genetic differentiation among Papuan groups is observed

today, with greater structure for the Y chromosome (paternal lineages) than mitochondrial DNA (maternal lineages), reflecting sex-specific cultural practices [22, 24–26].

These cultural and biological patterns result from the almost 50 kya of Sahul isolation leading to independent genetic and cultural evolution and diversification [21, 23, 27, 28]. New Guineans and Aboriginal Australians can be considered today as the descendants of the earliest modern human group(s) to leave Africa, and are the oldest locally continuous populations found outside Africa. They are also the living groups with the highest traces of archaic introgression from Denisovans (~4%), raising the possibility of a Denisovan presence east of the Wallace line [29], and possible genetic traces of a very early and elsewhere extinct expansion of modern humans out of Africa [27].

However, despite this exceptional context of New Guinean population history, fewer than 20 Pleistocene sites are known for all of New Guinea [30] and there are very limited genetic data available, compared with other regions of the world. Our understanding of the population dynamics that led to the current situation in New Guinea is therefore still largely unknown.

Here we report the largest mitogenome dataset for New Guinea, including 379 new genomes, and investigate four key points about the scenario of human arrival in Sahul that are still unclear and contentious: (1) the number of different groups of settlers involved, (2) the date of first human arrival (65 kya and/or 50 kya; [4, 5]), (3) the route(s) taken by the first settlers (northern and/or southern routes) [7–10], and (4) the nature of population substructure following the peopling of Sahul [22, 25, 26, 31–33].

Material and methods

Sampling and mtDNA sequence generation

The samples analyzed in this study were drawn from populations across Island South East Asia, Oceania and Australia. The dataset of 915 complete mtDNA genomes (Tables S1, S2, and S3) includes mtDNA sequences compiled from (1) newly collected samples from Papua New Guinea, (2) archival biological samples from the Institute of Medical Research (IMR) of Papua New Guinea, and (3) previously published studies.

A total of 123 DNA samples were collected during 2016 and 2017 field seasons in Papua New Guinea, and cover individuals from all 22 Papua New Guinea provinces (Table S1). All samples were collected from healthy unrelated adult donors after informed consent forms were signed. In each sampling location, a full presentation of the project was made, followed by discussion with each donor

to ensure that they fully understood the project. Participants were surveyed for language affiliation(s), current residence, date and place of birth, and a short genealogy up to three or four generations to establish regional ancestry. Saliva samples were collected using the Oragene DNA Collection Kit (DNA Genotek Inc., Ottawa, Canada). DNA was extracted according to the manufacturer's instructions.

A further 256 samples were selected from the IMR archival biobank, derived from blood samples collected in the 1980s, with new ethics approvals obtained in 2015. These samples cover four highland provinces and one coastal province (Table S1). DNA was extracted using the DNA Blood Mini Kit (QIAGEN, Hilden, Germany) and whole-genome amplified with the Illustra GenomiPhi V2 kit (GE Healthcare, Chicago, IL, USA) following the manufacturer's instructions.

Complete mitochondrial DNA sequences were generated for all 379 new and archival samples using two approaches. For 73 newly collected samples and 256 IMR samples, complete mtDNA sequences were generated following the protocol described in Brucato et al. [34]. Briefly, double bar-coded libraries were prepared and enriched for mtDNA, as described previously [35, 36]. For 50 newly collected samples, complete mtDNA sequences were extracted from whole genome sequencing performed at CNRGH, France (Table S1). Sequencing libraries were prepared using TruSeq DNA PCR-Free and TruSeq Nano DNA HT kits depending on DNA quantity. 150 bp paired-end sequencing was performed on the Illumina HiSeq X5 System (CNRGH). For all samples, consensus sequences were obtained after base-calling, quality filtering, and further quality control steps to obtain consensus sequences, as described previously [37]. The 379 new complete mtDNA sequences have been deposited in GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) under accession numbers MN849490–MN849868.

A comparative dataset of 539 individuals was built by compiling all of the published complete mtDNA sequences affiliated with haplogroup P, Q, and M (M27, M28, M29) described previously for people of Papuan ancestry (Table S4). These sequences were identified by screening the main web-based mtDNA databases: the DDBJ/EMBL/GenBank international nucleotide sequence database, PhyloTree [38], and Family Tree DNA (<https://www.familytreedna.com/>). The final dataset includes 915 mitochondrial genomes from haplogroup P, Q, and M, including the 379 new and archival sequences from Papua New Guinea generated in this study (Tables S1–S3).

All new, archival, and comparative sequences were analyzed and aligned against the revised Cambridge Reference Sequence [39] using the MAFFT aligner v.7 [40]. Mitochondrial haplogroups were determined using Haplogrep [41] based on PhyloTree Build 17 [38].

Mitochondrial DNA analysis

Phylogenetic relationships were analyzed by constructing maximum parsimony trees using the whole mtDNA sequences affiliated with haplogroups of Papuan ancestry M27 ($n = 144$ samples), M28 ($n = 73$), M29*Q ($n = 361$), and P ($n = 302$) (Figs. S1–S4), guided by published principles [38].

To estimate the time to the most recent common ancestor (TMRCA) of the clades, the maximum parsimony trees were used for maximum likelihood (ML) estimations (Table S4), by considering two widely used mutation rates so that our estimates are comparable with published ones: Fu et al. (mean = 2.67×10^{-8} substitutions per site per year, $SD = 2.6 \times 10^{-9}$) and Soares et al. [42] (mean = 1.67×10^{-8} substitutions per site per year). These authors extensively evaluated the effect of demographic effects, and, in the latter case, also of selection effects, on their estimations. It was not our aim to add another mutation rate to the ones already available and extensively used. The Soares clock accommodates the effects of purifying selection, tending to attribute older ages that will function as the maximum limit to our estimates. The mutation rate from Fu et al. [43] was calibrated on radiocarbon dated ancient sequences, and the used method provides inferences that minimize the effects of rate temporal dependency, and will work here as the minimum limit to our estimates. We performed the ML estimates of branch lengths using PAML v.4 [44], assuming the HKY85 mutation model with gamma-distributed rates, excluding indels and hotspot mutations, as reported previously [38].

The timing of modern human arrival in Northern Sahul, based on the mtDNA genomes analyzed in this study, was estimated from the TMRCA of the main Papuan haplogroups using both the Fu and Soares mutation rates. While confidence intervals estimated with either mutation rate tend to overlap, we favor Fu et al. [43] age estimates because they are based on directly dated fossils that are within the time range of modern human evolution considered here. However, older age estimates given by the Soares mutation rate cannot be ruled out. Thus, skeletal representations of the tree (Fig. 1) were drawn using Fig-Tree (tree.bio.ed.ac.uk/software/figtree/), with the tree scaled using the ML TMRCA estimates using Fu et al.'s [43] mutation rate.

The spatial frequency distribution of the main Papuan haplogroups was estimated with the ESRI ArcGIS software package (www.esri.com/software/arcview/), based on the frequency data for each haplogroup and the latitude and longitude of the center points for each of the seven regions considered in this study (Near Oceania, Remote Oceania, Coastal New Guinea, Highland New Guinea, Australia, Wallacea, and Sunda) (Table S5). The frequency maps were

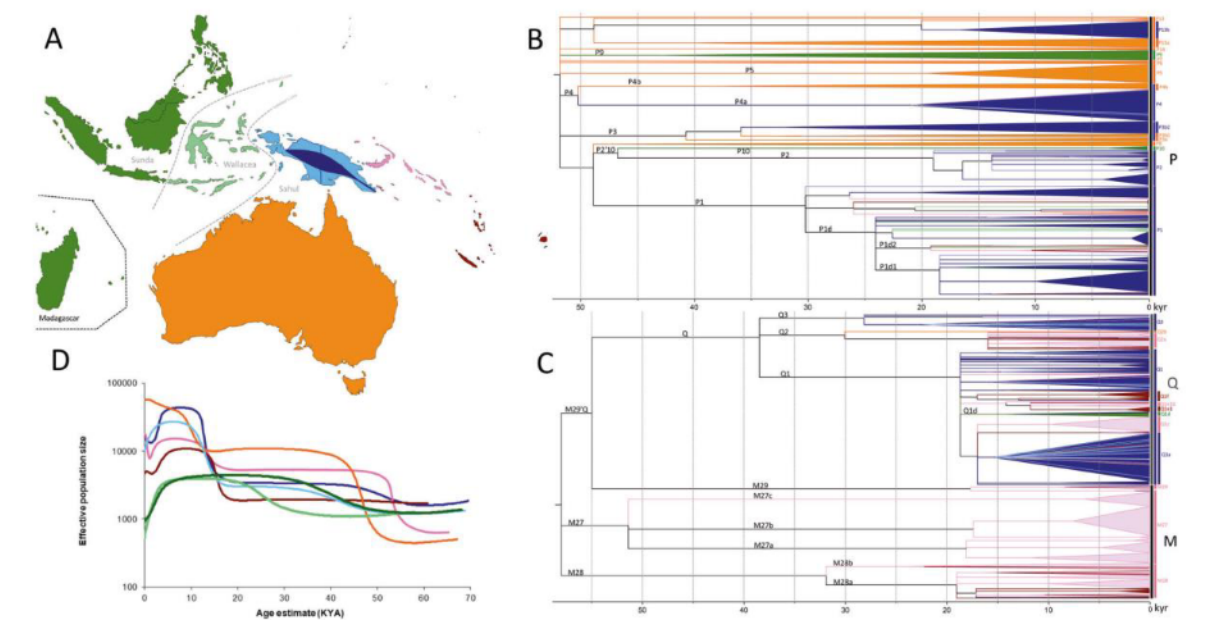


Fig. 1 Genetic relationships and increasing population sizes across Sunda and Sahul. **a** Seven geographical regions: coastal PNG (light blue), highland PNG (dark blue), Near Oceania (pink), Remote Oceania (red), Australia (Orange), Wallacea (light green), and Sunda (dark green). **b** Tree of haplogroup P. Subclades are represented by triangles, while single lineages are represented by lines. The tree is scaled to kya (thousands of years ago) using the maximum likelihood molecular clock for the whole mtDNA genome with the mutation rate of Fu et al. (details of age estimates are reported in Table S1). **c** Tree of haplogroups M and Q. **d** Bayesian skyline plot representing median estimates of effective population size for each the seven geographic regions based on P, M, and Q mtDNA lineages

created using the “Inverse Distance Weighted” (IDW) option, with a power value of two for the interpolation of the surface. IDW assumes that each input point has a local influence that decreases with distance based on the assumption that samples close to one another should be more alike than those that are farther apart. However, as it accounts only for the effects of distance, the color scale in regions with few or no samples (e.g., Australia for haplogroups M28/27/29, Indonesia for haplogroup P and Q) should not be considered as having high statistical confidence. Details regarding the different haplogroup frequencies in these regions should be verified in Table S5.

To assess effective population changes through time for Papuan haplogroups (M27, M28, M29/Q, P), Bayesian Skyline Plots (BSPs) were calculated using BEAST v.1.8 [45] and visualized with Tracer v. 1.6 (<http://beast.bio.ed.ac.uk/Tracer>). A 25-year generation time was assumed [46]. BSPs estimate the effective population size through time using random sequences from a given population but have also proved effective with individual haplogroup data. For this analysis, we used Fu’s mutation rate. BEAST uses a Markov chain Monte Carlo (MCMC) approach to sample from the posterior distributions of model parameters (branching times in the tree and substitution rates). Specifically, we ran 100,000,000 iterations, with samples drawn every 10,000 MCMC steps, after a discarded burn-in of

10,000,000 steps. We checked for convergence to the stationary distribution and sufficient sampling by inspection of posterior samples.

Results

Haplogroup affiliation in Northern Sahul

Of the 379 new complete mtDNA genomes, 28 individuals (7.4%) are affiliated to nonindigenous Northern Sahul haplogroups B4a1a1 (*n* = 25) and E1a (*n* = 3) and their descendant lineages. These haplogroups are only detected in the coastal region of New Guinea and the Bismarck Archipelago (Table S1 [47, 48]), while their ancestral lineages are thought to have entered the New Guinea region from mainland Asia following postglacial and/or Austronesian expansions in the early and mid-Holocene, respectively [48–50]. Overall, their current distribution demonstrates limited penetration inland into New Guinea itself in agreement with previous studies [47, 48, 51–53].

The remaining 351 mtDNA genomes include two minor indigenous haplogroups from Northern Sahul, R14 and M73 [54, 55]. Each present in only one highland PNG individual, they are too rare to be informative in the frame of this study and were not analyzed further.

The vast majority of the mtDNA genomes, 349 (92% of 379), are affiliated with the main indigenous Northern Sahul haplogroups—M27 (1.3%), M28 (<1%), M29 (<1%), Q (47%), and P (43%). Considering the entire dataset (Table S5), these indigenous lineages represent 1818 (57%) individuals in Northern Sahul (M27 8.2%, M28 8.8%, M29 1.5%, Q 25%, and P 13.5%); a number that rises to 2200 when all regions worldwide are considered. These indigenous Northern Sahul haplogroups are the focus of this study.

Northern Sahul phylogeography

The geographic distribution of the major indigenous haplogroups from northern Sahul (M27, M28, M29, Q, and P)

reveal strong geographic patterns for each of the five haplogroups (Fig. 2a–e), centered on northern Sahul, where they harbor—except for some P subhaplogroups—their greatest frequency and diversity (Figs. 1 and S1–S4). However, when we look in more detail at the geographic distribution of their subclades, associated phylogenetic trees and coalescence age estimates (Table S4, Figs. 1 and S1–S4), we can identify three main groups of indigenous Northern Sahul lineages.

The first group includes haplogroups M27, M28, and M29, which are almost exclusively found in Near Oceania (e.g., the Bismarck and Solomon archipelagos), and may have originated in this region as they are most diverse and frequent here (Tables S1, S2, S5 and Fig. 2a–c;

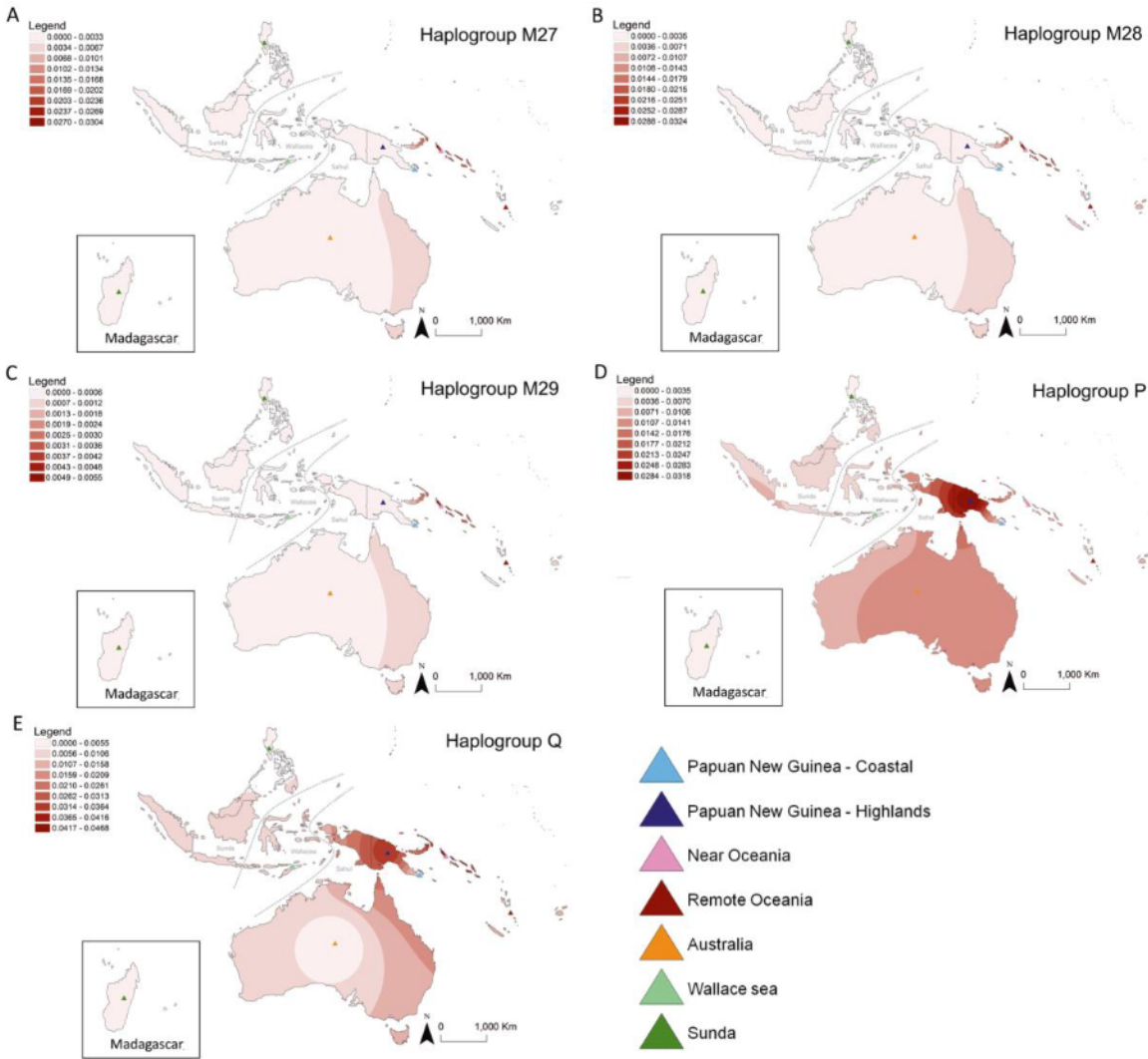


Fig. 2 Spatial distribution of main Sahul mtDNA haplogroups. Inverse distance weighted interpolation shows areas with higher frequencies in darker shading, taking into account only the effects of distance (color scale in regions with few or no samples should not be considered as

having high statistical confidence). Data details are provided in Table S5 and the triangles represent the central point for each region used in the interpolation. **a** Distribution of haplogroup M27. **b** Haplogroup M28. **c** Haplogroup M29. **d** Haplogroup P. **e** Haplogroup Q

[47, 48, 52, 56]). These haplogroups are absent from Australia, and are rare (<1%) in the highlands of Papua New Guinea and in Remote Oceania (Table S5), whose lineages root in haplotypes found only in the Bismarck and Solomon archipelago. From their age estimates (Table S4), haplogroups M27, M28, and M29 found outside Near Oceania appear to be related to increasing population interactions within the region during the mid-Holocene.

Coalescence ages estimated for M29*Q ~55 kya (95% CI 42–67 kya) and M27 ~51 kya (95% CI 40–62 kya) are in broad agreement with dates for the first settlement of Sahul from archeological and genetic data [4]. At ~32 kya (95% CI 22–42 kya), the M28 age estimate is close of the beginning of the Last Glacial Maximum (28 kya). Long branches on the phylogenetic tree for haplogroups M29, M27a,b,c, and M28 may suggest long-term isolation without expansion for these lineages after their initial arrival in the Bismarck and Solomon archipelagos (Fig. 1), while other explanations for this pattern—unsampled diversity or lineage disappearances—cannot be fully ruled out.

The next main lineage diversifications took place around the transition period between the Last Glacial Maximum (28–18 kya) and the postglacial warming period (18–10 kya) [18, 57] (M27a: 18 kya, M27b: 17 kya, M28a: 19 kya, M28b: 24 kya, M29: 18 kya) and extended into the Holocene period of increasing population interaction (Figs. S1–S3 and Table S4). All derived lineages arising from these two expansions periods are clustered geographically, suggesting limited dispersal over time despite major climatic and historical changes.

The second group includes haplogroup Q and its sub-haplogroups Q1, Q2, and Q3, all of which have their greatest frequency and diversity in Northern Sahul (highland and coastal New Guinea and Near Oceania) (Figs. 1, S3, and 2e; [48, 49]), suggesting probable origins within this region. Phylogenetic analysis indicates that the Q branch diverged from M29*Q around ~55 kya (95% CI 42–67 kya), and that after a period of isolation of nearly 15 kya (long branch, Fig. 1), the Q haplogroup diversified into three subhaplogroups around 38 kya (95% CI 28–52 kya), a time between initial settlement and the Last Glacial Maximum (28–18 kya). Haplogroup Q2 in Near Oceania and Q3 in highland New Guinea show diversification early in the Last Glacial Maximum period, ~30 kya (95% CI 20–40 kya) and 28 kya (95% CI 20–36 kya) respectively, while Q1 in Highland and coastal New Guinea and Near Oceania diversified at the end of Last Glacial Maximum ~19 kya (95% CI 15–22 kya). As seen for M27, M28, and M29, derived lineages of these subclades have tended to stay in the same geographical region with little evidence of spreading across Northern Sahul.

Haplogroup Q1, with a coalescence age estimate of ~18 kya, shows geographic clustering, with highest frequency

and diversity in highland New Guinea (Q1a, Q1f, and >10 related but unnamed subclades) and Near Oceania (Q1b, Q1c, and Q1e). Few lineages rooted in these sub-haplogroups are detected outside Northern Sahul. These are found in the Sunda and Wallacea islands (Q1d, also Taiwan, Philippines, and Madagascar), in northern Australia (Q1a), and in Remote Oceania (Q1b, Q1e, and Q1f subclades in the Solomon Islands, Vanuatu, Fiji, Samoa, and the Cook Islands). All have coalescence ages within the postglacial warming period (18–10 kya) or following the Holocene. The three Q1 lineages present in Australia were all identified in individuals with Torres Strait Islander ancestry, which are known for their close links with New Guinea [32], and are associated with the Q1 subclade expansion during the postglacial warming period (18–10 kya) (Fig. S3).

Haplogroup Q2 has a coalescence age estimate at ~30 kya, and is also strongly geographically structured, with greatest diversity and frequency in Near Oceania, particularly the Bismarck Archipelago. This is in agreement with proposed demographic expansions following the initial settlement of New Britain around 35 kya [58]. Diversification of this haplogroup began in the post-Last Glacial Maximum period and continue into the Holocene, but was largely restricted to Near Oceania. However, a few lineages have been observed in Remote Oceania with coalescence ages in the late Holocene (Fig. S3 and Table S4). We confirm the presence of Q2b in one western Indigenous Australian. This mtDNA haplotype branches deeply within the Q clade [54] and may reflect earlier connections (predating the Last Glacial Maximum, Fig. S3) between Northern and Southern Sahul populations.

Haplogroup Q3 shows a similar pattern, with a distribution center in the highlands of New Guinea, where it has its higher diversity and a coalescence age of ~28 kya (early in the Last Glacial Maximum period). However, several Q3a and Q3b subclades have been detected in coastal New Guinea (Q3–215 ~21 kya, preQ3b ~16 kya, Q3b ~7 kya), Near Oceania (preQ3b ~16 kya) and Timor (Q3–215 ~21 kya), all associated with coalescence ages from the end of the Last Glacial Maximum (28–18 kya), postglacial warming or Holocene periods (Table S4 and Fig. S3). Of particular note, the Q3 lineages shared between Timor and coastal New Guinea are deep branching and have a coalescence age of ~21 kya (Table S4 and Fig. S3) within the late Last Glacial Maximum period, reflecting a possible ancient connection between the island zone around Timor and continental Northern Sahul [55].

The third group includes haplogroup P, with a coalescence age of ~51 kya (95% CI 44–60 kya), and its numerous subhaplogroups (P1–P12) (Phylotree.org). Haplogroup P has clades with deep branches rooted in the basal P clade, distributed across Southern Sahul (P3a, P3b2, P4b,

P5, P6, P7, P8, new P13a (P-153G, Fig. S4)) and Northern Sahul (P1, P2, P3b1, P4a, new P13b1). This lineage has also been identified outside Sahul in the Philippines (P9 and P10 in the Aeta and Agta indigenous groups), suggesting that the P haplogroup may have evolved in Sunda or Wallacea (Fig. 1, Table S4, and Fig. S4) [55].

Both Northern and Southern Sahul P haplogroups have old coalescence ages (around 50–45 kya, 95% CI 32–62 kya) (Table S4) some probably related to the early settlement period P4 ~50 kya (95% CI 42–58 kya), P6 ~50 kya (95% CI 41–58 kya), Pre-P1 ~49 kya (P-16176T) (95% CI 38–60 kya), P-13: ~49 kya (95% CI 34–64 kya), pre-P8 ~48 kya Pre-P1 (95% CI 33–62 kya), P2’P10 ~47 kya (95% CI 35–58 kya), P9 ~43 kya (95% CI 30–55 kya), P3 ~41 kya (95% CI 32–50 kya). In general, P haplogroups diversified earlier in Southern Sahul (older coalescence age) than in Northern Sahul, which occurred from the end of the Last Glacial Maximum through to the postglacial and Holocene periods, followed by long-term isolation (long branches on the tree) (Table S4 and Fig. 1). While some rare sharing of P lineages within and outside Sahul suggest some level of population interaction (discussed below), most derived lineages are geographically clustered (Figs. 1 and S4; for Southern Sahul haplogroups, see [31–33, 54]).

Interestingly, haplogroup P includes some subhaplogroups restricted to Southern or Northern Sahul, but distantly connected as diverging from the same root P haplogroups (Fig. 1, Table S4, and Fig. S4). Coalescence ages and phylogenetic reconstructions suggest two unexpected pattern for these haplogroups, due either to the sharing of an ancestral population or a signature of ancient population connections between Southern and Northern Sahul.

On one hand, Northern Sahul hosts specific P haplogroups such P1, P2, P4a, and P13b, which are both more frequent and diverse in highland New Guinea, supporting their possible emergence/diversification in this region. These lineages have coalescence ages dating back to the early settlement phase, before 45 kya (Fig. 1, Table S4, and Fig. S4). These ancient P haplogroups specific to northern Sahul (P1, P2, P4a, and P13b) diverged from their sister Southern Sahul haplogroups within a very restricted time window (50–45 kya). Indeed, we observed that P1, P2, P10, and pre-P8 are related (Figs. 1 and S4). They diverged around 49 kya (95% CI 38–60 kya), in separate clades that today are geographically isolated: P10 in Wallacea, P2 and P1 in Northern Sahul, and pre-P8 in Southern Sahul. A similar pattern of an early split is observed for P4: ~50 kya (95% CI 42–58 kya) between the Northern (P4a) and Southern (P4b) Sahul subhaplogroups; and for P-13: ~49 kya (95% CI 34–64 kya) between the Northern (P13b) and Southern (P13a) subhaplogroups. This pattern suggests Northern and Southern Sahul populations shared an ancestral population that harbored high P diversity.

On the other hand, P1 haplogroups, the most frequent and diverse P haplogroup in Northern Sahul, are dominated by lineages diverged from highland New Guinea P1 haplogroups during the Last Glacial Maximum (18–28 kya) and spread geographically to other regions during the Last Glacial Maximum and postglacial periods: coastal New Guinea (Madang, P1d2 19 kya, P1d1 18 kya, and P1d5 16 kya), Near Oceania (Solomon Islands, P1g 26 kya, P1–152C 26 kya, P1d2 19 kya, and P1d1a 18 kya), Remote Oceania (Tonga, Fiji, Polynesia, P1d2 19 kya), Wallacea (Timor, P1–152C 26 kya, P1d 24 kya, P1d3 23 kya, P1d2 19 kya and Maranao in the Philippines, and P1d1c 17 kya) and Australia (P13b2 25 kya, P1–152C 26 kya; Torres strait ancestry individuals, Nagle et al.). Similarly Australian carriers of P3b1 have a Northern Sahul origin, supported by their Torres Strait ancestry and clustering with highland New Guinea P3b2 [31].

Demographic expansions

These settlement and expansion dates are corroborated by Bayesian skyline estimates obtained for each of the seven geographical regions when focusing only on P, M, and Q diversity (Fig. 1d). Near Oceania has the first population increase, starting soon after initial settlement (55 kya), while Australia displays a substantial population increase at a slightly later time but still in the same time frame. Coastal and highland New Guinea populations show smaller population increases at the same time as Near Oceania. We note that BSPs for Australia only include haplogroups shared with Northern Sahul (P haplogroups), which represents just a third of Indigenous Australian diversity [31–33]. However, this still represents a reasonable proxy for Australian diversity as all Australian lineages (O, P, S, and M) seem to show the same general process of expansion beginning ~50 kya following initial settlement of Australia [31–33].

All Sahul regions show population expansions in the postglacial warming period (18–10 kya), including within those lineages that later contributed to the initial settlement of Remote Oceania. Notice that Bayesian skyline estimates for Sunda and Wallacea are less reliable as these regions have many other haplogroups than those focused on here, and thus cannot be readily compared with inferences for northern Sahul where P, M, and Q comprise almost 100% of the local diversity.

Discussion

First phase of Sahul settlement

This study shows that the mtDNA diversity of Northern Sahul does not originate from Southern Sahul. This is

supported mainly by the fact that (1) mtDNA lineages found in New Guinea are not derived lineages of those found in Southern Sahul (Figs. 1 and S4), (2) the two regions of Northern and Southern Sahul host different deep rooted lineages with strong geographic structuring (Northern Sahul: M27, M28, M29, Q2, P1, P2, P10, P4a, and P13b; Southern Sahul: O, S, N13, M42a’c, P5, P6, P7, P8, and other P), and (3) all of these haplogroups are rooted in the age range of the initial settlement phase of Sahul (>50 kya), supported by most archeological and genetic evidence [4, 5, 30]. A similar rationale can be used to support the view that Southern Sahul mtDNA diversity is not derived from Northern Sahul diversity (Figs. 1 and S4 [31–33]).

The different origins of the Northern and Southern Sahul mtDNA profiles could be explained by two main hypothesis. Either (1) ancestral Sahul settlers carried all of the major haplogroups, and during their rapid dispersal within Sahul, carriers of specific haplogroup settled different regions, or (2) two (or more) groups of settlers carrying different haplogroup sets both reached Sahul during the early settlement phase (>50 kya), with one group settling Southern Sahul and the other group Northern Sahul.

The most parsimonious explanation, when considering the phylogeographic, phylogenetic, and coalescence age results, favors the second hypothesis: two (or more) groups of settlers likely originating from a common ancestral Sunda population. Indeed, (1) it is unlikely that a single group of settlers carrying all ancestral haplogroups could alone explain the strong geographic structure observed today—with two different profiles of ancestral lineages in Southern and Northern Sahul not derived from each other—either by a peculiar settlement pattern or genetic drift. (All haplogroups found in Northern Sahul would have to have been lost in Southern Sahul, and vice versa). (2) There are differences in age estimates of some of the Northern and Southern haplogroups suggesting different timings of arrival for various haplogroups. (3) Today, the oldest Sahul haplogroups M29’Q (~55 kya, 95% CI 42–67 kya) and M27 (~51 kya, 95% CI 40–62 kya) are restricted to the eastern part of Northern Sahul (Table S4 [47]). This is in line with different demographic expansion times in Northern and Southern Sahul, possibly related to at least two separate dispersal events (Fig. 1d). This finding is consistent with recent migration modeling suggesting entry points into both northern and southern Sahul were likely [7, 9].

Regarding the ancestral populations of Sahul, nuclear data support a common ancestral population of Northern and Southern Sahul settlers located in Sunda [21]. Strong population structure within this Sunda population can be postulated based on mtDNA diversity patterns (e.g., P haplogroup diversification with deep rooted P lineages found in Sunda, Northern, and Southern Sahul). Autosomal data is therefore more in agreement with the hypothesis of

two groups of Sahul settlers carrying different mtDNA lineages.

Regarding the routes used by the first settlers, our results do not allow us to disentangle the use of a northern (via Sulawesi and the Bird’s Head of New Guinea) or southern (via Timor and the northwest shelf of Australia) routes into Sahul, which from paleogeographic, environmental, and demographic reconstructions are both possible [7–10]. However, in the event of the arrival of two Sahul settler groups, the one leading to the settlement of Northern Sahul ending in the Bismarck archipelago may favor an origin within the maritime and coastal adapted cultures of Wallacea, and thus the use of the northern coastline route [10]. Similarly, the group leading to the settlement of Southern Sahul may favor the southern route across the savanna corridor running from Sunda (the Java plain) to Sahul (the Arafura plain) [59]. The use of both routes is finding some support in archeological evidence, with sites >40 kya old located along both paths [30]. However, lineages shared between Northern Sahul and Australia (Q2b, 30 kya), Timor (Q3–215, 21 kya) and Sunda (Q1d 14 kya) suggest that a southern route of interaction was active at a later period, possibly reflecting the more ancient settlement path [7].

Regarding the timing of the initial settlement of Sahul, while old age estimates of lineages agree with settlement over a rather narrow time window (55–45 kya), the detailed scenario is still unclear. The suggested 65 kya human occupation of Madjebebe (Northwestern Australia [5]) sits within the range of confidence intervals for some of the oldest Sahul haplogroups (M29’Q: 42–67 kya, M27: 40–62 kya) regardless of the mutation rate used (Table S4 and Fig. 1). On one hand, some of our results favor an early arrival of Near Oceania ancestors, based on the older age estimate of some northern Sahul haplogroups (M27, 51kya; M29’Q, and 55kya; Table S4) and early Near Oceania demographic expansions (Fig. 1d). On the other hand, the phylogeny of the P1, P2’P10, and pre-P8 cluster suggest that Southern Sahul haplogroup emerged earlier (pre-P8, 48kya) than Northern Sahul and Sunda related haplogroups (P2’10, 47kya). However, several potential biases—a limited number of southern haplogroups used for the Bayesian skyline analysis, demographic expansion occurring in Sunda rather than in Sahul, lineages lost in the last 50 kya and blurring the original demographic signal—do not allow clearer genetic timings.

The scenario we might postulate from these results suggests (1) the diversification of ancestral Sahul lineages in a population located in Sunda, (2) two (or more) dispersal events to Sahul within the same narrow time frame around 45–55 kya, (3) a dispersal to Sahul of a group of settlers carrying lineages observed today in Northern Sahul (M27, M28, M29’Q, P1, P2, P4a, P10, P13b), (4) another dispersal to Sahul of a group of settlers carrying the lineages

observed today in Southern Sahul (O, S, N13, M42a’c, P5, P6, P7, P8, and other P), and (5) an absence of detectable interactions (such as shared lineages and admixture) between Northern and Southern Sahul groups during the initial settlement phase and the following 20 ky years, until contacts increased after the Last Glacial Maximum.

Regionalism and isolation in Northern Sahul

Phylogeographic and phylogenetic analyses of northern Sahul haplogroups show strong geographic clustering of the different haplogroups and their derived lineages, suggesting that populations were structured early after the initial settlement of Northern Sahul (Figs. 1 and S1–S5). This is in good agreement with the nomadic sedentism and strong territoriality observed today in New Guinea societies [23, 60]. It also mirrors the “regionalism” observed in the mtDNA lineages of Indigenous Australians [33], suggesting that Northern and Southern Sahul populations exhibited similar regionalism patterns.

The main differences between Northern and Southern Sahul reflect the date at which this strong geographic structure appeared. Southern Sahul haplogroups (e.g., P, O, S, and M42a’c) show rapid diversification following initial arrival in their new environment, and then to a large extent, geographic stability with few large scale movements over the last 55 ky (Figs. 1 and S4 [33]). In Northern Sahul, haplogroups show more diverse patterns. In Near Oceania, haplogroups M27, M28, and M29 exhibit a geographic distribution restricted to a small geographical region over the last 55 ky, similar to that observed in Southern Sahul. On New Guinea, however, haplogroups Q and P and their derived lineages also have a restricted geographic distribution, but (1) this structure arose more recently in the Last Glacial Maximum, postglacial and Holocene periods, and (2) the structure resulted from both local lineage diversification and the arrival of new lineages from external regions, indicating some degree of movement within and beyond Sahul from the mid-Pleistocene onwards (Figs. 1 and S3, S4).

Another difference between Southern and Northern Sahul is the period of lineage isolation (longer branches on the tree with greater accumulation of mutations) before diversification. Following the initial settlement of Northern Sahul by 50 kya (Ivane valley 49 kya, Buang Merabak 45 kya [30]), we observe a pause of around 20 ky before most Northern Sahul lineages diversified (Figs. 1 and S1–S4). This suggests an unusual pattern: human occupation of a new territory without a concomitant genetic signal of demographic expansion. Diversification instead first occurred during the Last Glacial Maximum period (M28 32 kya, Q1 18 kya, Q2 30 kya, Q3 28 kya, P1 30 kya, and P2 19 kya), and may have been triggered by environmental changes such as the landmass increase during the Last Glacial Maximum (28–18 kya [57, 59]). This later

diversification period is also confirmed by distantly connected P subhaplogroups, which show later diversification in New Guinea (P4a 21 kya, P13b 19 kya) compared with their sister haplogroups in Southern Sahul (P4b 38 kya, P13a 40 kya), although all are present from the initial settlement period.

The reasons behind this observed longer isolation (Figs. 1 and S1–S4) and later diversification (Table S4) of many Northern Sahul haplogroups has still to be clarified, and is surprising considering the rapid adaptation of the first settlers to the new fauna, flora, and environments (tropical rainforests, semiarid plains, and upper mountain grasslands) encountered in New Guinea and Australia [30]. However, a combination of factors may have delayed population expansion in New Guinea itself—a harsh environment in northern mountainous Sahul, and greater shellfish coastal resources, competition with other occupants (e.g., mega-fauna or other members of the *Homo* genus)—or signals from this early period may have been lost in the current mtDNA gene pool.

Dispersal episodes from Northern Sahul

Northern and Southern Sahul lineages underwent a period of diversification immediately following initial settlement, and then subsequently during the Last Glacial Maximum, postglacial and Holocene periods (Figs. 1 and S1–S4). Some of the derived lineages resulting from these events are found outside their region of origin, and provide clues to ancient population interactions. This case is even more informative for Northern Sahul. While Southern Sahul haplogroups have very limited geographic ranges within Australia (expansion of the O2 derived lineage in the Holocene, and re-expansion in the Western central desert in postglacial period around 15 kya) [33], the Northern Sahul lineages show a much wider geographic range of expansion, also reaching Southern Sahul, the far east (Remote Oceania) and the far west (Sunda).

Northern Sahul hosts haplogroups that have locally diversified within New Guinea (P1 Q1, Q3) and Near Oceania (M27, M28, M29, and Q2), and from which a limited number of derived lineages have spread in neighboring regions (Figs. S1–S4 and 3).

First, our data support population interactions in both directions between New Guinea and Near Oceania, starting from the end of the Last Glacial Maximum and continuing into the postglacial and Holocene periods. New Guinea lineages are found in the neighboring region of Near Oceania, including preQ3b (16 kya), Q1b,c,e (14–5 kya), and several P1 subhaplogroups (26–18 kya). Lineage P1d2 was later carried into Remote Oceania during late Holocene. Near Oceania lineages spread within the same periods into New Guinea (M28a8 17 kya, Q2a4 12 kya, M29a,b 9–2 kya) and later to Remote Oceania (M29a,b around 2 kya).

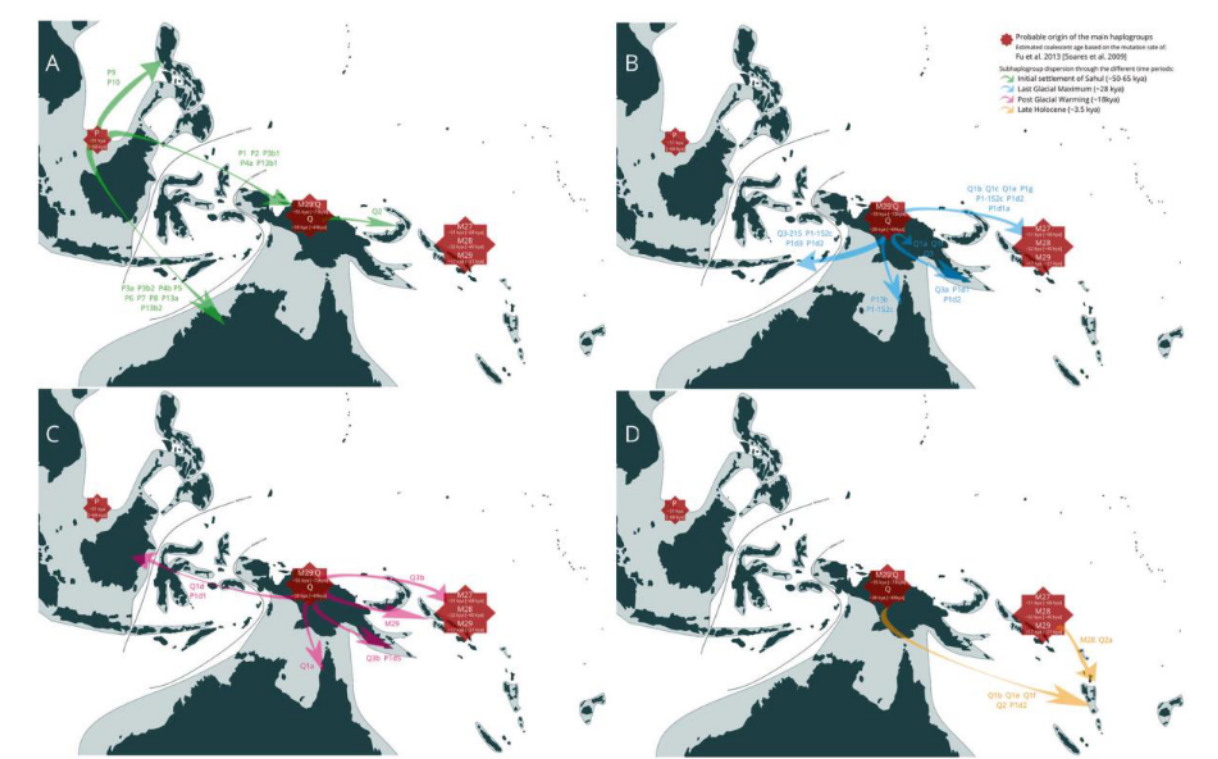


Fig. 3 Proposed movements of maternal lineages P, M, and Q. Dark shading represents modern coastlines; light shading illustrates the extent of the Sunda continent at the Last Glacial Maximum. Red octagrams represent the probable approximate origins of haplogroups. Arrows represent probable migration paths during: **a** the initial settlement of Sahul (~50 kya; green); **b** the Last Glacial Maximum (~28 kya; blue); **c** the postglacial warming period through to the Holocene (~18 kya; pink); and **d** the Late Holocene (~3.5 kya; orange)

This is in agreement with archeological evidence, which supports some level of connection between New Guinea and Near Oceania from the end of the Last Glacial Maximum (around 23 kya) based on animal, plant, and object (e.g., obsidian) translocation [23]. Interestingly, we did not detect genetic interactions for the earlier period lasting from the first settlement of Near Oceania 45 kya (Buang Merabak [13]) to the end of Last Glacial Maximum (18–30 kya). This may suggest that after being reached 45 kya by crossing the Vitiaz Strait from mainland New Guinea, this region of Near Oceania remained largely isolated from the rest of Sahul for more than 20 ky, despite the absence of known barriers that humans had already overcome to reach Near Oceania in the first place [61].

Second, our results also attest long-term interaction between Northern Sahul and Northern Australia from early in the Last Glacial Maximum (30 kya) to postglacial periods (18–10 kya), but in a unidirectional manner (from Northern to Southern Sahul). Indeed, haplogroups from New Guinea (P3b 36 kya, P1–152c 26 kya, P13b 20 kya, Q1a 15 kya) and Near Oceania (Q2b 30 kya) are present in Northern Australia. This result broadly matches those obtained from Y chromosome [26] and nuclear [21] data, although the Y chromosome

supports a much more recent split (9–12 kya) between Papuans and Indigenous Australians than nuclear DNA (10–32 kya [21]). Our mtDNA results indicate that interaction between New Guinea and southern Sahul stopped before the Holocene and before the geographical separation between Australia and New Guinea (6–8 kya) [6].

Third, Northern Sahul haplogroups from New Guinea are detected in Wallacea from the Last Glacial Maximum and postglacial warming periods onward (Timor Q3–215 21 kya, P1 subhaplogroups 19–26 kya), as well as on major regional islands (Q1d 14 kya in Taiwan and Q1d and P1d1c 17 kya in Philippines). Lineages moving during the last 30 kya from Northern Sahul to the Wallacea and Sunda regions would have been mediated by maritime interactions, in agreement with the “voyaging corridor” hypothesis and an increase in maritime interactions between Northern Sahul and Island Southeast Asia from the end of the Pleistocene [50, 62].

Conclusions

To summarize, our results suggest that lineage dispersals from Northern Sahul likely result from environmental

changes related to the Last Glacial Maximum and post-glacial periods. In the former period, colder, and dryer conditions in Northern Sahul led to increases in the Sahul landmass and may have motivated population movements within and outside this region without strong associated demographic expansions (Fig. 1d). Consistent with the limited range of lineages involved, this may reflect refugial movements (Figs. 1 and S1–S4). The more favorable conditions of the postglacial warming period (10–18kya) led to demographic expansion and accentuated lineage diversification (Fig. 1d), leading to the geographic dispersal of Northern Sahul lineages within and outside Sahul. Rising sea level during the Holocene saw an intensification of this pattern with geographically restricted demographic expansions, and ultimately, Pacific settlement from Near Oceania.

The maternal history of populations from Northern Sahul, one of the oldest continuous populations outside Africa, thus sheds light on the population history of this region. This study proposes an initial arrival to Sahul of two groups of settlers within the same broad time window (50–65 kya), each carrying a different set of maternal lineages, with one group settling Northern Sahul (New Guinea and Near Oceania), and one Southern Sahul (Australia). Following a period of a least 20 ky of relative isolation of Northern Sahul population, the cause of which is still unclear, the postglacial period after 30 kya stimulated lineage diversification and greater interactions within and beyond Northern Sahul, to Australia, Wallacea, and beyond. These lineage dispersals did not, however, erase the strong geographic structuring of the maternal lineages visible in Northern Sahul, which persists to the present.

Acknowledgements We thank Kylie Suseki, Roxanne Tsang, Jason Kariwiga, Kenneth Miampa, Tepsy Beni, Christopher Kinipi, and John Muke for assistance in collecting Papua New Guinea samples, and Alexander Hübner, Enrico Macholdt, and Roland Schröder for assistance with generating the mtDNA sequences. We also thank Pradiptajati Kusuma for helpful comments. We acknowledge support from the GenoToul bioinformatics facility of Genopole Toulouse Midi-Pyrénées the LabEx TULIP, France. We especially thank all of our study participants. We acknowledge the National Geographic Society’s support (Grant HJ-156R-17 to F-XR). This work was supported by the French Ministry of Research grant ANR-14-CE31-0013-01 (OCEOADAPTO) to F-XR, the French Ministry of Foreign and European Affairs (French Prehistoric Mission in Papua New Guinea to F-XR), and the French Embassy in Papua New Guinea; a fellowship from the Alexander von Humboldt Foundation to MPC; funding from the Max Planck Society to MS; and grant from COMPETE 2020 and Fundação para a Ciência e a Tecnologia (POCI-01-0145-FEDER-016609) to VF.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This study was approved by the Medical Research Advisory Committee of Papua New Guinea (National Department of

Health) under research ethics clearance MRAC 16.21 and by the French Ethics Committees (Committees of Protection of Persons IE-2015–837 (1)). Permission to conduct research in Papua New Guinea was granted by the National Research Institute of Papua New Guinea (permit 99902292358), with full support from the School of Humanities and Social Sciences, University of Papua New Guinea. Biological sampling was conducted by FXR, NB, and ML with the assistance of a team from the University of Papua New Guinea. All samples were collected with written informed consent.

Archive samples used in this study were collected between 1986 and 1988 under supervision and with regulatory approval from the Institute of Medical Research of Papua New Guinea and approval by the Medical Research Advisory Committee of Papua New Guinea (National Department of Health). Following a brief description of the project, a discussion with each individual willing to participate ensured that the project was fully understood. The description stated clearly that the aims were to use the blood of individuals to investigate settlement history and migration patterns in Papua New Guinea. Once participants had given oral consent, information was collected on their age, sex, date and place of birth, and their spoken language(s). All samples were anonymised and coded at the time of sampling. The DNA samples have been stored with approval from the IMR for future research. Contemporary approvals for the present work have been obtained from the French Ethics Committees (Committees of Protection of Persons IE-2015–837 (1)), and the Medical Research Advisory Committee of Papua New Guinea (National Department of Health) under research ethics clearance MRAC 16.21. (Revised paper: page 6, lines 136–149).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Loh J, Harmon D. A global index of biocultural diversity. *Ecol Indic.* 2005;5:231–41.
2. Bellwood PS, editor. First farmers: the origins of agricultural societies. Malden, USA: Blackwell Publishing Ltd; 2005.
3. Eberhard DM, Simons GF, and Fennig CD, editors. *Ethnologue: languages of the world*. 22nd ed. Dallas, Texas: SIL International; 2019.
4. O’Connell JF, Allen J, Williams MAJ, Williams NA, Turney CSM, Spooner NA, et al. When did Homo sapiens first reach Southeast Asia and Sahul? *Proc Natl Acad Sci USA.* 2018;115:8482–90.
5. Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L, Smith M, et al. Human occupation of northern Australia by 65,000 years ago. *Nature.* 2017;547:306–10.

6. Lewis SE, Sloss CR, Murray-Wallace CV, Woodroffe CD, Smithers SG. Post-glacial sea-level changes around the Australian margin: a review. *Quat Sci Rev.* 2013;74:115–38.

7. Norman K, Inglis J, Clarkson C, Faith JT, Shulmeister J, Harris D. An early colonisation pathway into northwest Australia 70-60,000 years ago. *Quat Sci Rev.* 2018;180:229e239.

8. Bird MI, Beaman RJ, Condie SA, Cooper A, Ulm S, Veth P. Palaeogeography and voyage modeling indicates early human colonization of Australia was likely from Timor-Roti. *Quat Sci Rev.* 2018;191:431e439.

9. Bird MI, Condie SA, O'Connor S, O'Grady D, Reepmeyer C, Ulm S, et al. Early human settlement of Sahul was not an accident. *Sci Rep.* 2019;9:8220.

10. Kealy S, Louys J, O'Connor S. Least-cost pathway models indicate northern human dispersal from Sunda to Sahul. *J Hum Evol.* 2018;125:59e70.

11. Bradshaw CJA, Ulm S, Williams AN, Bird MI, Roberts RG, Jacobs Z, et al. Minimum founding populations for the first peopling of Sahul. *Nat Ecol Evol.* 2019;3:1057–63.

12. Summerhayes GR, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, et al. Human adaptation and use of plants in highland New Guinea 49,000–44,000 years ago. *Science.* 2010;330:78e81.

13. Leavesley M, Chappell J. Buang Merabak: additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. *Antiquity.* 2004;78:301.

14. Wickler S. Prehistoric Melanesian exchange and interaction: recent evidence from the Northern Solomon Islands. *Asian Perspect.* 1990;29:135e154.

15. Kayser M, Lao O, Saar K, Brauer S, Wang X, Nürnberg P, et al. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet.* 2008;82:194–8.

16. Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, et al. Demographic history of Oceania inferred from genome-wide data. *Curr Biol.* 2010;20:1983–92.

17. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature.* 2016;538:510–3.

18. Haberle SG, Lentfer CI, Denham T. Palaeoecology. In: Golson J, Denham T, Hughes P, Swadling P, Muke J, editors. Ten thousand years of cultivation at kuk swamp in the highlands of Papua New Guinea. The Australian National University, Canberra, Australia: ANU Press; 2017, p. 145–62.

19. Donohue M, Tim Denham. Farming and language in island Southeast Asia reframing Austronesian history. *Curr Anthropol.* 2010;51:223.

20. Schapper A. Farming and the trans-new Guinea family: a consideration chapter. In: Robbeets M, Savelyev A, editors. Language dispersal beyond farming. Amsterdam: John Benjamins Publishing Compagny; 2017, p. 155–82.

21. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature.* 2016;538:207–14.

22. Bergstrom A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science.* 2017;357:1160–3.

23. Pawley A, Attenborough R, Golson J, Hide R, editors. Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples. Canberra, Australia: Australian National University; 2005.

24. Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol.* 2006;23:2234–44.

25. Mona S, Grunz KE, Brauer S, Pakendorf B, Castrì L, Sudoyo H, et al. Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol Biol Evol.* 2009;26:1865–77.

26. Bergström A, Nagle N, Chen Y, McCarthy S, Pollard MO, Ayub Q, et al. Deep roots for Aboriginal Australian Y chromosomes. *Curr Biol.* 2016;26:809–13.

27. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016;538:238–42.

28. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature.* 2016;538:201–6.

29. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple deeply divergent denisovan ancestries in Papuans. *Cell.* 2019;177:1010–21.

30. Summerhayes GR, Field JH, Shaw B, Gaffney D. The archaeology of forest exploitation and change in the tropics during the Pleistocene: the case of Northern Sahul (Pleistocene New Guinea). *Quat Int.* 2017;448:14e30.

31. Nagle N, van Oven M, Wilcox S, van Holst Pellekaan S, Tyler-Smith C, Xue Y, et al. Genographic Consortium. Aboriginal Australian mitochondrial genome variation—an increased understanding of population antiquity and diversity. *Sci Rep.* 2017;7:43041.

32. Nagle N, Ballantyne KN, van Oven M, Tyler-Smith C, Xue Y, Wilcox S, et al. Mitochondrial DNA diversity of present-day Aboriginal Australians and implications for human evolution in Oceania. *J Hum Genet.* 2016;62:343–53.

33. Tobler R, Rohrlach A, Soubrier J, Bover P, Llamas B, Tuke J, et al. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature.* 2017;544:180–4.

34. Brucato N, Fernandes V, Mazières S, Kusuma P, Cox MP, Ng'ang'a JW, et al. The Comoros shows the earliest Austronesian gene flow in East Africa. *Am J Hum Genet.* 2018;102:58–68.

35. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 2012;40:e3.

36. Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE.* 2010;5:e14004.

37. Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. High resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in North-west Amazonia. *Am J Phys Anthropol.* 2018;165:238–55.

38. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* 2009;30:E386–94.

39. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet.* 1999;23:147.

40. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.

41. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat.* 2011;32:25–32.

42. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet.* 2009;84:740–59.

43. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol.* 2013;23:553–9.

44. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS.* 1997;13:555–6.

45. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 2005;22:1185–92.

46. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol.* 2005;128:415–23.

47. Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G, et al. Melanesian mtDNA complexity. *PLoS ONE.* 2007;2:e248.

48. Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, et al. Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *Am J Hum Genet.* 2014;94:721–33.

49. Brandão A, Eng KK, Rito T, Cavadas B, Bulbeck D, Gandini F, et al. Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. *Hum Genet.* 2016;135:363–76.

50. Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, et al. Ancient voyaging and Polynesian origins. *Am J Hum Genet.* 2011;88:239–47.

51. Stoneking M, Jorde LB, Bhatia K, Wilson AC. Geographic variation in human mitochondrial DNA from Papua New Guinea. *Genetics.* 1990;124:717–33.

52. Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, et al. Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol.* 2005;22:1506–17.

53. Vilar MG, Kaneko A, Hombhanje FG, Tsukahara T, Hwaihwanje I, Lum JK. Reconstructing the origin of the Lapita Cultural Complex: mtDNA analyses of East Sepik Province, PNG. *J Hum Genet.* 2008;53:698–708.

54. Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci USA.* 2007;104:8726–30.

55. Gomes SM, Bodner M, Souto L, Zimmermann B, Huber G, Strobl C, et al. Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocene mtDNA diversity. *BMC Genomics.* 2015;14:16.70.

56. Ricaut FX, Thomas T, Mormina M, Cox MP, Bellatti M, Foley RA, et al. Ancient Solomon Islands mtDNA: Assessing Holocene settlement and the impact of European contact. *J Archaeol Sci.* 2010;37:1161–17.

57. Hope G. The sensitivity of the high mountain ecosystems of New Guinea to climatic change and anthropogenic impact. *Arct Antarct Alp Res.* 2014;46:777e786.

58. Pavlides C, Gosden C. 35,000 year-old sites in the rainforests of West New Britain, Papua New Guinea. *Antiquity.* 1994;69:604–10.

59. Wurster CM, Bird MI. Barriers and bridges: early human dispersals in equatorial SE Asia. In: Harff J, Bailey G, Lith F, editors. *Geology and Archaeology: submerged landscapes of the continental shelf. Special Publications (411).* London, UK: Geological Society of London; 2016. p. 235–50.

60. Keesing RM, Strathern AJ, editors. *Cultural anthropology: a contemporary perspective*, 3rd ed. USA: Wadsworth Publishing; 1997.

61. Irwin G, editors. *The prehistoric exploration and colonization of the Pacific.* Cambridge, UK: University Press; 1992.

62. Solheim WGII, Bulbeck D, Flavel A, editors. *Archaeology and culture in Southeast Asia: unraveling the Nusantara.* Diliman, Quezon City: University of Philippines Press, Philippines; 2006.

PAPER II. FIRST INSIGHT INTO ORAL MICROBIOME DIVERSITY IN PAPUA
NEW GUINEANS REVEALS A SPECIFIC REGIONAL SIGNATURE

Received: 13 April 2022 | Revised: 8 September 2022 | Accepted: 15 September 2022
DOI: 10.1111/mec.16702

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

First insight into oral microbiome diversity in Papua New Guineans reveals a specific regional signature

Nicole Pedro^{1,2,3} | Nicolas Brucato⁴ | Bruno Cavadas^{1,2} | Valentine Lisant⁴ | Rui Camacho^{5,6} | Christopher Kinipi⁷ | Matthew Leavesley^{8,9,10} | Luisa Pereira^{1,2} | François-Xavier Ricaut⁴

¹Instituto de Investigação e Inovação em Saúde, Universidade do Porto (i3S), Porto, Portugal
²Instituto de Patologia e Imunologia Molecular, Universidade do Porto (Ipatimup), Porto, Portugal
³Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto (ICBAS), Porto, Portugal
⁴Laboratoire Évolution and Diversité Biologique (EDB UMR5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France
⁵FEUP – Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
⁶INESC TEC-Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Universidade do Porto, Porto, Portugal
⁷Medical Clinic, University of Papua New Guinea, Port Moresby, Papua New Guinea
⁸Strand of Anthropology, Sociology and Archaeology, School of Humanities and Social Sciences, University of Papua New Guinea, National Capital District, Papua New Guinea
⁹College of Arts, Society and Education, James Cook University, Cairns, Queensland, Australia
¹⁰ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Wollongong, Wollongong, New South Wales, Australia

Correspondence
François-Xavier Ricaut, Laboratoire Évolution and Diversité Biologique (EDB UMR5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, 118 route de Narbonne, Bat 4R1, 31062 Toulouse, France.
Email: francois-xavier.ricaut@univ-tlse3.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: PAPUAEVOL n° ANR-20-CE12-0003-01; European Regional Development Fund, Grant/Award Number: POCI-01-0145-FEDER-007274 and POCI-01-0145-FEDER-022184; Fundação para a Ciência e a Tecnologia, Grant/Award Number: SFRH/BD/136299/2018

Handling Editor: Tatiana Giraud

Abstract

The oral microbiota is a highly complex and diversified part of the human microbiome. Being located at the interface between the human body and the exterior environment, this microbiota can deepen our understanding of the environmental impacts on the global status of human health. This research topic has been well addressed in Westernized populations, but these populations only represent a fraction of human diversity. Papua New Guinea hosts very diverse environments and one of the most unique human biological diversities worldwide. In this study we performed the first known characterization of the oral microbiome in 85 Papua New Guinean individuals living in different environments, using a qualitative and quantitative approach. We found a significant geographical structure of the Papua New Guineans oral microbiome, especially in the groups most isolated from urban spaces. In comparison to other global populations, two bacterial genera related to iron absorption were significantly more abundant in Papua New Guineans and Aboriginal Australians, which suggests a shared oral microbiome signature. Further studies will be needed to confirm and explore this possible regional-specific oral microbiome profile.

Luisa Pereira and François-Xavier Ricaut should be considered joint senior authors.

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd.

KEYWORDS
diet, environment, oral microbiome, Papua New Guinea, Sahul

1 | INTRODUCTION

In recent years the field of host-associated microbes has greatly increased due to a desire to understand how the microbiome impacts a broad array of hosts, such as animals, plants and algae (Parfrey et al., 2018). Microbes have been shown to be involved in host immunity, digestion and nutrient scavenging activities, facilitate metabolism and cellular growth, regulate behaviour, and release microbial toxins, among others (Miller et al., 2021). However, the first step in understanding host-microbiota dynamics is to identify the composition of the microbiome, and for that unique microbiomes across diverse host populations must be studied (Parfrey et al., 2018).

The human oral microbiota is one of the most complex microbial communities in the human body, colonized by around 700 different types of microorganisms (Lamont et al., 2018). Awareness of the role of oral microbes in health and disease has been increasing, not only in its obvious relation to oral diseases, such as periodontal disease (Costalonga & Herzberg, 2014) and oral cancer (Wang & Ganly, 2014), but also in how it relates to systemic (Graves et al., 2019), cardiovascular (Bryan et al., 2017) and gastrointestinal diseases (Flemer et al., 2018). The oral microbiome is formed mainly by bacteria, but it also contains lower amounts of fungi, viruses, archaea, and protozoa. The top bacterial phyla detected in the oral microbiome are Firmicutes, Fusobacteria, Proteobacteria, Actinobacteria, Bacteroidetes and Spirochaetes (Deo & Deshmukh, 2019), but abundance of these taxa differs between populations. Several studies on oral microbiome have been performed for Westernized populations (Burcham et al., 2020; Nasidze et al., 2009), which have been associated with a specific Western diet, low in fibre and fruit and high in refined sugars and fat (Cordain et al., 2005), but also in more remote populations, such as Native South Americans (Contreras et al., 2010), native Alaskans (Li et al., 2014), Philippines hunter-gatherers (Lassalle et al., 2018) and Batwa Pygmies (Nasidze et al., 2011). These studies revealed a high diversity in oral microbiome within and between the different populations. Different environments, lifestyles, eating habits, or genetic background are possible reasons for these variations (Li et al., 2014; Mousa et al., 2022).

Papua New Guinea (PNG) is one of the most culturally diverse and least urbanized countries in the world, with over 800 spoken languages and thousands of ethnic groups (Pawley et al., 2005). It also hosts a very high human genetic diversity (Brucato et al., 2021), including the highest proportion of archaic introgressions from Denisovans (~4%) in the world (Jacobs et al., 2019; Reich et al., 2011). The majority of Papua New Guineans live in a semi-traditional way with an agriculture-based diet, including sago, sweet potato, taro, cassava, yam, rice, maize, breadfruit, various nuts, sugarcane, pig, chicken and fruits, like coconut, banana and pineapple (Schapper, 2017). The PNG populations have a history

of low levels of protein consumption, which has been improving in the last years with an increase in the importation of poultry and fish (Schmidt & Fang, 2021). However, with the importation of animal-sourced foods there has also been an increased importation of ultra-processed foods, and meals containing higher saturated fat (Schmidt & Fang, 2021). PNG has poor general health due to low socioeconomic status and high burden of infectious diseases, such as pneumonia, malaria, tuberculosis and meningitis (Riley, 2009). While the gut microbiota of different regions within PNG showed higher bacterial diversity and different abundance profiles in comparison with Westernized populations (Martinez et al., 2015), a comprehensive analysis of the oral microbiome for diverse ethnic groups and regions in PNG has not been performed.

In this work we combined the results from two different methodologies (Figure 1) to characterize the microbial community from DNA extracted from PNG saliva samples: (1) Qualitative results from a microbial detection array, which contains specific probes to the different microbial species, and (2) quantitative from whole genome sequencing (WGS) reads unmapped to the human genome, which were aligned to a microbial database. The array methodology is cost-effective, highly sensitive and accurate (Thissen et al., 2014), and easy to analyse (Gardner et al., 2010; McLoughlin, 2011), being the ideal technique to validate an off-target microbiome inference from the human WGS. In fact, the recovery of off-target reads obtained from human WGS provides researchers with excellent secondary opportunities to study data beyond the intended target regions (Cavadas et al., 2020; Gomes et al., 2019; Samuels et al., 2013). This technique permits users to detect a larger number of microorganisms (Ranjan et al., 2016) and to obtain the abundance of the identified microbial communities. However, it is more expensive than the array methodology and involves much more complex bioinformatic analysis.

The quantitative oral microbiome data allowed us to assess differences between diverse PNG populations in light of diet and/or environmental factors, and against other worldwide populations. Our working hypothesis was that a specific oral microbiome signature might be found in these populations due to adaptation to their specific biodiversity, cultural lifestyle or genetic background, since their 50kya-establishment in the ancient Sahul region (New Guinea/Australia).

2 | MATERIALS AND METHODS

2.1 | Samples

Ninety-two unrelated individuals from different regions of PNG contributed saliva samples during field seasons between 2016 and 2019. Four different sampling sites were selected in order to

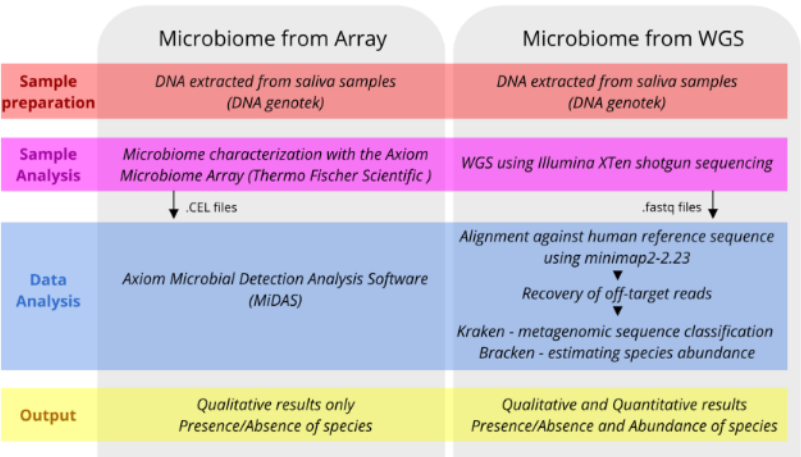


FIGURE 1 Schematic representation of the experimental strategy (sample preparation, sample analysis, data analysis and output) for the two microbial detection techniques.

access (1) population from different environments (coastal, highlands, lowlands), (2) different lifestyles (communities with a traditional subsistence or urban lifestyles and with variable access to health centers, grocery shops, medicine and Westernized food), and (3) being geographically distinct from each other. This sample set permits capture of some of the diversity of the PNG populations by defining five main groups: Port Moresby, includes individuals sampled in the east coastal lowlands ($n = 20$, Central, Oro and Milne Bay provinces); Simbu province, Highlanders sampled in the Highlands ($n = 20$, Simbu province); East Sepik province, individuals sampled in the north inland lowlands ($n = 20$); Western province, individuals sampled in the west coastal lowlands ($n = 20$); and East highlanders living in Port Moresby, Highlanders sampled in the east coastal lowlands ($n = 12$), (Figure 2a, Table S1). The saliva samples were collected with the Oragene DNA OG-500 collection kit (DNA Genotek Inc.), and the nucleic acids were extracted following the manufacturer's instructions.

2.2 | Ethics approval

This study was approved by the Medical Research Advisory Committee of PNG (National Department of Health) under research ethics clearance MRAC 16.21 and by the French Ethics Committees (Committees of Protection of Persons IE-2015-837 [1]). Permission to conduct research in PNG was granted by the National Research Institute of PNG (permit 99902292358), with full support from the School of Humanities and Social Sciences, University of PNG.

All samples were collected from healthy unrelated adult donors, all of whom provided written informed consent. In each sampling location, a full presentation of the project was made, followed by discussion with each donor to ensure that they fully understood the project. Participants were surveyed for language affiliation(s), current residence, date and place of birth, and a short genealogy up to three generations to establish regional ancestry.

2.3 | Microbiome inference from Axiom Microbiome Array

From the DNA extracted, we performed microbiome diversity characterization with the Axiom Microbial Detection Array (MDA) from Thermo Fisher Scientific. First, cDNA synthesis was performed using the SuperScript VILO cDNA synthesis kit from Invitrogen. Afterwards, microbiome diversity was characterized with the MDA, which can rapidly identify the microbial community at the species or strain level, giving a profile of which microorganisms are present in the analysed samples. The Microbiome array harbours a total of 1.38 million probes that detect over 12,513 microbial species including archaea, bacteria, fungi, protozoa, and viruses, and the probes were designed against both conserved regions (for the detection of higher taxonomic ranks) and unique regions (species/strain specific) of the microbial genomes with a sensitivity of 100–1000 genome copies and achieving 100% accuracy in species identification. The 96-well array plate version of the Microbiome array was used to run all the samples on the GeneTitan instrument (Thermo Fisher Scientific). A volume of 20 μ l of DNA from all samples was used as input into the MDA, including a no-template control and a positive control for quality control. The standard manufacturer's protocol was followed for all steps of the workflow. Quality control (QC) was performed using the QC option in the Axiom Microbial Detection Analysis Software (MiDAS), which analyses specific probes included in the Microbiome array to target regions of the human genome that are nonpolymorphic (sites that do not vary in sequence from one individual to the next). This QC analysis generates a metric termed Dish QC (DQC) and applies a threshold of DQC above 0.95 for the samples to be considered as having passed QC. Six samples were removed from the analysis after QC: PNG237 and PNG238 (East Sepik province), PNG128, PNG514 and PNG525 (Western province), and PNG205 (East highlanders living in Port Moresby). The remaining samples ($n = 86$) were then further analysed with MiDAS, which analyses all targets considered positive

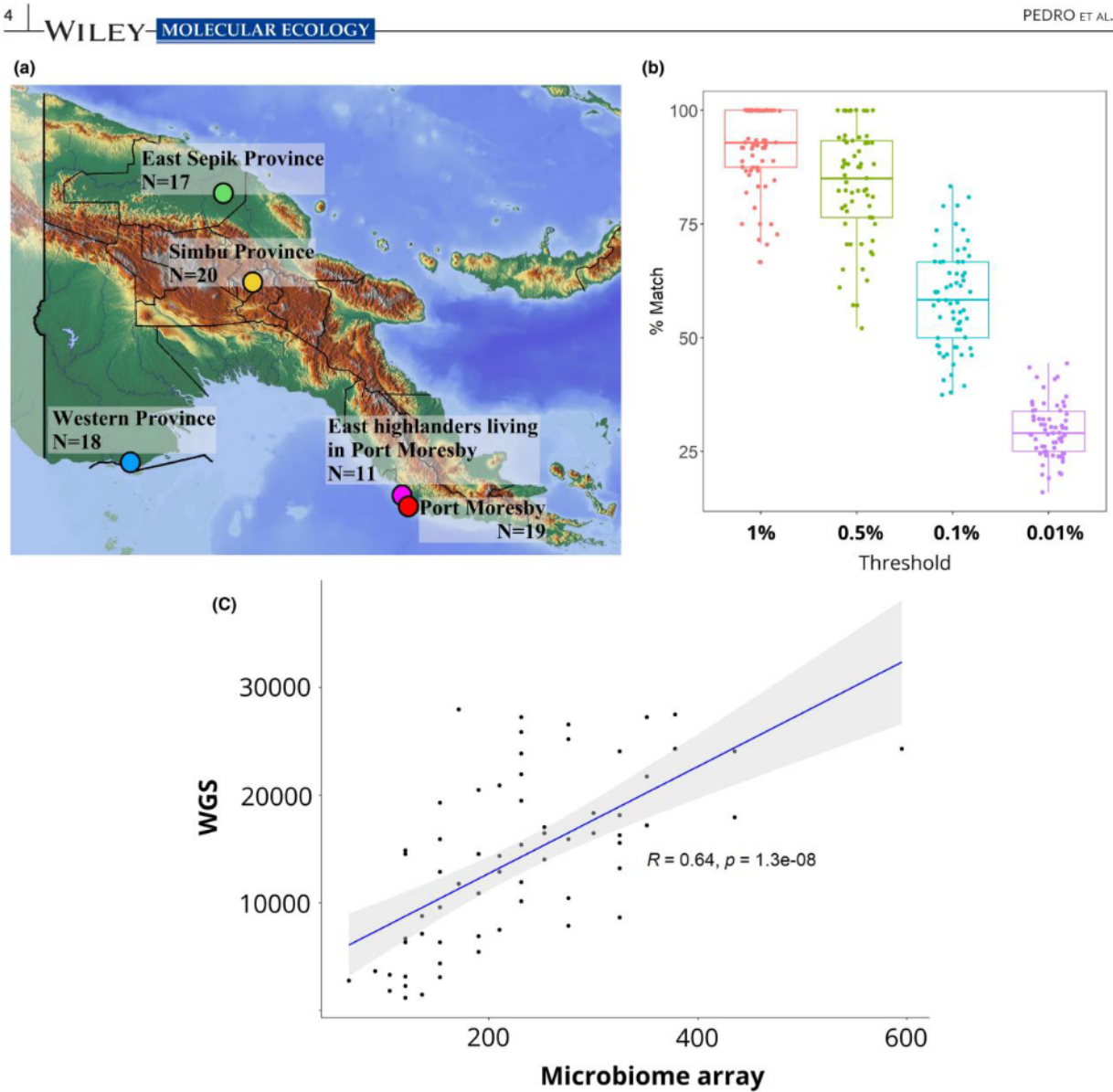


FIGURE 2 (a) Locations of the five sampled groups in PNG. (b) Box-plot representing the concordance in percentage of genera inferred in both methods by applying a threshold of abundance to the WGS results of 1%, 0.5%, 0.1% and 0.01% (Table S3). (c) Correlation of alpha-diversity values (Chao1 index) obtained for each individual oral microbiome characterized by both microbiome array and WGS methodologies. A regression line for the linear model (in blue) and the confidence interval (grey shadow) was added as well as the R and p -value for the model.

and reaches a consensus for the microbial community present in the samples. Positive targets are the probes with signal intensity above the 99th percentile of the random control probe intensities and with more than 20% of target-specific probes. These probe sequences are available from Thermo Fisher Scientific as part of the MiDAS software. To note, this technology is unable to provide a view of the varying prevalence of the microorganisms present in analysed samples.

2.4 | Microbiome inference from human whole genome sequences

For 68 individuals, also characterized by Microbiome array, we used reads extracted from WGS from (1) previously published studies (Brucato et al., 2021), and (2) from newly unpublished whole genome sequences data (Table S1). The sequencing procedure has been described previously (Brucato et al., 2021). Briefly,

sequencing libraries were prepared using the TruSeq DNA PCR-Free HT kit. About 150-bp paired-end sequencing was performed on the Illumina HiSeq X5 sequencer. We aligned the fastq files against the human reference GRCh38 (GCA 000001405.15) using MINIMAP2-2.23 (Li, 2018) and the unmapped reads were retrieved. For taxonomic assignment of the unmapped sequencing reads we resorted to KRAKEN 2 tool (Wood et al., 2019) using the standard plus protozoa and fungi database (PlusPF; includes archaea, bacteria, viral, plasmid, human, UniVec_Core, protozoa and fungi; updated on 27 January 2021). We then used Bracken (Lu et al., 2017; Bayesian Re-estimation of Abundance after Classification with KRAKEN) to estimate abundance at the genus level, using the taxonomic assignments made by Kraken. Since these PNG populations are understudied, we are unable to consider possible unidentified microbiota species that may be specific to them. We therefore performed all analysis on genus level to reduce the bias caused by using the Microbiome array and the WGS databases. We applied Bracken's default threshold of 10 reads as the minimum number of reads required for a classification at the specified rank. To remove low background noise on the WGS results due to minor contaminants, we applied a threshold for which all genera below a given percentage would be removed. This threshold was determined by calculating the best correlation of alpha-diversity values obtained from the Chao1 index between each sample using the qualitative and quantitative results, applying several thresholds, from 1% to 0.0001% (1%, 0.5%, 0.1%, 0.05%, 0.01%, 0.005%, 0.001%, 0.0005%, 0.0001%). Applying a threshold of 0.01% to the WGS results, the Chao1 values obtained from both methodologies presented the best correlation from all the tested thresholds.

2.5 | Microbial composition

We created phylogenetic trees and calculated the respective abundance of the genera found in the five groups from the qualitative results (Microbiome array). The abundance in this case refers to the percentage of samples in the group in which the specific taxon was detected. We used R STUDIO software (R Studio Team, 2020) to perform pairwise chi-squared tests and correction with Holm's method to see which genera were significantly different. Venn diagrams with the percentage of bacteria and virus were created to help to visualize the similarities and differences between the five groups resorting to the webtool Bioinformatics and Evolutionary Genomics (<https://bioinformatics.psb.ugent.be/webtools/Venn/>). Discrimination between the five different groups using quantitative data was performed with the linear discriminant analysis effect size (LEfSe) algorithm (Segata et al., 2011) with the calculation of a linear discriminant analysis (LDA) score, available at the Huttenhower laboratory Galaxy server (<http://huttenhower.sph.harvard.edu/galaxy/>) (Afgan et al., 2018). Statistical differences were found under a Kruskal-Wallis test which identified the features with significant variation across the five groups. Then, pairwise Wilcoxon tests were performed between each pair of groups, using the "one-against-all"

option for multiclass analysis, which only requires that one of the pairwise Wilcoxon tests to be significant in order to keep the feature in the list of significantly discriminating features (Segata et al., 2011). Only taxa with a *p*-value <.05 and with a LDA value >3 were considered enriched.

2.6 | Alpha-diversity

The alpha-diversity of each sample was calculated by estimating the genus and species richness, evenness, and diversity using the PHYLOSEQ package version 1.23.1 (McMurdie & Holmes, 2013). Alpha-diversity measures were calculated differently for qualitative and quantitative results. For the qualitative results (from the Microbiome array data) we calculated Chao1 (Hughes et al., 2001), which measures the richness of the samples and is more suitable for presence/absence data. We created a ggplot for each alpha-diversity metric and performed pairwise comparisons using t tests and correction with Holm's method using RSTUDIO (R Studio Team, 2020). After alpha-diversity analysis, one sample was considered an outlier, PNG54 (Port Moresby), as its Chao1 index value falls out of the mean ± 3SD for all samples (Figure S1), so it was removed from further analyses.

2.7 | Beta diversity and unrooted neighbour-joining tree

The phylogenetic information for the genera present in our samples was retrieved (on 20 October 2021) from the Time Tree website (<http://www.timetree.org/>). This information was used to calculate Jaccard (Koleff et al., 2003), unweighted and weighted UniFrac distances (Lozupone et al., 2011), beta diversity measures that analyse the similarities and differences among groups. The Jaccard and the unweighted UniFrac distances were applied to the qualitative data (Microbiome array) as they do not consider abundances in the calculations and the weighted UniFrac distances were used on the quantitative data (WGS results) The unrooted neighbour-joining tree was used to represent distances between the five groups. A permutation multivariate analysis (1000 permutations) of variance (PERMANOVA; Anderson, 2017) of the beta diversity distances (to reflect the phylogenetic diversity between genera) was also performed and plotted in a nonmetric multidimensional scale (nMDS). The null hypothesis tested by PERMANOVA was that the centroids of each group are equivalent, under the assumption of exchangeability of the samples among groups.

2.8 | Comparative analysis of oral microbiome data from other populations

The qualitative results (Microbiome array) were used to perform the analysis of the PNG oral microbiome (alpha-diversity) as we have more available samples. To have a global perspective of the oral

microbiome of PNG populations in the global oral microbiome we compared the abundance of the core microbiome obtained from the quantitative results (WGS) in this study with other published studies for Philippines hunter-gatherers and farmers, Batwa Pygmies, Aboriginal Australians and Westernized individuals (Handsley-Davis et al., 2021; Lassalle et al., 2018; Nasidze et al., 2011). For the hunter-gatherers and traditional farmer groups from the Philippines and their curated data of the Western controls from the Human Microbiome Project (Lassalle et al., 2018), we downloaded their published microbial metagenomic read data sets from ENA (www.ebi.ac.uk/ena; BioSample accession nos. ERS1202862–ERS1202885) and processed their data set with our workflow. These files were merged with our data set to perform beta-diversity analysis and to calculate the abundances of the core microbiome. For the Australian Aboriginals and Torres Strait Islanders (Handsley-Davis et al., 2021) and the Batwa Pygmies populations (Nasidze et al., 2011), we were not able to obtain WGS raw data, as they only sequenced a region of the microbial 16S rRNA gene (V4 for the Australian Aboriginals and Torres Strait Islanders and V1 and V2 for the Batwa pygmies). For these, we retrieved the values of abundance for the core microbiome that were described and available in their publications. Previous research has shown that microbiome results might differ quite substantially depending on the underlying method used (Durazzi et al., 2021; Hillmann et al., 2018), but a high correlation has been found between these different methods regarding the most abundant species in the samples. As such, we performed comparative analysis of only the most abundant taxon at genus level to limit inter-study variation due to the comparison between results from different methodologies (WGS and 16S).

3 | RESULTS

A total of 85 samples from the Microbiome array (qualitative results) were analysed in this study divided into five groups, based on the sampling locations and individual origins from: Port Moresby ($n = 19$); Simbu province ($n = 20$); East Sepik province ($n = 17$); Western province ($n = 18$); and East highlanders living in Port Moresby ($n = 11$) (Figure 2a). Our cohort includes 71 males (84%) and 14 females (16%), with an average age at time of collection of 38.6 years (range from 20–67 years; Table S1). After QC, 81 microbial genera were detected in the cohort. We controlled for substructure due to age or gender, and found no correlation (for age, p -value = .36; Figure S2A; for gender, p -value = .4; Figure S2B), confirming previous results (Nasidze et al., 2009).

A subset of the cohort was also characterized using available WGS to obtain quantitative results ($n = 65$): Port Moresby ($n = 13$); Simbu province ($n = 18$); East Sepik ($n = 9$); Western province ($n = 18$); and East highlanders living in Port Moresby ($n = 7$) (Table S1). Reads from WGS data not mapping to the human genome were analysed with KRAKEN 2 (Wood et al., 2019), resulting in an average of 18,652,535 reads matching microbial genomes per individual (ranging from 1,483,445–44,976,955 reads). As expected, the

WGS data identified a much higher number of genera (1757 genera; Table S2) than with the Microbiome array data. In fact, the BRAKEN method used in the WGS inference limits the number of reads to 10, while the Microbiome array has a sensitivity of 100–1000 genome copies. Hence, when comparing the concordance in inferred genera between the two methods, these values are high for the most abundant genera (Figure 2b; Table S3): an average of 92% match filtering to the genera with abundance above 1% (28 out of the 65 samples achieving 100% match); 84% when filtering to the genera with abundance above 0.5% (11 out of 65 samples reaching a 100% match); 58% when lowering the threshold to 0.1%; and 30% for threshold 0.01%. Even so, the overall diversity (alpha-diversity, Chao1) per sample is highly correlated between the two methods when using all genera above 0.01% of presence ($R = .64$; p -value = 1.3×10^{-08} ; Figure 2c). Further analyses based on WGS were done for this threshold.

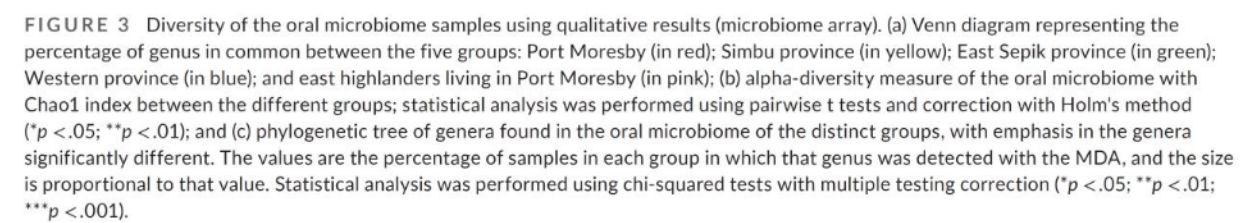
3.1 | Core oral microbiome in PNG

Our qualitative analysis (Microbiome array) of the PNG oral microbiome detected 75 bacterial genera and nine viral genera (Table S4). Although the array includes probes for RNA viruses and we performed RNA to cDNA reaction, we were not able to identify any RNA viruses in the PNG samples. This is most probably due to the storage of samples at room temperature, leading to the instability of RNA molecules. Around 43% of the detected bacteria genera are common to all the groups (Figure 3a). Of these, there are 12 genera that were present in more than 50% of the samples in all groups, namely *Veilonella*, *Streptococcus*, *Prevotella*, *Fusobacterium*, *Rothia*, *Haemophilus*, *Neisseria*, *Porphyromonas*, *Treponema*, *Leptotrichia*, *Actinomyces* and *Selenomonas* (Figure S3).

On exploration the quantitative results (WGS), the 12 genera mentioned above account for 80% of the oral microbiome of the PNG population (Table S5). *Prevotella* (~17%), *Neisseria* (~13%) and *Streptococcus* (~13%) are the three most abundant taxa. They represent more than 40% of the PNG oral microbiome, followed by *Haemophilus* (~9%), *Rothia* (~8%), *Veilonella* (~7%), *Fusobacterium* (~4%), *Porphyromonas* (~2%), *Actinomyces* (~2%), *Leptotrichia* (~2%), *Treponema* (~2%) and *Selenomonas* (~1%). These 12 taxa represent the core oral microbiome of PNG populations. The remaining 20% is composed by lower abundance taxa which differentiate PNG groups.

3.2 | Oral microbiome differences between PNG populations

We detected significant differences between PNG groups at the global level of oral microbiome, comparing their alpha-diversity (Chao1), based on qualitative data (Microbiome array; Figure 3b). The East Sepik group, the most remote (as in difficult to reach by transportation, and with less access to health care services, electricity,



unweighted UniFrac; Figure S4), although the PERMANOVA test was not significant for the East Sepik differentiation ($p > .05$). The East Sepik has one of the highest percentages of unique genera (8%; Figure 3a) and only shares 54% of genera with the four other groups, while the values are higher for the other groups: 57% for

Port Moresby; 67% for Simbu Province; 70% for Western Province; and 62% for East highlanders living in Port Moresby. This regional particularity is partly explained by the significantly higher frequency (in number of individuals of the group) of specific taxa such as *Vibrio*, *Capnocytophaga*, *Oribacterium*, *Peptostreptococcus*, *Megasphaera*, *Granulicatella*, *Gemella*, *Mycoplasma* and *Atopobium* when compared to each of the other PNG groups ($p < .05$, Figure 3c).

Quantitatively, the East Sepik oral microbiome is significantly enriched (medium abundance in the group) in several taxa such as *Haemophilus*, *Tannerella*, *Anabaena*, *Pseudonocardia*, *Aromatoleum*, *Actinobacillus*, *Histophilus*, *Dickeya*, *Aquitaea* and *Dietzia* (LDA>3, Figure 4).

The Western province group, another relatively remote coastal community, shows one of the highest alpha-diversities and the highest number of specific taxa in PNG (8%, Figure 3a and b). Although qualitatively, no genera were present in more samples in the Western province group than in the rest of PNG, a high number of taxa are quantitatively enriched (LDA>3, Figure 4) such as *Klebsiella*, *Fusobacterium*, *Starkeya*, *Corynebacterium*, *Anaerococcus*, *Mogibacterium*, *Liberibacter*, *Enterocloster*, *Parvimonas*, *Mediterraneibacter*, *Aeromonas*, *Anaerocolumna* and *Pseudomonas*.

Groups from Port Moresby, Simbu province and East highlanders living in Port Moresby show relatively similar alpha-diversity in the oral microbiome (Figure 3b). Qualitative data show that the Simbu province group has 6% of unique genera while the Port Moresby group has no unique genera (Figure 3a). No statistical difference in the presence of specific taxa was observed despite living in very different geographical regions (e.g., capital city on the southeast coast and east highlands respectively, Figure 3b). Correspondingly, the group of East highlanders living in Port Moresby shows no significant difference in the presence of taxa and only 3% of unique taxa (Figure 3a). However, each of these three groups can be characterized by the enrichment of some taxa like *Cronobacter* and *Flaviflexus* in Port Moresby, *Streptococcus* and *Gemella* in Simbu, and *Schaalia* in the East highlanders living in Port Moresby (LDA>3, Figure 4). The group of East highlanders living in Port Moresby shares 59% of their genera with the Simbu group but only 49% with the Port Moresby group (Figure 3a). The Alpha-diversity values in the East highlanders living in Port Moresby are more similar to the one from the Simbu group than to the Port Moresby group (Figure 3b). This lack of significant differences might be due to the incorporation of both microbiotas (initial oral microbiota from the highlands and the

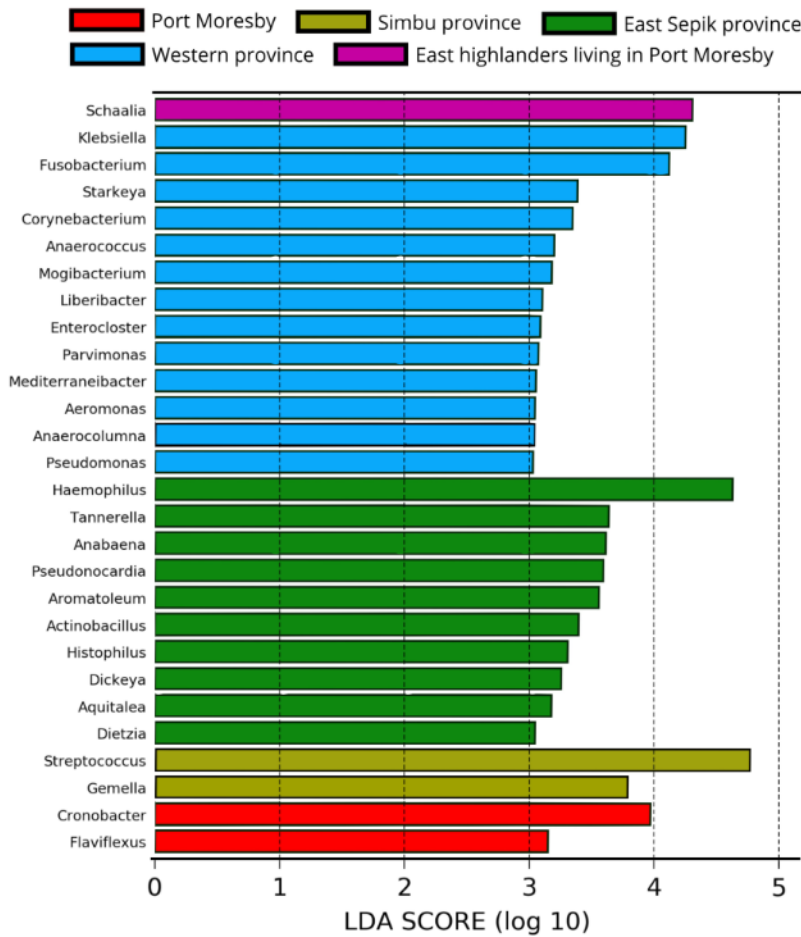


FIGURE 4 Profile of microbial composition of oral microbiome based on quantitative data (WGS results) with the taxa discriminating between the five distinct groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and east highlanders living in Port Moresby (in pink). We applied the LEfSe algorithm and LDA calculation in a "one-against-all" multi-class analysis, requiring only one pairwise Wilcoxon tests to be significant. Only discriminating features with $p < .05$ and LDA score over 3 are presented.

acquired microbiota from the coast) in the group of East Highlanders living in Port Moresby.

3.3 | Oral microbiome of PNG and other populations

To have a broader perspective of the oral microbiome of PNG population samples, we merged the WGS results from our PNG populations with the available WGS samples from Philippines hunter-gatherers, Philippines traditional farmers and Western controls (Lassalle et al., 2018) and processed their data set with our workflow. We generated an unrooted neighbour-joining for the eight populations (Figure 5a) which showed that the PNG populations are closer to one another and separated from the Philippines and the Western controls.

We also compared the abundance of the PNG core microbiome (Figure 5b; Table S5) obtained in this WGS-based data set (Human Microbiome Project, 2012; Lassalle et al., 2018) with other 16S rRNA-based published studies for Australian Aboriginals and Torres Strait Islanders (Handsley-Davis et al., 2021) and Batwa Pygmies (Nasidze et al., 2011). Interestingly, the *Rothia* genera abundances in the five PNG groups (average 8.24%) are significantly higher (p -value = .0013) than in all the other compared groups (Philippines hunter-gatherers, Philippines traditional farmers, Batwa Pygmies and Western Controls; average 3.87%), except for the Australian Aboriginals and Torres Strait Islanders (8%). The above pattern is also significantly different for the *Neisseria* genera (p -value = .0405), with high abundances in the PNG groups (average 15.05%) and Australian Aboriginals and Torres Strait Islanders (14%) than in the remaining groups (9.58%). This enrichment of *Rothia* and *Neisseria* appear to be limited to the New Guinea-Australia region (Figure 5b), which might indicate a specific trait of these genera on the oral microbiome, potentially influenced by environmental conditions found in this geographic region.

Veilonella appears in higher abundances in the Western controls (11.86%), the Philippines traditional farmers (11.75%), in the Port Moresby group (9.05%) and the East Highlanders living in Port Moresby group (9.07%). This genus has been associated with groups having a Westernized diet (Clemente et al., 2015; Lassalle et al., 2018). These studies also reported an enrichment of *Haemophilus* in hunter-gatherers, which is not apparent in our results (Table S5), as the Western controls have a higher abundance of *Haemophilus* than the Hunter-gatherer groups. (Lassalle et al., 2018).

4 | DISCUSSION

The core oral microbiome of PNG agrees with the normal microbiota in the oral cavity obtained in other studies (Aas et al., 2005; Deo & Deshmukh, 2019; Dewhirst et al., 2010). Bacterial genera such as *Prevotella*, *Neisseria* and *Streptococcus* represent most of the human oral microbiome diversity. However, as seen within our PNG

populations, the abundances of each taxon might vary, potentially due to differential lifestyles, diets or host genetics (Li et al., 2014; Mousa et al., 2022). Both quantitative (WGS) and qualitative (Microbiome array) results show significant differences between the East Sepik group and the other PNG groups. This might be explained by the relative geographic isolation and a more traditional form of subsistence, with limited access to Westernized food, contrary to the other four groups (Pawley et al., 2005; Roscoe, 2005). Similar results were described in the oral microbiomes of isolated groups across the globe, such as the Native American Yanomami population (Clemente et al., 2015), the African Batwa (Li et al., 2014), and in the Philippines hunter-gatherer group (Lassalle et al., 2018), when compared to groups with a Westernized diet. Apart from the East Sepik group, all the other PNG groups included in this study have a mixed traditional and Westernized lifestyle (Simbu and Western provinces and Port Moresby), which might explain the similarity of their respective oral microbiome compositions. Our results for the group of East Highlanders living in Port Moresby suggests an incorporation of both microbiotas (initial oral microbiota from the highlands and the acquired microbiota from the capital city), with higher similarity with the initial microbiota. There are no studies of the effect of human migration on oral microbiome, but for gut microbiome (Vangay et al., 2018), a complete change of gut microbiome diversity and function from the native to a new environment was observed. Our results indicate a more complex interaction between the original and the recently acquired oral microbiome, emphasizing the need for further study on the plasticity of the human oral microbiome.

An interesting pattern of genera enrichment was found in the populations from New Guinea and Australia (PNG groups and Australian Aboriginals and Torres Strait Islanders). We detected higher abundances for the *Neisseria* and *Rothia* genera than other populations of the world. New Guinean and Aboriginal Australian populations share a common history, since the early settlement of these territories 50,000 years ago (Allen & O'Connell, 2020; Brucato et al., 2021; Pedro et al., 2020), then connected into the ancient continent Sahul. A secondary gene flow linked to the Austronesian dispersal 3 kya led to a gradient of Asian genetic ancestry in coastal groups of the region (Brucato et al., 2021; Mallick et al., 2016). However, our limited data set (no group from New Guinea with a predominant Austronesian genetic ancestry) and current knowledge on microbiome research, do not support a strong link between a population genetic ancestry (e.g., Austronesian) and a specific oral microbiome signature, despite some encouraging new studies (Weyrich, 2021). It is, nevertheless, remarkable that our study revealed a specific pattern in the oral microbiome shared by New Guineans and Aboriginal Australian groups, that are genetically separated for at least 30,000 years (Brucato et al., 2021; Malaspinas et al., 2016). This specific pattern may have been accentuated by limited gene flow between southeast coastal New Guinean lowlanders and northeast Indigenous Australians until at least 10 kya, and during the Austronesian period of the last 3 kya (Bellwood, 2007; Brucato et al., 2021; Karmin et al., 2022; Lipson et al., 2014; Pedro et al., 2020). This may suggest a Sahul-specific oral microbiome diversity, caused by a shared environment and/or a shared ancient

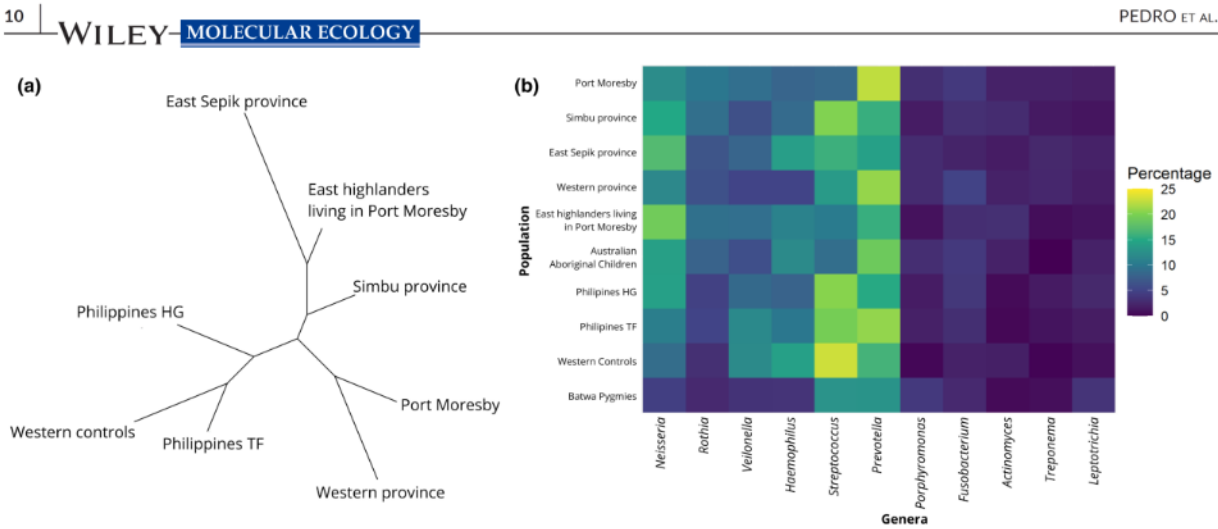


FIGURE 5 Comparison of oral microbiome profiles between PNG and other populations based on quantitative for review only data (WGS results). (a) Unrooted neighbour-joining tree of the mean weighted UniFrac distances between the five PNG groups (Port Moresby, Simbu province, East Sepik province, Western province and east highlanders living in Port Moresby), the Philippines hunter-gatherers (HG) and traditional farmers (TF) and the Western controls (HMP, 2012; Lassalle et al., 2018). (b) Heatmap for abundances of the top 11 genera detected with quantitative results in the eight groups also present in the unrooted neighbour-joining tree (>1% of abundance in all samples) and other published populations of aboriginal Australian children and Batwa pygmies (Handsley-Davis et al., 2021; Nasidze et al., 2011).

ancestry. Among the Sahul-specific oral microbiome, the *Rothia* genus has been shown to have positive effects on the host's iron pool, by the production of the iron-scavenging enterobactin (Uranga et al., 2020). Enterobactin is a siderophore protein released by bacteria into the surrounding environment to bind to available iron, which is then used in several biochemical pathways of the bacteria and the host (Raymond et al., 2003). Both *Rothia* and *Neisseria* genera appear to be depleted in the saliva of iron-deficient individuals, compared to healthy controls, which establishes their importance in the iron homeostasis (Xi et al., 2019). Iron deficiency is an important health matter in PNG and Indigenous Australian communities, associated with increased morbidity and mortality (Leonard et al., 2019; Manning et al., 2012). These genera might be more abundant in the human oral microbiome of populations from the ancient Sahul region to, in part, compensate a potential iron deficiency. Other worldwide studies of oral microbiome in healthy individuals did not report this *Rothia* and *Neisseria* pattern, except for Nasidze et al. (2009) that show a higher abundance of the *Rothia* genus in the German population, but no interpretation was provided. Additionally, in the gut microbiome of Papua New Guineans individuals, the *Lactobacillus* genus, which is able to sense iron deprivation and communicate this iron need to the host (Das et al., 2020), was found highly enriched when compared to US individuals (Martinez et al., 2015). Iron plays an important role in infection and immunity, for both pathogens and host cells (Nairz & Weiss, 2020). While it is well-known that pathogens and diet are motors for human adaptation, understudied populations still have the potential to shed light on new adaptations (Pedro et al., 2021). In this case, the environmental pressure (e.g., high pathogenic pressure, specific diet and plant use) observed in New Guinea and north Australia (Horwood et al., 2019) might have led to the development of a Sahul-specific oral microbiome.

This unexpected pattern of the PNG oral microbiome clearly needs further exploration. As this data was collected opportunistically, potential environmental and cultural cofactors were not controlled for. In particular, betel nut chewing (Hernandez et al., 2017; Zhong et al., 2021) has shown to be highly associated with variation in the oral microbiome in other populations and is a common habit to all the PNG locations. The impact of this specific cultural behaviour would presumably result in an alteration of the global oral microbiome in all PNG populations, which would not greatly affect our results. Nevertheless, the effect of betel nut chewing in the oral microbiome is an important question and should be further explored in the PNG populations. The time of sample collection might also influence the results obtained for the oral microbiome, as seasonality in sample collection should be considered in further studies, and might increase the power in discriminating between the oral microbiomes of the PNG populations (Clemente et al., 2015). Oral health is also one of the factors that influence oral microbiota (Chukkappalli et al., 2015; Costalonga & Herzberg, 2014) and is common to all PNG populations. We have no information available concerning oral hygiene habits of the participants, but the influence of dental hygiene should not be ignored. All these external factors should be integrated in future analyses to evaluate their impact on our results. However, the convergence of our results in PNG with those from another study on Australian Aboriginals and Torres Strait Islanders (Handsley-Davis et al., 2021) suggests that the influence of these potential cofactors is unlikely to be major.

The detection of a possible Sahul-specific oral microbiome diversity may be potentially related to adaptation to their specific biodiversity, cultural lifestyle or genetic background. The results obtained in this study highlight the value of studying the microbiomes of more isolated groups to fully encompass the ecology and biogeography of

the host-microbiota interactions. As the microbiota can affect the host's phenotypic traits, host-microbe relationships can evolve in selective environments, conferring fitness advantages to the host. The influence of the microbiome in host ecology and evolution is more easily studied in nonhuman mammals, which are subjected to variation in nutritional intake, due to the seasonal availability of foods, changes in their natural habitat and higher disease burden (Amato, 2016; Clayton et al., 2018). In fact, Amato et al. (2015) reported this relationship in black howler monkeys (*Alouatta pigra*). The researchers showed that the gut microbiota adapts its composition based on the higher consumption of available fruit versus a higher consumption of leaves. In most modern human populations, which have access to a broad availability of food and medicine year-round, there is limited selective pressure that could affect the microbiome of the host (Amato, 2016; Colquhoun & Lyon, 2001). However, human populations residing in rural or remote areas (as for PNG populations, especially the East Sepik province) are still subjected to the pressure of seasonal food availability. A potential study including more samples and more relevant data from these populations will be developed to further study, validate and understand how ecology, ancestry and human lifestyle impacts microbiome structure. Only by studying the different microbiotas in the ecosystem of diverse hosts, can we truly highlight the unique characteristics that defines the symbiotic relationship between microbe and host (Parfrey et al., 2018).

AUTHOR CONTRIBUTIONS

Nicole Pedro, Luisa Pereira and François-Xavier Ricaut designed the study; Christopher Kinipi and Matthew Leavesley performed sample collection; Nicole Pedro, Nicolas Brucato and Valentine Lisant performed laboratory procedures; Nicole Pedro, Bruno Cavadas and Rui Camacho performed the bioinformatics and computational analyses; Nicole Pedro, Nicolas Brucato, Luisa Pereira and François-Xavier Ricaut wrote the manuscript; all authors read and reviewed the manuscript.

ACKNOWLEDGEMENTS

We especially thank all of our study participants. We also thank Mathilde André and Georgi Hudjashov (University of Tartu) for their help in data analysis. We thank Roxanne Tsang, Jason Kariwiga, Teppsy Beni, Kylie Sesuki (University of Papua New Guinea), Alois Kuaso, and Kenneth Miamba (National Museum and Art Gallery, Papua New Guinea), and John Muke (Social Research Institute, Papua New Guinea) for their help during the sampling campaigns. We acknowledge support from the GenoToul bioinformatics facility of Genopole Toulouse Midi-Pyrénées, France; the LabEx TULIP, France; and the Centre National de Recherche en Génomique Humaine (CNRGH), Evry, France. This work was supported by the French National Research Agency (ANR; grant PAPUAEVOL n° ANR-20-CE12-0003-01), the French Ministry of Foreign and European Affairs (French Prehistoric Mission in Papua New Guinea), and the French Embassy in Papua New Guinea. The Portuguese

Foundation for Science and Technology (FCT) financed the PhD grant to NP (SFRH/BD/136299/2018). i3S is supported by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020 – Operational Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT/Ministério da Ciência, Tecnologia e Inovação in the framework of the project 'Institute for Research and Innovation in Health Sciences' (POCI-01-0145-FEDER-007274). The authors also acknowledge the support of the i3S Genomics Scientific Platform, member of the GenomePT project (POCI-01-0145-FEDER-022184), supported by FEDER funds through COMPETE 2020, POCI, Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Program (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement and FCT. Lastly, we would like to thank the reviewers for the time and effort spent in reviewing the manuscript. All their valuable comments and suggestions helped us improve the quality of the manuscript.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The MiDAS output files obtained from the Array methodology and the KRAKEN output files obtained from the WGS unmapped reads have been made available as Supporting Information files. The reads mapping to the microbial community (after removal of human reads) were submitted to ENA (www.ebi.ac.uk/ena), for the study number ERP139829 under the BioSample accession numbers ERS12516930 – ERS12516994.

OPEN RESEARCH BADGES



This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The MiDAS output files obtained from the Array methodology and the KRAKEN output files obtained from the WGS unmapped reads are available as supplementary files. The reads mapping to the microbial community (after removal of human reads) were submitted to ENA (www.ebi.ac.uk/ena), for the study number ERP139829 under the BioSample Accession numbers ERS12516930 – ERS12516994.

BENEFIT-SHARING

Benefits generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above.

ORCID

Nicole Pedro <https://orcid.org/0000-0001-5667-1549>

François-Xavier Ricaut <https://orcid.org/0000-0001-7609-7898>

REFERENCES

- Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., & Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *Journal of Clinical Microbiology*, 43(11), 5721–5732. <https://doi.org/10.1128/JCM.43.11.5721-5732.2005>
- Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltmann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1), W537–W544. <https://doi.org/10.1093/nar/ky379>
- Allen, J., & O'Connell, J. F. (2020). A different paradigm for the initial colonisation of Sahul. *Archaeology in Oceania*, 55(1), 1–14. <https://doi.org/10.1002/arco.5207>
- Amato, K. R. (2016). Incorporating the gut microbiota into models of human and non-human primate ecology and evolution. *American Journal of Physical Anthropology*, 159(Suppl 61), S196–S215. <https://doi.org/10.1002/ajpa.22908>
- Amato, K. R., Leigh, S. R., Kent, A., Mackie, R. I., Yeoman, C. J., Stumpf, R. M., Wilson, B. A., Nelson, K. E., White, B. A., & Garber, P. A. (2015). The gut microbiota appears to compensate for seasonal diet variation in the wild black howler monkey (*Alouatta pigra*). *Microbial Ecology*, 69(2), 434–443. <https://doi.org/10.1007/s00248-014-0554-7>
- Anderson, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels (Eds.), *Wiley StatsRef: Statistics reference online* (pp. 1–15). Wiley.
- Bellwood, P. (2007). *Prehistory of the Indo-Malaysian archipelago* (Revised ed.). ANU Press.
- Brucato, N., André, M., Tsang, R., Saag, L., Kariwiga, J., Sesuki, K., Beni, T., Pomat, W., Muke, J., Meyer, V., Boland, A., Deleuze, J. F., Sudoyo, H., Mondal, M., Pagani, L., Gallego Romero, I., Metspalu, M., Cox, M. P., Leavesley, M., & Ricaut, F. X. (2021). Papua new Guinean genomes reveal the complex settlement of north Sahul. *Molecular Biology and Evolution*, 38(11), 5107–5121. <https://doi.org/10.1093/molbev/msab238>
- Bryan, N. S., Tribble, G., & Angelov, N. (2017). Oral microbiome and nitric oxide: The missing link in the Management of Blood Pressure. *Current Hypertension Reports*, 19(4), 33. <https://doi.org/10.1007/s11906-017-0725-2>
- Burcham, Z. M., Garneau, N. L., Comstock, S. S., Tucker, R. M., Knight, R., Metcalf, J. L., & Genetics of Taste Lab Citizen, S. (2020). Patterns of Oral microbiota diversity in adults and children: A crowdsourced population study. *Scientific Reports*, 10(1), 2133. <https://doi.org/10.1038/s41598-020-59016-0>
- Cavadas, B., Camacho, R., Ferreira, J. C., Ferreira, R. M., Figueiredo, C., Brazma, A., Fonseca, N. A., & Pereira, L. (2020). Gastric microbiome diversities in gastric cancer patients from Europe and Asia mimic the human population structure and are partly driven by microbiome quantitative trait loci. *Microorganisms*, 8(8), E1196. <https://doi.org/10.3390/microorganisms8081196>
- Chukkapalli, S. S., Rivera-Kweh, M. F., Velsko, I. M., Chen, H., Zheng, D., Bhattacharyya, I., Gangula, P. R., Lucas, A. R., & Kesavalu, L. (2015). Chronic oral infection with major periodontal bacteria *Tannerella forsythia* modulates systemic atherosclerosis risk factors and inflammatory markers. *Pathogens and Disease*, 73(3), ftv009. <https://doi.org/10.1093/femspd/ftv009>
- Clayton, J. B., Gomez, A., Amato, K., Knights, D., Travis, D. A., Blekhman, R., Knight, R., Leigh, S., Stumpf, R., Wolf, T., Glander, K. E., Cabana, F., & Johnson, T. J. (2018). The gut microbiome of nonhuman primates: Lessons in ecology and evolution. *American Journal of Primatology*, 80(6), e22867. <https://doi.org/10.1002/ajp.22867>
- Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcón, Ó., Lander, O., McDonald, J., Cox, M., Walter, J., Oh, P. L., Ruiz, J. F., Rodriguez, S., Shen, N., Song, S. J., ... Dominguez-Bello, M. G. (2015). The microbiome of uncontacted Amerindians. *Science Advances*, 1(3), e1500183. <https://doi.org/10.1126/sciadv.1500183>
- Colquhoun, A., & Lyon, P. (2001). To everything there was a season: Deconstructing UK food availability. *Food Service Technology*, 1(2), 93–102. <https://doi.org/10.1046/j.1471-5740.2001.00010.x>
- Contreras, M., Costello, E. K., Hidalgo, G., Magris, M., Knight, R., & Dominguez-Bello, M. G. (2010). The bacterial microbiota in the oral mucosa of rural Amerindians. *Microbiology (Reading)*, 156(Pt 11), 3282–3287. <https://doi.org/10.1099/mic.0.043174-0>
- Cordain, L., Eaton, S. B., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B. A., O'Keefe, J. H., & Brand-Miller, J. (2005). Origins and evolution of the Western diet: Health implications for the 21st century. *The American Journal of Clinical Nutrition*, 81(2), 341–354. <https://doi.org/10.1093/ajcn.81.2.341>
- Costalonga, M., & Herzberg, M. C. (2014). The oral microbiome and the immunobiology of periodontal disease and caries. *Immunology Letters*, 162(2 Pt A), 22–38. <https://doi.org/10.1016/j.imlet.2014.08.017>
- Das, N. K., Schwartz, A. J., Barthel, G., Inohara, N., Liu, Q., Sankar, A., Hill, D. R., Ma, X., Lamberg, O., Schnizlein, M. K., Arqués, J. L., Spence, J. R., Nunez, G., Patterson, A. D., Sun, D., Young, V. B., & Shah, Y. M. (2020). Microbial metabolite signaling is required for systemic iron homeostasis. *Cell Metabolism*, 31(1), 115–130 e116. <https://doi.org/10.1016/j.cmet.2019.10.005>
- Deo, P. N., & Deshmukh, R. (2019). Oral microbiome: Unveiling the fundamentals. *Journal of Oral and Maxillofacial Pathology: JOMFP*, 23(1), 122–128. https://doi.org/10.4103/jomfp.JOMFP_304_18
- Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C., Yu, W. H., Lakshmanan, A., & Wade, W. G. (2010). The human oral microbiome. *Journal of Bacteriology*, 192(19), 5002–5017. <https://doi.org/10.1128/JB.00542-10>
- Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., & De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1), 3030. <https://doi.org/10.1038/s41598-021-82726-y>
- Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., Hurley, E., O'Riordain, M., Shanahan, F., & O'Toole, P. W. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, 67(8), 1454–1463. <https://doi.org/10.1136/gutjnl-2017-314814>
- Gardner, S. N., Jaing, C. J., McLoughlin, K. S., & Slezak, T. R. (2010). A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics*, 11(1), 668. <https://doi.org/10.1186/1471-2164-11-668>
- Gomes, S., Cavadas, B., Ferreira, J. C., Marques, P. I., Monteiro, C., Sucena, M., Sousa, C., Vaz Rodrigues, L., Teixeira, G., Pinto, P., Tavares de Abreu, T., Bárbara, C., Semedo, J., Mota, L., Carvalho, A. S., Matthiesen, R., Pereira, L., & Seixas, S. (2019). Profiling of lung microbiota discloses differences in adenocarcinoma and squamous cell carcinoma. *Scientific Reports*, 9(1), 12838. <https://doi.org/10.1038/s41598-019-49195-w>
- Graves, D. T., Correa, J. D., & Silva, T. A. (2019). The Oral microbiota is modified by systemic diseases. *Journal of Dental Research*, 98(2), 148–156. <https://doi.org/10.1177/0022034518805739>
- Handsley-Davis, M., Skelly, E., Johnson, N. W., Kapellas, K., Lalloo, R., Kroon, J., & Weyrich, L. S. (2021). Biocultural drivers of salivary microbiota in Australian aboriginal and Torres Strait islander children. *Frontiers in Oral Health*, 2, 641328. <https://doi.org/10.3389/froh.2021.641328>

Hernandez, B. Y., Zhu, X., Goodman, M. T., Gatewood, R., Mendiola, P., Quinata, K., & Paulino, Y. C. (2017). Betel nut chewing, oral pre-malignant lesions, and the oral microbiome. *PLoS One*, 12(2), e0172196. <https://doi.org/10.1371/journal.pone.0172196>

Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., Knight, R., & Knights, D. (2018). Evaluating the information content of shallow shotgun metagenomics. *mSystems*, 3(6), e00069-18. <https://doi.org/10.1128/mSystems.00069-18>

Horwood, P. F., Tarantola, A., Goarant, C., Matsui, M., Klement, E., Umezaki, M., Navarro, S., & Greenhill, A. R. (2019). Health challenges of the Pacific region: Insights from history, geography, social determinants, genetics, and the microbiome. *Frontiers in Immunology*, 10, 2184. <https://doi.org/10.3389/fimmu.2019.02184>

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10), 4399–4406. <https://doi.org/10.1128/AEM.67.10.4399-4406.2001>

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207–214. <https://doi.org/10.1038/nature11234>

Jacobs, G. S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C. C., Lawson, D. J., Mondal, M., Pagani, L., Ricaut, F. X., Stoneking, M., Metspalu, M., Sudoyo, H., Lansing, J. S., & Cox, M. P. (2019). Multiple deeply divergent Denisovan ancestries in Papuans. *Cell*, 177(4), 1010–1021 e1032. <https://doi.org/10.1016/j.cell.2019.02.035>

Karmin, M., Flores, R., Saag, L., Hudjashov, G., Brucato, N., Crenna-Darusallam, C., Larena, M., Endicott, P. L., Jakobsson, M., Lansing, J. S., Sudoyo, H., Leavesley, M., Metspalu, M., Ricaut, F. X., & Cox, M. P. (2022). Episodes of diversification and isolation in Island southeast Asian and near Oceanian male lineages. *Molecular Biology and Evolution*, 39(3), msac045. <https://doi.org/10.1093/molbev/msac045>

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence-absence data. *Journal of Animal Ecology*, 72(3), 367–382. <https://doi.org/10.1046/j.1365-2656.2003.00710.x>

Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: Dynamic communities and host interactions. *Nature Reviews. Microbiology*, 16(12), 745–759. <https://doi.org/10.1038/s41579-018-0089-x>

Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., Thomas, M. G., Bamberg Migliano, A., & Balloux, F. (2018). Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*, 27(1), 182–195. <https://doi.org/10.1111/mec.14435>

Leonard, D., Buttner, P., Thompson, F., Makrides, M., & McDermott, R. (2019). Anaemia in early childhood among aboriginal and Torres Strait islander children of far North Queensland: A retrospective cohort study. *Australian and New Zealand Journal of Public Health*, 43(4), 319–327. <https://doi.org/10.1111/1753-6405.12911>

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>

Li, J., Quinque, D., Horz, H. P., Li, M., Rzhetskaya, M., Raff, J. A., Hayes, M. G., & Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiology*, 14(1), 316. <https://doi.org/10.1186/s12866-014-0316-1>

Lipson, M., Loh, P. R., Patterson, N., Moorjani, P., Ko, Y. C., Stoneking, M., Berger, B., & Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. *Nature Communications*, 5, 4689. <https://doi.org/10.1038/ncomms5689>

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *The ISME Journal*, 5(2), 169–172. <https://doi.org/10.1038/ismej.2010.133>

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3, e104. [ARTNe10410.7717/peerj-cs.104](https://doi.org/10.1016/j.pccs.2017.03.004)

Malaspina, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., ... Willerslev, E. (2016). A genomic history of aboriginal Australia. *Nature*, 538(7624), 207–214. <https://doi.org/10.1038/nature18299>

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., ... Reich, D. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624), 201–206. <https://doi.org/10.1038/nature18964>

Manning, L., Laman, M., Rosanas-Urgell, A., Michon, P., Aipit, S., Bona, C., Siba, P., Mueller, I., & Davis, T. M. (2012). Severe anemia in Papua new Guinean children from a malaria-endemic area: A case-control etiologic study. *PLoS Neglected Tropical Diseases*, 6(12), e1972. <https://doi.org/10.1371/journal.pntd.0001972>

Martinez, I., Stegen, J. C., Maldonado-Gomez, M. X., Eren, A. M., Siba, P. M., Greenhill, A. R., & Walter, J. (2015). The gut microbiota of rural Papua new guineans: Composition, diversity patterns, and ecological processes. *Cell Reports*, 11(4), 527–538. <https://doi.org/10.1016/j.celrep.2015.03.049>

McLoughlin, K. S. (2011). Microarrays for pathogen detection and analysis. *Briefings in Functional Genomics*, 10(6), 342–353. <https://doi.org/10.1093/bfpg/blr027>

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>

Miller, A. K., Westlake, C. S., Cross, K. L., Leigh, B. A., & Bordenstein, S. R. (2021). The microbiome impacts host hybridization and speciation. *PLoS Biology*, 19(10), e3001417. <https://doi.org/10.1371/journal.pbio.3001417>

Mousa, W. K., Chehadeh, F., & Husband, S. (2022). Recent advances in understanding the structure and function of the human microbiome. *Frontiers in Microbiology*, 13, 825338. <https://doi.org/10.3389/fmicb.2022.825338>

Nairz, M., & Weiss, G. (2020). Iron in infection and immunity. *Molecular Aspects of Medicine*, 75, 100864. <https://doi.org/10.1016/j.mam.2020.100864>

Nasidze, I., Li, J., Quinque, D., Tang, K., & Stoneking, M. (2009). Global diversity in the human salivary microbiome. *Genome Research*, 19(4), 636–643. <https://doi.org/10.1101/gr.084616.108>

Nasidze, I., Li, J., Schroeder, R., Creasey, J. L., Li, M., & Stoneking, M. (2011). High diversity of the saliva microbiome in Batwa pygmies. *PLoS One*, 6(8), e23352. <https://doi.org/10.1371/journal.pone.0023352>

Parfrey, L. W., Moreau, C. S., & Russell, J. A. (2018). Introduction: The host-associated microbiome: Pattern, process and function. *Molecular Ecology*, 27(8), 1749–1765. <https://doi.org/10.1111/mec.14706>

Pawley, A., Attenborough, R., Golson, J., & Hide, R. (2005). *Papuan pasts: Cultural, linguistic and biological histories of the Papuan-speaking peoples* (pp. 555–584). Pacific Linguistics.

Pedro, N., Brucato, N., Fernandes, V., André, M., Saag, L., Pomat, W., Besse, C., Boland, A., Deleuze, J. F., Clarkson, C., Sudoyo, H., Metspalu, M., Stoneking, M., Cox, M. P., Leavesley, M., Pereira, L., & Ricaut, F. X. (2020). Papuan mitochondrial genomes and the

settlement of Sahul. *Journal of Human Genetics*, 65(10), 875–887. <https://doi.org/10.1038/s10038-020-0781-3>

Pedro, N., Pinto, R. J., Cavadas, B., & Pereira, L. (2021). Sub-Saharan African information potential to unveil adaptations to infectious disease. *Human Molecular Genetics*, 30(R1), R138–R145. <https://doi.org/10.1093/hmg/ddab001>

R Studio Team. (2020). *RStudio: Integrated development for R*. Boston, MA. <http://www.rstudio.com/>

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4), 967–977. <https://doi.org/10.1016/j.bbrc.2015.12.083>

Raymond, K. N., Dertz, E. A., & Kim, S. S. (2003). Enterobactin: An archetype for microbial iron transport. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7), 3584–3588. <https://doi.org/10.1073/pnas.0630018100>

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., Ko, A. M., Ko, Y. C., Jinam, T. A., Phipps, M. E., Saitou, N., Wollstein, A., Kayser, M., Pääbo, S., & Stoneking, M. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, 89(4), 516–528. <https://doi.org/10.1016/j.ajhg.2011.09.005>

Riley, I. (2009). *Demography and the epidemiology of disease in Papua New Guinea* (Vol. 52). Papua New Guinea Institute of Medical Research.

Roscoe, P. (2005). Foraging, ethnographic analogy and Papuan pasts: Contemporary models for the Sepik-Ramu past. In A. Pawley, R. Attenborough, J. Golson, R. Hide (Eds.), *Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples* (pp. 555–584). Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, 29(10), 593–599. <https://doi.org/10.1016/j.tig.2013.07.006>

Schapper, A. (2017). Farming and the trans-new Guinea family: A consideration chapter. In M. R. A. Savelyev (Ed.), *Language dispersal beyond farming* (pp. 155–181). John Benjamins Publishing Company.

Schmidt, E., & Fang, P. (2021). Papua New Guinea Agri-food trade and household consumption trends point towards dietary change and increased overweight and obesity prevalence. *Globalization and Health*, 17(1), 135. <https://doi.org/10.1186/s12992-021-00787-0>

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6), R60. <https://doi.org/10.1186/gb-2011-12-6-r60>

Thissen, J. B., McLoughlin, K., Gardner, S., Gu, P., Mabery, S., Slezak, T., & Jaing, C. (2014). Analysis of sensitivity and rapid hybridization of a multiplexed microbial detection microarray. *Journal of Virological Methods*, 201, 73–78. <https://doi.org/10.1016/j.jviro.2014.01.024>

Uranga, C. C., Arroyo, P., Jr., Duggan, B. M., Gerwick, W. H., & Edlund, A. (2020). Commensal Oral *Rothia mucilaginosa* produces Enterobactin, a metal-chelating siderophore. *mSystems*, 5(2), e00161-00120. doi:<https://doi.org/10.1128/mSystems.00161-20>, 5

Vangay, P., Johnson, A. J., Ward, T. L., Al-Ghalith, G. A., Shields-Cutler, R. R., Hillmann, B. M., Lucas, S. K., Beura, L. K., Thompson, E. A., Till, L. M., Batres, R., Paw, B., Pergament, S. L., Saenyakul, P., Xiong, M., Kim, A. D., Kim, G., Masopust, D., Martens, E. C., ... Knights, D. (2018). US immigration westernizes the human gut microbiome. *Cell*, 175(4), 962–972 e910. <https://doi.org/10.1016/j.cell.2018.10.029>

Wang, L., & Ganly, I. (2014). The oral microbiome and oral cancer. *Clinics in Laboratory Medicine*, 34(4), 711–719. <https://doi.org/10.1016/j.cll.2014.08.004>

Weyrich, L. S. (2021). The evolutionary history of the human oral microbiota and its implications for modern health. *Periodontology* 2000, 85(1), 90–100. <https://doi.org/10.1111/prd.12353>

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 257. <https://doi.org/10.1186/s13059-019-1891-0>

Xi, R., Wang, R., Wang, Y., Xiang, Z., Su, Z., Cao, Z., Xu, X., Zheng, X., & Li, J. (2019). Comparative analysis of the oral microbiota between iron-deficiency anaemia (IDA) patients and healthy individuals by high-throughput sequencing. *BMC Oral Health*, 19(1), 255. <https://doi.org/10.1186/s12903-019-0947-6>

Zhong, X., Lu, Q., Zhang, Q., He, Y., Wei, W., & Wang, Y. (2021). Oral microbiota alteration associated with oral cancer and areca chewing. *Oral Diseases*, 27(2), 226–239. <https://doi.org/10.1111/odi.13545>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pedro, N., Brucato, N., Cavadas, B., Lisant, V., Camacho, R., Kinipi, C., Leavesley, M., Pereira, L., & Ricaut, F.-X. (2022). First insight into oral microbiome diversity in Papua New Guineans reveals a specific regional signature. *Molecular Ecology*, 00, 1–14. <https://doi.org/10.1111/mec.16702>

PAPER III. HOST DIVERSITY VERSUS ENVIRONMENTAL INFLUENCE ON
THE PAPUA NEW GUINEAN ORAL MICROBIOME

Host diversity versus environmental influence on the Papua New Guinean oral microbiome

Nicole Pedro^{1,2,3} et al.

1 – i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

2 – IPATIMUP – Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal

3 – ICBAS – Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal

Keywords: Papua New Guinea, oral microbiome, microbiome quantitative trait loci, microbial metabolism, iron pathways.

Abstract

Papua New Guineans (PNG) are one of the most genetically diverse populations worldwide and are continuously subjected to unique cultural and environmental influences. Their oral microbiome is composed mainly by *Prevotella*, *Neisseria*, *Streptococcus*, *Haemophilus*, *Rothia* and *Veillonella*, but the abundance of *Rothia* and *Neisseria* in PNG is much higher than what is observed in other populations. To clarify the source of this Sahul-specific microbiome profile, we performed a microbiome quantitative trait analysis (mbQTL) on 252 individuals from several locations in PNG to uncover any ancestry-related influence on the oral microbiome. We also analyzed the abundance of bacterial pathways in the oral microbiome between PNG, Philippines and Western population groups, for a functional analysis of the underlying microbial community. Although the mbQTL analysis did not reveal any associations with *Neisseria* and *Rothia*, we were able to identify six interactions between host genomic variants and four bacterial genera, the more interesting being the associations of *Gemella* with the *BACE2* gene, and *Schaalia* with *ISG20* gene. The analysis of metabolic pathways allowed to confirm the statistically significant higher contribution of both *Neisseria* and *Rothia* in iron-related pathways in PNG versus *Veillonella* in Philippines and Western population groups. These results corroborate the hypothesis that iron deficiency might have been a motor for the development of this specific oral microbiome profile in PNG, as at least *Rothia* has been demonstrated as a very efficient scavenger of iron.

Introduction

The human oral microbiota is colonized by around 700 different types of microorganisms including bacteria, fungi, viruses, archaea, and protozoa (Lamont et al. 2018). In healthy conditions, the oral microbiota is composed mainly by Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria and Proteobacteria, but it can present different abundances between populations (Caselli et al. 2020). In Papua New Guinea (PNG), a country with high cultural diversity, a semi-traditional form of subsistence (Pawley et al. 2005), and a very high human genetic diversity (Brucato et al. 2021), the core oral microbiome was composed by *Prevotella* (~17%), *Neisseria* (~13%), *Streptococcus* (~13%), *Haemophilus* (~9%), *Rothia* (~8%), *Veillonella* (~7%), *Fusobacterium* (~4%), *Porphyromonas* (~2%), *Actinomyces* (~2%) and *Leptotrichia* (~2%) (Pedro et al. 2022). Interestingly, the abundance of *Rothia* and *Neisseria* in PNG was much higher than what is observed in other populations, which was suggested to be associated with a potential iron-related adaptation (Pedro et al. 2022). Iron is required for the growth of all microbes, however they possess distinct iron requirements and have unique abilities to persist in iron deficient conditions, making iron one of the modulating factors in shaping the host associated microbial community (Uranga et al. 2020). This specific oral microbiome profile can be an adaptation to different environments, lifestyles, eating habits, or genetic background (Li et al. 2014; Mousa et al. 2022).

The aim of quantitative trait loci (QTL) analysis is to identify variants which are significantly associated with certain phenotypes. In particular, the microbiome QTLs (mbQTLs) are human loci significantly associated with specific microbes in a certain microbiome of the human body. Recent works are beginning to show interesting mbQTLs, revealing essential biochemical processes underlying the interaction between both living systems. In healthy individuals from diverse population origins (but largely European descendants), Kurilshikov et al. (2021) identified 31 loci affecting the gut microbiome, with the lactase (*LCT*) gene locus showing an age-dependent association with *Bifidobacterium* abundance. The other associations were enriched for genes related to metabolism, as the *FUT2-FUT1* associated with the abundance of the *Ruminococcus*, or genes potentially involved in host-microbiome

cross-talk as the 9q21 locus (locus includes the genes *PCSK5*, *RFK* and *GCNT1*) associated with abundance of *Allisonella*. For oral microbiome in particular, few studies have been performed, and are more restricted to European populations. *IMMPL2* and *INHBA-AS1* genes were found associated with variation in the abundance of *Granulicatella* (Demmitt et al. 2017) and *AMY1* locus correlated with an increase in salivary *Porphyromonas* (Poole et al. 2019). A study by Liu et al. (2021) in a Chinese cohort found numerous associations of relative abundance with several taxa, but also reported genetic associations with α -diversity (Chao1 index) and with β -diversity (Bray-Curtis Similarity) in tongue dorsum and salivary samples.

Recently, the metagenomics data is also being used to infer the functional impact of microbe abundance in microbiota metabolism (Sharon et al. 2011). When applied to the Human Microbiome Project samples (Abubucker et al. 2012), more specifically in three oral sites (buccal mucosa, supragingival plaque, and tongue dorsum), the anterior nares, the retroauricular crease, and the stool of 102 individuals, they were able to identify the core metabolic pathways of each site, and the corresponding higher contributing bacteria. For instance, glutamate metabolism pathways were present in the multiple niches studied, but enriched in the stool, while arginine transport and methionine biosynthesis were found enriched in all three oral sites. By integrating functional and abundance information, it is possible to offer a different perspective on underlying microbial community function (Beghini et al. 2021).

In this work, we reanalyzed 252 human whole genome sequences obtained from PNG saliva samples, to catalog the human variants and infer the oral microbiome from the non-aligned human reads. We aim to identify promising host variants that may play a role modulating the oral microbiota in the PNG population and characterize the major contributing bacteria of the metabolic pathways enriched in PNG when compared to other worldwide populations. In doing so, we will be able to evaluate if the specific oral microbiome signature found in Pedro et al. (2022) for the Sahul populations (the higher abundance of *Rothia* and *Neisseria*) is associated with human diversity, possibly positively selected, or instead be better explained by some environmental factor related with microbial metabolism. By performing this evaluation at a global

microbiome level, we will also be able to uncover interesting new host-microbiome relations.

Materials and methods

Samples

Two hundred and fifty-two saliva samples were collected from individuals from three different regions of PNG during field seasons between 2016 and 2019. Sampling sites were selected in order to access different environments, different lifestyles and being geographically distinct from each other: Port Moresby, includes individuals sampled in the east coastal lowlands (n=115; a genetically heterogeneous groups as it includes individuals from across the country); Chimbu province, Highlanders sampled in the Highlands (n=58; Chimbu province); and Western province, individuals sampled in the west coastal lowlands (n=79; Western province). The saliva samples were collected with the Oragene DNA OG-500 collection kit (DNA Genotek Inc., Ottawa, Canada), and the nucleic acids were extracted following the manufacturer's instructions.

Ethics approval

This study was approved by the Medical Research Advisory Committee of PNG (National Department of Health) under research ethics clearance MRAC 16.21 and by the French Ethics Committees (Committees of Protection of Persons CPP 25/21_3, n°SI : 21.01.21.42754). Permission to conduct research in PNG was granted by the National Research Institute of PNG (permit 99902292358), with full support from the School of Humanities and Social Sciences, University of PNG. All samples were collected from healthy unrelated adult donors, all of whom provided written informed consent. In each sampling location, a full presentation of the project was made, followed by discussion with each donor to ensure that they fully understood the project. Participants were surveyed for language affiliation(s), current residence, date and place of birth, and a short genealogy up to three generations to establish regional ancestry.

Variant calling from human whole genome sequences

For the 252 samples we used reads extracted from WGS from (1) previously own published studies (Brucato et al. 2021), and (2) from newly unpublished data (Table S1). The sequencing procedure has been described previously (Brucato et al. 2021). Briefly, sequencing libraries were prepared using the TruSeq DNA PCR-Free HT kit, and the 150-bp paired-end reads were sequenced on the Illumina HiSeq X5 sequencer. Raw reads were processed simultaneously, starting with trimming of FASTQ files with fastp v0.23.2 (Chen et al. 2018) and converting to BAM using Picard Tools FastqToSam v2.26.2 (BroadInstitute 2022). The Broad Institute's GATK Germline short variant discovery (SNPs and Indels) Best Practices (Poplin et al. 2017) workflow was followed, under GATK v4.2.0.0 (Van der Auwera 2020). Individual GVCF files were created with the HaplotypeCaller tool and later combined to create multi-sample VCF files with the JointGenotyping workflow.

For retrieval of microbial reads from the WGS, we aligned the fastq files against the human reference GRCh38 (GCA 000001405.15) using minimap2-2.23 (Li 2018). The unmapped human reads were used in the microbiome analyses.

Microbiome inference from human WGS

For taxonomic assignment of the unmapped sequencing reads we resorted to KRAKEN 2 tool (Wood et al. 2019) using the Standard plus protozoa & fungi database (PlusPF; includes archaea, bacteria, viral, plasmid, human, UniVec_Core, protozoa and fungi; updated on January 27th, 2021). We then inputted the Kraken-taxonomic assignments in Bracken (Lu et al. 2017) (Bayesian Reestimation of Abundance after Classification with KRAKEN) to estimate abundance at the genus level, using the default threshold of 10 reads as the minimum number of reads required for a classification at the specified rank. To remove low background noise on the WGS results due to minor contaminants, we applied a threshold of 0.01% for which all bacterial genera below were removed, based on Pedro et al. (2022).

Microbiome genome-wide association study with Matrix eQTL

The association of host germline genotypes with microbiome genera abundance was tested using the Matrix eQTL software (v2.2) (Shabalín 2012). Before the analysis, the microbiome dataset was filtered for frequencies above 5% in at least one population and bacterial genera abundance was normalized by arc-sin transformation to obtain normality. The human variants were filtered for biallelic markers with MAF > 0.1 (minimum allele frequency) in the PNG cohort, and total sample count per genotype equal or above 6. The region of birth, region of sampling, sex, age, and principal component analysis values (PCA to account for population structure; the five first PCs were considered) were used as confounding variables in all analyses. The p-values were subsequently corrected for multiple testing using false discovery rate (FDR). Manhattan plots were created with the “qqman” library in RStudio (RStudio Team 2020).

Profiling the abundance of microbial metabolic pathways with HUMAnN 3.0

To efficiently and accurately determine the abundance of microbial pathways in a community from metagenomic data we resorted to the HUMAnN 3.0 pipeline (Beghini et al. 2021), for a total of 57 samples. From this study, 30 samples were selected based on sampling location (10 samples from Port Moresby, 10 samples from Chimbu, 10 samples from Western province) with the higher number of unmapped human reads, and for other oral microbiome samples from published studies (Lassalle et al. 2018) for Philippines hunter-gatherers (n=10) and farmers (n=10) and their curated data of the Western controls from the Human Microbiome Project (n=7). We downloaded Lassalle et al. (2018) published microbial metagenomic read data sets from ENA (www.ebi.ac.uk/ena; BioSample Accession nos. ERS1202862—ERS1202885) and processed their samples with the HUMAnN 3.0 pipeline. Samples were first mapped to detect the organisms present in the community of samples using MetaPhlan (Truong et al. 2017). For gene family annotation, we opted for the UniRef50 database, as it is preferable when dealing with very poorly characterized microbiomes. Pathway annotation was performed using MetaCyc (Caspi et al. 2016), a curated database of experimentally elucidated metabolic pathways from all domains of life.

These pathways were organized in two orders of hierarchy, class and superclass, also following MetaCyc definitions. When reads do not align to a pangenome, the “Comprehensive translated search” option was used, where a search of these unmapped reads is made against a comprehensive protein database to explain as many reads as possible. After, the output results for each sample were merged and the sum-normalizations to relative abundances were estimated. To visualize the stratified outputs, the `humann_barplot` function was used.

As recommend by the authors of HUMAnN 3.0 pipeline, the MaAsLin 2.0 tool (Mallick et al. 2021) was run in RStudio to identify significant differences in bacteria contribution to a specific bacterial pathway when comparing two population groups. This tool uses generalized linear and mixed models to accommodate the complex data in microbiome multi-omics. PNG was used as the reference group in MaAsLin 2.0 statistical evaluation.

For the representation of the overall results, the heatmaps were created with the RStudio `heatmap.2` package from the `gplot` library (RStudio Team 2020), and the charts of hierarchical data with KronaTools (Ondov et al. 2011).

Results

Oral microbiome inference from the non-human reads

A total of 252 samples were analyzed in this study divided into three groups, based on the sampling locations (Table S1): Port Moresby (n=115); Simbu province (n=58) and Western province (n=79). Reads from WGS data not mapping to the human genome were analyzed with KRAKEN 2 (Wood et al. 2019), resulting in an average of 10 892 960 reads matching microbial genomes per individual (ranging from 917 415 – 149 223 438 reads). The PNG oral microbiome detected a total of 427 microbial genera, with 389 genera belonging to the Bacteria kingdom, 20 belonging to the Viral kingdom, 7 genera belonging to the Fungi kingdom and the remaining belonging to the Archaea, Protista, Protozoa or Algae. The top 10 bacterial genera detected in our PNG samples accounted for around 80% of the oral microbiome of the PNG population: *Prevotella* (~19%), *Neisseria* (~16%) and *Streptococcus* (~13%) are the three most abundant taxa. They represent almost half (~48%) of the PNG oral microbiome, followed by *Rothia* (~8%), *Haemophilus* (~7%), *Veillonella* (~6%), *Fusobacterium* (~4%), *Schaalia* (~3%), *Porphyromonas* (~2%) and *Actinomyces* (~2%). This profile agrees with our previous results (Pedro et al. 2022), as expected as individuals from three of the locations are common to both works.

The use of the HUMAnN 3.0 tool for inference of microbe pathways gave the opportunity to compare this tool phylogenetic inference with the KRAKEN 2 in the subset of 57 samples inputted on both tools. The correlation between bacterial genera abundances inferred by the two tools was high in all the groups (r values above 0.95; Figure S1). HUMAnN 3.0 identified less very low-frequent bacterial genera than KRAKEN 2, probably due to the selection of its UniRef50 database.

Identification of host genetic variation associated with the microbiome

The association test between human variants and bacterial genera abundance in the oral microbiome of the 252 Papuan individuals led to the identification of 20

significantly associated variants after correction for multiple testing. The variants are located in six genomic regions and are significantly associated with four bacterial genera (Figure 1 and Table S2). There is a strong association ($p\text{-value} < 5.00 \times 10^{-08}$) between *Klebsiella* and Cadherin 4 (*CDH4*) gene, with 14 neighbor variants above the significance threshold after correction for multiple testing. The remaining ones, have one or two significant variants followed by other neighbor variants close to the significance threshold. Other isolated significant variants were considered as spurious associations.

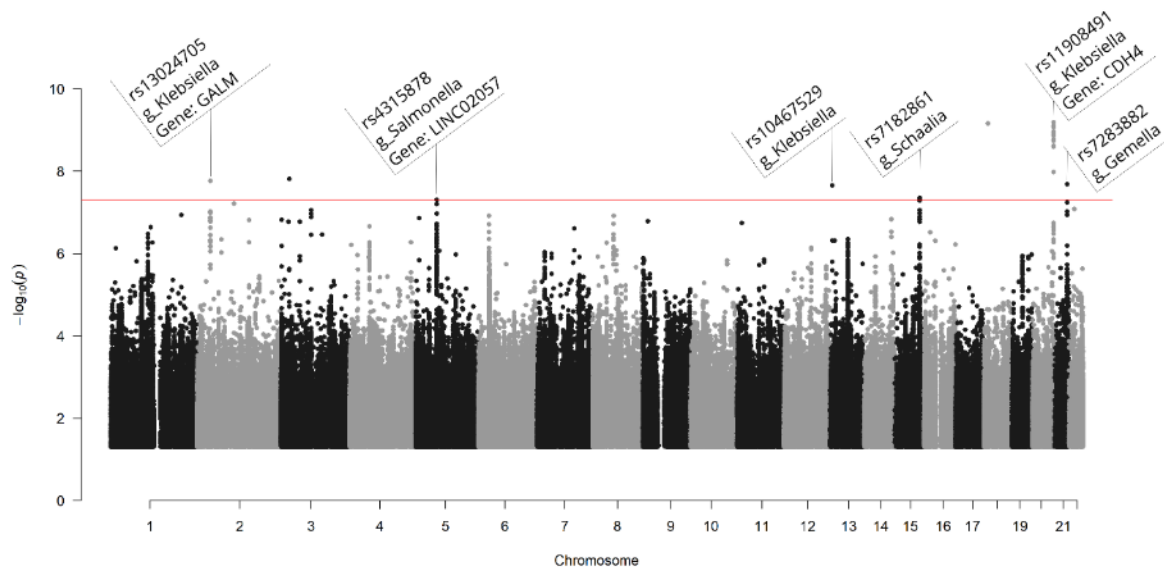


Figure 1. The 20 significant miQTLs, corresponding to six interactions between host genomic regions/variants and four bacterial genera of the oral microbiome.

Klebsiella genus, a bacterium common to the oral microbiome that can be pathogenic in situations of dysbiosis, was found associated with three host genomic regions. The strongest association is between *Klebsiella* and *CDH4* gene, a calcium-dependent cell adhesion protein from the superfamily of cadherins. Several cadherins are recognized by cell surface proteins of pathogens, thereby playing significant roles in host-pathogen interactions (Dash et al. 2021). *CDH4* is mainly expressed in the brain tissues. *Klebsiella* was also found associated with the Galactose Mutarotase (*GALM*)

gene that catalyzes the interconversion of beta-D-galactose and alpha-D-galactose during galactose metabolism. Beta-D-galactose is metabolized in the liver into glucose 1-phosphate, the primary metabolic fuel, by the action of four enzymes that constitute the Leloir pathway. It can also catalyze D-glucose although it shows a preference for galactose over glucose (Conte et al. 2021). The last association found for *Klebsiella* is with the rs10467529 SNP in an intergenic region of chromosome 13. The neighboring region for this SNP includes a selected genomic region in Papuan population containing the Zinc Finger DHHC-Type Palmitoyltransferase 20 (*ZDHHC20*) gene (Brucato et al. 2022), a palmitoyltransferase that catalyzes the addition of palmitate onto various protein substrates. It has links with microbial infection as many bacterial and viral proteins require palmitoylation by host enzymes to function properly. This includes the recently emerged SARS-CoV-2 virus (Li, Liu, Lu, et al. 2022).

Salmonella, a pathogenic bacterium responsible for infections in the gastrointestinal tract, was found in borderline significant association with an intron variant in the Long Intergenic Non-Protein Coding RNA 2057 (*LINC02057*), in chromosome 5. Additionally, this variant has been associated as an eQTL with the Elongation of Very Long Chain Fatty Acids Protein 7 (*ELOVL7*) gene in the Minor Salivary Gland (GTEx Portal). *ELOVL7* gene enables fatty acid elongase activity. During bacterial fatty acid biosynthesis, more specifically after the elongation step, a specific membrane fatty acid modification occurs, namely the formation of Cyclopropane fatty acid (*CFA*), which is suggested to adapt the bacterial cell for an adverse condition (Chen et al. 2014; Karlinsey et al. 2022).

We also found two associations between host diversity and two non-pathogenic bacterial genera involved in maintaining the oral microbiome balance: *Gemella* and an intergenic region in chromosome 21; and *Schaalia* and an intergenic region on chromosome 15. *Gemella* is a genus indigenous to the oral cavity with hemolytic activity, whose higher levels in the oral microbiome were associated with a healthier dental status, as opposed to the lower levels that were associated with patients with periodontitis (Miyoshi et al. 2021). The variant associated with *Gemella* abundance is located nearby several genes, namely Long Intergenic Non-Protein Coding RNA 323 (*LINC00323*), associated with Hereditary Hemorrhagic Telangiectasia (Tørring et al.

2014), and Beta-Secretase 2 (*BACE2*), associated with Down Syndrome (Chou 2004) and Myositis (Vattemi et al. 2001). *Schaalia* is a nitrite-producing bacterial genera common to the normal oral microbiome and has been recently associated with an increased risk of SARS-CoV-2 infection (Callahan et al. 2022). The variant associated with *Schaalia* abundance was found nearby Interferon Stimulated Exonuclease Gene 20 (*ISG20*) and Apoptosis Enhancing Nuclease (*AEN*) gene. *ISG20* is related to the innate immune system and Interferon gamma signaling pathways.

Interestingly, the variants found to be associated with the non-pathogenic *Gemella* and *Schaalia* displayed heterogeneity between population groups, thus explaining in part the microbiome heterogeneity related with human ancestry. For the rs7283882 SNP located in chromosome 21, the G allele is positively associated with a lower frequency of *Gemella* (Figure 2A), and the homozygous GG is relatively high in PNG (41.7%) and Africa (33.0%), but lower in Europe (24.5%) and more so in East Asia (10.3%). For the rs7182861 SNP located in chromosome 15, the A allele is positively associated with a higher frequency of *Schaalia* (Figure 2B), and East Asia is completely represented by the homozygous GG genotype (99.8%), while PNG presents some levels of both the homozygous AA (4.8%) and heterozygous (32.5%) genotypes, a pattern more similar to Africa (51.1% heterozygous).

The association signals detected for the pathogenic *Klebsiella* and *Salmonella* were driven by a few individuals having high abundance of these pathogens (three individuals having a mean of 49% of *Klebsiella* as opposed to 0.18% in the remaining individuals; and other three individuals having a mean of 0.05% *Salmonella* as opposed to $1.9 \times 10^{-04}\%$ in the remaining individuals. These may reflect active infection stages. Given the pathogenicity of these bacterial genera, the positively associated genotypes are low-frequent in the overall human population: rs11908491_CC 0.036 for *Klebsiella*; rs4315878_AA 0.052 for *Salmonella*.

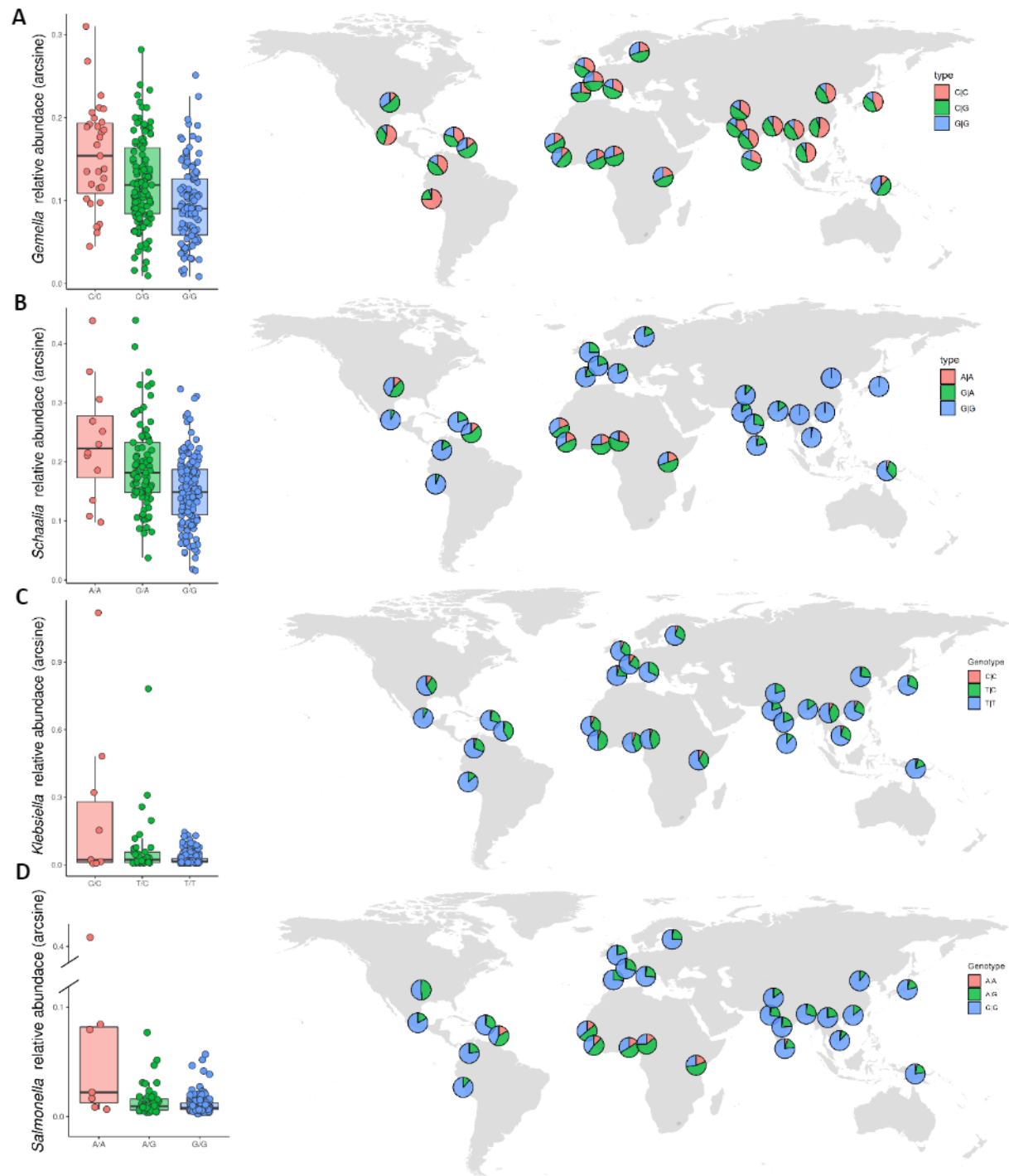


Figure 2. Relative abundance values per genotype and worldwide map with allele frequencies. A) *Gemella* and rs7283882 SNP located in the chromosome 21; B) *Schaalia* and rs7182861 SNP located in chromosome 15; C) *Klebsiella* and rs11908491 SNP located in chromosome 20; D) *Salmonella* and rs4315878 SNP located in chromosome 5.

Identification of bacterial pathways in the oral microbiome of PNG and other population groups

Oral microbiota have their own metabolism, which interacts with the host oral metabolism, either by synergy in digesting metabolites, competing for metabolites or producing xenobiotics. As the processing in HUMAnN 3.0 is time-consuming, and to homogenize sample sizes in the comparison between population groups, we limited the bacterial metabolism analysis to 30 samples from PNG (10 samples from Port Moresby, 10 samples from Chimbu, 10 samples from Western province), 20 samples from the Philippines (10 samples from hunter-gatherers, 10 samples from farmers) and seven samples from Western controls (Lassalle et al. 2018), for a total of 57 samples (Table S1). By restricting this analysis to a few individuals per population group, the metabolic impact of the bacterial genera present in most individuals is surveyed, but bacterial genera restricted to a few individuals (such as the pathogenic *Klebsiella* and *Salmonella*) are not impactful in the metabolic inference.

In total, 206 pathways were found to display significant differences in the bacterial genera that contribute to them when comparing PNG with the Philippines and Western (Table S3). These pathways belong to 34 classes, grouped in six superclasses. Biosynthesis superclass is the dominant one (63% of the pathways), followed by the Degradation/Utilization/Assimilation (D/U/A) superclass (23% of the pathways), and the Generation of Precursor Metabolites and Energy (M&E) superclass (13% of the pathways). The three last superclasses, Macromolecule Modification, Activation/Inactivation/Interconversion and the generalist Superpathways only account for 0.3% each.

From the two non-pathogenic bacterial genera found associated to host genomic variants (Figure 1), only the *Gemella* genus had significant impact in the metabolic pathways between PNG and the other two groups. *Gemella* genus significantly affected the contribution of 30 pathways from 11 classes when comparing with the Philippines (Figure 3A), and seven from three classes when comparing PNG and Western (Figure 3B). Contrary to what was observed with *Rothia* and *Neisseria*, *Gemella* contribution for the PNG pathways is lower than what is observed for the Philippines (Figure 3C-D).

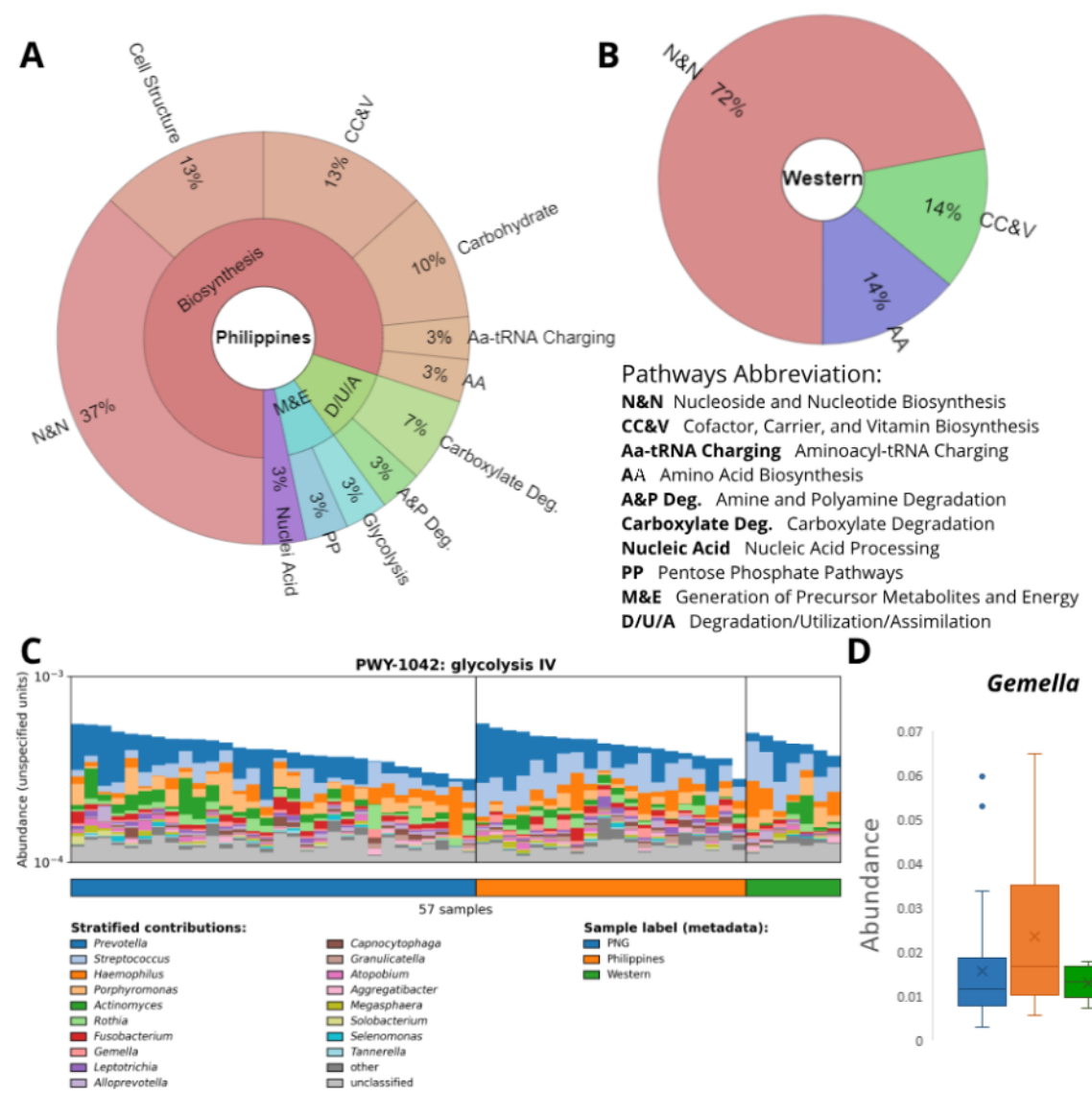


Figure 3. The impact of *Gemella* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Philippines group. B) Krona plot with the significantly different pathways between PNG and the Western controls. C) Bar-plot with abundance per individual samples of the several bacterial genera that contribute to the Glycolysis IV pathway, grouped by PNG (blue), Philippines (orange) and Western (green) and D) Boxplot for the abundance of *Gemella* in the three groups, PNG (blue), Philippines (orange) and Western (green).

This agrees to the worldwide genotype frequencies reported in Figure 2A, where we can observe that the most frequent genotypes for the PNG and Western populations are associated with lower abundance of *Gemella*, while the most frequent genotypes observed for the Asian populations (as Philippines) are the ones associated with higher abundance of *Gemella*. One of the top significant contributions of *Gemella* was in the Glycolysis IV pathway. This is one of the most important pathways, as it is used by all cells in the body for energy generation. In this pathway, the six-carbon glucose is converted into pyruvate in aerobic settings and lactate in anaerobic conditions (Puigserver 2018).

In light of the previous evidence we found that both *Neisseria* and *Rothia* are more abundant in the PNG oral microbiome (Pedro et al. 2022) we analyzed the microbial pathways significantly impacted by these two bacterial genera. *Neisseria* has a significant influence (q-value < 0.05) in PNG versus Philippines group, a total of 46 pathways from 14 classes, mostly (81%) grouped in the Biosynthesis superclass (Figure 4A) were affected. Less pathways were impacted when comparing PNG with the Western controls: 13 from seven classes, still mostly (86%) from the Biosynthesis group (Figure 4B). One of the top significant pathways between PNG and the other two groups was the superpathway of heme b biosynthesis from glutamate (Figure 4C), where *Neisseria* (and also *Actinomyces*) in PNG is compensated by *Veillonella* in Philippines and Western. This superpathway departs from glutamate to produce heme b, an essential iron containing metallo-organic cofactor, necessary as a subcellular compartment to perform diverse biological functions like hemoglobin and myoglobin (Heckl et al. 2021).

Rothia, also more abundant in PNG (Pedro et al. 2022), was found to significantly impact 39 pathways from 11 classes (Figure 5A), mostly in the Biosynthesis superclass (88%). When comparing the significantly different pathways between PNG and the Philippines, only the heme b biosynthesis I pathway reaches significance, which is also among the significantly different pathways found between PNG and Western controls.

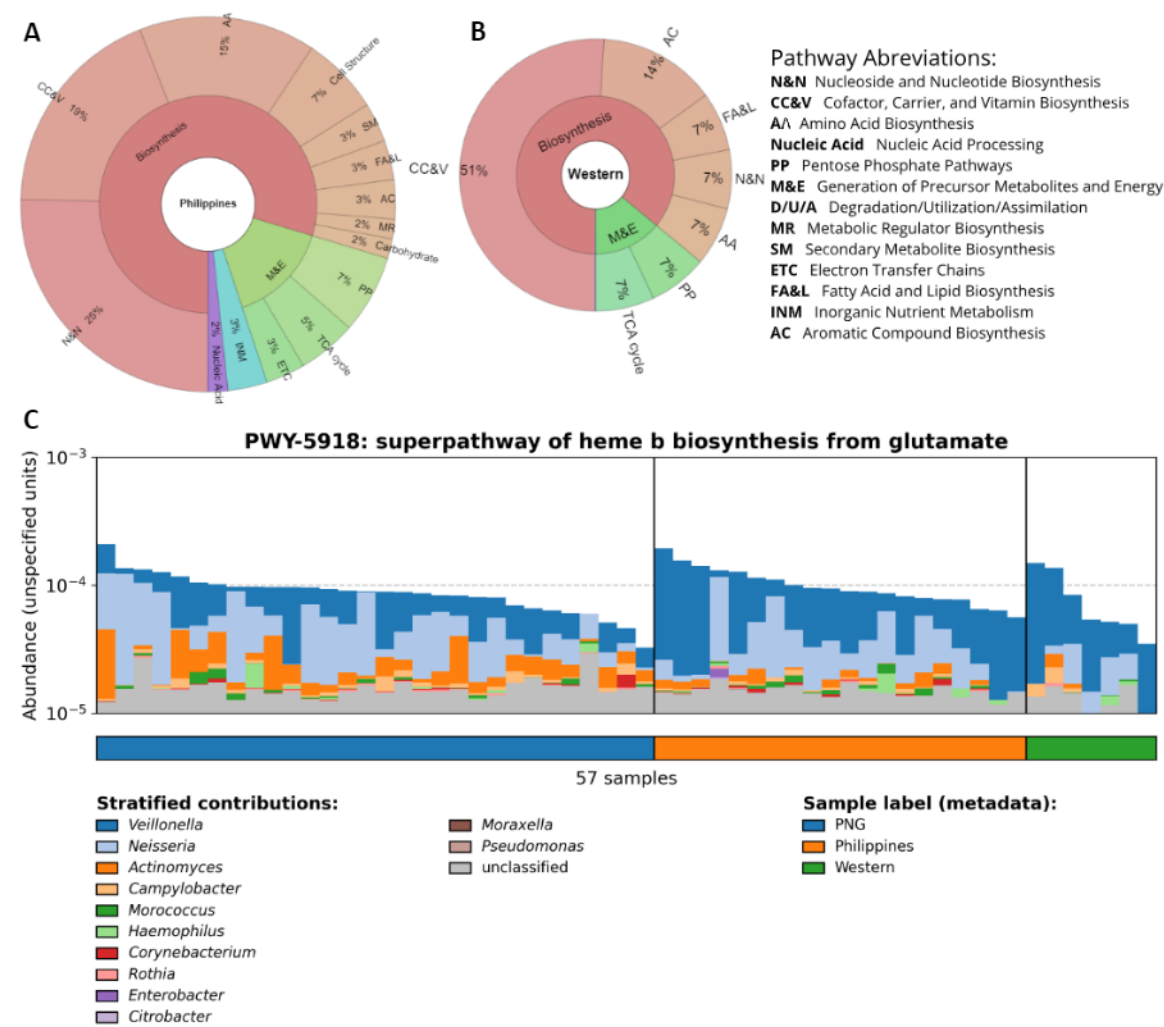


Figure 4. The impact of *Neisseria* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Philippines group. B) Krona plot with the significantly different pathways between PNG and the Western controls. C) Bar-plot with abundance of per individual samples of the several bacterial genera that contribute to the superpathway of heme b biosynthesis from glutamate, grouped by PNG (blue), Philippines (orange) and Western (green).

For the heme b biosynthesis I pathway (Figure 5B), all three population groups show identical overall proportion of microbes contributing to this pathway, but PNG displays higher contribution of *Rothia* and *Actinomyces* to this pathway, when compared to the other two populations, where *Veillonella* is the major contributor. The heme b biosynthesis I pathway is the last part of the superpathway of heme b biosynthesis. Its

synthesis requires eight enzymes to construct heme, that can be initiated by glycine or glutamate, until the last step that catalyzes the insertion of ferrous iron (Fe) into the porphyrin ring, resulting in heme b (Figure 5C). Heme biosynthesis is controlled by the availability of intracellular iron (Kafina and Paw 2017).

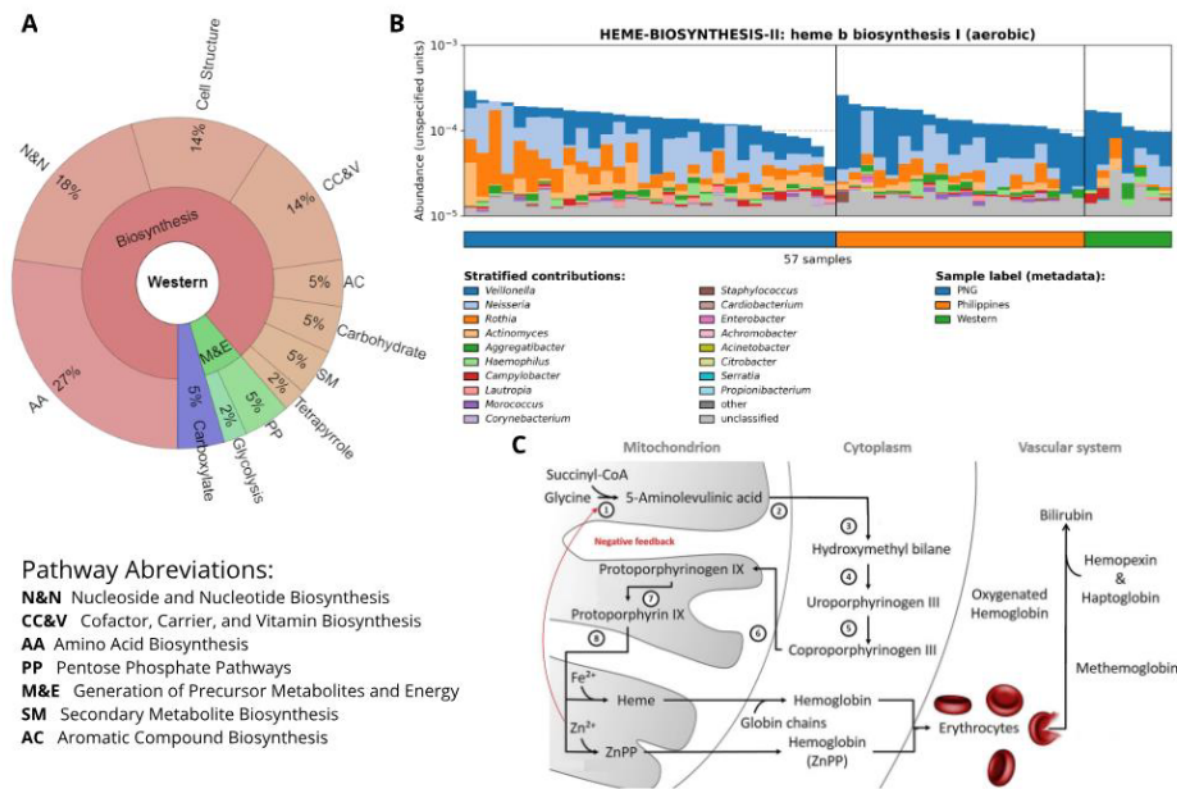


Figure 5. The impact of *Rothia* genus in the microbial pathways. A) Krona plot with the significantly different pathways between PNG and the Western controls. B) Bar-plot with abundance of per individual samples of the several bacterial genera that contribute to the heme b biosynthesis pathway, grouped by PNG (blue), Philippines (orange) and Western (green). C) Schematic representation of the eight enzymatic steps (arrows) of heme biosynthesis (enumerated next to the reaction) and the subsequent erythropoiesis and hemolysis in the vascular system. Figure adapted from Heckl et al. (2021).

Discussion

The indirect inference of the PNG oral microbiome from the non-aligned human sequenced reads allowed us to identify that *Prevotella*, *Neisseria* and *Streptococcus* represent almost half of the PNG oral microbiome diversity, as we previously described (Pedro et al. 2022). While *Prevotella* and *Streptococcus* are abundant in oral microbiomes from other populations, *Neisseria* (16%) and *Rothia* (8%) higher abundances seem to be a signature of the PNG oral microbiome. The higher abundances of the last two genera in PNG oral microbiome could be associated with human genetic variants, and for that reason we tested for this association at the whole genome and whole microbiome levels. However, we did not find any associations with *Neisseria* and *Rothia* and were only able to identify six interactions between host genomic variants and four other bacterial genera of the PNG oral microbiome. These four bacteria are low frequent, two are normally seen in oral microbiomes, *Gemella* (1.5% in PNG) and *Schaalia* (3%), and two are pathogenic, *Klebsiella* (0.78%) and *Salmonella* (0.18%). The variance of the microbiome composition explained by host genetics in our study is much lower than what has been reported in other studies (Liu et al. 2021; Wang et al. 2016), of at least 10%. One of the possible reasons for this low genetic-related association is the limited sample size in our study (n=252) when compared to other microbiome GWAS studies (~2000 samples), and hence the reduced statistical power after correction for multiple tests (Awany et al. 2018; Joo et al. 2016). Nevertheless, as *Neisseria* and *Rothia* are relatively frequent, we should have enough statistical power to detect associations with human variants in PNG, especially if these were positively selected, but that was not the case.

The detected association between a variant near the *BACE2* gene and the *Gemella* genus was supported in the independent bacterial pathway analysis. *BACE2* is a glycoprotein that functions as an aspartic protease involved in several biological functions, namely in glucose homeostasis, by the regulation of pancreatic β cell mass and function, and pigmentation (Vassar et al. 2014). And we found that the Glycolysis IV pathway was one of the top pathways affected by *Gemella*. According to the Genotype-Tissue Expression (GTEx 2013) project, the main tissue in which *BACE2* is expressed is the minor salivary gland, which is also an independent observation

supporting a possible involvement. Diseases associated with this gene are inclusion-body myositis (Vattemi et al. 2001) and type-2 diabetes (Alcarraz-Vizán et al. 2017). Also, as it is located in chromosome 21, the Down syndrome (DS) critical region, it might contribute to the pathogenesis of Alzheimer-related dementia Down Syndrome (Barbiero et al. 2003). *Gemella* has been associated with higher levels in the oral microbiome of individuals with Down Syndrome (Willis et al. 2020) and type-2 diabetes (Long et al. 2017). Those diseases are associated with the protein aggregation and deposition of intracellular amyloid deposits in specific tissues, muscles, pancreas islets and brain, respectively. Several publications support the hypothesis that amyloid structures formed by microorganisms may trigger host proteins aggregation (Seira Curto et al. 2022; Sureda et al. 2020). Although *Gemella* has not yet been associated with amyloid formation, some species from the same family (*Staphylococaceae*) have been associated with the formation of amyloid structures during biofilm formation (Sureda et al. 2020), a process in which *Gemella* is also suggested to impart (Heller et al. 2016). This is particularly interesting since the most prevalent genotype in PNG is associated with lower abundance of *Gemella*, implying some protection against these diseases. In the context of PNG epidemiology, type-2-diabetes is in fact on the rise (WHO 2020), but due mainly to a recent and drastic change in the food habits, with the adoption of a Westernized diet. So, any natural protection that PNG could have is being recently disrupted.

In the case of *Schaalia*, the variant associated was found nearby Interferon *ISG20* and *AEN* gene. This association is not detected in the bacterial pathway analysis, as it is not involved in metabolism. *ISG20* is related to the innate immune system and Interferon gamma signaling pathways, important in inhibiting a broad spectrum of viruses via direct degradation of viral RNAs due to its strong RNase properties (Deymier et al. 2022). Due to the high burden of infectious diseases in PNG, a mechanism to increase the activity of a viral RNA degrading interferon could have emerged in these populations, leading to a decrease in the bacteria more abundant during viral infections. The higher prevalence of the genotype associated with lower abundance of *Schaalia* in PNG may be an indicator of this possible adaptation. This hypothesis should be further explored in functional assays. Curiously, *Schaalia* has

been recently associated with an increased risk of SARS-CoV-2 infection (Callahan et al. 2022).

The variants detected in association for the pathogenic *Klebsiella* and *Salmonella* are low-frequent in the overall human population, and the association is driven by a small number of individuals having high abundance of these pathogens, which might reflect active infection conditions. It is difficult to interpret an association for such rare bacterial genera, as it is likely to be confounded with individual genetic profile.

No genetic associations were detected that could explain the higher abundance of these *Neisseria* and *Rothia* in PNG, although we should not rule out a possible ancestry-related influence based only in the current study. Given our results, it seems nevertheless more plausible that this pattern of the PNG oral microbiome has been driven by adaptation to environmental factors. The literature pointed to a possible iron-related adaptation due to *Rothia* producing an enterobactin that is the strongest known iron-binding siderophore, and this iron-scavenging role has positive effect on the host's iron pool (Uranga et al. 2020). In the case of *Neisseria*, several species from this bacterial genera use glutamic acid as source of carbon and nitrogen (McDonald and Johnson 1975), one of the precursors to the heme b biosynthesis pathway. Our bacterial pathway analysis confirmed that some of the top significant pathways in PNG versus Philippines and Western population groups, for both *Neisseria* and *Rothia* are iron-related pathways (Table S3). In the generated graphs it is clearly visible that *Neisseria* and *Rothia* are the main contributors to these iron-pathways in PNG, while their role is replaced by *Veillonella* in Philippines and Western population groups. It is plausible that *Rothia* and *Neisseria*, via higher efficiency in chelating iron or using alternative precursors, confer some advantage to individuals that have an iron-poor diet or under-nutrition overall. Concordantly, anemia is an important health matter in PNG, especially in children (Manning et al. 2012).

Our study has some limitations. The sampled individuals were genetically diverse, from distinct geographies and living under different environments and lifestyles, generating an incredibly diverse cohort of the PNG overall population. This diversity introduces artifacts and biases, which reduce statistical power of the microbiome

GWAS analysis (Li, Liu, Yang, et al. 2022), and were only counteracted by including some factors (region of birth, region of sampling, sex, age, and PCA values) as covariates in the analyses. Also, the microbiome data was inferred opportunistically, disregarding the impact of time of sampling and seasonality (Clemente et al. 2015), cultural factors (such as betel nut chewing; (Hernandez et al. 2017)), oral hygiene habits (Costalonga and Herzberg 2014), and nutritional status (Kato et al. 2017). Even so, our results provide first insights into how the symbiotic relationship between microbe and host can evolve in selective environments, conferring fitness advantages to the host. The detection of several host's genetic polymorphisms that interact with its microbiome and the confirmation of an iron-related oral microbiome profile in PNG, highlights the value of analyzing the microbiomes of understudied groups to fully understand how ecology, ancestry and human lifestyle impacts microbiome structure.

Acknowledgements

We appreciate the participation of all individuals of the study. We also thank Mathilde André and Georgi Hudjashov from the University of Tartu for their help in data analysis. We would like to thank all the people that contributed for the sampling campaigns: Roxanne Tsang, Jason Kariwiga, Teppsy Beni, Kylie Sesuki (University of Papua New Guinea), Alois Kuaso, and Kenneth Miamba (National Museum and Art Gallery, Papua New Guinea), and John Muke (Social Research Institute, Papua New Guinea). We acknowledge the support from the GenoToul bioinformatics facility of Genopole Toulouse Midi-Pyrénées, France; the LabEx TULIP, France; and the Centre National de Recherche en Génomique Humaine (CNRGH), Evry, France. This work was supported by the French National Research Agency (ANR; grant PAPUAEVOL n° ANR-20-CE12-0003-01), the French Ministry of Foreign and European Affairs (French Prehistoric Mission in Papua New Guinea), and the French Embassy in Papua New Guinea. The Portuguese Foundation for Science and Technology (FCT) financed the PhD grant to NP (SFRH/BD/136299/2018). i3S is supported by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020 – Operational Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT/Ministério da Ciência, Tecnologia e Inovação in the framework of the project ‘Institute for Research and Innovation in Health Sciences’ (POCI-01-0145-FEDER-007274).

References

ABUBUCKER S, SEGATA N, GOLL J, SCHUBERT AM, IZARD J, CANTAREL BL, RODRIGUEZ-MUELLER B, ZUCKER J, THIAGARAJAN M, HENRISSAT B, WHITE O, KELLEY ST, METHÉ B, SCHLOSS PD, GEVERS D, MITREVA M AND HUTTENHOWER C. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology*, 8(6), e1002358.

ALCARRAZ-VIZÁN G, CASTAÑO C, VISA M, MONTANE J, SERVITJA JM AND NOVIALS A. 2017. BACE2 suppression promotes β -cell survival and function in a model of type 2 diabetes induced by human islet amyloid polypeptide overexpression. *Cellular and Molecular Life Sciences*, 74(15), 2827-2838.

AWANY D, ALLALI I, DALVIE S, HEMMINGS S, MWAIKONO KS, THOMFORD NE, GOMEZ A, MULDER N AND CHIMUSA ER. 2018. Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. *Frontiers in Genetics*, 9, 637.

BARBIERO L, BENUSSI L, GHIDONI R, ALBERICI A, RUSSO C, SCHETTINI G, PAGANO SF, PARATI EA, MAZZOLI F, NICOSIA F, SIGNORINI S, FEUDATARI E AND BINETTI G. 2003. BACE-2 is overexpressed in Down's syndrome. *Experimental Neurology*, 182(2), 335-345.

BEGHINI F, MCIVER LJ, BLANCO-MÍGUEZ A, DUBOIS L, ASNICAR F, MAHARJAN S, MAILYAN A, MANGHI P, SCHOLZ M, THOMAS AM, VALLES-COLOMER M, WEINGART G, ZHANG Y, ZOLFO M, HUTTENHOWER C, FRANZOSA EA AND SEGATA N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10, e65088.

BROADINSTITUTE. 2022. *Picard Tools* [Accessed in 2022]. Available from <http://broadinstitute.github.io/picard>.

BRUCATO N, ANDRE M, TSANG R, SAAG L, KARIWIGA J, SESUKI K, BENI T, POMAT W, MUKEJ, MEYER V, BOLAND A, DELEUZE JF, SUDOYO H, MONDAL M, PAGANI L, GALLEGU ROMERO I, METSPALU M, COX MP, LEAVESLEY M AND RICAUT FX. 2021. Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Molecular Biology and Evolution*, 38(11), 5107-5121.

BRUCATO N, ANDRÉ M, HUDJASHOV G, MONDAL M, COX MP, LEAVESLEY M AND RICAUT FX. 2022. Chronology of natural selection in Oceanian genomes. *iScience*, 25(7), 104583.

CALLAHAN N, HATTAR M, BARBOUR T, ADAMI GR AND KAWAR N. 2022. Oral microbial taxa associated with risk for SARS-CoV-2 infection. *Frontiers in Oral Health*, 3, 886341.

CASELLI E, FABBRI C, D'ACCOLTI M, SOFFRITTI I, BASSI C, MAZZACANE S AND FRANCHI M. 2020. Defining the oral microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy picture. *BMC Microbiology*, 20(1), 120.

CASPI R, BILLINGTON R, FERRER L, FOERSTER H, FULCHER CA, KESELER IM, KOTHARI A, KRUMMENACKER M, LATENDRESSE M, MUELLER LA, ONG Q, PALEY S, SUBHRAVETI P, WEAVER DS AND KARP PD. 2016. The MetaCyc database of metabolic pathways and

enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1), D471-480.

CHEN S, ZHOU Y, CHEN Y AND GU J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884-i890.

CHEN W, GOLDEN DA AND CRITZER FJ. 2014. Salmonella survival and differential expression of fatty acid biosynthesis-associated genes in a low-water-activity food. *Letters in Applied Microbiology*, 59(2), 133-138.

CHOU KC. 2004. Insights from modeling the tertiary structure of human BACE2. *Journal of Proteome Research*, 3(5), 1069-1072.

CLEMENTE JC, PEHRSSON EC, BLASER MJ, SANDHU K, GAO Z, WANG B, MAGRIS M, HIDALGO G, CONTRERAS M, NOYA-ALARCON O, LANDER O, MCDONALD J, COX M, WALTER J, OH PL, RUIZ JF, RODRIGUEZ S, SHEN N, SONG SJ, METCALF J, KNIGHT R, DANTAS G AND DOMINGUEZ-BELLO MG. 2015. The microbiome of uncontacted Amerindians. *Science Advances*, 1(3), e1500183.

CONTE F, VAN BUURINGEN N, VOERMANS NC AND LEFEBER DJ. 2021. Galactose in human metabolism, glycosylation and congenital metabolic diseases: Time for a closer look. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1865(8), 129898.

COSTALONGA M AND HERZBERG MC. 2014. The oral microbiome and the immunobiology of periodontal disease and caries. *Immunology Letters*, 162(2 Pt A), 22-38.

DASH S, DURAIVELAN K AND SAMANTA D. 2021. Cadherin-mediated host-pathogen interactions. *Cellular Microbiology*, 23(5), e13316.

DEMMITT BA, CORLEY RP, HUIBREGTSE BM, KELLER MC, HEWITT JK, MCQUEEN MB, KNIGHT R, MCDERMOTT I AND KRAUTER KS. 2017. Genetic influences on the human oral microbiome. *BMC Genomics*, 18(1), 659.

DEYMIER S, LOUVAT C, FIORINI F AND CIMARELLI A. 2022. ISG20: an enigmatic antiviral RNase targeting multiple viruses. *FEBS Open Bio*, 12(6), 1096-1111.

GTEX. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580-585.

HECKL C, EISEL M, LANG A, HOMANN C, PAAL M, VOGESER M, RÜHM A AND SROKA R. 2021. Spectroscopic methods to quantify molecules of the heme-biosynthesis pathway: A review of laboratory work and point-of-care approaches. *Translational Biophotonics*, 3(2), e202000026.

HELLER D, HELMERHORST EJ, GOWER AC, SIQUEIRA WL, PASTER BJ AND OPPENHEIM FG. 2016. Microbial Diversity in the Early In Vivo-Formed Dental Biofilm. *Applied and Environmental Microbiology*, 82(6), 1881-1888.

HERNANDEZ BY, ZHU X, GOODMAN MT, GATEWOOD R, MENDIOLA P, QUINATA K AND PAULINO YC. 2017. Betel nut chewing, oral premalignant lesions, and the oral microbiome. *PloS One*, 12(2), e0172196.

JOO JW, HORMOZDIARI F, HAN B AND ESKIN E. 2016. Multiple testing correction in linear mixed models. *Genome Biology*, 17, 62.

KAFINA MD AND PAW BH. 2017. Intracellular iron and heme trafficking and metabolism in developing erythroblasts. *Metallomics*, 9(9), 1193-1203.

KARLINSEY JE, FUNG AM, JOHNSTON N, GOLDFINE H, LIBBY SJ AND FANG FC. 2022. Cyclopropane Fatty Acids Are Important for *Salmonella enterica* Serovar Typhimurium Virulence. *Infection and Immunity*, 90(1), e0047921.

KATO I, VASQUEZ A, MOYERBRAILEAN G, LAND S, DJURIC Z, SUN J, LIN HS AND RAM JL. 2017. Nutritional Correlates of Human Oral Microbiome. *Journal of the American College of Nutrition*, 36(2), 88-98.

KURILSHIKOV A, MEDINA-GOMEZ C, BACIGALUPE R, RADJABZADEH D, WANG J, DEMIRKAN A, LE ROY CI, RAYGOZA GARAY JA, FINNICUM CT, LIU X, ZHERNAKOVA DV, BONDER MJ, HANSEN TH, FROST F, RÜHLEMANN MC, TURPIN W, MOON JY, KIM HN, LÜLL K, BARKAN E, SHAH SA, FORNAGE M, SZOPINSKA-TOKOV J, WALLEN ZD, BORISEVICH D, AGREUS L, ANDREASSON A, BANG C, BEDRANI L, BELL JT, BISGAARD H, BOEHNKE M, BOOMSMA DI, BURK RD, CLARINGBOULD A, CROITORU K, DAVIES GE, VAN DUIJN CM, DUIJTS L, FALONY G, FU J, VAN DER GRAAF A, HANSEN T, HOMUTH G, HUGHES DA, IJZERMAN RG, JACKSON MA, JADDOE VWV, JOOSSENS M, JØRGENSEN T, KESZTHELYI D, KNIGHT R, LAAKSO M, LAUDES M, LAUNER LJ, LIEB W, LUSIS AJ, MASCLEE AAM, MOLL HA, MUJAGIC Z, QIBIN Q, ROTHSCHILD D, SHIN H, SØRENSEN SJ, STEVES

CJ, THORSEN J, TIMPSON NJ, TITO RY, VIEIRA-SILVA S, VÖLKER U, VÖLZKE H, VŐSA U, WADE KH, WALTER S, WATANABE K, WEISS S, WEISS FU, WEISSBROD O, WESTRA HJ, WILLEMSSEN G, PAYAMI H, JONKERS D, ARIAS VASQUEZ A, DE GEUS EJC, MEYER KA, STOKHOLM J, SEGAL E, ORG E, WIJMENGA C, KIM HL, KAPLAN RC, SPECTOR TD, UITTERLINDEN AG, RIVADENEIRA F, FRANKE A, LERCH MM, FRANKE L, SANNA S, D'AMATO M, PEDERSEN O, PATERSON AD, KRAAIJ R, RAES J AND ZHERNAKOVA A. 2021. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*, 53(2), 156-165.

LAMONT RJ, KOO H AND HAJISHENGALLIS G. 2018. The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology*, 16(12), 745-759.

LASSALLE F, SPAGNOLETTI M, FUMAGALLI M, SHAW L, DYBLE M, WALKER C, THOMAS MG, BAMBERG MIGLIANO A AND BALLOUX F. 2018. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*, 27(1), 182-195.

LI D, LIU Y, LU Y, GAO S AND ZHANG L. 2022. Palmitoylation of SARS-CoV-2 S protein is critical for S-mediated syncytia formation and virus entry. *Journal of Medical Virology*, 94(1), 342-348.

LI H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.

LI J, QUINQUE D, HORZ HP, LI M, RZHETSKAYA M, RAFF JA, HAYES MG AND STONEKING M. 2014. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiology*, 14(1), 316.

LI X, LIU Y, YANG X, LI C AND SONG Z. 2022. The Oral Microbiota: Community Composition, Influencing Factors, Pathogenesis, and Interventions. *Frontiers in Microbiology*, 13, 895537.

LIU X, TONG X, ZHU J, TIAN L, JIE Z, ZOU Y, LIN X, LIANG H, LI W, JU Y, QIN Y, ZOU L, LU H, ZHU S, JIN X, XU X, YANG H, WANG J, ZONG Y, LIU W, HOU Y, JIA H AND ZHANG T. 2021. Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. *Cell Discovery*, 7(1), 117.

LONG J, CAI Q, STEINWANDEL M, HARGREAVES MK, BORDENSTEIN SR, BLOT WJ, ZHENG W AND SHU XO. 2017. Association of oral microbiome with type 2 diabetes risk. *Journal of Periodontal Research*, 52(3), 636-643.

LU J, BREITWIESER FP, THIELEN P AND SALZBERG SL. 2017. Bracken: estimating species abundance in metagenomics data. *Peerj Computer Science*, 3(1), e104.

MALLICK H, RAHNAVARD A, MCIVER LJ, MA S, ZHANG Y, NGUYEN LH, TICKLE TL, WEINGART G, REN B, SCHWAGER EH, CHATTERJEE S, THOMPSON KN, WILKINSON JE, SUBRAMANIAN A, LU Y, WALDRON L, PAULSON JN, FRANZOSA EA, BRAVO HC AND HUTTENHOWER C. 2021. Multivariable association discovery in population-scale metagenomics studies. *PLoS Computational Biology*, 17(11), e1009442.

MANNING L, LAMAN M, ROSANAS-URGELL A, MICHON P, AIPIT S, BONA C, SIBA P, MUELLER I AND DAVIS TM. 2012. Severe anemia in Papua New Guinean children from a malaria-endemic area: a case-control etiologic study. *PLoS Neglected Tropical Diseases*, 6(12), e1972.

MCDONALD IJ AND JOHNSON KG. 1975. Nutritional requirements of some non-pathogenic *Neisseria* grown in simple synthetic media. *Canadian Journal of Microbiology*, 21(8), 1198-1204.

MIYOSHI T, OGE S, NAKATA S, UENO Y, UKITA H, KOUSAKA R, MIURA Y, YOSHINARI N AND YOSHIDA A. 2021. *Gemella haemolysans* inhibits the growth of the periodontal pathogen *Porphyromonas gingivalis*. *Scientific Reports*, 11(1), 11742.

MOUSA WK, CHEHADEH F AND HUSBAND S. 2022. Recent Advances in Understanding the Structure and Function of the Human Microbiome. *Frontiers in Microbiology*, 13, 825338.

ONDOV BD, BERGMAN NH AND PHILLIPPY AM. 2011. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12, 385.

PAWLEY A, ATTENBOROUGH R, GOLSON J AND HIDE R. 2005. Papuan Pasts: cultural, linguistic and biological histories of the Papuan-speaking peoples. The Australian National University, Canberra, Australia vols. Canberra: Pacific Linguistics Press, pp. 817

PEDRO N, BRUCATO N, CAVADAS B, LISANT V, CAMACHO R, KINIPI C, LEAVESLEY M, PEREIRA L AND RICAUT FX. 2022. First insight into oral microbiome diversity in Papua New Guineans reveals a specific regional signature. *Molecular Ecology*, 00, 1 – 14.

POOLE AC, GOODRICH JK, YOUNGBLUT ND, LUQUE GG, RUAUD A, SUTTER JL, WATERS JL, SHI Q, EL-HADIDI M, JOHNSON LM, BAR HY, HUSON DH, BOOTH JG AND LEY RE. 2019. Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host & Microbe*, 25(4), 553-564.e557.

POPLIN R, RUANO-RUBIO V, DEPRISTO M, FENNELL T, CARNEIRO M, VAN DER AUWERA G, KLING D, GAUTHIER L, LEVY-MOONSHINE A, ROAZEN D, SHAKIR K, THIBAUT J, CHANDRAN S, WHELAN C, LEK M, GABRIEL S, DALY M, NEALE B, MACARTHUR D AND BANKS E. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 00, 1-22.

PUIGSERVER P. 2018. Chapter 7 - Signaling Transduction and Metabolomics. In *Hematology (Seventh Edition)*, edited by R. HOFFMAN, E.J. BENZ, L.E. SILBERSTEIN, H.E. HESLOP, J.I. WEITZ, J. ANASTASI, M.E. SALAMA AND S.A. ABUTALIB: Elsevier, pp. 68-78.

RSTUDIO TEAM. RStudio: Integrated Development for R. 2020. RStudio PBC, Boston, MA, <http://www.rstudio.com/>. <http://www.rstudio.com/>.

SEIRA CURTO J, SURROCA LOPEZ A, CASALS SANCHEZ M, TIC I, FERNANDEZ GALLEGOS MR AND SANCHEZ DE GROOT N. 2022. Microbiome Impact on Amyloidogenesis. *Frontiers in Molecular Biosciences*, 9, 926702.

SHABALIN AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10), 1353-1358.

SHARON I, BERCOVICI S, PINTER RY AND SHLOMI T. 2011. Pathway-based functional analysis of metagenomes. *Journal of Computational Biology*, 18(3), 495-505.

SUREDA A, DAGLIA M, ARGÜELLES CASTILLA S, SANADGOL N, FAZEL NABAVI S, KHAN H, BELWAL T, JEANDET P, MARCHESE A, PISTOLLATO F, FORBES-HERNANDEZ T, BATTINO M, BERINDAN-NEAGOE I, D'ONOFRIO G AND NABAVI SM. 2020. Oral microbiota and Alzheimer's disease: Do all roads lead to Rome? *Pharmacological Research*, 151, 104582.

TØRRING PM, LARSEN MJ, KJELDSEN AD, OUSAGER LB, TAN Q AND BRUSGAARD K. 2014. Long non-coding RNA expression profiles in hereditary haemorrhagic telangiectasia. *PloS One*, 9(3), e90272.

TRUONG DT, TETT A, PASOLLI E, HUTTENHOWER C AND SEGATA N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, 27(4), 626-638.

URANGA CC, ARROYO P, JR., DUGGAN BM, GERWICK WH AND EDLUND A. 2020. Commensal Oral *Rothia mucilaginosa* Produces Enterobactin, a Metal-Chelating Siderophore. *mSystems*, 5(2), e00161-00120.

VAN DER AUWERA GAOGB. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. Sebastopol, CA: O'Reilly Media, pp. 467

VASSAR R, KUHN PH, HAASS C, KENNEDY ME, RAJENDRAN L, WONG PC AND LICHTENTHALER SF. 2014. Function, therapeutic potential and cell biology of BACE proteases: current status and future prospects. *Journal of Neurochemistry*, 130(1), 4-28.

VATTEMI G, ENGEL WK, MCFERRIN J, BUXBAUM JD, PASTORINO L AND ASKANAS V. 2001. Presence of BACE1 and BACE2 in muscle fibres of patients with sporadic inclusion-body myositis. *Lancet*, 358(9297), 1962-1964.

WANG J, THINGHOLM LB, SKIECEVICIENE J, RAUSCH P, KUMMEN M, HOV JR, DEGENHARDT F, HEINSEN FA, RUHLEMANN MC, SZYMCAK S, HOLM K, ESKO T, SUN J, PRICOP-JECKSTADT M, AL-DURY S, BOHOV P, BETHUNE J, SOMMER F, ELLINGHAUS D, BERGE RK, HUBENTHAL M, KOCH M, SCHWARZ K, RIMBACH G, HUBBE P, PAN WH, SHEIBANI-TEZERJI R, HASLER R, ROSENSTIEL P, D'AMATO M, CLOPPENBORG-SCHMIDT K, KUNZEL S, LAUDES M, MARSCHALL HU, LIEB W, NOTHLINGS U, KARLSEN TH, BAINES JF AND FRANKE A. 2016. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nature Genetics*, 48(11), 1396-1406.

WHO. 2022. *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*. World Health Organization [Accessed in 2022]. Available from

<https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>.

WILLIS JR, IRAOLA-GUZMÁN S, SAUS E, KSIEZOPOLSKA E, COZZUTO L, BEJARANO LA, ANDREU-SOMAVILLA N, ALLOZA-TRABADO M, PUIG-SOLA A, BLANCO A, BROGLIO E, CAROLIS C, HECHT J, PONOMARENKO J AND GABALDÓN T. 2020. Oral microbiome in down syndrome and its implications on oral health. *Journal of Oral Microbiology*, 13(1), 1865690.

WOOD DE, LU J AND LANGMEAD B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 257.

IV. CONCLUSION

The analysis of genetic (mtDNA and WGS) and oral microbiomic data of PNG, one of the most unique and least studied populations worldwide, allowed to shed light into the still debated scenario of human arrival and population substructure in Sahul and uncover a biological adaptation of the oral microbiome in these highly diverse population groups.

This study proposes that the initial arrival of the original settlers lies within a broad window, between 50 – 65 kya, in concordance to the confidence intervals obtained for some of the oldest Sahul mtDNA haplogroups, and also the oldest archeological remains (Clarkson et al. 2017). The mtDNA diversity characterization in Northern and Southern Sahul showed different deep-rooted lineages with strong geographic structuring for each region. The mtDNA lineages from Northern Sahul do not derive from the ones observed in Southern Sahul and all have roots in the same age range of the initial settlement phase of Sahul (>50 kya), which indicates different origins. The most likely scenario supports that in the early settlement phase (>50 kya), two (or more) groups of settlers, originating from a common ancestor in Sunda, carried different sets of mtDNA haplogroups and, as they reached Sahul separately, one group settled Southern Sahul and the other group Northern Sahul. The group of settlers that reached the Southern region most likely favored the previously described southern route (Wurster and Bird 2014), across the southern islands, via Timor and the northwest shelf of Australia. In contrast, the northern settlers may have benefited from the alternative northern route (Kealy et al. 2018) via Sulawesi and the Bird's Head of New Guinea. Our results of two possible routes are supported by archeological evidence of sites >40 kya old located along both paths (Summerhayes et al. 2017). More archeological and genetic data from Western New Guinea would help clarify the exact route of entry into Sahul (Brucato, Andre, et al. 2021; Taufik et al. 2022).

Previous studies (Tobler et al. 2017) have shown that after the initial settlement, Southern Sahul haplogroups displayed rapid diversification upon arrival to their new environment. During the last 55 ky, a limited number of large-scale movements occurred, indicating geographic stability and a rapid adaptation of the first settlers to the new fauna, flora, and environments. In opposition, we have found here that Northern Sahul haplogroups showed low sample sizes after settlement, with

expansion only occurring in the LGM, postglacial and Holocene periods (from 30 kya onwards), indicating a delay in diversification of around 20 ky after the initial settlement. Several factors may have hindered population expansion in New Guinea, namely a severe environment or the existence of competing inhabitants in the highlands of northern Sahul (e.g., megafauna or other members of the *Homo* genus) or the availability of better resources in the coastal areas. This high impact of the LGM was also observed in Island Southeast Asia (ISEA) maternal pool (Brandão et al. 2016). The inclusion of more genetic and archeological data in further studies will be important to better understand the initial lack of diversification observed in Northern Sahul.

Population movements within and outside Northern Sahul may have been driven by the increase of the Sahul landmass during the LGM (Hope 2014). The absence of demographic expansions and the limited number of lineages may reflect refugial movements due to the colder and harsher conditions. Starting early on the LGM (30 kya) and through the postglacial periods (18 – 10 kya), unidirectional interactions from Northern Sahul to Northern Australia were observed, in agreement with the scenarios obtained from Y chromosome (Bergström et al. 2016) and nDNA (Malaspinas et al. 2016) studies. Curiously, this Northern Sahul-Northern Australia interaction stopped before the geographical separation between Australia and New Guinea (6 – 8 kya) by submersion of the Arafura plain, resulting in the currently observed Torres Strait (Lewis et al. 2013). More comparative studies between Papuans and Australians are necessary to explain why dispersal was unidirectional from Northern to Southern Sahul.

Population interactions between New Guinea and Near Oceania could be observed since the end of the LGM (around 23 kya), in agreement with archeological evidence based on animal, plant, and object (e.g., obsidian) translocation (Pawley et al. 2005). During the postglacial warming period (18 - 10 kya), geographic dispersal of Northern Sahul lineages within and outside Sahul occurred probably due to the improved climatic conditions, a scenario also observed in the neighboring region of ISEA (Soares et al. 2008). Some of New Guinea P and Q subhaplogroups were detected in Wallacea and Sunda, from the LGM and postglacial warming periods onward (Purnomo et al.

2021), with increasing interactions between northern Sahul and ISEA at the end of the postglacial warming (~11 kya). Most probably, this stronger contact was due to the increase in maritime connections, matching the patterns of dispersal of the “voyaging corridor” hypothesis (Soares et al. 2011). Interactions between New Guinea and Remote Oceania only occurred during the late Holocene (around 2 kya), and ultimately, the Austronesian dispersal and subsequent population movement facilitated by the “voyaging corridor” led to the remote Pacific settlement from Near Oceania (Ioannidis et al. 2021).

The specific genetic diversity and long isolation (50 ky) of PNG makes them a valuable case study to investigate biological evolution and adaptation. The oral microbiome is an excellent tool to address questions of adaptation of the PNG populations to their very diverse ecosystem, as the mouth is the first organ to enter in contact with food, and where mechanical and minor chemical digestions as well as immune response must begin. Most interesting, the possibility of inferring the oral microbiome from the non-human aligned reads in WGS data obtained in saliva samples opens up new avenues. In this work, the parallel use of human WGS-inferred microbiome and microbiome microarray data allowed us to confirm high reliability on the most abundant microbes (an average of 92% match for genera with abundance above 1%). Overall, the bacterial genera that define the core oral microbiome of PNG resembles the normal microbiota in the oral cavity described in other studies (Aas et al. 2005; Deo and Deshmukh 2019; Dewhirst et al. 2010). However, differences in the abundances of these genera were observed between the PNG groups, and overall between PNG and other geographical regions.

The population of East Sepik, the most geographically isolated population included in this study, still relies on a more traditional form of subsistence due to limited access to Westernized food. It shows the most distinct oral microbiome profile of all PNG groups, with significantly less richness but more unique and highly abundant genera. Accordingly, Contreras et al. (2010) showed that South American Amerindians living in a remote village showed significantly lower richness than non-Amerindians, explained by the authors as reduced selective pressures of modern life. A relatively remote coastal community in the Western province, has a higher richness, but still a

higher number of enriched genera (13 genera). Clearly, Simbu province (Highlands), Port Moresby and East Highlanders living in Port Moresby populations, the more influenced by Westernized food, have a higher richness and only one or two highly abundant genera.

An interesting finding was revealed when we contextualized our results of the oral microbiome in PNG with other non-Westernized and Westernized worldwide populations, namely Australian Aboriginals (including Torres Strait Islanders) (Handsley-Davis et al. 2021), Batwa Pygmies (Nasidze et al. 2011), hunter-gatherer's and traditional farmers from the Philippines and Western controls (Lassalle et al. 2018). PNG and the Australian Aboriginals (including Torres Strait Islanders) presented higher abundances of *Neisseria* and *Rothia*, suggesting a possible Sahul-specific oral microbiome adaptation, despite our observation that they descend from at least two founder groups, and being isolated for a long period. It is well-known that pathogens and diet are motors for human adaptation, so this microbiome profile can be a result of several factors. Both *Rothia* and *Neisseria* show some associations with iron-related roles, *Rothia* genus by the production of iron-scavenging enterobactin (Uranga et al. 2020) and both *Rothia* and *Neisseria* were found decrease in saliva of iron-deficient individuals, proving their significance in the host's iron pool (Xi et al. 2019). Both populations have comparable subsistence styles, with a mixture of animal foods and traditional agriculture products, such as sweet and tangy fruits, fungi, starchy seeds and tubers (Brand-Miller and Holt 1998). The proximity to coastal areas also facilitated the addition of shellfish and seaweeds to their diets (O'Dea 1991). The exceptional environmental connection between these two regions should also be considered as, geographically, PNG and Australia are separated from Southeast Asia by a strong ecological barrier, that created a distinct biogeographic region with unique flora and fauna since the Pleistocene (Cámara-Leret et al. 2020).

The impact of small-scale human migration on the oral microbiome was briefly approached in this study, with the comparison between the East Highlanders living in the highlands and the East Highlanders living in Port Moresby. Contrary to what is observed in gut microbiome (Vangay et al. 2018) – a complete shift of gut microbiome diversity and function from the native to the new environment – the oral microbiome

suggests a merge of both the initial oral microbiota (highlands) and the acquired microbiota (Port Moresby). It would be of considerable interest to determine in more detail the plasticity of the oral microbiome over time and how it might be influenced by a change in diet and/or a change to a new location.

This signature of a higher abundance of *Neisseria* and *Rothia* in PNG could be a result of host-associated selection, or an environmental adaptation related to microbial metabolism. The mbQTL analysis performed in this work at the whole genome and whole microbiome levels uncovered no association between host diversity and *Neisseria* and *Rothia* in PNG. However, the microbial pathway analysis showed an enrichment in the contribution of *Neisseria* and *Rothia* in iron-related pathways, when comparing PNG with Philippines and Western population groups. In these two population groups, *Veillonella* appears as the major contributing genera to those pathways. These bacteria may have a higher ability in capturing iron or using alternative precursors, which would be advantageous to individuals with a known history of iron deficiency, either due to under-nutrition or disease, such as anemia (Manning et al. 2012). *Rothia* has a positive role on the host's iron pool (Uranga et al. 2020) by the production of an iron-scavenging enterobactin, that will bind to available iron and will then be used by the bacteria and the host in multiple functions for a variety of biochemical mechanisms (Raymond et al. 2003). *Neisseria* species are known to use glutamic acid, a precursor to the heme b biosynthesis pathway, as source of carbon and nitrogen (McDonald and Johnson 1975). Both genera are important in maintaining iron homeostasis. As iron deficit is a critical health matter in PNG and Indigenous Australian communities (Leonard et al. 2019; Manning et al. 2012), this pattern of the oral microbiome in these populations could be a regional adaptation to, in part, counteract this health issue. These results suggests that this pattern of the PNG oral microbiome has more likely been driven by adaptation to environmental factors, although a possible ancestry-related influence should not be ruled out based only in the current study. *In vitro* infection assays with controlled iron conditions could be performed to confirm the *Rothia* and *Neisseria* versus *Veillonella* role in iron-poor conditions.

The mbQTL analysis identified six interactions between host genomic variants and four low frequent bacterial genera of the PNG oral microbiome. The *Gemella* genus was found associated with a variant near the *BACE2* gene and was also altered in the bacterial pathway analysis (Glycolysis IV pathway). BACE2 plays important roles in the regulation of pancreatic β cell mass and function, and pigmentation (Vassar et al. 2014), and its dysfunction can lead to several diseases associated with amyloid deposits, namely inclusion-body myositis (Vattemi et al. 2001), type-2 diabetes (Alcarraz-Vizán et al. 2017) and Alzheimer-related dementia Down Syndrome (Barbiero et al. 2003). In individuals with these particular diseases, the oral microbiome has an enrichment of *Gemella* (Long et al. 2017; Willis et al. 2020), possibly due to a role several microorganisms (including species from the same family and similar functions of *Gemella*) have in the formation of amyloid structures, which is hypothesized to trigger host proteins aggregation (Heller et al. 2016; Seira Curto et al. 2022; Sureda et al. 2020). As PNG populations shows a genetic prevalence for a lower abundance of *Gemella*, it is possible that this variant might confer some protection against these or related diseases.

Schaalia was only found associated with a variant nearby *ISG20* and *AEN* genes. Higher levels of *Schaalia* have been recently associated with an increased risk of SARS-CoV-2 infection (Callahan et al. 2022), but PNG populations have a higher prevalence of the genotype associated with lower abundance of *Schaalia*. As *ISG20* is an important viral inhibitor (Deymier et al. 2022), a possible adaptation to increase viral inhibition could have emerged in the PNG populations due to their high burden of infectious diseases, which decrease the bacteria more abundant during viral infections, such as *Schaalia*.

Overall, the influence of host genetics in the PNG oral microbiome composition appears to be much lower than the 10% reported in other studies (Liu et al. 2021; Wang, Thingholm, et al. 2016). The limited sample size in our study (n=252) when compared to similar studies (~2000 samples) lead to a reduced statistical power after correction for multiple tests, which might explain this low number of genetic-related associations (Awany et al. 2018; Joo et al. 2016).

The long isolation and specific diversity observed in PNG created a unique and rich case study to explore the interaction between biology-culture-environment. We briefly approached the biological evolution of the oral microbiome of these populations, uncovering both genetic and environmental related adaptations. Our results highlight the importance to carry on the study of isolated populations to further uncover fitness advantages conferred to the host in selective environments. Globalization is quickly changing these remote groups which will ultimately lead to the loss of all the invaluable information not yet discovered.

REFERENCES

AAS JA, PASTER BJ, STOKES LN, OLSEN I AND DEWHIRST FE. 2005. Defining the normal bacterial flora of the oral cavity. *Journal of Clinical Microbiology*, 43(11), 5721-5732.

ADAMS SE, ARNOLD D, MURPHY B, CARROLL P, GREEN AK, SMITH AM, MARSH PD, CHEN T, MARRIOTT RE AND BRADING MG. 2017. A randomised clinical study to determine the effect of a toothpaste containing enzymes and proteins on plaque oral microbiome ecology. *Scientific Reports*, 7, 43344.

ADLER CJ, DOBNEY K, WEYRICH LS, KAIDONIS J, WALKER AW, HAAK W, BRADSHAW CJ, TOWNSEND G, SOŁTYSIAK A, ALT KW, PARKHILL J AND COOPER A. 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, 45(4), 450-455, 455e451.

ALBERS PK AND MCVEAN G. 2020. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biology*, 18(1), e3000586.

ALCARRAZ-VIZÁN G, CASTAÑO C, VISA M, MONTANE J, SERVITJA JM AND NOVIALS A. 2017. BACE2 suppression promotes β -cell survival and function in a model of type 2 diabetes induced by human islet amyloid polypeptide overexpression. *Cellular and Molecular Life Sciences*, 74(15), 2827-2838.

ALLEN J AND O'CONNELL J. 2014. Both half right: updating the evidence for dating first human arrivals in Sahul. *Australian Archaeology*, 79(1), 86-108.

ANDERSON S, BANKIER AT, BARRELL BG, DE BRUIJN MH, COULSON AR, DROUIN J, EPERON IC, NIERLICH DP, ROE BA, SANGER F, SCHREIER PH, SMITH AJ, STADEN R AND YOUNG IG. 1981. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806), 457-465.

ANDRÉ M, BRUCATO N, PLUTNIAK S, KARIWIGA J, MUKE J, MOREZ A, LEAVESLEY M, MONDAL M AND RICAUT FX. 2021. Phenotypic differences between highlanders and lowlanders in Papua New Guinea. *PloS One*, 16(7), e0253921.

ANDREWS RM, KUBACKA I, CHINNER PF, LIGHTOWLERS RN, TURNBULL DM AND HOWELL N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23(2), 147.

ARMOUR JA. 2009. Human genetics: measuring the raw material of evolution. *Current Biology*, 19(17), R736-738.

ASANTE EA, SMIDAK M, GRIMSHAW A, HOUGHTON R, TOMLINSON A, JEELANI A, JAKUBCOVA T, HAMDAN S, RICHARD-LONDT A, LINEHAN JM, BRANDNER S, ALPERS M, WHITFIELD J, MEAD S, WADSWORTH JD AND COLLINGE J. 2015. A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*, 522(7557), 478-481.

AWANY D, ALLALI I, DALVIE S, HEMMINGS S, MWAIKONO KS, THOMFORD NE, GOMEZ A, MULDER N AND CHIMUSA ER. 2018. Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. *Frontiers in Genetics*, 9, 637.

BAKER P, MACHADO P, SANTOS T, SIEVERT K, BACKHOLER K, HADJIKAKOU M, RUSSELL C, HUSE O, BELL C, SCRINIS G, WORSLEY A, FRIEL S AND LAWRENCE M. 2020. Ultra-processed foods and the nutrition transition: Global, regional and national trends, food systems transformations and political economy drivers. *Obesity Reviews*, 21(12), e13126.

BARBIERO L, BENUSSI L, GHIDONI R, ALBERICI A, RUSSO C, SCHETTINI G, PAGANO SF, PARATI EA, MAZZOLI F, NICOSIA F, SIGNORINI S, FEUDATARI E AND BINETTI G. 2003. BACE-2 is overexpressed in Down's syndrome. *Experimental Neurology*, 182(2), 335-345.

BARKER G, BARTON H, BIRD M, DALY P, DATAN I, DYKES A, FARR L, GILBERTSON D, HARRISSON B, HUNT C, HIGHAM T, KEALHOFFER L, KRIGBAUM J, LEWIS H, MCLAREN S, PAZ V, PIKE A, PIPER P, PYATT B, RABETT R, REYNOLDS T, ROSE J, RUSHWORTH G, STEPHENS M, STRINGER C, THOMPSON J AND TURNEY C. 2007. The 'human revolution'

in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *Journal of Human Evolution*, 52(3), 243-261.

BARREIRO LB AND QUINTANA-MURCI L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1), 17-30.

BARROSO E, MARTÍN V, MARTÍNEZ-CUESTA MC, PELÁEZ C AND REQUENA T. 2015. Stability of saliva microbiota during moderate consumption of red wine. *Archives of Oral Biology*, 60(12), 1763-1768.

BAUMGARTNER S, IMFELD T, SCHICHT O, RATH C, PERSSON RE AND PERSSON GR. 2009. The impact of the stone age diet on gingival conditions in the absence of oral hygiene. *Journal of Periodontology*, 80(5), 759-768.

BEGHINI F, MCIVER LJ, BLANCO-MÍGUEZ A, DUBOIS L, ASNICAR F, MAHARJAN S, MAILYAN A, MANGHI P, SCHOLZ M, THOMAS AM, VALLES-COLOMER M, WEINGART G, ZHANG Y, ZOLFO M, HUTTENHOWER C, FRANZOSA EA AND SEGATA N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10, e65088.

BERGSTRÖM A, NAGLE N, CHEN Y, MCCARTHY S, POLLARD MO, AYUB Q, WILCOX S, WILCOX L, VAN OORSCHOT RA, MCALLISTER P, WILLIAMS L, XUE Y, MITCHELL RJ AND TYLER-SMITH C. 2016. Deep Roots for Aboriginal Australian Y Chromosomes. *Current Biology*, 26(6), 809-813.

BERGSTRÖM A, OPPENHEIMER SJ, MENTZER AJ, AUCKLAND K, ROBSON K, ATTENBOROUGH R, ALPERS MP, KOKI G, POMAT W, SIBA P, XUE Y, SANDHU MS AND TYLER-SMITH C. 2017. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science*, 357(6356), 1160-1163.

BEVERIDGE TJ. 2001. Use of the gram stain in microbiology. *Biotechnic and Histochemistry*, 76(3), 111-118.

BIRD MI, BEAMAN RJ, CONDIE SA, COOPER A, ULM S AND VETH P. 2018. Palaeogeography and voyage modeling indicates early human colonization of Australia was likely from Timor-Roti. *Quaternary Science Reviews*, 191, 431-439.

BIRD MI, CONDIE SA, O'CONNOR S, O'GRADY D, REEPMEYER C, ULM S, ZEGA M, SALTRÉ F AND BRADSHAW CJA. 2019. Early human settlement of Sahul was not an accident. *Scientific Reports*, 9(1), 8220.

BIRKY CW, JR. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics*, 35, 125-148.

BLEKHMANN R, GOODRICH JK, HUANG K, SUN Q, BUKOWSKI R, BELL JT, SPECTOR TD, KEINAN A, LEY RE, GEVERS D AND CLARK AG. 2015. Host genetic variation impacts microbiome composition across human body sites. *Genome Biology*, 16, 191-203.

BLEVINS SM AND BRONZE MS. 2010. Robert Koch and the 'golden age' of bacteriology. *International Journal of Infectious Diseases*, 14(9), e744-751.

BLUM HE. 2017. The human microbiome. *Advances in Medical Sciences*, 62(2), 414-420.

BOCQUET-APPEL JP. 2011. When the world's population took off: the springboard of the Neolithic Demographic Transition. *Science*, 333(6042), 560-561.

BOMBA L, WALTER K AND SORANZO N. 2017. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, 18(1), 77-93.

BOSTANCI N, KROG MC, HUGERTH LW, BASHIR Z, FRANSSON E, BOULUND F, BELIBASAKIS GN, WANNERBERGER K, ENGSTRAND L, NIELSEN HS AND SCHUPPE-KOISTINEN I. 2021. Dysbiosis of the Human Oral Microbiome During the Menstrual Cycle and Vulnerability to the External Exposures of Smoking and Dietary Sugar. *Frontiers in Cellular and Infection Microbiology*, 11, 625229.

BOTTALICO D, CHEN Z, DUNNE A, OSTOLOZA J, MCKINNEY S, SUN C, SCHLECHT NF, FATAHZADEH M, HERRERO R, SCHIFFMAN M AND BURK RD. 2011. The oral cavity contains abundant known and novel human papillomaviruses from the Betapapillomavirus and Gammapapillomavirus genera. *The Journal of Infectious Diseases*, 204(5), 787-792.

BOURKE R AND HARWOOD T. 2009. *Food and Agriculture in Papua New Guinea*. Canberra: ANU Press, pp. 638

BRADSHAW CJA, ULM S, WILLIAMS AN, BIRD MI, ROBERTS RG, JACOBS Z, LAVIANO F, WEYRICH LS, FRIEDRICH T, NORMAN K AND SALTRÉ F. 2019. Minimum founding populations for the first peopling of Sahul. *Nature Ecology & Evolution*, 3(7), 1057-1063.

BRAND-MILLER JC AND HOLT SHA. 1998. Australian Aboriginal plant foods: a consideration of their nutritional composition and health implications. *Nutrition Research Reviews*, 11(1), 5-23.

BRANDÃO A, ENG KK, RITO T, CAVADAS B, BULBECK D, GANDINI F, PALA M, MORMINA M, HUDSON B, WHITE J, KO TM, SAIDIN M, ZAFARINA Z, OPPENHEIMER S, RICHARDS MB, PEREIRA L AND SOARES P. 2016. Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. *Human Genetics*, 135(4), 363-376.

BRAY M AND SMITH P. 1985. *Education and social stratification in Papua New Guinea*. Melbourne: Longman Cheshire, pp. 227

BRUCATO N, ANDRE M, TSANG R, SAAG L, KARIWIGA J, SESUKI K, BENI T, POMAT W, MUKEJ, MEYER V, BOLAND A, DELEUZE JF, SUDOYO H, MONDAL M, PAGANI L, GALLEGO ROMERO I, METSPALU M, COX MP, LEAVESLEY M AND RICAUT FX. 2021. Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Molecular Biology and Evolution*, 38(11), 5107-5121.

BRUCATO N, ANDRÉ M, TSANG R, SAAG L, KARIWIGA J, SESUKI K, BENI T, POMAT W, MUKEJ, MEYER V, BOLAND A, DELEUZE JF, SUDOYO H, MONDAL M, PAGANI L, GALLEGO ROMERO I, METSPALU M, COX MP, LEAVESLEY M AND RICAUT FX. 2021. Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Molecular Biology and Evolution*, 38(11), 5107-5121.

BRUCATO N, ANDRÉ M, HUDJASHOV G, MONDAL M, COX MP, LEAVESLEY M AND RICAUT FX. 2022. Chronology of natural selection in Oceanian genomes. *iScience*, 25(7), 104583.

BRUMM A, OKTAVIANA AA, BURHAN B, HAKIM B, LEBE R, ZHAO JX, SULISTYARTO PH, RIRIMASSE M, ADHITYATAMA S, SUMANTRI I AND AUBERT M. 2021. Oldest cave art found in Sulawesi. *Science Advances*, 7(3).

BUKIN YS, GALACHYANTS YP, MOROZOV IV, BUKIN SV, ZAKHARENKO AS AND ZEMSKAYA TI. 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, 6, 190007.

BURCHAM ZM, GARNEAU NL, COMSTOCK SS, TUCKER RM, KNIGHT R, METCALF JL AND GENETICS OF TASTE LAB CITIZEN S. 2020. Patterns of Oral Microbiota Diversity in Adults and Children: A Crowdsourced Population Study. *Scientific Reports*, 10(1), 2133.

CALLAHAN N, HATTAR M, BARBOUR T, ADAMI GR AND KAWAR N. 2022. Oral microbial taxa associated with risk for SARS-CoV-2 infection. *Frontiers in Oral Health*, 3, 886341.

CÁMARA-LERET R, FRODIN DG, ADEMA F, ANDERSON C, APPELHANS MS, ARGENT G, ARIAS GUERRERO S, ASHTON P, BAKER WJ, BARFOD AS, BARRINGTON D, BOROSOVA R, BRAMLEY GLC, BRIGGS M, BUERKI S, CAHEN D, CALLMANDER MW, CHEEK M, CHEN CW, CONN BJ, COODE MJE, DARBYSHIRE I, DAWSON S, DRANSFIELD J, DRINKELL C, DUYFJES B, EBIHARA A, EZEDIN Z, FU LF, GIDEON O, GIRMANSYAH D, GOVAERTS R, FORTUNE-HOPKINS H, HASSEMER G, HAY A, HEATUBUN CD, HIND DJN, HOCH P, HOMOT P, HOVENKAMP P, HUGHES M, JEBB M, JENNINGS L, JIMBO T, KESSLER M, KIEW R, KNAPP S, LAMEI P, LEHNERT M, LEWIS GP, LINDER HP, LINDSAY S, LOW YW, LUCAS E, MANCERA JP, MONRO AK, MOORE A, MIDDLETON DJ, NAGAMASU H, NEWMAN MF, NIC LUGHADHA E, MELO PHA, OHLSEN DJ, PANNELL CM, PARRIS B, PEARCE L, PENNEYS DS, PERRIE LR, PETOE P, POULSEN AD, PRANCE GT, QUAKENBUSH JP, RAES N, RODDA M, ROGERS ZS, SCHUITEMAN A, SCHWARTSBURD P, SCOTLAND RW, SIMMONS MP, SIMPSON DA, STEVENS P, SUNDUE M, TESTO W, TRIAS-BLASI A, TURNER I, UTTERIDGE T, WALSINGHAM L, WEBBER BL, WEI R, WEIBLEN GD, WEIGEND M, WESTON P, DE WILDE W, WILKIE P, WILMOT-DEAR CM, WILSON HP, WOOD JRI, ZHANG LB AND VAN WELZEN PC. 2020. New Guinea has the world's richest island flora. *Nature*, 584(7822), 579-583.

CARLHOFF S, DULI A, NÄGELE K, NUR M, SKOV L, SUMANTRI I, OKTAVIANA AA, HAKIM B, BURHAN B, SYAHDAR FA, MCGAHAN DP, BULBECK D, PERSTON YL, NEWMAN K, SAIFUL AM, RIRIMASSE M, CHIA S, HASANUDDIN, PULUBUHU DAT, SURYATMAN, SUPRIADI, JEONG C, PETER BM, PRÜFER K, POWELL A, KRAUSE J, POSTH C AND BRUMM A. 2021. Genome of a middle Holocene hunter-gatherer from Wallacea. *Nature*, 596(7873), 543-547.

CARTER K, TOVU V, LANGATI JT, BUTTSWORTH M, DINGLEY L, CALO A, HARRISON G, RAO C, LOPEZ AD AND TAYLOR R. 2016. Causes of death in Vanuatu. *Population Health Metrics*, 14(1), 1-9.

CASELLI E, FABBRI C, D'ACCOLTI M, SOFFRITTI I, BASSI C, MAZZACANE S AND FRANCHI M. 2020. Defining the oral microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy picture. *BMC Microbiology*, 20(1), 120.

CASPI R, BILLINGTON R, FERRER L, FOERSTER H, FULCHER CA, KESELER IM, KOTHARI A, KRUMMENACKER M, LATENDRESSE M, MUELLER LA, ONG Q, PALEY S, SUBHRAVETI P, WEAVER DS AND KARP PD. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1), D471-480.

CAVALLI-SFORZA LL, PIAZZA A, MENOZZI P AND MOUNTAIN J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the United States of America*, 85(16), 6002-6006.

CHANG H-SC, MULLEN B, ANJEN J AND BE'SOER L. 2014. Assessing the Socioeconomic Impact of a Rural Market: The Case of Mt Hagen Market in the Papua New Guinea Highlands. *Australasian Agribusiness Perspectives*, 17, 1-20.

CHAO A, CHAZDON RL, COLWELL RK AND SHEN TJ. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, 62(2), 361-371.

CHAO A AND CHIU C-H. 2016. Species Richness: Estimation and Comparison. In *Wiley StatsRef: Statistics Reference Online*, edited by T.C. N. BALAKRISHNAN, B. EVERITT, W. PIEGORSCH, F. RUGGERI AND J.L. TEUGELS. Hoboken, NJ: John Wiley & Sons, Inc, pp. 1-26.

CHATTOPADHYAY I, VERMA M AND PANDA M. 2019. Role of Oral Microbiome Signatures in Diagnosis and Prognosis of Oral Cancer. *Technology in Cancer Research & Treatment*, 18, 1533033819867354.

CHO I AND BLASER MJ. 2012. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4), 260-270.

CHOTIVANICH K, UDOMSANGPETCH R, PATTANAPANYASAT K, CHIERAKUL W, SIMPSON J, LOOAREESUWAN S AND WHITE N. 2002. Hemoglobin E: a balanced polymorphism protective against high parasitemias and thus severe *P falciparum* malaria. *Blood*, 100(4), 1172-1176.

CLAESSON MJ, JEFFERY IB, CONDE S, POWER SE, O'CONNOR EM, CUSACK S, HARRIS HM, COAKLEY M, LAKSHMINARAYANAN B, O'SULLIVAN O, FITZGERALD GF, DEANE J, O'CONNOR M, HARNEDY N, O'CONNOR K, O'MAHONY D, VAN SINDEREN D, WALLACE M, BRENNAN L, STANTON C, MARCHESI JR, FITZGERALD AP, SHANAHAN F, HILL C, ROSS RP AND O'TOOLE PW. 2012. Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410), 178-184.

CLARK PU, DYKE AS, SHAKUN JD, CARLSON AE, CLARK J, WOHLFARTH B, MITROVICA JX, HOSTETLER SW AND MCCABE AM. 2009. The Last Glacial Maximum. *Science*, 325(5941), 710-714.

CLARKSON C, JACOBS Z, MARWICK B, FULLAGAR R, WALLIS L, SMITH M, ROBERTS RG, HAYES E, LOWE K, CARAH X, FLORIN SA, MCNEIL J, COX D, ARNOLD LJ, HUA Q, HUNTLEY J, BRAND HEA, MANNE T, FAIRBAIRN A, SHULMEISTER J, LYLE L, SALINAS M, PAGE M, CONNELL K, PARK G, NORMAN K, MURPHY T AND PARDOE C. 2017. Human occupation of northern Australia by 65,000 years ago. *Nature*, 547(7663), 306-310.

CLEARY E, HETZEL MW AND CLEMENTS ACA. 2022. A review of malaria epidemiology and control in Papua New Guinea 1900 to 2021: Progress made and future directions. *Frontiers in Epidemiology*, 2, 980795.

COGHLAN B, MILLAN J, MALAU C, KALDOR J AND TOOLE M. 2011. The HIV epidemic in Papua New Guinea. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 58(2), e48-e51.

CONTRERAS M, COSTELLO EK, HIDALGO G, MAGRIS M, KNIGHT R AND DOMINGUEZ-BELLO MG. 2010. The bacterial microbiota in the oral mucosa of rural Amerindians. *Microbiology (Reading)*, 156(Pt 11), 3282-3287.

CORDAIN L, MILLER JB, EATON SB, MANN N, HOLT SH AND SPETH JD. 2000. Plant-animal subsistence ratios and macronutrient energy estimations in worldwide hunter-gatherer diets. *The American Journal of Clinical Nutrition*, 71(3), 682-692.

COUNTRYREPORTS. 2022. *Country Reports - Papua New Guinea Geography* [October 9th 2022]. Available from <https://www.countryreports.org/country/PapuaNewGuinea/geography.htm>.

DAVID LA, MAURICE CF, CARMODY RN, GOOTENBERG DB, BUTTON JE, WOLFE BE, LING AV, DEVLIN AS, VARMA Y, FISCHBACH MA, BIDDINGER SB, DUTTON RJ AND TURNBAUGH PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484), 559-563.

DAVIS IM. 2020. Antoni van Leeuwenhoek and measuring the invisible: The context of 16th and 17th century micrometry. *Studies in History and Philosophy of Science*, 83, 75-85.

DEMENOCAL PB AND STRINGER C. 2016. Human migration: Climate and the peopling of the world. *Nature*, 538(7623), 49-50.

DEMMITT BA, CORLEY RP, HUIBREGTSE BM, KELLER MC, HEWITT JK, MCQUEEN MB, KNIGHT R, MCDERMOTT I AND KRAUTER KS. 2017. Genetic influences on the human oral microbiome. *BMC Genomics*, 18(1), 659.

DENHAM TP, HABERLE SG, LENTFER C, FULLAGAR R, FIELD J, THERIN M, PORCH N AND WINSBOROUGH B. 2003. Origins of agriculture at Kuk Swamp in the highlands of New Guinea. *Science*, 301(5630), 189-193.

DEO PN AND DESHMUKH R. 2019. Oral microbiome: Unveiling the fundamentals. *Journal of Oral and Maxillofacial Pathology*, 23(1), 122-128.

DESHPANDE NP, RIORDAN SM, CASTANO-RODRIGUEZ N, WILKINS MR AND KAAKOUSH NO. 2018. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome*, 6(1), 227-241.

DÉTROIT F, DIZON E, FALGUÈRES C, HAMEAU S, RONQUILLO W AND SÉMAH F. 2004. Upper Pleistocene *Homo sapiens* from the Tabon cave (Palawan, The Philippines): description and dating of new discoveries. *Comptes Rendus Palevol*, 3(8), 705-712.

DEWHIRST FE, CHEN T, IZARD J, PASTER BJ, TANNER AC, YU WH, LAKSHMANAN A AND WADE WG. 2010. The human oral microbiome. *Journal of Bacteriology*, 192(19), 5002-5017.

DEYMIER S, LOUVAT C, FIORINI F AND CIMARELLI A. 2022. ISG20: an enigmatic antiviral RNase targeting multiple viruses. *FEBS Open Bio*, 12(6), 1096-1111.

ELIZABETH L, MACHADO P, ZINÖCKER M, BAKER P AND LAWRENCE M. 2020. Ultra-Processed Foods and Health Outcomes: A Narrative Review. *Nutrients*, 12(7), 1955.

ENATTAH NS, JENSEN TG, NIELSEN M, LEWINSKI R, KUOKKANEN M, RASINPERA H, EL-SHANTI H, SEO JK, ALIFRANGIS M, KHALIL IF, NATAH A, ALI A, NATAH S, COMAS D, MEHDI SQ, GROOP L, VESTERGAARD EM, IMTIAZ F, RASHED MS, MEYER B, TROELSEN J AND PELTONEN L. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1), 57-72.

ESBERG A, HAWORTH S, KUJA-HALKOLA R, MAGNUSSON PKE AND JOHANSSON I. 2020. Heritability of Oral Microbiota and Immune Responses to Oral Bacteria. *Microorganisms*, 8(8), 1126.

ESCOBAR-ZEPEDA A, VERA-PONCE DE LEÓN A AND SANCHEZ-FLORES A. 2015. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics*, 6, 348.

EVERSHED RP, DAVEY SMITH G, ROFFET-SALQUE M, TIMPSON A, DIEKMANN Y, LYON MS, CRAMP LJE, CASANOVA E, SMYTH J, WHELTON HL, DUNNE J, BRYCHOVA V, ŠOBERL L, GERBAULT P, GILLIS RE, HEYD V, JOHNSON E, KENDALL I, MANNING K, MARCINIAK A, OUTRAM AK, VIGNE J-D, SHENNAN S, BEVAN A, COLLEDGE S, ALLASON-JONES L, AMKREUTZ L, ANDERS A, ARBOGAST R-M, BĂLĂȘESCU A, BÁNFFY E, BARCLAY A, BEHRENS A, BOGUCKI P, CARRANCHO ALONSO Á, CARRETERO JM, CAVANAGH N, CLAËN E, COLLADO GIRALDO H, CONRAD M, CSENGERI P, CZERNIAK L, DĘBIEC M, DENAIRE A, DOMBORÓCZKI L, DONALD C, EBERT J, EVANS C, FRANCÉS-NEGRO M, GRONENBORN D, HAACK F, HALLE M, HAMON C, HÜLSHOFF R, ILETT M, IRIARTE E, JAKUCS J, JEUNESSE C, JOHNSON M, JONES AM, KARUL N, KIOSAK D, KOTOVA N, KRAUSE R, KRETSCHMER S, KRÜGER M, LEFRANC P, LELONG O, LENNEIS E, LOGVIN A, LÜTH F, MARTON T, MARLEY J, MORTIMER R, OOSTERBEEK L, OROSS K, PAVÚK J, PECHTL J, PÉTREQUIN P, POLLARD J, POLLARD R, POWLESLAND D, PYZEL J, RACZKY P, RICHARDSON A, ROWE P, ROWLAND S, ROWLANDSON I, SAILE T, SEBŐK K, SCHIER W, SCHMALFUß G, SHARAPOVA S, SHARP H, SHERIDAN A, SHEVNINA I, SOBKOWIAK-TABAKA I, STADLER P, STÄUBLE H, STOBBE A, STOJANOVSKI D, TASIĆ N, VAN WIJK I, VOSTROVSKÁ I, VUKOVIĆ J, WOLFRAM S, ZEEB-LANZ A AND THOMAS MG. 2022. Dairying, diseases and the evolution of lactase persistence in Europe. *Nature*, 608(7922), 336-345.

EZKURDIA I, JUAN D, RODRIGUEZ JM, FRANKISH A, DIEKHANS M, HARROW J, VAZQUEZ J, VALENCIA A AND TRESS ML. 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Human Molecular Genetics*, 23(22), 5866-5878.

FÁBIÁN TK, HERMANN P, BECK A, FEJÉRDY P AND FÁBIÁN G. 2012. Salivary defense proteins: their network and role in innate and acquired oral immunity. *International Journal of Molecular Sciences*, 13(4), 4295-4320.

FAN S, HANSEN ME, LO Y AND TISHKOFF SA. 2016. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308), 54-59.

FERREIRA JC, ALSHAMALI F, PEREIRA L AND FERNANDES V. 2022. Characterization of Arabian Peninsula whole exomes: Contributing to the catalogue of human diversity. *iScience*, 25(11), 105336.

FINN RD, MISTRY J, SCHUSTER-BÖCKLER B, GRIFFITHS-JONES S, HOLLICH V, LASSMANN T, MOXON S, MARSHALL M, KHANNA A, DURBIN R, EDDY SR, SONNHAMMER EL AND BATEMAN A. 2006. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl_1), D247-251.

FLINT J, HARDING RM, BOYCE AJ AND CLEGG JB. 1998. The population genetics of the haemoglobinopathies. *Bailliere's Clinical Haematology*, 11(1), 1-51.

FOLEY WA. 2000. The Languages of New Guinea. *Annual Review of Anthropology*, 29(1), 357-404.

FRIEDLAENDER J, SCHURR T, GENTZ F, KOKI G, FRIEDLAENDER F, HORVAT G, BABB P, CERCHIO S, KAESTLE F, SCHANFIELD M, DEKA R, YANAGIHARA R AND MERRIWETHER DA. 2005. Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Molecular Biology and Evolution*, 22(6), 1506-1517.

FUMAGALLI M, MOLTKE I, GRARUP N, RACIMO F, BJERREGAARD P, JORGENSEN ME, KORNELIUSSEN TS, GERBAULT P, SKOTTE L, LINNEBERG A, CHRISTENSEN C, BRANDSLUND I, JORGENSEN T, HUERTA-SANCHEZ E, SCHMIDT EB, PEDERSEN O, HANSEN T, ALBRECHTSEN A AND NIELSEN R. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254), 1343-1347.

GALLEGO ROMERO I, BASU MALLICK C, LIEBERT A, CRIVELLARO F, CHAUBEY G, ITAN Y, METSPALU M, EAASWARKHANTH M, PITCHAPPAN R, VILLEMS R, REICH D, SINGH L, THANGARAJ K, THOMAS MG, SWALLOW DM, MIRAZÓN LAHR M AND KIVISILD T. 2012.

Herders of Indian and European cattle share their predominant allele for lactase persistence. *Molecular Biology and Evolution*, 29(1), 249-260.

GERBAULT P, LIEBERT A, ITAN Y, POWELL A, CURRAT M, BURGER J, SWALLOW DM AND THOMAS MG. 2011. Evolution of lactase persistence: an example of human niche construction. *Philosophical Transactions of The Royal Society B Biological Sciences*, 366(1566), 863-877.

GHANNOUM MA, JUREVIC RJ, MUKHERJEE PK, CUI F, SIKAROODI M, NAQVI A AND GILLEVET PM. 2010. Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals. *PLOS Pathogens*, 6(1), e1000713.

GILBERT JA, BLASER MJ, CAPORASO JG, JANSSON JK, LYNCH SV AND KNIGHT R. 2018. Current understanding of the human microbiome. *Nature Medicine*, 24(4), 392-400.

GIOVANNONI SJ, DELONG EF, SCHMIDT TM AND PACE NR. 1990. Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Applied and Environmental Microbiology*, 56(8), 2572-2575.

GOMEZ A, ESPINOZA JL, HARKINS DM, LEONG P, SAFFERY R, BOCKMANN M, TORRALBA M, KUELBS C, KODUKULA R, INMAN J, HUGHES T, CRAIG JM, HIGHLANDER SK, JONES MB, DUPONT CL AND NELSON KE. 2017. Host Genetic Control of the Oral Microbiome in Health and Disease. *Cell Host & Microbe*, 22(3), 269-278.e263.

GOMEZ A AND NELSON KE. 2017. The Oral Microbiome of Children: Development, Disease, and Implications Beyond Oral Health. *Microbial Ecology*, 73(2), 492-503.

GOODRICH JK, DAVENPORT ER, CLARK AG AND LEY RE. 2017. The Relationship Between the Human Genome and Microbiome Comes into View. *Annual Review of Genetics*, 51, 413-433.

GOTELLI NJ AND CHAO A. 2013. Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data. In *Encyclopedia of Biodiversity, second edition*, edited by L. S.A. Waltham, MA: Academic Press, pp. 195-211.

GOUDA H, LEHMANN D AND RILEY I. 2013. Measurement of cause of death in Papua New Guinea populations using locally developed verbal autopsy instrument: 1970-2001. Vol. 36, Working Paper: Health Information systems Knowledge Hub, pp. 1-45

GREENHILL AR, TSUJI H, OGATA K, NATSUHARA K, MORITA A, SOLI K, LARKINS JA, TADOKORO K, ODANI S, BABA J, NAITO Y, TOMITSUKA E, NOMOTO K, SIBA PM, HORWOOD PF AND UMEZAKI M. 2015. Characterization of the gut microbiota of Papua New Guineans using reverse transcription quantitative PCR. *PLoS One*, 10(2), e0117427.

GRICE EA AND SEGRE JA. 2011. The skin microbiome. *Nature Reviews Microbiology*, 9(4), 244-253.

GUPTA PC AND WARNAKULASURIYA S. 2002. Global epidemiology of areca nut usage. *Addiction Biology*, 7(1), 77-83.

HAFT DH, SELENGUT JD AND WHITE O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1), 371-373.

HANDSLEY-DAVIS M, SKELLY E, JOHNSON NW, KAPELLAS K, LALLOO R, KROON J AND WEYRICH LS. 2021. Biocultural Drivers of Salivary Microbiota in Australian Aboriginal and Torres Strait Islander Children. *Frontiers in Oral Health*, 2, 641328.

HANSEN TH, KERN T, BAK EG, KASHANI A, ALLIN KH, NIELSEN T, HANSEN T AND PEDERSEN O. 2018. Impact of a vegan diet on the human salivary microbiota. *Scientific Reports*, 8(1), 5847.

HARDISON RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trend in Genetics*, 16(9), 369-372.

HAWKINS S, O'CONNOR S, MALONEY TR, LITSTER M, KEALY S, FENNER JN, APLIN K, BOULANGER C, BROCKWELL S, WILLAN R, PIOTTO E AND LOUYS J. 2017. Oldest human occupation of Wallacea at Laili Cave, Timor-Leste, shows broad-spectrum foraging responses to late Pleistocene environments. *Quaternary Science Reviews*, 171, 58-72.

HELLENTHAL G, BUSBY GBJ, BAND G, WILSON JF, CAPELLI C, FALUSH D AND MYERS S. 2014. A genetic atlas of human admixture history. *Science*, 343(6172), 747-751.

HELLER D, HELMERHORST EJ, GOWER AC, SIQUEIRA WL, PASTER BJ AND OPPENHEIM FG. 2016. Microbial Diversity in the Early In Vivo-Formed Dental Biofilm. *Applied and Environmental Microbiology*, 82(6), 1881-1888.

HERNANDEZ BY, ZHU X, GOODMAN MT, GATEWOOD R, MENDIOLA P, QUINATA K AND PAULINO YC. 2017. Betel nut chewing, oral premalignant lesions, and the oral microbiome. *PloS One*, 12(2), e0172196.

HEYER E, BRAZIER L, SÉGUREL L, HEGAY T, AUSTERLITZ F, QUINTANA-MURCI L, GEORGES M, PASQUET P AND VEUILLE M. 2011. Lactase persistence in central Asia: phenotype, genotype, and evolution. *Human Biology*, 83(3), 379-392.

HGP IHGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-945.

HONG BY, HOARE A, CARDENAS A, DUPUY AK, CHOQUETTE L, SALNER AL, SCHAUER PK, HEGDE U, PETERSON DE, DONGARI-BAGTZOGLOU A, STRAUSBAUGH LD AND DIAZ PI. 2020. The Salivary Mycobiome Contains 2 Ecologically Distinct Mycotypes. *Journal of Dental Research*, 99(6), 730-738.

HOPE G. 2014. The Sensitivity of the High Mountain Ecosystems of New Guinea to Climatic Change and Anthropogenic Impact. *Arctic, Antarctic, and Alpine Research*, 46(4), 777-786.

HUA X, GOEDERT JJ, LANDI MT AND SHI J. 2016. Identifying Host Genetic Variants Associated with Microbiome Composition by Testing Multiple Beta Diversity Matrices. *Human Heredity*, 81(2), 117-126.

HUDJASHOV G, KIVISILD T, UNDERHILL PA, ENDICOTT P, SANCHEZ JJ, LIN AA, SHEN P, OEFNER P, RENFREW C, VILLEMS R AND FORSTER P. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8726-8730.

HUGHES JB, HELLMANN JJ, RICKETTS TH AND BOHANNAN BJ. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10), 4399-4406.

HUMAN MICROBIOME PROJECT C. 2012. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207-214.

HUOPONEN K, SCHURR TG, CHEN Y-S AND WALLACE DC. 2001. Mitochondrial DNA variation in an Aboriginal Australian population: evidence for genetic isolation and regional differentiation. *Human Immunology*, 62(9), 954-969.

ILARDO MA, MOLTKE I, KORNELIUSSEN TS, CHENG J, STERN AJ, RACIMO F, DE BARROS DAMGAARD P, SIKORA M, SEGUIN-ORLANDO A, RASMUSSEN S, VAN DEN MUNCKHOF ICL, TER HORST R, JOOSTEN LAB, NETEA MG, SALINGKAT S, NIELSEN R AND WILLERSLEV E. 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, 173(3), 569-580.e515.

INGMAN M, KAESMANN H, PÄÄBO S AND GYLLENSTEN U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813), 708-713.

INTERNATIONAL HAPMAP C. 2003. The International HapMap Project. *Nature*, 426(6968), 789-796.

IOANNIDIS AG, BLANCO-PORTILLO J, SANDOVAL K, HAGELBERG E, BARBERENA-JONAS C, HILL AVS, RODRÍGUEZ-RODRÍGUEZ JE, FOX K, ROBSON K, HAOA-CARDINALI S, QUINTO-CORTÉS CD, MIQUEL-POBLETE JF, AUCLAND K, PARKS T, SOFRO ASM, ÁVILA-ARCOS MC, SOCKELL A, HOMBURGER JR, ENG C, HUNTSMAN S, BURCHARD EG, GIGNOUX CR, VERDUGO RA, MORAGA M, BUSTAMANTE CD, MENTZER AJ AND MORENO-ESTRADA A. 2021. Paths and timings of the peopling of Polynesia inferred from genomic networks. *Nature*, 597(7877), 522-526.

JACOBS GS, HUDJASHOV G, SAAG L, KUSUMA P, DARUSALLAM CC, LAWSON DJ, MONDAL M, PAGANI L, RICAUT FX, STONEKING M, METSPALU M, SUDOYO H, LANSING JS AND COX MP. 2019. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*, 177(4), 1010-1021 e1032.

JIA YJ, LIAO Y, HE YQ, ZHENG MQ, TONG XT, XUE WQ, ZHANG JB, YUAN LL, ZHANG WL AND JIA WH. 2021. Association Between Oral Microbiota and Cigarette Smoking in the Chinese Population. *Frontiers in Cellular and Infection Microbiology* 11, 658203.

JOBLING MA, KIVISILD T AND TYLER-SMITH C. 2014. *Human Evolutionary Genetics*. London: Garland Science, pp. 650

JOO JW, HORMOZDIARI F, HAN B AND ESKIN E. 2016. Multiple testing correction in linear mixed models. *Genome Biology*, 17, 62.

KANEHISA M AND GOTO S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27-30.

KARLSSON EK, KWIATKOWSKI DP AND SABETI PC. 2014. Natural selection and infectious disease in human populations. *Nature Review Genetics*, 15(6), 379-393.

KAYSER M, BRAUER S, WEISS G, UNDERHILL PA, ROEWER L, SCHIEFENHÖVEL W AND STONEKING M. 2000. Melanesian origin of Polynesian Y chromosomes. *Current Biology*, 10(20), 1237-1246.

KAYSER M, BRAUER S, CORDAUX R, CASTO A, LAO O, ZHIVOTOVSKY LA, MOYSE-FAURIE C, RUTLEDGE RB, SCHIEFENHOEVEL W, GIL D, LIN AA, UNDERHILL PA, OEFNER PJ, TRENT RJ AND STONEKING M. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Molecular Biology and Evolution*, 23(11), 2234-2244.

KAYSER M, LAO O, SAAR K, BRAUER S, WANG X, NÜRNBERG P, TRENT RJ AND STONEKING M. 2008. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *The American Journal of Human Genetics*, 82(1), 194-198.

KEALY S, LOUYS J AND O'CONNOR S. 2015. Islands Under the Sea: A Review of Early Modern Human Dispersal Routes and Migration Hypotheses Through Wallacea. *The Journal of Island and Coastal Archaeology*, 1-21.

KEALY S, LOUYS J AND O'CONNOR S. 2018. Least-cost pathway models indicate northern human dispersal from Sunda to Sahul. *Journal of Human Evolution*, 125, 59-70.

KIM E, MAGEN A AND AST G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1), 125-131.

KIMURA M. 1971. Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2(2), 174-208.

KITAU R, DATTA SS, PATEL MK, HENNESSEY K, WANNEMUEHLER K, SUI G AND LAGANI W. 2015. Hepatitis B surface antigen seroprevalence among children in Papua New Guinea, 2012-2013. *American Journal of Tropical Medicine and Hygiene*, 92(3), 501-506.

KITUR U, ADAIR T, RILEY I AND LOPEZ AD. 2019. Estimating the pattern of causes of death in Papua New Guinea. *BMC Public Health*, 19(1), 1322.

KO WY, RAJAN P, GOMEZ F, SCHEINFELDT L, AN P, WINKLER CA, FROMENT A, NYAMBO TB, OMAR SA, WAMBEBE C, RANCIARO A, HIRBO JB AND TISHKOFF SA. 2013. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. *The American Journal of Human Genetics*, 93(1), 54-66.

LAMONT RJ, KOO H AND HAJISHENGALLIS G. 2018. The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology*, 16(12), 745-759.

LANGERGRABER KE, SIEDEL H, MITANI JC, WRANGHAM RW, REYNOLDS V, HUNT K AND VIGILANT L. 2007. The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. *PloS One*, 2(10), e973.

LARSSON WEXELL C, RYBERG H, SJÖBERG ANDERSSON WA, BLOMQVIST S, COLIN P, VAN BOCXLAER J AND DAHLÉN G. 2016. Antimicrobial Effect of a Single Dose of Amoxicillin on the Oral Microbiota. *Clinical Implant Dentistry and Related Research*, 18(4), 699-706.

LASSALLE F, SPAGNOLETTI M, FUMAGALLI M, SHAW L, DYBLE M, WALKER C, THOMAS MG, BAMBERG MIGLIANO A AND BALLOUX F. 2018. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*, 27(1), 182-195.

LEAVESLEY MG, BIRD MI, FIFIELD LK, HAUSLADEN PA, G.M. S AND M.L. DT. 2002. Buang Merabak: Additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. *Australian Archaeology*, 54(1), 55-57.

LEE YH, CHUNG SW, AUH QS, HONG SJ, LEE YA, JUNG J, LEE GJ, PARK HJ, SHIN SI AND HONG JY. 2021. Progress in Oral Microbiome Related to Oral and Systemic Diseases: An Update. *Diagnostics (Basel)*, 11(7), 1283.

LEONARD D, BUTTNER P, THOMPSON F, MAKRIDES M AND MCDERMOTT R. 2019. Anaemia in early childhood among Aboriginal and Torres Strait Islander children of Far North Queensland: a retrospective cohort study. *Australian and New Zealand Journal of Public Health*, 43(4), 319-327.

LEWIS SE, SLOSS CR, MURRAY-WALLACE CV, WOODROFFE CD AND SMITHERS SG. 2013. Post-glacial sea-level changes around the Australian margin: a review. *Quaternary Science Reviews*, 74, 115-138.

LI X, LIU Y, YANG X, LI C AND SONG Z. 2022. The Oral Microbiota: Community Composition, Influencing Factors, Pathogenesis, and Interventions. *Frontiers in Microbiology*, 13, 895537.

LIPSON M, SKOGLUND P, SPRIGGS M, VALENTIN F, BEDFORD S, SHING R, BUCKLEY H, PHILLIP I, WARD GK, MALLICK S, ROHLAND N, BROOMANDKHOSHBAHT N, CHERONET O, FERRY M, HARPER TK, MICHEL M, OPPENHEIMER J, SIRAK K, STEWARDSON K, AUCKLAND K, HILL AVS, MAITLAND K, OPPENHEIMER SJ, PARKS T, ROBSON K, WILLIAMS TN, KENNETT DJ, MENTZER AJ, PINHASI R AND REICH D. 2018. Population Turnover in Remote Oceania Shortly after Initial Settlement. *Current Biology*, 28(7), 1157-1165.e1157.

LIU X, TONG X, ZHU J, TIAN L, JIE Z, ZOU Y, LIN X, LIANG H, LI W, JU Y, QIN Y, ZOU L, LU H, ZHU S, JIN X, XU X, YANG H, WANG J, ZONG Y, LIU W, HOU Y, JIA H AND ZHANG T. 2021. Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. *Cell Discovery*, 7(1), 117.

LOGSDON GA, VOLLGER MR AND EICHLER EE. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597-614.

LONG T, HICKS M, YU HC, BIGGS WH, KIRKNESS EF, MENNI C, ZIERER J, SMALL KS, MANGINO M, MESSIER H, BREWERTON S, TURPAZ Y, PERKINS BA, EVANS AM, MILLER LA, GUO L, CASKEY CT, SCHORK NJ, GARNER C, SPECTOR TD, VENTER JC AND TELENTI

A. 2017. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature Genetics*, 49(4), 568-578.

LOU RN, JACOBS A, WILDER AP AND THERKILDSEN NO. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23), 5966-5993.

LOZUPONE C, HAMADY M AND KNIGHT R. 2006. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, 7, 371.

LUO S, VALENCIA CA, ZHANG J, LEE NC, SLONE J, GUI B, WANG X, LI Z, DELL S, BROWN J, CHEN SM, CHIEN YH, HWU WL, FAN PC, WONG LJ, ATWAL PS AND HUANG T. 2018. Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences of the United States of America*, 115(51), 13039-13044.

LUZZATTO L. 2012. Sick cell anaemia and malaria. *Mediterranean Journal of Hematology and Infectious Diseases*, 4(1), e2012065.

MALASPINAS AS, WESTAWAY MC, MULLER C, SOUSA VC, LAO O, ALVES I, BERGSTROM A, ATHANASIADIS G, CHENG JY, CRAWFORD JE, HEUPINK TH, MACHOLDT E, PEISCHL S, RASMUSSEN S, SCHIFFELS S, SUBRAMANIAN S, WRIGHT JL, ALBRECHTSEN A, BARBIERI C, DUPANLOUP I, ERIKSSON A, MARGARYAN A, MOLTKE I, PUGACH I, KORNELIUSSEN TS, LEVKIVSKYI IP, MORENO-MAYAR JV, NI S, RACIMO F, SIKORA M, XUE Y, AGHAKHANIAN FA, BRUCATO N, BRUNAK S, CAMPOS PF, CLARK W, ELLINGVAG S, FOURMILE G, GERBAULT P, INJIE D, KOKI G, LEAVESLEY M, LOGAN B, LYNCH A, MATISOO-SMITH EA, MCALLISTER PJ, MENTZER AJ, METSPALU M, MIGLIANO AB, MURGH A, PHIPPS ME, POMAT W, REYNOLDS D, RICAUT FX, SIBA P, THOMAS MG, WALES T, WALL CM, OPPENHEIMER SJ, TYLER-SMITH C, DURBIN R, DORTCH J, MANICA A, SCHIERUP MH, FOLEY RA, LAHR MM, BOWERN C, WALL JD, MAILUND T, STONEKING M, NIELSEN R, SANDHU MS, EXCOFFIER L, LAMBERT DM AND WILLERSLEV E. 2016. A genomic history of Aboriginal Australia. *Nature*, 538(7624), 207-214.

MANNING L, LAMAN M, ROSANAS-URGELL A, MICHON P, AIPIT S, BONA C, SIBA P, MUELLER I AND DAVIS TM. 2012. Severe anemia in Papua New Guinean children from a malaria-endemic area: a case-control etiologic study. *PLoS Neglected Tropical Diseases*, 6(12), e1972.

MAROTI Z, BOLDOGKOI Z, TOMBACZ D, SNYDER M AND KALMAR T. 2018. Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis. *BMC Genomics*, 19(1), 778-791.

MARTINEZ I, STEGEN JC, MALDONADO-GOMEZ MX, EREN AM, SIBA PM, GREENHILL AR AND WALTER J. 2015. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Reports*, 11(4), 527-538.

MATISOO-SMITH EA. 2015. Tracking Austronesian expansion into the Pacific via the paper mulberry plant. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44), 13432-13433.

MCDONALD IJ AND JOHNSON KG. 1975. Nutritional requirements of some non-pathogenic *Neisseria* grown in simple synthetic media. *Canadian Journal of Microbiology*, 21(8), 1198-1204.

MEAD S, WHITFIELD J, POULTER M, SHAH P, UPHILL J, CAMPBELL T, AL-DUJAILY H, HUMMERICH H, BECK J, MEIN CA, VERZILLI C, WHITTAKER J, ALPERS MP AND COLLINGE J. 2009. A novel protective prion protein variant that colocalizes with kuru exposure. *The New England Journal of Medicine*, 361(21), 2056-2065.

MELVIN RG AND BALLARD JWO. 2017. Cellular and population level processes influence the rate, accumulation and observed frequency of inherited and somatic mtDNA mutations. *Mutagenesis*, 32(3), 323-334.

MENON RK, GOMEZ A, BRANDT BW, LEUNG YY, GOPINATH D, WATT RM, CRIELAARD W, NELSON KE AND BOTELHO MG. 2019. Long-term impact of oral surgery with or without

amoxicillin on the oral microbiome-A prospective cohort study. *Scientific Reports*, 9(1), 18761.

MERRIWETHER DA, HODGSON JA, FRIEDLAENDER FR, ALLABY R, CERCHIO S, KOKI G AND FRIEDLAENDER JS. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37), 13034-13039.

MIJARES AS, DÉTROIT F, PIPER P, GRÜN R, BELLWOOD P, AUBERT M, CHAMPION G, CUEVAS N, DE LEON A AND DIZON E. 2010. New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. *Journal of Human Evolution*, 59(1), 123-132.

MONA S, GRUNZ KE, BRAUER S, PAKENDORF B, CASTRÌ L, SUDOYO H, MARZUKI S, BARNES RH, SCHMIDTKE J, STONEKING M AND KAYSER M. 2009. Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Molecular Biology and Evolution*, 26(8), 1865-1877.

MONROY-PÉREZ E, RODRÍGUEZ-BEDOLLA RM, GARZÓN J, VACA-PANIAGUA F, ARTURO-ROJAS JIMÉNEZ E AND PANIAGUA-CONTRERAS GL. 2020. Marked virulence and azole resistance in *Candida albicans* isolated from patients with periodontal disease. *Microbial Pathogenesis*, 148, 104436.

MORAES LC, LANG PM, ARCANJO RA, RAMPELOTTO PH, FATTURI-PAROLO CC, FERREIRA MBC AND MONTAGNER F. 2020. Microbial ecology and predicted metabolic pathways in various oral environments from patients with acute endodontic infections. *International Endodontic Journal*, 53(12), 1603-1617.

NASIDZE I, LI J, SCHROEDER R, CREASEY JL, LI M AND STONEKING M. 2011. High diversity of the saliva microbiome in Batwa Pygmies. *PloS One*, 6(8), e23352.

NDOH. 2018 STI/HIV & AIDS program annual report, National Department of Health. Port Moresby, Papua New Guinea: 2019.

NG SB, TURNER EH, ROBERTSON PD, FLYGARE SD, BIGHAM AW, LEE C, SHAFFER T, WONG M, BHATTACHARJEE A, EICHLER EE, BAMSHAD M, NICKERSON DA AND SHENDURE J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272-276.

NOBELPRIZE.ORG. 2022. The Nobel Prize in Physiology or Medicine 2022. In.: Nobel Prize Outreach AB 2022.

NURK S, KOREN S, RHIE A, RAUTIAINEN M, BZIKADZE AV, MIKHEENKO A, VOLLGER MR, ALTEMOSE N, URALSKY L, GERSHMAN A, AGANEZOV S, HOYT SJ, DIEKHANS M, LOGSDON GA, ALONGE M, ANTONARAKIS SE, BORCHERS M, BOUFFARD GG, BROOKS SY, CALDAS GV, CHEN NC, CHENG H, CHIN CS, CHOW W, DE LIMA LG, DISHUCK PC, DURBIN R, DVORKINA T, FIDDES IT, FORMENTI G, FULTON RS, FUNGTAMMASAN A, GARRISON E, GRADY PGS, GRAVES-LINDSAY TA, HALL IM, HANSEN NF, HARTLEY GA, HAUKNES M, HOWE K, HUNKAPILLER MW, JAIN C, JAIN M, JARVIS ED, KERPEDJIEV P, KIRSCH M, KOLMOGOROV M, KORLACH J, KREMITZKI M, LI H, MADURO VV, MARSCHALL T, MCCARTNEY AM, MCDANIEL J, MILLER DE, MULLIKIN JC, MYERS EW, OLSON ND, PATEN B, PELUSO P, PEVZNER PA, PORUBSKY D, POTAPOVA T, ROGAEV EI, ROSENFELD JA, SALZBERG SL, SCHNEIDER VA, SEDLAZECK FJ, SHAFIN K, SHEW CJ, SHUMATE A, SIMS Y, SMIT AFA, SOTO DC, SOVIĆ I, STORER JM, STREETS A, SULLIVAN BA, THIBAUD-NISSEN F, TORRANCE J, WAGNER J, WALENZ BP, WENGER A, WOOD JMD, XIAO C, YAN SM, YOUNG AC, ZARATE S, SURTI U, MCCOY RC, DENNIS MY, ALEXANDROV IA, GERTON JL, O'NEILL RJ, TIMP W, ZOOK JM, SCHATZ MC, EICHLER EE, MIGA KH AND PHILLIPPY AM. 2022. The complete sequence of a human genome. *Science*, 376(6588), 44-53.

O'CONNELL JF, ALLEN J, WILLIAMS MAJ, WILLIAMS AN, TURNEY CSM, SPOONER NA, KAMMINGA J, BROWN G AND COOPER A. 2018. When did Homo sapiens first reach

Southeast Asia and Sahul? Proceedings of the National Academy of Sciences of the United States of America, 115(34), 8482-8490.

O'DEA K. 1991. Traditional diet and food preferences of Australian aboriginal hunter-gatherers. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 334(1270), 233-241.

O'CONNELL JF AND ALLEN J. 2012. Forum The Restaurant At The End Of The Universe: Modelling the colonisation of Sahul. Australian Archaeology, 74(1), 5-31.

OBA PM, HOLSCHER HD, MATHAI RA, KIM J AND SWANSON KS. 2020. Diet Influences the Oral Microbiota of Infants during the First Six Months of Life. Nutrients, 12(11), 3400.

OKUDA K, ADACHI M AND IJIMA K. 1998. The efficacy of antimicrobial mouth rinses in oral health care. Bulletin of Tokyo Dental College, 39(1), 7-14.

OLIVEIRA S, NÄGELE K, CARLHOFF S, PUGACH I, KOESBARDIATI T, HÜBNER A, MEYER M, OKTAVIANA AA, TAKENAKA M, KATAGIRI C, MURTI DB, PUTRI RS, MAHIRTA, PETCHEY F, HIGHAM T, HIGHAM CFW, O'CONNOR S, HAWKINS S, KINASTON R, BELLWOOD P, ONO R, POWELL A, KRAUSE J, POSTH C AND STONEKING M. 2022. Ancient genomes from the last three millennia support multiple human dispersals into Wallacea. Nature Ecology & Evolution, 6(7), 1024-1034.

OLIVER JD, NILSSON L AND KJELLEBERG S. 1991. Formation of nonculturable *Vibrio vulnificus* cells and its relationship to the starvation state. Applied and Environmental Microbiology, 57(9), 2640-2644.

PAGANI L, LAWSON DJ, JAGODA E, MÖRSEBURG A, ERIKSSON A, MITT M, CLEMENTE F, HUDJASHOV G, DEGIORGIO M, SAAG L, WALL JD, CARDONA A, MÄGI R, WILSON SAYRES MA, KAEWERT S, INCHLEY C, SCHEIB CL, JÄRVE M, KARMIN M, JACOBS GS, ANTAO T, ILIESCU FM, KUSHNIAREVICH A, AYUB Q, TYLER-SMITH C, XUE Y, YUNUSBAYEV B,

TAMBETS K, MALLICK CB, SAAG L, POCHESHKHOVA E, ANDRIADZE G, MULLER C, WESTAWAY MC, LAMBERT DM, ZORAQI G, TURDIKULOVA S, DALIMOVA D, SABITOV Z, SULTANA GNN, LACHANCE J, TISHKOFF S, MOMYNALIEV K, ISAKOVA J, DAMBA LD, GUBINA M, NYMADAWA P, EVSEEVA I, ATRAMENTOVA L, UTEVSKA O, RICAUT FX, BRUCATO N, SUDOYO H, LETELLIER T, COX MP, BARASHKOV NA, SKARO V, MULAHASANOVIC L, PRIMORAC D, SAHAKYAN H, MORMINA M, EICHSTAEDT CA, LICHMAN DV, ABDULLAH S, CHAUBEY G, WEE JTS, MIHAILOV E, KARUNAS A, LITVINOV S, KHUSAINOVA R, EKOMASOVA N, AKHMETOVA V, KHIDIYATOVA I, MARJANOVIĆ D, YEPISKOPOSYAN L, BEHAR DM, BALANOVSKA E, METSPALU A, DERENKO M, MALYARCHUK B, VOEVODA M, FEDOROVA SA, OSIPOVA LP, LAHR MM, GERBAULT P, LEAVESLEY M, MIGLIANO AB, PETRAGLIA M, BALANOVSKY O, KHUSNUTDINOVA EK, METSPALU E, THOMAS MG, MANICA A, NIELSEN R, VILLEMS R, WILLERSLEV E, KIVISILD T AND METSPALU M. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624), 238-242.

PAGE CM, BARANZINI SE, MEVIK BH, BOS SD, HARBO HF AND ANDREASSEN BK. 2015. Assessing the Power of Exome Chips. *PloS One*, 10(10), 1-13.

PASANIUC B, ROHLAND N, MCLAREN PJ, GARIMELLA K, ZAITLEN N, LI H, GUPTA N, NEALE BM, DALY MJ, SKLAR P, SULLIVAN PF, BERGEN S, MORAN JL, HULTMAN CM, LICHTENSTEIN P, MAGNUSSON P, PURCELL SM, HAAS DW, LIANG L, SUNYAEV S, PATTERSON N, DE BAKKER PI, REICH D AND PRICE AL. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, 44(6), 631-635.

PAWLEY A, ATTENBOROUGH R, GOLSON J AND HIDE R. 2005. Papuan Pasts: cultural, linguistic and biological histories of the Papuan-speaking peoples. The Australian National University, Canberra, Australia vols. Canberra: Pacific Linguistics Press, pp. 817

PEDRO N, PINTO RJ, CAVADAS B AND PEREIRA L. 2021. Sub-Saharan African information potential to unveil adaptations to infectious disease. *Human Molecular Genetics*, 30(R1), R138-R145.

PENNARUN E, KIVISILD T, METSPALU E, METSPALU M, REISBERG T, MOISAN JP, BEHAR DM, JONES SC AND VILLEMS R. 2012. Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evolutionary Biology*, 12, 234.

PEPRAH E, XU H, TEKOLA-AYELE F AND ROYAL CD. 2015. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics*, 18(1), 40-51.

PETERSON D, BONHAM KS, ROWLAND S, PATTANAYAK CW AND KLEPAC-CERAJ V. 2021. Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Frontiers in Microbiology*, 12, 670336.

POOLE AC, GOODRICH JK, YOUNGBLUT ND, LUQUE GG, RUAUD A, SUTTER JL, WATERS JL, SHI Q, EL-HADIDI M, JOHNSON LM, BAR HY, HUSON DH, BOOTH JG AND LEY RE. 2019. Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host & Microbe*, 25(4), 553-564.e557.

POPEJOY AB AND FULLERTON SM. 2016. Genomics is failing on diversity. *Nature*, 538(7624), 161-164.

PRUFER K, RACIMO F, PATTERSON N, JAY F, SANKARARAMAN S, SAWYER S, HEINZE A, RENAUD G, SUDMANT PH, DE FILIPPO C, LI H, MALLICK S, DANNEMANN M, FU Q, KIRCHER M, KUHLWILM M, LACHMANN M, MEYER M, ONGYERTH M, SIEBAUER M, THEUNERT C, TANDON A, MOORJANI P, PICKRELL J, MULLIKIN JC, VOHR SH, GREEN RE, HELLMANN I, JOHNSON PL, BLANCHE H, CANN H, KITZMAN JO, SHENDURE J, EICHLER EE, LEIN ES, BAKKEN TE, GOLOVANOVA LV, DORONICHEV VB, SHUNKOV MV, DEREVIANKO AP, VIOLA B, SLATKIN M, REICH D, KELSO J AND PAABO S. 2014. The

complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481), 43-49.

PURNOMO GA, MITCHELL KJ, O'CONNOR S, KEALY S, TAUFIK L, SCHILLER S, ROHRLACH A, COOPER A, LLAMAS B, SUDOYO H, TEIXEIRA JC AND TOBLER R. 2021. Mitogenomes Reveal Two Major Influxes of Papuan Ancestry across Wallacea Following the Last Glacial Maximum and Austronesian Contact. *Genes (Basel)*, 12(7), 965.

QUINTANA-MURCI L. 2016. Understanding rare and common diseases in the context of human evolution. *Genome Biology*, 17(1), 225.

RAJENDHRAN J AND GUNASEKARAN P. 2011. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. *Microbiological Research*, 166(2), 99-110.

RANJAN R, RANI A, METWALLY A, MCGEE HS AND PERKINS DL. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4), 967-977.

RARAU P, VENGIAU G, GOUDA H, PHUANUKOONON S, KEVAU IH, BULLEN C, SCRAGG R, RILEY I, MARKS G AND UMEZAKI M. 2017. Prevalence of non-communicable disease risk factors in three sites across Papua New Guinea: a cross-sectional study. *BMJ global health*, 2(2), e000221.

RAYMOND KN, DERTZ EA AND KIM SS. 2003. Enterobactin: an archetype for microbial iron transport. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7), 3584-3588.

REICH D, GREEN RE, KIRCHER M, KRAUSE J, PATTERSON N, DURAND EY, VIOLA B, BRIGGS AW, STENZEL U, JOHNSON PL, MARICIC T, GOOD JM, MARQUES-BONET T, ALKAN C, FU Q, MALLICK S, LI H, MEYER M, EICHLER EE, STONEKING M, RICHARDS M, TALAMO S, SHUNKOV MV, DEREVIANKO AP, HUBLIN JJ, KELSO J, SLATKIN M AND PAABO

S. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327), 1053-1060.

RITO T, RICHARDS MB, FERNANDES V, ALSHAMALI F, CERNY V, PEREIRA L AND SOARES P. 2013. The first modern human dispersals across Africa. *PloS One*, 8(11), e80031.

ROSZAK DB, GRIMES DJ AND COLWELL RR. 1984. Viable but nonrecoverable stage of *Salmonella enteritidis* in aquatic systems. *Canadian Journal of Microbiology*, 30(3), 334-338.

SAADAOU M, SINGH P AND AL KHODOR S. 2021. Oral microbiome and pregnancy: A bidirectional relationship. *Journal of Reproductive Immunology*, 145, 103293.

SAMPAIO-MAIA B AND MONTEIRO-SILVA F. 2014. Acquisition and maturation of oral microbiome throughout childhood: An update. *Dental Research Journal*, 11(3), 291-301.

SAMPAIO-MAIA B, CALDAS IM, PEREIRA ML, PÉREZ-MONGIOVI D AND ARAUJO R. 2016. The Oral Microbiome in Health and Its Implication in Oral and Systemic Diseases. *Advances in Applied Microbiology*, 97, 171-210.

SÁNCHEZ MC, RIBEIRO-VIDAL H, ESTEBAN-FERNÁNDEZ A, BARTOLOMÉ B, FIGUERO E, MORENO-ARRIBAS MV, SANZ M AND HERRERA D. 2019. Antimicrobial activity of red wine and oenological extracts against periodontal pathogens in a validated oral biofilm model. *BMC Complementary and Alternative Medicine*, 19(1), 145.

SANGER F, NICKLEN S AND COULSON AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463-5467.

SANTONOCITO S, GIUDICE A, POLIZZI A, TROIANO G, MERLO EM, SCLAFANI R, GROSSO G AND ISOLA G. 2022. A Cross-Talk between Diet and the Oral Microbiome: Balance of Nutrition on Inflammation and Immune System's Response during Periodontitis. *Nutrients*, 14(12), 2426.

SCHAPPER A. 2017. Farming and the trans-new Guinea family: a consideration chapter. In *Language Dispersal Beyond Farming*, edited by M. ROBBEETS AND A. SAVELYEV. Amsterdam: John Benjamins Publishing Company, pp. 155–181.

SCHMIDT E AND FANG P. 2021. Papua New Guinea agri-food trade and household consumption trends point towards dietary change and increased overweight and obesity prevalence. *Global Health*, 17(1), 135.

SEDGHI L, DIMASSA V, HARRINGTON A, LYNCH SV AND KAPILA YL. 2021. The oral microbiome: Role of key organisms and complex networks in oral health and disease. *Periodontology 2000*, 87(1), 107-131.

SEGUREL L, GUARINO-VIGNON P, MARCHI N, LAFOSSE S, LAURENT R, BON C, FABRE A, HEGAY T AND HEYER E. 2020. Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biology*, 18(6), e3000742.

SEIRA CURTO J, SURROCA LOPEZ A, CASALS SANCHEZ M, TIC I, FERNANDEZ GALLEGOS MR AND SANCHEZ DE GROOT N. 2022. Microbiome Impact on Amyloidogenesis. *Frontiers in Molecular Biosciences*, 9, 926702.

SENDER R, FUCHS S AND MILO R. 2016. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3), 337-340.

SHANG Q, GAO Y, QIN T, WANG S, SHI Y AND CHEN T. 2020. Interaction of Oral and Toothbrush Microbiota Affects Oral Cavity Health. *Frontiers in Cellular and Infection Microbiology*, 10, 17.

SHARON I, BERCOVICI S, PINTER RY AND SHLOMI T. 2011. Pathway-based functional analysis of metagenomes. *Journal of Computational Biology*, 18(3), 495-505.

SHIPTON C, O'CONNOR S, REEPMAYER C, KEALY S AND JANKOWSKI N. 2020. Shell Adzes, Exotic Obsidian, and Inter-Island Voyaging in the Early and Middle Holocene of Wallacea. *The Journal of Island and Coastal Archaeology*, 15(4), 525-546.

SIERRA B, TRISKA P, SOARES P, GARCIA G, PEREZ AB, AGUIRRE E, OLIVEIRA M, CAVADAS B, REGNAULT B, ALVAREZ M, RUIZ D, SAMUELS DC, SAKUNTABHAI A, PEREIRA L AND GUZMAN MG. 2017. OSBPL10, RXRA and lipid metabolism confer African-ancestry protection against dengue haemorrhagic fever in admixed Cubans. *PLoS Pathogens*, 13(2), 1-21.

SKOGLUND P, POSTH C, SIRAK K, SPRIGGS M, VALENTIN F, BEDFORD S, CLARK GR, REEPMAYER C, PETCHEY F, FERNANDES D, FU Q, HARNEY E, LIPSON M, MALLICK S, NOVAK M, ROHLAND N, STEWARDSON K, ABDULLAH S, COX MP, FRIEDLAENDER FR, FRIEDLAENDER JS, KIVISILD T, KOKI G, KUSUMA P, MERRIWETHER DA, RICAUT FX, WEE JT, PATTERSON N, KRAUSE J, PINHASI R AND REICH D. 2016. Genomic insights into the peopling of the Southwest Pacific. *Nature*, 538(7626), 510-513.

SOARES P, TREJAUT JA, LOO JH, HILL C, MORMINA M, LEE CL, CHEN YM, HUDJASHOV G, FORSTER P, MACAULAY V, BULBECK D, OPPENHEIMER S, LIN M AND RICHARDS MB. 2008. Climate change and postglacial human dispersals in southeast Asia. *Molecular Biology and Evolution*, 25(6), 1209-1218.

SOARES P, ERMINI L, THOMSON N, MORMINA M, RITO T, ROHL A, SALAS A, OPPENHEIMER S, MACAULAY V AND RICHARDS MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *American Journal of Human Genetics*, 84(6), 740-759.

SOARES P, RITO T, TREJAUT J, MORMINA M, HILL C, TINKLER-HUNDAL E, BRAID M, CLARKE DJ, LOO JH, THOMSON N, DENHAM T, DONOHUE M, MACAULAY V, LIN M,

OPPENHEIMER S AND RICHARDS MB. 2011. Ancient voyaging and Polynesian origins. *American Journal of Human Genetics*, 88(2), 239-247.

SOARES P, ALSHAMALI F, PEREIRA JB, FERNANDES V, SILVA NM, AFONSO C, COSTA MD, MUSILOVA E, MACAULAY V, RICHARDS MB, CERNY V AND PEREIRA L. 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Molecular Biology and Evolution*, 29(3), 915-927.

SOARES P, ABRANTES D, RITO T, THOMSON N, RADIVOJAC P, LI B, MACAULAY V, SAMUELS DC AND PEREIRA L. 2013. Evaluating purifying selection in the mitochondrial DNA of various mammalian species. *PloS One*, 8(3), e58993.

SOUILMI Y, LAUTERBUR ME, TOBLER R, HUBER CD, JOHAR AS, MORADI SV, JOHNSTON WA, KROGAN NJ, ALEXANDROV K AND ENARD D. 2021. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *Current Biology*, 31(16), 3504-3514.e3509.

SPOR A, KOREN O AND LEY R. 2011. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews: Microbiology*, 9(4), 279-290.

STOJANOV S, BERLEC A AND ŠTRUKELJ B. 2020. The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel disease. *Microorganisms*, 8(11), 1715.

SUMMERHAYES GR, LEAVESLEY M, FAIRBAIRN A, MANDUI H, FIELD J, FORD A AND FULLAGAR R. 2010. Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. *Science*, 330(6000), 78-81.

SUMMERHAYES GR, FIELD JH, SHAW B AND GAFFNEY D. 2017. The archaeology of forest exploitation and change in the tropics during the Pleistocene: The case of Northern Sahul (Pleistocene New Guinea). *Quaternary International*, 448, 14-30.

SUREDA A, DAGLIA M, ARGÜELLES CASTILLA S, SANADGOL N, FAZEL NABAVI S, KHAN H, BELWAL T, JEANDET P, MARCHESE A, PISTOLLATO F, FORBES-HERNANDEZ T, BATTINO M, BERINDAN-NEAGOE I, D'ONOFRIO G AND NABAVI SM. 2020. Oral microbiota and Alzheimer's disease: Do all roads lead to Rome? Pharmacological Research, 151, 104582.

TATUSOV RL, FEDOROVA ND, JACKSON JD, JACOBS AR, KIRYUTIN B, KOONIN EV, KRYLOV DM, MAZUMDER R, MEKHEDOV SL, NIKOLSKAYA AN, RAO BS, SMIRNOV S, SVERDLOV AV, VASUDEVAN S, WOLF YI, YIN JJ AND NATALE DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4, 41.

TAUFIK L, TEIXEIRA JC, LLAMAS B, SUDOYO H, TOBLER R AND PURNOMO GA. 2022. Human Genetic Research in Wallacea and Sahul: Recent Findings and Future Prospects. Genes (Basel), 13(12), 2373.

TEREZHALMY GT, BSOUK SA, BARTIZEK RD AND BIESBROCK AR. 2005. Plaque removal efficacy of a prototype manual toothbrush versus an ADA reference manual toothbrush with and without dental floss. Journal of Contemporary Dental Practice, 6(3), 1-13.

THISSEN JB, MCLOUGHLIN K, GARDNER S, GU P, MABERY S, SLEZAK T AND JAING C. 2014. Analysis of sensitivity and rapid hybridization of a multiplexed Microbial Detection Microarray. Journal of Virological Methods, 201, 73-78.

THISSEN JB, BE NA, MCLOUGHLIN K, GARDNER S, RACK PG, SHAPERO MH, ROWLAND RRR, SLEZAK T AND JAING CJ. 2019. Axiom Microbiome Array, the next generation microarray for high-throughput pathogen and microbiome analysis. PloS One, 14(2), e0212045.

THUKRAL A. 2017. A review on measurement of Alpha diversity in biology. Agricultural Research Journal, 54, 1-10.

TISHKOFF S. 2015. GENETICS. Strength in small numbers. *Science*, 349(6254), 1282-1283.

TISHKOFF SA, VARKONYI R, CAHINHINAN N, ABBES S, ARGYROPOULOS G, DESTRO-BISOL G, DROUSIOTOU A, DANGERFIELD B, LEFRANC G, LOISELET J, PIRO A, STONEKING M, TAGARELLI A, TAGARELLI G, TOUMA EH, WILLIAMS SM AND CLARK AG. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, 293(5529), 455-462.

TOBLER R, ROHRLACH A, SOUBRIER J, BOVER P, LLAMAS B, TUKE J, BEAN N, ABDULLAH-HIGHFOLD A, AGIUS S, O'DONOGHUE A, O'LOUGHLIN I, SUTTON P, ZILIO F, WALSHE K, WILLIAMS AN, TURNEY CSM, WILLIAMS M, RICHARDS SM, MITCHELL RJ, KOWAL E, STEPHEN JR, WILLIAMS L, HAAK W AND COOPER A. 2017. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature*, 544(7649), 180-184.

TOMMASEO-PONZETTA M, ATTIMONELLI M, DE ROBERTIS M, TANZARIELLO F AND SACCONI C. 2002. Mitochondrial DNA variability of West New Guinea populations. *American Journal of Physical Anthropology*, 117(1), 49-67.

TORRONI A, ACHILLI A, MACAULAY V, RICHARDS M AND BANDELT HJ. 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*, 22(6), 339-345.

URANGA CC, ARROYO P, JR., DUGGAN BM, GERWICK WH AND EDLUND A. 2020. Commensal Oral *Rothia mucilaginosa* Produces Enterobactin, a Metal-Chelating Siderophore. *mSystems*, 5(2), e00161-00120.

VALM AM. 2019. The Structure of Dental Plaque Microbial Communities in the Transition from Health to Dental Caries and Periodontal Disease. *Journal of Molecular Biology*, 431(16), 2957-2969.

VAN OPSTAL EJ AND BORDENSTEIN SR. 2015. MICROBIOME. Rethinking heritability of the microbiome. *Science*, 349(6253), 1172-1173.

VANGAY P, JOHNSON AJ, WARD TL, AL-GHALITH GA, SHIELDS-CUTLER RR, HILLMANN BM, LUCAS SK, BEURA LK, THOMPSON EA, TILL LM, BATRES R, PAW B, PERGAMENT SL, SAENYAKUL P, XIONG M, KIM AD, KIM G, MASOPUST D, MARTENS EC, ANGKURAWARANON C, MCGREADY R, KASHYAP PC, CULHANE-PERA KA AND KNIGHTS D. 2018. US Immigration Westernizes the Human Gut Microbiome. *Cell*, 175(4), 962-972 e910.

VASSAR R, KUHN PH, HAASS C, KENNEDY ME, RAJENDRAN L, WONG PC AND LICHTENTHALER SF. 2014. Function, therapeutic potential and cell biology of BACE proteases: current status and future prospects. *Journal of Neurochemistry*, 130(1), 4-28.

VATTEMI G, ENGEL WK, MCFERRIN J, BUXBAUM JD, PASTORINO L AND ASKANAS V. 2001. Presence of BACE1 and BACE2 in muscle fibres of patients with sporadic inclusion-body myositis. *Lancet*, 358(9297), 1962-1964.

WADE WG. 2013. The oral microbiome in health and disease. *Pharmacological Research*, 69(1), 137-143.

WALLACE DC. 2015. Mitochondrial DNA variation in human radiation and disease. *Cell*, 163(1), 33-38.

WANG J, SANTIAGO E AND CABALLERO A. 2016. Prediction and estimation of effective population size. *Heredity*, 117(4), 193-206.

WANG J, THINGHOLM LB, SKIECEVICIENE J, RAUSCH P, KUMMEN M, HOV JR, DEGENHARDT F, HEINSEN FA, RUHLEMANN MC, SZYMCZAK S, HOLM K, ESKO T, SUN J, PRICOP-JECKSTADT M, AL-DURY S, BOHOV P, BETHUNE J, SOMMER F, ELLINGHAUS D, BERGE RK, HUBENTHAL M, KOCH M, SCHWARZ K, RIMBACH G, HUBBE P, PAN WH, SHEIBANI-TEZERJI R, HASLER R, ROSENSTIEL P, D'AMATO M, CLOPPENBORG-SCHMIDT K, KUNZEL S, LAUDES M, MARSCHALL HU, LIEB W, NOTHLINGS U, KARLSEN TH, BAINES JF AND FRANKE A. 2016. Genome-wide association analysis identifies variation in

vitamin D receptor and other host factors influencing the gut microbiota. *Nature Genetics*, 48(11), 1396-1406.

WESTAWAY KE, LOUYS J, AWE RD, MORWOOD MJ, PRICE GJ, ZHAO JX, AUBERT M, JOANNES-BOYAU R, SMITH TM, SKINNER MM, COMPTON T, BAILEY RM, VAN DEN BERGH GD, DE VOS J, PIKE AWG, STRINGER C, SAPTOMO EW, RIZAL Y, ZAIM J, SANTOSO WD, TRIHASCARYO A, KINSLEY L AND SULISTYANTO B. 2017. An early modern human presence in Sumatra 73,000-63,000 years ago. *Nature*, 548(7667), 322-325.

WHITTAKER RH. 1972. Evolution and Measurement of Species Diversity. *Taxon*, 21(2/3), 213-251.

WHO. *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*. World Health Organization. Available from <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>.

WICKLER S. 1990. Prehistoric Melanesian exchange and interaction: recent evidence from the Northern Solomon Islands. *Asian Perspectives*, 29(2), 135-154.

WILLIAMS AN. 2013. A new population curve for prehistoric Australia. *Proceedings: Biological Sciences*, 280(1761), 20130486.

WILLIS JR, IRAOLA-GUZMÁN S, SAUS E, KSIEZOPOLSKA E, COZZUTO L, BEJARANO LA, ANDREU-SOMAVILLA N, ALLOZA-TRABADO M, PUIG-SOLA A, BLANCO A, BROGLIO E, CAROLIS C, HECHT J, PONOMARENKO J AND GABALDÓN T. 2020. Oral microbiome in down syndrome and its implications on oral health. *Journal of Oral Microbiology*, 13(1), 1865690.

WOESE CR AND FOX GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11), 5088-5090.

WU J, PETERS BA, DOMINIANNI C, ZHANG Y, PEI Z, YANG L, MA Y, PURDUE MP, JACOBS EJ, GAPSTUR SM, LI H, ALEKSEYENKO AV, HAYES RB AND AHN J. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. *International Society for Microbial Ecology Journal*, 10(10), 2435-2446.

WURSTER C AND BIRD M. 2014. Barriers and bridges: Early human dispersals in equatorial SE Asia. *Geological Society, London, Special Publications*, 411(1), 235-250.

WYLIE KM, MIHINDUKULASURIYA KA, ZHOU Y, SODERGREN E, STORCH GA AND WEINSTOCK GM. 2014. Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biology*, 12, 71.

XI R, WANG R, WANG Y, XIANG Z, SU Z, CAO Z, XU X, ZHENG X AND LI J. 2019. Comparative analysis of the oral microbiota between iron-deficiency anaemia (IDA) patients and healthy individuals by high-throughput sequencing. *BMC Oral Health*, 19(1), 255.

XU X, HE J, XUE J, WANG Y, LI K, ZHANG K, GUO Q, LIU X, ZHOU Y, CHENG L, LI M, LI Y, LI Y, SHI W AND ZHOU X. 2015. Oral cavity contains distinct niches with dynamic microbial communities. *Environmental Microbiology*, 17(3), 699-710.

YANG B, WANG Y AND QIAN PY. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, 17, 135.

YANG Y, ZHENG W, CAI QY, SHRUBSOLE MJ, PEI Z, BRUCKER R, STEINWANDEL MD, BORDENSTEIN SR, LI Z, BLOT WJ, SHU XO AND LONG J. 2019. Cigarette smoking and oral microbiota in low-income and African-American populations. *Journal of Epidemiology and Community Health*, 73(12), 1108-1115.

YUEN LKW, LITTLEJOHN M, DUCHÊNE S, EDWARDS R, BUKULATJPI S, BINKS P, JACKSON K, DAVIES J, DAVIS JS, TONG SYC AND LOCARNINI S. 2019. Tracing Ancient Human

Migrations into Sahul Using Hepatitis B Virus Genomes. *Molecular Biology and Evolution*, 36(5), 942-954.

YUSSOF A, YOON P, KRKLJES C, SCHWEINBERG S, COTTRELL J, CHU T AND CHANG SL. 2020. A meta-analysis of the effect of binge drinking on the oral microbiome and its relation to Alzheimer's disease. *Scientific Reports*, 10(1), 19872.

ZHANG Y, WANG X, LI H, NI C, DU Z AND YAN F. 2018. Human oral microbiota and its modulation for oral health. *Biomedicine and Pharmacotherapy*, 99, 883-893.

ZIMMERMAN PA, WOOLLEY I, MASINDE GL, MILLER SM, MCNAMARA DT, HAZLETT F, MGONE CS, ALPERS MP, GENTON B, BOATIN BA AND KAZURA JW. 1999. Emergence of FY*A(null) in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24), 13973-13977.

APPENDICES

Appendix A - Supplementary information from Paper I

Papuan mitochondrial genomes and the settlement of Sahul

Nicole Pedro, Nicolas Brucato, Veronica Fernandes, Mathilde André, Lauri Saag,
William Pomat, Céline Besse, Anne Boland, Jean-François Deleuze, Chris Clarkson,
Herawati Sudoyo, Mait Metspalu, Mark Stoneking, Murray P. Cox, Matthew Leavesley,
Luisa Pereira, François-Xavier Ricaut

Table S3. Total number of newly (n=379) and published (n=539) mitogenomes used in this study.

	New Guinea Highlands	New Guinea Coastal	Near Oceania	Remote Oceania	Australia	Sunda	Wallace Sea (Timor)	ND (New Guinea)
This study	296	57	26	0	0	0	0	0
Published	61	18	282	68	69	25	13	3
Total	357	75	308	68	69	25	13	0

Note. Near oceania include the islands settled during Pleistocene: islands from Bismark and Solomon Archipelagos up to Malaita, the southernmost island of the main Solomon islands group.

Table S4. TMRCA age estimates using ML for all M, P and Q Papuan subclades.

Haplogroups nomenclature			Soares et al 2009 mutation rate		Fu et al 2013 mutation rate	
			Maximum Likelihood		Maximum Likelihood	
			Whole mtDNA		Whole mtDNA	
Nagle et al 2017	This study	N	Age	95% CI	Age	95% CI
M27		144	68192	[66427.352601 - 69782.533548]	51451	[40368.221158 - 62533.221962]
M27a		49	21804	[21247.552667 - 22326.376700]	18011	[10451.910773 - 25571.083507]
M27a1		20	13935	[13537.514670 - 14307.796358]	11758	[3904.862623 - 19611.187697]
M27a1a		4	7532	[7377.631937 - 7679.624513]	6478	[1120.382525 - 11835.912715]
M27a1a1		2	2570	[2542.424833 - 2597.288667]	2247	[-468.823169 - 4962.335969]
M27a1a2		2	1880	[1856.551995 - 1903.526053]	1648	[-1508.322463 - 4803.565183]

M27a1b	16	3176	[3092.299975 - 3256.241025]	2771	[-3834.464256 - 9376.463376]
M27a2	22	5405	[5328.336863 - 5479.908234]	4681	[1011.040380 - 8350.444620]
M27a2a	11	1665	[1653.700025 - 1676.971849]	1460	[-301.065358 - 3221.848678]
M27a2a+709A	6	383	[381.134196 - 384.068250]	337	[-617.109091 - 1291.136011]
M27a2b	11	3480	[3444.205019 - 3515.832996]	3033	[390.935614 - 5675.306666]
M27a2b+189C!	8	1065	[1060.019118 - 1070.628393]	936	[-311.550221 - 2183.847221]
M27a3	7	1751	[1731.775838 - 1770.490620]	1535	[-1253.690071 - 4324.257151]
M27b	66	20942	[20332.684129 - 21508.891094]	17337	[8823.761302 - 25851.179138]
M27b1	6	1065	[1057.493372 - 1073.095303]	936	[-898.702560 - 2770.999560]
M27b2	60	8695	[8587.575427 - 8800.057466]	7452	[4149.010152 - 10754.473968]
M27b2a	6	7755	[7653.693312 - 7853.440284]	6665	[3215.857327 - 10114.897313]
M27b2a1	5	3567	[3521.972199 - 3611.627869]	3108	[-121.324846 - 6337.350886]
M27b2a1+11016A	2	2441	[2414.202468 - 2466.368230]	2134	[-581.160989 - 4849.998149]
M27b2b	29	6598	[6507.157267 - 6686.356157]	5692	[2095.474802 - 9288.090958]
M27b2b1	28	1880	[1873.747038 - 1886.855819]	1648	[766.892851 - 2528.349869]
M27b2b1+4216C	2	766	[762.175591 - 770.279819]	674	[-647.065843 - 1995.119683]
M27b2b1+4639C	2	766	[762.402918 - 770.056908]	674	[-573.671801 - 1921.725641]
M27b2c	15	425	[423.673097 - 426.681523]	374	[-506.269109 - 1255.187909]
M27b2+8387A	2	1580	[1562.802932 - 1595.939822]	1385	[-1256.685746 - 4027.685306]
M27c	29	7398	[7290.723578 - 7502.289317]	6366	[2549.319595 - 10182.300005]
M27c+204C!	7	3133	[3056.630847 - 3205.818651]	2734	[-3358.151899 - 8825.259139]
M27c+204C!+150T	6	1451	[1440.164482 - 1461.328448]	1273	[-561.689100 - 3108.013020]

M27c+204C!+150T+151T!	2	425	[423.673097 - 426.681523]	374	[-506.269109 - 1255.187909]
M27c+186T!	15	5361	[5300.350034 - 5420.670661]	4643	[1707.534864 - 7579.058256]
M27c+186T!+204C!	3	468	[459.930388 - 475.369706]	412	[-3698.161034 - 4521.971714]
M27c+186T!+2272T	3	852	[846.614202 - 856.613710]	749	[-718.962048 - 2216.799648]
M27c+151T!	4	5802	[5729.629364 - 5872.238102]	5018	[1788.418094 - 8247.093826]
M27c+151T!+4994C	2	1023	[1015.038634 - 1030.020776]	899	[-936.148500 - 2733.553620]
M28	73	40277	[39125.691873 - 41332.253615]	31904	[22069.139198 - 41738.742562]
M28a	58	23104	[22683.006744 - 23505.537664]	19023	[13517.984340 - 24527.090700]
M28a1	2	15140	[14696.782620 - 15555.223926]	12732	[4584.880894 - 20878.358306]
M28a+204	19	20655	[20238.868435 - 21050.904443]	17113	[11167.877146 - 23057.712014]
M28a2	8	14676	[14359.564903 - 14976.876344]	12357	[6338.848723 - 18375.471677]
M28a2a	5	8336	[8207.121710 - 8461.522293]	7152	[3042.108166 - 11262.240914]
M28a2a+4135C	3	6509	[6412.207497 - 6603.598337]	5617	[1727.006753 - 9506.775247]
M28a2a+4135C+1503A	2	3133	[3081.854694 - 3182.506717]	2734	[-1376.512754 - 6843.619994]
M28a3	7	3611	[3540.486379 - 3678.616940]	3145	[-1771.941881 - 8062.859801]
M28a3+7873T	3	937	[916.326392 - 957.003308]	824	[-4607.348458 - 6254.969818]
M28a4	3	3220	[3178.648840 - 3259.850991]	2808	[-420.892366 - 6037.783366]
M28a5	18	11912	[11665.355205 - 12148.322640]	10110	[4459.062535 - 15761.745065]
M28a5a	12	1923	[1820.769179 - 2016.295477]	1685	[-11158.890120 - 14529.024720]
M28a5a+574C	10	809	[804.401102 - 813.428235]	711	[-683.013946 - 2105.959666]
M28a5b	4	5449	[5371.765363 - 5524.504603]	4718	[1048.486320 - 8387.890560]
M28a5b+3278	2	1622	[1605.286866 - 1639.309578]	1423	[-1219.239806 - 4065.131246]

M28a6	8	13151	[12856.285565 - 13431.153733]	11121	[4956.344618 - 17286.543742]
M28a6a	5	3176	[3136.723882 - 3215.045019]	2771	[-384.944263 - 5926.943383]
M28a6a+9554A	2	1194	[1180.314993 - 1206.853146]	1048	[-1740.487291 - 3837.459931]
M28a7	11	20130	[19707.374486 - 20532.047655]	16701	[10535.789678 - 22865.988802]
M28a7a	4	2786	[2716.552658 - 2852.907362]	2434	[-3804.507504 - 8672.479704]
M28a7a+16086C!	2	1622	[1583.814633 - 1659.425017]	1423	[-4448.577672 - 7294.469112]
M28a7b	6	15792	[15438.745622 - 16128.195182]	13256	[6943.975114 - 19567.750406]
M28a7b+16506C	5	14769	[14426.777595 - 15092.882755]	12432	[5973.376349 - 18890.727811]
M28a7b+16506C+16471A!	2	14305	[13984.106907 - 14610.019931]	12058	[5819.099076 - 18296.086284]
M28b	15	29761	[28895.852979 - 30560.755938]	24115	[14940.930060 - 33289.440660]
M28b1	3	4134	[4051.306498 - 4213.317549]	3595	[-1469.378686 - 8658.999166]
M28b1+2280	2	1023	[1013.821887 - 1031.201212]	899	[-1229.724670 - 3027.129790]
M28b+203A	3	27243	[26202.854418 - 28185.589514]	22205	[10535.789678 - 33875.095162]
M28b+189	9	3567	[3521.972199 - 3611.627869]	3108	[-121.324846 - 6337.350886]
M28b+189+8027A	2	1279	[1268.821546 - 1289.760565]	1123	[-931.654987 - 3178.411387]
M28b+189+10454C	2	1279	[1268.821546 - 1289.760565]	1123	[-931.654987 - 3178.411387]
M29'Q	361	73529	[71510.523281 - 75325.969131]	55046	[42862.120762 - 67228.942838]
M29	15	21325	[20640.138750 - 21957.264656]	17637	[8242.600313 - 27031.475167]
M29a	6	1880	[1866.572339 - 1893.882451]	1648	[-187.229700 - 3482.472420]
M29b	9	10955	[10682.773499 - 11212.517388]	9324	[2645.181202 - 16002.896918]
M29b1	6	2139	[1972.026959 - 2283.294192]	1872	[-16549.607642 - 20294.201642]
Q	346	49463	[47580.146864 - 51123.279584]	38494	[24622.952306 - 52365.900334]

Q1'2	313	49463	[47821.690897 - 50932.471317]	38494	[26311.015282 - 50677.837358]
Q1	275	22622	[22364.427508 - 22871.273222]	18648	[15198.558127 - 22097.598113]
Q1+@16223	139	20416	[20030.174003 - 20784.907000]	16926	[11347.617658 - 22503.512102]
Q1+16223!+5154	6	10637	[10338.836168 - 10917.158295]	9062	[1575.725155 - 16548.109805]
Q1+16223!+5154+5201	5	3090	[3038.255722 - 3139.331064]	2696	[-1487.352737 - 6879.568097]
Q1a	98	18469	[18094.642517 - 18826.298067]	15390	[9518.757948 - 21261.804732]
Q1a1	87	16915	[16558.237574 - 17255.028744]	14155	[8136.253843 - 20172.876797]
preQ1a1a	51	15326	[15063.403086 - 15579.106385]	12881	[8037.396562 - 17725.410158]
preQ1a1a+92A!	2	8695	[8511.796762 - 8870.679365]	7452	[1873.794838 - 13029.689282]
preQ1a1a+7765C	11	8247	[7852.863395 - 8604.345386]	7077	[-5179.522421 - 19334.087741]
preQ1a1a+246	19	2096	[2028.587255 - 2158.586553]	1835	[-6018.311477 - 9688.013597]
preQ1a1a+246+12208	2	0	[0.000000 - 0.000000]	0	[-8733.891046 - 8733.891046]
preQ1a1a+246+92A!	10	1408	[1394.252869 - 1421.375563]	1236	[-1186.287379 - 3657.719419]
Q1a1a	17	13704	[13487.375325 - 13913.002518]	11571	[7167.152916 - 15974.438004]
Q1a1a+4703C	5	9370	[9192.087311 - 9540.908358]	8013	[2949.242234 - 13077.620086]
Q1a1a+4703C+92A!	4	4178	[4101.520309 - 4250.939238]	3632	[-991.568491 - 8256.080851]
Q1a1a+4703C+92A!+7948	2	1408	[1393.834170 - 1421.778785]	1236	[-1259.681422 - 3731.113462]
Q1a1a+208!	6	12508	[12297.524276 - 12710.067544]	10597	[5973.376349 - 15221.025691]
Q1a1a+208!+3714G	3	2225	[2172.816507 - 2274.579603]	1947	[-3850.940470 - 7745.318230]
Q1a1a+208!+3714G+16093C	2	894	[888.814887 - 899.836197]	786	[-754.910150 - 2327.639630]
Q1a1+16223T	34	13658	[13386.668937 - 13917.165901]	11533	[6028.796340 - 17037.902700]
Q1a1+16223T+92A!	4	11228	[11012.157465 - 11434.644726]	9549	[4337.737690 - 14759.691710]

Q1a1+16223T+195	2	6244	[6070.528267 - 6407.426706]	5392	[-1727.006753 - 12511.437473]
Q1a1+16223T+16093C	2	10591	[10285.596119 - 10878.652344]	9024	[1318.097088 - 16730.845992]
Q1a1+16223T+6272G	6	10818	[10612.739522 - 11015.660685]	9212	[4074.118272 - 14349.284208]
Q1a1+16223T+6272+10203	4	809	[804.401102 - 813.428235]	711	[-683.013946 - 2105.959666]
Q1a1+16223T+7061	19	9596	[9362.377388 - 9816.992019]	8201	[1741.985129 - 14659.336591]
Q1a1+16223T+7061+1018	17	170	[168.552979 - 171.265214]	150	[-1831.855385 - 2131.422905]
Q1a+15498	3	4396	[4331.661556 - 4458.667378]	3819	[76.389718 - 7562.582042]
Q1a+16223	8	14908	[14551.229160 - 15245.639968]	12544	[5865.532042 - 19223.247758]
Q1a+16223+12509	5	638	[635.424648 - 641.057405]	562	[-539.221536 - 1662.599736]
Q1b	3	5890	[5781.262130 - 5994.960629]	5093	[322.035084 - 9863.260596]
Q1b+16172C	2	2916	[2876.329278 - 2955.131993]	2546	[-903.196073 - 5995.843913]
Q1c	31	11365	[11137.101281 - 11582.223064]	9661	[4229.893382 - 15092.211658]
Q1c1	19	5714	[5612.812313 - 5810.883573]	4943	[392.433451 - 9493.294709]
Q1c1a	17	1322	[1315.228400 - 1329.133834]	1161	[-160.268623 - 2481.916903]
Q1c1a+7388G	14	681	[679.299877 - 682.503369]	599	[11.982701 - 1186.287379]
Q1c1+16172C	2	1665	[1646.298426 - 1684.115547]	1460	[-1401.975994 - 4322.759314]
Q1c2	9	5890	[5759.010627 - 6015.456125]	5093	[-632.087467 - 10817.383147]
Q1c2a	3	2225	[2159.965858 - 2286.208639]	1947	[-5245.427275 - 9139.805035]
Q1d	10	17291	[16922.698903 - 17641.147083]	14454	[8362.427321 - 20545.838359]
Q1d+238G	7	9776	[9432.898482 - 10095.023456]	8350	[-897.204722 - 17598.093962]
Q1d+238G+92A!	5	1194	[1185.288773 - 1202.049553]	1048	[-712.970698 - 2809.943338]
Q1d+3316A	2	10319	[10038.110766 - 10584.323561]	8800	[1533.785702 - 16065.806098]

Q1e	22	16868	[16583.023297 - 17142.510710]	14117	[9273.112582 - 18961.126178]
Q1e1	19	13889	[13669.559878 - 14100.173381]	11721	[7316.936676 - 16124.221764]
Q1e1a	9	2613	[2592.863143 - 2633.554112]	2284	[302.563195 - 4265.841485]
Q1e1a1	3	1794	[1781.717352 - 1806.757723]	1573	[-188.727538 - 3334.186498]
Q1e1a1+8507T	2	766	[762.402918 - 770.056908]	674	[-573.671801 - 1921.725641]
Q1e1b	7	9100	[8944.422466 - 9249.607699]	7789	[3238.324891 - 12339.186149]
Q1e1b1	4	4746	[4688.040161 - 4803.252054]	4119	[963.109577 - 7274.997223]
Q1e1c	2	3133	[3084.634244 - 3179.893797]	2734	[-1156.330627 - 6623.437867]
Q1f	18	20464	[20165.221740 - 20752.209116]	16963	[12632.762318 - 21293.259322]
preQ1f1	8	15280	[14980.988220 - 15565.507080]	12844	[7339.404240 - 18348.510600]
preQ1f1+92!	2	3133	[3093.866083 - 3171.151188]	2734	[-422.390203 - 5889.497443]
Q1f1	6	3654	[3574.507294 - 3730.928804]	3183	[-2321.648280 - 8687.458080]
Q1f1+16293C	4	1665	[1609.395786 - 1718.012308]	1460	[-6759.741089 - 9680.524409]
Q1f2	10	6376	[6148.978182 - 6588.278834]	5505	[-3596.308078 - 14605.414438]
Q1f2+16189C	9	2786	[2744.049791 - 2827.462111]	2434	[-1382.504105 - 6250.476305]
Q1f2+16189C+16169T	8	1023	[1018.068704 - 1027.057951]	899	[-202.208076 - 1999.613196]
Q1+16172C	8	18469	[17954.365648 - 18951.388143]	15390	[7390.330718 - 23390.231962]
Q1+16172C+8911	2	3220	[3171.065327 - 3267.032453]	2808	[-1008.044705 - 6624.935705]
Q1+16172C+92A!	5	15699	[15065.028343 - 16277.898004]	13181	[2025.076435 - 24336.865325]
Q1+16172C+92A!+208	3	4352	[4294.884341 - 4408.347791]	3782	[405.913990 - 7158.165890]
Q1+15908C	26	19369	[19083.761153 - 19644.945500]	16102	[11771.505698 - 20432.002702]
Q1+15908C+5843!	11	11365	[11162.158590 - 11559.150127]	9661	[4817.045722 - 14505.059318]

Q1+15908C+5843!+15731A	3	4134	[3918.155930 - 4329.197014]	3595	[-9249.147180 - 16438.767660]
Q1+15908C+5843!+11422	6	1279	[1268.441642 - 1290.128496]	1123	[-1005.049030 - 3251.805430]
Q1+15908C+534	15	17620	[17351.622651 - 17879.218385]	14716	[10312.611876 - 19119.896964]
Q1+15908C+534+16319	7	15699	[15422.361902 - 15964.724183]	13181	[8190.175997 - 18171.765763]
Q1+15908C+534+16319+7954	4	11319	[11082.946163 - 11544.461078]	9624	[3972.265315 - 15274.947845]
Q1+15908C+534+16319+7954+5321	3	3654	[3603.649650 - 3703.755014]	3183	[-340.009135 - 6705.818935]
Q1+15908C+534+16319+5984	3	9190	[8996.879871 - 9374.363527]	7864	[2285.700178 - 13441.594622]
Q1+15908C+534+234	8	12187	[11958.217112 - 12406.214790]	10335	[5197.496472 - 15472.662408]
Q1+15908C+534+234+15172	5	5934	[5838.357114 - 6027.073999]	5130	[946.633363 - 9313.554197]
Q1+15908C+534+234+13444	2	3437	[3381.000722 - 3491.080994]	2996	[-1114.391174 - 7105.741574]
Q1+14305	4	17667	[17250.572115 - 18061.460680]	14754	[8001.448459 - 21505.952261]
Q1+16093	7	19227	[18598.223306 - 19808.754245]	15989	[6594.978953 - 25383.853807]
Q1+16093+92!	5	12278	[11909.103918 - 12624.327384]	10410	[2263.232614 - 18556.710026]
Q1+16093+92!+4356	4	8471	[8196.868387 - 8727.125888]	7265	[-1175.802516 - 15704.827236]
Q1+16093+92!+4356+4345	3	1537	[1460.735771 - 1605.903323]	1348	[-10541.781029 - 13237.888709]
Q1+7456	6	17714	[17254.646760 - 18146.885049]	14791	[7378.348018 - 22203.944582]
Q1+7456+7349	3	4878	[4776.861909 - 4974.872222]	4231	[-1052.979833 - 9515.762273]
Q1+7456+7349+16173	2	1537	[1520.788762 - 1552.143442]	1348	[-1220.737644 - 3916.845324]
Q1+7456+143	3	9686	[9323.004132 - 10021.296390]	8276	[-1559.248942 - 18110.354422]
Q1+1462	8	3567	[3524.059724 - 3609.639942]	3108	[25.463239 - 6190.562801]
Q1+1462+9214	4	1494	[1476.093995 - 1510.942673]	1311	[-1625.153796 - 4246.369596]
Q1+1943	6	13519	[13132.389706 - 13882.638498]	11421	[3567.849163 - 19274.174237]

Q1+1943+92!	3	5581	[5490.852279 - 5668.966303]	4831	[647.065843 - 9013.986677]
Q1+208C	4	15839	[14753.305566 - 16773.239571]	13293	[-5128.595942 - 31715.213342]
Q1+208C+92!	3	13013	[12414.672530 - 13554.345381]	11009	[-1321.092763 - 23339.305483]
Q1+4500	12	19132	[18795.595614 - 19454.711964]	15915	[10776.941532 - 21052.107468]
Q1+4500+195	2	5273	[5162.867147 - 5379.048926]	4568	[-789.360415 - 9926.169775]
Q1+4500+92!	9	9505	[9263.343162 - 9734.568765]	8126	[1373.517079 - 14878.020881]
Q1+4500+92!+6512	8	1108	[1102.587011 - 1113.617497]	974	[-274.104281 - 2221.293161]
Q2	38	37765	[36619.745230 - 38812.270974]	30069	[19867.317926 - 40270.861714]
Q2a	37	19132	[18839.663035 - 19414.017884]	15915	[11437.487914 - 20391.561086]
Q2a1	3	8292	[8137.207188 - 8440.082733]	7115	[2197.327759 - 12032.129441]
Q2a1+8999C	2	2786	[2748.995788 - 2822.783070]	2434	[-942.139850 - 5810.112050]
Q2a2	4	14676	[14398.946314 - 14940.956720]	12357	[7072.789147 - 17641.531253]
Q2a2a	2	8740	[8550.875736 - 8920.945603]	7489	[1764.452693 - 13213.923307]
Q2a2b	2	3654	[3593.983904 - 3712.860513]	3183	[-1000.555517 - 7366.365317]
Q2a3	13	15979	[15617.631677 - 16322.118837]	13406	[7020.364831 - 19790.928209]
preQ2a3a	11	8516	[8313.942651 - 8707.733974]	7302	[1063.464696 - 13540.451904]
Q2a3a	10	638	[635.424648 - 641.057405]	562	[-539.221536 - 1662.599736]
Q2a3b	2	5405	[5139.886479 - 5646.324130]	4681	[-7576.062581 - 16937.547581]
Q2a4	4	17244	[16944.282523 - 17531.513218]	14417	[9425.892017 - 19407.481783]
Q2a4+143A	3	14815	[14436.214994 - 15172.459887]	12469	[5350.275907 - 19588.720133]
Q2a+474	6	1322	[1256.780936 - 1381.965076]	1161	[-10729.010729 - 13050.659009]
Q2a+7220	2	1880	[1856.551995 - 1903.526053]	1648	[-1508.322463 - 4803.565183]

Q3	33	35173	[34317.947049 - 35969.754264]	28159	[20086.002216 - 36232.691544]
Q3a	27	28129	[27368.426699 - 28836.974135]	22879	[14439.154464 - 31319.784216]
Q3a+6l	26	16026	[15748.386325 - 16292.231007]	13443	[8525.691619 - 18360.493301]
Q3a1	9	12095	[11824.983859 - 12352.671819]	10260	[4168.482041 - 16351.893079]
Q3a1+453C	8	8202	[8009.461053 - 8385.428148]	7040	[874.737158 - 13204.936282]
Q3a1+453+5177	4	3829	[3732.493476 - 3920.084090]	3333	[-2979.198986 - 9644.576306]
Q3a1+453+228	4	2484	[2454.732275 - 2512.100702]	2172	[-763.897176 - 5107.626216]
Q3a+6l+8733	3	3176	[3102.800138 - 3246.700540]	2771	[-3027.129790 - 8569.128910]
Q3a+6l+8733+16150	2	1194	[1169.889383 - 1216.681888]	1048	[-3868.914521 - 5965.887161]
Q3a+6l+8489	4	1023	[944.099657 - 1091.014720]	899	[-17082.837828 - 18880.242948]
Q3a+6l+8489+143!	2	1751	[1720.258989 - 1781.389098]	1535	[-2868.359004 - 5938.926084]
Q3a+6l+189	4	10909	[10708.139734 - 11102.492282]	9287	[4295.798237 - 14277.388003]
Q3a+6l+189+8489	3	2527	[2497.435799 - 2555.776507]	2209	[-726.451236 - 5145.072156]
pre-Q3b	3	19797	[18994.819875 - 20525.708934]	16439	[4842.508961 - 28035.026359]
Q3b*	2	8202	[8014.153653 - 8381.167032]	7040	[1021.525243 - 13058.148197]
Q3+215	2	26263	[25554.005437 - 26923.730084]	21457	[13162.996829 - 29750.050411]
P	308	68746	[67521.038663 - 69883.583079]	51825	[44045.412466 - 59604.949454]
P+16176	170	64392	[62663.515972 - 65949.973568]	48867	[37711.057255 - 60022.846145]
P1	122	37918	[36838.631026 - 38910.466469]	30181	[20566.808086 - 39796.047194]
P1d	89	29612	[29144.696233 - 30059.983125]	24003	[18938.658614 - 29067.036466]
P1d1	47	22236	[21867.182456 - 22590.575635]	18349	[13357.715717 - 23339.305483]
P1d1a	4	18422	[18057.741306 - 18769.584018]	15353	[9628.100093 - 21077.570707]

P1d1a+1462	2	4571	[4489.569502 - 4649.839582]	3969	[-581.160989 - 8519.700269]
P1d1+13071	24	11593	[11370.672671 - 11805.367760]	9848	[4637.305210 - 15059.259230]
P1d1+13071+212!	3	4790	[4635.400067 - 4935.468505]	4156	[-3990.239366 - 12303.238046]
P1d1+13071+212!+2483	2	2527	[2496.689812 - 2556.489097]	2209	[-799.845278 - 5218.466198]
P1d1+13071+15667	7	1322	[1277.999436 - 1363.760887]	1161	[-6985.914566 - 9307.562846]
P1d1+13071+14233G	6	9821	[9621.905000 - 10012.323246]	8388	[2956.731422 - 13819.049698]
P1d1+13071+16093	7	7264	[7031.043839 - 7483.154717]	6253	[-2040.054811 - 14546.998771]
P1d1+13071+16093+212!	2	2527	[2495.943411 - 2557.201278]	2209	[-873.239321 - 5291.860241]
P1d1+13071+16093+16111	5	809	[756.899219 - 855.290248]	711	[-14481.093917 - 15904.039637]
P1d1+6366A	7	20846	[20500.761684 - 21177.441188]	17263	[12345.177499 - 22179.979181]
P1d1+6366A+3699T	6	19512	[19189.721659 - 19821.293302]	16214	[11370.085222 - 21058.098818]
P1d1+6366A+3699T+8269	5	13243	[13018.445522 - 13459.094424]	11196	[6499.117346 - 15893.554774]
P1d1+6366A+3699T+8269+480	4	9055	[8910.269685 - 9194.475057]	7751	[3494.455121 - 12008.164039]
P1d1+6366A+3699T+8269+480+4994	2	1065	[1056.226286 - 1074.324568]	936	[-1192.278730 - 3064.575730]
P1d1+6366A+3699T+8269+480+8251	2	5978	[5890.448432 - 6063.815544]	5168	[1351.049515 - 8984.029925]
P1d1+5460	6	4484	[4208.144650 - 4728.647584]	3894	[-11151.400932 - 18940.156452]
P1d1+5460+16189	5	2830	[2794.218387 - 2864.234149]	2471	[-684.511783 - 5627.375863]
P1d1+5460+16189+16093C	2	1065	[1055.909076 - 1074.631448]	936	[-1265.672772 - 3137.969772]
P1d2	9	23249	[22767.945260 - 23704.812035]	19135	[12896.381736 - 25373.368944]
P1d2a	6	13612	[13345.009530 - 13866.890384]	11496	[6064.744442 - 16927.062718]
P1d2a+16176!	2	10501	[10244.605572 - 10743.388129]	8950	[2417.509886 - 15481.649434]
P1d2+212C!	2	20989	[20313.828545 - 21614.215710]	17375	[7980.478733 - 26769.353587]

P1d+16311C	P1d3	19	27637	[27073.727973 - 28169.824508]	22505	[16119.728251 - 28890.291629]
P1d+16311C+5987		15	1665	[1656.643418 - 1674.097236]	1460	[139.298897 - 2781.484423]
P1d+16311C+5987+150T		2	1023	[1017.766455 - 1027.354988]	899	[-275.602118 - 2073.007238]
P1d+16311C+5987+16266T!		2	724	[720.152314 - 726.957803]	637	[-537.723698 - 1810.885658]
P1d+16311+212C!		2	2527	[2485.450272 - 2567.128793]	2209	[-1900.755914 - 6319.376834]
P1d+16311C+152C		2	2096	[2074.697632 - 2116.000840]	1835	[-660.546382 - 4330.248502]
P1d+93	P1d4	5	18422	[17898.456511 - 18911.627187]	15353	[7206.096694 - 23499.574106]
P1d+93+212C!		4	5714	[5554.340657 - 5864.252428]	4943	[-2176.358033 - 12062.086193]
P1d+93+212C!+16189		3	1665	[1368.919037 - 1886.147432]	1460	[-37585.238897 - 40506.022217]
P1d+13708A		5	18943	[18545.221938 - 19320.391127]	15765	[9673.035221 - 21856.446259]
P1d+13708A+783G		2	9100	[8890.541028 - 9299.118096]	7789	[1697.050001 - 13880.461039]
P1+152	pre-P1eP1f	17	32203	[30891.681624 - 33376.061244]	25950	[13032.684958 - 38867.387882]
P1f		3	4047	[3972.846082 - 4117.770134]	10635	[4322.759314 - 16946.534606]
P1+152+13722G	P1e	9	25043	[23951.492344 - 26020.887685]	20520	[7529.629615 - 33511.120625]
P1+152+13722G+5330A	P1e1	4	11091	[10831.716962 - 11337.921119]	9436	[3124.489234 - 15748.264526]
P1+152+13722G+5330A+15927A		2	4047	[3972.846082 - 4117.770134]	3520	[-1103.906311 - 8143.743031]
P1+152+13722G+5330A+		2	6376	[6259.193568 - 6489.431842]	5505	[733.940424 - 10275.165936]
P1+125C	P1g	14	32704	[31746.415564 - 33586.714802]	26324	[16856.664350 - 35792.327290]
P1+125C+152		2	2527	[2493.701722 - 2559.335359]	2209	[-1093.421448 - 5512.042368]
P1+125C+127C		12	25092	[24413.500215 - 25724.165203]	20558	[12337.688311 - 28777.953809]
P1+125C+127+2359T		11	2657	[2631.873442 - 2680.876403]	2322	[-26.961077 - 4670.257637]
P1+125C+127+2359T+8865A		2	2657	[2609.875297 - 2701.758764]	2322	[-2081.994264 - 6725.290824]

P1+125C+127+2359T+150		2	1023	[1015.342399 - 1029.725248]	899	[-862.754458 - 2660.159578]
P1+1719	P1h	2	10319	[10032.267985 - 10589.519662]	8800	[1386.997618 - 16212.594182]
P2'10		43	61277	[59555.512833 - 62824.917172]	46733	[35356.456548 - 58108.609692]
P2		39	22959	[22402.280618 - 23482.760811]	18910	[11644.189502 - 26176.209898]
P2+143A		34	19654	[19214.374043 - 20071.223057]	16326	[9794.360066 - 22858.499614]
P2+143A+185		13	4571	[4502.916329 - 4637.334233]	3969	[152.779435 - 7785.759845]
P2+143A+185+228		7	2786	[2758.838488 - 2813.376263]	2434	[-61.411342 - 4929.383542]
P2+143A+185+228+9438		6	1708	[1696.863391 - 1719.730825]	1498	[-190.225375 - 3185.900575]
P2+143A+185+2483		4	766	[717.185487 - 809.973718]	674	[-14445.145814 - 15793.199654]
P2+143A+16278		21	16400	[16060.725658 - 16722.967439]	13743	[7871.136588 - 19614.183372]
P2+143A+16278+5046		7	2096	[2082.101681 - 2108.827074]	1835	[220.182127 - 3449.519993]
P2+143A+16278+103A		6	4790	[4723.042763 - 4855.486668]	4156	[560.191262 - 7752.807418]
P2+143A+16278+103A+195C		3	2570	[2539.393698 - 2600.188989]	2247	[-762.399338 - 5255.912138]
P2+143A+16278+146		4	638	[604.190830 - 669.174907]	562	[-12135.480235 - 13258.858435]
P2+143A+16278+7521A		3	11137	[10709.278075 - 11530.639986]	9474	[-727.949074 - 19675.594714]
P2+152C		3	2311	[2281.993850 - 2339.478149]	2022	[-1133.863063 - 5178.024583]
P2+16256		2	20464	[19940.708878 - 20955.694060]	16963	[9476.818495 - 24449.203145]
P10		4	4178	[4111.327096 - 4241.770501]	3632	[-404.416152 - 7668.928512]
P+16176+152C	pre-P8	5	62586	[60379.315765 - 64518.368633]	47631	[33392.791454 - 61869.679906]
P8		2	17009	[16510.249327 - 17475.093232]	14229	[5935.930409 - 22522.983991]
P+16176+152C+153G		3	59971	[57441.780739 - 62141.833348]	45834	[29320.171020 - 62347.490100]
P+16176+152C+153G+571T		2	15932	[15482.021629 - 16354.467103]	13368	[5441.644001 - 21294.757159]

P3		23	52596	[51374.629123 - 53723.231547]	40704	[31823.057650 - 49584.415910]
P3a		4	28771	[27983.099369 - 29502.667532]	23366	[14779.163599 - 31953.369521]
P3b		19	45780	[44693.883490 - 46784.598259]	35873	[27286.107559 - 44460.313481]
P3b+152C	P3b2	5	42187	[40748.670796 - 43482.135258]	33289	[21472.999834 - 45105.881486]
P3b1		3	26410	[25665.131281 - 27101.634993]	21569	[12908.364437 - 30229.358443]
P3b1+125C		2	24459	[23757.877107 - 25111.665378]	20071	[11410.526837 - 28731.520843]
P3b+152C+125C		2	42187	[40242.304624 - 43878.639932]	33289	[17583.115586 - 48995.765734]
P3b+15454C	P3b1	14	40792	[39613.010893 - 41871.350178]	32278	[22296.810514 - 42259.990046]
P3b+15454C+127!		4	5273	[5147.395750 - 5393.197120]	4568	[-1523.300839 - 10660.110199]
P3b+15454C+127!+251		2	2657	[2600.342920 - 2710.605225]	2322	[-2962.722773 - 7606.019333]
P3b+15454C+127!+150		2	2397	[2357.977890 - 2435.570346]	2097	[-2013.093734 - 6207.039014]
P3b+15454+942		10	4265	[4149.457868 - 4374.460123]	3707	[-3118.497883 - 10532.794003]
P3b+15454+942+195		7	1408	[1021.630548 - 1661.261315]	1236	[-55644.666840 - 58116.098880]
P4	P11	42	66426	[65174.169983 - 67586.993713]	50252	[42179.106816 - 58325.796144]
P4a	P4	35	25823	[25307.462630 - 26312.358458]	21120	[14954.410598 - 27284.609722]
P4a1+146C		26	18611	[18306.192842 - 18904.412801]	15503	[10732.006404 - 20273.231916]
P4a1+35A		12	14815	[14559.619959 - 15060.511760]	12469	[7625.491222 - 17313.504818]
P4a1+35A+64		2	2311	[2266.145729 - 2354.380091]	2022	[-2821.926038 - 6866.087558]
P4a1+35A+153		3	852	[846.614202 - 856.613710]	749	[-718.962048 - 2216.799648]
P4a1+178		3	937	[926.531802 - 947.419053]	824	[-1965.162931 - 3612.784291]
P4a1+5426C		10	1365	[1355.844558 - 1374.180559]	1198	[-489.792895 - 2886.333055]
P4a1+146+507C!		3	596	[577.406853 - 612.819339]	524	[-6888.555122 - 7937.041442]

P4a+185		4	2052	[2023.521899 - 2080.659514]	1797	[-1725.508915 - 5320.319155]
P4a+507C!		4	23781	[22300.199702 - 25053.481573]	19547	[1565.240292 - 37528.321068]
P4a+146		2	1537	[1520.331881 - 1552.582434]	1348	[-1294.131686 - 3990.239366]
P4b	P11	7	49198	[47748.588215 - 50512.514858]	38307	[27444.878345 - 49169.514895]
P4b1		6	20035	[19536.281621 - 20504.533780]	16626	[9359.987162 - 23892.007558]
P4b1+12007A		5	16775	[16393.429337 - 17135.973367]	14042	[7583.551769 - 20500.903231]
P4b1+12007A+3421A		2	11638	[11341.426364 - 11918.997312]	9886	[2986.688174 - 16784.768146]
P4b1+12007A+16258C		3	12324	[12076.496691 - 12561.279963]	10447	[4942.864080 - 15951.970440]
P4b1+12007A+16258C+16263C		2	8247	[8081.486394 - 8405.351483]	7077	[1792.911607 - 12361.653713]
P5		21	23781	[22195.987901 - 25129.770689]	19547	[390.935614 - 38702.625746]
P5+16192T!		20	23781	[22607.232567 - 24819.712972]	19547	[5088.154327 - 34005.407033]
P5+16192T!+7419A!		19	23781	[22979.278880 - 24516.990811]	19547	[9491.796871 - 29601.764489]
P5+16192T!+7419A!+1438G		18	18375	[17783.960701 - 18922.584603]	15315	[6141.134160 - 24489.644760]
P5+16192T!+7419A!+1438G+125C		15	14722	[14159.235627 - 15239.038786]	12395	[1899.258077 - 22889.954203]
P5+16192T!+7419A!+1438G+125C+16311C!		14	14722	[14507.295102 - 14929.987456]	12395	[8284.539766 - 16504.672514]
P5+16192T!+7419A!+1438G+125C+16311C!+203A		10	13381	[13204.877018 - 13552.259905]	11309	[7638.971760 - 14978.376000]
P5+16192T!+7419A!+1438G+125C+16311C!+203A+151T		2	10183	[10022.471683 - 10338.634476]	8687	[4430.603621 - 12944.312539]
P5+16192T!+7419A!+1438G+125C+16311C!+203A+16145A		6	9280	[9129.515810 - 9425.164678]	7939	[3608.290778 - 12268.787782]
P5+16192T!+7419A!+1438G+125C+16311C!+203A+16145A+9802C		2	4878	[4799.670573 - 4953.674056]	4231	[121.324846 - 8341.457594]
P5+16192T!+7419A!+1438G+125C+16311C!+203A+16145A+204C		2	4047	[3982.357649 - 4108.877020]	3520	[-516.753972 - 7556.590692]
P5+16192T!+7419A!+1438G+125C+16311C!+16274A		2	12508	[12294.126469 - 12713.219169]	10597	[5899.982306 - 15294.419734]
P5+16192T!+7419A!+1438G+5435T		2	5846	[5714.088640 - 5971.988296]	5055	[-742.927450 - 10853.331250]

P6			3	65711	[64358.986723 - 66956.055773]	49766	[41031.763214 - 58499.545306]
P6+271T			2	14259	[13841.543651 - 14649.367639]	12020	[3946.802076 - 20093.491404]
P9			11	55756	[53974.970127 - 57346.353015]	42913	[30582.848117 - 55243.246363]
P9a			8	49886	[48142.088014 - 51438.577865]	38794	[25950.036420 - 51637.951260]
P9a+152C+146			6	2916	[2875.468347 - 2955.947811]	2546	[-976.590115 - 6069.237955]
P9a+152C+146C+2831A			5	553	[550.698013 - 555.256890]	487	[-540.719374 - 1514.313814]
P9a+6392C			2	3785	[3688.846086 - 3876.540592]	3295	[-3090.038969 - 9680.524409]
P+6848T	P14		3	39454	[38194.702587 - 40599.101122]	31305	[20442.487565 - 42167.124115]
P+6848T+195C			2	1966	[1946.203589 - 1986.152957]	1723	[-846.278244 - 4291.304724]
P+153G	P13		33	64392	[62104.733222 - 66389.519274]	48867	[34334.931305 - 63398.972095]
P+153G+5117T	P12	P13a	10	52010	[50309.026909 - 53531.788509]	40292	[28035.026359 - 52548.636521]
P+153G+5117T+709A	P12b	P13a1	9	15047	[14659.295516 - 15413.152614]	12657	[5464.111565 - 19849.343875]
P+153G+5117T+709A+16258C			8	10047	[9871.874521 - 10216.546400]	8575	[3877.901546 - 13272.338974]
P+153G+5117T+709A+16258C+295T			2	3133	[3095.706301 - 3169.396592]	2734	[-275.602118 - 5742.709358]
P+153G+5117T+709A+16258C+8704C			3	6288	[6183.047938 - 6389.308877]	5430	[1099.412798 - 9759.909802]
P+153G+5117T+4829G+709A+16258C+8704C+533G			2	3046	[3009.084355 - 3082.546385]	2659	[-423.888041 - 5741.211521]
P+153G+1888	P13b		21	24362	[23619.889089 - 25049.611684]	19996	[10821.876660 - 29170.387260]
P+153G+1888+6755	P13b1		19	20083	[19488.876076 - 20635.515917]	16663	[8076.340339 - 25250.546261]
P+153G+1888+6755+4502C	P13b1a		12	1408	[1380.739397 - 1434.164746]	1236	[-3534.896736 - 6006.328776]
P+153G+1888+6755+4502C+5			11	212	[208.602053 - 216.134109]	187	[-4216.412844 - 4590.872244]
P+153G+1888+6755+195			7	5405	[5300.127735 - 5506.271318]	4681	[-310.052383 - 9671.537383]
P+153G+1888+6755+195+16284			2	1194	[1183.871627 - 1203.425915]	1048	[-1006.546867 - 3103.519507]

P+153G+195C	P13b2	2	4659	[4560.631875 - 4752.789463]	4044	[-1313.603575 - 9401.926615]
P+153G+152C		2	1665	[1646.298426 - 1684.115547]	1460	[-1401.975994 - 4322.759314]

To note that the timing of the modern human arrival in northern Sahul based on the mitogenomes analysed in this study was estimated from the TMRCA of the main Papuan haplogroups using two mutation rate (Fu et al and Soares et al). While the confidence interval from both mutation rate tend to overlap, we favor Fu et al age estimates, because (1) it is based on directly-dated fossils that are within the time range of modern human evolution (and which span the last 40k) and (2) provide results comparable to recent study on southern Sahul mitogenomes (Nagle et al 2017; Tobler et al 2017). However, older age estimates provided by Soares et al mutation rate cannot be ruled out.

Table S5. Region frequencies of haplogroups M27, M28, M29, Q and P.

Regions	Haplogroups						Total
	M27	M28	M29	P	Q	Others	
Remote Oceania	1	51	2	50	80	548	732
Near Oceania	264	281	47	75	407	1211	2285
Coastal New Guinea	0	0	0	79	85	107	271
Highland New Guinea	0	1	2	277	300	54	634
Australia	0	0	0	106	4	190	300
Wallace Sea (Timor)	0	0	0	14	47	263	324
Sunda	0	0	0	19	8	4117	4144
Total	265	333	51	620	931	6490	8690

Remote Oceania: Temoto Province (Solomon Islands), Reef Islands, Vanuatu, New Caledonia, Fiji, and Polynesia. Near Oceania: Bismarck Archipelago, and the Solomon Islands (as far east as the Makira-Ulawa Province). Coastal New Guinea: National Capital District, Central province, East Sepik, Madang (Bogia district), Milne Bay, Morobe, Oro, Gulf, Bougainville and the Milne Bay. Highlands New Guinea: Jiwaka, Chimbu, Enga, Hela, Madang (Middle district), Western and Southern Highlands and Indonesian Papua province. Australia. Sunda: Philippines and Madagascar. Wallace Sea: East Timor.

n	References
379	This study
260	Delfin F, Min-Shan Ko A, Li M, Gunnarsdóttir ED, Tabbada KA, Salvador JM, Calacal GC, Sagum MS, Datar FA, Padilla SG, De Ungria MC, Stoneking M. Complete mtDNA genomes of Filipino ethnolinguistic groups: melting pot of recent and ancient lineages in the Asia-Pacific region. <i>Eur J Hum Genet.</i> 2014 Feb;22(2):228-37
795	Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, Kayser M, Stoneking M. Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. <i>Am J Hum Genet.</i> 2014 May 1;94(5):721-33.
878	Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G, Zhadanov S, Schurr TG, Merriwether DA. Melanesian mtDNA complexity. <i>PLoS One.</i> 2007 Feb 28;2(2):e248.
1743	Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, Merriwether DA, Weber JL. The genetic structure of Pacific Islanders. <i>PLoS Genet.</i> 2008 Jan;4(1):e19.
324	Gomes SM, Bodner M, Souto L, Zimmermann B, Huber G, Strobl C, Röck AW, Achilli A, Olivieri A, Torroni A, Côrte-Real F, Parson W. Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocene mtDNA diversity. <i>BMC Genomics.</i> 2015 Feb 14;16:70.
109	Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. <i>Genome Res.</i> 2011 Jan;21(1):1-11.
80	Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, Forster .Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. <i>Proc Natl Acad Sci U S A.</i> 2007 May 22;104(21):8726-30.
52	Ingman M, Gyllensten U. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. <i>Genome Res.</i> 2003 Jul;13(7):1600-6.
31	Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, Mondal M, Pagani L, Ricaut FX, Stoneking M, Metspalu M, Sudoyo H, Lansing JS, Cox MP. Multiple Deeply Divergent Denisovan Ancestries in Papuans. <i>Cell.</i> 2019 May 2;177(4):1010-1021.e32.
550	Ko AM, Chen CY, Fu Q, Delfin F, Li M, Chiu HL, Stoneking M, Ko YC. Early Austronesians: into and out of Taiwan. <i>Am J Hum Genet.</i> 2014 Mar 6;94(3):426-36.
108	Malaspina AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, Heupink TH, Macholdt E, Peischl S, Rasmussen S, Schiffels S, Subramanian S, Wright JL, Albrechtsen A, Barbieri C, Dupanloup I, Eriksson A, Margaryan A, Moltke I, Pugach I, Korneliussen TS, Levkivskyi IP, Moreno-Mayar JV, Ni S, Racimo F, Sikora M, Xue Y, Aghakhanian FA, Brucato N, Brunak S, Campos PF, Clark W, Ellingvåg S, Fourmile G, Gerbault P, Injie D, Koki G,

	Leavesley M, Logan B, Lynch A, Matisoo-Smith EA, McAllister PJ, Mentzer AJ, Metspalu M, Migliano AB, Murgha L, Phipps ME, Pomat W, Reynolds D, Ricaut FX, Siba P, Thomas MG, Wales T, Wall CM, Oppenheimer SJ, Tyler-Smith C, Durbin R, Dortch J, Manica A, Schierup MH, Foley RA, Lahr MM, Bowern C, Wall JD, Mailund T, Stoneking M, Nielsen R, Sandhu MS, Excoffier L, Lambert DM, Willerslev E. A genomic history of Aboriginal Australia. <i>Nature</i> . 2016 Oct 13;538(7624):207-214.
127	Nagle N, van Oven M, Wilcox S, van Holst Pellekaan S, Tyler-Smith C, Xue Y, Ballantyne KN, Wilcox L, Papac L, Cooke K, van Oorschot RA, McAllister P, Williams L, Kayser M, Mitchell RJ; Genographic Consortium. Aboriginal Australian mitochondrial genome variation - an increased understanding of population antiquity and diversity. <i>Sci Rep</i> . 2017 Mar 13;7:43041.
2790	Pierron D, Heiske M, Razafindrazaka H, Rakoto I, Rabetokotany N, Ravololomanga B, Rakotozafy LM, Rakotomalala MM, Razafiarivony M, Rasoarifetra B, Raharijesy MA, Razafindralambo L, Ramilisonina, Fanony F, Lejambale S, Thomas O, Mohamed Abdallah A, Rocher C, Arachiche A, Tonaso L, Pereda-Loth V, Schiavinato S, Brucato N, Ricaut FX, Kusuma P, Sudoyo H, Ni S, Boland A, Deleuze JF, Beaujard P, Grange P, Adelaar S, Stoneking M, Rakotoarisoa JA, Radimilahy C, Letellier T. Genomic landscape of human diversity across Madagascar. <i>Proc Natl Acad Sci U S A</i> . 2017 Aug 8;114(32):E6498-E6506.
423	Tabbada KA, Trejaut J, Loo JH, Chen YM, Lin M, Mirazón-Lahr M, Kivisild T, De Ungria MC. Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? <i>Mol Biol Evol</i> . 2010 Jan;27(1):21-31.
41	van Holst Pellekaan SM, Ingman M, Roberts-Thomson J, Harding RM. Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. <i>Am J Phys Anthropol</i> . 2006 Oct;131(2):282-94.
8690	

Appendix B - Supplementary information from Paper II

First insight into oral microbiome diversity in Papua New Guineans reveals a specific regional signature

Nicole Pedro, Nicolas Brucato, Bruno Cavadas, Valentine Lissant, Rui Camacho, Christopher Kinipi, Matthew Leavesley, Luisa Pereira, François-Xavier Ricaut

The MiDAS output files obtained from the Array methodology and the KRAKEN output files obtained from the WGS unmapped reads have been made available as Supporting Information files at <https://cloud.i3s.up.pt/index.php/s/64z3Mx2EA6BDgWk>.

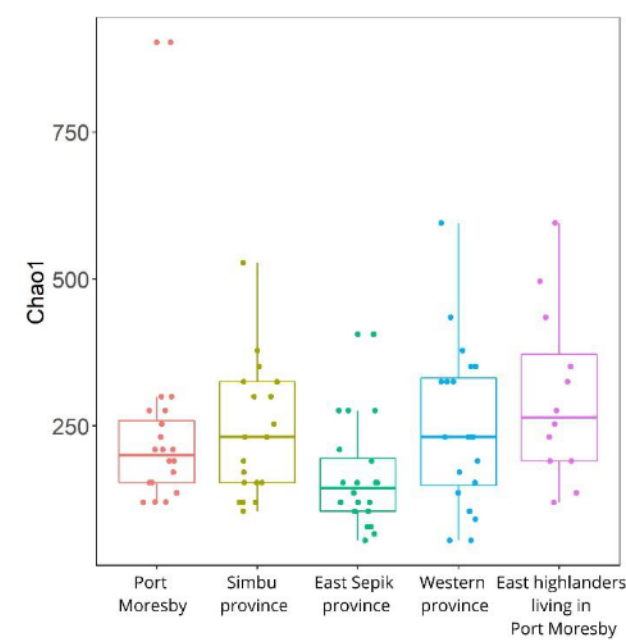


Figure S1. Alpha-diversity plot for all samples with Chao1 index. Sample PNG54 from Port Moresby (in red) was removed as an outlier for Alpha diversity (Chao1 = 903). Sample is considered an outlier if value falls out of the mean +/- 3SD (637) for all samples.

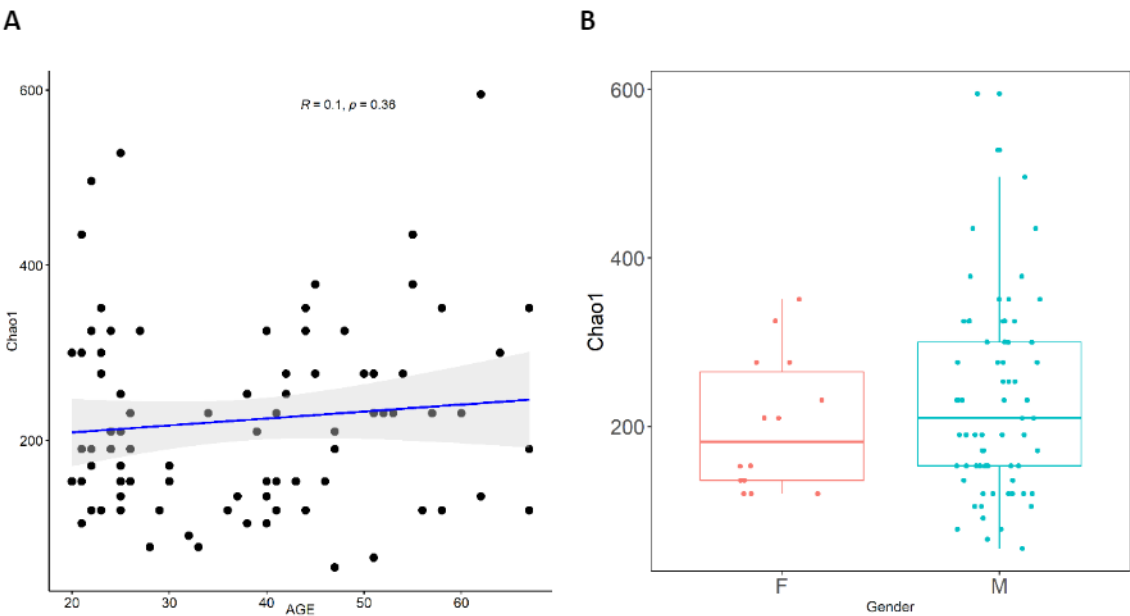


Figure S2. Correlation between alpha-diversity and (A) Age and (B) Gender of the individuals (F=Female and M=Male). A Pearson correlation was performed for the Age correlation, obtaining a p-value of 0.36, and a t-test was performed in the alpha-diversity between Female and Male (p-value = 0.4).

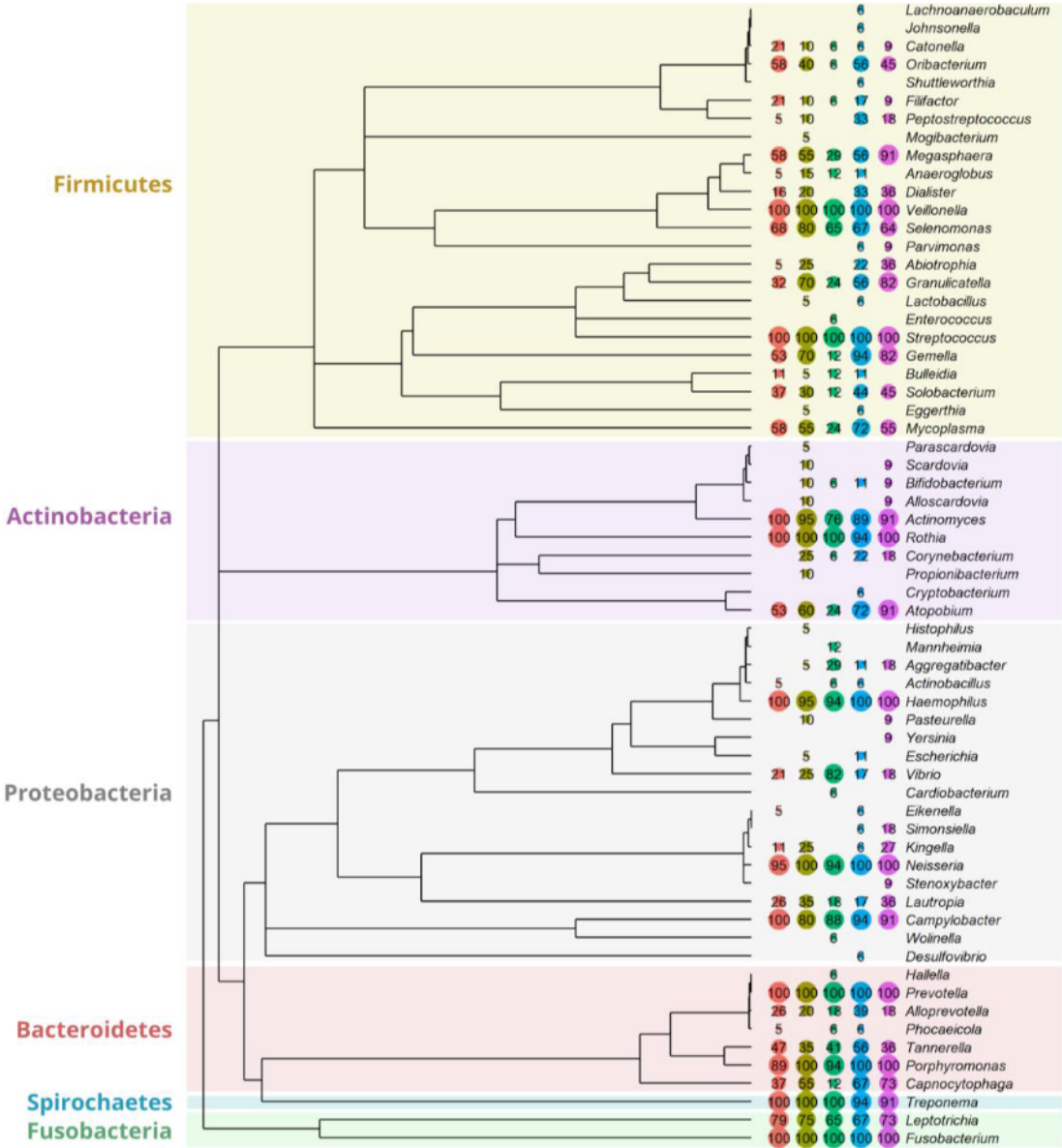


Figure S3. Profile of microbial composition of samples with phylogenetic tree of all genera found in the oral microbiome of the different groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and East highlanders living in Port Moresby (in pink).

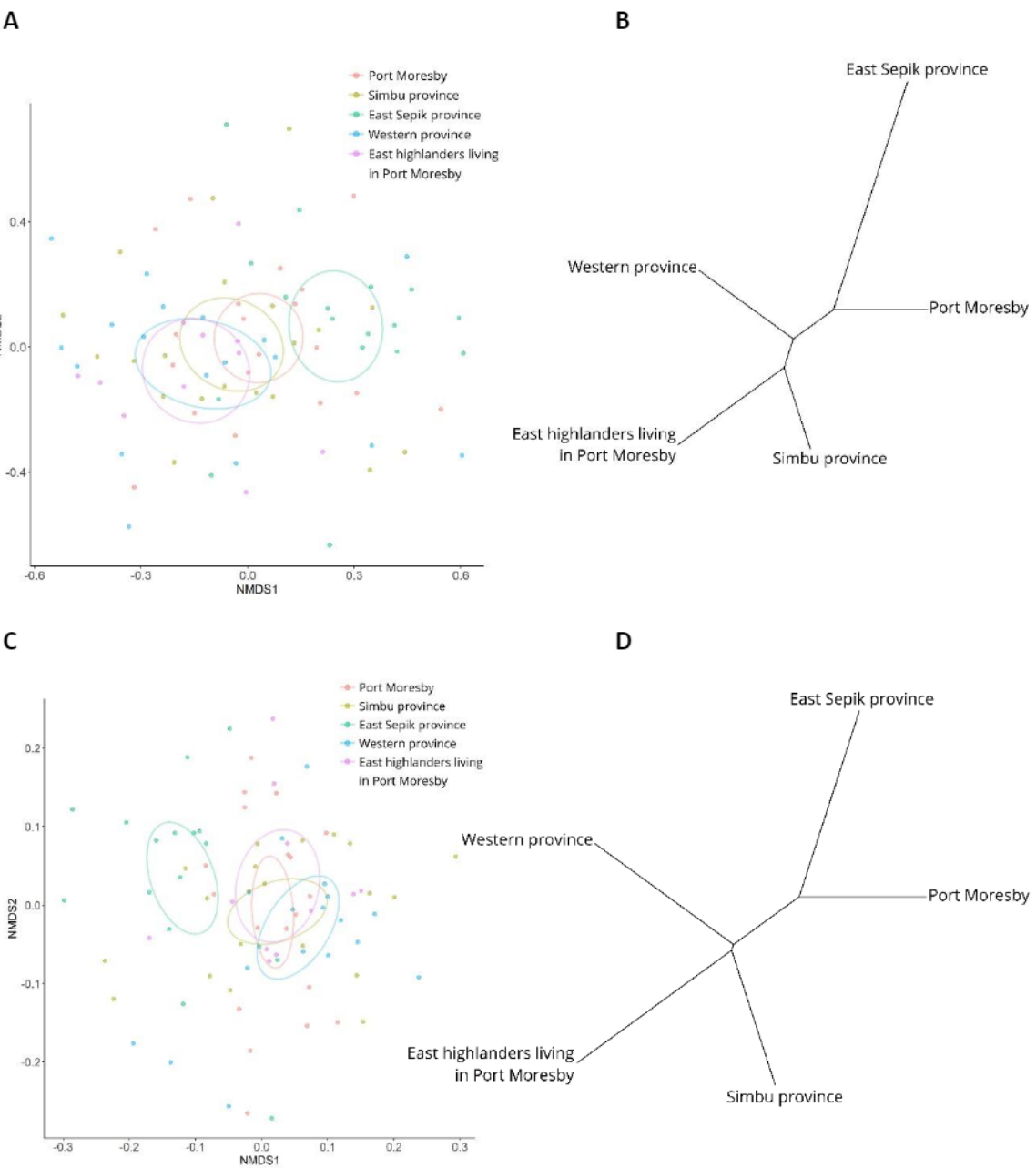


Figure S4. Beta diversity and microbiome profiling of the different groups. A nMDS plot and PERMANOVA analysis for the (A) Jaccard distance and the (C) unweighted UniFrac distance between the five tested groups: Port Moresby (in red); Simbu province (in yellow); East Sepik province (in green); Western province (in blue); and East highlanders living in Port Moresby (in pink). Unrooted neighbor-joining trees of the mean (B) Jaccard distance and (D) unweighted UniFrac distances between the 5 groups. Stress value for the NMDS in Jaccard is 0.2182 and for the unweighted UniFrac is 0.2187.

Table S1. Demographic information (Age (A), Gender (G), Origin for both parents, Sampling place) for all the individuals in this study, indication of which samples we obtained Array microbiome results and WGS microbiome results. For these last, we also added the number of unmapped reads used for microbiome inference and the sequencing depth of the WGS.

Project ID	Group	A	G	Origin for both parents	Sampling place	Region	Array	WGS	WGS unmapped reads	% Reads mapping to the oral microbiome	Sequencing depth	Reference
PNG104	Port Moresby	51	F	Oro	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	4511582	60.28%	WGS 10X	This study
PNG174	Port Moresby	24	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	11797997	73.61%	WGS 10X	This study
PNG177	Port Moresby	25	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG18	Port Moresby	25	M	Oro	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	11211278	77.14%	WGS 10X	This study
PNG200	Port Moresby	21	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG201	Port Moresby	22	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	5425199	63.57%	WGS 10X	This study
PNG221	Port Moresby	26	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	5660918	67.72%	WGS 10X	This study
PNG24	Port Moresby	26	M	Oro	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	11721020	67.50%	WGS 30X	Brucato et al. (2021)
PNG304	Port Moresby	64	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	7156056	57.37%	WGS 10X	This study
PNG316	Port Moresby	58	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG323	Port Moresby	46	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG327	Port Moresby	45	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	3058858	55.79%	WGS 10X	This study
PNG343	Port Moresby	37	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG53	Port Moresby	47	M	Milne Bay	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	8619875	64.68%	WGS 30X	Brucato et al. (2021)
PNG61	Port Moresby	34	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	7670395	55.36%	WGS 30X	Brucato et al. (2021)
PNG65	Port Moresby	25	M	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	NO	-	-	-	This study
PNG68	Port Moresby	29	M	Milne Bay	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	3052005	71.70%	WGS 30X	Brucato et al. (2021)
PNG70	Port Moresby	25	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	2580272	60.35%	WGS 30X	Brucato et al. (2021)
PNG79	Port Moresby	25	F	Central	UPNG	Lowlands coastal east (sampled in lowland coastal)	YES	YES	2748776	60.98%	WGS 10X	This study
PNG365	Simbu province	67	M	Simbu	Mont Whitem	Highlands (sampled in Highlands)	YES	YES	2070095	73.16%	WGS 10X	This study

PNG366	Simbu province	41	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	7488572	67.57%	WGS 10X	This study
PNG367	Simbu province	44	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	2481150	67.88%	WGS 10X	This study
PNG370	Simbu province	42	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	16893824	73.94%	WGS 10X	This study
PNG379	Simbu province	41	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	10945715	80.06%	WGS 10X	This study
PNG380	Simbu province	40	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	NO	-	-	-	This study
PNG382	Simbu province	55	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	10075442	62.69%	WGS 10X	This study
PNG384	Simbu province	57	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	5257392	67.24%	WGS 10X	This study
PNG385	Simbu province	67	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	3104156	40.15%	WGS 10X	This study
PNG388	Simbu province	67	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	4202801	25.88%	WGS 10X	This study
PNG391	Simbu province	20	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	5602756	58.01%	WGS 10X	This study
PNG392	Simbu province	25	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	NO	-	-	-	This study
PNG393	Simbu province	22	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	6233582	63.81%	WGS 10X	This study
PNG394	Simbu province	24	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	6530778	62.86%	WGS 10X	This study
PNG395	Simbu province	27	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	11737937	69.87%	WGS 10X	This study
PNG399	Simbu province	21	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	2690961	66.55%	WGS 10X	This study
PNG401	Simbu province	22	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	2523427	69.33%	WGS 10X	This study
PNG402	Simbu province	21	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	2408021	61.42%	WGS 10X	This study
PNG404	Simbu province	20	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	4480594	65.04%	WGS 10X	This study
PNG405	Simbu province	23	M	Simbu	Mont Whilem	Highlands (sampled in Highlands)	YES	YES	5763730	57.36%	WGS 10X	This study
PNG121	East Sepik province	33	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-	-	-	This study
PNG122	East Sepik province	43	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-	-	-	This study
PNG124	East Sepik province	40	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	13215849	72.53%	WGS 10X	This study
PNG125	East Sepik province	38	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-	-	-	This study
PNG129	East Sepik province	50	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	44976955	71.66%	WGS 10X	This study
PNG130	East Sepik province	40	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	16117736	69.81%	WGS 10X	This study
PNG131	East Sepik province	23	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	31931587	61.19%	WGS 10X	This study

PHD THESIS

APPENDICES

NICOLE PEDRO

PNG132	East Sepik province	30	M	East Sepik	Awin, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	4834533	81.01%	WGS 10X	This study
PNG222	East Sepik province	56	M	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-		-	This study
PNG223	East Sepik province	25	F	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-		-	This study
PNG224	East Sepik province	28	M	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-		-	This study
PNG225	East Sepik province	47	M	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-		-	This study
PNG226	East Sepik province	36	F	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	NO	-		-	This study
PNG228	East Sepik province	51	M	East Sepik	Awim, Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	3666747	62.39%	WGS 10X	This study
PNG234	East Sepik province	39	M	East Sepik	Imboim Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	10024105	53.83%	WGS 10X	This study
PNG235	East Sepik province	54	M	East Sepik	Imboim Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	4834533	52.39%	WGS 10X	This study
PNG239	East Sepik province	41	M	East Sepik	Imboim Angoram district, East Sepik	Lowlands inland north (sampled in lowlands inland north)	YES	YES	3004850	48.30%	WGS 10X	This study
PNG511	Western province	44	F	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	29537421	43.75%	WGS 10X	This study
PNG513	Western province	32	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	1606510	59.12%	WGS 10X	This study
PNG515	Western province	40	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	1483445	63.96%	WGS 10X	This study
PNG517	Western province	30	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	9246242	57.80%	WGS 10X	This study
PNG518	Western province	55	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	6910165	64.40%	WGS 10X	This study
PNG519	Western province	48	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	7053474	65.53%	WGS 10X	This study
PNG521	Western province	60	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	8664017	58.75%	WGS 10X	This study
PNG523	Western province	51	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	6568630	64.73%	WGS 10X	This study
PNG524	Western province	45	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	8608011	27.28%	WGS 10X	This study
PNG526	Western province	44	F	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	4986156	64.90%	WGS 10X	This study
PNG527	Western province	62	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	9673261	38.12%	WGS 10X	This study
PNG529	Western province	53	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	14262885	74.33%	WGS 10X	This study
PNG541	Western province	47	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	3613197	72.12%	WGS 10X	This study
PNG531	Western province	40	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	5442665	72.46%	WGS 10X	This study
PNG532	Western province	58	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	13198200	50.03%	WGS 10X	This study
PNG535	Western province	62	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	1505044	62.50%	WGS 10X	This study

PHD THESIS

APPENDICES

NICOLE PEDRO

PNG538	Western province	52	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	8566115	70.50%	WGS 10X	This study
PNG540	Western province	40	M	Western Fly	Daru, Western Province, PNG	Lowlands coastal west (sampled in lowland coastal west)	YES	YES	3512490	68.75%	WGS 10X	This study
PNG10	East highlanders living in Port Moresby	38	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	9841634	63.58%	WGS 30X	Brucato et al. (2021)
PNG101	East highlanders living in Port Moresby	26	F	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	13152119	60.07%	WGS 30X	Brucato et al. (2021)
PNG188	East highlanders living in Port Moresby	21	M	Eastern Highland	UPNG	Highlands (sampled in lowlands coastal east)	YES	NO	-	-	-	This study
PNG193	East highlanders living in Port Moresby	23	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	NO	-	-	-	This study
PNG219	East highlanders living in Port Moresby	22	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	NO	-	-	-	This study
PNG220	East highlanders living in Port Moresby	24	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	6551834	73.45%	WGS 10X	This study
PNG324	East highlanders living in Port Moresby	42	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	NO	-	-	-	This study
PNG436	East highlanders living in Port Moresby	21	M	Simbu	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	12801203	50.31%	WGS 10X	This study
PNG45	East highlanders living in Port Moresby	22	M	Eastern Highland	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	19586335	57.36%	WGS 10X	This study
PNG100	East highlanders living in Port Moresby	25	F	Eastern Highland	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	6033612	71.30%	WGS 30X	Brucato et al. (2021)
PNG50	East highlanders living in Port Moresby	23	F	Eastern Highland	UPNG	Highlands (sampled in lowlands coastal east)	YES	YES	11761679	73.23%	WGS 30X	Brucato et al. (2021)

Brucato, Nicolas et al. 2021. 'Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul', Molecular Biology and Evolution, 38: 5107-21.

Table S2. Frequency of all the genera detected in all samples with the WGS microbiome results and the Array microbiome results.

Genera	WGS microbiome	Array microbiome
Fusobacterium	1	1
Prevotella	1	1
Streptococcus	1	1
Veillonella	1	1
Rothia	1	0.984615
Haemophilus	1	0.969231
Neisseria	1	0.969231
Treponema	1	0.969231
Actinomyces	1	0.907692
Campylobacter	1	0.876923
Leptotrichia	1	0.692308
Selenomonas	1	0.692308
Gemella	1	0.676923
Vibrio	1	0.276923
Dialister	1	0.215385
Aggregatibacter	1	0.153846
Corynebacterium	1	0.153846
Filifactor	1	0.138462
Bifidobacterium	1	0.076923
Actinobacillus	1	0.046154
Lactobacillus	1	0.030769
Parvimonas	1	0.030769
Pasteurella	1	0.030769
Schaalia	1	0.030769
Enterococcus	1	0.015385
Acinetobacter	1	0
Arthrobacter	1	0
Bacillus	1	0
Barnesiella	1	0
Chryseobacterium	1	0
Clostridium	1	0
Finegoldia	1	0
Flavobacterium	1	0
Halomonas	1	0
Klebsiella	1	0
Microbacterium	1	0
Mogibacterium	1	0.011765
Mycobacterium	1	0
Mycolicibacterium	1	0
Olsenella	1	0
Paenibacillus	1	0
Pseudomonas	1	0
Rhizobium	1	0
Shewanella	1	0
Sphingobacterium	1	0
Staphylococcus	1	0
Streptomyces	1	0
Xanthomonas	1	0
Porphyromonas	0.985714	0.969231
Mycoplasma	0.985714	0.569231
Capnocytophaga	0.985714	0.476923
Tannerella	0.985714	0.415385
Alloprevotella	0.985714	0.276923
Pseudoleptotrichia	0.985714	0.169231
Kingella	0.985714	0.107692
Bulleidia	0.985714	0.092308
Lancefieldella	0.985714	0.061538
Eikenella	0.985714	0.030769
Phocaeicola	0.985714	0.030769
Lachnoanaerobaculum	0.985714	0.015385
Aeromonas	0.985714	0
Alistipes	0.985714	0
Bacteroides	0.985714	0
Blautia	0.985714	0
Bradyrhizobium	0.985714	0
Burkholderia	0.985714	0
Butyrivimonas	0.985714	0
Candidatus Nanosynbacter	0.985714	0
Chitinophaga	0.985714	0
Cutibacterium	0.985714	0
Helicobacter	0.985714	0
Hymenobacter	0.985714	0
Kocuria	0.985714	0
Lachnoclostridium	0.985714	0
Lactococcus	0.985714	0
Listeria	0.985714	0

Mucilaginibacter	0.985714	0
Muribaculum	0.985714	0
Nocardioides	0.985714	0
Odoribacter	0.985714	0
Parabacteroides	0.985714	0
Paraprevotella	0.985714	0
Pedobacter	0.985714	0
Plasmodium	0.985714	0
Polaribacter	0.985714	0
Pseudoalteromonas	0.985714	0
Rhodococcus	0.985714	0
Lautropia	0.971429	0.261538
Mycoplasma	0.971429	0.030769
Cardiobacterium	0.971429	0.015385
Desulfovibrio	0.971429	0.015385
Mannheimia	0.971429	0.015385
Aminipila	0.971429	0
Arachnia	0.971429	0
Blattabacterium	0.971429	0
Brevibacterium	0.971429	0
Buchnera	0.971429	0
Butyrivibrio	0.971429	0
Clostridioides	0.971429	0
Cupriavidus	0.971429	0
Duncaniella	0.971429	0
Jeotgalibaca	0.971429	0
Legionella	0.971429	0
Ligilactobacillus	0.971429	0
Lysinibacillus	0.971429	0
Mesorhizobium	0.971429	0
Sodaliophilus	0.971429	0
Sphingomonas	0.971429	0
Spiroplasma	0.971429	0
Spirosoma	0.971429	0
Synechococcus	0.971429	0
Vagococcus	0.971429	0
Propionibacterium	0.957143	0.015385
Carnobacterium	0.957143	0
Enterobacter	0.957143	0
Eubacterium	0.957143	0.552941
Fretibacterium	0.957143	0
Mammaliicoccus	0.957143	0
Nostoc	0.957143	0
Pauljensenia	0.957143	0
Serratia	0.957143	0

Virgibacillus	0.957143	0
Megasphaera	0.942857	0.584615
Abiotrophia	0.942857	0.169231
Aerococcus	0.942857	0
Anaerococcus	0.942857	0
Anaerocolumna	0.942857	0
Anaerostipes	0.942857	0
Coproccoccus	0.942857	0
Elizabethkingia	0.942857	0
Lacrimispora	0.942857	0
Roseburia	0.942857	0
Weissella	0.942857	0
Limosilactobacillus	0.928571	0.015385
Aliarcobacter	0.928571	0
Bartonella	0.928571	0
Brachybacterium	0.928571	0
Citrobacter	0.928571	0
Dysgonomonas	0.928571	0
Micromonospora	0.928571	0
Nonlabens	0.928571	0
Pantoea	0.928571	0
Paraburkholderia	0.928571	0
Pontibacter	0.928571	0
Streptobacillus	0.928571	0
Histophilus	0.914286	0.015385
Comanilactobacillus	0.914286	0
Desulfohalobium	0.914286	0
Francisella	0.914286	0
Glaesserella	0.914286	0
Sneathia	0.914286	0
Stenotrophomonas	0.914286	0
Yersinia	0.9	0.015385
Coproccoccus	0.9	0
Erysipelothrix	0.9	0
Faecalibacterium	0.9	0
Geobacter	0.9	0
Kaistella	0.9	0
Leuconostoc	0.9	0
Mediterraneibacter	0.9	0
Nocardia	0.9	0
Ottowia	0.9	0
Paludibacter	0.9	0
Planococcus	0.9	0
Priestia	0.9	0
Romboutsia	0.9	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Ruminococcus	0.9	0
Toxoplasma	0.9	0
Achromobacter	0.885714	0
Bergeyella	0.885714	0
Bordetella	0.885714	0
Brevibacillus	0.885714	0
Erysipelatoclostridium	0.885714	0
Gramella	0.885714	0
Massilia	0.885714	0
Myroides	0.885714	0
Proteus	0.885714	0
Sporosarcina	0.885714	0
Tessaracoccus	0.885714	0
Variovorax	0.885714	0
Winogradskyella	0.885714	0
Simonsiella	0.871429	0.046154
Acholeplasma	0.871429	0
Acidovorax	0.871429	0
Amycolatopsis	0.871429	0
Cellulomonas	0.871429	0
Moraxella	0.871429	0
Providencia	0.871429	0
Pseudarthrobacter	0.871429	0
Escherichia	0.857143	0.046154
Rodentibacter	0.857143	0.046154
Alkalihalobacillus	0.857143	0
Arcobacter	0.857143	0
Gordonia	0.857143	0
Latilactobacillus	0.857143	0
Lysobacter	0.857143	0
Macrococcus	0.857143	0
Methylobacterium	0.857143	0
Petrimonas	0.857143	0
Deinococcus	0.842857	0
Draconibacterium	0.842857	0
Georgenia	0.842857	0
Micrococcus	0.842857	0
Pediococcus	0.842857	0
Peribacillus	0.842857	0
Proteiniphilum	0.842857	0
Psychrobacter	0.842857	0
Riemerella	0.842857	0
Tenacibaculum	0.842857	0
Trueperella	0.842857	0
Alteromonas	0.828571	0

Dolosigranulum	0.828571	0
Enterocloster	0.828571	0
Exiguobacterium	0.828571	0
Formosa	0.828571	0
Glutamicibacter	0.828571	0
Mageebacillus	0.828571	0
Maribacter	0.828571	0
Rathayibacter	0.828571	0
Agrobacterium	0.814286	0
Arcanobacterium	0.814286	0
Candidatus Saccharimonas	0.814286	0
Curtobacterium	0.814286	0
Dickeya	0.814286	0
Megamonas	0.814286	0
Paracoccus	0.814286	0
Peptacetobacter	0.814286	0
Pseudonocardia	0.814286	0
Azospirillum	0.8	0
Echinicola	0.8	0
Erwinia	0.8	0
Lachnospira	0.8	0
Leucobacter	0.8	0
Oceanobacillus	0.8	0
Acidipropionibacterium	0.785714	0
Agromyces	0.785714	0
Brachyspira	0.785714	0
Caldicellulosiruptor	0.785714	0
Dietzia	0.785714	0
Flaviflexus	0.785714	0
Geobacillus	0.785714	0
Leifsonia	0.785714	0
Sphingobium	0.785714	0
Cryptobacterium	0.771429	0.015385
Anoxybacillus	0.771429	0
Desulfosporosinus	0.771429	0
Lentilactobacillus	0.771429	0
Ornithinimicrobium	0.771429	0
Polynucleobacter	0.771429	0
Sebaldella	0.771429	0
Tetragenococcus	0.771429	0
Xylanimonas	0.771429	0
Actinoplanes	0.757143	0
Aeromicrobium	0.757143	0
Fusarium	0.757143	0
Gardnerella	0.757143	0

Janibacter	0.757143	0
Lactiplantibacillus	0.757143	0
Malassezia	0.757143	0
Marinobacter	0.757143	0
Pseudobutyrvibrio	0.757143	0
Ruania	0.757143	0
Sulfurimonas	0.757143	0
Tenuifilum	0.757143	0
Thermoanaerobacterium	0.757143	0
Acetivibrio	0.742857	0
Arachidicoccus	0.742857	0
Azoarcus	0.742857	0
Comamonas	0.742857	0
Dokdonia	0.742857	0
Lacticaseibacillus	0.742857	0
Pectobacterium	0.742857	0
Sulfurospirillum	0.742857	0
Brevundimonas	0.728571	0
Cellulosimicrobium	0.728571	0
Dermabacter	0.728571	0
Desulfitobacterium	0.728571	0
Janthinobacterium	0.728571	0
Microlunatus	0.728571	0
Runella	0.728571	0
Actinomadura	0.714286	0
Avibacterium	0.714286	0
Calothrix	0.714286	0
Hydrogenophaga	0.714286	0
Liquorilactobacillus	0.714286	0
Marinomonas	0.714286	0
Empedobacter	0.7	0
Epilithonimonas	0.7	0
Fastidiosipila	0.7	0
Flammeovirga	0.7	0
Serinicoccus	0.7	0
Collinsella	0.685714	0
Eggerthella	0.685714	0
Frankia	0.685714	0
Pandoraea	0.685714	0
Phascolarctobacterium	0.685714	0
Prochlorococcus	0.685714	0
Anaerobutyricum	0.671429	0
Dermacoccus	0.671429	0
Lacinutrix	0.671429	0
Mycobacteroides	0.671429	0

Parolsenella	0.671429	0
Pelosinus	0.671429	0
Peptoniphilus	0.671429	0
Salmonella	0.671429	0
Sinorhizobium	0.671429	0
Sorangium	0.671429	0
Winkia	0.671429	0
Acetobacter	0.657143	0
Amedibacterium	0.657143	0
Caulobacter	0.657143	0
Desulfotomaculum	0.657143	0
Loigolactobacillus	0.657143	0
Luteimonas	0.657143	0
Malaciobacter	0.657143	0
Pseudodesulfovibrio	0.657143	0
Ralstonia	0.657143	0
Saccharomonospora	0.657143	0
Sphingopyxis	0.657143	0
Thermoanaerobacter	0.657143	0
Aquimarina	0.642857	0
Bibersteinia	0.642857	0
Cellulophaga	0.642857	0
Metabacillus	0.642857	0
Rhodoferax	0.642857	0
Sanguibacter	0.642857	0
Sulfitobacter	0.642857	0
Aequorivita	0.628571	0
Basfia	0.628571	0
Clavibacter	0.628571	0
Colwellia	0.628571	0
Methylomonas	0.628571	0
Rufibacter	0.628571	0
Salinibacterium	0.628571	0
Bdellovibrio	0.614286	0
Chromobacterium	0.614286	0
Fibrella	0.614286	0
Nakamurella	0.614286	0
Ruminiclostridium	0.614286	0
Turicibacter	0.614286	0
Scardovia	0.6	0.030769
Acidaminococcus	0.6	0
Algoriphagus	0.6	0
Candidatus Planktophila	0.6	0
Cellulosilyticum	0.6	0
Crassaminicella	0.6	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Herbaspirillum	0.6	0
Hungatella	0.6	0
Isoptericola	0.6	0
Levilactobacillus	0.6	0
Mobiluncus	0.6	0
Murdochiella	0.6	0
Novosphingobium	0.6	0
Alkaliphilus	0.585714	0
Bisgaardia	0.585714	0
Candida	0.585714	0
Kurthia	0.585714	0
Leptospira	0.585714	0
Muricauda	0.585714	0
Neobacillus	0.585714	0
Sarcina	0.585714	0
Alcaligenes	0.571429	0
Caproiciproducens	0.571429	0
Cloacibacterium	0.571429	0
Cruoricaptor	0.571429	0
Herbinix	0.571429	0
Kordia	0.571429	0
Paeniclostridium	0.571429	0
Paucilactobacillus	0.571429	0
Rhodopseudomonas	0.571429	0
Actinobaculum	0.557143	0
Agrococcus	0.557143	0
Anaerotignum	0.557143	0
Candidatus Phytoplasma	0.557143	0
Gordonibacter	0.557143	0
Kitasatospora	0.557143	0
Microbulbifer	0.557143	0
Photobacterium	0.557143	0
Planobacterium	0.557143	0
Propionimicrobium	0.557143	0
Cohnella	0.542857	0
Devosia	0.542857	0
Massilistercora	0.542857	0
Myxococcus	0.542857	0
Oenococcus	0.542857	0
Ornithobacterium	0.542857	0
Phycococcus	0.542857	0
Saccharothrix	0.542857	0
Thauera	0.542857	0
Wolbachia	0.542857	0
Xenorhabdus	0.542857	0

Algibacter	0.528571	0
Auritidibacter	0.528571	0
Boudabousia	0.528571	0
Fructilactobacillus	0.528571	0
Gallibacterium	0.528571	0
Kytococcus	0.528571	0
Labilibaculum	0.528571	0
Nitrosomonas	0.528571	0
Nonomuraea	0.528571	0
Saccharopolyspora	0.528571	0
Salinivirga	0.528571	0
Sinomonas	0.528571	0
Akkermansia	0.514286	0
Alicyclobacillus	0.514286	0
Caloramator	0.514286	0
Christensenella	0.514286	0
Desulfomicrobium	0.514286	0
Ezakiella	0.514286	0
Nocardiosis	0.514286	0
Parascardovia	0.514286	0.011765
Streptacidiphilus	0.514286	0
Bosea	0.5	0
Brochothrix	0.5	0
Candidatus Pelagibacter	0.5	0
Chlamydia	0.5	0
Cryobacterium	0.5	0
Ehrlichia	0.5	0
Intestinibaculum	0.5	0
Oscillibacter	0.5	0
Psychroflexus	0.5	0
Ureibacillus	0.5	0
Roseolovirus	0.485714	0.015385
Alkalitalea	0.485714	0
Borrelia	0.485714	0
Dyadobacter	0.485714	0
Erythrobacter	0.485714	0
Gottschalkia	0.485714	0
Kosakonia	0.485714	0
Moineauvirus	0.485714	0
Paenarthrobacter	0.485714	0
Paraclostridium	0.485714	0
Protaetiibacter	0.485714	0
Ruthenibacterium	0.485714	0
Antarcticibacterium	0.471429	0
Apibacter	0.471429	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Brucella	0.471429	0
Chlorobium	0.471429	0
Diaphorobacter	0.471429	0
Frederiksenia	0.471429	0
Labrenzia	0.471429	0
Mesoplasma	0.471429	0
Photorhabdus	0.471429	0
Rickettsia	0.471429	0
Rubrobacter	0.471429	0
Weizmannia	0.471429	0
Chlorobaculum	0.457143	0
Cytobacillus	0.457143	0
Fannyhessea	0.457143	0
Gimesia	0.457143	0
Halanaerobium	0.457143	0
Jeotgalicoccus	0.457143	0
Melissococcus	0.457143	0
Mycetocola	0.457143	0
Sedimentibacter	0.457143	0
Actinoalloteichus	0.442857	0
Actinotignum	0.442857	0
Adhaeribacter	0.442857	0
Altererythrobacter	0.442857	0
Candidatus Arthromitus	0.442857	0
Cellvibrio	0.442857	0
Croceimicrobium	0.442857	0
Edwardsiella	0.442857	0
Flavisolibacter	0.442857	0
Halobacillus	0.442857	0
Methanosarcina	0.442857	0
Otariodibacter	0.442857	0
Solibacillus	0.442857	0
Sphaerochaeta	0.442857	0
Tamlana	0.442857	0
Viridibacillus	0.442857	0
Acetobacterium	0.428571	0
Adlercreutzia	0.428571	0
Alysiella	0.428571	0
Amniculibacterium	0.428571	0
Delftia	0.428571	0
Ensifer	0.428571	0
Marinilactibacillus	0.428571	0
Phytohabitans	0.428571	0
Thiothrix	0.428571	0
Oerskovia	0.414286	0

Paenisporosarcina	0.414286	0
Pontibacillus	0.414286	0
Pseudactinotalea	0.414286	0
Sutterella	0.414286	0
Vespertiliibacter	0.414286	0
Aquitalea	0.4	0
Archangium	0.4	0
Aurantimicrobium	0.4	0
Brenneria	0.4	0
Cyclobacterium	0.4	0
Dictyostelium	0.4	0
Geminocystis	0.4	0
Hathewayia	0.4	0
Humibacter	0.4	0
Ndongobacter	0.4	0
Neptunomonas	0.4	0
Propioniciclava	0.4	0
Salegentibacter	0.4	0
Slackia	0.4	0
Streptosporangium	0.4	0
Apilactobacillus	0.385714	0
Desulfosarcina	0.385714	0
Filimonas	0.385714	0
Ilyobacter	0.385714	0
Jiangella	0.385714	0
Leadbetterella	0.385714	0
Leptolyngbya	0.385714	0
Lutibacter	0.385714	0
Olivibacter	0.385714	0
Rhodoluna	0.385714	0
Solitalea	0.385714	0
Thalassospira	0.385714	0
Actinosynnema	0.371429	0
Aliivibrio	0.371429	0
Beutenbergia	0.371429	0
Candidatus Kinetoplastibacterium	0.371429	0
Conchiformibius	0.371429	0
Fron dih abitans	0.371429	0
Idiomarina	0.371429	0
Luteimicrobium	0.371429	0
Mesobacillus	0.371429	0
Microterricola	0.371429	0
Miniimonas	0.371429	0
Owenweeksia	0.371429	0
Pradoshia	0.371429	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Prosthecochloris	0.371429	0
Pseudoclostridium	0.371429	0
Pyricularia	0.371429	0
Tsukamurella	0.371429	0
Acetoanaerobium	0.357143	0
Alcanivorax	0.357143	0
Amphibacillus	0.357143	0
Anaeromyxobacter	0.357143	0
Berryella	0.357143	0
Borrelia	0.357143	0
Corallococcus	0.357143	0
Cronobacter	0.357143	0
Hydrogenovibrio	0.357143	0
Leishmania	0.357143	0
Nitratiruptor	0.357143	0
Polaromonas	0.357143	0
Roseomonas	0.357143	0
Sporomusa	0.357143	0
Thermosipho	0.357143	0
Weeksella	0.357143	0
Acidithiobacillus	0.342857	0
Aureimonas	0.342857	0
Botrytis	0.342857	0
Caldibacillus	0.342857	0
Candidatus Methylopusillus	0.342857	0
Fibrobacter	0.342857	0
Fictibacillus	0.342857	0
Intrasporangium	0.342857	0
Maridesulfovibrio	0.342857	0
Niabella	0.342857	0
Plantactinospora	0.342857	0
Plantibacter	0.342857	0
Raoultella	0.342857	0
Spirochaeta	0.342857	0
Succinivibrio	0.342857	0
Suicoccus	0.342857	0
Thermus	0.342857	0
Anaerobacillus	0.328571	0
Bernardetia	0.328571	0
Caloranaerobacter	0.328571	0
Chryseolinea	0.328571	0
Conexibacter	0.328571	0
Denitrobacterium	0.328571	0
Ethanoligenens	0.328571	0
Gilliamella	0.328571	0

Granulicella	0.328571	0
Keratinibaculum	0.328571	0
Komagataeibacter	0.328571	0
Methylobacterium	0.328571	0
Niastella	0.328571	0
Paraliobacillus	0.328571	0
Phaeobacter	0.328571	0
Pseudobacter	0.328571	0
Rhodothermus	0.328571	0
Roseivirga	0.328571	0
Roseovarius	0.328571	0
Secundilactobacillus	0.328571	0
Terriglobus	0.328571	0
Thermococcus	0.328571	0
Thiomonas	0.328571	0
Xylanimicrobium	0.328571	0
Aneurinibacillus	0.314286	0
Arcticibacterium	0.314286	0
Austwickia	0.314286	0
Cedecea	0.314286	0
Collimonas	0.314286	0
Desulfonema	0.314286	0
Desulfopila	0.314286	0
Dyella	0.314286	0
Lelliottia	0.314286	0
Methanococcus	0.314286	0
Mixta	0.314286	0
Mobilicoccus	0.314286	0
Morganella	0.314286	0
Mycobacter	0.314286	0
Paraflavitalea	0.314286	0
Peptoclostridium	0.314286	0
Petrocella	0.314286	0
Psychrobacillus	0.314286	0
Psychromicrobium	0.314286	0
Rhizobacter	0.314286	0
Seonamhaeicola	0.314286	0
Thiomicrobacter	0.314286	0
Verrucomicrobium	0.314286	0
Vitreoscilla	0.314286	0
Arsenophonus	0.3	0
Candidatus Blochmannia	0.3	0
Citricoccus	0.3	0
Desulfuromonas	0.3	0
Flagellimonas	0.3	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Flaviumibacter	0.3	0
Flintibacter	0.3	0
Furfurilactobacillus	0.3	0
Gemmatimonas	0.3	0
Geosporobacter	0.3	0
Hafnia	0.3	0
Kineococcus	0.3	0
Kribbella	0.3	0
Lawsonella	0.3	0
Leclercia	0.3	0
Methylocystis	0.3	0
Micropruina	0.3	0
Microvirga	0.3	0
Olleya	0.3	0
Pectinatus	0.3	0
Pseudopedobacter	0.3	0
Rahnella	0.3	0
Stanieria	0.3	0
Actinotalea	0.285714	0
Azospira	0.285714	0
Brevilactibacter	0.285714	0
Celeribacter	0.285714	0
Devriesea	0.285714	0
Flavonifractor	0.285714	0
Gracilibacillus	0.285714	0
Hyphomicrobium	0.285714	0
Kushneria	0.285714	0
Marivirga	0.285714	0
Marteella	0.285714	0
Neomicrococcus	0.285714	0
Panacibacter	0.285714	0
Pseudocnuella	0.285714	0
Rhodanobacter	0.285714	0
Salinimonas	0.285714	0
Syntrophotalea	0.285714	0
Theileria	0.285714	0
Thioalkalivibrio	0.285714	0
Tumebacillus	0.285714	0
Ureaplasma	0.285714	0
Wenyingzhuangia	0.285714	0
Zobellia	0.285714	0
Cepunavirus	0.271429	0.061538
Altibacter	0.271429	0
Andhravirus	0.271429	0
Belliella	0.271429	0

Candidatus Aquiluna	0.271429	0
Candidatus Izimaplasma	0.271429	0
Candidatus Thioglobus	0.271429	0
Dehalococcoides	0.271429	0
Fervidobacterium	0.271429	0
Flavivirga	0.271429	0
Gilvibacter	0.271429	0
Glaciecola	0.271429	0
Haliscomenobacter	0.271429	0
Halobacteriovorax	0.271429	0
Labilithrix	0.271429	0
Leeuwenhoekella	0.271429	0
Mariprofundus	0.271429	0
Methanobrevibacter	0.271429	0
Nitrosococcus	0.271429	0
Opitutus	0.271429	0
Phreatobacter	0.271429	0
Rheinheimera	0.271429	0
Rhodocytophaga	0.271429	0
Rummeliibacillus	0.271429	0
Saccharomyces	0.271429	0
Salicibibacter	0.271429	0
Salinibacter	0.271429	0
Sporolactobacillus	0.271429	0
Thalassotalea	0.271429	0
Thermoactinomyces	0.271429	0
Thermomonospora	0.271429	0
Thermotoga	0.271429	0
Tissierella	0.271429	0
Urechidicola	0.271429	0
Actinocatenispora	0.257143	0
Acutalibacter	0.257143	0
Aminomonas	0.257143	0
Anaerolinea	0.257143	0
Anoxybacter	0.257143	0
Anseongella	0.257143	0
Aquirufa	0.257143	0
Arenibacter	0.257143	0
Babesia	0.257143	0
Brussowvirus	0.257143	0
Cereibacter	0.257143	0
Changchengzhania	0.257143	0
Dehalobacter	0.257143	0
Entomomonas	0.257143	0
Fischerella	0.257143	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Gemmata	0.257143	0
Gluconobacter	0.257143	0
Halarcobacter	0.257143	0
Jonesia	0.257143	0
Jonquetella	0.257143	0
Lapidilactobacillus	0.257143	0
Liberibacter	0.257143	0
Magnetospirillum	0.257143	0
Methylococcus	0.257143	0
Microcystis	0.257143	0
Monoglobus	0.257143	0
Oceanihabitans	0.257143	0
Octadecabacter	0.257143	0
Oscillatoria	0.257143	0
Pelodictyon	0.257143	0
Phoenicibacter	0.257143	0
Pimelobacter	0.257143	0
Renibacterium	0.257143	0
Rhodobacter	0.257143	0
Rhodovulum	0.257143	0
Shinella	0.257143	0
Sphingorhabdus	0.257143	0
Stieleria	0.257143	0
Tatumella	0.257143	0
Thermaerobacter	0.257143	0
Thermoclostridium	0.257143	0
Thermomonas	0.257143	0
Undibacterium	0.257143	0
Acidihalobacter	0.242857	0
Algicoccus	0.242857	0
Aminobacter	0.242857	0
Arsenicicoccus	0.242857	0
Caballeronia	0.242857	0
Cobetia	0.242857	0
Fulvivirga	0.242857	0
Galbibacter	0.242857	0
Gillisia	0.242857	0
Gudongella	0.242857	0
Kangiella	0.242857	0
Lacibacter	0.242857	0
Lacipirellula	0.242857	0
Marinitoga	0.242857	0
Marinobacterium	0.242857	0
Meiothermus	0.242857	0
Mesoflavibacter	0.242857	0

Muriicola	0.242857	0
Mycolicibacillus	0.242857	0
Naumovozyma	0.242857	0
Neorhizobium	0.242857	0
Nibribacter	0.242857	0
Nitrospira	0.242857	0
Robiginitalea	0.242857	0
Ruegeria	0.242857	0
Saprospira	0.242857	0
Scytonema	0.242857	0
Sediminispirochaeta	0.242857	0
Sodalis	0.242857	0
Tepiditoga	0.242857	0
Thiomicrospira	0.242857	0
Tomitella	0.242857	0
Zymomonas	0.242857	0
Abyssicoccus	0.228571	0
Agreia	0.228571	0
Arabia	0.228571	0
Aromatoleum	0.228571	0
Croceibacter	0.228571	0
Dechloromonas	0.228571	0
Dehalobacterium	0.228571	0
Desulfococcus	0.228571	0
Duganella	0.228571	0
Entomoplasma	0.228571	0
Faecalitalea	0.228571	0
Fluviicola	0.228571	0
Gequatrovirus	0.228571	0
Halobacteroides	0.228571	0
Halocella	0.228571	0
Ignavibacterium	0.228571	0
Luteibacter	0.228571	0
Luteipulveratus	0.228571	0
Microcella	0.228571	0
Moorella	0.228571	0
Moritella	0.228571	0
Nitrincola	0.228571	0
Paraglaciecola	0.228571	0
Pseudothermotoga	0.228571	0
Psychromonas	0.228571	0
Psychrosphaera	0.228571	0
Rimavirus	0.228571	0
Schnuerera	0.228571	0
Sediminicola	0.228571	0

Skermanella	0.228571	0
Streptomonospora	0.228571	0
Sulfuriferula	0.228571	0
Thermanaerovibrio	0.228571	0
Thiospirochaeta	0.228571	0
Catenovulum	0.214286	0
Denitratisoma	0.214286	0
Dermatophilus	0.214286	0
Diaminobutyricimonas	0.214286	0
Endomicrobium	0.214286	0
Epidermidibacterium	0.214286	0
Ereboglobus	0.214286	0
Friedmanniella	0.214286	0
Halioglobus	0.214286	0
Halothermothrix	0.214286	0
Heliobacterium	0.214286	0
Hypericibacter	0.214286	0
Ichthyobacterium	0.214286	0
Kutzneria	0.214286	0
Lentibacillus	0.214286	0
Melioribacter	0.214286	0
Neurospora	0.214286	0
Oceanispirochaeta	0.214286	0
Paraoerskovia	0.214286	0
Pedococcus	0.214286	0
Salipiger	0.214286	0
Sphingosinicella	0.214286	0
Tetrasphaera	0.214286	0
Usitatibacter	0.214286	0
Wenzhouxiangella	0.214286	0
Zunongwangia	0.214286	0
Acidobacterium	0.2	0
Anabaena	0.2	0
Aquiflexum	0.2	0
Aquisphaera	0.2	0
Aureitalea	0.2	0
Azotobacter	0.2	0
Blastococcus	0.2	0
Candidatus Cardinium	0.2	0
Candidatus Cyclonatrum	0.2	0
Candidatus Syntrophocurvum	0.2	0
Chivirus	0.2	0
Cloacibacillus	0.2	0
Coriobacterium	0.2	0
Cytophaga	0.2	0

Dehalogenimonas	0.2	0
Desulfobacter	0.2	0
Desulfocurvibacter	0.2	0
Dictyoglomus	0.2	0
Euzebyella	0.2	0
Hahella	0.2	0
Haliangium	0.2	0
Halorubrum	0.2	0
Hyphobacterium	0.2	0
Intestinimonas	0.2	0
Lederbergia	0.2	0
Limnohabitans	0.2	0
Marmoricola	0.2	0
Methylovimicrobium	0.2	0
Minicystis	0.2	0
Moorea	0.2	0
Pigmentiphaga	0.2	0
Prauserella	0.2	0
Pseudanabaena	0.2	0
Pusillimonas	0.2	0
Radiobacillus	0.2	0
Raineyella	0.2	0
Rhodospirillum	0.2	0
Roseiflexus	0.2	0
Roseimicrobium	0.2	0
Roseobacter	0.2	0
Rouxella	0.2	0
Salinispira	0.2	0
Salinivibrio	0.2	0
Siansivirga	0.2	0
Symbiobacterium	0.2	0
Syntrophobotulus	0.2	0
Teredinibacter	0.2	0
Tetrapisispora	0.2	0
Thermosynechococcus	0.2	0
Yimella	0.2	0
Wolinella	0.185714	0.015385
Aspergillus	0.185714	0
Auraticoccus	0.185714	0
Bombilactobacillus	0.185714	0
Costertonia	0.185714	0
Ectothiorhodospira	0.185714	0
Edaphobacter	0.185714	0
Geodermatophilus	0.185714	0
Heliorestis	0.185714	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Kinneretia	0.185714	0
Mariniflexile	0.185714	0
Methanocaldococcus	0.185714	0
Methylophaga	0.185714	0
Methylophilus	0.185714	0
Nesterenkonia	0.185714	0
Nitrosopumilus	0.185714	0
Novibacillus	0.185714	0
Orrella	0.185714	0
Parashewanella	0.185714	0
Pelistega	0.185714	0
Pepyhexavirus	0.185714	0
Persicimonas	0.185714	0
Pseudoxanthomonas	0.185714	0
Psychroserpens	0.185714	0
Ramlibacter	0.185714	0
Saccharibacillus	0.185714	0
Salinispora	0.185714	0
Sandaracinus	0.185714	0
Thermoanaeroseptrum	0.185714	0
Thermodesulfobacterium	0.185714	0
Tsuneonella	0.185714	0
Acidianus	0.171429	0
Alloactinosynnema	0.171429	0
Caldithrix	0.171429	0
Candidatus Nardonella	0.171429	0
Candidatus Promineofilum	0.171429	0
Carboxydothermus	0.171429	0
Chelatococcus	0.171429	0
Chloroflexus	0.171429	0
Chryseoglobus	0.171429	0
Cystobacter	0.171429	0
Deferribacter	0.171429	0
Desulfotalea	0.171429	0
Gemmatirosa	0.171429	0
Gloeotheca	0.171429	0
Granulibacter	0.171429	0
Herbiconiux	0.171429	0
Hyphomonas	0.171429	0
Iodobacter	0.171429	0
Jeongeupia	0.171429	0
Kluyvera	0.171429	0
Kroppenstedtia	0.171429	0
Limnochorda	0.171429	0
Luteithermobacter	0.171429	0

Methylomicrobium	0.171429	0
Mumia	0.171429	0
Natranaerobius	0.171429	0
Oceanisphaera	0.171429	0
Oligella	0.171429	0
Planctomyces	0.171429	0
Planomicrobium	0.171429	0
Plesiomonas	0.171429	0
Porphyrobacter	0.171429	0
Pseudolabrys	0.171429	0
Starkeya	0.171429	0
Thalassolituus	0.171429	0
Thermincola	0.171429	0
Thioflavicoccus	0.171429	0
Trichodesmium	0.171429	0
Urbifossilla	0.171429	0
Xanthobacter	0.171429	0
Xylella	0.171429	0
Pahexavirus	0.157143	0.030769
Aeribacillus	0.157143	0
Alkalibacter	0.157143	0
Allokutzneria	0.157143	0
Asticcacaulis	0.157143	0
Atlantibacter	0.157143	0
Basilea	0.157143	0
Bradymonas	0.157143	0
Buttiauxella	0.157143	0
Catenulispora	0.157143	0
Croceicoccus	0.157143	0
Cryptococcus	0.157143	0
Curvibacter	0.157143	0
Denitrovibrio	0.157143	0
Desulfallas	0.157143	0
Faecalibaculum	0.157143	0
Ferrimonas	0.157143	0
Glaciihabitans	0.157143	0
Gryllotalpicola	0.157143	0
Hydrogenophilus	0.157143	0
Hylemonella	0.157143	0
Jatrophihabitans	0.157143	0
Jeotgalibacillus	0.157143	0
Kibdelosporangium	0.157143	0
Leisingera	0.157143	0
Limnoglobus	0.157143	0
Luteitalea	0.157143	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Luteolibacter	0.157143	0
Lysinimonas	0.157143	0
Marisediminicola	0.157143	0
Modestobacter	0.157143	0
Parageobacillus	0.157143	0
Paraphotobacterium	0.157143	0
Phenylobacterium	0.157143	0
Phyllobacterium	0.157143	0
Phytobacter	0.157143	0
Planktothrix	0.157143	0
Pluralibacter	0.157143	0
Poseidonibacter	0.157143	0
Pseudogulbenkiania	0.157143	0
Pseudolysobacter	0.157143	0
Roseimaritima	0.157143	0
Roseococcus	0.157143	0
Rosistilla	0.157143	0
Saccharophagus	0.157143	0
Salimicrobium	0.157143	0
Schleiferilactobacillus	0.157143	0
Serpentinomonas	0.157143	0
Singulisphaera	0.157143	0
Sphingosinithalassobacter	0.157143	0
Stackebrandtia	0.157143	0
Stella	0.157143	0
Sugiyamaella	0.157143	0
Sulfurivermis	0.157143	0
Sulfurovum	0.157143	0
Swingsia	0.157143	0
Tardiphaga	0.157143	0
Tepidanaerobacter	0.157143	0
Terribacillus	0.157143	0
Thermodesulfobium	0.157143	0
Thiodictyon	0.157143	0
Verminephrobacter	0.157143	0
Actinomarinicola	0.142857	0
Advenella	0.142857	0
Alicycliphilus	0.142857	0
Allofrancisella	0.142857	0
Anaerohalosphaera	0.142857	0
Anatilimnocola	0.142857	0
Aquaspirillum	0.142857	0
Aurantiacibacter	0.142857	0
Bythopirellula	0.142857	0
Candidatus Koribacter	0.142857	0

Chondromyces	0.142857	0
Colletotrichum	0.142857	0
Congregibacter	0.142857	0
Cyanobium	0.142857	0
Cylindrospermum	0.142857	0
Desulfobacula	0.142857	0
Desulfofarcimen	0.142857	0
Desulfohalobium	0.142857	0
Frigoribacterium	0.142857	0
Geoalkalibacter	0.142857	0
Geovibrio	0.142857	0
Glaciimonas	0.142857	0
Gloeobacter	0.142857	0
Haloplanus	0.142857	0
Hydrogenimonas	0.142857	0
Jejubacter	0.142857	0
Ketogulonicigenium	0.142857	0
Koinonema	0.142857	0
Kyrpidia	0.142857	0
Labrys	0.142857	0
Laceyella	0.142857	0
Lacunisphaera	0.142857	0
Lentzea	0.142857	0
Lonsdalea	0.142857	0
Mameliella	0.142857	0
Methyloceanibacter	0.142857	0
Methylocella	0.142857	0
Methylosinus	0.142857	0
Methylotenera	0.142857	0
Methylovorus	0.142857	0
Nautilia	0.142857	0
Neosasaia	0.142857	0
Neospora	0.142857	0
Nibricoccus	0.142857	0
Nitratireductor	0.142857	0
Oceanithermus	0.142857	0
Qipengyuania	0.142857	0
Salinicoccus	0.142857	0
Sedimentisphaera	0.142857	0
Silvanigrella	0.142857	0
Solimonas	0.142857	0
Spiribacter	0.142857	0
Stappia	0.142857	0
Sulfurifustis	0.142857	0
Sulfurihydrogenibium	0.142857	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Synechocystis	0.142857	0
Syntrophobacter	0.142857	0
Terasakiella	0.142857	0
Thermacetogenium	0.142857	0
Thermobifida	0.142857	0
Thermomicrobium	0.142857	0
Thermovirga	0.142857	0
Vulgatibacter	0.142857	0
Yokenella	0.142857	0
Actinopolymorpha	0.128571	0
Actinopolyspora	0.128571	0
Ahniella	0.128571	0
Allochromatium	0.128571	0
Aquicella	0.128571	0
Bacteriovorax	0.128571	0
Baekduia	0.128571	0
Beggiatoa	0.128571	0
Blastochloris	0.128571	0
Caldanaerobacter	0.128571	0
Caminibacter	0.128571	0
Candidatus Desulfofervidus	0.128571	0
Candidatus Hamiltonella	0.128571	0
Candidatus Methylospira	0.128571	0
Candidatus Nanopelagicus	0.128571	0
Candidatus Sulcia	0.128571	0
Castellaniella	0.128571	0
Chelativorans	0.128571	0
Cycloclasticus	0.128571	0
Dichelobacter	0.128571	0
Drechmeria	0.128571	0
Elusimicrobium	0.128571	0
Filomicrobium	0.128571	0
Frischella	0.128571	0
Grimontia	0.128571	0
Haloferax	0.128571	0
Halorhodospira	0.128571	0
Haloterrigena	0.128571	0
Herminiimonas	0.128571	0
Hoyosella	0.128571	0
Humisphaera	0.128571	0
Kazachstania	0.128571	0
Kiritimatiella	0.128571	0
Kosmotoga	0.128571	0
Ktedonosporobacter	0.128571	0
Laribacter	0.128571	0

Litorilituus	0.128571	0
Longibaculum	0.128571	0
Mahella	0.128571	0
Mesotoga	0.128571	0
Methylocaldum	0.128571	0
Methylomusa	0.128571	0
Microvirgula	0.128571	0
Nitrosospira	0.128571	0
Oceanimonas	0.128571	0
Oleiphilus	0.128571	0
Oryzomicrobium	0.128571	0
Paludibaculum	0.128571	0
Pannonibacter	0.128571	0
Parasaccharibacter	0.128571	0
Piscirickettsia	0.128571	0
Poriferisphaera	0.128571	0
Pragia	0.128571	0
Pseudoarcobacter	0.128571	0
Rhodoplanes	0.128571	0
Roseateles	0.128571	0
Roseivivax	0.128571	0
Rugosibacter	0.128571	0
Salaquimonas	0.128571	0
Sporisorium	0.128571	0
Sterolibacterium	0.128571	0
Syntrophus	0.128571	0
Tautonia	0.128571	0
Taylorella	0.128571	0
Thermobispora	0.128571	0
Thermocrinis	0.128571	0
Thermothielavioides	0.128571	0
Xylophilus	0.128571	0
Alkalilimnicola	0.114286	0
Aminobacterium	0.114286	0
Amylolactobacillus	0.114286	0
Anabaenopsis	0.114286	0
Aquabacter	0.114286	0
Azorhizobium	0.114286	0
Besnoitia	0.114286	0
Botrimarina	0.114286	0
Calditerrivibrio	0.114286	0
Candidatus Accumulibacter	0.114286	0
Candidatus Cloacimonas	0.114286	0
Candidatus Hepatoplasma	0.114286	0
Candidatus Portiera	0.114286	0

Candidatus Reidiella	0.114286	0
Candidatus Symbiobacter	0.114286	0
Chamaesiphon	0.114286	0
Chitinibacter	0.114286	0
Chroococcidiopsis	0.114286	0
Cnuibacter	0.114286	0
Coxiella	0.114286	0
Crinalium	0.114286	0
Crocospaera	0.114286	0
Cyanobacterium	0.114286	0
Defluviimonas	0.114286	0
Defluviitoga	0.114286	0
Desulfarculus	0.114286	0
Desulfatibacillum	0.114286	0
Desulfurispirillum	0.114286	0
Dysosmobacter	0.114286	0
Faecalibacillus	0.114286	0
Hartmannibacter	0.114286	0
Hoeflea	0.114286	0
Ketobacter	0.114286	0
Kineobactrum	0.114286	0
Kluyveromyces	0.114286	0
Koleobacter	0.114286	0
Leptospirillum	0.114286	0
Leptothrix	0.114286	0
Melittangium	0.114286	0
Methylibium	0.114286	0
Micavibrio	0.114286	0
Mitsuaria	0.114286	0
Nitrobacter	0.114286	0
Nitrospirillum	0.114286	0
Niveibacterium	0.114286	0
Niveispirillum	0.114286	0
Orbus	0.114286	0
Oxalobacter	0.114286	0
Paenalcaligenes	0.114286	0
Paremcibacter	0.114286	0
Pelagibacterium	0.114286	0
Pseudopuniceibacterium	0.114286	0
Pyrococcus	0.114286	0
Pyxidicoccus	0.114286	0
Rubrivivax	0.114286	0
Schlegelella	0.114286	0
Sediminibacillus	0.114286	0
Shimwellia	0.114286	0

Sinimarinibacterium	0.114286	0
Snodgrassella	0.114286	0
Sphaerotilus	0.114286	0
Spongiibacter	0.114286	0
Sulfurimicrobium	0.114286	0
Symmachiella	0.114286	0
Tepidiforma	0.114286	0
Thalassiosira	0.114286	0
Thermanaeromonas	0.114286	0
Tistrella	0.114286	0
Turicimonas	0.114286	0
Vogesella	0.114286	0
Acetohalobium	0.1	0
Agarilytica	0.1	0
Amphritea	0.1	0
Ancylobacter	0.1	0
Bingvirus	0.1	0
Candidatus Desulforudis	0.1	0
Candidatus Halobonum	0.1	0
Candidatus Nitrosocosmicus	0.1	0
Candidatus Solibacter	0.1	0
Candidatus Vampirococcus	0.1	0
Chloroherpeton	0.1	0
Chondrocystis	0.1	0
Desulfoglaeba	0.1	0
Dokdonella	0.1	0
Dolichospermum	0.1	0
Dongshaea	0.1	0
Egicoccus	0.1	0
Euzebya	0.1	0
Fluoribacter	0.1	0
Gallaecimonas	0.1	0
Geitlerinema	0.1	0
Gemmobacter	0.1	0
Haloarcula	0.1	0
Halosimplex	0.1	0
Halothece	0.1	0
Halothiobacillus	0.1	0
Jinshanibacter	0.1	0
Kaistia	0.1	0
Methanobacterium	0.1	0
Methanosphaera	0.1	0
Methylacidimicrobium	0.1	0
Methylacidiphilum	0.1	0
Methylobacillus	0.1	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Methylovirgula	0.1	0
Mycoavidus	0.1	0
Natronomonas	0.1	0
Nitratifractor	0.1	0
Nodularia	0.1	0
Paludisphaera	0.1	0
Paraoceanicella	0.1	0
Parasphingopyxis	0.1	0
Pelolinea	0.1	0
Phycisphaera	0.1	0
Pirellula	0.1	0
Pirellulimonas	0.1	0
Pochonia	0.1	0
Rivularia	0.1	0
Sagittula	0.1	0
Scandinavium	0.1	0
Spartinivivinus	0.1	0
Stigmatella	0.1	0
Sulfolobus	0.1	0
Sulfuricella	0.1	0
Talaromyces	0.1	0
Tatlockia	0.1	0
Thermobacillus	0.1	0
Thermochromatium	0.1	0
Thermothelomyces	0.1	0
Thermovibrio	0.1	0
Thioclava	0.1	0
Thiohalobacter	0.1	0
Thiolapillus	0.1	0
Youhaiella	0.1	0
Acaryochloris	0.085714	0
Acidibrevibacterium	0.085714	0
Acidothermus	0.085714	0
Aeoliella	0.085714	0
Afipia	0.085714	0
Anaplasma	0.085714	0
Aquabacterium	0.085714	0
Aquirhabdus	0.085714	0
Asaia	0.085714	0
Bacterioplanes	0.085714	0
Beijerinckia	0.085714	0
Biomaibacter	0.085714	0
Brasilonema	0.085714	0
Breoghania	0.085714	0
Candidatus Amoebophilus	0.085714	0

Candidatus Cytomitobacter	0.085714	0
Candidatus Fukatsuia	0.085714	0
Candidatus Moranella	0.085714	0
Candidatus Nitrosoglobus	0.085714	0
Candidatus Nitrotoga	0.085714	0
Candidatus Tachikawaea	0.085714	0
Carboxydocella	0.085714	0
Cercospora	0.085714	0
Chania	0.085714	0
Chitinimonas	0.085714	0
Chitinolyticbacter	0.085714	0
Chromohalobacter	0.085714	0
Chthonomonas	0.085714	0
Commensalibacter	0.085714	0
Coralimargarita	0.085714	0
Cryptosporidium	0.085714	0
Desulfobacterium	0.085714	0
Effusibacillus	0.085714	0
Endozoicomonas	0.085714	0
Ephemeropterocola	0.085714	0
Eremothecium	0.085714	0
Ewingella	0.085714	0
Fimbriimonas	0.085714	0
Frateuria	0.085714	0
Frigoriglobus	0.085714	0
Georhizobium	0.085714	0
Gibbsiella	0.085714	0
Gluconacetobacter	0.085714	0
Granulosicoccus	0.085714	0
Gyruella	0.085714	0
Haematobacter	0.085714	0
Hpunavirus	0.085714	0
Hydrogenobaculum	0.085714	0
Immundisolibacter	0.085714	0
Indioceanicola	0.085714	0
Kozakia	0.085714	0
Lacimicrobium	0.085714	0
Lawsonia	0.085714	0
Lentilitoribacter	0.085714	0
Marivivens	0.085714	0
Nissabacter	0.085714	0
Nordella	0.085714	0
Obesumbacterium	0.085714	0
Oblitimonas	0.085714	0
Orientia	0.085714	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Pararhodospirillum	0.085714	0
Paraurantiacibacter	0.085714	0
Parvibaculum	0.085714	0
Paucibacter	0.085714	0
Pelagibaca	0.085714	0
Pelobacter	0.085714	0
Pseudescherichia	0.085714	0
Pyruvatibacter	0.085714	0
Rhodopirellula	0.085714	0
Saliniradius	0.085714	0
Salinisphaera	0.085714	0
Schizosaccharomyces	0.085714	0
Sedimenticola	0.085714	0
Sulfuricurvum	0.085714	0
Sulfuriflexus	0.085714	0
Sulfuritortus	0.085714	0
Syntrophomonas	0.085714	0
Thalassococcus	0.085714	0
Thermotomaculum	0.085714	0
Thiocystis	0.085714	0
Thiosulfativibrio	0.085714	0
Torulaspora	0.085714	0
Trichormus	0.085714	0
Trypanosoma	0.085714	0
Variibacter	0.085714	0
Verrucosispora	0.085714	0
Zobellella	0.085714	0
Alphapapillomavirus	0.071429	0.092308
Acetilactobacillus	0.071429	0
Acetomicrobium	0.071429	0
Acidiphilium	0.071429	0
Acidisarcina	0.071429	0
Aerosticca	0.071429	0
Ammonifex	0.071429	0
Aquisalimonas	0.071429	0
Arenimonas	0.071429	0
Blastomonas	0.071429	0
Caenibius	0.071429	0
Caldimicrobium	0.071429	0
Caldisericum	0.071429	0
Candidatus Atelocyanobacterium	0.071429	0
Candidatus Deianiraea	0.071429	0
Candidatus Midichloria	0.071429	0
Candidatus Nitrosacidococcus	0.071429	0
Candidatus Nucleicultrix	0.071429	0

Candidatus Profftella	0.071429	0
Candidatus Protochlamydia	0.071429	0
Candidatus Vesicomysocius	0.071429	0
Caulifigura	0.071429	0
Chloracidobacterium	0.071429	0
Ciceribacter	0.071429	0
Citromicrobium	0.071429	0
Cognatishimia	0.071429	0
Cohaesibacter	0.071429	0
Crenobacter	0.071429	0
Deefgea	0.071429	0
Defluviicoccus	0.071429	0
Desulfobacca	0.071429	0
Desulfocapsa	0.071429	0
Desulfolutivibrio	0.071429	0
Desulfurivibrio	0.071429	0
Desulfurobacterium	0.071429	0
Egibacter	0.071429	0
Euhalotheca	0.071429	0
Flexistipes	0.071429	0
Gallionella	0.071429	0
Gloeocapsa	0.071429	0
Halomicronema	0.071429	0
Halorussus	0.071429	0
Halotalea	0.071429	0
Hirschia	0.071429	0
Hydrocarboniclastica	0.071429	0
Inhella	0.071429	0
Isosphaera	0.071429	0
Kerstersia	0.071429	0
Leminorella	0.071429	0
Limihaloglobus	0.071429	0
Limnobaculum	0.071429	0
Litoricola	0.071429	0
Magnetospira	0.071429	0
Maoricimonas	0.071429	0
Maricaulis	0.071429	0
Marichromatium	0.071429	0
Marinicauda	0.071429	0
Marinovum	0.071429	0
Maritalea	0.071429	0
Melaminivora	0.071429	0
Methyloligella	0.071429	0
Methyloversatilis	0.071429	0
Monaibacterium	0.071429	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Mycetohabitans	0.071429	0
Natrinema	0.071429	0
Neokomagataea	0.071429	0
Nitrosophilus	0.071429	0
Novosphingopyxis	0.071429	0
Oceanicoccus	0.071429	0
Oceanicola	0.071429	0
Oxynema	0.071429	0
Parasedimentitalea	0.071429	0
Persephonella	0.071429	0
Planktomarina	0.071429	0
Pleurocapsa	0.071429	0
Pseudohongiella	0.071429	0
Pseudorhodobacter	0.071429	0
Pseudorhodoplanes	0.071429	0
Pseudovibrio	0.071429	0
Rhizorhabdus	0.071429	0
Rhodomicrobium	0.071429	0
Rippkaea	0.071429	0
Sideroxydans	0.071429	0
Simkania	0.071429	0
Simplexvirus	0.071429	0
Sphaerobacter	0.071429	0
Steroidobacter	0.071429	0
Sulfuritalea	0.071429	0
Tabrizicola	0.071429	0
Tateyamaia	0.071429	0
Thermoleptolyngbya	0.071429	0
Thermosediminibacter	0.071429	0
Thermosulfidibacter	0.071429	0
Thiobacillus	0.071429	0
Truepera	0.071429	0
Yoonia	0.071429	0
Zymobacter	0.071429	0
Acidimicrobium	0.057143	0
Allosphingosinicella	0.057143	0
Archaeoglobus	0.057143	0
Arthrospira	0.057143	0
Bermanella	0.057143	0
Boseongicola	0.057143	0
Brevefilum	0.057143	0
Caldichromatium	0.057143	0
Caldilinea	0.057143	0
Candidatus Babela	0.057143	0
Candidatus Carsonella	0.057143	0

Candidatus Endolissoclinum	0.057143	0
Candidatus Fokinia	0.057143	0
Candidatus Hodgkinia	0.057143	0
Candidatus Ishikawaella	0.057143	0
Candidatus Puniceispirillum	0.057143	0
Candidatus Riesia	0.057143	0
Candidatus Ruthia	0.057143	0
Candidatus Sneabacter	0.057143	0
Casimicrobium	0.057143	0
Cyanotheca	0.057143	0
Dactylococcopsis	0.057143	0
Debaryomyces	0.057143	0
Dissulfurispira	0.057143	0
Enterovibrio	0.057143	0
Fluviibacter	0.057143	0
Fromanvirus	0.057143	0
Fuerstia	0.057143	0
Gillianvirus	0.057143	0
Gordonvirus	0.057143	0
Halapricum	0.057143	0
Ilumatobacter	0.057143	0
Jannaschia	0.057143	0
Lichenicola	0.057143	0
Lichenihabitans	0.057143	0
Lignipirellula	0.057143	0
Mariniblastus	0.057143	0
Marinithermus	0.057143	0
Methanocella	0.057143	0
Methanococcoides	0.057143	0
Methylovulum	0.057143	0
Mimivirus	0.057143	0
Noviherbaspirillum	0.057143	0
Parachlamydia	0.057143	0
Parasphingorhabdus	0.057143	0
Parvularcula	0.057143	0
Petrotoga	0.057143	0
Polymorphum	0.057143	0
Pulveribacter	0.057143	0
Reinekea	0.057143	0
Roseibacterium	0.057143	0
Roseicitreum	0.057143	0
Simiduia	0.057143	0
Sphaerospermopsis	0.057143	0
Staphylospora	0.057143	0
Thalassobius	0.057143	0

Thermobaculum	0.057143	0
Thermosulfurimonas	0.057143	0
Tolumonas	0.057143	0
Tolypothrix	0.057143	0
Turneriella	0.057143	0
Wigglesworthia	0.057143	0
Woeseia	0.057143	0
Zetavirus	0.057143	0
Zhihengliuella	0.057143	0
Zhongshania	0.057143	0
Zymoseptoria	0.057143	0
Alienimonas	0.042857	0
Antarctobacter	0.042857	0
Aquifex	0.042857	0
Aureliella	0.042857	0
Bombella	0.042857	0
Brettanomyces	0.042857	0
Brevirhabdus	0.042857	0
Calycomorphotria	0.042857	0
Candidatus Chazhemtobacterium	0.042857	0
Candidatus Fonsibacter	0.042857	0
Candidatus Kuenenia	0.042857	0
Candidatus Nesciobacter	0.042857	0
Candidatus Nitrosocaldus	0.042857	0
Candidatus Prometheoarchaeum	0.042857	0
Candidatus Xiphinematobacter	0.042857	0
Cetia	0.042857	0
Cheoctovirus	0.042857	0
Coprothermobacter	0.042857	0
Desulfomonile	0.042857	0
Ferrovibrio	0.042857	0
Gaiavirus	0.042857	0
Halobacterium	0.042857	0
Halocynthiibacter	0.042857	0
Halogeometricum	0.042857	0
Halomicrobium	0.042857	0
Halorhabdus	0.042857	0
Halorientalis	0.042857	0
Halovivax	0.042857	0
Hydrogenobacter	0.042857	0
Kordiimonas	0.042857	0
Methanolobus	0.042857	0
Mudcatvirus	0.042857	0
Neochlamydia	0.042857	0
Oricola	0.042857	0

Permianibacter	0.042857	0
Phaeodactylum	0.042857	0
Pistricoccus	0.042857	0
Planctopirus	0.042857	0
Polystyrenella	0.042857	0
Pontimonas	0.042857	0
Pontivivens	0.042857	0
Rubinisphaera	0.042857	0
Salinicola	0.042857	0
Sandaracinobacter	0.042857	0
Scapunavirus	0.042857	0
Simplicispira	0.042857	0
Thermodesulfatator	0.042857	0
Thermogutta	0.042857	0
Tropheryma	0.042857	0
Ustilago	0.042857	0
Waddlia	0.042857	0
Yarrowia	0.042857	0
Zygosaccharomyces	0.042857	0
Zygotorulaspota	0.042857	0
Acidiferrobacter	0.028571	0
Actibacterium	0.028571	0
Agarivorans	0.028571	0
Baoshanvirus	0.028571	0
Bigelowiella	0.028571	0
Candidatus Annandia	0.028571	0
Candidatus Azoamicus	0.028571	0
Candidatus Bipolaricaulis	0.028571	0
Candidatus Gullanella	0.028571	0
Candidatus Hoaglandella	0.028571	0
Candidatus Paracaedibacter	0.028571	0
Candidatus Phaeomarinobacter	0.028571	0
Candidatus Phycorickettsia	0.028571	0
Candidatus Walczuchella	0.028571	0
Capripoxvirus	0.028571	0
Cryptomonas	0.028571	0
Cylindrospermopsis	0.028571	0
Cymopoleiavirus	0.028571	0
Derbicusvirus	0.028571	0
Dinoroseobacter	0.028571	0
Encephalitozoon	0.028571	0
Ferriphaselus	0.028571	0
Gloeomargarita	0.028571	0
Guillardia	0.028571	0
Haematospirillum	0.028571	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Halobiforma	0.028571	0
Halopiger	0.028571	0
Halog quadratum	0.028571	0
Hippea	0.028571	0
Magnetococcus	0.028571	0
Maribius	0.028571	0
Metakosakonia	0.028571	0
Metallosphaera	0.028571	0
Methanoculleus	0.028571	0
Methanoplanus	0.028571	0
Methanothermobacter	0.028571	0
Microchaete	0.028571	0
Microcoleus	0.028571	0
Mosigvirus	0.028571	0
Nakaseomyces	0.028571	0
Neorickettsia	0.028571	0
Nitrosarchaeum	0.028571	0
Obolenskvirus	0.028571	0
Oecophyllibacter	0.028571	0
Oneupvirus	0.028571	0
Pajaroellobacter	0.028571	0
Pandoravirus	0.028571	0
Pelagerythrobacter	0.028571	0
Pelagovum	0.028571	0
Pikminvirus	0.028571	0
Polerovirus	0.028571	0
Profundibacter	0.028571	0
Pseudooceanicola	0.028571	0
Raphidiopsis	0.028571	0
Rickettsiella	0.028571	0
Roseitalea	0.028571	0
Segniliparus	0.028571	0
Sextaecvirus	0.028571	0
Shigella	0.028571	0
Silicimonas	0.028571	0
Skunavirus	0.028571	0
Staphylothermus	0.028571	0
Sulfurisphaera	0.028571	0
Syntrophothermus	0.028571	0
Terricaulis	0.028571	0
Thalassoglobus	0.028571	0
Thermosulfuriphilus	0.028571	0
Thiosulfatimonas	0.028571	0
Tuwongella	0.028571	0
Lymphocryptovirus	0.014286	0.2

Acidilobus	0.014286	0
Aeropyrum	0.014286	0
Alcyoneusvirus	0.014286	0
Alphatorquevirus	0.014286	0
Athalassotoga	0.014286	0
Avunavirus	0.014286	0
Ayakvirus	0.014286	0
Baikalvirus	0.014286	0
Betaentomopoxvirus	0.014286	0
Betatorquevirus	0.014286	0
Bremerella	0.014286	0
Caldisphaera	0.014286	0
Candidatus Doolittlea	0.014286	0
Candidatus Methanoplasma	0.014286	0
Candidatus Mikella	0.014286	0
Candidatus Nitrosomarinus	0.014286	0
Candidatus Nitrosotenuis	0.014286	0
Candidatus Purcelliella	0.014286	0
Certrevirus	0.014286	0
Chlorovirus	0.014286	0
Coccolithovirus	0.014286	0
Crateriforma	0.014286	0
Cyprinivirus	0.014286	0
Drulisvirus	0.014286	0
Eneladusvirus	0.014286	0
Entomobacter	0.014286	0
Epibacterium	0.014286	0
Gamaleyavirus	0.014286	0
Gammaretrovirus	0.014286	0
Glycocaulis	0.014286	0
Guyparkeria	0.014286	0
Halalkaliarchaeum	0.014286	0
Halalkalicoccus	0.014286	0
Halarchaeum	0.014286	0
Harbinvirus	0.014286	0
Ilzatvirus	0.014286	0
Jasminevirus	0.014286	0
Jedunavirus	0.014286	0
Jiaodavirus	0.014286	0
Johnsonvirus	0.014286	0
Kagunavirus	0.014286	0
Karamvirus	0.014286	0
Kayvirus	0.014286	0
Komagataella	0.014286	0
Koutsourovirus	0.014286	0

Lachancea	0.014286	0
Lederbergvirus	0.014286	0
Lightbulbvirus	0.014286	0
Limnobacter	0.014286	0
Lymphocystivirus	0.014286	0
Mastadenovirus	0.014286	0
Methanohalophilus	0.014286	0
Methanomassiliicoccus	0.014286	0
Methanoregula	0.014286	0
Methanothermococcus	0.014286	0
Methanotorris	0.014286	0
Natrarchaeobaculum	0.014286	0
Natronococcus	0.014286	0
Natronorubrum	0.014286	0
Nipunavirus	0.014286	0
Nitrososphaera	0.014286	0
Ochrobactrum	0.014286	0
Orthopoxvirus	0.014286	0
Orthotospovirus	0.014286	0
Papirusvirus	0.014286	0
Paradevosia	0.014286	0
Peatvirus	0.014286	0
Phistoryvirus	0.014286	0
Pichia	0.014286	0
Pseudohalocynthiibacter	0.014286	0
Qingshengfaniella	0.014286	0
Rhodobaca	0.014286	0
Roufvirus	0.014286	0
Salinarchaeum	0.014286	0
Scheffersomyces	0.014286	0
Sepunavirus	0.014286	0
Slopekvirus	0.014286	0
Taipeivirus	0.014286	0
Tardibacter	0.014286	0
Tegunavirus	0.014286	0
Tequatrovirus	0.014286	0
Vulcanisaeta	0.014286	0
Wifcevirus	0.014286	0
Wphvirus	0.014286	0
Atopobium	0	0.6
Granulicatella	0	0.523077
Oribacterium	0	0.461538
Solobacterium	0	0.353846
Peptostreptococcus	0	0.138462
Anaeroglobus	0	0.076923

Catonella	0	0.076923
Alloscardovia	0	0.030769
Eggerthia	0	0.030769
Hallella	0	0.015385
Johnsonella	0	0.015385
Shuttleworthia	0	0.015385
Stenoxybacter	0	0.015385
unclassified Anelloviridae	0	0.015385
unclassified Lachnospiraceae	0	0.015385
Orthohepadnavirus	0	0.011765
unclassified Myoviridae	0	0.011765
unclassified Siphoviridae	0	0.023529
unclassified Saccharibacteria	0	0.047059
uncultured_bacterium	0	0.011765
unclassified Bacteroidetes	0	0.023529
unclassified Candidatus Absconditabacteria	0	0.047059

PHD THESIS

APPENDICES

NICOLE PEDRO

Table S3. Comparison of the genera detected in both methods per sample by applying a threshold of abundance to the WGS results of 0.01%, 0.1%, 0.5% and 1%.

Project ID	Threshold abundance above 1%			Threshold abundance above 0.5%			Threshold abundance above 0.1%			Threshold abundance above 0.01%		
	Genera detected from both methods	Genera detected only on WGS	%Match	Genera detected from both methods	Genera detected only on WGS	%Match	Genera detected from both methods	Genera detected only on WGS	%Match	Genera detected from both methods	Genera detected only on WGS	%Match
PNG104	11	1	91.67	13	1	92.86	15	10	60.00	21	39	35.00
PNG174	11	0	100.00	12	0	100.00	15	3	83.33	17	22	43.59
PNG18	8	0	100.00	11	0	100.00	14	5	73.68	16	31	34.04
PNG201	9	0	100.00	13	1	92.86	14	5	73.68	15	32	31.91
PNG221	11	0	100.00	13	0	100.00	13	11	54.17	14	43	24.56
PNG24	11	1	91.67	14	2	87.50	15	9	62.50	19	38	33.33
PNG304	14	0	100.00	15	0	100.00	16	8	66.67	20	45	30.77
PNG327	15	0	100.00	17	3	85.00	18	15	54.55	21	52	28.77
PNG53	13	0	100.00	14	2	87.50	15	12	55.56	18	33	35.29
PNG61	11	1	91.67	12	4	75.00	12	14	46.15	16	50	24.24
PNG68	12	2	85.71	13	7	65.00	15	16	48.39	15	46	24.59
PNG70	13	1	92.86	13	7	65.00	13	20	39.39	14	56	20.00
PNG79	14	2	87.50	14	4	77.78	15	13	53.57	17	48	26.15
PNG365	12	1	92.31	14	3	82.35	14	11	56.00	14	30	31.82
PNG366	12	1	92.31	13	1	92.86	14	10	58.33	18	37	32.73
PNG367	14	1	93.33	14	3	82.35	14	13	51.85	15	37	28.85
PNG370	11	0	100.00	13	0	100.00	15	4	78.95	18	26	40.91
PNG379	7	1	87.50	9	2	81.82	12	6	66.67	14	25	35.90
PNG382	13	0	100.00	16	1	94.12	19	5	79.17	22	39	36.07
PNG384	12	0	100.00	14	1	93.33	17	8	68.00	17	24	41.46
PNG385	16	2	88.89	19	4	82.61	20	20	50.00	22	78	22.00
PNG388	10	0	100.00	13	4	76.47	15	19	44.12	17	54	23.94
PNG391	13	2	86.67	16	2	88.89	18	11	62.07	19	45	29.69
PNG393	11	0	100.00	14	0	100.00	14	8	63.64	15	46	24.59
PNG394	13	0	100.00	17	1	94.44	18	9	66.67	20	52	27.78
PNG395	8	0	100.00	11	0	100.00	17	4	80.95	20	25	44.44
PNG399	10	0	100.00	14	1	93.33	14	13	51.85	15	37	28.85
PNG401	9	2	81.82	11	3	78.57	12	13	48.00	13	37	26.00
PNG402	9	3	75.00	11	5	68.75	11	13	45.83	13	39	25.00
PNG404	10	2	83.33	10	6	62.50	11	18	37.93	13	52	20.00
PNG405	11	0	100.00	14	3	82.35	15	10	60.00	18	52	25.71
PNG124	12	0	100.00	14	2	87.50	14	7	66.67	16	29	35.56
PNG129	12	0	100.00	15	2	88.24	16	12	57.14	20	54	27.03
PNG130	15	1	93.75	16	1	94.12	16	10	61.54	17	24	41.46

PNG131	10	0	100.00	14	0	100.00	18	10	64.29	20	46	30.30
PNG132	8	1	88.89	12	5	70.59	12	14	46.15	16	50	24.24
PNG228	10	2	83.33	10	4	71.43	10	11	47.62	11	35	23.91
PNG234	14	0	100.00	14	0	100.00	15	10	60.00	21	47	30.88
PNG235	9	0	100.00	15	2	88.24	15	11	57.69	18	48	27.27
PNG239	11	2	84.62	11	7	61.11	12	20	37.50	14	73	16.09
PNG511	14	1	93.33	14	3	82.35	19	14	57.58	21	57	26.92
PNG513	10	5	66.67	12	11	52.17	13	19	40.63	13	55	19.12
PNG515	10	4	71.43	12	9	57.14	14	16	46.67	14	55	20.29
PNG517	12	4	75.00	13	6	68.42	13	15	46.43	16	48	25.00
PNG518	11	0	100.00	16	3	84.21	18	13	58.06	21	41	33.87
PNG519	7	1	87.50	13	1	92.86	16	9	64.00	20	31	39.22
PNG521	13	2	86.67	16	4	80.00	16	7	69.57	19	40	32.20
PNG523	13	1	92.86	15	2	88.24	16	10	61.54	18	45	28.57
PNG524	13	2	86.67	15	4	78.95	19	19	50.00	22	71	23.66
PNG526	12	4	75.00	15	5	75.00	17	15	53.13	20	49	28.99
PNG527	16	2	88.89	17	4	80.95	22	14	61.11	28	59	32.18
PNG529	11	1	91.67	15	4	78.95	15	10	60.00	16	46	25.81
PNG531	12	0	100.00	14	3	82.35	19	8	70.37	20	34	37.04
PNG532	10	0	100.00	10	0	100.00	15	6	71.43	20	46	30.30
PNG535	12	5	70.59	12	9	57.14	13	18	41.94	13	44	22.81
PNG538	14	1	93.33	18	2	90.00	18	10	64.29	19	37	33.93
PNG540	8	3	72.73	12	5	70.59	12	13	48.00	16	37	30.19
PNG541	11	3	78.57	12	5	70.59	13	14	48.15	17	44	27.87
PNG10	11	1	91.67	13	3	81.25	14	10	58.33	18	44	29.03
PNG100	12	1	92.31	13	4	76.47	13	11	54.17	14	36	28.00
PNG101	9	1	90.00	15	2	88.24	18	8	69.23	19	43	30.65
PNG220	10	0	100.00	12	2	85.71	14	7	66.67	17	36	32.08
PNG436	12	0	100.00	17	1	94.44	20	10	66.67	23	53	30.26
PNG45	9	0	100.00	11	0	100.00	18	6	75.00	19	34	35.85
PNG50	10	0	100.00	10	1	90.91	10	8	55.56	13	24	35.14
	AVERAGE		91.96	AVERAGE		83.87	AVERAGE		58.44	AVERAGE		29.75

PHD THESIS

APPENDICES

NICOLE PEDRO

Table S4. Qualitative results indicating the number of samples for each group in which the different genus were detected.

Genera		Port Moresby	Simbu province	East Sepik province	Western province	East highlanders living in Port Moresby
Bacteria	Abiotrophia	1	5	0	4	4
	Actinobacillus	1	0	1	1	0
	Actinomyces	19	19	13	16	10
	Aggregatibacter	0	1	5	2	2
	Alloprevotella	5	4	3	7	2
	Alloscardovia	0	2	0	0	1
	Anaeroglobus	1	3	2	2	0
	Atopobium	10	12	4	13	10
	Bifidobacterium	0	2	1	2	1
	Bulleidia	2	1	2	2	0
	Campylobacter	19	16	15	17	10
	Capnocytophaga	7	11	2	12	8
	Cardiobacterium	0	0	1	0	0
	Catonella	4	2	1	1	1
	Corynebacterium	0	5	1	4	2
	Cryptobacterium	0	0	0	1	0
	Desulfovibrio	0	0	0	1	0
	Dialister	3	4	0	6	4
	Eggerthia	0	1	0	1	0
	Eikenella	1	0	0	1	0
	Enterococcus	0	0	1	0	0
	Escherichia	0	1	0	2	0
	Eubacterium	9	14	3	14	7
	Filifactor	4	2	1	3	1
	Fusobacterium	19	20	17	18	11
	Gemella	10	14	2	17	9
	Granulicatella	6	14	4	10	9
	Haemophilus	19	19	16	18	11
	Hallella	0	0	1	0	0
	Histophilus	0	1	0	0	0
	Johnsonella	0	0	0	1	0
	Kingella	2	5	0	1	3
	Lachnoanaerobaculum	0	0	0	1	0
	Lactobacillus	0	1	0	1	0
	Lancefieldella	3	1	0	0	0
	Lautropia	5	7	3	3	4
	Leptotrichia	15	15	11	12	8
	Limosilactobacillus	0	1	0	0	0
	Mannheimia	0	0	2	0	0
	Megasphaera	11	11	5	10	10

	Mogibacterium	0	1	0	0	0
	Mycoplasma	11	11	4	13	6
	Mycoplasma mopsis	0	0	0	2	0
	Neisseria	18	20	16	18	11
	Oribacterium	11	8	1	10	5
	Parascardovia	0	1	0	0	0
	Parvimonas	0	0	0	1	1
	Pasteurella	0	2	0	0	1
	Peptostreptococcus	1	2	0	6	2
	Phocaeicola	1	0	1	1	0
	Porphyromonas	17	20	16	18	11
	Prevotella	19	20	17	18	11
	Propionibacterium	0	2	0	0	0
	Pseudoleptotrichia	1	4	1	5	0
	Rodentibacter	1	1	1	0	0
	Rothia	19	20	17	17	11
	Scardovia	0	2	0	0	1
	Schaalia	2	2	0	0	2
	Selenomonas	13	16	11	12	7
	Shuttleworthia	0	0	0	1	0
	Simonsiella	0	0	0	1	2
	Solobacterium	7	6	2	8	5
	Stenoxymbacter	0	0	0	0	1
	Streptococcus	19	20	17	18	11
	Tannerella	9	7	7	10	4
	Treponema	19	20	17	17	10
	unclassified Bacteroidetes	1	0	0	1	0
	unclassified Candidatus Absconditabacteria	1	1	0	0	2
	unclassified Lachnospiraceae	0	0	0	0	2
	unclassified Saccharibacteria	0	0	0	4	0
	uncultured_bacterium	0	0	0	1	0
	Veillonella	19	20	17	18	11
	Vibrio	4	5	14	3	2
	Wolinella	0	0	1	0	0
	Yersinia	0	0	0	0	1
Virus	Alphapapillomavirus	0	1	0	4	2
	Cepunavirus	0	2	0	2	0
	Lymphocryptovirus	3	6	3	5	3
	Orthohepadnavirus	1	0	0	0	0
	Pahexavirus	2	0	0	0	0
	Roseolovirus	0	0	0	0	2
	unclassified Anelloviridae	1	0	0	0	0
	unclassified Myoviridae	0	1	0	0	0
	unclassified Siphoviridae	0	1	0	1	0

PHD THESIS

APPENDICES

NICOLE PEDRO

Table S5. Relative abundance of each of the top Phyla and Genera for the five groups in this study (Port Moresby, Simbu province, East Sepik province, Western province, East highlanders living in Port Moresby) and in comparison with other published populations (Australian Aboriginal Children, Batwa Pygmies, Philipines HG, Philippine TF, Western controls).

	Port Moresby	Simbu province	East Sepik province	Western province	East highlanders living in Port Moresby	Australian Aboriginal Children ¹	Batwa Pygmies ²	Philipines HG ³	Philipines TF ³	Western Controls ^{3,4}
Proteobacteria	25.05%	28.15%	35.54%	25.65%	35.70%	31.00%	46.60%	28.82%	25.74%	30.55%
<i>Neisseria</i>	12.16%	14.96%	17.14%	11.73%	19.24%	14.00%	4.60%	14.19%	10.73%	8.81%
<i>Haemophilus</i>	8.15%	8.73%	13.83%	5.03%	11.01%	12.00%	3.56%	7.93%	9.93%	14.21%
Bacteroidetes	27.56%	19.12%	19.11%	26.12%	18.37%	26.00%	18.50%	19.97%	25.22%	18.75%
<i>Prevotella</i>	22.51%	15.72%	14.19%	20.88%	15.71%	19.00%	12.89%	15.14%	20.81%	16.29%
<i>Porphyromonas</i>	3.06%	1.66%	2.74%	2.75%	0.96%	3.00%	3.96%	1.59%	1.93%	0.16%
Firmicutes	20.67%	30.57%	27.33%	23.67%	21.78%	25.00%	20.80%	35.06%	35.35%	39.12%
<i>Veillonella</i>	9.05%	6.14%	8.13%	5.04%	9.07%	6.00%	3.44%	8.57%	11.75%	11.86%
<i>Streptococcus</i>	8.56%	20.20%	15.86%	13.41%	10.27%	9.00%	12.80%	20.41%	19.57%	23.00%
Actinobacteria	18.50%	16.19%	10.71%	14.92%	19.13%	12.00%	3.50%	7.70%	7.60%	8.01%
<i>Rothia</i>	9.94%	9.15%	6.62%	6.26%	9.21%	8.00%	2.59%	4.75%	4.99%	3.15%
<i>Actinomyces</i>	2.00%	2.84%	1.57%	2.08%	3.19%	2.00%	0.50%	0.50%	0.33%	1.86%
Fusobacteria	5.79%	4.27%	4.28%	6.68%	4.00%	6.00%	6.50%	6.58%	4.75%	2.90%
<i>Fusobacterium</i>	4.01%	3.16%	2.20%	4.94%	3.00%	4.00%	2.62%	3.87%	3.08%	1.99%
<i>Leptotrichia</i>	1.76%	1.08%	2.06%	1.73%	0.98%	2.00%	3.52%	2.48%	1.55%	0.84%
Spirochaetes	1.99%	1.40%	2.50%	2.37%	1.00%	1.00%	0.70%	1.53%	1.05%	0.28%
<i>Treponema</i>	1.99%	1.40%	2.50%	2.37%	0.66%	1.00%	0.72%	1.49%	1.02%	0.27%

¹Handsley-Davis, Matilda et al. 2021. 'Biocultural Drivers of Salivary Microbiota in Australian Aboriginal and Torres Strait Islander Children', *Frontiers in Oral Health*, 2:641328

²Nasidze, Ivan et al. 2011. 'High Diversity of the Saliva Microbiome in Batwa Pygmies', *PLoS One*, 6(8):e23352.

³Lassalle, Florent et al. 2018. 'Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet', *Mol Ecol*, 27: 182–95

⁴Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, 486(7402): 215–221.

Appendix C - Supplementary information from Paper III

Host diversity versus environmental influence on the Papua New Guinean oral microbiome

Nicole Pedro et al.

The KRAKEN output files obtained from the WGS unmapped reads have been made available as Supporting files at: <https://cloud.i3s.up.pt/index.php/s/PHLEkHddPPSp5dJ>.

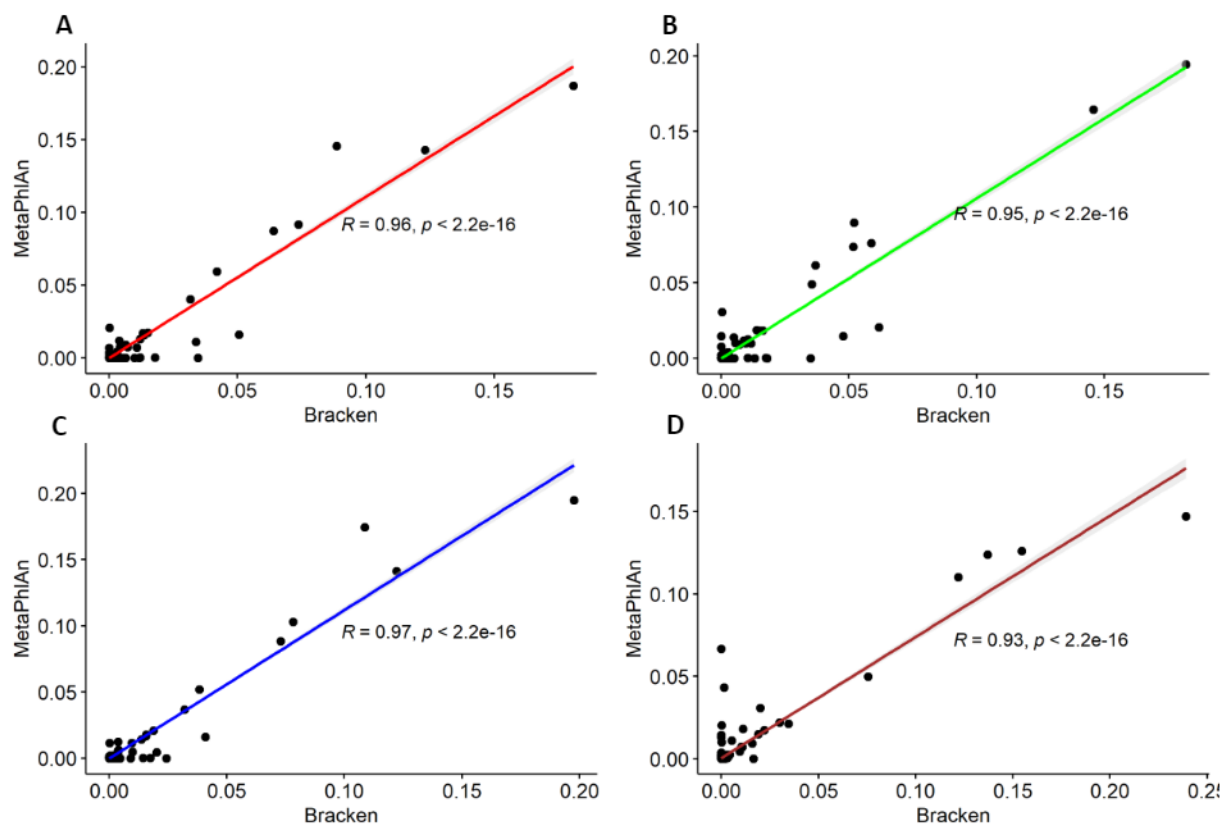


Figure S1. Correlation between bacterial genera abundances inferred through Bracken 2 and MetaPhlAn algorithms, in a) the total 57 samples from PNG, Philippines and Western (in red); B) the samples from PNG (in green); C) the samples from the Philippines (in blue); and D) the samples from the Western population (in brown).

Table S1. Demographic information (Age, Gender, Province and Region of origin of the father, Sampling place) for all the individuals in this study, the number of reads mapped to the oral microbiome inference from the WGS and indication of which samples were used for the HUMAnN analysis.

Sample ID	Age	Gender	Province father	Region father	Sampling place	Reads mapping to the oral microbiome	HUMAnN
B00IEV8	26	M	Bougainville	Islands	Port Moresby	7106882	NO
B00IEV9	26	F	Bougainville	Islands	Port Moresby	8356964	NO
B00IEVA	25	F	Bougainville	Islands	Port Moresby	8473522	NO
B00IEVB	67	M	Bougainville	Islands	Port Moresby	39500120	YES
B00IEVD	24	F	Enga	Highlands West	Port Moresby	21535774	NO
B00IEVE	29	M	Central	Lowlands Southeast	Port Moresby	18735568	YES
B00IEVF	22	F	National Capital	Lowlands Southeast	Port Moresby	12608104	YES
B00IEVG	34	M	Central	Lowlands Southeast	Port Moresby	9394486	YES
B00IEVH	25	F	Central	Lowlands Southeast	Port Moresby	3193091	NO
B00IEVI	27	F	East Sepik	Lowlands North	Port Moresby	12278363	NO
B00IEVJ	40	M	East Sepik	Lowlands North	Port Moresby	21822736	NO
B00IEVK	26	F	East Sepik	Lowlands North	Port Moresby	31253266	NO
B00IEVL	32	M	East Sepik	Lowlands North	Port Moresby	17637876	NO
B00IEVM	25	F	Eastern Highlands	Highlands East	Port Moresby	8006138	NO
B00IEVN	25	F	Eastern Highlands	Highlands East	Port Moresby	14988392	NO
B00IEVO	28	M	Enga	Highlands West	Port Moresby	10286788	NO
B00IEVQ	27	M	Enga	Highlands West	Port Moresby	24025036	NO
B00IEVR	33	M	Gulf	Lowlands South	Port Moresby	17534844	YES
B00IEVW	25	M	Jiwaka	Highlands West	Port Moresby	10055815	NO
B00IEW0	N/A	F	Madang	Highlands East	Port Moresby	14978452	NO
B00IEW1	23	F	Milne Bay	Lowlands Southeast	Port Moresby	14762603	NO
B00IEW3	25	F	Milne Bay	Lowlands Southeast	Port Moresby	6748874	NO
B00IEW5	29	M	Milne Bay	Lowlands Southeast	Port Moresby	7504904	NO
B00IEW8	26	M	Oro	Lowlands Southeast	Port Moresby	20013343	YES
B00IEW9	26	F	Chimbu	Highlands East	Port Moresby	17327901	NO
B00IEWA	27	F	Chimbu	Highlands East	Port Moresby	10931124	NO
B00IEWD	24	F	Southern Highlands	Highlands West	Port Moresby	16809327	NO

B00IEWG	22	M	Western Highlands	Highlands West	Port Moresby	17832788	NO
B00IEWH	23	F	Eastern Highlands	Highlands East	Port Moresby	15098695	NO
B00IEWJ	24	M	Western Highlands	Highlands West	Port Moresby	40994287	NO
B00IEWM	26	M	Bougainville	Islands	Port Moresby	17815044	YES
B00IEWP	24	F	National Capital	Lowlands Southeast	Port Moresby	11746048	YES
B00IEWR	25	M	Enga	Highlands West	Port Moresby	8238514	NO
B00IEWV	26	F	Western Highlands	Highlands West	Port Moresby	34109931	NO
B00IWTN	25	M	Western Highlands	Highlands West	Port Moresby	8620324	NO
B00IWTQ	23	M	Southern Highlands	Highlands West	Port Moresby	32702637	NO
B00IWTR	34	M	Gulf	Lowlands South	Port Moresby	7907356	NO
B00IWTT	47	M	Milne Bay	Lowlands Southeast	Port Moresby	28430281	YES
B00IWTV	45	M	Southern Highlands	Highlands West	Port Moresby	6798030	NO
B00IWTW	47	M	Milne Bay	Lowlands Southeast	Port Moresby	9431654	YES
B00IWTX	35	M	Madang	Highlands East	Port Moresby	17710574	NO
B00IWU1	22	M	Enga	Highlands West	Port Moresby	37008267	NO
B00IWU2	28	F	Hela	Highlands West	Port Moresby	15650620	NO
B00IWU3	27	M	Sandaun	Lowlands North	Port Moresby	12670425	NO
B00IWU5	25	F	Western Highlands	Highlands West	Port Moresby	3888640	NO
B00IWU7	23	F	Morobe	Highlands East	Port Moresby	11170845	NO
B00IWU8	27	F	Jiwaka	Highlands West	Port Moresby	5133175	NO
B00IWU9	24	F	Oro	Lowlands Southeast	Port Moresby	24525716	NO
B00IWUA	27	F	Western Highlands	Highlands West	Port Moresby	36924816	NO
B00IWUE	47	M	Chimbu	Highlands East	Mt Wilhelm	57637803	YES
B00IWUK	35	M	Jiwaka	Highlands West	Port Moresby	51596871	NO
C001KL2	53	M	Chimbu	Highlands East	Mt Wilhelm	7639700	YES
C001KL3	20	M	Jiwaka	Highlands West	Mt Wilhelm	3072433	NO
C001KL4	41	M	Chimbu	Highlands East	Mt Wilhelm	11159877	YES
C001KL7	55	M	Chimbu	Highlands East	Mt Wilhelm	10363142	YES
C001KL9	57	M	Chimbu	Highlands East	Mt Wilhelm	5985518	NO
C001KLA	67	M	Chimbu	Highlands East	Mt Wilhelm	3214358	NO
C001KLC	68	M	Chimbu	Highlands East	Mt Wilhelm	5929496	NO
C001KLD	67	M	Chimbu	Highlands East	Mt Wilhelm	4965862	NO

C001KLE	35	M	Chimbu	Highlands East	Mt Wilhelm	5749084	NO
C001KLF	50	M	Chimbu	Highlands East	Mt Wilhelm	4751051	NO
C001KLG	20	M	Chimbu	Highlands East	Mt Wilhelm	5773043	NO
C001KLH	22	M	Chimbu	Highlands East	Mt Wilhelm	6458120	NO
C001KLI	24	M	Chimbu	Highlands East	Mt Wilhelm	7386210	NO
C001KLK	19	M	Chimbu	Highlands East	Mt Wilhelm	7651313	YES
C001KLL	20	F	Eastern Highlands	Highlands East	Mt Wilhelm	6859029	NO
C001KLM	22	M	Chimbu	Highlands East	Mt Wilhelm	6036831	NO
C001KLN	22	M	Chimbu	Highlands East	Mt Wilhelm	2839090	NO
C001KLO	21	M	Chimbu	Highlands East	Mt Wilhelm	2504076	NO
C001KLP	20	M	Chimbu	Highlands East	Mt Wilhelm	4905224	NO
C001KLQ	23	M	Chimbu	Highlands East	Mt Wilhelm	5844593	NO
C001KLR	20	F	Chimbu	Highlands East	Mt Wilhelm	8509416	NO
C001KLS	20	F	Chimbu	Highlands East	Mt Wilhelm	5543500	NO
C001KLT	21	F	Eastern Highlands	Highlands East	Mt Wilhelm	9272791	NO
C001KLU	21	M	Chimbu	Highlands East	Mt Wilhelm	6771507	NO
C001KLV	28	M	Chimbu	Highlands East	Mt Wilhelm	15639730	NO
C001KLW	20	M	Chimbu	Highlands East	Mt Wilhelm	9246372	NO
C001KLX	20	M	Chimbu	Highlands East	Mt Wilhelm	13750693	YES
C001KLY	20	M	Chimbu	Highlands East	Mt Wilhelm	10444249	YES
C001KLZ	20	M	Chimbu	Highlands East	Mt Wilhelm	23823609	YES
C001KM0	21	M	Jiwaka	Highlands West	Mt Wilhelm	7803423	NO
C001KM2	45	M	Chimbu	Highlands East	Mt Wilhelm	6544953	NO
C001KM4	21	F	Chimbu	Highlands East	Mt Wilhelm	1867925	NO
C001KM5	20	F	Chimbu	Highlands East	Mt Wilhelm	5794742	NO
C001KM6	20	M	Chimbu	Highlands East	Mt Wilhelm	5791735	NO
C001KM8	48	M	Western Fly	Lowlands South	Daru	4576022	NO
C001KMA	27	M	Western Fly	Lowlands South	Daru	5917723	NO
C001KMC	24	M	Western Fly	Lowlands South	Daru	10581719	NO
C001KMD	35	M	Western Fly	Lowlands South	Daru	2263123	NO
C001KME	37	F	Western Fly	Lowlands South	Daru	8793751	NO
C001KMF	53	M	Western Fly	Lowlands South	Daru	11782565	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

C001KMG	26	F	Western Fly	Lowlands South	Daru	4312779	NO
C001KMH	44	F	Western Fly	Lowlands South	Daru	14617806	NO
C001KMI	44	M	Western Fly	Lowlands South	Daru	8782603	NO
C001KMJ	32	M	Western Fly	Lowlands South	Daru	1773461	NO
C001KMK	45	F	Western Fly	Lowlands South	Daru	2662828	NO
C001KML	40	M	Western Fly	Lowlands South	Daru	2372503	NO
C001KMN	55	M	Western Fly	Lowlands South	Daru	8882618	NO
C001KMO	48	M	Western Fly	Lowlands South	Daru	7470358	NO
C001KMP	67	M	South Fly	Lowlands South	Daru	2884316	NO
C001KMR	51	M	Western Fly	Lowlands South	Daru	7058785	NO
C001KMU	44	F	Western Fly	Lowlands South	Daru	6918805	NO
C001KMX	27	M	Western Fly	Lowlands South	Daru	8038730	NO
C001KMY	40	M	Western Fly	Lowlands South	Daru	6638801	NO
C001KMZ	58	M	Western Fly	Lowlands South	Daru	13674503	YES
C001KN0	62	M	Western Fly	Lowlands South	Daru	1595215	NO
C001KN2	40	M	Western Fly	Lowlands South	Daru	4658371	NO
C001KN3	47	M	Western Fly	Lowlands South	Daru	5586644	NO
C001KN4	20	F	Western Fly	Lowlands South	Daru	5233427	NO
C001KN6	26	M	Western Fly	Lowlands South	Daru	27270129	YES
C001KN7	42	M	Western Fly	Lowlands South	Daru	6758287	NO
C001KN8	39	M	Western Fly	Lowlands South	Daru	7584911	NO
C001KN9	18	M	Western Fly	Lowlands South	Daru	3132253	NO
C001KNA	45	M	Western Fly	Lowlands South	Daru	4727974	NO
C001KNB	44	M	Western Fly	Lowlands South	Daru	4572142	NO
C001KNC	32	M	Western Fly	Lowlands South	Daru	4347866	NO
C001KNE	51	M	Western Fly	Lowlands South	Daru	10416112	YES
C001KNG	24	M	Western Fly	Lowlands South	Daru	4196471	NO
C001KNH	53	M	Western Fly	Lowlands South	Daru	2332331	NO
C001KNJ	49	M	Western Fly	Lowlands South	Daru	3491069	NO
C001KNK	55	M	Western Fly	Lowlands South	Daru	15427684	NO
C001KNM	52	M	Chimbu	Highlands East	Mt Wilhelm	28943830	YES
C001KNN	54	M	Chimbu	Highlands East	Mt Wilhelm	3379436	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

C001KNO	67	M	Chimbu	Highlands East	Mt Wilhelm	2026406	NO
C001KNP	31	F	Chimbu	Highlands East	Mt Wilhelm	5474763	NO
C001KNS	64	M	Chimbu	Highlands East	Mt Wilhelm	29631779	YES
C001KNT	41	M	Chimbu	Highlands East	Mt Wilhelm	7739209	NO
C001PIT	44	M	Chimbu	Highlands East	Mt Wilhelm	21265490	NO
C001PIU	27	M	Chimbu	Highlands East	Mt Wilhelm	11876455	NO
C001PIV	21	M	Chimbu	Highlands East	Mt Wilhelm	8248840	NO
C001PIY	62	M	Western Fly	Lowlands South	Daru	11128356	YES
C001PJ5	41	M	Chimbu	Highlands East	Mt Wilhelm	7611723	NO
C001PJ6	67	M	Chimbu	Highlands East	Mt Wilhelm	2069807	NO
C001PJ7	32	M	Chimbu	Highlands East	Mt Wilhelm	4659398	NO
C002DAC	23	F	Bougainville	Islands	Port Moresby	14025528	NO
C002DAD	24	F	East New Britain	Islands	Port Moresby	14699722	NO
C002DAE	38	M	Chimbu	Highlands East	Port Moresby	9948803	NO
C002DAG	24	M	Western Highlands	Highlands West	Port Moresby	10695177	NO
C002DAI	25	M	Oro	Lowlands Southeast	Port Moresby	11324642	NO
C002DAK	25	M	Western Highlands	Highlands West	Port Moresby	24350468	NO
C002DAM	26	F	Central	Lowlands Southeast	Port Moresby	4008991	NO
C002DAN	23	F	East Sepik	Lowlands North	Port Moresby	2971181	NO
C002DAO	24	F	Central	Lowlands Southeast	Port Moresby	8770029	NO
C002DAP	20	M	West Sepik	Lowlands North	Port Moresby	9729579	NO
C002DAQ	25	F	East Sepik	Lowlands North	Port Moresby	9091717	NO
C002DAR	24	F	East New Britain	Islands	Port Moresby	27521173	NO
C002DAS	27	M	New Ireland	Islands	Port Moresby	5317508	NO
C002DAT	25	F	National Capital	Lowlands Southeast	Port Moresby	6040546	NO
C002DAU	22	M	Eastern Highlands	Highlands East	Port Moresby	19764082	NO
C002DAV	35	M	Southern Highlands	Highlands West	Daru	2269345	NO
C002DAW	24	M	West New Britain	Islands	Port Moresby	24930094	NO
C002DAX	27	F	New Ireland	Islands	Port Moresby	13563071	NO
C002DAZ	27	F	East New Britain	Islands	Port Moresby	1359568	NO
C002DB0	27	F	Eastern Highlands	Highlands East	Port Moresby	1326462	NO
C002DB2	25	F	Manus	Highlands East	Port Moresby	7278041	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

C002DB6	28	M	East New Britain	Islands	Port Moresby	7093340	NO
C002DB7	25	F	East New Britain	Islands	Port Moresby	3410111	NO
C002DB8	27	M	New Ireland	Islands	Port Moresby	3103867	NO
C002DB9	25	F	Central	Lowlands Southeast	Port Moresby	2838775	NO
C002DBA	27	M	Western Highlands	Highlands West	Port Moresby	8337329	NO
C002DBB	64	F	New Ireland	Islands	Port Moresby	26258626	NO
C002DBC	25	F	Hela	Highlands West	Port Moresby	4645011	NO
C002DBD	21	F	West Sepik	Lowlands North	Port Moresby	1710928	NO
C002DBE	46	F	West New Britain	Islands	Port Moresby	4515482	NO
C002DBF	27	M	Jiwaka	Highlands West	Port Moresby	4094986	NO
C002DBH	23	M	Eastern Highlands	Highlands East	Port Moresby	17041670	NO
C002DBI	21	F	East New Britain	Islands	Port Moresby	6823178	NO
C002DBK	45	M	Chimbu	Highlands East	Mt Wilhelm	12471321	NO
C002DBL	50	M	Chimbu	Highlands East	Mt Wilhelm	15691440	NO
C002DBN	44	M	Chimbu	Highlands East	Mt Wilhelm	2581801	NO
C002DBP	32	M	Chimbu	Highlands East	Mt Wilhelm	2147355	NO
C002DBQ	42	M	Chimbu	Highlands East	Mt Wilhelm	17022324	NO
C002DBS	55	M	Chimbu	Highlands East	Mt Wilhelm	7383836	NO
C002DBT	36	F	West Papua	Highlands West	Port Moresby	2960839	NO
C002DBV	N/A	F	Enga	Highlands West	Port Moresby	17575988	NO
C002DBW	20	M	Chimbu	Highlands East	Mt Wilhelm	1836341	NO
C002DBX	21	M	Chimbu	Highlands East	Mt Wilhelm	2784312	NO
C002DBY	38	M	Sandaun	Lowlands North	Port Moresby	9179920	NO
C002DBZ	22	F	Morobe	Highlands East	Port Moresby	1471638	NO
C002DC0	22	F	Eastern Highlands	Highlands East	Port Moresby	4824468	NO
C002DC1	20	F	New Ireland	Islands	Port Moresby	8567412	NO
C002DC2	21	M	West New Britain	Islands	Port Moresby	3313900	NO
C002DC3	77	M	Gulf	Lowlands South	Port Moresby	5833253	NO
C002DC4	22	F	East Sepik	Lowlands North	Port Moresby	7878171	NO
C002DC5	23	F	Western Highlands	Highlands West	Port Moresby	6380921	NO
C002DC6	19	F	Western Highlands	Highlands West	Port Moresby	5730685	NO
C002DC7	21	F	Hela	Highlands West	Port Moresby	9239033	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

C002DC8	23	F	Morobe	Highlands East	Port Moresby	7183619	NO
C002DC9	21	M	Chimbu	Highlands East	Port Moresby	12903487	NO
C002DCA	19	F	National Capital	Lowlands Southeast	Port Moresby	4096368	NO
C002DCB	19	M	National Capital	Lowlands Southeast	Port Moresby	2262820	NO
C002DCC	58	M	Chimbu	Highlands East	Mt Wilhelm	15538869	NO
C002DCE	51	F	Oro	Lowlands Southeast	Port Moresby	4588850	NO
C002DCF	23	M	Jiwaka	Highlands West	Port Moresby	10325063	NO
C002DCH	22	M	Central	Lowlands Southeast	Port Moresby	5516017	NO
C002DCI	24	M	East Sepik	Lowlands North	Port Moresby	3114760	NO
C002DCM	24	M	Chimbu	Highlands East	Port Moresby	6646020	NO
C002DCN	26	M	Central	Lowlands Southeast	Port Moresby	5787799	NO
C002DCO	22	M	Enga	Highlands West	Port Moresby	16077073	NO
C002DCQ	24	M	East Sepik	Lowlands North	Port Moresby	19127940	NO
C002DCR	N/A	F	Southern Highlands	Highlands West	Port Moresby	8746729	NO
C002DCS	22	M	Enga	Highlands West	Port Moresby	3281644	NO
C002DCV	22	M	Western Highlands	Highlands West	Port Moresby	7658220	NO
C002DCW	30	M	Chimbu	Highlands East	Mt Wilhelm	3072073	NO
C002DCY	26	N/A	East Sepik	Lowlands North	Port Moresby	9075437	NO
C002DCZ	83	M	Chimbu	Highlands East	Mt Wilhelm	7457051	NO
C002DD1	27	F	Madang	Highlands East	Port Moresby	11703752	NO
C002DD2	N/A	M	Milne Bay	Lowlands Southeast	Port Moresby	3521779	NO
C002DD6	64	M	Central	Lowlands Southeast	Port Moresby	7260028	NO
C002DDE	45	F	Central	Lowlands Southeast	Port Moresby	3137097	NO
C002DDF	24	F	Central	Lowlands Southeast	Port Moresby	11882160	NO
FD09254746	60	M	Western Fly	Lowlands South	Daru	28775945	YES
FD09254747	62	M	Western Fly	Lowlands South	Daru	3905117	NO
FD09254748	69	M	Western Fly	Lowlands South	Daru	17579386	YES
FD09254754	27	M	Western Fly	Lowlands South	Daru	6320403	NO
FD09254756	30	M	Western Fly	Lowlands South	Daru	9493427	NO
FD09254762	60	M	Western Fly	Lowlands South	Daru	8925625	NO
FD09254763	45	M	Western Fly	Lowlands South	Daru	8848455	YES
FD09254764	53	M	Western Fly	Lowlands South	Daru	14510605	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

FD09254770	64	F	Middle Fly	Lowlands South	Daru	1.49E+08	NO
FD09254771	52	M	Western Fly	Lowlands South	Daru	8779783	NO
FD09254772	25	F	Western Fly	Lowlands South	Daru	16245611	NO
FD09254777	37	M	Western Fly	Lowlands South	Daru	51854071	YES
FD09254778	77	M	Western Fly	Lowlands South	Daru	17953643	NO
FD09254779	43	M	Western Fly	Lowlands South	Daru	5358390	NO
FD09254780	27	M	Western Fly	Lowlands South	Daru	13436500	NO
FD09254785	53	M	Western Fly	Lowlands South	Daru	30219238	YES
FD09254786	49	M	Western Fly	Lowlands South	Daru	5360350	NO
FD09254787	59	M	Western Fly	Lowlands South	Daru	7443367	NO
FD09254788	67	M	Western Fly	Lowlands South	Daru	1256626	NO
FD09254793	67	M	Western Fly	Lowlands South	Daru	3447658	NO
FD09254794	65	M	Western Fly	Lowlands South	Daru	35032834	YES
FD09254795	66	M	Western Fly	Lowlands South	Daru	4517455	NO
FD09254796	22	M	Western Fly	Lowlands South	Daru	2385764	NO
FD09254801	52	M	Western Fly	Lowlands South	Daru	917415	NO
FD09254802	49	M	Western Fly	Lowlands South	Daru	6265293	NO
FD09254803	61	M	Western Fly	Lowlands South	Daru	3945350	NO
FD09254804	29	M	Western Fly	Lowlands South	Daru	4923177	NO
FD09254809	63	M	Western Fly	Lowlands South	Daru	2318346	NO
FD09254810	63	M	Western Fly	Lowlands South	Daru	9168703	NO
FD09254811	65	M	Western Fly	Lowlands South	Daru	4070376	NO
FD09254812	57	M	Western Fly	Lowlands South	Daru	4753777	NO
FD09254817	55	M	Western Fly	Lowlands South	Daru	14044413	NO
FD09254818	60	M	Middle Fly	Lowlands South	Daru	1723212	NO
FD09254819	44	F	National Capital	Lowlands Southeast	Daru	2335175	NO
FD09254820	47	M	Western Fly	Lowlands South	Daru	3972020	NO
FD09254825	21	M	Western Fly	Lowlands South	Daru	2283808	NO
FD09254826	43	M	Western Fly	Lowlands South	Daru	5520659	NO
FD09254827	27	F	Western Fly	Lowlands South	Daru	2622223	NO
FD09254833	63	M	Western Fly	Lowlands South	Daru	1704031	NO
FD09254834	38	M	Western Fly	Lowlands South	Daru	3256744	NO

PHD THESIS

APPENDICES

NICOLE PEDRO

FD09254835	27	M	Western Fly	Lowlands South	Daru	1832461	NO
------------	----	---	-------------	----------------	------	---------	----

Table S2- Statistically significant (p-value < 5 x 10-08) miQTLs in the PNG groups. For each SNP, information is provided for the reference (R) and alternative (A) alleles, the associated genus, genomic region (gene or intergenic region), the p-value and the genotype frequencies for European, East Asian and African population groups obtained from the 1000 genomes data and for PNG calculated from the genetic data of this study.

Genus	Gene	SNP	R	A	p-value	Genotype Frequency			
						MAF_EUR	MAF_EAS	MAF_AFR	MAF_PNG
Klebsiella	CDH4	rs11908491	T	C	6.53 x 10 ⁻¹⁰	T T: 0.680	T T: 0.667	T T: 0.573	T T: 0.806
						C T: 0.274	C T: 0.292	C T: 0.369	C T: 0.155
						C C: 0.046	C C: 0.042	C C: 0.057	C C: 0.040
Klebsiella	CDH4	rs62198854	A	G	7.86 x 10 ⁻¹⁰	A A: 0.680	A A: 0.665	A A: 0.573	A A: 0.805
						A G: 0.274	A G: 0.294	A G: 0.369	A G: 0.155
						G G: 0.046	G G: 0.042	G G: 0.057	G G: 0.040
Klebsiella	CDH4	rs28407743	G	A	8.22 x 10 ⁻¹⁰	G G: 0.042	G G: 0.032	G G: 0.141	G G: 0.032
						A G: 0.272	A G: 0.282	A G: 0.477	A G: 0.143
						A A: 0.686	A A: 0.687	A A: 0.383	A A: 0.825
Klebsiella	CDH4	rs4239687	A	C	8.47 x 10 ⁻¹⁰	A A: 0.678	A A: 0.659	A A: 0.433	A A: 0.806
						A C: 0.276	A C: 0.298	A C: 0.452	A C: 0.155
						C C: 0.046	C C: 0.044	C C: 0.115	C C: 0.040
Klebsiella	CDH4	rs12624814	C	G	1.09 x 10 ⁻⁰⁹	C C: 0.676	C C: 0.667	C C: 0.543	C C: 0.802
						C G: 0.278	C G: 0.290	C G: 0.384	C G: 0.159
						G G: 0.046	G G: 0.044	G G: 0.073	G G: 0.040
Klebsiella	CDH4	rs62201898	A	G	1.16 x 10 ⁻⁰⁹	A A: 0.674	A A: 0.657	A A: 0.460	A A: 0.810
						A G: 0.278	A G: 0.298	A G: 0.434	A G: 0.151
						G G: 0.048	G G: 0.046	G G: 0.106	G G: 0.040
Klebsiella	CDH4	rs60828599	C	A	1.19 x 10 ⁻⁰⁹	C C: 0.596	C C: 0.657	C C: 0.419	C C: 0.798
						A C: 0.330	A C: 0.300	A C: 0.451	A C: 0.162
						A A: 0.074	A A: 0.044	A A: 0.130	A A: 0.040
Klebsiella	CDH4	rs12624833	G	A	1.27 x 10 ⁻⁰⁹	G G: 0.676	G G: 0.667	G G: 0.543	G G: 0.806
						A G: 0.278	A G: 0.290	A G: 0.386	A G: 0.155
						A A: 0.046	A A: 0.044	A A: 0.071	A A: 0.040
Klebsiella	CDH4	rs62198855	A	G	1.35 x 10 ⁻⁰⁹	A A: 0.680	A A: 0.659	A A: 0.570	A A: 0.802

						A G: 0.274 G G: 0.046	A G: 0.298 G G: 0.044	A G: 0.371 G G: 0.059	A G:0.159 G G:0.040
<i>Klebsiella</i>	<i>CDH4</i>	rs12624815	C	T	1.70 x 10 ⁻⁰⁹	C C: 0.676 C T: 0.278 T T: 0.046	C C: 0.657 C T: 0.298 T T: 0.046	C C: 0.461 C T: 0.433 T T: 0.106	C C:0.798 C T:0.163 T T:0.040
<i>Klebsiella</i>	<i>CDH4</i>	rs62201875	C	G	1.82 x 10 ⁻⁰⁹	C C: 0.674 C G: 0.280 G G: 0.046	C C: 0.667 C G: 0.290 G G: 0.044	C C: 0.543 C G: 0.386 G G: 0.071	C C:0.797 C G:0.163 G G:0.040
<i>Klebsiella</i>	<i>CDH4</i>	rs28401694	G	A	2.41 x 10 ⁻⁰⁹	G G: 0.062 A G: 0.362 A A: 0.577	G G: 0.032 A G: 0.284 A A: 0.685	G G: 0.189 A G: 0.486 A A: 0.325	G G:0.036 A G:0.139 A A:0.825
<i>Klebsiella</i>	<i>CDH4</i>	rs7261942	C	A	2.54 x 10 ⁻⁰⁹	C C: 0.670 A C: 0.284 A A: 0.046	C C: 0.611 A C: 0.333 A A: 0.056	C C: 0.277 A C: 0.510 A A: 0.213	C C:0.794 A C:0.167 A A:0.040
<i>Klebsiella</i>	<i>CDH4</i>	rs28415571	C	T	1.04 x 10 ⁻⁰⁸	C C: 0.042 C T: 0.270 T T: 0.688	C C: 0.032 C T: 0.278 T T: 0.690	C C: 0.144 C T: 0.469 T T: 0.387	C C:0.044 C T:0.151 T T:0.806
<i>Klebsiella</i>	<i>GALM</i>	rs13024705	A	G	1.71 x 10 ⁻⁰⁸	A A: 0.417 A G: 0.447 G G: 0.135	A A: 0.349 A G: 0.486 G G: 0.165	A A: 0.595 A G: 0.371 G G: 0.035	A A:0.821 A G:0.155 G G:0.024
<i>Klebsiella</i>	Intergenic Variant	rs10467529	A	G	2.22 x 10 ⁻⁰⁸	A A: 0.789 A G: 0.199 G G: 0.012	A A: 0.988 A G: 0.012	A A: 0.605 A G: 0.342 G G: 0.053	A A:0.790 A G:0.187 G G:0.024
<i>Gemella</i>	Intergenic Variant	rs7283882	C	G	2.07 x 10 ⁻⁰⁸	C C: 0.284 C G: 0.471 G G: 0.245	C C: 0.466 C G: 0.431 G G: 0.103	C C: 0.165 C G: 0.505 G G: 0.330	C C:0.123 C G:0.460 G G:0.417
<i>Schaalia</i>	Intergenic Variant	rs7182861	G	A	4.43 x 10 ⁻⁰⁸	G G: 0.791 A G: 0.197 A A: 0.012	G G: 0.998 A G: 0.002	G G: 0.301 A G: 0.511 A A: 0.188	G G:0.615 A G:0.337 A A:0.048
<i>Schaalia</i>	Intergenic Variant	rs7174379	T	C	4.59 x 10 ⁻⁰⁸	T T: 0.791 C T: 0.197 C C: 0.012	T T: 0.998 C T: 0.002	T T: 0.265 C T: 0.511 C C: 0.224	T T:0.623 C T:0.325 C C:0.052
<i>Salmonella</i>	Intron Variant	rs4315878	A	G	5.01 x 10 ⁻⁰⁸	A A: 0.014 A G: 0.241 G G: 0.746	A A: 0.016 A G: 0.145 G G: 0.839	A A: 0.151 A G: 0.496 G G: 0.352	A A: 0.028 A G: 0.198 G G: 0.774