

Genomic approaches to stop the spread of bacterial infectious diseases

Andrew Harris Buultjens

<https://orcid.org/0000-0002-5984-1328>

Doctor of Philosophy

March 2020

Department of Microbiology and Immunology
Faculty of Medicine, Dentistry and Health Sciences
The University of Melbourne

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

Synopsis

This thesis investigated the utility of genomics in Australian public health microbiology through case studies of three bacterial pathogens of public health importance: *Enterococcus faecium*, for genomic characterization of an emerging, health care-associated bacterial pathogen; *Legionella pneumophila* for developing high resolution genomic-based source tracking tools for public health investigations; and *Mycobacterium ulcerans*, for explaining epidemiological trends and assessing transmission pathways through genomic insights. Outcomes from this research reinforce the emergence of microbial genomics as the cornerstone technology of public health microbiology.

Abstract

The advent of affordable whole genome sequencing has been a disruptive force that is changing the biological sciences, with public health microbiology being no exception. As microbial genomics is a new frontier for public health microbiology, a substantial amount of development and validation research is required. Genomic-based investigations in response to bacterial public health problems are gaining traction, but optimal approaches to bioinformatics and data analysis are still being developed. In order to help address these needs, this thesis investigated the utility of microbial genomics for Australian public health microbiology through three case studies of three distinct bacterial pathogens.

Hospital adapted pathogens pose a significant health threat through their ability to rapidly emerge with adaptive phenotypes that permit them to circumvent infection control practices. In the first results chapter, comparative bacterial population genomics was used to characterise and better understand the genomic changes that have accompanied and driven the emergence of a novel and highly hospital associated lineage of *Enterococcus faecium* called ST796. These analyses revealed that the novel strain belonged to a highly hospital adapted lineage and shared a close evolutionary origin with another newly emerged hospital associated lineage of *E. faecium*. Complete assembly and analysis of a representative ST796 genome identified several key genomic events that were distinctive in the newly emerged lineage.

Investigations of bacterial disease outbreaks caused by environmental pathogens have traditionally relied upon insights delivered through epidemiology supported by pathogen genotyping methods. Genotyping can be critical for identifying the source of an outbreak, but methods vary in their resolving power and genotyping data can be difficult to compare between investigations. In the second results chapter, a genomics-based machine learning approach was

developed and used to predict the source of clinical *Legionella pneumophila* isolates using genomic patterns obtained prospectively through environmental sampling of bacteria. A proof-of-concept method was developed for assigning outbreak sources without using phylogenetic trees. Notably too, this study demonstrated phylogenetics can be misleading when assessing *L. pneumophila* environmental origins. The predictive statistical genomic method presented here has utility as a powerful tool for use alongside conventional *L. pneumophila* disease outbreak epidemiological investigations.

Understanding a pathogen's mode of transmission and infectious reservoir is a key goal in order to control the spread of disease, as such knowledge permits deployment of evidence-based strategies to manage outbreaks and prevent future occurrences. In the third results chapter, genomics was used to conduct high-resolution, micro-epidemiology population analyses on clinical isolates of another environmental bacterial pathogen, *Mycobacterium ulcerans*, the cause of Buruli ulcer. Bacterial population genomics and phylogeographic modeling suggested a westward migration of the pathogen across south east Australia that aligned well with the disease epidemiology. In addition, demographic models inferred from the genomic data provided a plausible explanation for a recent increase in cases of Buruli ulcer in south-eastern Australia. These insights demonstrate that *M. ulcerans* can be introduced to a new environment and then expands, rather than the awakening of a quiescent pathogen reservoir. Such discoveries inform our understanding of Buruli ulcer transmission and control.

Collectively, these studies further demonstrate that microbial genomics significantly improves our understanding of the transmission dynamics behind clinically important pathogens over standard epidemiological investigations. This research demonstrates the advantages of the implementation of microbial genomics in public health microbiology.

Declaration

This is to certify that:

- This thesis, entitled “Genomic approaches to stop the spread of bacterial infectious diseases”, comprises only my original work towards this PhD except where otherwise indicated;
- Due acknowledgement has been made in the text to all other material used;
- The thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

.....

Mr Andrew Buultjens

Department of Microbiology and Immunology, The University of Melbourne
at the Peter Doherty Institute for Infection and Immunity

Preface

Chapter overview

This PhD thesis presents three published studies with the overarching theme of investigating the utility of microbial genomics in public health microbiology. Provided below is a summary of the chapters that are included in this thesis. Each summary indicates the contributions of collaborators for each of the chapters.

Chapter 1: Literature review

An outline of the relevant background literature, fundamental concepts and aims of the project.

Contributions:	
Andrew H. Buultjens	Writing and revision
Timothy P. Stinear	Critical review
Sacha J. Pidot	Critical review

Chapter 2: Genomic characterisation of novel bacterial pathogens

This chapter includes the following publication:

Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*.

Andrew H. Buultjens, Margaret M.C. Lam, Susan Ballard, Ian R. Monk, Andrew A. Mahony, Elizabeth A. Grabsch, M. Lindsay Grayson, Stanley Pang, Geoffrey W. Coombs, J. Owen Robinson, Torsten Seemann, Paul D.R. Johnson, Benjamin P. Howden, Timothy P. Stinear. *PeerJ*, 2017 January; 5: e2916. doi: <https://doi.org/10.7717/peerj.2916>.

Contributions:	
Andrew H. Buultjens	Study design, coordination, bioinformatic analysis, data interpretation, writing and revision
Margaret M. C. Lam	Isolate characterisation and critical review
Susan Ballard	Isolate characterisation and critical review
Ian R. Monk	Isolate characterisation
Andrew A. Mahony	Isolate characterisation and critical review
Elizabeth A. Grabsch	Provided bacterial strains and data interpretation
M. Lindsay Grayson	Provided bacterial strains and data interpretation
Stanley Pang	Provided bacterial strains and data interpretation
Geoffrey W. Coombs	Provided bacterial strains and data interpretation
J. Owen Robinson	Provided bacterial strains and data interpretation
Torsten Seemann	Study design, bioinformatic analysis
Paul D.R. Johnson	Data interpretation, critical review
Benjamin P. Howden	Provided bacterial strains and data interpretation
Timothy P. Stinear	Study design, data interpretation, critical review

Chapter 3: Developing useful genomic based tools for public health investigations

This chapter includes the following publication:

A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks.

Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha

J. Pidot, Lucinda Franklin, Torsten Seemann, Anders Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. *J Appl Environ Microbiol.* 2017 August; 83(21): e01482-17. doi: 10.1128/AEM.01482-17.

Contributions:	
Andrew H. Buultjens	Study design, coordination, bioinformatic analysis, data interpretation, writing and revision
Kyra Y. L. Chua	Study design, isolate characterisation and critical review
Sarah L. Baines	Data interpretation, critical review
Jason Kwong	Data interpretation, critical review
Wei Gao	Data interpretation, critical review
Zoe Cutcher	Provided bacterial strains and data interpretation
Stuart Adcock	Provided bacterial strains and data interpretation
Susan Ballard	Provided bacterial strains and data interpretation
Mark B. Schultz	Data interpretation, critical review
Takehiro Tomita	Provided bacterial strains and data interpretation
Nela Subasinghe	Provided bacterial strains and data interpretation
Glen P. Carter	Provided bacterial strains and data interpretation
Sacha J. Pidot	Data interpretation, critical review
Lucinda Franklin	Provided bacterial strains and data interpretation
Torsten Seemann	Study design, bioinformatic analysis, and critical review

Anders Gonçalves Da Silva	Study design, bioinformatic analysis, and critical review
Benjamin P. Howden	Provided bacterial strains and data interpretation
Timothy P. Stinear	Study design, data interpretation, critical review

Chapter 4: Explaining epidemiological trends through genomic insights

This chapter includes the following publication:

Comparative genomics shows *Mycobacterium ulcerans* migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia.

Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. *J Appl Environ Microbiol.* 2018 April; 2; 84(8). doi: 10.1128/AEM.02612-17.

Contributions:

Andrew H. Buultjens	Study design, coordination, bioinformatic analysis, data interpretation, writing and revision
Koen Vandelannoote	Study design, bioinformatic analysis, data interpretation, and critical review
Conor J. Meehan	Bioinformatic analysis, data interpretation, and critical review
Miriam Eddyani	Provided bacterial strains and data interpretation
Bouke C. de Jong	Provided bacterial strains and data interpretation

Janet A. M. Fyfe	Provided bacterial strains, data interpretation and critical review
Maria Globan	Provided bacterial strains and data interpretation
Nicholas J. Tobias	Provided bacterial strains and data interpretation
Jessica L. Porter	Provided bacterial strains and data interpretation
Takehiro Tomita	Provided bacterial strains and data interpretation
Ee Laine Tay	Provided bacterial strains and data interpretation
Torsten Seemann	Study design, data interpretation, bioinformatic analysis
Benjamin P. Howden	Provided bacterial strains and data interpretation
Paul D. R. Johnson	Study design and critical review
Timothy P. Stinear	Study design, data interpretation, critical review

Chapter 5: Discussion and conclusions

An overview of the existing limitations, knowledge gaps, with prospective directions for application and further research.

Contributions:	
Andrew H. Buultjens	Writing and revision
Timothy P. Stinear	Critical review
Sacha J. Pidot	Critical review

Funding declaration

Mr Andrew Buultjens is supported by an Australian Postgraduate Award Scholarship from the Australian Government.

List of publications arising during candidature

1. Andrew H. Buultjens, Margaret M.C. Lam, Susan Ballard, Ian R. Monk, Andrew A. Mahony, Elizabeth A. Grabsch, M. Lindsay Grayson, Stanley Pang, Geoffrey W. Coombs, J. Owen Robinson, Torsten Seemann, Paul D.R. Johnson, Benjamin P. Howden, Timothy P. Stinear . Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. *PeerJ*, 2017 January; 5: e2916. doi: <https://doi.org/10.7717/peerj.2916>.
2. Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks. *J Appl Environ Microbiol*. 2017 August, 83(21): e01482-17. doi: 10.1128/AEM.01482-17.
3. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Comparative genomics shows *Mycobacterium ulcerans* migration

- and expansion has preceded the rise of Buruli ulcer in south-eastern Australia. *J Appl Environ Microbiol*. 2018 April; 2: 84(8). doi: 10.1128/AEM.02612-17.
4. Nana Ama Amissah, Andrew H. Buultjens (joint first author), Anthony Ablordey, Lieke van Dam, Ampomah Opoku-Ware, Sarah L. Baines, Dieter Bulach, Caitlin S. Tetteh, Isaac Prah, Tjip S. van der Werf, Alexander W. Friedrich, Torsten Seemann, Jan Maarten van Dijl, Ymkje Stienstra, Timothy P. Stinear and John W. Rossen. Methicillin resistant *Staphylococcus aureus* transmission in a Ghanaian burn unit: the importance of active surveillance in resource-limited settings. *Front. Microbiol*, 2017; 8: 1906. doi: <https://doi.org/10.3389/fmicb.2017.01906>.
 5. Andrew A. Mahony, Andrew H. Buultjens (joint first author), Susan A. Ballard, Elizabeth A. Grabsch, Shirley Xie, Torsten Seemann, Rhonda L. Stuart, Despina Kotsanas, Allen Cheng, Helen Heffernan, Sally A. Roberts, Geoffrey W. Coombs, Narin Bak, John K. Ferguson, Glen C. Carter, Benjamin P. Howden, Timothy P. Stinear and Paul D. R. Johnson. Vancomycin-resistant *Enterococcus faecium* sequence type 796 - rapid international dissemination of a new epidemic clone. *Antimicrob Resist Infect Control*, 2018; 7. doi: 10.1186/s13756-018-0335-z.
 6. Sacha J. Pidot, Wei Gao, Andrew H. Buultjens (joint first author), Ian R. Monk, Romain Guerillot, Glen P. Carter, Jean Y. H. Lee, Margaret M. C. Lam, M. Lindsay Grayson, Susan A. Ballard, Andrew A. Mahony, Elizabeth A. Grabsch, Despina Kotsanas, Tony M. Korman, Geoffrey W. Coombs, J. Owen Robinson, Anders Gonçalves da Silva, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Increasing tolerance of

- hospital *Enterococcus faecium* to handwash alcohols. *Sci Transl Med*. 2018; Aug 1; 10 (452). doi: 10.1126/scitranslmed.aar6115.
7. Glen P. Carter, Andrew H. Buultjens, Susan A. Ballard, Sarah L. Baines, Takehiro Tomita, Janet Strachan, Paul D. R. Johnson, John K. Ferguson, Torsten Seemann, Timothy P. Stinear, Benjamin P. Howden. Emergence of endemic MLST non-typeable vancomycin-resistant *Enterococcus faecium*. *J Antimicrob Chemother*. 2016; 71(12): 3367-3371. doi: <https://doi.org/10.1093/jac/dkw314>.
 8. Grace Kpeli, Andrew H. Buultjens, Stefano Giulieri, Evelyn Owusu-Mireku, Samuel Y. Aboagye, Sarah L. Baines, Torsten Seemann, Dieter Bulach, Anders Gonçalves da Silva, Ian R. Monk, Benjamin P. Howden, Gerd Pluschke, Dorothy Yeboah-Manu, Timothy Stinear. Genomic analysis of ST88 community-acquired methicillin resistant *Staphylococcus aureus* in Ghana. *PeerJ*. 5; e3047. doi: 10.7717/peerj.3047.
 9. Grace Kpeli, Isaac Otchere, Araceli Lamelas, Andrew Buultjens, Dieter Bulach, Sarah Baines, Torsten Seemann, Stefano Giulieri, Zuliehatu Nakobu, Samuel Aboagye, Evelyn Owusu-Mireku, Emelia Danso, Julia Hauser, Vladimira Hinic, Gerd Pluschke, Timothy Stinear, Dorothy Yeboah-Manu. Drug resistance and genetic profile of bacterial species associated with Buruli ulcer wound infections in two districts of Ghana. *BMJ Global Health* 2017; 2: A50. doi: 10.1136/bmjgh-2016-000260.133.
 10. Daniel P. O'Brien, James W. Wynne, Andrew H. Buultjens, Wojtek P. Michalski, Timothy P. Stinear, N. Deborah Friedman, Andrew Hughes, and Eugene Athan. Exposure Risk for

- Infection and Lack of Human-to-Human Transmission of *Mycobacterium ulcerans* Disease, Australia. *Emerg Infect Dis.* 23(5); 2017. doi: 10.3201/eid2305.160809.
11. G. Kpeli, I. Darko Otchere, A. Lamelas, A.H. Buultjens, D. Bulach, S.L. Baines, T. Seemann, S. Giulieri, Z. Nakobu, S.Y. Aboagye, E. Owusu-Mireku, G. Pluschke, T.P. Stinear and D. Yeboah-Manu. Possible healthcare-associated transmission as a cause of secondary infection and population structure of *Staphylococcus aureus* isolates from two wound treatment centres in Ghana. *New Microbes New Infect.* 2016. 13: 92–101. doi: 10.1016/j.nmni.2016.07.001.
 12. Anthony S. Ablordey, Koen Vandelanoot, Isaac A. Frimpong, Evans K. Ahoritor, Nana Ama Amissah, Miriam Eddyani, Lies Durnez, Françoise Portaels, Bouke C. de Jong, Herwig Leirs, Jessica L. Porter, Kirstie M. Mangas, Margaret M. C. Lam, Andrew Buultjens, Torsten Seemann, Nicholas J. Tobias and Timothy P. Stinear. Whole genome comparisons suggest random distribution of *Mycobacterium ulcerans* genotypes in a Buruli ulcer endemic region of Ghana. *PLoS Negl Trop Dis.* 2015; 9: 3. doi: 10.1371/journal.pntd.0003681.
 13. Md. Nazim Uddin, Randall William Robinson, Andrew Buultjens, Md. Abdullah Yousuf Al Harun, Shahana Haque Shampa. Role of allelopathy of *Phragmites australis* in its invasion processes. *J Exp Mar Bio Ecol.* 2017. 486: 237-244. doi: <https://doi.org/10.1016/j.jembe.2016.10.016>.
 14. Jose Manuel Guerra Laso, Teresa Nebreda Mayoral, Juan José Palacios Gutiérrez, Élia Samaniego González, Bárbara Rodríguez Martín, Leticia Barrio Rodríguez, Nieves Alonso

Orcajo, Elena Magaz García, Timothy Stinear, Andrew H. Buultjens, Jose Antonio Pérez Molina. Buruli ulcers in a Spanish aid worker after a stay in Peru. *Int J Infect Dis*. 2019. 78: 99–102. doi: <https://doi.org/10.1016/j.ijid.2018.10.012>.

15. Koen Vandelannoote, Miriam Eddyani, Andrew Buultjens & Timothy P. Stinear. Population Genomics and Molecular Epidemiology of *Mycobacterium ulcerans*. Chapter in Buruli Ulcer – 1st Edition. 2019. doi: <https://doi.org/10.1007/978-3-030-11114-4>.

List of conference communications arising during candidature

Abstracts presented in international and national meetings

1. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Using genomics to understand the ecology of *Mycobacterium ulcerans*. *Bio21 Genomics Symposium*, Melbourne, Australia 2014 (oral presentation).
2. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Exploring the evolution of *Mycobacterium ulcerans* and emergence of Buruli ulcer in Australia with whole genome sequencing. *WHO Meeting on Buruli ulcer Control and Research*, Geneva, Switzerland 2015 (oral presentation).

3. Sacha J. Pidot, Wei Gao, Andrew H. Buultjens (joint first author), Ian R. Monk, Romain Guerillot, Glen P. Carter, Jean Y. H. Lee, Margaret M. C. Lam, M. Lindsay Grayson, Susan A. Ballard, Andrew A. Mahony, Elizabeth A. Grabsch, Despina Kotsanas, Tony M. Korman, Geoffrey W. Coombs, J. Owen Robinson, Anders Gonçalves da Silva, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Increasing tolerance of hospital *Enterococcus faecium* to hand-wash alcohols. *BacPath 13*, San Remo, Australia 2015 (poster).
4. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Exploring the evolution of *Mycobacterium ulcerans* and emergence of Buruli ulcer in Australia with whole genome sequencing. *Combine Student Symposium*, Sydney, Australia 2015 (poster).
5. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Exploring the evolution of pathogens that cause emerging diseases with whole genome sequencing. *SPASIM Postgraduate Retreat*, Lanefield, Australia 2015 (oral presentation).
6. Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders

- Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. Exploiting extremes of *Legionella pneumophila* genomic diversity for accurate source attribution. *ABACBS Conference*, Brisbane, Australia 2016 (oral presentation).
7. Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. Exploiting *Legionella pneumophila* genomic diversity for accurate source attribution. *Combine Student Symposium*, Brisbane, Australia 2016 (poster).
8. Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. Exploiting *Legionella pneumophila* genomic diversity for accurate source attribution. *SPASIM Postgraduate Retreat*, Lancefield, Australia 2017 (oral presentation).
9. Andrew H. Buultjens, Koen Vandellannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Comparative genomics shows *Mycobacterium ulcerans* migration has preceded the rise of Buruli ulcer in south-eastern Australia. *Combine Student Symposium*, Adelaide, Australia 2017 (oral presentation).

10. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Comparative genomics shows *Mycobacterium ulcerans* migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia. *ABACBS Conference*, Adelaide, Australia 2017 (poster).

Invited speaker presentations

11. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Using genomics to understand the ecology of *Mycobacterium ulcerans*. Department of Health Buruli ulcer meeting, Melbourne, Australia 2015 (invited presentation).
12. Andrew H. Buultjens, Koen Vandelannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Exploring the evolution of *Mycobacterium ulcerans* and emergence of Buruli ulcer in Victoria with whole genome sequencing. *Mycobacterial Special Interest Group*, Melbourne, Australia 2016 (invited presentation).
13. Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders

Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. Exploiting *Legionella pneumophila* genomic diversity for accurate source attribution. *The Australian Society for Microbiology Public Health Night*, Melbourne, Australia 2017 (invited presentation).

Acknowledgements

As I write this, I reflect on how I have changed during the course of my life up until now. I had never thought I'd end up doing a PhD, in fact, there was a time where I thought I'd never finish secondary school, let alone go to university! My journey towards completing this PhD has only been possible through persistence, hard work, Google searches and, most importantly, the assistance of others.

I would firstly like to thank Tim Stinear, Torsten Seemann and Anders Gonçalves da Silva for their academic guidance and friendship which has enabled me to develop as scientist and as an individual. They have undoubtedly been the most knowledgeable and inspirational supervisors that I could have wished for. What I have valued immensely was their support and encouragement in me pursuing my own ideas and willingness to make themselves available for regular discussions despite their busy schedules. In particular, I would like to thank Tim for providing me with the life changing opportunity to do my Honours in his lab. That opportunity then developed into me undertaking this PhD - a new path that has dramatically and positively transformed my life.

I also wish to extend my thanks to the numerous colleagues and collaborators who have contributed in innumerable ways to make this PhD project work possible. In particular I would like to acknowledge the contributions of those in the Stinear, Howden and Williamson Laboratories. I am grateful to the guidance and support provided by my PhD Advisory Committee, made up of Paul Johnson, Janet Fyfe, Susan Ballard, Anders Gonçalves da Silva, John Wallace and Grant Jenkin. I am also indebted to the financial provision of the Australian Federal Government and would otherwise have not been able to undertake my PhD education without this assistance.

Taking a step back – all the way to secondary school - I would like to thank Cathy O'Brien, Judith Clarke and Vicki Poke for their tireless efforts to keep me in school. In particular, I would like to thank them for eventuating a special allowance for me to undertake a TAFE course in Conservation and Land Management during my year 10 studies – the perfect motivation to stay in school. At the time I had stopped attending and was certain that leaving was the better option and it was their efforts that eventually encouraged me to return. Once convinced to stay in school, it was due to the persistence of a dedicated and passionate English teacher, Bernadette Harris, that I became proficient academically and made it through my final VCE year. If not for them, I would certainly not have embarked on this challenging, yet incredibly rewarding journey.

On a personal note, I would like to thank my girlfriend Tina for all the thought-provoking discussions about my work and science in general. I'm always astonished as to how fast you can soundly comprehend the exceedingly detailed aspects of my work after me briefly explaining them just once – things that took me much longer to get my head around the first time! I would also like to thank my father for providing a home and my sister for keeping noise intensity at an appropriate level for my study.

In a more general sense, I feel honoured to stand at a point in biological science where the sequencing of an entire genome, a task that just over a decade ago would have taken several years and millions of dollars to undertake, can be insignificantly performed in only a few hours and at little expense. For this, I thank the systematic enterprise that is science and the pure chance of my existence at this point in time.

Table of contents

Synopsis.....	i
Abstract.....	iii
Declaration	vi
Preface	vii
Chapter overview	vii
Chapter 1: Literature review	vii
Chapter 2: Genomic characterisation of novel bacterial pathogens.....	vii
Chapter 3: Developing useful genomic based tools for public health investigations	viii
Chapter 4: Explaining epidemiological trends through genomic insights.....	x
Chapter 5: Discussion and conclusions.....	xi
Funding declaration	xii
List of publications arising during candidature.....	xii
List of conference communications arising during candidature	xvi
Abstracts presented in international and national meetings.....	xvi
Invited speaker presentations	xix
Acknowledgements.....	xxi
List of figures.....	xxviii
List of abbreviations.....	xxix
Chapter One: Literature review.....	1
The genomics revolution.....	3
WGS in the microbiology research domain.....	3
Public health microbiology.....	3
WGS in public health microbiology.....	4

Implementation of WGS by Public Health England.....	5
WGS in Australian public health microbiology	6
The evolution of genome sequencing technology	6
First-generation DNA sequencing	6
Second-generation sequencing.....	7
Third-generation sequencing.....	8
Single molecule real-time sequencing	8
Nanopore sequencing	8
Bioinformatics.....	9
Bioinformatic developments	9
Bioinformatic software	9
Bioinformatic databases.....	10
Sequence alignment.....	12
Reference based methods for genome assembly.....	12
Reference independent methods for genome assembly	12
Determination of core and pan genomes	13
SNP comparisons and clustering	14
Phylogenomics.....	14
Maximum parsimony	14
Distance based methods.....	15
Models of molecular evolution	15
Bayesian phylogenomics	16
Multivariate statistical analyses.....	16
Infectious diseases in Australia	18
<i>Enterococcus faecium</i>	18

Pathogenesis of <i>E. faecium</i>	19
Vancomycin resistant enterococci	19
Genotyping of <i>E. faecium</i>	20
WGS of <i>E. faecium</i>	20
<i>E. faecium</i> sequence type 17	21
<i>E. faecium</i> sequence type 203.....	21
<i>E. faecium</i> sequence type 796.....	22
<i>Legionella pneumophila</i>	22
Transmission of <i>L. pneumophila</i>	23
The role of cooling towers in the spread of <i>L. pneumophila</i>	23
Pathogenesis of <i>L. pneumophila</i>	24
Pathology of <i>L. pneumophila</i> infections.....	24
Genotyping of <i>L. pneumophila</i>	24
<i>L. pneumophila</i> Sequence Based Typing	25
WGS of <i>L. pneumophila</i>	25
Current issues with WGS-based outbreak investigations	26
The Melbourne Aquarium outbreak of Legionnaires' disease	26
<i>Mycobacterium ulcerans</i>	27
Pathogenesis of <i>M. ulcerans</i>	27
Pathology of Buruli ulcer	28
Transmission and infectious reservoir of <i>M. ulcerans</i>	28
<i>M. ulcerans</i> in animals.....	29
Molecular typing of <i>M. ulcerans</i>	30
WGS of <i>M. ulcerans</i>	30
Buruli ulcer in Southeastern Australia	32

The epidemiology of Buruli ulcer in Victoria.....	32
Research objectives	33
Chapter Two: Genomic characterisation of emerging hospital bacterial pathogens.....	35
Introduction.....	37
Published manuscript: “Evolutionary origins of the emergent ST796 clone of vancomycin resistant <i>Enterococcus faecium</i> ”	38
Summary	61
Chapter Three: Developing useful genomic based tools for public health investigations	63
Introduction.....	65
Published manuscript: “A supervised statistical learning approach for accurate <i>Legionella pneumophila</i> source attribution during outbreaks”	66
Summary	80
Chapter Four: Explaining epidemiological trends through genomic insights	82
Introduction.....	84
Published manuscript: “Comparative genomics shows <i>Mycobacterium ulcerans</i> migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia”	85
Summary	101
Chapter Five: Discussion and conclusions	103
Discussion	105
Key Findings	105
Strengths and limitations.....	105
Importance and implications for public health microbiology	109
Future directions	110
Conclusions	111

References 112

List of figures

Figure 1: Overview of genome analysis

List of abbreviations

BLAST	Basic Local Alignment Search Tool
bp	Base Pair
CC	Clonal complex
cgMLST	Core genome MLST
contig	Contiguous DNA sequence
DAPC	Discriminate analysis of principal components
DNA	Deoxyribonucleic acid
MLST	Multilocus Sequence Typing
MRCA	Most Recent Common Ancestor
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
PFGE	Pulse field gel electrophoresis
PhD	Doctor of Philosophy
rRNA	Ribosomal ribonucleic acid
SBT	Sequence Based Typing
SNP	Single nucleotide polymorphism
ST	Sequence Type
USB	Universal Serial Bus
WGS	Whole genome sequencing

Chapter One:
Literature review

The genomics revolution

Every now and then a technological breakthrough is achieved that changes the field of biological sciences forever. Currently, we stand at a point in time where technological advances, together with decreasing cost of genome sequencing and analysis, denote the genomics revolution. In just the past few years there has been substantial competition among sequencing technologies that has driven the emergence of enhanced DNA sequencing instruments equating to dramatically reduced sequencing costs. The per-base cost of DNA sequencing has reduced by approximately 100,000-fold over the preceding decade, translating to a rate of development that exceeds that of Moore's law of advancements seen in the semiconductor industry (1). Such affordability has decentralized whole genome sequencing (WGS), as it is now available as a research tool to even small research laboratories.

WGS in the microbiology research domain

WGS technology is emerging to become the cornerstone technology for microbiological investigations. WGS is now being used extensively to study the evolution and dissemination of bacterial pathogens, including detecting person-to-person transmission incidents and propagation of pathogens through communities, hospital outbreaks or transcontinental spread of pathogens (2-10).

Public health microbiology

Public health microbiology (PHM) is a discipline focused on observing the epidemiology of infectious diseases for the purpose of detecting outbreaks, informing infection control practices and gaining intelligence on emerging infectious threats (11). Conventional PHM practices are to grow isolates from environmental and clinical specimens and apply a multitude of

specialized and frequently species-specific methods (11). Examples of such procedures include species identification, antibiotic resistance testing, outbreak detection and surveillance, of which are often labor intensive, complex and require clinical microbiologists with extensive knowledge bases (11). Another disadvantage of conventional methods is the slow turnaround time for generating data. For example, tests can take several days for rapidly growing bacteria like *Escherichia coli* or months for relatively slowly growing bacteria such as *Mycobacterium tuberculosis* (11). Ideally, the same (or better) diagnostic yield as obtained from a variety of microbiological tests could be attained with WGS in a single step and with a rapid turn-around time.

WGS in public health microbiology

Despite microbial genomics being well established in the research domain, it is just beginning to be integrated into the realm of PHM. WGS now represents a viable option for public health practice owing to the drastic reduction in both capital expenditure and turn-around time (12). The reduction in price of WGS is so substantial that it is becoming a cheaper option than that of conventional strain genotyping approaches. Beyond being a more financially feasible option, WGS also delivers greater genotyping resolution compared to that of conventional methods (13). Given the advantages of WGS, it is obvious that this technology will play a major role to modernise PHM.

Potential applications of WGS for PHM include antibiotic susceptibility testing, pathogen surveillance, assessing relatedness to other organisms and identifying and tracing emerging infectious hazards (13-15). In terms of workflow, another advantage of WGS is that it offers a common procedure for all organisms and that genome data holds all, or virtually all, the

information previously obtained through numerous lengthy procedures (11, 13). This common workflow greatly simplifies laboratory practices and translates to reduced handling of pathogens, lessening the risk of health and safety incidents (13).

Further to routine public health service, microbial genomics can also be used in a retrospective sense to indirectly enlighten public health actions and practices. Such analyses can make use of comparative genomics to reveal deep insights into the epidemiology, evolution and population dynamics of important pathogens (16-21). An example of this is a comparative genomics study that exposed the emergence and spread of 19A pneumococcal capsular variants succeeding the introduction of the pneumococcal vaccine in the USA (22). This study demonstrated that WGS based insights can identify the emergence of novel infectious threats as well as appraisal of public health interventions. Microbial genomics can also be used to conduct studies that provide direct public health relevancy, such as those that guide infection control and inform public health responses in a timely manner (10, 23-26).

Implementation of WGS by Public Health England

A leading example of the implementation of WGS into PHM is that of Public Health England, an institution that has been working to adopt such technologies since 2012 (13). Here, the value of WGS on the service was realised through enhanced disambiguation of outbreaks, speedier resolution of outbreak investigations and more robust linking of cases to outbreak sources (13). It was also found that WGS provided superior evolutionary context, allowing for what were beforehand alleged as sporadic cases to be associated through common sources (13). With respect to the laboratory, the implementation of WGS was found to improve laboratory management, streamline staff training and enhance health and safety (13).

WGS in Australian public health microbiology

As has been demonstrated by various public health institutions worldwide, the adoption of WGS into Australian public health practice is imperative. The implementation of this technology will better equip Australian public health agencies to respond to outbreaks and modernise laboratory processes. Such an alteration will keep Australia up to date with the genomics revolution. In particular, the adoption of WGS will preserve compatibility of Australia's public health intelligence to that of public health institutions abroad that have implemented WGS. Sustaining international compatibility on infectious threats is exceptionally relevant in today's highly mobile global population.

The evolution of genome sequencing technology

First-generation DNA sequencing

Despite the genomics revolution being relatively recent, the initial development of genome sequencing commenced many decades ago. The first DNA sequencing technique was the chain termination method, established by Frederick Sanger and associates during the 1970s, known as "Sanger sequencing" (27). Using this method, the bacteriophage phi-X174 genome was sequenced, representing the first genome to be determined (28). During the 1990s, Sanger sequencing was combined with a method that used numerous overlapping sequences known as "shotgun sequencing" to improve efficiency (29). This improved method allowed for the sequencing of the first bacterial genome, *Haemophilus influenza* in 1995 (30). The development of fluorescently labeled dye-terminator nucleotides, combined with automated detection, allowed for a new level of sequencing throughput that facilitated the human genome project (1). The human genome project was a substantial undertaking that realised for the first time,

the abundant potential of genome sequencing and drove many innovations of sequencing technology. These developments resulted with the dawn of “second-generation sequencing” technologies.

Second-generation sequencing

The advent of second-generation sequencing was characterised by instruments capable of massive parallelization of sequencing reactions, and thus greater throughput and enhanced sequence yield. Such technologies made use of distinct properties of DNA synthesis such as (i) the release of pyrophosphate during DNA synthesis to generate ATP - a substrate for luciferase that in turn creates light, or (ii) the release of H⁺ ions during DNA synthesis that cause a pH shift that can in turn be measured using an ion-sensitive field-effect transistor (31). These improvements were further developed by Illumina with the incorporation of reversible dye terminators as well as bridge-amplification to yield large amounts of high-quality data with minute quantities of input DNA. This technology was used with the Illumina HiSeq and MiSeq machines that provided greater throughput and longer read lengths than earlier platforms (32, 33). The higher efficiency and lower costs of second-generation sequencing have allowed this technology to transfer from high-end sequencing centers to that of smaller laboratories (34). Within the field of microbiology, studies using second-generation sequencing have uncovered substantial micro-evolution within bacterial species and demonstrated the potential to use genomic variation to elucidate the origins of disease outbreaks (35, 36).

Some limitations of second-generation sequencing technologies include high error rates near the terminal ends of reads and read lengths of only a few hundred bp (37). In addition, these technologies have an inherent source of bias and sequencing error as they are contingent upon

PCR-based amplification to generate sufficient template DNA (38). More recent developments in sequencing technology have made use of innovative methods to infer nucleotide identity and substantially increased read length.

Third-generation sequencing

Single molecule real-time sequencing

Single molecule real-time sequencing technology was established by Pacific Biosciences (PacBio). Here, synthesis of the DNA occurs with fluorescently labeled nucleotides that don't terminate the replication process. In this way, the synthesis occurs uninterrupted as the fluorescence signals are identified (39). Another point of distinction is that replication occurs using single DNA molecules (no clonal amplification step) within millions of discrete nanostructures known as zero-mode waveguides (40). Through this method, PacBio instruments are able to produce read lengths in the order of several thousand bases or greater, which are considerably longer than that of second-generation technologies (36, 39). Despite providing long read lengths, such datasets can possess higher error rates than that of second-generation technology (41).

Nanopore sequencing

Another major third-generation technology is nanopore sequencing of Oxford Nanopore. Nanopore sequencing is further differentiated from second-generation technologies as base identities are inferred through variations in electric current rather than detecting of fluorescent signals. In this way, strands of DNA are passed through nano-scale pores in a lipid membrane and specific changes in pore voltage provide signals that can be correlated to a specific DNA base (A, G, C or T) passing through the pore (42). Nanopore sequencing technology is capable

of producing reads lengths of approximately a million bp (43). Owing to the fact that such an approach eliminates the need for optical measurement, the physical footprint of the instrument is vastly smaller. In fact, the current size of the Oxford nanopore MinION is comparable to that of an ordinary portable USB storage device (42). As with PacBio technology, despite very long read lengths, current nanopore sequencing data is hampered by a high error rate (42).

Bioinformatics

Bioinformatic developments

Bioinformatics is the computational handling and processing of genetic information (44). This field has benefited from parallel advancements in sequencing and computing technology. The vast amount of data generated through second and third generation sequencing projects, sometimes in the realm of terabytes, has driven the development of novel bioinformatic tools for data analytics.

Bioinformatic software

Bioinformatic software is largely publicly available as “open-source”, generated predominantly from research activity, however commercial software packages are also becoming increasingly available (15, 45). The majority of bioinformatic software are modular tools that perform a specific task. Dedicated scripts are generally developed to automate the movement of data through many discrete programs, known as “pipelines” (46). Such pipelines empower researchers to focus on biological investigations rather than needing to have an in-depth knowledge of how to run each required software module. Pipelines also allow for many large datasets to be batched at once, permitting high throughput analyses.

Bioinformatic databases

With an increasing number of sequencing projects there came a demand to collate the resulting information in centralized and accessible databases. With the advent of more efficient sequencing technologies, a number of central databases were established including the National Center for Biotechnology Information's (NCBI) GenBank database, the European Molecular Biology Laboratory's (EMBL) European Nucleotide Archive and the DNA Data Bank of Japan (DDBJ). Collectively these major databases comprise the International Nucleotide Sequence Database Collaboration. With an abundance of sequence data accessible, studies began to focus on comparing sequences.

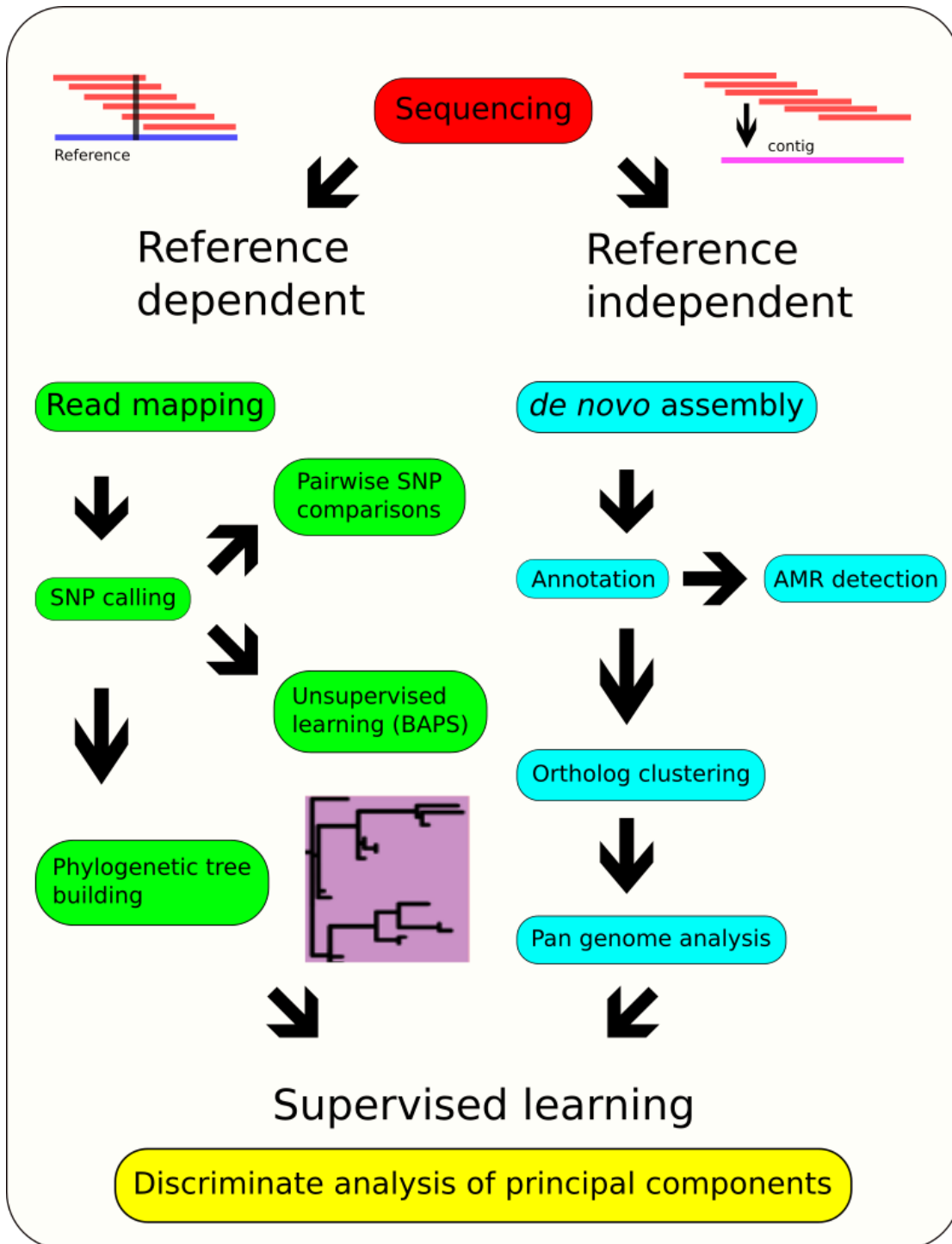


Figure 1: Overview of genome analysis

Sequence alignment

One of the most straightforward sequence based comparative techniques is alignment. Sequence alignment involves searching a query sequence against a database of sequences for matches and ranking of best hits. The mainstay of sequence searching is the Basic Local Alignment Search Tool (BLAST) algorithm, due to its' speed and easy to interpret output (47). More advanced analyses broadly fall into those that are reference dependent and reference independent (Figure 1). Generally speaking, reference dependent procedures are more accurate for calling the smallest genomic variants, Single Nucleotide Polymorphisms (SNP) (2, 48, 49).

Reference based methods for genome assembly

Reference based methods largely involve the alignment of reads against a reference sequence. The layering alignment of reads allows for multiple estimates of nucleotide identity at each reference site. This redundancy in estimates permits for a consensus to be established for the most probable base at a given site, overcoming the known random error rate specific to the dataset. Once bases for the genomes of interest have been called in respect to the reference, comparisons of high-resolution variants such as SNPs can be undertaken (Figure 1).

Reference independent methods for genome assembly

Reference independent methods make use of *de novo* assembly that build overlapping sequence reads into contiguous stretches of larger sequence, known as contigs (Figure 1) (50-53). The foremost value of reference independent methods is that novel genomic elements that are specific to the isolate in question can be examined. In the case of reference dependent methods, only genomic regions that share direct homology with the reference sequence can be assessed.

The latter approach is problematic when using short-read data as some repeat regions can be larger than the read-lengths, leading to assembly failure for such sections (52, 54).

The longer read lengths of third-generation sequencing technologies make genome assembly a much less challenging undertaking, however such data has higher error rates. In order to produce intact and high quality assemblies, data from both second and third-generation technologies are being combined (41, 55-57). Gene annotation software can be used to predict and infer the function of genes from contigs (Figure 1). Such annotation information can then be used to perform comparative gene content analyses. One such approach is to cluster the predicted genes according to their homology using a reciprocating sequence alignment analysis. Here genes are clustered into orthologous groups based on translated amino-acid homology. With such information, gene content questions such as “which genes are unique to particular genomes” and “which genes do these genomes share?” can be investigated.

Determination of core and pan genomes

The sequencing of multiple isolates from within the same species has allowed for interesting insights into within species genomic diversity. Appreciable amounts of plasticity have been identified in the presence and absence of orthologous genes among genomes. Genes shared among all isolates of a species generally encode for essential functions and are collectively referred to as the “core genome”. Conversely, genes that are found in some but not all members of a species generally encode for functions required for survival in continuously dynamic environments. These dispensable genes are referred to as the “accessory genome” that, when complemented with the core genome, make up the “pan genome”. Generally, bacteria that are ecological “specialists” tend to have streamlined pan genomes, in that the core genome is rather

large with a smaller complement of accessory genes. On the other hand, ecological “generalists” tend to have comparably open pan genomes, with less than half of genes comprising their core genome (58).

SNP comparisons and clustering

One of the most basic comparative techniques performed upon SNP data is the pairwise comparison of differences between and within that of specified groups (Figure 1). This approach therefore requires that isolates be assigned to groupings according to some particular attribute. A more sophisticated approach that does not rely upon prior groupings is to cluster genomes using an unsupervised method. Here, genomes are classified according to underlying patterns of population structure (59).

Phylogenomics

Traditionally, phylogenies based on DNA differences were built upon variation in one or more genes and are known as phylogenetic trees (60-62). With access to genome wide data, SNP discovery is not limited to gene(s) as variant detection can extend to virtually any genomic position. In this way, phylogenies built using genome wide SNP data are termed “phylogenomic trees”, of which tend to better depict the evolutionary history of the entire genome, rather than that of phylogenetic trees (63-65).

Maximum parsimony

The most elementary methods to build phylogenies are based on maximum parsimony. These procedures seek to infer a tree that requires the least amount of evolutionary events in order to be consistent with the original sequence data (66, 67). Limitations to maximum parsimony

methods are that they are susceptible to long branch attraction, in which faster evolving genomes are erroneously congregated together despite being divergent (68). To overcome the issue of long branch attraction and improve phylogeny accuracy, alternative distance-based methods were developed (69, 70).

Distance based methods

The Neighbor Joining method is a popular distance-based technique that converts a SNP alignment into a distance matrix of estimated evolutionary distances between genomes (70, 71). The major limitation of Neighbor Joining is that it fails to take into consideration the fact that multiple substitutions can occur within a site, obscuring the true evolutionary distances among isolates. In this way, the observed differences between isolates are not an accurate reflection of the evolutionary distances, with sequences appearing to be exaggeratedly adjacent to each other (72).

Models of molecular evolution

In order to account for the occurrence of unobserved substitutions, models of molecular evolution were developed. Such models work to correct the pairwise differences among isolates to account for multiple unobserved substitutions. One of the most traditional and robust methods that employ such models is that of Maximum Likelihood (73). This method works by judging the hypothesis of how well the tree predicts the observed sequence, otherwise known as the likelihood of the tree (74, 75). Further advantages of Maximum Likelihood approaches are the incorporation of additional evolutionally relevant model parameters including empiric base frequencies, probability of transitions and transversions as well as rates and sites of heterogeneity (73, 76-78). Like the previously discussed phylogeny

estimating methods, the Maximum Likelihood technique produces a single output tree, in this case the tree with the greatest likelihood. The reliability of a maximum likelihood tree can be appraised through a bootstrapping process in which the resulting tree is compared to a large population of pseudo-replicate trees (79, 80). Such an approach provides a level of confidence for specific nodes in the tree.

Bayesian phylogenomics

The most sophisticated development in phylogeny estimation has been the expansion of methods that make use of Bayesian theory. Unlike the underlying hypothesis employed by Maximum Likelihood methods, Bayesian approaches combine prior information with the likelihood of the sequence data to inform the posterior, which is a probability distribution of the resulting output (81, 82). The benefit of such an increasingly complex approach is that it allows for the modeling of additional parameters such as epidemiological information, disease transmission, divergence-time dating and geographical dynamics (83-85).

Studies have made use of Bayesian phylogenomics to provide high-resolution insights into the evolutionary dynamics of bacterial pathogens. One such example has been the evolutionary investigation of a large collection of *M. tuberculosis* genomes. Here the use of temporal and genomic data generated a time-based phylogeny that demonstrates that, for tens of thousands of years, *M. tuberculosis* has been co-evolving with anatomically modern humans (86).

Multivariate statistical analyses

Multivariate statistical techniques have attracted attention in recent years for the analysis of genetic data (87-91). One such multivariate procedure that is particularly amenable to

genomic datasets is Discriminant Analysis of Principal Components (DAPC). This approach is designed to identify and describe groups among individuals that are genetically related (92). DAPC can be used to unearth rich information from genomic data as it makes use of the multidimensionality of complex datasets. In order to increase between group discrimination, sample variation is partitioned into “within-group” and “between-group” components (92). The data is first transformed and dimensionality reduced using PCA and the prior groupings of individuals are separated using discriminant analysis (92). If the data comprises sufficient richness to separate individuals according to group-specific priors, the genetic variations that contribute to such separation can be identified (92).

Numerous studies have employed DAPC to conduct comparative genomic analyses of bacterial isolates (93-95). A study of *Streptococcus suis* genomes used DAPC to assess the separation of isolates according to assigned phenotype groups. In this case the groups were defined by disease and non-clinical statuses and the model demonstrated strong separation, indicating a genomic basis for such traits (93). Going beyond simply separating according to groupings, DAPC has also been used to identify the specific variants that contribute to genomic structuring of populations. An example of this has been the use of DAPC to identify SNPs with a role in the host adaptation of *Staphylococcus aureus* (94).

Another feature of DAPC is that once a model with well-defined separation of individuals according to group priors has been built, it can be used to probabilistically assign new individuals into the model groupings (92). This is particularly advantageous because it can be exploited to predict the probability of genomes belonging to specific groupings of interest.

Infectious diseases in Australia

When compared to other nations, Australia has been privileged with its ability to combat infectious diseases over the previous century. The elevation in living standards, implementation of vaccines and antimicrobial treatment in Australia, up until the 1960s, has translated to the control of many infectious diseases (96). In addition, the work of Australian medical scientists has driven major advances that have led to improved management of public health. Despite the decline of many historically significant infectious diseases, evolutionary forces have been ongoing, with infectious threats continuing to emerge. Three particular emerging infectious diseases of public health importance that will be investigated in this thesis are *Enterococcus faecium*, *Legionella pneumophila* and *Mycobacterium ulcerans*.

Enterococcus faecium

Hospitals provide unique ecological niches that certain bacterial pathogens have evolved and thrived within. In particular, two species from the genus *Enterococcus* have risen to become leading hospital-acquired pathogens (*E. faecalis* and *E. faecium*) (97-100). Enterococci are characterised as Gram-positive, catalase-negative, non-spore forming facultative anaerobes and are found in a variety of habitats including water, soil, plants, food and animals (101-105). The presence of *E. faecium* in a wide range of animal hosts suggests that the species play an important role within a host's gut microbiota (106). A study investigating the incidence of enterococcal bacteremia isolates in Australia during 2014 revealed that 39.9% of unique episodes of bacteremia were caused by *E. faecium* (107). In addition, the proportion of resistance to antibiotics such as ampicillin and vancomycin were found in 90.5% and 46.1% of *E. faecium* isolates, respectively (107). Both the high bacteremia burden and propensity for antibiotic resistance implicate this pathogen as a major threat to Australian public health.

Pathogenesis of E. faecium

The development of *E. faecium* infections requires a combination of host and bacterial factors. Host factors leading to infection include having an impaired immune system, being moderately to seriously ill as well as admission to an intensive care unit (108, 109). The most common types of infection caused by *E. faecium* include wound infections, endocarditis, intra-abdominal infections, urinary tract infections and bacteremia (105, 110). Bacterial factors that contribute to pathogenic potential include capacity to resist a variety of antibiotics, particularly vancomycin, as well as genes that enhance a strains' virulence (111-113).

Vancomycin resistant enterococci

Vancomycin resistance is conferred by a cluster of genes comprising a glycopeptide resistance locus. Enterococci possessing these loci are known as Vancomycin Resistant Enterococci (VRE) (114, 115). The key gene in this system encodes a ligase that substitutes the D-Ala-D-Ala dipeptide moiety of peptidoglycan (the target of vancomycin) for a D-Ala-D-Lac depsipeptide substituent, thereby conferring resistance (114, 115). These loci are located on transposons, which are mobile transposable elements that can mobilize resistance via horizontal gene transfer (114). Of the nine distinct types of glycopeptide resistance loci so far discovered, the VanA and VanB variants are the most common (116). VanA is the most frequent type found in northern hemisphere countries including the USA and across Europe (117). In Australia, VanB is the most predominant type of VRE, however there has been a recent escalation of the VanA genotype (118). Some of the main concerns with VRE are the added complexity to treat infections, associations with greater mortality and greater economic burden than infections with Vancomycin Susceptible Enterococci (VSE) (119). Another risk posed by VRE is that they

can act as a reservoir that can spread vancomycin resistance to other bacteria with pathogenic potential (20).

Genotyping of E. faecium

Traditional means to assess genetic variation among *E. faecium* populations originally included pulse field gel electrophoresis and later Multi Locus Sequence Typing (MLST) (120, 121). MLST was a major improvement in genotyping technology. Some of its' advantages were that it was fully portable and the allele data could be stored in central databases accessible remotely via the internet and allowing for comparisons around the globe (122). This method considered sequence variation in multiple loci, affording improved strain-resolving capacity relative to previous methods. Despite such advancements, MLST only interrogates very small regions of the genome and therefore is hampered by a lack of resolution compared to genomic methods that consider variation across whole genomes.

WGS of E. faecium

More recent studies of *E. faecium* have turned to genomics to provide high resolution insights into the population structure and evolutionary drivers for the emergence of novel clones (20, 123-125). Phylogenomic analyses have revealed that hospital associated isolates cluster as a single clade, termed Clade-A1, that is distinct from isolates that are not generally associated with disease (125, 126). Strains belonging to Clade A-1 frequently carry genes conferring resistance to several antibiotics, harbor greater numbers of mobile genetic elements and have altered metabolic systems (125).

***E. faecium* sequence type 17**

Within the greater *E. faecium* population, the clonal complex 17 (CC17) is particularly associated with clinical infections and responsible for outbreaks across several countries (126, 127). MLST designations have suggested that ST17 is the ancestral progenitor that founded CC17 and that all subsequent lineages have evolved from this clone. In an effort to better understand the genetic basis of the highly hospital adapted ST17, a representative ST17 genome (strain name Ef_aus0004) was fully assembled (123). Analysis of this genome revealed that it comprised of a chromosome and three plasmids and contained several prophage (123). Comparison of the fully assembled genome amongst a collection of ST17 draft genomes depicted that this clone had a large accessory genome (38%), highlighting the ability of this lineage to horizontally acquire novel genetic material (123).

***E. faecium* sequence type 203**

Molecular epidemiological studies assessing genotypic diversity of *E. faecium* bacteremia isolates revealed that, prior to 2005, the Austin Hospital, Victoria, Australia, was dominated by ST17, after which time it was largely replaced by ST203 (128). Compared to ST17, ST203 was associated with a substantial upsurge in incidents of bacteremia (128). In order to better understand the genetic basis of the newly emerged ST203, a comparative genomics study was undertaken (124). Here, again, a fully assembled genome of a ST203 representative strain (Ef_aus0085) was established and used for subsequent comparative genomic analyses. High resolution comparisons of ST203 and ST17 genomes revealed approximately 500 kb of sequence in ST203 that was absent in ST17 genomes (124). Phenotypic growth assays showed that ST203 isolates grew significantly faster than that of ST17 (124). The genomic findings

suggest that the acquisition of specific genomic factors had enabled ST203 to largely replace predecessor clones in the healthcare environment.

***E. faecium* sequence type 796**

From early 2012, a novel clone of vancomycin resistant *E. faecium* (assigned the multi locus sequence type ST796) was concurrently isolated from geographically separate hospitals in south-eastern Australia and New Zealand. This newly emerged clone was found to be accountable for 40% of *E. faecium* VRE BSI across five geographically distinct Melbourne hospitals, fundamentally supplanting its ST203 predecessor strains (129). The ST796 lineage was also the causative agent of a sizeable outbreak of colonisation in a Melbourne hospital's neonatal intensive care unit (130). There was a striking increase in the number of notifications in Tasmania from 2015 and ST796 was identified as one of the most important VRE causing agents (131). During 2016, ST796 was responsible for the majority of Australian VRE bacteraemic episodes (132). At the time of commencement of this project, no genomic studies had been conducted to better understand the evolutionary emergence of the ST796 lineage.

Legionella pneumophila

In 1976, an intense epidemic of relentless pneumonia occurred in the USA, for which the causative agent was unknown (133). Epidemiological investigations revealed that majority of cases had attended an American Legion convention held in Philadelphia prior to the onset of disease (133). Further investigation exposed that the causative agent was a Gram-negative bacterium, of which was later designated as *Legionella pneumophila* (134).

Legionella is a sole representative of the family Legionellaceae, which belongs to the phylum Proteobacteria. *L. pneumophila* are ubiquitously present in freshwater systems worldwide (135-137). When residing in the environment, *L. pneumophila* can survive in a variety of forms including as a free-living bacterium, as an intracellular parasite of protozoa (which is required for replication) as well as a component of biofilms (138-143).

Despite causing disease in humans, human hosts are not considered to be natural hosts of the bacterium. *Legionella* are thought to have co-evolved with unicellular protozoa, which are also found in natural environments (144). Due to the infection of protozoa with *L. pneumophila*, it is thought that such unicellular hosts have armed this bacterium with the capability to replicate within human alveolar macrophages (145).

Transmission of L. pneumophila

The general model of transmission is that humans breathe in aerosols from an environmental source that is contaminated with the bacterium (140, 146). It is thought that the use of artificial water systems has driven the advent of Legionnaires' disease in the 20th century (147). This notion is based on the fact that sediment accumulation, temperature and the company of other microbiota enhance the growth of *L. pneumophila* (148).

The role of cooling towers in the spread of *L. pneumophila*

One particular man-made environment fostering the growth of *L. pneumophila* and often associated with outbreaks is that of cooling towers. In the majority of cases, a deficiency of control measures, inadequately maintained systems and unqualified laborers are correlated with disease acquisition from cooling towers (149).

Pathogenesis of *L. pneumophila*

The pathogenesis of *L. pneumophila* is largely due to its ability to infect and grow within alveolar macrophages. Once endocytosed, *L. pneumophila* are able to modify the vacuolar compartment to circumvent endosomal and lysosomal bactericidal processes (150). This procedure of invasion and host modification then establishes an environment that is safeguarded from immune cells and supportive of replication (151).

Pathology of *L. pneumophila* infections

Legionnaires' disease exhibits as atypical pneumonia, Pontiac fever or a self-limited flu-like infection (147, 152). Patient outcomes are contingent on a range of factors including the underlying health of the patient, if the patient is a smoker and promptness of treatment (153). Commonly, individuals with a functioning and robust immune response can clear the infection (133, 154). The efficiency of infection is somewhat low, as it has been found that only a minor subset of individuals exposed to the bacteria progress to develop Legionnaires' disease (150, 155, 156).

Genotyping of *L. pneumophila*

Owing to the fact that a single case of Legionnaires' disease means that a source of contamination exists that could potentially infect more people, timely identification and control of such a contaminating source is of primary importance. Investigations make use of patient information to trace recent movements so that presumed sources can be identified. In addition to source tracing, genotyping is employed to assess if a patients' strain matches that of other

patients as well as that of environmental strains from potential sources. Traditionally, PFGE and later Sequence Based Typing (SBT) have been used to genotype *L. pneumophila* (157, 158).

***L. pneumophila* Sequence Based Typing**

Sequence Based Typing (SBT) is analogous to that of MLST, in that isolates are allocated a “sequence type” established on the allelic profile of seven genes (159-162). One notable difference between SBT and MLST is that the latter uses housekeeping genes while the former makes use of both housekeeping and virulence genes to enhance genotypic discrimination. Genotyping studies have revealed that a great diversity of *L. pneumophila* genotypes exist in the environment, however only a fraction of such genotypes are commonly found to cause clinically relevant disease (163).

WGS of L. pneumophila

In more recent times, a growing number of studies have made use of WGS to investigate outbreaks of Legionnaires’ disease (26, 164-167). Initial genomic-based investigations of *L. pneumophila* outbreaks used SNP difference metrics to rule-in and rule-out related strains. A study by Reuter, et al. (26) found that a collection of environmental and clinical isolates, that were linked through an epidemiological investigation, had less than 15 SNP differences (26). Further comparisons with other epidemiologically unrelated isolates revealed much larger SNP differences, demonstrating a solid concordance between genomic data and traditional outbreak investigation practices (26).

Genomics has also revealed appreciable amounts of diversity within single outbreaks. A study by McAdam, et al. (167) showed that an outbreak in Edinburgh, UK, was caused by multiple

subtypes of *L. pneumophila* (167). Such within outbreak diversity is burdensome as it affects the assurance of source attribution.

The use of WGS to investigate *L. pneumophila* outbreaks has also brought about innovative means to type isolates. One such approach is that of core genome MLST (cgMLST), which is an expanded MLST scheme that exploits 1521 core genes (168). This method combines the enhanced resolution of WGS with the portability and standardisation of traditional MLST, as it delivers a single “type” for an isolate defined from a set of alleles.

Current issues with WGS-based outbreak investigations

In order for WGS based investigations to be effectively integrated into public health practice, intuitive methods are required that can surmount the unusual population structure of *L. pneumophila* outbreaks. Inferring disease transmission from phylogenies can be ambiguous and confounded when outbreaks contain several distinct genotypes, leading to possible misidentifications of outbreak sources. An alternative approach might be to implement a statistical model that takes into account both location and genome wide variation to predict the geographical source of clinical isolates.

The Melbourne Aquarium outbreak of Legionnaires’ disease

Melbourne, Victoria, Australia, is considered the ninth largest city in the southern hemisphere, with a population approaching five million inhabitants. During the April of 2000, a severe outbreak of Legionnaires’ disease occurred, resulting with 125 confirmed cases and representing the largest single episode of Legionellosis reported in Australia (169). Traditional outbreak investigative practices were employed including epidemiological tracing and PFGE

genotyping (169). At the commencement of this project, no genomic-based investigations had been conducted to provide additional insights into the population dynamics of such an important outbreak of Legionellosis.

Mycobacterium ulcerans

Buruli ulcer (BU) is an emerging disease, first described by Australian scientists in 1948 when infections in six patients were reported in the Gippsland region of Victoria, Southeastern Australia (170). The disease has since been reported in more than 30 countries (171), across Asia, Western Pacific, Latin America and in Australia, with the main encumbrance of disease situated in regions of West and Central Africa (172). Following the 1980s, West African countries including Ghana (173), Benin (174) and Cote d'Ivoire (175) have experienced major upsurges in the prevalence of BU cases with detection rates in Southern Benin surpassing that of both tuberculosis and leprosy (174). In Australia, the occurrence of BU is also becoming increasingly common, with infections documented in the Northern Territory (176), Queensland (177) and Victoria (178).

Pathogenesis of M. ulcerans

BU is caused by infection with *Mycobacterium ulcerans* (170), an atypical mycobacterial pathogen that predominantly exists as extracellular microcolonies during infection (179) and produces the lipid toxin mycolactone (180). The capacity of *M. ulcerans* to produce mycolactone is conferred by the *mlsA1*, *mlsA2* and *mlsB* polyketide synthase genes harboured on its pMUM001 virulence plasmid (181). As the pathogens' principal virulence factor, mycolactone is largely responsible for the devastating pathology of BU (180).

Pathology of Buruli ulcer

Distinct clinical characteristics describe both pre-ulcerative and ulcerative disease. Pre-ulcerative stages include plaques, papules, nodules or edema (182), which can develop to ulcers resulting in extensive impairment to skin, soft tissue and in some cases bone (183). When left untreated, the disease may result in scarring, contractures and associated functional deformities (173, 180).

The use of antibiotic treatment regimens alone are efficacious (184, 185), while antibiotic therapy with adjunctive surgery may increase the success of treating BU (186). Timely treatment of BU generally translates to better patient outcomes, however prompt detection is rare in low-income communities due to technical inabilities in making accurate diagnoses (187).

Despite being a non-life threatening disease, BU causes major social and economic upheaval to low-income communities (188, 189) and therefore needs to be controlled. Among the numerous aspects to controlling BU, understanding the mode of transmission and infectious reservoir of *M. ulcerans* is a research priority for this disease.

Transmission and infectious reservoir of M. ulcerans

Epidemiological evidence suggests *M. ulcerans* is spread from environment to humans with no human-to-human transmission (175, 190, 191). Despite research efforts aimed at delineating the ecology of *M. ulcerans*, both its mode of transmission and infectious reservoir remain poorly understood. One major obstacle to defining the natural ecology of *M. ulcerans* has been an

inability to cultivate this pathogen from the environment, although the bacteria can be isolated from clinical material.

During the 1990s, the development of a polymerase chain reaction (PCR) assay targeting the high copy number IS2404 insertion sequence of the *M. ulcerans* genome (192, 193) was a landmark for investigations, as it permitted detection of *M. ulcerans* specific DNA in environmental samples. Using the IS2404 PCR assay, *M. ulcerans* DNA was discovered in aquatic insects from a BU endemic area (194), sparking interest that aquatic habitats may be involved in the environmental supply of *M. ulcerans*. In accordance with this notion, several publications have advocated that aquatic insect bites may have facilitated the inoculation of *M. ulcerans* (194-197), with water-associated activities identified as BU risk factors (175, 191).

***M. ulcerans* in animals**

Despite the established concept that the source of *M. ulcerans* is likely to exist within aquatic environments, *M. ulcerans* DNA has also been detected in terrestrial mammals. *M. ulcerans* has been reported to infect a number of domestic animal species including horses (*Equus ferus caballus*) (198) and cats (*Felis catus*) (199), as well as native marsupials of Southeastern Australia, including koalas (*Phascolarctos cinereus*) (200), common ring tail possums (*Pseudocheirus peregrinus*) and common brush-tail possums (*Trichosurus vulpecula*) (201). In particular, possums were found to shed *M. ulcerans* in their faecal matter, suggesting that possums may act as environmental reservoirs in Southeastern Australia (202).

Molecular typing of M. ulcerans

Previous efforts to distinguish *M. ulcerans* strains have included the use of conventional genotyping technologies (203-212), however these methods provided insufficient typing resolution for *M. ulcerans* as they are unable to discriminate between closely related strains. The low resolution offered by these techniques can be explained as they sample extremely minor genomic regions, a setback that is compounded by the fact that *M. ulcerans* populations demonstrate minimal genetic diversity (213).

WGS of M. ulcerans

Analyses of the *M. ulcerans* genome have delivered insights into the natural ecology of this pathogen (214). Significant multiplication of the IS2404 and IS2606 insertion sequences have caused extensive gene decay (214), analogous to the reductive evolution of *Mycobacterium leprae* (215). In particular, the loss of genes associated with environmental survival suggests that *M. ulcerans* may be evolving to a more privileged and host-adapted niche (214).

Genomic analyses have also favored the hypothesis that *M. ulcerans* acquired the pMUM001 plasmid, necessary for the synthesis of mycolactone (181), through horizontal gene transfer with another bacterium (214). Such processes of horizontal gene transfer and reductive evolution are in agreement with other lines of evidence that suggest *M. ulcerans* evolved from a *Mycobacterium marinum*-like ancestor (203, 204, 216, 217). Owing to these evolutionary processes, *M. ulcerans* is thought to have arisen through one or more population bottlenecks (likely a single cell) (214, 218), which may explain the low genetic diversity exhibited by *M. ulcerans* populations (205, 213).

The potential of WGS to discriminate *M. ulcerans* strains was first assessed by Doig, et al. (218) among a collection of Ghanaian disease isolates. As has been reported with other pathogens such as *Mycobacterium tuberculosis* (5, 219), which also exhibits little genetic diversity (220), the WGS genotyping of *M. ulcerans* was found to afford unmatched typing resolution compared to that offered by conventional genotyping methods (218).

Assessments of genomic diversity in the Asante Akim North District, Ashanti region, Ghana, revealed that there were two dissimilar genotypes causing clinical disease (221). One of these genotypes was formerly identified in the Ashanti region while the second closely matched isolates from Nigeria. The uncovering of the Nigerian genotype suggests that there had been a second introduction of the bacterium, presumably from Nigeria (221).

A recent study made use of more sophisticated population genomic techniques to provide temporal insights into the evolutionary histories of *M. ulcerans*. The authors used a Bayesian phylogenomic approach to estimate a time-calibrated phylogeny for a collection of disease causing isolates originating from 11 endemic countries across Africa (222). The analysis revealed that there have been two distinct introductions of *M. ulcerans* into the African continent, the first of which occurred several hundred years ago and the second during the 19th century (222). Temporal estimates for the second introduction suggest that it concurred with the Neo-imperialism period, implying that European occupation of Africa may have played a role in pathogen dispersal (222).

Buruli ulcer in Southeastern Australia

In temperate Southeastern Australia, outbreaks of BU have been found to occur in localised areas, with new disease foci continuing to arise along coastal areas of Victoria (178). Victorian BU disease concentrations include Gippsland (223), Phillip Island (224), Frankston/Langwarrin and the Mornington and Bellarine Peninsulas (178, 225, 226). As well as having among the highest BU disease incidence rate in the world, Victoria also boasts what is likely to be the most accurate associated patient information available (227), thus this region presents as an ideal microcosm for delineating the ecology of *M. ulcerans*.

The epidemiology of Buruli ulcer in Victoria

Originating in the Gippsland region of Victoria, where BU first caused disease in Australia in 1939 (178, 228), there has been a gradually westward trend of BU case reporting into the Western Port (178) and Port Phillip Bay regions since the 1980s. Given that *M. ulcerans* has been detected in both endemic and non-endemic environments in both Africa (229) and Victoria (202), the question arises as to whether *M. ulcerans* was existent in Western Port and Port Phillip Bay precursory endemic environments throughout the 1930s, or it has subsequently spread westward from Gippsland. Additionally, succeeding the advent of cases west of Gippsland, there has been a substantial escalation in cases reported over time, with a 100-fold growth in the number of notified incidents per year since the year 2000. Prior to the commencement of this project, no micro-epidemiological population genomic studies had been conducted to better understand the emergence of Buruli ulcer in Australia.

Research objectives

This is a revolutionary era of progress for public health microbiology and investigations into the value and implications of WGS are required. Microbial genomic-based investigations in response to public health problems are emerging in Australia, but details on utility for specific pathogens are lacking. The overarching aim of this project was to meet this need by providing highly detailed assessments for the utility of microbial genomics in the analysis of three bacterial pathogens of public health significance.

The specific aims of this project were to:

- Assess the ability of microbial genomics to deconstruct the emergence of a novel and highly hospital-adapted pathogen *E. faecium*.
- Development of a powerful genomic approach to guide source-tracking efforts during outbreak investigations of *Legionella pneumophila*.
- Employ genomics to understand the spread of *Mycobacterium ulcerans* around Melbourne.

Chapter Two:

Genomic characterisation of emerging hospital bacterial pathogens

*Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium**

Introduction

In biology, evolution and adaptation are dynamic processes where organisms acquire mutations and other genetic changes that may confer resilience to specific environments. Those organisms might then expand their population to occupy that specific environment. With respect to the emergence of hospital bacterial pathogens, strains can rapidly emerge to occupy new niches. The regimented nature of healthcare environments, the repeated use of antibiotics and biocides in these settings supports the development of ecosystems of microbes that take advantage of these specific conditions. Into these environments we place our most sick and vulnerable people. Every so often, new pathogens that are highly adapted to the healthcare environment emerge and cause outbreaks of disease. Understanding the origins and distinguishing features of novel pathogens is necessary to provide effective control measures. From early 2012, a novel clone of vancomycin resistant *E. faecium* (assigned the multi locus sequence type ST796) was simultaneously isolated from geographically separate hospitals in south-eastern Australia and New Zealand.

In this chapter, genomics was employed to better understand the nature of ST796 with the hypothesis that genomics would provide superior resolution than that of traditional methods used to study novel pathogens. The following research questions were investigated:

- How does a fully assembled ST796 genome compare to that of other fully assembled *E. faecium* genomes?
- What is the evolutionary relationship of ST796 isolates to the greater *E. faecium* population?
- What is the genomic variation within the ST796 population?

Published manuscript: “Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*”

The methods and results for this research study were reported in a manuscript published in PeerJ:

Andrew H. Buultjens, Margaret M.C. Lam, Susan Ballard, Ian R. Monk, Andrew A. Mahony, Elizabeth A. Grabsch, M. Lindsay Grayson, Stanley Pang, Geoffrey W. Coombs, J. Owen Robinson, Torsten Seemann, Paul D.R. Johnson, Benjamin P. Howden, Timothy P. Stinear. Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. *PeerJ*, 2017 January; 5:e2916. Doi: <https://doi.org/10.7717/peerj.2916>.



Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*

Andrew H. Buultjens¹, Margaret M.C. Lam¹, Susan Ballard², Ian R. Monk¹, Andrew A. Mahony³, Elizabeth A. Grabsch³, M. Lindsay Grayson³, Stanley Pang^{4,5}, Geoffrey W. Coombs^{4,5}, J. Owen Robinson^{4,6}, Torsten Seemann⁷, Paul D.R. Johnson^{3,8}, Benjamin P. Howden² and Timothy P. Stinear¹

¹ Department of Microbiology and Immunology, Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

² Microbiology Diagnostic Unit, Department of Microbiology and Immunology, Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

³ Infectious Diseases Department, Austin Health, Heidelberg, Victoria, Australia

⁴ School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western Australia, Australia

⁵ Department of Microbiology, Pathwest Laboratory Medicine-WA, Fiona Stanley Hospital, Murdoch, Western Australia, Australia

⁶ Department of Infectious Diseases, Fiona Stanley Hospital, Murdoch, Western Australia, Australia

⁷ Victorian Life Sciences Computation Initiative, University of Melbourne, Carlton, Victoria, Australia

⁸ Department of Medicine, University of Melbourne, Heidelberg, Victoria, Australia

ABSTRACT

From early 2012, a novel clone of vancomycin resistant *Enterococcus faecium* (assigned the multi locus sequence type ST796) was simultaneously isolated from geographically separate hospitals in south eastern Australia and New Zealand. Here we describe the complete genome sequence of Ef_aus0233, a representative ST796 *E. faecium* isolate. We used PacBio single molecule real-time sequencing to establish a high quality, fully assembled genome comprising a circular chromosome of 2,888,087 bp and five plasmids. Comparison of Ef_aus0233 to other *E. faecium* genomes shows Ef_aus0233 is a member of the epidemic hospital-adapted lineage and has evolved from an ST555-like ancestral progenitor by the accumulation or modification of five mosaic plasmids and five putative prophage, acquisition of two cryptic genomic islands, accrued chromosomal single nucleotide polymorphisms and a 80 kb region of recombination, also gaining Tn1549 and Tn916, transposons conferring resistance to vancomycin and tetracycline respectively. The genomic dissection of this new clone presented here underscores the propensity of the hospital *E. faecium* lineage to change, presumably in response to the specific conditions of hospital and healthcare environments.

Subjects Evolutionary Studies, Genomics, Microbiology, Epidemiology, Infectious Diseases

Keywords *Enterococcus faecium*, Vancomycin resistant, Genome sequence, PacBio, Evolution, Hospital adapted, Comparative genomics, Recombination, Accessory genome, Antibiotic resistance

Submitted 28 October 2016
Accepted 16 December 2016
Published 24 January 2017

Corresponding authors
Paul D.R. Johnson,
Paul.Johnson@austin.org.au
Timothy P. Stinear,
tstinear@unimelb.edu.au

Academic editor
Nicholas Loman

Additional Information and
Declarations can be found on
page 15

DOI 10.7717/peerj.2916

© Copyright
2017 Buultjens et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Buultjens et al. (2017), Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. *PeerJ* 5:e2916; DOI 10.7717/peerj.2916

INTRODUCTION

Enterococcus faecium is a human and animal gastrointestinal tract (GIT) commensal but a lineage within the species has rapidly evolved to become a significant opportunistic pathogen (Coombs et al., 2014; Deshpande et al., 2007).

Early genotyping methods such as amplified fragment length polymorphism (Willems et al., 2000), restriction endonuclease analysis (Quednau, Ahrne & Molin, 1999), multi-locus sequence typing (MLST) (Homan et al., 2002; Leavis, Bonten & Willems, 2006; Top, Willems & Bonten, 2008) and more recent analyses using whole genome datasets (Lebreton et al., 2013), have shown the *E. faecium* population separates into two major clades largely according to source origin, designated as clades A and B. Clade B strains are community-associated and mostly of non-clinical origin while clade A strains are hospital-associated and mostly of clinical origin (Galloway-Peña et al., 2012; Leavis et al., 2007; Palmer et al., 2012). Clade A has been found to further divide into clade A1, which contains epidemic hospital strains, and clade A2, which encompasses animal strains and strains linked to sporadic human infections (Lebreton et al., 2013).

Clade A1 or clonal complex 17 (CC17, a MLST designation) has adapted to the hospital environment and is adept at GIT colonization with the potential to cause invasive disease (Top, Willems & Bonten, 2008; Willems et al., 2005). Members of clade A1 are characterised by larger genomes and harbour a greater abundance of virulence factors and genes conferring antibiotic resistance compared to non-A1 *E. faecium* lineages, a reflection of adaptation to healthcare environments (Galloway-Peña et al., 2012; Guzman Prieto et al., 2016; Top, Willems & Bonten, 2008).

In Australia, as in other countries, we have observed the sequential emergence of new *E. faecium* clones within the clade A1 hospital lineage which spread rapidly and displace previously endemic clones. For example, from 1994 to 2005, Australian hospital acquired *E. faecium* VRE was uncommon and mostly caused by ST17 strains. The situation changed suddenly from 2005 when there was a nationwide wave of by *E. faecium* ST203 blood stream infections (BSI), a significant and rising proportion of which are *vanB* VRE (Coombs et al., 2014; Johnson et al., 2010; Lam et al., 2012). Previous work comparing ST17 and ST203 genomes revealed that ST203 possesses 40 unique genes with inferred functions of riboflavin metabolism, ion transport and phosphorylation, and harboured a larger vancomycin resistance-conferring Tn1549 transposon (Lam et al., 2013).

At the Austin Hospital in Melbourne, improved cleaning protocols following our local ST203 outbreak were associated with a reduction in VRE BSI between 2009 and 2011 (Grabsch et al., 2012). However, despite retaining these protocols we once again observed an abrupt increase in *vanB* VRE *E. faecium* BSI from 2012 onwards that was caused by a completely new ST. We originally recognised the change in strain using PFGE and a high-resolution melt method (Tong et al., 2011) but have now switched to whole genome sequencing for epidemiological typing. We lodged the alleles of the new ST with the MLST Database and received the new designation ST796 in September 2012 (Mahony et al., 2014). ST796 was unknown before 2011 but by 2013 *vanB* ST796 *E. faecium* had caused a large outbreak of colonisation in a Melbourne Neonatal Intensive Care (Lister et al., 2015) and in

the same year was responsible for 40% of *E. faecium* VRE BSI in 5 geographically separate Melbourne hospitals, largely replacing its ST203 predecessor strains. In 2015, ST796 *vanB* *E. faecium* was responsible for 62 of 117 (53%) of all patient episodes of all *E. faecium* bacteraemia in Melbourne Hospitals, compared with 10 of 117 (8.5%) for ST203.

In the current study, we used single molecule real-time sequencing to establish a high quality, fully assembled genome sequence of ST796 *E. faecium* isolate Ef_aus0233, a representative of this emerging clone and then employed population based comparative genomics to better understand the genetic changes that have accompanied the emergence.

METHODS

Bacterial strains

A list of the isolates examined in the study is provided (Table S1). *E. faecium* were cultured as previously described (Johnson *et al.*, 2010). Four isolates were randomly selected and sequenced twice, included as technical replicates to ensure analytical reproducibility.

Whole genome sequencing

Short fragment DNA libraries were generated using the Illumina NexteraXT DNA preparation kit and fragment sequencing was undertaken with the Illumina NextSeq 500 platform using 2 × 150 bp chemistry. Highly intact and high quality genomic DNA was extracted from Ef_aus0233 and subjected to Pacific Biosciences SMRT sequencing according to the manufacturer's instructions and sequenced with two SMRT cells on the RS II platform (Pacific Biosciences) using P5-C3 chemistry. Genome assembly was performed using the SMRT Analysis System v2.3.0.140936 (Pacific Biosciences). Raw sequence data were *de novo* assembled using the HGAP v3 protocol with a genome size of 3 Mb. Polished contigs were error corrected using Quiver v1. The resulting assembly was then checked using BridgeMapper v1 in the SMRT Analysis System, and the consensus sequence corrected with short-read Illumina data, using the program Snippy (<https://github.com/tseemann/snippy>). The final chromosome assembly was validated by reference to a high-resolution NcoI optical map using MapSolver (version 3.10; OpGen, Gaithersburg, MD, USA). Common bacterial DNA base modifications and methyltransferase motifs were assessed using the protocol, RS_Modification_and_Motif_Analysis in the SMRT Analysis System v2.3.0.140936 (Pacific Biosciences).

Plasmid copy number

The approximate number of plasmid copies per cell for the Ef_aus0233 genome was inferred using differences in Illumina sequence read depth. The read depth of plasmid sequences was compared to the average chromosomal coverage to estimate copy number multiplicity.

Comparison of completed genomes

Artemis Comparison Tool (Carver *et al.*, 2005) was used to align the chromosomes of four fully assembled *E. faecium* genomes. The Ef_aus0233 chromosome was compared against other fully assembled *E. faecium* chromosomes using BLASTn DNA:DNA

comparisons that were undertaken and visualized using Blast Ring Image Generator (Alikhan *et al.*, 2011).

De novo assembly and genome annotation

Illumina sequence reads were *de novo* assembled into contigs using Spades v3.6.1 (Nurk *et al.*, 2013). The closed Ef_aus0233 genome and Spades contigs were annotated with Prokka (v1.12b) (Seemann, 2014) using the Enterococcus database (<https://github.com/tseemann/prokka/blob/master/db/genus/Enterococcus>) as well as manually annotated protein files derived from two fully assembled *E. faecium* genomes (Lam *et al.*, 2012; Lam *et al.*, 2013). Multilocus sequence types (STs) were determined using an *in silico* tool (<https://github.com/tseemann/mlst>). CRISPR databases were used to search for CRISPR sequences (<http://crispi.genouest.org> and <http://crispr.u-psud.fr/Server/>) (accessed 19th of May 2016). Sequence files were uploaded to the web based ISSaga (Varani *et al.*, 2011) (accessed 11th of February 2016) to detect both the abundance and diversity of insertion elements. Phage discovery was undertaken using the web based resource PHAST (accessed 15th of February 2016) (Zhou *et al.*, 2011).

Variant detection and Bayesian population clustering

Snippy was used to map short read data against the full-assembled Ef_aus0233 genome to call core genome single nucleotide polymorphism (SNP) differences. Hierarchical Bayesian clustering was performed upon a core SNP alignment to assign genomes into discrete populations using hierBAPS with BAPS6 (a prior of 10 depth levels and a maximum of 20 clusters were specified) (Cheng *et al.*, 2013). Nested clustering analyses were undertaken upon subsets of the original SNP alignment to a total depth of three levels or until no further clustering could be achieved.

Recombination and phylogenomic analysis

Recombination within the core genome was inferred using ClonalFrameML v1.7 (Didelot & Wilson, 2015) using the whole genome alignment generated by Snippy. The ML tree generated with FastTree v2.1.8 was used as a guide tree for ClonalFrameML. Positions in the reference genome that were not present in at least one genome (non-core) were omitted from the analysis using the “ignore_incomplete_sites true” option and providing ClonalFrameML with a list of all non-core positions. Maximum likelihood trees with bootstrap support were constructed using a recombination free SNP alignment with the program FastTree (Price, Dehal & Arkin, 2010). Bootstrap support was derived from comparisons between the original tree against 1,000 trees that were built upon pseudo-alignments (sampled from the original alignment with replacement).

Pan genome analysis

Orthologous proteins were identified through reciprocal blast using Proteinortho5 v5.11 (Lechner *et al.*, 2011). A blast cutoff of 95% identity and alignment coverage of 30% were used. The resulting matrix of ortholog presence and absence was visualized using Fripan (<https://github.com/drpowell/FriPan>) (downloaded on the 28th of April 2016). The General Feature Format files have been deposited in Figshare

Table 1 Characteristics of *E. faecium* ST796 Ef_aus0233 complete genome.

	Length (bp)	% G + C	Copy number	No of CDS	No of non-paralogous CDS	No of tRNA	No of rRNA
Chromosome	2,888,087	38.2	1	2,726	2,644	70	6
Ef_aus0233_p1	197,153	35.4	1	210	197	–	–
Ef_aus0233_p2	79,293	33.8	1	96	96	–	–
Ef_aus0233_p3	77,977	35.2	1	84	84	–	–
Ef_aus0233_p4	22,080	35.6	2	28	23	–	–
Ef_aus0233_p5	7,837	33.5	8	8	8	–	–

(https://figshare.com/articles/Evolutionary_origins_of_the_emergent_ST796_clone_of_vancomycin_resistant_Enterococcus_faecium/4007760).

Sequence alignment and visualization

The alignment of homologous sequences was undertaken using Mauve (Darling et al., 2004). Alignments were performed using MUSCLE (Edgar, 2004). Sequences and alignments were visualized using Geneious Pro (version 8.1.8, Biomatters Ltd. (www.geneious.com)).

RESULTS AND DISCUSSION

Genome overview

Assembly of the 158,885 sequence reads from PacBio SMRT sequencing of Ef_aus0233 (N50 read length 8,952 bp) resulted in reconstruction of a 3,272,427 bp genome, comprising a circular chromosome and five circular plasmids (Table 1). Remaining small homopolymer insertion errors were corrected using Illumina reads. The structural integrity of the chromosome assembly was confirmed correct by reference to a *NcoI* optical map (Fig. 1A). DNA base modification analysis indicated an absence of adenine methylation.

Antimicrobial resistance gene content

One of the major drivers behind the success of *E. faecium* in the clinical environment is its ability to acquire genes conferring antibiotic resistance (Handwerger et al., 1993; Iwen et al., 1997; Murray, 2000). *In silico* antibiotic resistance screening of the Ef_aus0233 genome confirmed the presence of seven loci, conferring resistance to many major classes of antibiotics including trimethoprim and vancomycin (Table 2). Vancomycin resistance in Ef_aus0233 is conferred by the Tn1549 transposon, harbouring the *vanB* operon that has integrated into the chromosome. Here, Tn1549 was 57 Kb in size and inserted into a signal peptidase 1 gene, which is the larger of the two reported versions and is an insertion site that was previously reported in a comparative analysis of ST203 genomes (Howden et al., 2013; Lam et al., 2013). Of particular interest, the Tn1549 transposon in Ef_aus0233 and that of the fully assembled ST203 representative genome (Ef_aus0085) (Lam et al., 2013), shared 100% pairwise nucleotide identity across the full length of the element, implying a common Tn1549 origin for these two clones. The majority of ST796 isolates exhibit vancomycin resistance. However, three ST796 vancomycin susceptible

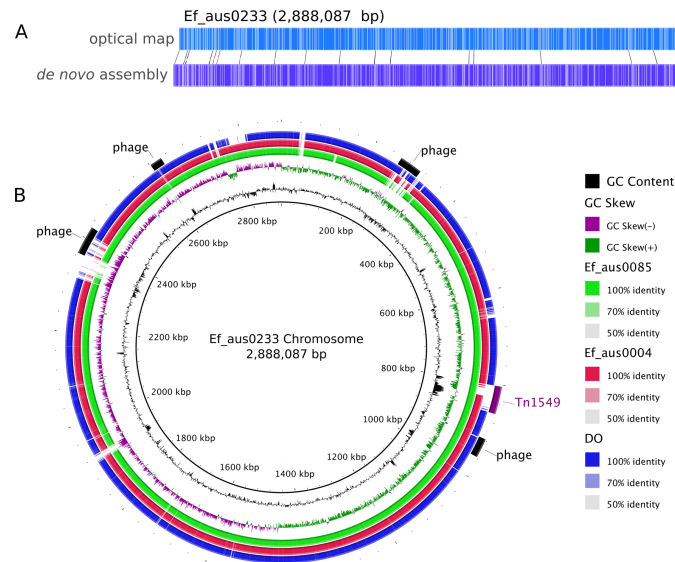


Figure 1 Ef_aus0233 chromosomal optical map and BRIG plot. (A) Optical map of the Ef_aus0233 chromosome. (B) Referenced based alignment of blast hits of Ef_aus0085, Ef_aus0004 and DO genomes against the aus0233 chromosome. Prophage elements and the Tn1549 *VanB* containing transposon are annotated as arcs in the outermost ring.

enterococci (VSE) have been isolated to date, one of which (Ef_aus1016) is included in our study and is discussed later.

Virulence gene content

In addition to antimicrobial resistance genes, virulence related genes are particularly enriched among hospital adapted *E. faecium* strains and are thought to enhance fitness in the hospital environment (Rice et al., 2003). *In silico* comparative analysis (see methods) of the Ef_aus0233 genome revealed the presence of several genes associated with virulence including collagen-binding adhesin (Rice et al., 2003) (chromosome coordinates 2,235,651–2,233,486), enterococcal surface protein (Shankar et al., 1999; Van Wamel et al., 2007) (chromosome coordinates 2,786,822–2,780,895), hemolysin (Cox, Coburn & Gilmore, 2005) (chromosome coordinates 1,025,987–1,027,363), all of which are also present in the fully assembled genomes of ST17 and ST203 isolates (Lam et al., 2012; Lam et al., 2013).

The Ef_aus0233 enterococcal surface protein (length = 5,928 nt; 9 rib repeats) shared 83.3% and 82.9% nucleotide identity with the ortholog in Ef_aus0004 (length = 4,938 nt; 6 Rib repeats), a fully assembled ST17 genome (Lam et al., 2012) and Ef_aus0085

Table 2 Antibiotic resistance genes and mutations present in Ef_aus0233 and other ST796.

Resistance	Product	Gene	Location (nucleotide positions)	Reference
Trimethoprim	Dihydrofolate reductase	<i>dhfrG</i>	331,475–331,972 (chromosome)	<i>Sekiguchi et al. (2005)</i>
Tetracycline	Tetracycline resistance protein	<i>tetM</i> (Tn916)	652,734–654,653 (chromosome)	<i>Burdett, Inamine & Rajagopalan (1982)</i>
Macrolides	ABC transporter protein	<i>msrC</i>	2,711,468–2,712,946 (chromosome)	<i>Portillo et al. (2000)</i>
	rRNA adenine N-6-methyltransferase	<i>ermB</i>	13,080–13,842 (plasmid 4)	<i>Trieu-Cuot et al. (1990)</i>
Aminoglycosides	Bifunctional aminoglycoside modifying enzyme	<i>aac(6')-aph2^r</i>	60,698–62,008 (plasmid 1)	<i>Patterson & Zervos (1990)</i>
			60,366–61,805 (plasmid 3)	
Vancomycin	VanB ligase	<i>vanB</i> (Tn1549)	803,567–861,054 (chromosome)	<i>Arthur, Reynolds & Courvalin (1996)</i>

(length = 5,199 nt; 7 Rib repeats), a fully assembled ST203 genome (*Lam et al., 2013*), respectively (*Fig. S1*). The hemolysin of Ef_aus0233 shared complete identity with orthologs in Ef_aus0004 and Ef_aus0085, suggesting that this CDS may be under strong selection.

Insertion sequence content

The Ef_aus0233 chromosome was found to contain 80 distinct elements (9 families) while Ef_aus0233_p1 had 41 (6 families), Ef_aus0233_p3 had 8 (5 families), Ef_aus0233_p4 had 8 (2 families) and no IS elements were detected on Ef_aus0233_p2 or Ef_aus0233_p5. Several of these IS families have been found not only in enterococci but additionally in species of other genera, including *Carnobacterium* and *Lysinibacillus*, reflecting the easy by which *E. faecium* can acquire exogenous DNA (*Guzman Prieto et al., 2016*).

CRISPR content

Akin to an adaptive immune system, the clustered regularly interspaced short palindromic repeats (CRISPR) systems of prokaryotes function as a sequence-specific security to defend genomes against viral predation and exposure to invading nucleic acid (*Horvath & Barrangou, 2010*). Unlike members of the community-associated *E. faecium* lineage, genomes belonging to the CC17 *E. faecium* have been found to lack CRISPR systems (*Van Schaik et al., 2010*). Given the advantages associated with the acquisition of extraneous DNA carrying antimicrobial resistance genes, CRISPRs are thought to be under negative selection among multi-drug resistant enterococci (*Palmer & Gilmore, 2010*). Despite this, two distinct CRISPR loci were detected on Ef_aus0233_p1 (chromosome coordinates 168197–168396) and Ef_aus0233_p2 (chromosome coordinates 2,630–2,860), both containing three spacers and imperfect direct repeats. A single CRISPR associated gene (*cas2*) was detected. However, no *cas1* ortholog was detected. Due to the practical necessity of *cas1* for the operation of CRISPR systems (*Yosef, Goren & Qimron, 2012*), it is unlikely that these detected CRISPR systems are functional.

Prophage content

The Ef_aus0233 genome was found to contain five putative prophages. Prophages Ef_aus0233_chr_phage-1 (chromosome coordinates 260,208–308,229: 52 CDS), Ef_aus0233_chr_phage-2 (chromosome coordinates 916,578–956,914: 61 CDS), Ef_aus0233_chr_phage-3 (chromosome coordinates 2,366,359–2,425,249: 78 CDS) and Ef_aus0233_chr_phage-4 (chromosome coordinates 2,601,769–2,627,494: 19 CDS) were located on the chromosome (Fig. 1B) while Ef_aus0233_p1_phage-1 (plasmid-1 coordinates 62,059–94,736: 30 CDS) was identified on Ef_aus0233_p1. Alignment of these prophage elements signified that several common blocks of co-linearity existed (Fig. S2); however, an overall lack of prophage genome conservation implies that these phage represent five distinct elements. Prophage gene content among a diverse collection of *E. faecium* genomes is discussed below.

Comparisons with other completed *E. faecium* genomes

In addition to diversity within the core and accessory genome, structural rearrangements represent an additional layer of genomic variation that may contribute to *E. faecium* phenotypic differences (Lam et al., 2012; Lam et al., 2013; Matthews & Maloy, 2010). To assess how the genomic organization of the Ef_aus0233 chromosome compared to that of other *E. faecium* genomes, a whole chromosome alignment of Ef_aus0233, Ef_aus0004, Ef_aus0085 and DO was undertaken. The BLASTn based alignment revealed substantial conservation of genome content (Fig. 1B) and chromosomal architecture (Fig. 2A). Like Ef_aus0085 and DO, Ef_aus0233 does not exhibit the replicore inversion observed in Ef_aus0004 (Lam et al., 2012; Lam et al., 2013).

E. faecium population genomic comparisons

In order to contextualize the 21 ST796 genomes within the global diversity of *E. faecium* as a species, we compared these genome data with a diverse collection of 89 published, fully assembled and draft *E. faecium* genomes (Table S1). To investigate the structure and evolutionary relatedness of the strains, we employed an unsupervised Bayesian clustering technique (BAPS) to distinguish distinct genomic populations and estimated a rooted phylogenomic tree using maximum likelihood.

Here, we found that our BAPS groups unambiguously classified genomes into the two previously reported A and B clades (Lebreton et al., 2013) (Fig. 3A). BAPS-1 corresponded with clade A and BAPS-2 corresponded with clade B, while BAPS-1, BAPS-1.1-4 and BAPS-1.5 overlapped with clades A1 and A2 (Fig. 3B). All ST796 isolates clustered within clade A1.

When inspecting the phylogeny of the ST796 genomes, it was noted that ST796 and ST555 share a most recent common ancestor (MRCA) (Fig. 3C). Without exception, ST555 and ST796 genomes formed distinct monophyletic clades that were in agreement with the BAPS groupings (BAPS 1.3.3 and 1.3.4, respectively). The ST555 clone is another recently-emerged hospital adapted ST, however its discovery in the hospital environment predates that of ST796 (Coombs et al., 2014). Another major difference between these two STs is that ST796 appears to have been localized to south east Australia and New Zealand

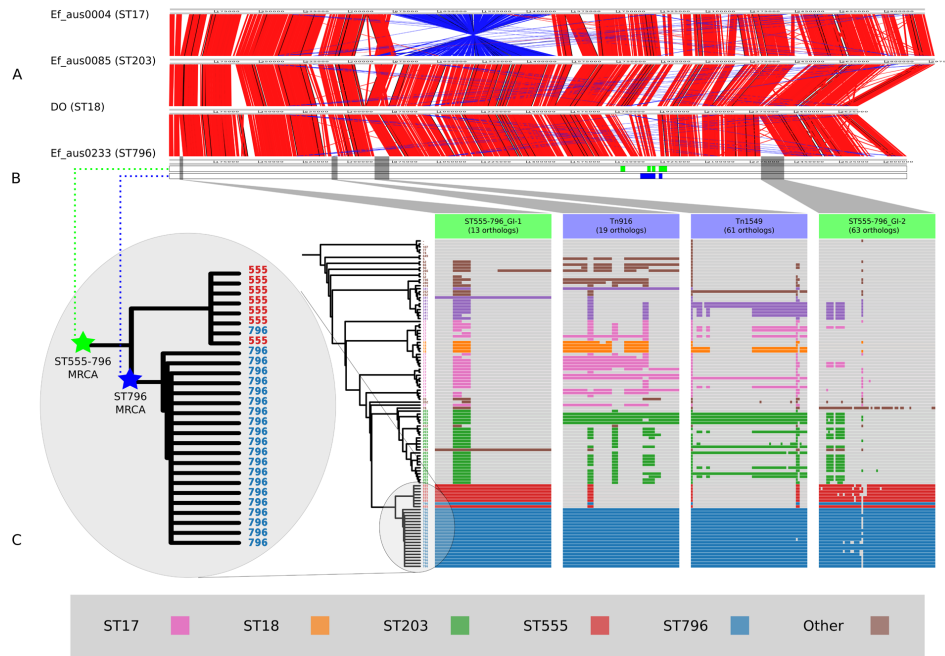


Figure 2 Comparisons of chromosomal architecture, genomic islands and recombining segments associated with the ST555-796 and ST796 clades. (A) Alignment of fully assembled chromosomes of Ef_aus0233, Ef_aus0085, Ef_aus0004 and DO. (B) Conservation of recombining segments in the ST555-796 and ST796 MRCA. (C) Core genome phylogeny aligned with gene content blocks for identified genomic islands. Colours indicate the MLST designations.

(Carter *et al.*, 2016; Coombs *et al.*, 2014), while ST555 has been reported nationally in the Northern Territory, South Australia and Western Australia (Coombs *et al.*, 2014), in China (Liu *et al.*, 2011) and among wild birds in the United States (Oravcova *et al.*, 2014). The national and international pervasiveness of ST555 and relatively limited geographical dispersal of ST796 in southeast Australia is consistent with a scenario in which the evolutionary emergence of ST555 predates that of ST796.

Recombination analyses indicated that both the ST555-796 and ST796 MRCA have evolved in part by recombination. Inspection of the inferred recombining segments for these two ancestors revealed a single hotspot of 170 kb that contained two overlapping clusters of increased SNP density (ST555-796_MRCA: chromosome coordinates 1,783,249–1,953,029, ST796_MRCA: chromosome coordinates 1,857,926–1,937,284) (Fig. 2B). The spatial clustering of these inferred ancestral recombination events suggests that this region

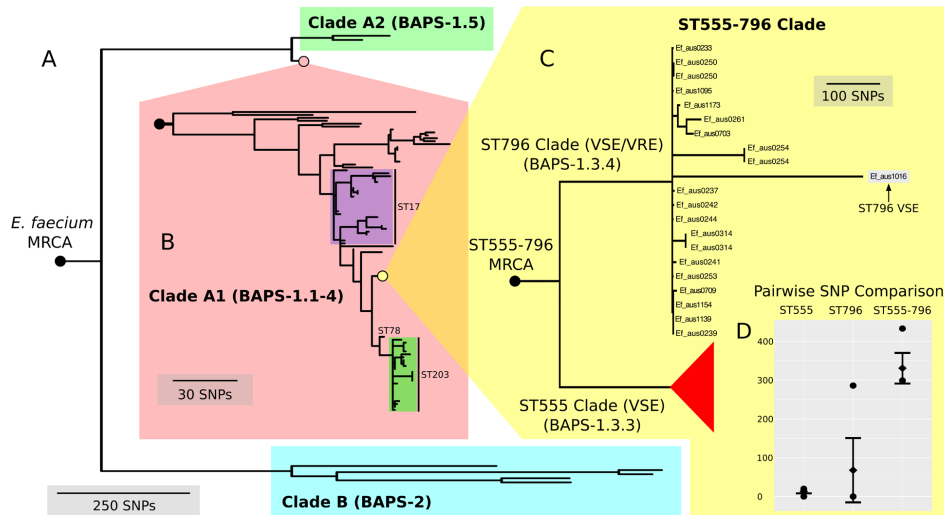


Figure 3 Nested core genome SNP phylogeny of the greater *E. faecium* population. (A) *E. faecium* population tree containing the major division between the community and hospital associated clades (B and A). (B) Nested tree focusing on the A1 clade containing the hospital associated isolates. (C) Nested tree focusing on the ST555-796 containing clade. Nodes with less than 70% bootstrap support have been collapsed. The VSE ST796 isolate is located within the ST796 clade. (D) Pairwise core SNP comparisons of within and between group diversity for the ST555 and ST796 clades (ST555: 13 SNPs, ST796: 68 SNPs, ST555-796: 330 SNPs). Y-axis depicts the number of SNP differences, error bars indicate one standard deviation above and below the mean and points represent the minimum and maximum values.

may contain non-clonally derived alleles (particularly those in the ST796 MRCA) that may have been under positive selection and perhaps contributed to the emergence of ST796. Inspection of this region revealed a preponderance of cell-wall associated transport CDS, including CDS encoding putative copper and cadmium-translocating P-type ATPases, amino acid permeases, OxaA-like membrane protein, as well as housekeeping CDS such as Glycyl-tRNA synthetase subunits, RecO DNA repair proteins and a GTP-binding protein.

In order to assess the evolutionary divergence between the ST555 and ST796 clades, a ST555-796 specific core genome was established and pairwise SNP distances were calculated. As we were primarily interested in SNPs derived through clonal evolution, again we removed SNPs inferred to have arisen through recombination. Inspection of SNP distribution in VSE Ef_au1016 revealed several dense clusters, indicating substantial recombination, which was subsequently removed by a second round of recombination detection. Consistent with two distinct groups, within clade comparisons revealed smaller mean SNP differences (within-ST555: 13 SNPs, within-ST796: 68 SNPs) than that between clades (between ST555 and ST796: 330 SNPs) (Fig. 3D).

Ancestral single nucleotide polymorphisms

Forty-one core-genome SNPs differentiated the ST555–ST796 MRCA from its predecessors, while only two core-genome SNPs were predicted in the ST796 MRCA compared to ST555 (Table S2). Analysis of these SNPs showed a range of CDS impacted. Among the 41 SNPs, 22 were predicted to change amino-acid sequence and potentially alter protein function, including non-synonymous mutations in four CDS encoding putative regulatory proteins. While the function of these regulatory proteins and the consequence of the predicted mutations is unknown, such changes can have profound impacts on phenotype (Howden et al., 2011; Howden et al., 2008).

Accessory gene content comparisons

The clustering of all predicted CDSs into orthologous groups allowed for inter-ST comparisons at the gene-content level. In total, there were 10,740 orthologous clusters among the 110 genomes, of which 1,437 were core and 9,303 were variably present (accessory)—representing the *E. faecium* pan genome (Table S3). Using this approach, orthologous clusters that were diagnostic of the ST555–ST796 and ST796 populations were identified.

Lineage specific genomic islands

The acquisition of genomic islands has been reported in previous studies that compared the genomes of hospitalized and non-hospital derived isolates, suggesting that such novel elements may offer possessing strains a competitive advantage (Heikens et al., 2008). In this study, subsets of the accessory genome that were associated with ST555–ST796 and ST796 genomes were found to cluster on the Ef_aus0233 chromosome (Fig. 2C). The contiguous location of these CDS and conserved inheritance patterns, indicated that these elements collectively formed larger genomic islands and were likely to have been acquired through horizontal gene transfer events. Given the ancestral relationships among the genomes and the conservation of ortholog presence amongst these lineages, it is reasonable to infer that these events occurred at various stages along the evolutionary paths of the ST555–ST796 and ST796 ancestries.

Two genomic islands were conserved among ST555–ST796 genomes while being almost entirely absent from genomes of other STs (ST555–ST796_GI-1, chromosome coordinates 39,093–53,122: 13 CDS, and ST555–ST796_GI-2 chromosome coordinates 2,316,643–2,373,309: 63 CDS), suggesting that these elements are likely to have been acquired by the ST555–ST796 MRCA. Assessment of the CDS annotation for ST555–ST796_GI-1 (56 kb) suggests it is a mosaic integrative element. Two 3 kb regions spanning a site-specific tyrosine recombinase and excisionase at the 5' end of this element and replication proteins at the 3' extreme were identical to a previously described *Enterococcus faecalis* pathogenicity island (Shankar, Baghdayan & Gilmore, 2002). A 13 kb region harbouring a putative beta-galactosidase and other sugar modifying CDS was identical to a region in ST203 Ef_aus0085 (Lam et al., 2013). The function of CDS in the remaining 33 kb was more difficult to infer with few database matches to indicate function. However, a role for this region in cell wall modification (potentially DNA transfer) was suggested by the presence

of CDS encoding cell-wall binding proteins, peptidases, ATP/GTP-binding proteins and peptidoglycan-binding proteins. ST555-ST796_GI-2 (14 kb) is another integrative element with CDS encoding a site-specific tyrosine recombinase, replication proteins, and putative sugar kinases, hydrolases and permeases. This potential carbohydrate utilization/transport locus shared complete nucleotide sequence identity with a region of the *Enterococcus gallinarum* genome (strain ID: FDA ARGOS_163, NCBI BioProject: [PRJNA231221](#)). The phage CDS content and their predicted products is provided ([Table S4](#)).

Two other elements were found to be exclusively present among ST796 genomes in this comparison. One of these was the Tn916-like transposon (chromosome coordinates 635,179–658,601: 19 CDS), carrying tetracycline resistance ([Franke & Clewell, 1981](#)) and the second was the Tn1549 transposon (chromosome coordinates 803,567–861,054: 61 CDS), carrying vancomycin resistance ([Garnier et al., 2000](#)). Given the phylogeny, it appears likely that the ST796 MRCA acquired these two transposons and then spread ([Fig. 2C](#)). The exception to this pattern was the single VSE ST796 genome (Ef_aus1016) that lacked these elements that were universally conserved among VRE ST796 genomes. Ef_aus1016 exhibits the same genomic island presence/absence profile that is observed among the ST555 genomes. Given this, it appears that Ef_aus1016 might be an extant descendent of an ST796 evolutionary intermediate that had not yet acquired the ST796 specific GIs, such as Tn1549, suggesting that this lineage was originally VSE. The horizontal acquisition of Tn1549 has been demonstrated ([Launay et al., 2006](#)) and evidence for a VSE version of an emergent *E. faecium* clone to precede the VRE version has been previously documented with the emergence of ST17, ST203 and ST252 ([Johnson et al., 2010](#)).

Prophage gene content comparisons

Alignment of the orthologs found within the prophages that were identified in the Ef_aus0233 genome revealed the extent to which these elements are conserved among the greater *E. faecium* population ([Fig. 4](#)). The majority of orthologs within these prophages were found to exist in non-ST796 genomes, suggesting that at the gene-content level, such prophages are not unique to ST796. Prophages Ef_aus0233_chr_phage-2, Ef_aus0233_chr_phage-4 and Ef_aus0233_p1_phage-1 showed the greatest degree of ortholog conservation with non-ST796 genomes. Prophages Ef_aus0233_chr_phage-1 and Ef_aus0233_chr_phage-3 did contain orthologs present in non-ST796 genomes, however the presence of these orthologs outside of ST796 genomes was limited. Overall the prophages in Ef_aus0233 form a substantial contribution to the accessory genome but do not contain CDS that are unique to ST796 ([Table S4](#)).

Plasmid gene content comparisons

Plasmids form an important component of the *E. faecium* accessory genome that can spread horizontally through a population and carry genetic elements that may confer enhanced fitness ([Fiedler et al., 2016](#)). Approximately 12% of the Ef_aus0233 genome (384,340 bp) is comprised of plasmid DNA. In order to assess the conservation of plasmid gene content among ST796 genomes and across the greater *E. faecium* population, the presence and absence of plasmid genes within the ortholog clusters were inspected ([Fig. 5](#)). Patterns of

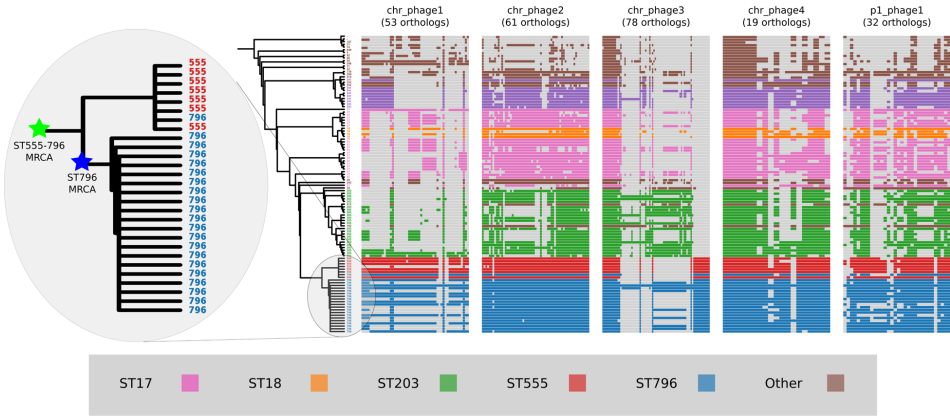


Figure 4 Prophage gene content comparisons: the presence and absence of orthologs within each of the five phages that were identified in the *Ef_aus0233* genome. The phylogeny depicts the evolutionary relationships among the genomes. Colours indicate the MLST designations.

individual ortholog presence and absence demonstrated that all plasmid orthologs were found in non-ST796 genomes, however in varying degrees. Plasmids *Ef_aus0233_p1*, *Ef_aus0233_p4* and *Ef_aus0233_p5* contain orthologs found outside ST796, however they are rarely seen elsewhere in their entirety. A list of all plasmid CDS content and their predicted products is provided (Table S5).

Given these gene content patterns and the aforementioned phylogenomic relationships between ST555 and ST796 genomes, it appears likely that the ST555–ST796 MRCA acquired these plasmids, as they are not observed in their entirety in surrounding clades. Interestingly, *Ef_aus0233_p2* was not only scarce among non-ST796 genomes but lacked conservation among the ST796 genomes. Overall, no single plasmid ortholog was specific to ST796, however given the set of isolates analyzed in this study, such plasmids, in their entirety, appear to be diagnostic of the ST555–ST796 lineage. Furthermore, the intra-ST796 differences in plasmid gene content, particularly in *Ef_aus0233_p2*, indicate there are appreciable amounts of diversity within the ST796 accessory genome, variation that might be useful during outbreak investigations involving this clone (Lister et al., 2015).

CONCLUSION

The hospital environment presents a challenging ecological niche for the adaptation of bacterial pathogens. Historically, *E. faecalis* was the leading causative agent of enterococcal nosocomial infections, however *E. faecium* infections have escalated in the last decade (Galloway-Peña et al., 2009; Guzman Prieto et al., 2016; Leavis, Bonten & Willems, 2006; Willems et al., 2011; Willems & Van Schaik, 2009). Following this apparent interspecies replacement, population-based studies have observed substantial intraspecies dynamics

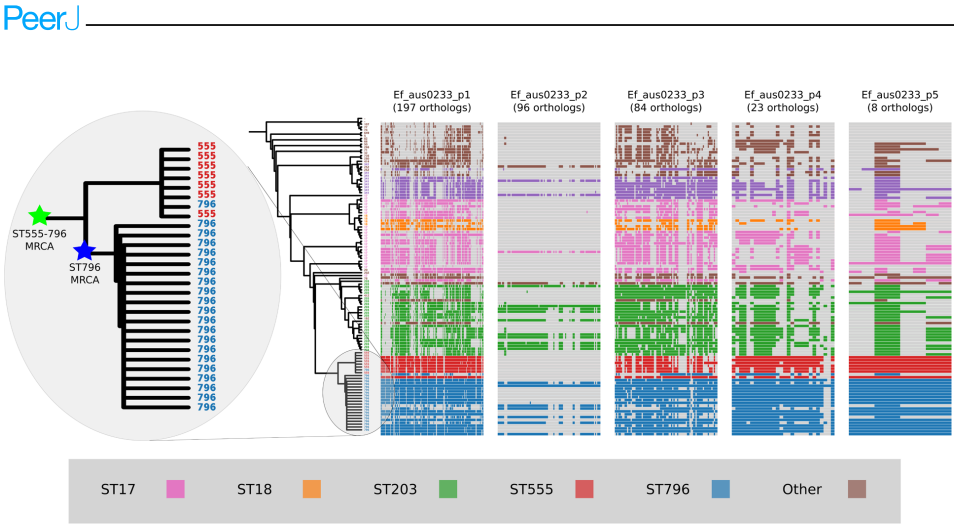


Figure 5 Plasmid gene content comparisons: the presence and absence of orthologs within each of the five plasmids that were identified in the Ef_aus0233 genome. The phylogeny depicts the evolutionary relationships among the genomes. Colours indicate the MLST designations.

with clonal replacement of *E. faecium* STs in hospitals (Bender et al., 2016; Johnson et al., 2010). Here we have described the genomic basis for the emergence of a new highly hospital adapted *E. faecium* ST early in its evolutionary history. The preparation of a fully assembled ST796 genome facilitated a comprehensive genomic analysis of this lineage and enabled detailed comparisons among other clinically relevant draft and fully assembled *E. faecium* genomes.

We demonstrate that the emergence of ST796 was preceded by several genomic events including the acquisition of two genomic islands, plasmid and phage activity, modest SNV accumulation and recombination. These analyses highlight genetic elements within the *E. faecium* core and accessory genome that may have been important drivers for the evolution of the ST555–ST796 and ST796 lineages. Given the likely significance of genomic island acquisition for the emergence of CC17 (Heikens et al., 2008), the GIs presented in this study presumably reflect adaptive responses to the clinical environment, either through acquired antibiotic resistance or perhaps enhanced capacity to utilize carbohydrates and thus augment gastrointestinal colonization. Our finding that ST796 shares a MRCA with ST555 and may have emerged from a VSE evolutionary intermediate is another example of newly emergent VRE arising from a VSE MRCA, although ST796 itself is almost exclusively VRE when identified in human BSIs unlike ST555 which causes both VSE and VRE BSI in Australian hospitals in about equal proportions (Coombs et al., 2014).

This analysis focused upon providing an overview of the first fully assembled ST796 genome and genomic differences that were assessed at the inter-ST population level. In order to explore specific diversity within the ST796 lineage, an intra-ST population study

focusing upon diversity among a large collection ST796 genomes is currently underway. Our observation of substantial variation within the ST796 accessory genome, in particular plasmid presence and absence, suggest a means for effective intra-ST796 genotyping that could potentially be more useful than core genome analysis in the tracking of outbreaks.

Hospitals are controlled but nonetheless dynamic environments. Examining pathogen genomic changes in these environments is important to understanding the bacterial response to human interventions, such as changing infection control or antibiotic stewardship practices. In this study we deconstructed the genomic events that have shaped the evolution of a highly successful *E. faecium* clone. With our current state of knowledge, we know that there is rapid genomic change occurring however we don't fully understand the consequences of such changes, as a large percentage of identified *E. faecium* genes have unknown functions. The promise of genomics will only be realized when we can combine our insights on genomic evolution with the functional consequences of such changes on pathogen phenotypes. The research presented here incrementally builds our understanding and provides a solid basis for future studies, as both clinical and public health microbiology transition into the genomic era.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research was supported by the National Health and Medical Research Council (NHMRC) of Australia (1027874, 1084015). BPH and TPS are recipients of NHMRC fellowships (1023526 and 1008549, respectively). SG and TS was supported in part by the VLSCI Life Sciences Computation Centre, a collaboration between Melbourne, Monash and La Trobe Universities, and an initiative of the Victorian Government, Australia. AAM is a recipient of an Australian Postgraduate Award, University of Melbourne. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Health and Medical Research Council (NHMRC) of Australia: 1027874, 1084015.
NHMRC fellowships: 1023526, 1008549.
VLSCI Life Sciences Computation Centre.
Australian Postgraduate Award.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Andrew H. Buultjens and Timothy P. Stinear conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

- Margaret M.C. Lam conceived and designed the experiments, performed the experiments, analyzed the data, reviewed drafts of the paper.
- Susan Ballard, Elizabeth A. Grabsch and Torsten Seemann performed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Ian R. Monk performed the experiments, analyzed the data, reviewed drafts of the paper.
- Andrew A. Mahony, M. Lindsay Grayson, Geoffrey W. Coombs and J. Owen Robinson contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Stanley Pang contributed reagents/materials/analysis tools.
- Paul D.R. Johnson conceived and designed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.
- Benjamin P. Howden conceived and designed the experiments, analyzed the data, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

DNA sequence data deposited in ENA and GenBank:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJEB14733/>.

Data Availability

The following information was supplied regarding data availability:

Buultjens, Andrew (2016): Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. figshare.

<https://dx.doi.org/10.6084/m9.figshare.4007760.v1>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.2916#supplemental-information>.

REFERENCES

- Alikhan N-F, Petty NK, Zakour NLB, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:1 DOI 10.1186/1471-2164-12-1.
- Arthur M, Reynolds P, Courvalin P. 1996. Glycopeptide resistance in enterococci. *Trends in Microbiology* 4:401–407 DOI 10.1016/0966-842X(96)10063-9.
- Bender JK, Kalmbach A, Fleige C, Klare I, Fuchs S, Werner G. 2016. Population structure and acquisition of the *vanB* resistance determinant in German clinical isolates of *Enterococcus faecium* ST192. *Scientific Reports* 6:Article 21847 DOI 10.1038/srep21847.
- Burdett V, Inamine J, Rajagopalan S. 1982. Heterogeneity of tetracycline resistance determinants in *Streptococcus*. *Journal of Bacteriology* 149:995–1004.
- Carter GP, Buultjens AH, Ballard SA, Baines S, Takehiro T, Strachan J, Johnson PDR, Ferguson JK, Seemann T, Stinear TP, Howden BP. 2016. Emergence of endemic MLST non-typeable vancomycin-resistant *Enterococcus faecium*. *Journal of Antimicrobial Chemotherapy* 71:3367–3371 DOI 10.1093/jac/dkw314.

- Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the artemis comparison tool. *Bioinformatics* 21:3422–3423 DOI 10.1093/bioinformatics/bti553.
- Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution* 30:1224–1228 DOI 10.1093/molbev/mst028.
- Coombs GW, Pearson JC, Daley DA, Le T, Robinson OJ, Gottlieb T, Howden BP, Johnson PD, Bennett CM, Stinear TP, Turnidge JD, Australian Group on Antimicrobial R. 2014. Molecular epidemiology of enterococcal bacteremia in Australia. *Journal of Clinical Microbiology* 52:897–905 DOI 10.1128/JCM.03286-13.
- Cox CR, Coburn PS, Gilmore MS. 2005. Enterococcal cytolysin: a novel two component peptide system that serves as a bacterial defense against eukaryotic and prokaryotic cells. *Current Protein & Peptide Science* 6:77–84 DOI 10.2174/1389203053027557.
- Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14:1394–1403 DOI 10.1101/gr.2289704.
- Deshpande LM, Fritsche TR, Moet GJ, Biedenbach DJ, Jones RN. 2007. Antimicrobial resistance and molecular epidemiology of vancomycin-resistant enterococci from North America and Europe: a report from the SENTRY antimicrobial surveillance program. *Diagnostic Microbiology and Infectious Disease* 58:163–170 DOI 10.1016/j.diagmicrobio.2006.12.022.
- Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Computational Biology* 11:e1004041–e1004041 DOI 10.1371/journal.pcbi.1004041.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792–1797 DOI 10.1093/nar/gkh340.
- Fiedler S, Bender JK, Klare I, Halbedel S, Grohmann E, Szewzyk U, Werner G. 2016. Tigecycline resistance in clinical isolates of *Enterococcus faecium* is mediated by an upregulation of plasmid-encoded tetracycline determinants *tet* (L) and *tet* (M). *Journal of Antimicrobial Chemotherapy* 71:871–881 DOI 10.1093/jac/dkv420.
- Franke AE, Clewell DB. 1981. Evidence for a chromosome-borne resistance transposon (Tn916) in *Streptococcus faecalis* that is capable of “conjugal” transfer in the absence of a conjugative plasmid. *Journal of Bacteriology* 145:494–502.
- Galloway-Peña J, Roh JH, Latorre M, Qin X, Murray BE. 2012. Genomic and SNP analyses demonstrate a distant separation of the hospital and community-associated clades of *Enterococcus faecium*. *PLOS ONE* 7:e30187 DOI 10.1371/journal.pone.0030187.
- Galloway-Peña JR, Nallapareddy SR, Arias CA, Eliopoulos GM, Murray BE. 2009. Analysis of clonality and antibiotic resistance among early clinical isolates of *Enterococcus faecium* in the United States. *Journal of Infectious Diseases* 200:1566–1573 DOI 10.1086/644790.

- Garnier F, Taourit S, Glaser P, Courvalin P, Galimand M. 2000. Characterization of transposon Tn1549, conferring VanB-type resistance in *Enterococcus* spp. *Microbiology* 146:1481–1489 DOI 10.1099/00221287-146-6-1481.
- Grabsch EA, Mahony AA, Cameron DR, Martin RD, Heland M, Davey P, Petty M, Xie S, Grayson ML. 2012. Significant reduction in vancomycin-resistant enterococcus colonization and bacteraemia after introduction of a bleach-based cleaning-disinfection programme. *Journal of Hospital Infection* 82:234–242 DOI 10.1016/j.jhin.2012.08.010.
- Guzman Prieto AM, Van Schaik W, Rogers MR, Coque TM, Baquero F, Corander J, Willems RJ. 2016. Global emergence and dissemination of enterococci as nosocomial pathogens: attack of the clones? *Frontiers in Microbiology* 7:Article 788 DOI 10.3389/fmicb.2016.00788.
- Handwerger S, Raucher B, Altarac D, Monka J, Marchione S, Singh KV, Murray BE, Wolff J, Walters B. 1993. Nosocomial outbreak due to *Enterococcus faecium* highly resistant to vancomycin, penicillin, and gentamicin. *Clinical Infectious Diseases* 16:750–755 DOI 10.1093/clind/16.6.750.
- Heikens E, Van Schaik W, Leavis HL, Bonten MJ, Willems RJ. 2008. Identification of a novel genomic island specific to hospital-acquired clonal complex 17 *Enterococcus faecium* isolates. *Applied and Environmental Microbiology* 74:7094–7097 DOI 10.1128/AEM.01378-08.
- Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, Van Embden JD, Willems RJ. 2002. Multilocus sequence typing scheme for *Enterococcus faecium*. *Journal of Clinical Microbiology* 40:1963–1971 DOI 10.1128/JCM.40.6.1963-1971.2002.
- Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327:167–170 DOI 10.1126/science.1179555.
- Howden BP, Holt KE, Lam MM, Seemann T, Ballard S, Coombs GW, Tong SY, Grayson ML, Johnson PD, Stinear TP. 2013. Genomic insights to control the emergence of vancomycin-resistant enterococci. *MBio* 4:e00412-13 DOI 10.1128/mBio.00412-13.
- Howden BP, McEvoy CR, Allen DL, Chua K, Gao W, Harrison PF, Bell J, Coombs G, Bennett-Wood V, Porter JL, Robins-Browne R, Davies JK, Seemann T, Stinear TP. 2011. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. *PLoS Pathogens* 7:e1002359 DOI 10.1371/journal.ppat.1002359.
- Howden BP, Stinear TP, Allen DL, Johnson PD, Ward PB, Davies JK. 2008. Genomic analysis reveals a point mutation in the two-component sensor gene *graS* that leads to intermediate vancomycin resistance in clinical *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy* 52:3755–3762 DOI 10.1128/AAC.01613-07.
- Iwen PC, Kelly DM, Linder J, Hinrichs SH, Dominguez EA, Rupp ME, Patil KD. 1997. Change in prevalence and antibiotic resistance of *Enterococcus* species isolated from blood cultures over an 8-year period. *Antimicrobial Agents and Chemotherapy* 41:494.

- Johnson PD, Ballard SA, Grabsch EA, Stinear TP, Seemann T, Young HL, Grayson ML, Howden BP. 2010. A sustained hospital outbreak of vancomycin-resistant *Enterococcus faecium* bacteremia due to emergence of *vanB* *E. faecium* sequence type 203. *Journal of Infectious Diseases* 202:1278–1286 DOI 10.1086/656319.
- Lam MM, Seemann T, Bulach DM, Gladman SL, Chen H, Haring V, Moore RJ, Ballard S, Grayson ML, Johnson PD, Howden BP, Stinear TP. 2012. Comparative analysis of the first complete *Enterococcus faecium* genome. *Journal of Bacteriology* 194:2334–2341 DOI 10.1128/JB.00259-12.
- Lam MM, Seemann T, Tobias NJ, Chen H, Haring V, Moore RJ, Ballard S, Grayson LM, Johnson PD, Howden BP, Stinear TP. 2013. Comparative analysis of the complete genome of an epidemic hospital sequence type 203 clone of vancomycin-resistant *Enterococcus faecium*. *BMC Genomics* 14:595 DOI 10.1186/1471-2164-14-595.
- Launay A, Ballard SA, Johnson PD, Grayson ML, Lambert T. 2006. Transfer of vancomycin resistance transposon Tn1549 from *Clostridium symbiosum* to *Enterococcus* spp. in the gut of gnotobiotic mice. *Antimicrobial Agents and Chemotherapy* 50:1054–1062 DOI 10.1128/AAC.50.3.1054-1062.2006.
- Leavis HL, Bonten MJM, Willems RJL. 2006. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. *Current Opinion in Microbiology* 9:454–460 DOI 10.1016/j.mib.2006.07.001.
- Leavis HL, Willems RJ, Van Wamel WJ, Schuren FH, Caspers MP, Bonten MJ. 2007. Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. *PLoS Pathogens* 3:e7 DOI 10.1371/journal.ppat.0030007.
- Lebreton F, Van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S, Zeng Q, Wortman J, Birren B, Willems RJ, Earl AM, Gilmore MS. 2013. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *MBio* 4 DOI 10.1128/mBio.00534-13.
- Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12:124 DOI 10.1186/1471-2105-12-124.
- Lister DM, Kotsanas D, Ballard SA, Howden BP, Carse E, Tan K, Scott C, Gillespie EE, Mahony AA, Doherty R, Korman TM, Johnson PD, Stuart RL. 2015. Outbreak of *vanB* vancomycin-resistant *Enterococcus faecium* colonization in a neonatal service. *American Journal of Infection Control* 43:1061–1065 DOI 10.1016/j.ajic.2015.05.047.
- Liu Y, Cao B, Gu L, Wang H. 2011. Molecular characterization of vancomycin-resistant enterococci in a Chinese hospital between 2003 and 2009. *Microbial Drug Resistance* 17:449–455 DOI 10.1089/mdr.2011.0025.
- Mahony AA, Grabsch EA, Ballard SA, Wang J, Xie S, Roberts SA, Stuart RL, Bak N, Stinear TP, Johnson PDR. 2014. Vancomycin-resistant *Enterococcus faecium* sequence type 796, the new trans-Tasman epidemic clone. In: *Australian society for antimicrobials—antimicrobials, Melbourne; 2014*. Melbourne: Australian Society for Antimicrobials.

- Matthews TD, Maloy S. 2010. Fitness effects of replicore imbalance in *Salmonella enterica*. *Journal of Bacteriology* 192:6086–6088 DOI 10.1128/JB.00649-10.
- Murray BE. 2000. Vancomycin-resistant enterococcal infections. *New England Journal of Medicine* 342:710–721 DOI 10.1056/NEJM200003093421007.
- Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology* 20:714–737 DOI 10.1089/cmb.2013.0084.
- Oravcova V, Zurek L, Townsend A, Clark AB, Ellis JC, Cizek A, Literak I. 2014. American crows as carriers of vancomycin-resistant enterococci with *vanA* gene. *Environmental Microbiology* 16:939–949 DOI 10.1111/1462-2920.12213.
- Palmer KL, Gilmore MS. 2010. Multidrug-resistant enterococci lack CRISPR-cas. *MBio* 1:e00227–00210.
- Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, Cerqueira G, Gevers D, Walker S, Wortman J. 2012. Comparative genomics of enterococci: variation in *Enterococcus faecalis*, clade structure in *E. faecium*, and defining characteristics of *E. gallinarum* and *E. casseliflavus*. *MBio* 3:e00318–00311 DOI 10.1128/mBio.00318-11.
- Patterson JE, Zervos MJ. 1990. High-level gentamicin resistance in *Enterococcus*: microbiology, genetic basis, and epidemiology. *Review of Infectious Diseases* 12:644–652 DOI 10.1093/clinids/12.4.644.
- Portillo A, Ruiz-Larrea F, Zarazaga M, Alonso A, Martinez JL, Torres C. 2000. Macrolide Resistance Genes in *Enterococcus* spp. *Antimicrobial Agents and Chemotherapy* 44:967–971 DOI 10.1128/AAC.44.4.967-971.2000.
- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS ONE* 5:e9490 DOI 10.1371/journal.pone.0009490.
- Quednau M, Ahrne S, Molin G. 1999. Genomic relationships between *Enterococcus faecium* strains from different sources and with different antibiotic resistance profiles evaluated by restriction endonuclease analysis of total chromosomal DNA using EcoRI and PvuII. *Applied and Environmental Microbiology* 65:1777–1780.
- Rice LB, Carias L, Rudin S, Vael C, Goossens H, Konstabel C, Klare I, Nallapareddy SR, Huang W, Murray BE. 2003. A potential virulence gene, *hylEfm*, predominates in *Enterococcus faecium* of clinical origin. *Journal of Infectious Diseases* 187:508–512 DOI 10.1086/367711.
- Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069 DOI 10.1093/bioinformatics/btu153.
- Sekiguchi J-i, Tharavichitkul P, Miyoshi-Akiyama T, Chupia V, Fujino T, Araake M, Irie A, Morita K, Kuratsuji T, Kirikae T. 2005. Cloning and characterization of a novel trimethoprim-resistant dihydrofolate reductase from a nosocomial isolate of *Staphylococcus aureus* CM. S2 (IMCJ1454). *Antimicrob Agents Chemother* 49:3948–3951.

- Shankar N, Baghdayan AS, Gilmore MS. 2002. Modulation of virulence within a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*. *Nature* 417:746–750 DOI 10.1038/nature00802.
- Shankar V, Baghdayan AS, Huycke MM, Lindahl G, Gilmore MS. 1999. Infection-derived *Enterococcus faecalis* strains are enriched in *esp*, a gene encoding a novel surface protein. *Infection and Immunity* 67:193–200.
- Tong SY, Xie S, Richardson LJ, Ballard SA, Dakh F, Grabsch EA, Grayson ML, Howden BP, Johnson PD, Giffard PM. 2011. High-resolution melting genotyping of *Enterococcus faecium* based on multilocus sequence typing derived single nucleotide polymorphisms. *PLOS ONE* 6:e29189 DOI 10.1371/journal.pone.0029189.
- Top J, Willems R, Bonten M. 2008. Emergence of CC17 *Enterococcus faecium*: from commensal to hospital-adapted pathogen. *FEMS Immunology and Medical Microbiology* 52:297–308 DOI 10.1111/j.1574-695X.2008.00383.x.
- Trieu-Cuot P, Poyart-Salmeron C, Carlier C, Courvalin P. 1990. Nucleotide sequence of the erythromycin resistance gene of the conjugative transposon Tn1545. *Nucleic Acids Research* 18:3660–3660 DOI 10.1093/nar/18.12.3660.
- Van Schaik W, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, Hendrickx AP, Nijman IJ, Bonten MJ, Tettelin H. 2010. Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island. *BMC Genomics* 11:239 DOI 10.1186/1471-2164-11-239.
- Van Wamel WJ, Hendrickx AP, Bonten MJ, Top J, Posthuma G, Willems RJ. 2007. Growth condition-dependent *Esp* expression by *Enterococcus faecium* affects initial adherence and biofilm formation. *Infection and Immunity* 75:924–931 DOI 10.1128/IAI.00941-06.
- Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. 2011. ISSaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology* 12:Article R30 DOI 10.1186/gb-2011-12-3-r30.
- Willems RJ, Hanage WP, Bessen DE, Feil EJ. 2011. Population biology of Gram-positive pathogens: high-risk clones for dissemination of antibiotic resistance. *FEMS Microbiology Reviews* 35:872–900 DOI 10.1111/j.1574-6976.2011.00284.x.
- Willems RJ, Top J, Van Belkum A, Endtz H, Mevius D, Stobberingh E, Van den Bogaard A, Van Embden JD. 2000. Host specificity of vancomycin-resistant *Enterococcus faecium*. *Journal of Infectious Diseases* 182:816–823 DOI 10.1086/315752.
- Willems RJ, Top J, Van Santen M, Robinson DA, Coque TM, Baquero F, Grundmann H, Bonten MJ. 2005. Global spread of vancomycin-resistant *Enterococcus faecium* from distinct nosocomial genetic complex. *Emerging Infectious Diseases* 11:821–828 DOI 10.3201/1106.041204.
- Willems RJ, Van Schaik W. 2009. Transition of *Enterococcus faecium* from commensal organism to nosocomial pathogen. *Future Microbiology* 4:1125–1135 DOI 10.2217/fmb.09.82.

PeerJ

Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research* **40**:5569–5576 DOI [10.1093/nar/gks216](https://doi.org/10.1093/nar/gks216).

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. *Nucleic Acids Research* **39**:W347–W352 DOI [10.1093/nar/gkr485](https://doi.org/10.1093/nar/gkr485).

Summary

From this work the key findings were that:

- That the emergence of ST796 was preceded by several genomic events including the acquisition of two genomic islands, plasmid and phage activity, modest SNP accumulation and recombination.
- That ST796 belongs to the A1 clade of hospital adapted *E. faecium*.
- That ST796 shares a Most Recent Common Ancestor (MRCA) with ST555 and may have emerged from a VSE evolutionary intermediate and is therefore another example of newly emergent VRE arising from a VSE MRCA.

Chapter Three:

Developing useful genomic based tools for public health investigations

A supervised statistical learning approach for accurate Legionella pneumophila source attribution during outbreaks

Introduction

Public health agencies are responsible for carrying out investigations into disease outbreaks and the implementation of control measures to limit further disease. Determining the source of an outbreak is often a difficult task that requires extensive epidemiological investigations of which can take considerable time. The time that such an investigation takes to establish a common exposure setting is particularly important as during this time further members of the public are at risk of disease acquisition. Conventional methods to determine source attribution involve compiling a list of cases from a particular geographical area and looking for associations with a point of common exposure. Public health authorities have also made use of genotyping technologies such as Sequence Based Typing (SBT) to assess the relatedness of clinical and environmental isolates. SBT is a seven-gene scheme analogous to the Multi-locus Typing Scheme (MLST) that uses SNP patterns to provide a single allelic profile. Despite having the ability to clearly distinguish divergent isolates, this genotyping scheme lacks sufficient resolution to discern among genomes that are closely related.

In this chapter, genomics was used to better understand the population structure of isolates of a large outbreak of Legionnaire's disease and to develop a tool for the prediction of outbreak source attribution. The following research questions were investigated:

- What are the evolutionary relationships of the outbreak isolates to the greater *L. pneumophila* population?
- What are the evolutionary relationships among isolates from the outbreak?
- Can a machine learning approach be used to predict the origins of clinical isolates?
- How would such a probabilistic approach compare to that of the recently proposed whole-genome MLST scheme?

Published manuscript: “A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks”

The methods and results for this research study were reported in a manuscript published in Applied and Environmental Microbiology:

Andrew H. Buultjens, Kyra Y. L. Chua, Sarah L. Baines, Jason Kwong, Wei Gao, Zoe Cutcher, Stuart Adcock, Susan Ballard, Mark B. Schultz, Takehiro Tomita, Nela Subasinghe, Glen P. Carter, Sacha J. Pidot, Lucinda Franklin, Torsten Seemann, Anders Gonçalves Da Silva, Benjamin P. Howden and Timothy P. Stinear. A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks. *J Appl Environ Microbiol.* 2017 August, 83(21): e01482-17. doi: 10.1128/AEM.01482-17.



Applied and Environmental
Microbiology®

PUBLIC AND ENVIRONMENTAL
HEALTH MICROBIOLOGY



A Supervised Statistical Learning Approach for Accurate *Legionella pneumophila* Source Attribution during Outbreaks

Andrew H. Buultjens,^{a,b} Kyra Y. L. Chua,^c Sarah L. Baines,^a Jason Kwong,^{a,b,c} Wei Gao,^a Zoe Cutcher,^{d,e} Stuart Adcock,^d Susan Ballard,^c Mark B. Schultz,^c Takehiro Tomita,^c Nela Subasinghe,^c Glen P. Carter,^{a,b} Sacha J. Pidot,^a Lucinda Franklin,^d Torsten Seemann,^{c,f} Anders Gonçalves Da Silva,^{b,c} Benjamin P. Howden,^{a,b,c} Timothy P. Stinear^{a,b}

Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Parkville, Victoria, Australia^a; Doherty Applied Microbial Genomics, The Peter Doherty Institute for Infection and Immunity, Parkville, Victoria, Australia^b; Microbiological Diagnostic Unit Public Health Laboratory at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Parkville, Victoria, Australia^c; Health Protection Branch, Department of Health and Human Services, Melbourne, Victoria, Australia^d; National Centre for Epidemiology and Population Health, Australian National University, Canberra, ACT, Australia^e; Victorian Life Sciences Computational Initiative, The University of Melbourne, Carlton, Victoria, Australia^f

ABSTRACT Public health agencies are increasingly relying on genomics during Legionnaires' disease investigations. However, the causative bacterium (*Legionella pneumophila*) has an unusual population structure, with extreme temporal and spatial genome sequence conservation. Furthermore, Legionnaires' disease outbreaks can be caused by multiple *L. pneumophila* genotypes in a single source. These factors can confound cluster identification using standard phylogenomic methods. Here, we show that a statistical learning approach based on *L. pneumophila* core genome single nucleotide polymorphism (SNP) comparisons eliminates ambiguity for defining outbreak clusters and accurately predicts exposure sources for clinical cases. We illustrate the performance of our method by genome comparisons of 234 *L. pneumophila* isolates obtained from patients and cooling towers in Melbourne, Australia, between 1994 and 2014. This collection included one of the largest reported Legionnaires' disease outbreaks, which involved 125 cases at an aquarium. Using only sequence data from *L. pneumophila* cooling tower isolates and including all core genome variation, we built a multivariate model using discriminant analysis of principal components (DAPC) to find cooling tower-specific genomic signatures and then used it to predict the origin of clinical isolates. Model assignments were 93% congruent with epidemiological data, including the aquarium Legionnaires' disease outbreak and three other unrelated outbreak investigations. We applied the same approach to a recently described investigation of Legionnaires' disease within a UK hospital and observed a model predictive ability of 86%. We have developed a promising means to breach *L. pneumophila* genetic diversity extremes and provide objective source attribution data for outbreak investigations.

IMPORTANCE Microbial outbreak investigations are moving to a paradigm where whole-genome sequencing and phylogenetic trees are used to support epidemiological investigations. It is critical that outbreak source predictions are accurate, particularly for pathogens, like *Legionella pneumophila*, which can spread widely and rapidly via cooling system aerosols, causing Legionnaires' disease. Here, by studying hundreds of *Legionella pneumophila* genomes collected over 21 years around a major Australian city, we uncovered limitations with the phylogenetic approach that could lead to a misidentification of outbreak sources. We implement instead a statistical

Received 7 July 2017 Accepted 11 August 2017

Accepted manuscript posted online 18 August 2017

Citation Buultjens AH, Chua KYL, Baines SL, Kwong J, Gao W, Cutcher Z, Adcock S, Ballard S, Schultz MB, Tomita T, Subasinghe N, Carter GP, Pidot SJ, Franklin L, Seemann T, Gonçalves Da Silva A, Howden BP, Stinear TP. 2017. A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks. Appl Environ Microbiol 83:e01482-17. <https://doi.org/10.1128/AEM.01482-17>.

Editor Donald W. Schaffner, Rutgers, The State University of New Jersey

Copyright © 2017 American Society for Microbiology. All Rights Reserved.

Address correspondence to Benjamin P. Howden, bhowden@unimelb.edu.au, or Timothy P. Stinear, tstinear@unimelb.edu.au.

Downloaded from <http://aem.asm.org/> on November 22, 2017 by University of Melbourne

learning technique that eliminates the ambiguity of inferring disease transmission from phylogenies. Our approach takes geolocation information and core genome variation from environmental *L. pneumophila* isolates to build statistical models that predict with high confidence the environmental source of clinical *L. pneumophila* during disease outbreaks. We show the versatility of the technique by applying it to unrelated Legionnaires' disease outbreaks in Australia and the UK.

KEYWORDS *Legionella pneumophila*, comparative studies, genomics, microbial source tracking, phylogeography

Legionellae are Gram-negative bacteria that replicate within free-living aquatic amoebae and are present in aquatic environments worldwide. These bacteria can proliferate in man-made water systems and cause large outbreaks of pneumonia, known as Legionnaires' disease, when contaminated water is aerosolized and inhaled (1). The majority of human infections are caused by *Legionella pneumophila* serogroup 1 (2). Public health investigations of Legionnaires' disease outbreaks are typically supported by molecular typing methods to establish the likely source of the bacteria and the extent of the outbreak. Investigations usually proceed with the assumption that a single *Legionella* genotype is responsible for an environmental point source reservoir (3). Traditional molecular typing methods described for fingerprinting legionellae include pulsed-field gel electrophoresis (PFGE) and sequence-based typing (SBT) (4). Increasingly, whole-genome sequencing (WGS) is being employed to investigate individual *Legionella* outbreaks, and the insights obtained from these high-resolution comparisons are challenging our expectations regarding common-source outbreaks, which usually are characterized by a single strain or genotype (5–9). It is becoming evident that outbreaks can be caused by multiple cocirculating *L. pneumophila* genotypes (5, 10) and that *L. pneumophila* core genomes can be surprisingly conserved across space and time (8, 11–13).

Melbourne is in the state of Victoria and is the second largest city in Australia, with a population approaching five million inhabitants, and it is considered the ninth largest city in the Southern Hemisphere. Legionellosis has been a notifiable disease in Victoria since 1979, and there are 50 to 100 cases reported each year, most occurring in the greater metropolitan region of Melbourne (14). The Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) is Victoria's State Reference Laboratory for the characterization and typing of *Legionella* species. The laboratory's collection includes isolates from a particularly noteworthy outbreak at the Melbourne Aquarium in April 2000. This was the largest single episode of legionellosis reported in Australia (15), approximately 3 months after the aquarium was opened to visitors, with construction of the site completed in December 1999. It resulted in 125 confirmed cases, with positive cultures obtained from 11 patients. Our isolate collection also spanned 28 other potential legionellosis outbreaks or infection clusters for which at least one culture isolate had been obtained.

In this study, we used comparative genomics to explore the population structure of 234 *Legionella pneumophila* isolates recovered from human and environmental sources submitted to the MDU PHL in Melbourne over a 21-year period. This collection included 11 clinical and 14 environmental isolates from the Aquarium outbreak and 42 clinical and 50 environmental isolates from 28 other likely point source case clusters. We also assessed genomic data from a recently described investigation of Legionnaires' disease cases at a UK hospital (8). The aim of this project was to develop a robust genomic approach that would surmount the unusual population structure of *L. pneumophila* and assist in the identification of case clusters and source tracking efforts during Legionnaires' disease outbreak investigations.

RESULTS

Isolates and epidemiology. There were 234 *Legionella pneumophila* serogroup 1 (Lpn-SG1) isolates obtained across a 21-year period between 1994 and 2014. Initial

multilocus sequence type (MLST) analysis indicated that 180 isolates (77%) belonged to sequence type 30 (ST30). The collection comprised 170 clinical isolates of respiratory origin (sputum or bronchoscopy specimens) and 64 environmental isolates recovered from cooling tower water samples. All isolates were collected in the state of Victoria, with the exception of six isolates from patients who were exposed elsewhere. Further information for each isolate is available in Table S1 in the supplemental material, including NCBI SRA accession numbers. One hundred ten of the 234 isolates were epidemiologically associated with 29 formally investigated case clusters or outbreaks, designated outbreaks A to AC (Table S1). The majority of these cases occurred within a 42-km radius of Melbourne city center and over a 16-year period. Outbreak A, the Melbourne Aquarium outbreak, was the largest (15).

Complete genome sequence of *Legionella pneumophila* serogroup 1 isolate Lpm7613. Before this study, there were no closed fully assembled ST30 *L. pneumophila* genomes. Thus, to ensure the identification of the maximum genetic variation among this dominant ST in our collection, we first established a ST30 reference genome sequence, selecting a clinical isolate from the Melbourne Aquarium outbreak (Lpm7613). The finished genome consisted of a single circular 3,261,562-bp chromosome (38.3% GC content) and a 129,875-bp circular plasmid (pLpm7613) (Fig. S1). Although the chromosome indicated that this genome belonged to the same lineage as *L. pneumophila* Philadelphia (Fig. 1A), the plasmid shared 100% nucleotide identity with pLPP reported in *L. pneumophila* Paris but was 2 kb shorter in length (16). A total of 2,891 chromosomal protein-coding sequences (CDSs), 43 tRNA genes, and nine rRNA loci were predicted using Prokka (17). Clustered regularly interspaced short palindromic repeat (CRISPR)-Cas regions were not detected (18).

Assessment of *L. pneumophila* population structure. Sequence reads from the other 233 genomes and the 10 selected publicly available completed genomes were mapped to the chromosome of reference strain Lpm7613. Approximately 90% of the Lpm7613 genome was present in all genomes (i.e., core), with 188,049 variable core nucleotide positions identified. Population structure analyses using an unsupervised Bayesian clustering approach revealed six distinct groups (Bayesian analysis of population structure [BAPS] groups) (Fig. 1A). Comparison of intra- and inter-BAPS group pairwise SNP distances confirmed the validity of these clusters and highlighted the extensive genetic variation among this Lpn-SG1 population (Fig. 1C). The exceptions were BAPS groups 3 and 4, as this distinction classified isolates across two clades and is likely explained by recombination. Most striking, however, was the lack of diversity within the 186 genomes comprising BAPS group 5 (here referred to as BAPS-5), with a median core SNP distance of only 5 SNPs (interquartile range [IQR], 3 to 7 SNPs). Isolates dispersed across time and space (including isolates from England, New South Wales, South Australia, and Tasmania) were scattered throughout the phylogeny. All 180 ST30 isolates were encompassed by BAPS-5, as were ST37 *L. pneumophila* Philadelphia (Philadelphia, PA, USA), ST211 *L. pneumophila* ATCC 43290 (Denver, CO, USA), and ST733 *L. pneumophila* Thunder Bay (Ontario, Canada) (Fig. 1A and B). The median inter-BAPS group distances ranged from 27,506 to 63,136 SNPs (Fig. 1C), highlighting that there is also substantial genetic diversity among Lpn-SG1 isolates circulating in Melbourne.

A rooted maximum likelihood phylogeny of the population was then inferred using the 181,633 nonrecombining core SNP loci. The phylogenomic tree reflected the BAPS clusters, with BAPS-5 forming a distinct well-supported lineage (Fig. 1A). The separation of the three North American reference isolates from the Melbourne ST30 isolates is suggestive of contemporaneous global dispersal of this BAPS-5 lineage (Fig. 1A and B). All BAPS groups displayed monophyletic origins, with the exception of BAPS-3 and BAPS-4. BAPS-3 had a single isolate of paraphyletic origin that shared a most recent common ancestor (MRCA) with BAPS-2, while BAPS-4 contained two paraphyletic subclades, one of which shared a MRCA with the majority of BAPS-3 isolates.

Impact of recombination. Recombination is a driving force in the evolution of the legionellae (5, 7, 19–22). Therefore, to further understand the structure and evolution

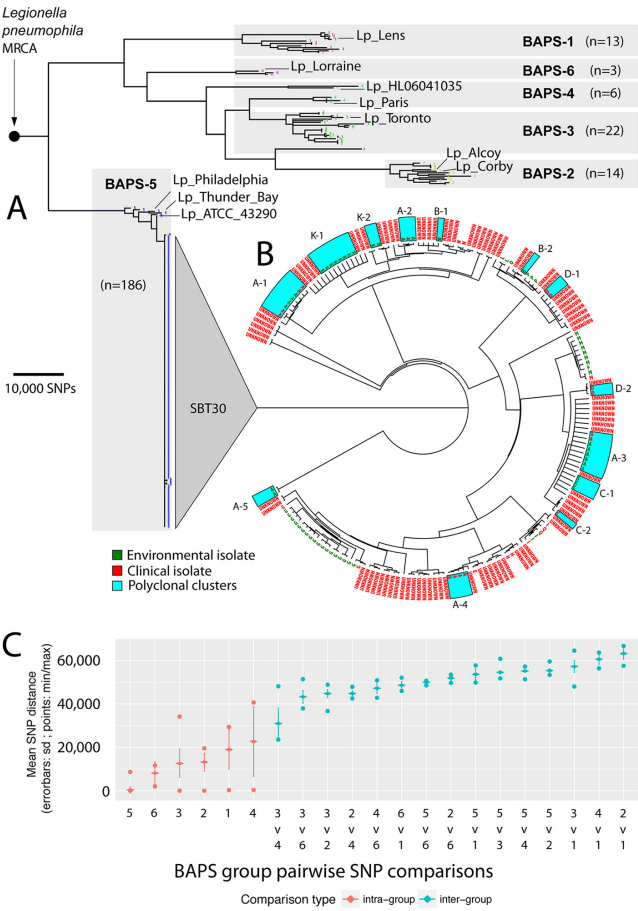


FIG 1 Global *Legionella pneumophila* population clustering, phylogenomics, and genomic molecular epidemiology of local outbreaks. (A) Core genome phylogeny estimated using maximum likelihood corresponds with six BAPS groups. Branches with less than 70% bootstrap support were collapsed, and scale indicates the number of core SNPs. The locations of the 10 international genomes are labeled. (B) ST30 core genome phylogeny. Tree tips are labeled with outbreak codes. Environmental and clinical isolates are colored according to the key. Polyclonal outbreaks/case clusters are highlighted with blue boxes. Branch lengths have been transformed and are proportional to the number of nodes under each parent node. (C) Core genome pairwise SNP comparisons of within and between BAPS groups. All groups had smaller within-group diversity than between-group diversity. sd, standard deviation; min, minimum; max, maximum.

of this Lpn-SG1 population, we assessed the impact of DNA exchange. There was evidence of extensive recombination among isolates across BAPS groups 1 to 4 and 6, with approximately 3% of the variable nucleotide sites impacted relative to the Lpm7613 reference chromosome. The detection of two paraphyletic groups (BAPS-3 and BAPS-4) is likely explained by ancestral recombination among the component subclades. In comparison, there was little recombination evident among BAPS-5 iso-

lates (Fig. S2), in accord with the core SNP phylogeny described above, and suggesting the relatively recent emergence of this *L. pneumophila* lineage. After removal of putative sequences affected by recombination, tree branch lengths showed no correlation with isolation dates ($r^2 = 0.116$). This observation indicates that nucleotide substitutions in the population have not been evolving under a molecular clock model, thus limiting estimates for dates of emergence for particular lineages.

Genomic molecular epidemiology of local outbreaks. We next compared only the 180 ST30 genomes to our Lpm7613 reference genome and again confirmed the very restricted genomic diversity within this lineage (median core SNP distance, 6 SNPs [IQR, 4 to 9 SNPs]), with five outlier genomes, impacted by recombination (Fig. 1B and C and S2). Within this reconstructed core genome ST30-specific phylogeny, many but not all epidemiologically related isolates formed distinct well-supported monophyletic clades. In some instances, epidemiologically associated isolates spanned multiple clades (outbreaks A, B, C, D, and K) (Fig. 1B). In addition, outbreak A (the Melbourne Aquarium outbreak), which was previously considered to represent infections caused by a single clone (Table S1) (15), actually contained five distinct genotypes (A1 to A5) (Fig. 1B and C).

The analysis of environmental surveillance isolates provided an ideal means to gain insights into the diversity within potential reservoirs of Lpn-SG1, diversity that might enable prospective source tracking. A phylogeographic analysis was therefore undertaken to assess the relationships between 64 environmental Melbourne metropolitan isolates against their 11 cooling tower sampling locations. Based on variation in the core SNPs, striking geographical structure was observed, with the majority of isolates from common cooling towers tightly clustering in the phylogeny (Fig. 2A). Comparisons of pairwise core SNPs depicted smaller within-group diversity and larger between-location group diversity, further indicating the existence of geographical population structure (Fig. 2B). This structure among the environmental Lpn-SG1 isolates suggested it might be possible to use the genome data to build models predictive of the environmental source to assist epidemiological efforts during outbreak investigations.

A multivariate statistical model for source attribution. To enhance resolution and try to detect outbreak-specific genomic signals, a supervised statistical learning approach called discriminant analysis of principal components (DAPC) (23) was employed. DAPC is a linear discriminant analysis (LDA) that accommodates discrete genetic-based predictors by first transforming the genetic data into continuous principal components (PCs) and building predictive classification models. The PCs are used to build discriminant functions (DFs) under the constraint that they must minimize within-group variance and maximize variance between groups. Infection clusters were defined *a priori* from the epidemiological findings, and training (environmental) isolates were used to establish the discriminant functions. The model was then used to estimate the posterior probability of membership for an unknown (e.g., clinical) isolate for each prespecified infection cluster given the training data. Here, we used 43 of the 64 environmental isolates in the training set (cooling tower isolates originating from epidemiologically defined infection clusters that possessed at least one environmental and clinical representative), under the assumption that each outbreak was caused by exposure to a point source of Lpn-SG1. We used core genome SNPs from only environmental Lpn-SG1 genomes to build the classifier (24).

Outbreak-associated environmental Lpn-SG1 strains were grouped *a priori* into training set groups based on the origin of the cooling towers from which they were isolated (see model building details in Materials and Methods). The DFs were then used to classify 15 clinical isolates that had been independently assigned based on epidemiological data to the training set groups, here referred to as the validation genomes (Table S1). The input matrix for DAPC was an alignment of 714 nonrecombinogenic SNPs that varied among the 43 environmental genomes. Plots depicting the separation of isolates according to the first two discriminant functions and the amount of variation explained are shown (Fig. 3). A model was trained using the first four principal

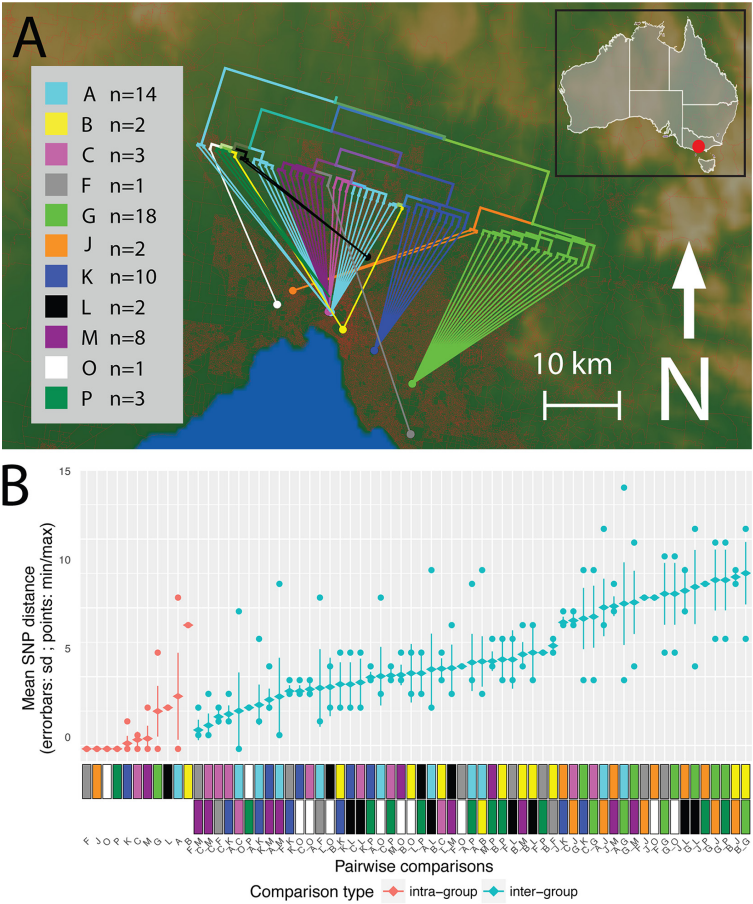


FIG 2 Phylogeography of 64 Lpn-SG1 environmental isolate genomes. (A) Map of the greater Melbourne area, showing the location of the 11 cooling towers assessed during legionellosis outbreaks, designated by colored circles. "A" (light blue) represents the location Melbourne Aquarium outbreak and is close to the center of Melbourne. The inset shows the location of Melbourne (red circle) within the state of Victoria in Southeast Australia. Overall, the phylogeny aligns closely with the geography of originating cooling towers. For several outbreak codes, polyclonality is apparent, as some common origins have connecting lines drawn from different subclades of the phylogeny. Red coloration on the base map represents population density within the greater Melbourne region. The branch lengths of the tree have been transformed and are proportional to the number of nodes under each parent node. (B) Core genome pairwise SNP comparisons of within, and between, cooling tower isolate groups. Comparisons of specific epidemiologically defined groups (infection clusters) are indicated with color codes as defined in the key. All groups had smaller within-group diversity than between-group diversity.

components (PCs), as this was found to be optimal (see Materials and Methods). We next classified our clinical validation genomes using the model and found a 93% match between our model's assignment and that proposed by the epidemiological data (Fig. 4A and B). These data show that despite the high level of genome conservation and the presence of multiple genotypes within a single environmental source, it is possible to utilize signature differences in core genome SNPs to build predictive probabilistic

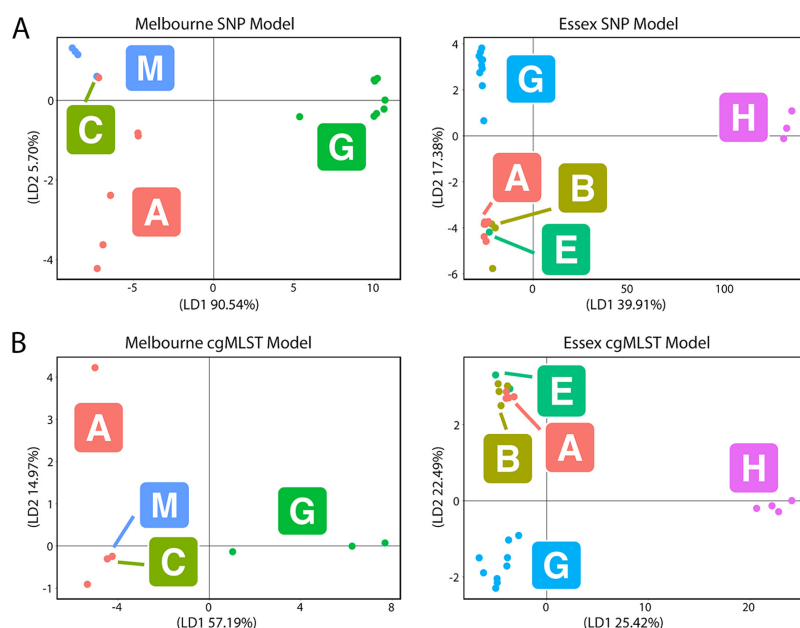


FIG 3 Scatterplots resulting from DAPC. (A) Core genome SNP-based models of the Melbourne (top left) and Essex (top right) data sets. (B) Core genome MLST (cgMLST)-based models of the Melbourne (bottom left) and Essex (bottom right) data sets. The membership of each point within an epidemiologically defined cluster (e.g., "A" is the Melbourne Aquarium outbreak) is indicated by the colored circles and the corresponding letters labeled within squares. The amount of variation explained by the first and second discriminant (LD1 and LD2, respectively) functions are specified on the axes of each plot.

classification models. The single discrepancy between model predictions and epidemiological groupings was an infection cluster C genome that was predicted to originate from the Melbourne Aquarium. Interestingly, cluster C was located closest to the Melbourne Aquarium, at a distance of approximately 500 meters. Given the proximity of clusters A and C, these data may indicate that the cooling towers were seeded from a common *L. pneumophila* source. In order to appraise the utility of this method beyond a large urban setting and the ST30 genotype, we built a sister model using 31 ST1 environmental *L. pneumophila* genomes from a previously published hospital investigation in Essex, UK, and used it to predict the origins of seven nosocomial clinical isolates (Fig. 3A and Table S1) (8). Here, the model was trained using an alignment of 59 nonrecombinogenic SNPs among the 31 environmental genomes and retaining the first 15 PCs, as this was found to be optimal. As with the Melbourne disease clusters, the model performed very well. For 86% of the clinical isolates, there was a match between the model's ward assignment and the origin suggested by epidemiology (Fig. 4A and B). Again, a single discrepancy occurred with a ward G genome predicted to originate from ward A. Wards A and G were colocated on the same corner and level of a common building, again suggesting a common *L. pneumophila* source (8). As before, isolates from a common source would be misassigned by the model, owing to the lack of location-specific genomic variants.

cgMLST has reduced discrimination. In order to evaluate the utility of the recently described core genome multilocus sequence typing (cgMLST) scheme for source tracking (25, 26), we trained new DAPC models for both the Melbourne and Essex

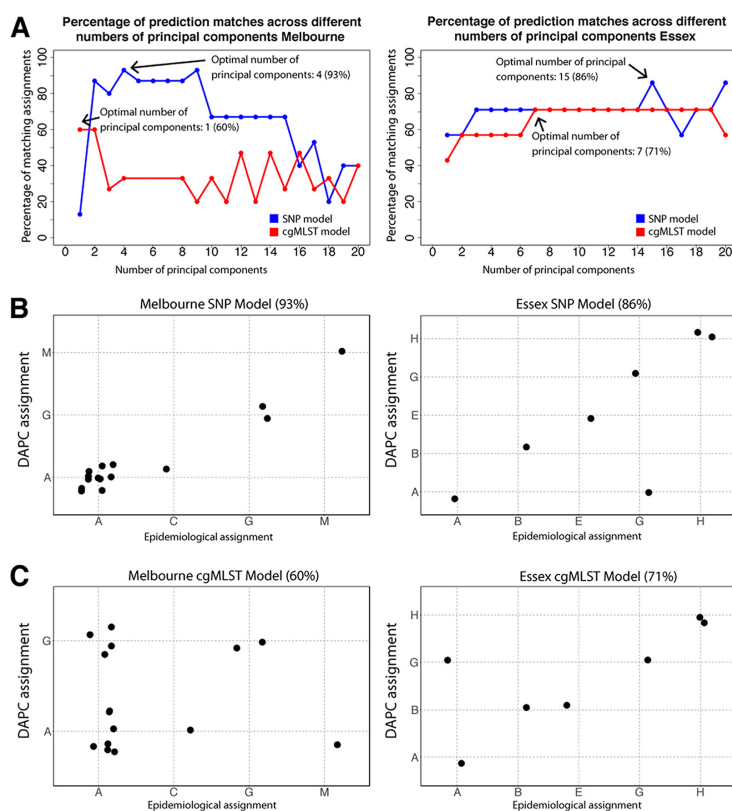


FIG 4 DAPC modeling of Lpn-SG1 genomic data. (A) Model comparison plots depicting the percentage of matches between the predicted and epidemiologically determined groupings of the validation set genomes across a range of 1 to 20 principal components for SNP and cgMLST DAPC models for the Melbourne and Essex data sets. The retention of four and one principal component was found to be optimal for the SNP (93% match) and cgMLST (60% match) models in Melbourne, respectively, while 15 and seven principal components were found to be optimal for the SNP (86% match) and cgMLST (71% match) models in the Essex hospital, respectively. (B) Assignment plots depicting the ability of the SNP models to predict the source attribution of the validation set clinical isolate genomes for the Melbourne and Essex hospital data sets. (C) Assignment plots depicting the ability of the cgMLST models to predict the source attribution of the validation set clinical isolate genomes for the Melbourne and Essex hospital data sets.

hospital data sets using a matrix of allelic integers derived from SNP profiles of the 1,529 cgMLST loci (Fig. 3B). When using the first PC and seven PCs, we observed only 60% and 71% concordance between our model's assignment and that predicted by the epidemiological data for the Melbourne and Essex hospital data sets, respectively (Fig. 4A and C).

DISCUSSION

In this study, we have retrospectively examined a large collection of 234 clinical and environmental Lpn-SG1 isolates spanning 29 defined outbreaks. Isolates were collected over wide temporal and spatial scales, and detailed genomic comparisons revealed wide extremes of Lpn-SG1 genetic diversity among distinct genomic populations, a

phenomenon not fully appreciated from previous genomic investigations that have sampled less extensively and focused on single outbreaks (5, 6, 27). Most striking in our collection was the high sequence conservation and dominance of a single genotype (BAPS-5, ST30) shared by 77% of isolates, with a median core SNP distance of only 5 SNPs across 21 years. In agreement with our findings, two recent population genomic investigations of Lpn-SG1 also describe the unusual restriction in core genome diversity (8, 12).

Based on our previous experience with other bacterial pathogens (28) and reports in the literature of Legionnaires' disease outbreak investigations using genomics (27, 29), we expected to be able to use Lpn-SG1 genomic comparisons and develop genetic rule-in or rule-out criteria to guide outbreak assessment and source attribution. For example, we recently proposed a "traffic light" system for *Listeria monocytogenes* based on SNP difference cutoffs of "likely related," "possibly related," and "not related" (28). This approach has also been proposed for *L. pneumophila* (25). A comparison of genotyping approaches using 335 *L. pneumophila* isolates, including 106 isolates from the European Society for Clinical Microbiology Study Group's *Legionella* Typing Panel, proposed an escalating hierarchical approach to genotyping, beginning with an extended 50-gene MLST scheme up to a 1,529-gene cgMLST (25, 26).

The analysis of the population structure of Lpn-SG1 presented here indicates that SNP-based typing with threshold cutoffs, whether they are based on seven genes, 50 genes, 1,500 genes, or whole genomes, will not necessarily provide sufficient discriminative power. These genotyping approaches will be confounded by the presence of (i) indistinguishable Lpn-SG1 genotypes present in unrelated cases and (ii) polyclonal outbreaks. Our retrospective analysis of the Melbourne Aquarium outbreak illustrates both of these issues clearly, where five distinct subtypes were recovered from 25 clinical and environmental isolates (Fig. 1B). There is a growing awareness of single-source polyphyletic Lpn-SG1 outbreaks (8, 10, 13, 30, 31). These data all point to the need for a different approach in order to use molecular epidemiology and genomics in support of *Legionella* outbreak investigations.

We address this issue by exploiting all core genome information to train probabilistic classification models. Our DAPC analysis demonstrates that it is possible to build predictive models based on Lpn-SG1 environmentally derived genomes that help in identifying the source of clinical isolates during complex outbreak investigations in both the community and hospital environments (Fig. 4). By including all core SNP variations, DAPC was able to identify outbreak-specific genotypes, even when the source of the outbreak was polyclonal. This enabled us to build robust models that assigned validation set genomes, with known provenance, back to their original groupings with high concordance. The fact that this model was built purely from environmental surveillance isolates demonstrates that such approaches can be developed prospectively and be preexisting, ready to deploy at the onset of outbreaks.

In contrast to the high performance of the DAPC model developed from core genome SNPs, the model built using variants identified by cgMLST scheme had a lower matching rate when assigning validation genomes back to their putative epidemiological groupings (Fig. 4C). Despite cgMLST being a useful tool for broad Lpn-SG1 population structure assessment, our analysis suggests it may have insufficient resolution and thus predictive capacity for outbreak investigations.

The DAPC approach, however, while promising, does not permit discrimination among isolates that do not belong to defined clusters. This is because the model assumes that the world is composed of only the k groups used to train it, and it therefore assigns unknown isolates to one of these groups, even if the isolate is known not to be part of any of the groups. One way to address this issue would be to create a single group classifier that is trained with environmental samples. Isolates with a low probability of membership in this single large group would then be excluded before being analyzed with the multigroup model. Future models could be further improved by adding epidemiological evidence (e.g., patient zip codes) and assessing how that improves our assignment of a clinical isolate to a particular location. An advantage of

a classification-based model is that its output could be distilled down to a zip code (or group of zip codes) and a probability that a clinical isolate is associated with the zip code (indicating uncertainty about the classification). This would obviate the need to interpret and explain phylogenetic trees. Interpreting trees is often not intuitive, and trees may fail to communicate what action is required from a public health perspective. Crucial to such a classification approach to work, however, is an extensive and temporally dynamic database of environmental Lpn-SG1 genotypes. That is, there would need to be ongoing surveillance and isolation of Lpn-SG1 from environmental sources. We are currently investigating how to implement such models.

The modeling approach is not intended to be used in isolation but rather employed as an adjunct to traditional epidemiological investigations. In this way, insights gained through epidemiological investigations can be informed by microbiological evidence from our predictive models. A limitation of our current models is the relatively small sample sizes. Performance measures for validation sets this small are often sensitive to slight perturbations in the data and may be influenced by small features of the data. However, as a proof-of-concept implementation of our approach, we have built two models from independent data sets, and both demonstrate high predictive capacity. More robust appraisals of model performance will require validation with larger data sets, collected prospectively.

From a biological perspective, the lack of genetic diversity in Lpn-SG1 over such coarse temporal and spatial scales is potentially explained by a reservoir of latent-state bacteria intermittently seeding warm-water sources in the greater Melbourne region and is supported by the frequently reported and widespread presence of *Legionella* species in drinking water supply systems (DWSS) (32–34). Independent studies propose similar hypotheses to explain the surprisingly high sequence conservation among some *L. pneumophila* genomes (8, 12).

This study is, to our knowledge, the largest genomic investigation of environmental and clinical *Legionella* spp. reported to date from a single jurisdiction and confirms that Lpn-SG1 is an unusual “edge case” in the application of genomics in public health microbiology. In the absence of a deep understanding of local *L. pneumophila* population structure (both clinical and environmental), the combination of extreme genomic monomorphism combined with outbreaks caused by mixed pathogen populations could easily lead to erroneous conclusions regarding source attribution. Thus, we require new approaches that can better utilize the genomic information available and harmoniously combine it with epidemiological evidence in order to provide public health officials with useful and timely information.

MATERIALS AND METHODS

Bacterial strains, growth conditions, and case definitions. *Legionella pneumophila* serogroup 1 isolates were resuscitated from -80°C storage and assessed. Duplicate isolates from the same patient were excluded from the study. Isolates were cultured for 48 to 72 h at 37°C on buffered charcoal yeast extract (BCYE) agar and reconfirmed to be serogroup 1 by latex agglutination (Oxoid). Metadata collected on all isolates included the year of isolation and country or city of isolation. Cases residing in the state of Victoria, Australia, were assessed by the Victorian State Government public health unit, in accordance with national guidelines, and an outbreak investigation was initiated when common exposures were reported by different cases whose onset dates occurred within a 2-week window (<http://www.health.gov.au/internet/main/publishing.nsf/content/cdna-song-legionella.htm>). In this manner, we were able to determine the human cases that were epidemiologically linked to each other. Many of the outbreaks/infection clusters contained a greater number of cases than there were isolates, as the diagnosis of legionellosis was made by culture-independent methods. Complete closed genomes of *L. pneumophila* that were publicly available were obtained from GenBank for inclusion in the analysis (Table S1).

Sequence-based typing. Sequence-based typing was performed as previously described, according to the European Legionnaires' Disease Surveillance Network (ELDSNet) method (http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) (35).

DNA sequencing. DNA libraries were prepared using the Nextera XT DNA preparation kit (Illumina), and whole-genome sequencing was performed on the NextSeq platform (Illumina) with 2×150 -bp chemistry. For single-molecule real-time (SMRT) sequencing (Pacific Biosciences), genomic DNA was extracted from agarose plugs using the CDC PulseNet protocol to allow for recovery of high-molecular-weight intact DNA (<http://www.cdc.gov/pulsenet/pathogens>). Size-selected 10-kb DNA libraries were prepared according to the manufacturers' instructions and sequenced on the RSII platform (Pacific Biosciences) using P6-C4 chemistry.

Legionella pneumophila serogroup 1 isolate Lpm7613 assembly and closure. A high-quality finished ST30 reference genome was established for *L. pneumophila* serogroup 1 clinical isolate Lpm7613 using the SMRT Analysis system version 2.3.0.140936 (Pacific Biosciences). Raw sequence data were *de novo* assembled using the HGAP version 3 protocol, with a genome size of 4 Mb. Polished contigs were error corrected using Quiver version 1. The resulting assembly was then checked using BridgeMapper version 1 in the SMRT Analysis system, and the consensus sequence was corrected with short-read Illumina data using the program Snippy (<https://github.com/tseemann/snippy>). Whole-genome annotation was performed using Prokka (17), preferentially using the *L. pneumophila* Paris strain annotation (16). BRIG was used to visualize BLASTn DNA:DNA comparisons of *L. pneumophila* Lpm7613 against other *L. pneumophila* genomes (36). Nomenclature of the genomic islands demonstrated in *L. pneumophila* Lpm7613 was based on previously described islands (37). CRISPR databases were used to search for CRISPR sequences (<http://crispi.genouest.org> and <http://crispr.i2bc.paris-saclay.fr/>).

Variant detection and phylogenetic analysis. The genomes of 10 publicly available complete *L. pneumophila* strains (Table S1) were shredded to generate short *in silico* sequence reads of 250 bp, and all 244 *L. pneumophila* reads sets were mapped against the Lpm7613 reference genome using Snippy version 3.2. An alignment file from pairwise comparisons of core genome SNPs (with inferred recombining sites removed) was used as input to FastTree version 2.1.8 with double precision (38) to infer a maximum likelihood phylogenetic tree using the general time reversible model of nucleotide substitution. Branch support was estimated using 1,000 bootstrap replicates. The resulting trees were visualized in FigTree version 1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>). Single nucleotide polymorphism (SNP) differences between isolates were tabulated and visualized using a custom R script (https://github.com/MDU-PHL/pairwise_snp_differences). The core genome SNPs were also used as the input into a Bayesian analysis of population structure (BAPS) using iterative clustering to a depth of 10 levels and a prespecified maximum of 20 clusters (39).

Recombination and molecular clock analysis. Recombination detection was performed using ClonalFrameML (40), taking as input a full-genome alignment (included invariant sites) prepared using Snippy, as described above, and the maximum likelihood (ML) phylogeny as a guide tree, with polymorphisms removed from the FastTree using a custom python script (https://github.com/kwongj/nw_multi2bifurcation). The results were visualized using a custom Python script to render separate and superposable images of extant and ancestral inferred recombination regions (<https://github.com/kwongj/cfml-maskrc>). Molecular clock-likeness of the ML tree with ClonalFrameML-adjusted branch lengths was assessed using TempEst version 1.5 (<http://tree.bio.ed.ac.uk/software/tempest/>).

Phylogeographic analysis. Variant detection for the 64 environmental genomes was undertaken by running snippy-core. Core SNPs were used to reconstruct a phylogenomic tree with FastTree that was overlaid upon a base map in GenGIS (41). Victorian population mesh data were downloaded from the Australia Bureau of Statistics webpage (<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument>), and Local Government Area data were downloaded from the Victorian Government Data Directory webpage (<https://www.data.vic.gov.au/data/dataset/lga-geographical-profiles-2014-beta/resource/f6c49074-0679-4c79-a0db-04dac8eda364>).

DAPC model building using core SNPs. Discriminant analysis of principal components (DAPC) is a multivariate method that tries to reconstruct hypothesized subdivisions in a given population (typically formed from demographic or phenotypic information) using genomic data (42). DAPC was implemented in the R package *adegenet* version 2.0.1 (42). For input, we used a matrix of SNPs for all genomes originating from infection clusters that possessed at least one environmental and clinical representative (Table S1). SNP detection was undertaken by running Snippy, and sites that were recombinogenic and/or invariant among the environmental genomes were discarded. An input SNP matrix of exclusively environmental isolates (here referred to as the training set) was used to develop a DAPC model. The training set subdivisions were based on the geographic origin of the environmental isolates (Table S1) (23). The resultant model was then tested using clinical isolates (here referred to as the validation set). The ability of the model to predict the environmental source of the validation set genomes was simulated across the 1st to the 20th principal components, allowing an optimal number of principal components to be identified. The optimized model was then used to predict the environmental origins of the clinical isolate genomes.

DAPC model building using cgMLST variation. In order to detect variants within the recently described cgMLST regions, reads were mapped to the *L. pneumophila* Philadelphia chromosome (accession no. [NC_002942.5](https://ncbi.nlm.nih.gov/nucl/NC_002942.5)) using Snippy. SNP profiles from within the cgMLST regions were reduced to allelic integers, with all genes containing zero coverage or uncertain base calls excluded. Allelic integers were concatenated into a matrix and, using the same DAPC model building method as mentioned above, models were established using the training set environmental genomes and used to predict the origins of the validation set clinical isolate genomes.

Accession number(s). All sequence reads and the completed genome are available under GenBank BioProject no. [PRJEB13594](https://ncbi.nlm.nih.gov/bioproject/PRJEB13594) and [PRJNA319136](https://ncbi.nlm.nih.gov/bioproject/PRJNA319136).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/AEM.01482-17>.

SUPPLEMENTAL FILE 1, PDF file, 1.9 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.1 MB.

REFERENCES

- Fields BS, Benson RF, Besser RE. 2002. *Legionella* and Legionnaires' disease: 25 years of investigation. Clin Microbiol Rev 15:506–526. <https://doi.org/10.1128/CMR.15.3.506-526.2002>.
- Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schoushoe M, Widmer A, Summersgill J, File T, Heath CM, Paterson DL, Cheresky A. 2002. Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. J Infect Dis 186:127–128. <https://doi.org/10.1086/341087>.
- Luck C, Fry NK, Helbig JH, Jarraud S, Harrison TG. 2013. Typing methods for *Legionella*. Methods Mol Biol 954:119–148. https://doi.org/10.1007/978-1-62703-161-5_6.
- Mercante JW, Winchell JM. 2015. Current and emerging *Legionella* diagnostics for laboratory and outbreak investigations. Clin Microbiol Rev 28:95–133. <https://doi.org/10.1128/CMR.00029-14>.
- McAdam PR, Vander Broek CW, Lindsay DS, Ward MJ, Hanson MF, Gillies M, Watson M, Stevens JM, Edwards GF, Fitzgerald JR. 2014. Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biol 15:504. <https://doi.org/10.1186/PREACCEPT-1675723368141690>.
- Reuter S, Harrison TG, Koser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Torok ME. 2013. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open 3. <https://doi.org/10.1136/bmjopen-2012-002175>.
- Sánchez-Busó L, Comas I, Jorques G, Gonzalez-Candelas F. 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nat Genet 46:1205–1211. <https://doi.org/10.1038/ng.3114>.
- David S, Afshar B, Mentasti M, Ginevra C, Podglajen I, Harris SR, Chalker VJ, Jarraud S, Harrison TG, Parkhill J. 2017. Seeding and establishment of *Legionella pneumophila* in hospitals: implications for genomic investigations of nosocomial Legionnaires' disease. Clin Infect Dis 64: 1251–1259. <https://doi.org/10.1093/cid/cix153>.
- Weiss D, Boyd C, Rakeman JL, Greene SK, Fitzhenry R, McProud T, Musser K, Huang L, Kornblum J, Nazarian EJ, Fine AD, Braunstein SL, Kass D, Landman K, Lapiere P, Hughes S, Tran A, Taylor J, Baker D, Jones L, Kornstein L, Liu B, Perez R, Lucero DE, Peterson E, Benowitz I, Lee KF, Ngai S, Stripling M, Varma JK, South Bronx Legionnaires' Disease Investigation Team. 2017. A large community outbreak of Legionnaires' disease associated with a cooling tower in New York City, 2015. Public Health Rep 132:241–250. <https://doi.org/10.1177/0033354916689620>.
- Sánchez-Busó L, Guiral S, Crespi S, Moya V, Camaro ML, Olmos MP, Adrian F, Morera V, Gonzalez-Moran F, Vanaclocha H, Gonzalez-Candelas F. 2015. Genomic investigation of a legionellosis outbreak in a persistently colonized hotel. Front Microbiol 6:1556. <https://doi.org/10.3389/fmicb.2015.01556>.
- Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG. 2013. Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol 13:302. <https://doi.org/10.1186/1471-2180-13-302>.
- David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, Lees J, Ginevra C, Glaser C, Ma L, Bouchier C, Underwood A, Jarraud S, Harrison TG, Parkhill J, Buchrieser C. 2016. Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. Genome Res 26:1555–1564. <https://doi.org/10.1101/gr.209536.116>.
- Schjorring S, Stegger M, Kjelsø C, Lilje B, Bangsbo JM, Petersen RF, David S, Uldum SA, ESCMID Study Group for Legionella Infections. 2017. Genomic investigation of a suspected outbreak of *Legionella pneumophila* ST82 reveals undetected heterogeneity by the present gold-standard methods, Denmark, July to November 2014. Euro Surveill 22:pil=30558. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22821>.
- Anonymous. 2014. Communicable disease surveillance, Victoria, Oct–Dec 2014. Vic Infect Dis Bull 17:22–23.
- Greig JE, Carnie JA, Tallis GF, Ryan NJ, Tan AG, Gordon IR, Zwolak B, Leydon JA, Guest CS, Hart WG. 2004. An outbreak of Legionnaires' disease at the Melbourne Aquarium, April 2000: investigation and case-control studies. Med J Aust 180:566–572.
- Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C. 2004. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet 36:1165–1173. <https://doi.org/10.1038/ng1447>.
- Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
- Rao C, Guyard C, Pelaz C, Wasserscheid J, Bondy-Denomy J, Dewar K, Ensinger AW. 2016. Active and adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell Microbiol 18:1319–1338. <https://doi.org/10.1111/cmi.12586>.
- Coscollá M, Gonzalez-Candelas F. 2007. Population structure and recombination in environmental isolates of *Legionella pneumophila*. Environ Microbiol 9:643–656. <https://doi.org/10.1111/j.1462-2920.2006.01184.x>.
- Coscollá M, Gonzalez-Candelas F. 2009. Comparison of clinical and environmental samples of *Legionella pneumophila* at the nucleotide sequence level. Infect Genet Evol 9:882–888. <https://doi.org/10.1016/j.meegid.2009.05.013>.
- Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J, Buchrieser C. 2011. Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. BMC Genomics 12:536. <https://doi.org/10.1186/1471-2164-12-536>.
- Costa J, Teixeira PG, d'Avo AF, Junior CS, Verissimo A. 2014. Intragenic recombination has a critical role on the evolution of *Legionella pneumophila* virulence-related effector *sidL*. PLoS One 9:e109840. <https://doi.org/10.1371/journal.pone.0109840>.
- Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94. <https://doi.org/10.1186/1471-2156-11-94>.
- McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Valimaki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaeffer K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet 12: e1006280. <https://doi.org/10.1371/journal.pgen.1006280>.
- David S, Mentasti M, Tewolde R, Aslett M, Harris SR, Afshar B, Underwood A, Fry NK, Parkhill J, Harrison TG. 2016. Evaluation of an optimal epidemiologic typing scheme for *Legionella pneumophila* with whole genome sequence data using validation guidelines. J Clin Microbiol 54: 2135–2148. <https://doi.org/10.1128/JCM.00432-16>.
- Moran-Gilad J, Prior K, Yakunin E, Harrison TG, Underwood A, Lazarovitch T, Valinsky L, Luck C, Krux F, Agmon V, Grotto I, Harmsen D. 2015. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. Euro Surveill 20: pii=21186. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=21186>.
- Bartley PB, Ben Zakour NL, Stanton-Cook M, Muguli R, Prado L, Garnys V, Taylor K, Barnett TC, Pinna G, Robson J, Paterson DL, Walker MJ, Schembri MA, Beatson SA. 2016. Hospital-wide eradication of a nosocomial *Legionella pneumophila* serogroup 1 outbreak. Clin Infect Dis 62: 273–279. <https://doi.org/10.1093/cid/civ870>.
- Kwong JC, Mercouliou K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP. 2016. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol 54:333–342. <https://doi.org/10.1128/JCM.02344-15>.
- Graham RM, Doyle CJ, Jennison AV. 2014. Real-time investigation of a *Legionella pneumophila* outbreak using whole genome sequencing. Epidemiol Infect 142:2347–2351. <https://doi.org/10.1017/S0950268814000375>.
- Sánchez-Busó L, Olmos MP, Camaro ML, Adrian F, Calafat JM, Gonzalez-Candelas F. 2015. Phylogenetic analysis of environmental *Legionella pneumophila* isolates from an endemic area (Alcoy, Spain). Infect Genet Evol 30:45–54. <https://doi.org/10.1016/j.meegid.2014.12.008>.
- Coscollá M, Fernandez C, Colomina J, Sanchez-Buso L, Gonzalez-Candelas F. 2014. Mixed infection by *Legionella pneumophila* in outbreak patients. Int J Med Microbiol 304:307–313. <https://doi.org/10.1016/j.jmm.2013.11.002>.
- Diederer BM, de Jong CM, Aarts I, Peeters MF, van der Zee A. 2007. Molecular evidence for the ubiquitous presence of *Legionella* species in Dutch tap water installations. J Water Health 5:375–383. <https://doi.org/10.2166/wh.2007.033>.
- Rivera JM, Aguilar L, Granizo JJ, Vos-Arenilla A, Gimenez MJ, Aguilar JM,

- Prieto J. 2007. Isolation of *Legionella* species/serogroups from water cooling systems compared with potable water systems in Spanish healthcare facilities. *J Hosp Infect* 67:360–366. <https://doi.org/10.1016/j.jhin.2007.07.022>.
34. Storey MV, Langmark J, Ashbolt NJ, Stenstrom TA. 2004. The fate of legionellae within distribution pipe biofilms: measurement of their persistence, inactivation and detachment. *Water Sci Technol* 49:269–275.
35. Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TG. 2005. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol* 43:2047–2052. <https://doi.org/10.1128/JCM.43.5.2047-2052.2005>.
36. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. <https://doi.org/10.1186/1471-2164-12-402>.
37. D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A. 2010. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. *BMC Genomics* 11:181. <https://doi.org/10.1186/1471-2164-11-181>.
38. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. <https://doi.org/10.1371/journal.pone.0009490>.
39. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* 30:1224–1228. <https://doi.org/10.1093/molbev/mst028>.
40. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11:e1004041. <https://doi.org/10.1371/journal.pcbi.1004041>.
41. Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, Langille MG, Beiko RG. 2013. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLoS One* 8:e69885. <https://doi.org/10.1371/journal.pone.0069885>.
42. Jombart T, Ahmed I. 2011. *adegenet* 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071. <https://doi.org/10.1093/bioinformatics/btr521>.

Summary

From this work the key findings were that:

- That the Melbourne Aquarium outbreak isolates had high sequence conservation and dominance of a single genotype.
- That SNP-based typing with threshold cutoffs will not necessarily provide sufficient discriminative power as such approaches will be confounded by the presence of indistinguishable genotypes present in unrelated cases and polyclonal outbreaks.
- That it is possible to build predictive models based on environmentally derived genomes that help in identifying the source of clinical isolates during complex outbreak investigations in both the community and hospital environments.

Chapter Four:

Explaining epidemiological trends through genomic insights

Comparative genomics shows Mycobacterium ulcerans migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia

Introduction

Determining a pathogen's mode of transmission and infectious reservoir is a major goal in the fight against infectious diseases, as such knowledge allows for the deployment of evidence-based strategies to manage outbreaks and prevent future occurrences. Buruli ulcer is a neglected tropical disease for which both the mode of transmission and infectious reservoirs remain unknown. Studies have employed conventional genotyping technologies to assess the relatedness of disease isolates, however they have failed to provide sufficient detail to conduct micro-epidemiological studies (204-208, 210). This shortcoming can be explained as *M. ulcerans* populations have little genetic diversity and that such technologies only interrogate minute regions of the genome.

In this chapter, genomics was used to provide greater strain discriminating ability and to model the historical spatial and temporal as well as population demographic dynamics of *M. ulcerans* in Australia. The following research questions were investigated:

- What are the evolutionary relationships among Australian *M. ulcerans* isolates?
- Can genomics provide insights into unexplained epidemiological trends of case reporting?
- Can genomics reveal a deeper truth on the origins of *M. ulcerans* in Australia?

Published manuscript: “Comparative genomics shows *Mycobacterium ulcerans* migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia”

The methods and results for this research study were reported in a manuscript published in Applied and Environmental Microbiology:

Andrew H. Buultjens, Koen Vandellannoote, Conor J. Meehan, Miriam Eddyani, Bouke C. de Jong, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Ee Laine Tay, Torsten Seemann, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear. Comparative genomics shows *Mycobacterium ulcerans* migration and expansion has preceded the rise of Buruli ulcer in south-eastern Australia. *J Appl Environ Microbiol.* 2018 Feb; doi:10.1128/AEM.02612-17.



Applied and Environmental
Microbiology®

PUBLIC AND ENVIRONMENTAL
HEALTH MICROBIOLOGY



Comparative Genomics Shows That *Mycobacterium ulcerans* Migration and Expansion Preceded the Rise of Buruli Ulcer in Southeastern Australia

Andrew H. Buultjens,^{a,b} Koen Vandelannootte,^c Conor J. Meehan,^c Miriam Eddyani,^c Bouke C. de Jong,^c Janet A. M. Fyfe,^d Maria Globan,^d Nicholas J. Tobias,^a Jessica L. Porter,^a Takehiro Tomita,^e Ee Laine Tay,^f Torsten Seemann,^{b,g} Benjamin P. Howden,^{b,e,h} Paul D. R. Johnson,^h Timothy P. Stinear^{a,b}

^aDepartment of Microbiology and Immunology, Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

^bDoherty Applied Microbial Genomics, Department of Microbiology and Immunology, Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

^cUnit of Mycobacteriology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

^dVictorian Infectious Diseases Reference Laboratory, Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

^eMicrobiology Diagnostic Unit, Department of Microbiology and Immunology, Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

^fHealth Protection Branch, Department of Health & Human Services, Melbourne, Victoria, Australia

^gMelbourne Bioinformatics, University of Melbourne, Melbourne, Victoria, Australia

^hDepartment of Infectious Diseases, Austin Health, Heidelberg, Victoria, Australia

ABSTRACT Since 2000, cases of the neglected tropical disease Buruli ulcer, caused by infection with *Mycobacterium ulcerans*, have increased 100-fold around Melbourne (population 4.4 million), the capital of Victoria, in temperate southeastern Australia. The reasons for this increase are unclear. Here, we used whole-genome sequence comparisons of 178 *M. ulcerans* isolates obtained primarily from human clinical specimens, spanning 70 years, to model the population dynamics of this pathogen from this region. Using phylogeographic and advanced Bayesian phylogenetic approaches, we found that there has been a migration of the pathogen from the east end of the state, beginning in the 1980s, 300 km west to the major human population center around Melbourne. This move was then followed by a significant increase in *M. ulcerans* population size. These analyses inform our thinking around Buruli ulcer transmission and control, indicating that *M. ulcerans* is introduced to a new environment and then expands, rather than it being from the awakening of a quiescent pathogen reservoir.

IMPORTANCE Buruli ulcer is a destructive skin and soft tissue infection caused by *Mycobacterium ulcerans* and is characterized by progressive skin ulceration, which can lead to permanent disfigurement and long-term disability. Despite the majority of disease burden occurring in regions of West and central Africa, Buruli ulcer is also becoming increasingly common in southeastern Australia. Major impediments to controlling disease spread are incomplete understandings of the environmental reservoirs and modes of transmission of *M. ulcerans*. The significance of our research is that we used genomics to assess the population structure of this pathogen at the Australian continental scale. We have then reconstructed a historical bacterial spread and modeled demographic dynamics to reveal bacterial population expansion across southeastern Australia. These findings provide explanations for the observed epidemiological trends with Buruli ulcer and suggest possible management to control disease spread.

Received 6 December 2017 Accepted 25 January 2018

Accepted manuscript posted online 9 February 2018

Citation Buultjens AH, Vandelannootte K, Meehan CJ, Eddyani M, de Jong BC, Fyfe JAM, Globan M, Tobias NJ, Porter JL, Tomita T, Tay EL, Seemann T, Howden BP, Johnson PDR, Stinear TP. 2018. Comparative genomics shows that *Mycobacterium ulcerans* migration and expansion preceded the rise of Buruli ulcer in southeastern Australia. *Appl Environ Microbiol* 84:e02612-17. <https://doi.org/10.1128/AEM.02612-17>.

Editor Andrew J. McBain, University of Manchester

Copyright © 2018 American Society for Microbiology. All Rights Reserved.

Address correspondence to Andrew H. Buultjens, abuultjens@student.unimelb.edu.au, or Timothy P. Stinear, tstinear@unimelb.edu.au.

Downloaded from <http://aem.asm.org/> on August 9, 2018 by University of Melbourne

KEYWORDS Australia, Buruli ulcer, *Mycobacterium ulcerans*, genomics, phylogenetics, population genetics, public health

Buruli ulcer (BU) is a recently emerged disease, first described by Australian scientists in 1948, when it appeared in six patients in the Gippsland region of Victoria, southeastern Australia (1). BU is caused by *Mycobacterium ulcerans* (1), an atypical mycobacterial pathogen that has evolved from a *Mycobacterium marinum* progenitor (2, 3). The evolution of this pathogen is characterized by the acquisition of the pMUM001 virulence plasmid (4, 5), allowing production of the polyketide lipid toxin mycolactone (6) and the expansion of two distinct insertion sequence elements (IS2404 and IS2606) (7). These evolutionary events have facilitated genome plasticity and size reduction, likely indicating an adaptive response toward a more stable protected niche environment (7). Mycolactone is largely responsible for the debilitating pathology of BU (6). Although typically presenting as a deeply undermined skin ulcer, there are also preulcerative stages, such as plaques, papules, nodules, and edema (8), which can progress to ulcers. Buruli ulcer may result in extensive damage to skin, soft tissue, and in some cases, bone (9). When left untreated, the disease can result in scarring, contractures, and associated functional deformities (6, 10, 11).

BU has been reported in more than 30 countries (12) across Asia, the Western Pacific, and the Americas, with the main burden of disease situated in regions of West and central Africa (13). In Australia, the occurrence of BU has been recorded in the Northern Territory (NT) (14), Queensland (QLD) (15–17), and Victoria (18) (Fig. 1A). In Victoria, BU disease foci include Gippsland (19), Phillip Island (20), Frankston/Langwarrin (18), and the Mornington (18) and Bellarine (21–23) Peninsulas (Fig. 1B). In Victoria, the disease has been intensively studied and was made notifiable in 2004, allowing access to detailed patient demographics (24); thus, this region presents as an attractive option for delineating the ecology of *M. ulcerans*.

Commencing in the Gippsland region of Victoria, where BU first caused disease in Australia in 1939 (18, 25), there has been a westward trend of BU case reporting into Western Port and Port Phillip regions beginning in the 1980s (18). Given that *M. ulcerans* has been detected in Africa (26) and Victoria (27) both in areas where the disease is endemic and in areas where it is not endemic, the question arises as to whether *M. ulcerans* was already dispersed as a quiescent pathogen reservoir in the Western Port and Port Phillip environments or has been introduced westward from Gippsland, causing ulcerative disease along its migratory front. Furthermore, following the emergence of cases west of Gippsland, there has been a substantial increase in cases reported over time, with a 100-fold increase in the number of notified cases per year since the year 2000. This may be related to increased human population density in areas in which the disease is endemic or, alternatively, that the bacterial population is growing and humans are exposed more frequently.

Due to the rarity of interhuman transmission and frequent onset of disease following environmental contact, it is widely held that the source of *M. ulcerans* is of environmental origin (28–32). In southeastern Australia, native marsupials, in particular common ringtail (*Pseudocheirus peregrinus*) and brushtail (*Trichosurus vulpecula*) possums, and mosquitos have been identified as playing roles in the ecology and transmission of *M. ulcerans* (23, 27, 33–35). However, despite intense research efforts aimed at delineating the ecology of *M. ulcerans*, both its mode of transmission and exact environmental reservoir(s) remain elusive, representing a major research priority for this disease. One avenue to further investigate the ecology of *M. ulcerans* is to trace the spatial and temporal history of isolates cultured from human and animal cases using genotyping with sufficient resolution to differentiate closely related isolates. Past attempts to examine the ecology of *M. ulcerans* in Victoria have used conventional genotyping technologies (36–45); however, these methods provided inadequate resolution for *M. ulcerans*, as they are unable to distinguish between closely related strains. The low resolution afforded by these techniques can be explained in that they sample



FIG 1 Map of Australia and Victorian areas of endemicity for BU. (A) Map of Australia and Papua New Guinea. Australian states and territories endemic for BU are labeled. (B) Map of Victorian areas of endemicity for BU, represented by colors as indicated in the key. (Maps adapted from reference 81.)

relatively small genomic regions and that *M. ulcerans* populations exhibit very limited genetic diversity (46).

Genotyping methods based upon whole-genome-derived single nucleotide polymorphisms (SNPs) have enabled major breakthroughs in *M. ulcerans* strain disambiguation. The first of its kind, Röltgen et al. (47) developed a limited SNP genotyping assay that provided high-resolution rendering of the pathogen population structure in Ghana; however, this method was susceptible to phylogenetic discovery bias, as its

canonical SNP panel was based upon whole-genome comparisons of only seven *M. ulcerans* strains (47). One approach to utilize the high genotyping resolution offered by SNP markers while reducing the complications associated with biased discovery is to undertake whole-genome sequencing of all isolates in a study population (48). This approach involves the sequencing of entire genomes and aligning sequence reads to a high-quality reference genome, allowing for *in silico* identification of SNP variants (49, 50). Owing to the nature of this approach, microbial genomics examines variation at virtually all positions across a bacterial genome, hence offering the greatest state of discriminatory power achievable.

The potential of microbial genomics to study *M. ulcerans* populations was first examined using a relatively small collection of Ghanaian disease isolates (3). Several studies have since made use of genomics to conduct high-resolution molecular epidemiological investigations (51, 52), discerning between relapse and reinfections (53) and revealing deep insights into reconstructed historical evolutionary trajectories (54). In this study, we have sequenced and compared the genomes of 178 *M. ulcerans* isolates to reveal more detail about the evolutionary history and consequent population structure of *M. ulcerans* in Australia. Here, we place a special focus on disease incidence in southeastern Australia, a region that claims both the highest disease incidence in the Southern Hemisphere and the place of longest-standing enquiry aimed to understand this enigmatic pathogen, with initial investigations dating back to the 1930s (1).

RESULTS

Australian *M. ulcerans* ML phylogeny. In order to understand the evolutionary dynamics of *M. ulcerans* in Australia, we sequenced and analyzed the genomes of 176 *M. ulcerans* isolates from several major Australian locations in which the disease is endemic, as well as two isolates from Papua New Guinea (PNG). Comparisons of 6,206 core-genome SNPs allowed for high-resolution insights into the population structure at the continental scale. Comparisons of core SNP diversity both within and between Australian states revealed fewer within-state differences than between-state differences, indicating strong geographical structure (Fig. 2A). A maximum likelihood (ML) phylogeny inferred using the same polymorphisms also revealed the separation of isolates into strongly supported clades linked to broad geographic origins (Fig. 2B). Interestingly, the two isolates from PNG were phylogenetically closest to the *M. ulcerans* most recent common ancestor (MRCA), followed by QLD and then Victorian isolates, suggesting that *M. ulcerans* arrived in Australia from the north and moved progressively south, indicating that PNG might be the ancestral origin of Australian *M. ulcerans* (Fig. 2B). Victoria reports substantially more BU cases than other Australian states and was thus represented by more *M. ulcerans* isolates. Within-state SNP diversity was higher in QLD (median, 35 SNPs) than in Victoria (median, 11 SNPs). The Victorian *M. ulcerans* population was separated from all other Australian and PNG isolates by a median of 2,180 SNPs.

Assessing the temporal and spatial history of *M. ulcerans* in Victoria. Comparisons among the 165 Victorian isolate genomes revealed a total of 417 core-genome SNPs. Included in our isolate panel were two replicate isolates of historical cultures, "RS" and "RT," from the first clinical description of BU (1). Despite being isolates derived from the same patient, there were 19 and 23 core SNP differences between the RT and RS duplicates, respectively. Using both the diversity detected in the core genome and the year of isolation of the disease isolates, a rooted time-measured phylogeny was reconstructed using the BEAST 2 software (Fig. 3). As was observed at the continental scale, *M. ulcerans* genomes were found to associate with the originating geography at the regional scale. Given the geographical locations of clusters of disease endemicity depicted on the phylogenomic time tree, it was apparent that *M. ulcerans* has spread from east to west in Victoria. The branching order of the rooted Victorian time-measured phylogeny indicated that *M. ulcerans* in western regions, such as the Western Port Bay and the Port Phillip Bay, has evolved from a common progenitor originating

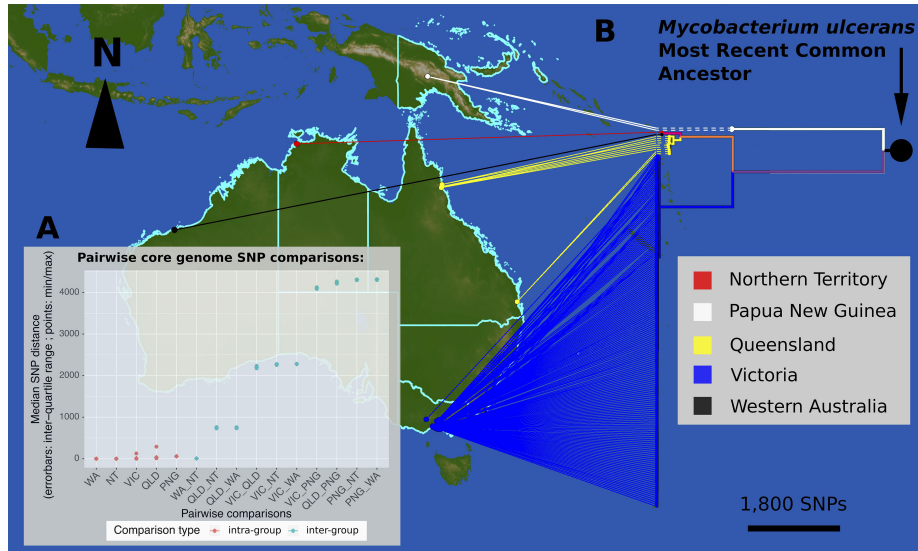


FIG 2 Genome comparisons of Australian isolate genomes. (A) Pairwise SNP comparisons of isolates from Australian states and territories and from PNG. VIC, Victoria; QLD, Queensland; WA, Western Australia; NT, Northern Territory. (B) Phylogeography of Australian and PNG isolates. Note that the *M. marinum* outgroup is not depicted in the ML tree. Australian states and territories and PNG country boundaries are indicated on the map and the colors specified in the key. The map was obtained from The Oak Ridge National Laboratory Distributed Active Archive Center (81).

from an ancestral Gippsland *M. ulcerans* population. The historical RT isolate genomes originating from Gippsland displayed substantial divergence from all other Victorian isolates, separated by a median of 123 SNPs.

To further examine the relationship between tree topology and the geography of isolate origins, the time tree was laid out on an approximately east-west cline along the Victorian coastline, with connecting lines drawn between leaf nodes and isolate origin locations (Fig. 3). This graphical representation showed that the order in which the Gippsland, Phillip Island, Frankston/Langwarrin, and Bellarine Peninsula MRCA emerged from the Victorian MRCA was the same order in which these areas of endemicity are geographically situated from east to west (Fig. 1 and 3), suggesting that *M. ulcerans* had evolved along the east-to-west cline. After 1,000 random rearrangements of connecting lines, it was found that the tree topology produced significantly fewer crossings than that which would be probable by chance alone, indicating that the relationship between tree topology and geography was statistically significant ($P < 0.05$) (Fig. S1). A single isolate (DMG1701364) was obtained from a patient living in Redan, located near Ballarat, which is approximately 100 km west of Melbourne. Patient interviews reported no exposure to areas in which the disease is likely endemic; however, the isolate's genotype strongly suggested that it had originated from either the Bellarine Peninsula or Mornington Peninsula region (Fig. 3).

As BEAST 2 can associate sampling dates directly to the sequences representing the tips (terminal nodes) of the tree, the resulting tip-calibrated timed phylogeny can allow for the inference of emergence times for key ancestral nodes. The MRCA for *M. ulcerans* in Victoria was estimated to have existed in 1766 (95% highest posterior density [HPD], 1646 to 1877), indicating that *M. ulcerans* has been extant in Victoria for at least 250 years, likely originating in the Gippsland region. This time period coincides with the first

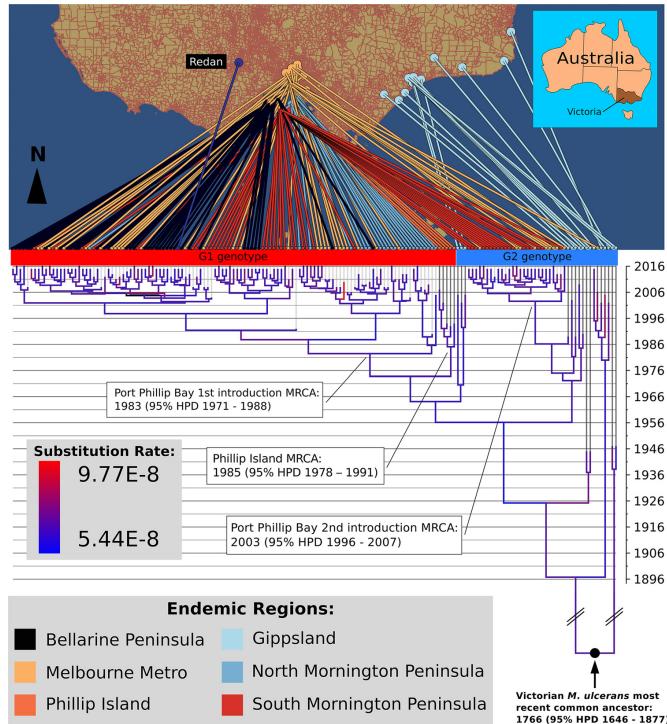


FIG 3 Victorian phylogeographic time tree. Tree tips are aligned with a map of southeastern Australia. Population density is depicted in red on the map. The horizontal lines and scale depict time in the past, starting from 2016. Branches are colored according to the rate of nucleotide substitutions, as indicated in the key. The ancestors of the first and second introductions into the Port Phillip Bay region and Phillip Island are highlighted, and the G1 and G2 genotypes are labeled. The colors of regions of endemicity in Victoria are specified in the key. The map was obtained from The Oak Ridge National Laboratory Distributed Active Archive Center (81), and state population mesh data were downloaded from the Australia Bureau of Statistics (82).

arrival of Europeans in the area, with the first European settlement of Victoria occurring in 1803 (55, 56).

The time tree indicates that since the existence of the Victorian MRCA, descendants continued to exclusively occupy the Gippsland region until a Gippsland-like ancestral progenitor, estimated to have emerged in 1985 (95% HPD, 1978 to 1991), migrated westward, giving rise to an outbreak on Phillip Island (Fig. 1 and 3). The emergence of this ancestor is estimated to have occurred 8 years before the first BU case was reported from Phillip Island in 1993 (20). A sister ancestor of the Phillip Island ancestral progenitor migrated further west into the Port Phillip Bay region (Fig. 3). This ancestor is estimated to have emerged in 1983 (95% HPD, 1971 to 1988), which is 7 years before the first BU case was reported in that region within the Frankston/Langwarrin area (Fig. 1 and 3) in 1990 (18). The lineage that migrated west out of Gippsland into the Western Port and Port Phillip regions is referred to here as the G1 genotype (Fig. 3).

The phylogeny also illustrates that there has been a more recent reintroduction of *M. ulcerans* into the Port Phillip Bay region, here referred to as the G2 genotype (Fig. 3).

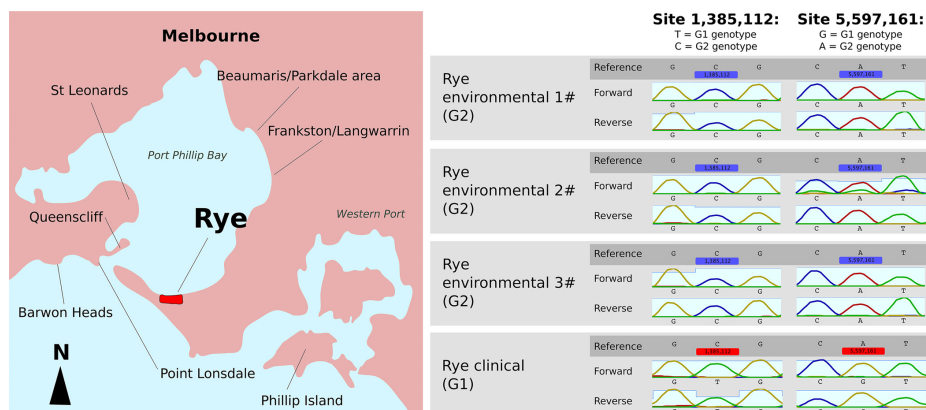


FIG 4 Direct detection of clade-specific SNPs from environmental specimens. Alignment of sequenced PCR amplicons for two canonical genomic positions revealed the presence of the G2 genotype in three environmental specimens and the G1 genotype in one clinical sample from Rye. The reference, forward, and reverse amplicon electropherograms are labeled for both genomic markers and all three samples. Major Port Phillip region areas of BU endemicity are labeled. (Map adapted from reference 81.)

This second introduction was observed in the South Mornington Peninsula region among clinical isolates from Rye in 2012 (Fig. 1 and 3). The ancestor of the G2 genotype in the South Mornington Peninsula region is estimated to have emerged in 2003 (95% HPD, 1996 to 2007), which is 9 years before the second introduction was observed among clinical isolates from Rye BU patients on the Mornington Peninsula.

Direct detection of the G2 genotype from environmental specimens. In order to further validate the existence of the Gippsland genotype in the Rye environment, a specific PCR-based genotyping assay was developed and used to interrogate the genotypes of *M. ulcerans* directly from DNA extracted from environmental specimens. Genomic DNA purified from three possum fecal samples positive for *M. ulcerans*, as verified by low cycle threshold (C_t) values with the IS2404 quantitative PCR (qPCR) assay, was used as the template for a nested PCR genotyping assay (refer to Materials and Methods). Sanger sequencing and alignment of the sequenced amplicons back to a reference sequence confirmed the presence of the Gippsland-associated genotype (C at position 1385112 and A at position 5597161) in the Rye environment (Fig. 4).

Assessing the demographic history of *M. ulcerans* in Victoria. Through the use of an Extended Bayesian Skyline Plot (EBSP) coalescent tree prior, we reconstructed the estimated demographic history of *M. ulcerans* in Victoria. An examination of the historical mycobacterial population size suggested that the Victorian *M. ulcerans* population had remained stable until approximately 2007, after which it increased dramatically (Fig. 5). This profile of a stable population size followed with rapid population expansion at approximately 2007 aligns with the increasing number of Victorian BU cases across the same timespan (Fig. 5).

Nucleotide substitution rate. Using a relaxed molecular clock model, we were able to assess branch-specific rates of nucleotide substitution (57) (Fig. 3). This approach allowed us to estimate a mean genome-wide substitution rate of 7.50×10^{-8} per site per year (95% HPD, 5.44×10^{-8} to 9.77×10^{-8}), corresponding to the accumulation of 0.39 SNPs per chromosome per year (excluding insertion sequence elements). To test the legitimacy of the estimated temporal signal in the data, we undertook 40 permutation tests by generating a set of 40 substitution rate distributions of reshuffled "null sets" of tip date and sequence associations. Assessment of the random and clustered sets depicted

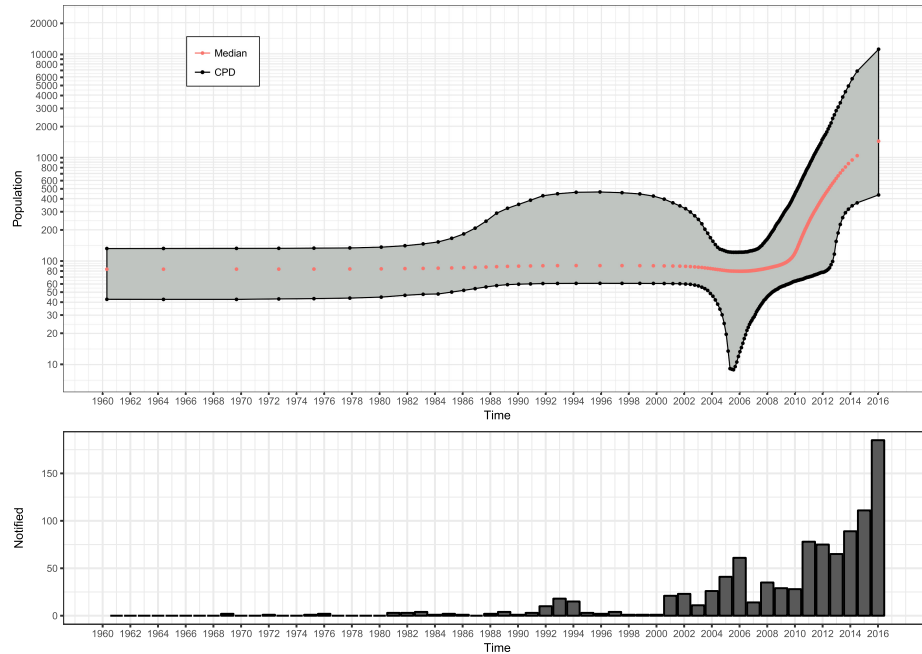


FIG 5 EBSP of the Victorian *M. ulcerans* demographic history combined with a histogram of BU cases in Victoria over the same time period. The central red line represents the median mycobacterial population size, with its 95% central posterior density (CPD) interval represented by the upper and lower black lines. Note that the y axis of the EBSP does not represent the absolute population size but a scaled one.

that the results achieved using the actual tip dates were significantly different from those expected by random chance (see Fig. S2).

DISCUSSION

Australian phylogeny. We describe the use of microbial comparative genomics to gain a better understanding of the emergence of BU in Australia. Our use of population-based analyses allowed us to explore the evolutionary dynamics of *M. ulcerans* in southeastern Australia through the study of genome changes. Comparisons of core-genome diversity between Victorian isolates against isolates from other Australian states and territories and PNG revealed evolutionary insights at the continental scale. The clear geographical clustering indicates that *M. ulcerans* has likely had single introduction events into each of the major Australian areas of endemicity. Since these introductions, the founding populations have accumulated location-specific SNPs that cluster them according to originating geography. This broad-scale phylogeographic clustering has been observed in studies focusing on BU disease in regions of Africa (54). The basal position of the PNG isolates to that of the Australian isolates suggests that Australian *M. ulcerans* may have evolved from a PNG-like ancestral progenitor. PNG isolates were also found to be ancestral to *M. ulcerans* isolates from Africa, with all isolates belonging to the more virulent "classical lineage" (54, 58). The greater genetic variation observed among QLD isolates suggests that *M. ulcerans* may have been extant in QLD for longer than it has been in Victoria. Such a scenario accords with the

pathogen progressively moving south from PNG, reaching QLD before progressing further south to Victoria.

East-to-west migration. The Victorian time-measured phylogeny depicted a close agreement between phylogenetic tree topology and the originating geography. The relationship between population structure and geography was less clear-cut than that observed at the broader continental scale, with several phylogeographic incongruences observed. This breakdown of geographic structure at more local scales has been reported in studies employing whole-genome sequencing to assess *M. ulcerans* population structure in regions of Africa (51). Nonetheless, the geographical structure in the Victorian population was quantitatively assessed and found to be significantly different from what is expected by random chance alone.

Assessment of the topology also revealed that *M. ulcerans* has evolved from east to west in Victoria, leaving Gippsland and migrating into the Western Port and Port Phillip regions. Such an east-to-west migration is fitting with a well-documented epidemiological trend in which cases were detected out of the Gippsland region and through time emerged in the Western Port and then Port Phillip regions. This finding suggests that both the Western Port and Port Phillip regions were not previously colonized with quiescent *M. ulcerans* but rather became colonized as part of the east-to-west migration out of Gippsland. In this way, our findings suggest that the epidemiological trend of case reporting is likely due to bacterial spread. Implicit in this conclusion is the notion that if we had knowledge of the reservoir moving the pathogen, biosecurity interventions could have prevented the westward pathogen migration out of Gippsland and averted the rise of endemicity in both the Western Port and Port Phillip regions. Infected humans have recently been implicated as a possible infectious reservoir (K. Vandellannoote, D. M. Phanzu, K. Kibadi, M. Eddyani, C. J. Meehan, K. Jordaens, H. Leirs, F. Portael, T. P. Stinear, S. R. Harris, and B. C. de Jong, unpublished data); however, given the prompt treatment and adequate wound management in Australia, humans with active infection are unlikely to play a major role in the ecology of *M. ulcerans* in southeastern Australia.

Fitting with the findings of a recent study that used modern population genomics to explore the evolution of *M. ulcerans* in Africa (54), spread events in southeastern Australia appear to have recently occurred in the late 19th and the early 20th centuries. The exact means by which *M. ulcerans* was introduced into the Port Phillip region remains unknown. Given the abundance of *M. ulcerans* DNA detected in possum excreta (27), the translocation of an infected possum or contaminated fecal material from Gippsland into the Rye area provides a plausible means for the spread of *M. ulcerans* into a new environment.

The likely scenario in which *M. ulcerans* spread west, rather than the awakening of quiescent populations, is also consistent with genomic insights into the pathogen's biology. Genomic analyses have indicated that *M. ulcerans* has evolved through a process of reductive evolution toward a privileged niche and away from the generalist environmental lifestyle of its ancestral progenitor (3, 7). In this way, it would be unlikely that *M. ulcerans* would exist in an unprotected environment without a host reservoir, which is thought to be marsupials in southeastern Australia (27). Furthermore, the concentration of bacteria in positive environmental samples is almost always low, suggesting that *M. ulcerans* is unlikely to be multiplying in the environment (59, 60).

Second introduction from Gippsland. There was a striking exception to the congruency between phylogeny and geography, where we observed *M. ulcerans* isolates with the Gippsland genotype (G2 genotype) cocirculating with the Port Phillip Bay region isolates (G1 genotype), particularly in the Rye area. This incongruence strongly suggests that there has been a second introduction of *M. ulcerans* into the Rye region from Gippsland. Given that disease acquisition origins are conditional upon accurate patient recollection, and the mean incubation time is 4.5 months (61), accurate source attribution is not always ensured. In this way, there is the possibility that the patients thought to have acquired *M. ulcerans* in Rye may have actually been infected

in Gippsland, explaining the apparent incongruence between the phylogeny and the epidemiology.

To eliminate this possibility and provide more ensured source attribution, *M. ulcerans* genotyping directly from environmental specimens was undertaken. Given the difficulties of culturing *M. ulcerans* from the environment (62), a PCR-based SNP typing assay that could directly assess environmental specimens was developed. Possum excreta is a spatially trustworthy analyte for environmental SNP genotyping. This is because both common ringtail and brushtail possums do not generally range further than a few kilometers (63, 64). Here, the Gippsland region is over 300 km away from Rye, well beyond the natural range of an individual possum. Possum excreta positive for *M. ulcerans* from the Rye environment were screened for specific *M. ulcerans* genetic signatures and were found to contain SNPs associated with the G2 genotype, confirming that *M. ulcerans* of the Gippsland clade indeed was circulating in the Rye environment on the Mornington Peninsula.

Introductions of a second genotype into environments already colonized with *M. ulcerans* have been reported in two previous studies in Africa (51, 54). In one study, the estimate of the temporal emergence of the second introduction is thought to have coincided with a major human activity: the start of colonial rule in Africa (54). Our temporal estimates of the ancestor of the second introduction of *M. ulcerans* into the Port Phillip Bay region in the 1980s did not align with any particular movement of people or social upheaval. However, given that marsupials native to southeastern Australia have been found to excrete *M. ulcerans* in fecal material (27), an infected possum translocated from Gippsland to the Mornington Peninsula area in Melbourne may have provided a source for subsequent pathogen dissemination. An alternative explanation would be a low-probability introduction from a long-range migratory bird or bat into a receptive population, such as colonies of ringtail possums. In this regard, individual gray-headed flying foxes (*Pteropus poliocephalus*) undertake long journeys of hundreds of kilometers (65, 66). Although there is so far no direct evidence linking birds or bats to the movement of *M. ulcerans* across long distances, this possibility has been proposed since the first description of *M. ulcerans* (67).

Our timing estimates for the temporal emergence of the Phillip Island, Port Phillip region, and second introduction from Gippsland ancestral progenitors were eight (Phillip Island MRCA, 1985; first case, 1993), seven (Port Phillip region MRCA, 1983; first case, 1990), and nine (Gippsland second introduction MRCA, 2003; first case, 2012) years prior to when cases were initially reported in those areas. Given this, it might be that *M. ulcerans* requires an approximately 7- to 9-year period to establish to sufficient levels in a new environment before it can cause disease in humans. Given the slow growth of this bacterium (68), during this establishment period, the mycobacterial population may be growing toward a critical threshold in population size that permits a spillover effect to humans.

The estimated time to the Victorian ancestor in 1766 coincides with the first arrival of Europeans in southeastern Australia, with the first European settlement in Victorian occurring in 1803 (55, 56). This temporal association suggests that the activity of nonindigenous Australians, around the time of Victorian European settlement, might have facilitated the introduction of *M. ulcerans* into southeastern Australia at Gippsland. In this way, maritime transport may have facilitated the movement of *M. ulcerans* along the east coast of Australia to Victoria, possibly south from northern Australia.

RS and RT isolates. Included in our isolate panel were two replicate isolates of two cultures derived from patients originally described in the original medical report by MacCallum et al. in 1948 (1). Interestingly, SNP differences existed between these duplicate isolate genomes from the same patients. These SNP differences are thought to have arisen due to differential sequential lab passaging of the isolates. Most interesting was the substantial genetic variation between the RT isolates and that of all other Victorian isolate genomes, separated by a median of 123 SNPs and the Victorian MRCA.

The existence of a substantially distinct genotype from the Gippsland region suggests that perhaps the westbound lineage had a survival advantage over the more ancestral RT lineage. This hypothesis would explain why only a single ancestral lineage isolate was detected and why *M. ulcerans* rose to become a prolific pathogen so late in its occupation of southeastern Australia. This hypothesis could be tested by assessing phenotypic differences between the lineages that are linked to pathogenesis, such as ability to colonize and produce mycolactone. Distinct lineages with different abilities to cause disease have been previously reported (58).

Demographic history. We identified a striking relationship between the demographic history inferred from the genomic data and the number of Victorian BU cases over time. Given that humans are likely a spillover host in southeastern Australia, the amount of BU cases is likely to be influenced by the abundance of the pathogen. In this way, the observed relationship might be linked through cause and effect, in that the increase in bacterial population size may be causing more BU cases, providing an explanation for the apparent recent rise in cases. The drivers behind the rise in bacterial population density in Victoria remain unknown. Further studies could investigate environmental factors that display a similar pattern of increase that temporally align with the rise in cases and profile of population size growth.

Nucleotide substitution rate. Our uncorrelated relaxed molecular modeling allowed us to estimate a genome-wide rate of nucleotide substitutions per chromosome per year. The Victorian genome-wide rate of nucleotide substitution is 6.18×10^{-2} SNPs per chromosome per year, or 1.18×10^{-8} more than that estimated for the African *M. ulcerans* population (54). Potential explanations for the higher substitution rate in the Victorian population may be ecological differences, possibly the presence of marsupial reservoirs (27). The rate variation observed among branches in the Victorian time tree is likely to be a function of variation in bacterial replication cycles per unit time, changes in selection pressures through time, or a combination of these factors (54).

Limitations. One potential limitation of this study is that only a subset of BU cases yielded cultured isolates, a further subset of which was then forwarded for whole-genome sequencing. Despite this, we had carefully selected isolates based on their epidemiology to maximize both temporal and spatial coverage, reducing the likelihood of basing our analyses on a biased isolate panel.

Conclusion. In this study, we have demonstrated the power of microbial genomics to reveal the evolutionary dynamics underpinning the emergence of *M. ulcerans* in Australia. On a continental scale, Australian *M. ulcerans* has evolved from a PNG-like ancestral progenitor and since established disease foci around Australia, developing region-specific SNP patterns. Our data depict that the westward progression of cases out of Gippsland and into both the Western Port and Port Phillip regions is likely due to pathogen migration. Our observation on the likely causation of new areas of BU endemicity suggests that interventions to prevent pathogen spread will prevent the rise of future endemicity and the associated public health burden. Through our Bayesian coalescence approach, we were able to estimate historical bacterial population sizes, revealing a pattern of population expansion that aligns with a temporally matched escalation in Victorian BU cases. Through this genomic insight, we postulate that the rise in Victorian BU cases in humans is a response to a growing pathogen mycobacterial population in the environment. Furthermore, our use of genomics has revealed a second and more recent westward spread of *M. ulcerans* out of Gippsland and into the Port Phillip Bay region. The development of a novel PCR-based SNP typing assay allowed us to explore the validity for a second pathogen introduction by detecting Gippsland-associated genomic markers directly from environmental specimens. Overall, our findings shine light on the evolutionary history of *M. ulcerans* in Australia and provide testable hypotheses aimed at preventing the rise of future areas of endemicity.

MATERIALS AND METHODS

Study area. Australia is a country and continent delimited by both the Pacific and Indian oceans (Fig. 1A). The majority of Australian BU cases are from the states of QLD and Victoria, located in the north and southeastern corner of Australia, respectively. Geographically speaking, Victoria is bordered by New South Wales to the north, South Australia to the west, the Tasman Sea to the east, and the Bass Strait and Tasmania to the south (Fig. 1A). In this study, the Victorian areas of Gippsland, Phillip Island, Frankston/Langwarrin, Mornington Peninsula, and Bellarine Peninsula were considered areas of endemicity, as the majority of Victorian BU cases have originated from these regions (Fig. 1B).

Mycobacterium ulcerans strains and sequencing. One hundred seventy-eight disease-causing *M. ulcerans* isolates, for which detailed information was available, were chosen for comparative genomic analyses. One hundred sixty-five isolates were from Victoria, while 11 and two isolates were from other Australian states/territories and Papua New Guinea (PNG), respectively. The mycobacterial disease isolates were selected to boost temporal and spatial diversity, were derived from both animal and human BU infections, and are described in Table S1. Isolates were cultivated on Brown and Buckle slopes as described by MacCallum et al. (1). Long-term storage of isolates consisted of preservation in 20% (vol/vol) glycerol in LB broth at -80°C , with the exception of isolates collected prior to 1950, as these were preserved by freeze-drying and storage in glass ampoules. Genomic DNA was prepared as previously described (69) from at least 50 mg (wet weight) of mycobacterial biomass, and paired-end DNA libraries were created using the Nextera XT DNA preparation kit (Illumina, San Diego, CA). Sequencing was undertaken using an Illumina NextSeq 500 (Illumina), Illumina MiSeq, or Illumina GAIIx DNA sequencing platform. The quality of raw reads was assessed using the fq tool of the nullarbor package (<https://github.com/tseemann/nullarbor>). All *M. ulcerans* isolates and their respective sequencing reactions are described in Table S1.

Variant detection and maximum likelihood phylogenetics. Repetitive regions of the Agy99 reference genome (GenBank accession no. CP000325), including IS2404 and IS2606 genomic elements, were hard masked and consequently excluded from the analyses to eliminate the possibility of calling false-positive variants and unreliable mapping (70). The snippy version 3.1 program (<https://github.com/tseemann/snippy>) was used to align reads against the reference and call single nucleotide polymorphisms (SNPs). An alignment of core-genome SNPs was used to generate a maximum likelihood (ML) phylogenetic tree using PhyML version 3.3.20170830 (71) under the general time-reversible model of nucleotide substitution. Bootstrapping was performed with 1,000 replicates to assess the reliability of the phylogeny, collapsing nodes with less than 70% support using TreeCollapseCL version 4 (<http://emmahodcroft.com/TreeCollapseCL.html>). All phylogenies were rooted using *M. marinum* (GenBank accession no. GCA_000018345.1) as an outgroup to determine the direction of evolution.

Bayesian phylogenetic analysis. BEAST version 2.4.4 was employed to determine the substitution rate, date evolutionary events, estimate the demographic history, and generate a time tree of the 165 Victorian *M. ulcerans* isolates. BEAUti xml files were modified manually to specify the number of invariant sites in the genome. An uncorrelated log-normal relaxed molecular clock (57) was used with the Extended Bayesian Skyline Plot (EBSF) demographic method (72) and bModelTest (73) to infer a genome scale Victorian *M. ulcerans* time tree, with a tip date defined as the year of isolation (Table S1). Analysis was performed in BEAST 2 using a total of 5 independent chains of 900 million Markov chain Monte Carlo (MCMC) generations, with samples taken every 90,000 generations. Log files were inspected in Tracer version 1.6 (<http://tree.bio.ed.ac.uk/software/tracer/>) for convergence and proper mixing, as well as to see whether the chain length produced an effective sample size (ESS) for all parameters larger than 300, indicating sufficient sampling. LogCombiner version 2.4.4 (74) was then used to combine log and tree files of the 5 independent runs, with a 30% burn-in for each run. TreeAnnotator version 2.4.4 (74) was used to summarize the posterior sample of time trees in order to produce a maximum clade credibility tree with the posterior estimates of node heights visualized on it.

In order to assess the legitimacy of the temporal signal in the data, a permutation assessment was undertaken by performing 40 additional BEAST 2 runs (of 900 million MCMC generations each), with substitution, clock, and demographic models identical to those detailed before, but with tip dates reshuffled to sequences. The reshuffling was done both randomly (75) and using clusters (76) to check for potential confounding between temporal and genetic structures (this can occur when closely related isolates are more likely to have been sampled during similar times). These two reshuffled “null sets” of tip date and sequence correlations were then compared with the substitution rate estimate achieved using the actual tip dates (75, 76). All temporal estimates reported in this study are parameter medians and 95% highest posterior density (HPD) intervals.

Phylogeographic analysis. The open-source software GenGIS version 2.4.1 (77) was used to complement phylogenies with geography. A geographical axis running approximately east to west along the Victorian coastline was drawn to provide the layout for a geographically oriented phylogeny. To assess whether the fit between ordered leaf nodes and geographic points was significantly better than that expected by chance alone, a Monte Carlo permutation test with 1,000 replicates was performed (78), with a probability value of less than or equal to 0.05 considered significant.

Direct genotyping of environmental specimens. Two SNPs associated with the Gippsland clade (Agy99 chromosome positions 1385112 and 5597161) were used as targets for the development of a PCR-based SNP genotyping assay. Site 1385112 is located in an oxidoreductase gene (MUL_1288), and site 5597161 is within a conserved hypothetical transmembrane protein gene (MUL_5043). Primers specific to genomic regions flanking the Gippsland SNP positions were designed so that the amplicons contained the bases of interest (Table 1). In order to eliminate nonspecific binding, a fully nested PCR

TABLE 1 PCR primers used for direct SNP genotyping of environmental specimens in this study

Primer name	Forward sequence (5'-3')	Reverse sequence (5'-3')	Function	Amplicon size (bp)
1385112_outer	GCCATCATGGGTTGAGTGA	ACGACCTGCAGTTCGAGAAG	Amplify outer region of position 1385112	701
1385112_inner	TTTTCGAACAGCGTGATGGC	CACCCAGGACTGACCAAGAC	Amplify inner region of position 1385112	468
5597161_outer	TGCACCAAGTCCCTCGATGAG	GACCGCAAGAATCAGCGTTC	Amplify outer region of position 5597161	875
5597161_inner	TTCTACCAAAAATCGCGGCT	GTCGGTGCTCTCTCGTTGGT	Amplify inner region of position 5597161	446

approach was undertaken, in which a second reaction was performed using amplicons from the first reaction as the template. PCR primers were designed using the Primer3 software version 2.3.4 (79).

For PCR, 50- μ l reaction mixtures were prepared with 1- μ l samples of purified genomic DNA, 32 μ l of H_2O , 10 μ l of 5 $\times$ Phire buffer, 1 μ l of 10 mM dinucleoside triphosphates (dNTPs), 2.5 μ l of 0.5 μ M primer F, 2.5 μ l of 0.5 μ M primer R, and 1 μ l of Phire Hot Start DNA polymerase (Thermo Fisher). The reactions were performed in an automated thermal cycler (Applied Biosystems Veriti thermal cycler). After an initial denaturation at 98°C for 30 s, DNA was amplified by 35 cycles of 1-min steps at 98°C, 60°C, and 72°C. Amplicons were cleaned in a buffer of 0.5 μ l of Pellet Paint, 12.5 μ l of 100% ethyl alcohol (EtOH), and 0.5 μ l of 3 M NaOAc, incubated at room temperature for 5 min, and spun at top speed in a benchtop centrifuge for 5 min. Pellets were then washed with 500 μ l of 70% EtOH, spun for 5 min, and air-dried for 10 min. Amplicons were Sanger sequenced using the BigDye sequencing kit on an ABI instrument (Applied Biosystems 3130xl genetic analyzer), and the identity of the base of interest was undertaken by aligning the amplicon sequence against a reference in Geneious version 8.1.8. The assay was validated using pure genomic DNA from known Gippsland and non-Gippsland isolates to serve as positive and negative controls, respectively. Nested PCR was then performed on genomic DNA extracted from possum fecal material with a high abundance of *M. ulcerans* genomic units, as inferred from low cycle threshold (C_T) values with the IS2404 qPCR assay (80).

Accession number(s). Short-read data generated through this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number [PRJNA421048](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA421048).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/AEM.02612-17>.

SUPPLEMENTAL FILE 1, PDF file, 0.1 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This project was supported in part by the National Health and Medical Research Council of Australia (NHMRC), an NHMRC Senior Research Fellowship to T.P.S. (grant GNT1105525); and an NHMRC Practitioner Fellowship to B.P.H. (GNT1105905). A.H.B. was supported by an Australian Postgraduate Award Ph.D. scholarship.

A.H.B. and T.P.S. conceived and directed the project. Bioinformatic analyses were conducted by A.H.B., K.V., T.S., and C.J.M. and J.A.M.F. provided bacterial isolates for analysis. E.L.T. provided epidemiological information on the isolates. A.H.B., T.P.S., and P.D.R.J. prepared the manuscript. J.L.P., N.J.T., T.T., and B.P.H. performed and supported DNA extraction and whole-genome sequencing on the isolates. All authors reviewed and approved the manuscript.

REFERENCES

- MacCallum P, Tolhurst JC, Buckle G, Sissons HA. 1948. A new *Mycobacterium* infection in man. *J Pathol Bacteriol* 60:93–122. <https://doi.org/10.1002/path.1700600111>.
- Yip MJ, Porter JL, Fyfe JA, Lavender CJ, Portals F, Rhodes M, Kator H, Colomni A, Jenkin GA, Stinear TP. 2007. Evolution of *Mycobacterium ulcerans* and other mycolactone-producing mycobacteria from a common *Mycobacterium marinum* progenitor. *J Bacteriol* 189:2021–2029. <https://doi.org/10.1128/JB.01442-06>.
- Doig KD, Holt KE, Fyfe JA, Lavender CJ, Eddyani M, Portals F, Yeboah-Manu D, Pluschke G, Seemann T, Stinear TP. 2012. On the origin of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *BMC Genomics* 13:258. <https://doi.org/10.1186/1471-2164-13-258>.
- Stinear TP, Mve-Obiang A, Small PL, Frigui W, Pryor MJ, Brosch R, Jenkin GA, Johnson PD, Davies JK, Lee RE, Adusumilli S, Garnier T, Haydock SF, Leadlay PF, Cole ST. 2004. Giant plasmid-encoded polyketide synthases produce the macrolide toxin of *Mycobacterium ulcerans*. *Proc Natl Acad Sci U S A* 101:1345–1349. <https://doi.org/10.1073/pnas.0305877101>.
- Stinear TP, Hong H, Frigui W, Pryor MJ, Brosch R, Garnier T, Leadlay PF, Cole ST. 2005. Common evolutionary origin for the unstable virulence plasmid pMUM found in geographically diverse strains of *Mycobacterium ulcerans*. *J Bacteriol* 187:1668–1676. <https://doi.org/10.1128/JB.187.5.1668-1676.2005>.
- George KM, Chatterjee D, Gunawardana G, Welty D, Hayman J, Lee R, Small PL. 1999. Mycolactone: a polyketide toxin from *Mycobacterium ulcerans* required for virulence. *Science* 283:854–857. <https://doi.org/10.1126/science.283.5403.854>.
- Stinear TP, Seemann T, Pidot S, Frigui W, Reyset G, Garnier T, Meurice G, Simon D, Bouchier C, Ma L, Tichit M, Porter JL, Ryan J, Johnson PD, Davies JK, Jenkin GA, Small PL, Jones LM, Tekala F, Laval F, Daffe M, Parkhill J, Cole ST. 2007. Reductive evolution and niche adaptation

- inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res* 17:192–200. <https://doi.org/10.1101/gr.5942807>.
8. van der Werf TS, van der Graaf WT, Tappero JW, Asiedu K. 1999. *Mycobacterium ulcerans* infection. *Lancet* 354:1013–1018. [https://doi.org/10.1016/S0140-6736\(99\)01156-3](https://doi.org/10.1016/S0140-6736(99)01156-3).
 9. Portaels F, Aguiar J, Debacker M, Guedenon A, Steunou C, Zinsou C, Meyers WM. 2004. *Mycobacterium bovis* BCG vaccination as prophylaxis against *Mycobacterium ulcerans* osteomyelitis in Buruli ulcer disease. *Infect Immun* 72:62–65. <https://doi.org/10.1128/IAI.72.1.62-65.2004>.
 10. Portaels F, Johnson P, Meyers WM. 2001. Buruli ulcer: diagnosis of *Mycobacterium ulcerans* disease. World Health Organization, Geneva, Switzerland. http://apps.who.int/iris/bitstream/10665/67000/1/WHO_CDS_CPE_GBUI_2001.4.pdf.
 11. Amofah G, Bonsu F, Tetteh C, Okrah J, Asamoah K, Asiedu K, Addy J. 2002. Buruli ulcer in Ghana: results of a national case search. *Emerg Infect Dis* 8:167–170. <https://doi.org/10.3201/eid0802.010119>.
 12. Johnson PD, Stinear T, Small PL, Pluschke G, Merritt RW, Portaels F, Huygen K, Hayman JA, Asiedu K. 2005. Buruli ulcer (*M. ulcerans* infection): new insights, new hope for disease control. *PLoS Med* 2:e108. <https://doi.org/10.1371/journal.pmed.0020108>.
 13. World Health Organization. 2009. Buruli ulcer: first programme review meeting for West Africa—summary report. *Wkly Epidemiol Rec* 84:41–48.
 14. Hayman J. 1985. Clinical features of *Mycobacterium ulcerans* infection. *Australas J Dermatol* 26:67–73. <https://doi.org/10.1111/j.1440-0960.1985.tb01819.x>.
 15. Francis G, Whitby M, Woods M. 2006. *Mycobacterium ulcerans* infection: a rediscovered focus in the Capricorn Coast region of central Queensland. *Med J Aust* 185:179–180.
 16. Steffen CM, Freeborn H. 2016. *Mycobacterium ulcerans* in the Daintree 2009–2015 and the mini-epidemic of 2011. *Aust N Z J Surg* <https://doi.org/10.1111/ans.13817>.
 17. Steffen CM, Smith M, McBride WJ. 2010. *Mycobacterium ulcerans* infection in North Queensland: the 'Daintree ulcer'. *Aust N Z J Surg* 80: 732–736. <https://doi.org/10.1111/j.1445-2197.2010.05338.x>.
 18. Johnson PD, Veitch MG, Leslie DE, Flood PE, Hayman JA. 1996. The emergence of *Mycobacterium ulcerans* infection near Melbourne. *Med J Aust* 164:76–78.
 19. Jackson K, Edwards R, Leslie DE, Hayman J. 1995. Molecular method for typing *Mycobacterium ulcerans*. *J Clin Microbiol* 33:2250–2253.
 20. Veitch MG, Johnson PD, Flood PE, Leslie DE, Street AC, Hayman JA. 1997. A large localized outbreak of *Mycobacterium ulcerans* infection on a temperate southern Australian island. *Epidemiol Infect* 119:313–318. <https://doi.org/10.1017/S0950268897008273>.
 21. Boyd SC, Athan E, Friedman ND, Hughes A, Walton A, Callan P, McDonald A, O'Brien DP. 2012. Epidemiology, clinical features and diagnosis of *Mycobacterium ulcerans* in an Australian population. *Med J Aust* 196: 341–344. <https://doi.org/10.5694/mja12.10087>.
 22. Tiong A. 2005. The epidemiology of a cluster of *Mycobacterium ulcerans* infections in Point Lonsdale. *Vic Infect Dis Bull* 8:2–4.
 23. Johnson PD, Azuolas J, Lavender CJ, Wishart E, Stinear TP, Hayman JA, Brown L, Jenkin GA, Fyfe JAM. 2007. *Mycobacterium ulcerans* in mosquitoes captured during outbreak of Buruli ulcer, southeastern Australia. *Emerg Infect Dis* 13:1653–1660. <https://doi.org/10.3201/eid1311.061369>.
 24. van Ravensway J, Benbow ME, Tsonis AA, Pierce SJ, Campbell LP, Fyfe JA, Hayman JA, Johnson PD, Wallace JR, Qi J. 2012. Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. *PLoS One* 7:e51074. <https://doi.org/10.1371/journal.pone.0051074>.
 25. Hayman J. 1991. Postulated epidemiology of *Mycobacterium ulcerans* infection. *Int J Epidemiol* 20:1093–1098. <https://doi.org/10.1093/ije/20.4.1093>.
 26. Williamson HR, Benbow ME, Nguyen KD, Beachboard DC, Kimbirauskas RK, McIntosh MD, Quayle C, Ampadu EO, Boakye D, Merritt RW, Small PLC. 2008. Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana. *PLoS Negl Trop Dis* 2:e205. <https://doi.org/10.1371/journal.pntd.0000205>.
 27. Fyfe JA, Lavender CJ, Handasyde KA, Legione AR, O'Brien CR, Stinear TP, Pidot SJ, Seemann T, Benbow ME, Wallace JR, McCowan C, Johnson PD. 2010. A major role for mammals in the ecology of *Mycobacterium ulcerans*. *PLoS Negl Trop Dis* 4:e791. <https://doi.org/10.1371/journal.pntd.0000791>.
 28. Barker D, Clancy J, Morrow R, Rao S. 1970. Transmission of Buruli disease. *Br Med J* 4:558.
 29. Marston BJ, Diallo MO, Horsburgh CR, Jr, Diomande I, Saki MZ, Kanga J-M, Patrice G, Lipman HB, Ostroff SM, Good RC. 1995. Emergence of Buruli ulcer disease in the Daloa region of Cote d'Ivoire. *Am J Trop Med Hyg* 52:219–224. <https://doi.org/10.4269/ajtmh.1995.52.219>.
 30. Raghunathan PL, Whitney EA, Asamoah K, Stienstra Y, Taylor TH, Jr, Amofah GK, Ofori-Adjei D, Dobos K, Guarner J, Martin S, Pathak S, Klutse E, Etsuful S, van der Graaf WT, van der Werf TS, King CH, Tappero JW, Ashford DA. 2005. Risk factors for Buruli ulcer disease (*Mycobacterium ulcerans* infection): results from a case-control study in Ghana. *Clin Infect Dis* 40:1445–1453. <https://doi.org/10.1086/429623>.
 31. Aliga H, Amano T, Cairncross S, Adomako J, Nanas OK, Coleman S. 2004. Assessing water-related risk factors for Buruli ulcer: a case-control study in Ghana. *Am J Trop Med Hyg* 71:387–392.
 32. O'Brien DP, Wynne JW, Buultjens AH, Michalski WP, Stinear TP, Friedman ND, Hughes A, Athan E. 2017. Exposure risk for infection and lack of human-to-human transmission of *Mycobacterium ulcerans* disease, Australia. *Emerg Infect Dis* 23:837. <https://doi.org/10.3201/eid2305.160809>.
 33. O'Brien CR, Handasyde KA, Hibble J, Lavender CJ, Legione AR, McCowan C, Globan M, Mitchell AT, McCracken HE, Johnson PD, Fyfe JA. 2014. Clinical, microbiological and pathological findings of *Mycobacterium ulcerans* infection in three Australian possum species. *PLoS Negl Trop Dis* 8:e2666. <https://doi.org/10.1371/journal.pntd.0002666>.
 34. Quek TY, Athan E, Henry MJ, Pasco JA, Redden-Hoare J, Hughes A, Johnson PDR. 2007. Risk factors for *Mycobacterium ulcerans* infection, southeastern Australia. *Emerg Infect Dis* 13:1661–1666. <https://doi.org/10.3201/eid1311.061206>.
 35. Johnson PDR, Lavender CJ. 2009. Correlation between Buruli ulcer and vector-borne notifiable diseases, Victoria, Australia. *Emerg Infect Dis* 15:614–615. <https://doi.org/10.3201/eid1504.081162>.
 36. Portaels F, Fonteyne PA, de Beenhouwer H, de Rijk P, Guedenon A, Hayman J, Meyers MW. 1996. Variability in 3' end of 16S rRNA sequence of *Mycobacterium ulcerans* is related to geographic origin of isolates. *J Clin Microbiol* 34:962–965.
 37. Chemlal K, De Ridder K, Fonteyne PA, Meyers WM, Swings J, Portaels F. 2001. The use of IS2404 restriction fragment length polymorphisms suggests the diversity of *Mycobacterium ulcerans* from different geographical areas. *Am J Trop Med Hyg* 64:270–273. <https://doi.org/10.4269/ajtmh.2001.64.270>.
 38. Chemlal K, Huys G, Fonteyne PA, Vincent V, Lopez AG, Rigouts L, Swings J, Meyers WM, Portaels F. 2001. Evaluation of PCR-restriction profile analysis and IS2404 restriction fragment length polymorphism and amplified fragment length polymorphism fingerprinting for identification and typing of *Mycobacterium ulcerans* and *M. marinum*. *J Clin Microbiol* 39:3272–3278. <https://doi.org/10.1128/JCM.39.9.3272-3278.2001>.
 39. Abiordey A, Swings J, Hubans C, Chemlal K, Loch C, Portaels F, Supply P. 2005. Multilocus variable-number tandem repeat typing of *Mycobacterium ulcerans*. *J Clin Microbiol* 43:1546–1551. <https://doi.org/10.1128/JCM.43.4.1546-1551.2005>.
 40. Stragier P, Abiordey A, Meyers WM, Portaels F. 2005. Genotyping *Mycobacterium ulcerans* and *Mycobacterium marinum* by using mycobacterial interspersed repetitive units. *J Bacteriol* 187:1639–1647. <https://doi.org/10.1128/JB.187.5.1639-1647.2005>.
 41. Hilty M, Yeboah-Manu D, Boakye D, Mensah-Quainoo E, Rondini S, Schelling E, Ofori-Adjei D, Portaels F, Zinsstag J, Pluschke G. 2006. Genetic diversity in *Mycobacterium ulcerans* isolates from Ghana revealed by a newly identified locus containing a variable number of tandem repeats. *J Bacteriol* 188:1462–1465. <https://doi.org/10.1128/JB.188.4.1462-1465.2006>.
 42. Stinear TP, Jenkin GA, Johnson PD, Davies JK. 2000. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. *J Bacteriol* 182:6322–6330. <https://doi.org/10.1128/JB.182.22.6322-6330.2000>.
 43. Huys G, Rigouts L, Chemlal K, Portaels F, Swings J. 2000. Evaluation of amplified fragment length polymorphism analysis for inter- and intraspecific differentiation of *Mycobacterium bovis*, *M. tuberculosis*, and *M. ulcerans*. *J Clin Microbiol* 38:3675–3680.
 44. Abiordey A, Kotlowski R, Swings J, Portaels F. 2005. PCR amplification with primers based on IS2404 and GC-rich repeated sequence reveals polymorphism in *Mycobacterium ulcerans*. *J Clin Microbiol* 43:448–451. <https://doi.org/10.1128/JCM.43.1.448-451.2005>.
 45. Stinear T, Davies JK, Jenkin GA, Portaels F, Ross BC, Oppedisano F, Purcell M, Hayman JA, Johnson PD. 2000. A simple PCR method for rapid genotype analysis of *Mycobacterium ulcerans*. *J Clin Microbiol* 38: 1482–1487.
 46. Käser M, Gutmann O, Hauser J, Stinear T, Cole S, Yeboah-Manu D,

- Derrick G, Certa U, Pluschke G. 2009. Lack of insertional-deletional polymorphism in a collection of *Mycobacterium ulcerans* isolates from Ghanaian Buruli ulcer patients. *J Clin Microbiol* 47:3640–3646. <https://doi.org/10.1128/JCM.00760-09>.
47. Röttgen K, Qi W, Ruf M-T, Mensah-Quainoo E, Pidot SJ, Seemann T, Stinear TP, Käser M, Yeboah-Manu D, Pluschke G. 2010. Single nucleotide polymorphism typing of *Mycobacterium ulcerans* reveals focal transmission of Buruli ulcer in a highly endemic region of Ghana. *PLoS Negl Trop Dis* 4:e751. <https://doi.org/10.1371/journal.pntd.0000751>.
 48. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML. 2004. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci U S A* 101:13536–13541. <https://doi.org/10.1073/pnas.0403844101>.
 49. Bryant J, Chewapreecha C, Bentley SD. 2012. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol* 7:1283–1296. <https://doi.org/10.2217/fmb.12.108>.
 50. Flicek P, Birney E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 6:56–512. <https://doi.org/10.1038/nmeth.1376>.
 51. Abiordey AS, Vandelannoote K, Krimpong IA, Ahortor EK, Amisshah NA, Eddyani M, Durnez L, Portals F, de Jong BC, Leirs H, Porter JL, Mangas KM, Lam MMC, Buultjens A, Seemann T, Tobias NJ, Stinear TP. 2015. Whole genome comparisons suggest random distribution of *Mycobacterium ulcerans* genotypes in a Buruli ulcer endemic region of Ghana. *PLoS Negl Trop Dis* 9:e0003681. <https://doi.org/10.1371/journal.pntd.0003681>.
 52. Bolz M, Bratschi MW, Kerber S, Minyem JC, Boock AU, Vogel M, Bayi PF, Junghans T, Brites D, Harris SR, Parkhill J, Pluschke G, Lamelas Cabello A. 2015. Locally confined clonal complexes of *Mycobacterium ulcerans* in two Buruli ulcer endemic regions of Cameroon. *PLoS Negl Trop Dis* 9:e0003802. <https://doi.org/10.1371/journal.pntd.0003802>.
 53. Eddyani M, Vandelannoote K, Meehan CJ, Bhujji S, Porter JL, Aguiar J, Seemann T, Jarek M, Singh M, Portals F, Stinear TP, de Jong BC. 2015. A genomic approach to resolving relapse versus reinfection among four cases of Buruli ulcer. *PLoS Negl Trop Dis* 9:e0004158. <https://doi.org/10.1371/journal.pntd.0004158>.
 54. Vandelannoote K, Meehan CJ, Eddyani M, Affolabi D, Phanuz DM, Eyangoh S, Jordans K, Portals F, Mangas K, Seemann T, Marsollier L, Marion E, Chauly A, Landier J, Fontanet A, Leirs H, Stinear TP, de Jong BC. 2017. Multiple introductions and recent spread of the emerging human pathogen *Mycobacterium ulcerans* across Africa. *Genome Biol Evol* 9:414–426. <https://doi.org/10.1093/gbe/evx003>.
 55. Moorhead L. 2011. *Morington in the wake of flinders: historical survey*, 3rd ed, vol 1. Synergy Press, Preston, Victoria, Australia.
 56. Shaw AGL. 1996. *A history of the Port Phillip District: Victoria before separation*, vol 1. Megunyah Press, Victoria, Australia.
 57. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88. <https://doi.org/10.1371/journal.pbio.0040088>.
 58. Käser M, Rondini S, Naegeli M, Stinear T, Portals F, Certa U, Pluschke G. 2007. Evolution of two distinct phylogenetic lineages of the emerging human pathogen *Mycobacterium ulcerans*. *BMC Evol Biol* 7:177. <https://doi.org/10.1186/1471-2148-7-177>.
 59. Vandelannoote K, Durnez L, Amisshah D, Gryseels S, Dodo A, Yeboah S, Addo P, Eddyani M, Leirs H, Abiordey A, Portals F. 2010. Application of real-time PCR in Ghana, a Buruli ulcer-endemic country, confirms the presence of *Mycobacterium ulcerans* in the environment. *FEMS Microbiol Lett* 304:191–194. <https://doi.org/10.1111/j.1574-6968.2010.01902.x>.
 60. Tian RBD, Niamké S, Tissot-Dupont H, Drancourt M. 2016. Detection of *Mycobacterium ulcerans* DNA in the environment, Ivory Coast. *PLoS One* 11:e0151567. <https://doi.org/10.1371/journal.pone.0151567>.
 61. Trubiano JA, Lavender CJ, Fyfe JA, Bittmann S, Johnson PDR. 2013. The incubation period of Buruli ulcer (*Mycobacterium ulcerans* infection). *PLoS Negl Trop Dis* 7:e2463. <https://doi.org/10.1371/journal.pntd.0002463>.
 62. Portals F, Meyers WM, Abiordey A, Castro AG, Chemlal K, de Rijk P, Elsen P, Fissette K, Fraga AG, Lee R, Mahrous E, Small PL, Stragier P, Torrado E, Van Aerde A, Silva MT, Pedrosa J. 2008. First cultivation and characterization of *Mycobacterium ulcerans* from the environment. *PLoS Negl Trop Dis* 2:e178. <https://doi.org/10.1371/journal.pntd.0000178>.
 63. Thomson JA, Owen WH. 1964. A field study of the Australian ringtail possum *Pseudocheirus peregrinus* (Marsupialia: Phalangeridae). *Ecol Monogr* 34:27–52. <https://doi.org/10.2307/1948462>.
 64. How RA. 1981. Population parameters of two congeneric possums, *Trichosurus* spp., in north-eastern New South Wales. *Aust J Zool* 29: 205–215. <https://doi.org/10.1071/ZO9810205>.
 65. Tidemann CR, Nelson JE. 2004. Long-distance movements of the grey-headed flying fox (*Pteropus poliocephalus*). *J Zool* 263:141–146. <https://doi.org/10.1017/S0952836904004960>.
 66. Roberts BJ, Catterall CP, Eby P, Kanowski J. 2012. Long-distance and frequent movements of the flying-fox *Pteropus poliocephalus*: implications for management. *PLoS One* 7:e42532. <https://doi.org/10.1371/journal.pone.0042532>.
 67. Buckle G. 1969. Notes on *Mycobacterium ulcerans*. *Aust N Z J Surg* 38:320–323.
 68. Portals F, Chemlal K, Elsen P, Johnson PD, Hayman JA, Hibble J, Kirkwood R, Meyers WM. 2001. *Mycobacterium ulcerans* in wild animals. *Rev Sci Tech* 20:252–264. <https://doi.org/10.20506/rst.20.1.1270>.
 69. Pidot SJ, Porter JL, Tobias NJ, Anderson J, Catmull D, Seemann T, Kidd S, Davies JK, Reynolds E, Dashper S, Stinear TP. 2010. Regulation of the 18 kDa heat shock protein in *Mycobacterium ulcerans*: an alpha-crystallin orthologue that promotes biofilm formation. *Mol Microbiol* 78: 1216–1231. <https://doi.org/10.1111/j.1365-2958.2010.07401.x>.
 70. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* 40:987–993. <https://doi.org/10.1038/ng.195>.
 71. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696–704. <https://doi.org/10.1080/10635150390235520>.
 72. Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evol Biol* 8:289. <https://doi.org/10.1186/1471-2148-8-289>.
 73. Bouckaert RR, Drummond AJ. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol Biol* 17:42. <https://doi.org/10.1186/s12862-017-0890-6>.
 74. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>.
 75. Duchêne S, Duchêne D, Holmes EC, Ho SY. 2015. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol* 32:1895–1906. <https://doi.org/10.1093/molbev/msv056>.
 76. Murray GG, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, Holmes MA, Rambaut A, Welch JJ. 2016. The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol* 7:80–89. <https://doi.org/10.1111/2041-210X.12466>.
 77. Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG. 2009. GenGIS: a geospatial information system for genomic data. *Genome Res* 19:1896–1904. <https://doi.org/10.1101/gr.095612.109>.
 78. Parks DH, Beiko RG. 2009. Quantitative visualizations of hierarchically organized data in a geographic context, p 1–6. The 17th Int Conf Geoinformatics, 12 to 14 August 2009, Fairfax, VA. <https://doi.org/10.1109/GEOINFORMATICS.2009.5293552>.
 79. Rozen S, Skaletsky H. 1999. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 365–386.
 80. Fyfe JA, Lavender CJ, Johnson PD, Globan M, Sievers A, Aзуolas J, Stinear TP. 2007. Development and application of two multiplex real-time PCR assays for the detection of *Mycobacterium ulcerans* in clinical and environmental samples. *Appl Environ Microbiol* 73:4733–4740. <https://doi.org/10.1128/AEM.02971-06>.
 81. Distributed Active Archive Center for Biogeochemical Dynamics (DAAC). 2000. Shuttle radar topography mission (SRTM) near-global digital elevation models (DEMs). Oak Ridge National Laboratory, Oak Ridge, TN. https://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1.
 82. Australian Bureau of Statistics. 2016. 1270.0.55.001—Australian statistical geography standard (ASGS): volume 1—main structure and greater capital city statistical areas, July 2016. Australian Bureau of Statistics, Belconnen, Australian Capital Territory, Australia. <http://www.abs.gov.au/AUSSTATS/abs@nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument>.

Summary

From this work the key findings were that:

- That there has likely been a single introduction event of *M. ulcerans* into each of the major Australian endemic areas.
- That Australian *M. ulcerans* may have evolved from a Papua New Guinea ancestral progenitor.
- That the estimated time to the Victorian MRCA in 1766 coincides with the first arrival of Europeans in south-eastern Australia.
- That *M. ulcerans* spread west in Victoria, rather than the awakening of quiescent populations.
- That there has been a second introduction of *M. ulcerans* into the Rye region from Gippsland.
- That the increase in bacterial population size may be causing more disease in humans, providing an explanation for the apparent recent rise in cases.

Chapter Five:

Discussion and conclusions

Discussion

In this new and exciting era of big data in biology, due investigation must be carried out to evaluate the utility of genomics to assist investigations into confounding public health situations. This project arose to meet this need, as it demonstrates through case-studies of three pathogens how microbial genomics can provide useful insights for bacterial pathogens of public health importance. The three case studies consider several key avenues relevant to infectious diseases including understanding pathogen evolutionary origins, the development of tools for prospective management of outbreaks and historical population dynamics.

Key Findings

The key findings in this thesis are that:

- Genomic-based studies aimed at deconstructing and better understanding novel pathogens of public health significance provide greater insights than those conducted with conventional microbiological methods.
- Whole genome SNP based analyses provide greater strain resolution than other contemporary approaches based on the MLST framework.
- Genomic-based machine learning approaches developed using prospective environmental sampling have utility in outbreak investigations.
- Genomic insights can provide plausible explanations for observed epidemiological trends and deliver greater understanding of disease transmission.

Strengths and limitations

Genomic analysis is a complex task that can be undertaken through a variety of approaches, each of which have specific strengths and weaknesses. A strength of the work presented in

Chapter 2 and Chapter 3 was the use of real-time single molecule sequencing technology. Here genomes were fully assembled for *E. faecium* and *L. pneumophila*. These complete genome reconstructions served as critical resources for high-resolution population-based analyses. Without such high quality, finished genome sequences, alternative and less related reference genomes would have been used. Recent research has shown that the accuracy of variant calling by techniques such as read-mapping is contingent on the use of a reference genome that is closely related to the bacterial population under investigation (230).

Analyses presented in Chapter 2 provide high-resolution insights into the genomic differences of the recently emerged *E. faecium* ST796 to that of other *E. faecium* clones. Despite this, without having a solid understanding for how such genomic changes might impact pathogen phenotype, such insights are of limited practical value. In this way, it is not until enough knowledge of functional consequences are attained that the true promise of genomics can be afforded.

Since publishing the work in Chapter 2, *E. faecium* ST796 has been detected in Swiss hospitals (231). Investigators made use of the finished reference genome, established in Chapter 2, to conduct their comparative genome analyses (231). Analyses based on this reference exposed a strong genomic association among isolates from the outbreak in Switzerland and those from Australia (231). This highlights the value of Australia's use of WGS for investigation purposes, as such work allowed for high-resolution compatibility and comparability with outbreak investigations in a European institution.

The supervised machine learning approach presented in Chapter 3 demonstrates a powerful, alternative method to phylogeny interpretations to accurately predict the source attribution of *L. pneumophila* outbreaks. However, one issue with this approach is its lack of negative

predictive capacity. This issue is largely due to the simplicity of the model, in that it behaves as if the world is only composed of the cooling tower groups used for training. More specifically, this problem is known as closed set classification, in that there is no allowance for an 'other' or 'unknown' possible assignments. This means that if this method were used to classify a *L. pneumophila* isolate originating outside the locations used for model training it would incorrectly assign it to one of the training locations. At present, the approach does not provide any indications for such a misattribution. Efforts to potentially overcome this issue include the use of a collection of single group classifiers that assign isolates to individual groups based on probability values.

The population genomics analyses of *M. ulcerans* reported in Chapter 4 makes the assumption that the pathogen population infecting humans is derived from and is representative of the greater population, including those residing within environmental reservoirs. While this is likely a safe assumption, it has not been demonstrated, as no environmentally derived genomes were sequenced, owing to the technical challenges of culturing *M. ulcerans* from environmental specimens.

An additional strength of this study was the availability of relevant bacterial isolates for sequencing. The partnerships at the Doherty Institute between the Department of Microbiology, the Microbiological Diagnostic Unit (MDU) and the Victorian Infectious Disease Reference Laboratory (VIDRL) allowed for the availability of many contemporary and historical isolates that were used in this research. In particular, VIDRL provided access to early *M. ulcerans* clinical isolates that were originally prepared by Alfred Hospital Pathology Department in the 1940s. Having access to such temporally diverse isolates permitted a greater ability to perform the temporal dating analyses that were reported in Chapter 4.

The genomic analyses undertaken as part of this investigation made use of both read mapping and *de novo* assembly of short read data to detect genomic variants. While these are pragmatic and widely accepted methods for comparative genomic analysis, alternative raw-read based approaches have the potential to expand the detection of genomic variants. One such approach is to count short DNA oligomers with length k ; entities known as k -mers. k -mers counted from reads sets allow for comparisons among isolates based upon either the number of counts or simply just the presence and absence of the k -mers for an isolate. The advantage of using k -mer-based methods in genomic variant discovery is that they do not necessitate read mapping and are therefore not susceptible to misassembly issues, as they are applied to original sequence data. In this way, k -mers have the potential to simultaneously detect variations caused by SNPs, indels, copy number and larger DNA structural changes. In addition to the detection of variants, k -mer counts depict the abundances of variants, allowing for comparisons of k -mer frequency between isolates. Future work along the lines presented in this investigation could make use of k -mer variation to delve deeper into the associations between bacterial genomes and phenotypes, or to find additional genome variations that can be used to implicate pathogen sources.

It is important to acknowledge that all analyses conducted as part of this project were undertaken in a research rich environment that was well equipped with sequencing and bioinformatic facilities. This infrastructure was critical to carry out such investigations. Such resources and analyses might be limited in routine diagnostic laboratories, and so redressing this deficiency will be critical to ensure genomics becomes more widely accessible.

Importance and implications for public health microbiology

The research in this thesis demonstrates the utility of microbial genomics to generate important insights into pathogens of public health importance. The value of such insights is likely to translate into public health benefits, including reduced infectious disease burden and public health expenditure.

Work from Chapter 2 has provided insights into the likely origins of a novel and highly hospital-adapted lineage of *E. faecium*. Such findings highlight the ability of novel hospital-adapted lineages to emerge through genomic changes, presumably in response to varying conditions in the health care environment. Work undertaken in Chapter 3 exposed weaknesses with traditional phylogenetic approaches that might lead to the misidentification of *L. pneumophila* outbreak sources. Given these limitations, a statistical learning alternative was developed that removes the uncertainty of deducing disease transmission from phylogenies. The findings in Chapter 4 showed that there has been a westward migration of *M. ulcerans* that was followed by a substantial increase in pathogen population size. These analyses demonstrate that *M. ulcerans* is introduced to a new environment and then expands, rather than the awakening of a quiescent pathogen reservoir. Such findings inform our understanding of Buruli ulcer transmission and provide a different perspective on how to control this disease. The importance of this finding is that the emergence of an endemic is predicated on the spread of the bacterium and suggests that efforts to limit the environmental movement of this pathogen could prevent the rise of new disease endemic areas.

Future directions

The genomic changes observed to have accompanied, or perhaps driven, the emergence of the *E. faecium* ST796 lineage in Chapter 2 provides the groundwork for further studies to assess for functional consequences on pathogen phenotypes. Here, phenotypes and genotypes could be linked through gene knockout and targeted mutation-based experiments. An example of this approach to find mutations linked to a particular phenotype is provided with the research arising during this thesis on alcohol tolerance in *E. faecium* (232).

Considering the limitations of the machine learning approach developed in Chapter 3, further work could assess the use of alternative modeling methods. One alternative could be to train a single group classifier to firstly evaluate if a particular isolate will be suitable for classification with the model, as assessed through membership probability. Further alternatives may be to incorporate epidemiological information such as patient residential areas as additional model covariates.

Given that the population genomic analyses in Chapter 4 only considered the genomes of clinical isolates, innovative ways to incorporate environmentally derived genomes should be investigated. One such means may be through the development of a novel isolation technique that can successfully separate *M. ulcerans* from other competing environmental microorganisms. With such a method, environmentally derived *M. ulcerans* isolates could be incorporated in the WGS workflow just as that of their clinical isolate counterparts. An alternative approach might be to employ deep sequencing or metagenomics to directly sequence *M. ulcerans* genomes from environmental specimens.

Conclusions

The advent of affordable microbial genomics is a disruptive force that is changing biological research, with public health microbiology being no exception. Public health problems pose unique challenges, in that they draw from numerous scientific disciplines to advance crucial understandings that can translate into practical interventions. As microbial genomics is largely uncharted territory for public health microbiology, significant development and validation work needs to be undertaken. Using the information rich datasets provided by microbial genomes affords deep insights into evolutionary drivers, disambiguation of clinical source attribution and tracking of pathogen movement patterns. Together with well-established epidemiological approaches, microbial genomics can augment these traditional methods to afford deeper understandings for the dynamics behind clinically relevant pathogens. This thesis showcases the application of microbial genomics to public health problems with three distinct bacterial pathogens. In this work and that of others, microbial genomics can support and extend what was formerly achievable with traditional approaches - findings that substantiate the vision of genomics as a cornerstone of public health microbiology practice.

References

1. **Consortium IHGS.** 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860.
2. **Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD.** 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**:469-474.
3. **Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M.** 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**:462.
4. **Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W.** 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104: H4 outbreak by rapid next generation sequencing technology. *PLoS One* **6**:e22751.
5. **Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P.** 2011. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med.* **364**:730-739.
6. **Bosch T, Verkade E, van Luit M, Pot B, Vauterin P, Burggrave R, Savelkoul P, Kluytmans J, Schouls L.** 2013. High resolution typing by whole genome mapping enables discrimination of LA-MRSA (CC398) strains and identification of transmission events. *PLoS One* **8**:e66493.
7. **Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ.** 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* **10**:599.
8. **Ståhl PL, Lundeberg J.** 2012. Toward the single-hour high-quality genome. *Annu Rev Biochem* **81**:359-378.
9. **Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, Pallen MJ.** 2010. High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J Hosp Infect* **75**:37-41.

10. **Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR.** 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine* **366**:2267-2275.
11. **Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW.** 2012. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* **13**:601.
12. **Bertelli C, Greub G.** 2013. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clinical Microbiology and Infection* **19**:803-813.
13. Grant K, Jenkins C, Cath A, Green J, Zambon M. 2018. Implementing pathogen genomics: a case study. *Public Health England*.
14. **Didelot X, Méric G, Falush D, Darling AE.** 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**:256.
15. **Kwong JC, McCallum N, Sintchenko V, Howden BP.** 2015. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47**:199-210.
16. **Chua KY, Seemann T, Harrison PF, Monagle S, Korman TM, Johnson PD, Coombs GW, Howden BO, Davies JK, Howden BP.** 2011. The dominant Australian community-acquired methicillin-resistant *Staphylococcus aureus* clone ST93-IV [2B] is highly virulent and genetically distinct. *PLoS One* **6**:e25887.
17. **Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'connor L, Ip CL, Golubchik T, Batty EM, Finney JM.** 2013. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *New England Journal of Medicine* **369**:1195-1205.
18. **Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G.** 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*:1259657.
19. **Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S.** 2014. Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *The Lancet infectious diseases* **14**:220-226.
20. **Howden BP, Holt KE, Lam MM, Seemann T, Ballard S, Coombs GW, Tong SY, Grayson ML, Johnson PD, Stinear TP.** 2013. Genomic Insights to Control the Emergence of Vancomycin-Resistant Enterococci. *mBio* **4**.

21. **Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, Bowden R, Walker AS, Peto TE, Paul J.** 2013. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clinical Infectious Diseases* **58**:609-618.
22. **Golubchik T, Brueggemann AB, Street T, Gertz Jr RE, Spencer CC, Ho T, Giannoulatou E, Link-Gelles R, Harding RM, Beall B.** 2012. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nature genetics* **44**:352.
23. **Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L.** 2012. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ open* **2**:e001124.
24. **Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, Ellington MJ, Quail MA, Bentley SD, Parkhill J.** 2013. Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *The Lancet infectious diseases* **13**:130-136.
25. **Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T.** 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome biology* **16**:114.
26. **Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME.** 2013. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ open* **3**:e002175.
27. **Sanger F, Coulson AR.** 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**:441-448.
28. **Sanger F, Nicklen S, Coulson AR.** 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**:5463-5467.
29. **Green ED.** 2001. Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics* **2**:573.
30. **Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*:496-512.
31. **Nyrén P, Lundin A.** 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem* **151**:504-509.

32. **Balasubramanian S.** 2011. Sequencing nucleic acids: from chemistry to medicine. *Chemical Communications* **47**:7281-7286.
33. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**:341.
34. **Loman NJ, Pallen MJ.** 2015. Twenty years of bacterial genome sequencing. *Nature reviews. Microbiology* **13**:787.
35. **Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D.** 2011. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *New England Journal of Medicine* **365**:709-717.
36. **Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P.** 2011. The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* **364**:33-42.
37. **Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C.** 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**:e37-e37.
38. **Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A.** 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* **12**:R18.
39. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B.** 2008. Real-time DNA sequencing from single polymerase molecules. *Science*.
40. **Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW.** 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**:682-686.
41. **Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED.** 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* **30**:693.
42. **Jain M, Olsen HE, Paten B, Akeson M.** 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology* **17**:239.

43. **Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Diltthey AT, Fiddes IT.** 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**:338.
44. **Ouzounis CA, Valencia A.** 2003. Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics* **19**:2176-2190.
45. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C.** 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647-1649.
46. **Leipzig J.** 2017. A review of bioinformatic pipeline frameworks. *Briefings in bioinformatics* **18**:530-536.
47. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. *J Mol Biol* **215**:403-410.
48. **Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW.** 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature genetics* **44**:1056-1059.
49. **Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS.** 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**:430-434.
50. **Liao Y-C, Lin S-H, Lin H-H.** 2015. Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep* **5**:8747.
51. **Pop M.** 2009. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics* **10**:354-366.
52. **Miller JR, Koren S, Sutton G.** 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**:315-327.
53. **Nagarajan N, Pop M.** 2013. Sequence assembly demystified. *Nature Reviews Genetics* **14**:157.
54. **Phillippy AM, Schatz MC, Pop M.** 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology* **9**:R55.
55. **Antipov D, Korobeynikov A, McLean JS, Pevzner PA.** 2015. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**:1009-1015.
56. **Deshpande V, Fung ED, Pham S, Bafna V.** 2013, p 349-363. International workshop on algorithms in bioinformatics.

57. **Wick RR, Judd LM, Gorrie CL, Holt KE.** 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**:e1005595.
58. **Georgiades K, Raoult D.** 2011. Defining pathogenic bacterial species in the genomic era. *Frontiers in Microbiology* **1**:151.
59. **Corander J, Marttinen P, Sirén J, Tang J.** 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**:539.
60. Hillis DM, Allard MW, Miyamoto MM. 1993. Analysis of DNA sequence data: Phylogenetic inference, p. 456-487, *Methods Enzymol*, vol. 224. Elsevier.
61. **Swofford D, Olsen G, Waddell P, Hillis D.** 1996. Phylogenetic inference. *Phylogenetic inference*:407-514.
62. **de Queiroz A, Gatesy J.** 2007. The supermatrix approach to systematics. *Trends Ecol Evol* **22**:34-41.
63. **Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P.** 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**:1283-1287.
64. **Daubin V, Gouy M, Perriere G.** 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome research* **12**:1080-1090.
65. **Teichmann SA, Mitchison G.** 1999. Is there a phylogenetic signal in prokaryote proteins? *Journal of molecular evolution* **49**:98-107.
66. **Farris JS.** 1970. Methods for computing Wagner trees. *Systematic biology* **19**:83-92.
67. **Fitch WM.** 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic biology* **20**:406-416.
68. **Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F.** 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC evolutionary biology* **5**:50.
69. **Fitch WM, Margoliash E.** 1967. Construction of phylogenetic trees. *Science* **155**:279-284.
70. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**:406-425.
71. **Studier JA, Keppler KJ.** 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular biology and evolution* **5**:729-731.
72. **Steel M, Penny D.** 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular biology and evolution* **17**:839-850.
73. **Felsenstein J.** 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**:368-376.

74. **Holder M, Lewis PO.** 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* **4**:275.
75. **Whelan S, Liò P, Goldman N.** 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS in Genetics* **17**:262-272.
76. **Arenas M.** 2015. Trends in substitution models of molecular evolution. *Front Genet* **6**:319.
77. **Kimura M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **16**:111-120.
78. **Tamura K.** 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. *Molecular biology and evolution* **9**:678-687.
79. **Efron B, Halloran E, Holmes S.** 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences* **93**:13429-13429.
80. **Felsenstein J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.
81. **Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ.** 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10**:e1003537.
82. **Drummond AJ, Rambaut A.** 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**:214.
83. **Boskova V, Bonhoeffer S, Stadler T.** 2014. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLoS Comput Biol* **10**:e1003913.
84. **De Maio N, Wu C-H, Wilson DJ.** 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput Biol* **12**:e1005130.
85. **Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A.** 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**:vey016.
86. **Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G.** 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics* **45**:1176.

87. **Jombart T.** 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**:1403-1405.
88. **Jombart T, Pontier D, Dufour A.** 2009. Genetic markers in the playground of multivariate analysis. *Heredity* **102**:330.
89. **Jombart T, Devillard S, Dufour A, Pontier D.** 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**:92.
90. **McVean G.** 2009. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**:e1000686.
91. **Novembre J, Stephens M.** 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40**:646.
92. **Jombart T, Devillard S, Balloux F.** 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics* **11**:94.
93. **Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Välimäki N.** 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nature communications* **6**:6740.
94. **Harrison EM, Weinert LA, Holden MT, Welch JJ, Wilson K, Morgan FJ, Harris SR, Loeffler A, Boag AK, Peacock SJ.** 2014. A shared population of epidemic methicillin-resistant *Staphylococcus aureus* 15 circulates in humans and companion animals. *mBio* **5**:e00985-00913.
95. **Montano V, Didelot X, Foll M, Linz B, Reinhardt R, Suerbaum S, Moodley Y, Jensen JD.** 2015. Worldwide population structure, long term demography, and local adaptation of *Helicobacter pylori*. *Genetics:genetics*. 115.176404.
96. **Cossart YE.** 2014. The rise and fall of infectious diseases: Australian perspectives, 1914-2014. *The Medical journal of Australia* **201**:11-14.
97. **Inweregbu K, Dave J, Pittard A.** 2005. Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care & Pain* **5**:14-17.
98. **Richards MJ, Edwards JR, Culver DH, Gaynes RP, System NNIS.** 2000. Nosocomial infections in combined medical-surgical intensive care units in the United States. *Infection Control & Hospital Epidemiology* **21**:510-515.
99. **Tendolkar P, Baghdayan A, Shankar N.** 2003. Pathogenic enterococci: new developments in the 21st century. *Cellular and Molecular Life Sciences CMLS* **60**:2622-2636.

100. **Huycke MM, Sahm DF, Gilmore MS.** 1998. Multiple-drug resistant enterococci: the nature of the problem and an agenda for the future. *Emerging Infectious Diseases* **4**:239.
101. **Moreno MF, Sarantinopoulos P, Tsakalidou E, De Vuyst L.** 2006. The role and application of enterococci in food and health. *Int J Food Microbiol* **106**:1-24.
102. **Mundt JO.** 1963. Occurrence of enterococci in animals in a wild environment. *Applied microbiology* **11**:136-140.
103. **Mundt JO.** 1963. Occurrence of enterococci on plants in a wild environment. *Applied microbiology* **11**:141-144.
104. **Mundt JO.** 1961. Occurrence of enterococci: bud, blossom, and soil studies. *Applied microbiology* **9**:541-544.
105. **Fisher K, Phillips C.** 2009. The ecology, epidemiology and virulence of *Enterococcus*. *Microbiology* **155**:1749-1757.
106. **Klein G.** 2003. Taxonomy, ecology and antibiotic resistance of enterococci from food and the gastro-intestinal tract. *Int J Food Microbiol* **88**:123-131.
107. **Coombs GW, Daley DA, Thin Lee Y, Pang S, Pearson JC, Robinson JO, Johnson PD, Kotsanas D, Bell JM, Turnidge JD.** Australian Group on Antimicrobial Resistance (AGAR) Australian Enterococcal Sepsis Outcome Programme (AESOP) Annual Report 2014.
108. **Weinstein RA, Hayden MK.** 2000. Insights into the epidemiology and control of infection with vancomycin-resistant enterococci. *Clinical Infectious Diseases* **31**:1058-1065.
109. **Handwerger S, Raucher B, Altarac D, Monka J, Marchione S, Singh KV, Murray BE, Wolff J, Walters B.** 1993. Nosocomial outbreak due to *Enterococcus faecium* highly resistant to vancomycin, penicillin, and gentamicin. *Clinical Infectious Diseases* **16**:750-755.
110. **Poh C, Oh H, Tan A.** 2006. Epidemiology and clinical outcome of enterococcal bacteraemia in an acute care hospital. *Journal of Infection* **52**:383-386.
111. **Hidron AI, Edwards JR, Patel J, Horan TC, Sievert DM, Pollock DA, Fridkin SK.** 2008. Antimicrobial-resistant pathogens associated with healthcare-associated infections: annual summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. *Infection Control & Hospital Epidemiology* **29**:996-1011.

112. **de Been M, van Schaik W, Cheng L, Corander J, Willems RJ.** 2013. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium*. *Genome biology and evolution* **5**:1524-1535.
113. **Gao W, Howden BP, Stinear TP.** 2018. Evolution of virulence in *Enterococcus faecium*, a hospital-adapted opportunistic pathogen. *Current opinion in microbiology* **41**:76-82.
114. **Courvalin P.** 2006. Vancomycin resistance in gram-positive cocci. *Clinical Infectious Diseases* **42**:S25-S34.
115. **Cetinkaya Y, Falk P, Mayhall C.** 2000. Vancomycin-resistant enterococci. *Clin Microbiol Rev* **13**:686–707.
116. **Ribeiro T, Abrantes M, Lopes MdFS, Crespo MTB.** 2007. Vancomycin-susceptible dairy and clinical enterococcal isolates carry vanA and vanB genes. *Int J Food Microbiol* **113**:289-295.
117. **Bell J, Tumidge J, Coorobs G, O'Brien F.** 1998. Emergence and epidemiology of vancomycin-resistant enterococci in Australia. *Commun Dis Intell* **22**:249-249.
118. **Lee RS, Gonçalves da Silva A, Baines SL, Strachan J, Ballard S, Carter GP, Kwong JC, Schultz MB, Bulach DM, Seemann T.** 2018. The changing landscape of vancomycin-resistant *Enterococcus faecium* in Australia: a population-level genomic study. *Journal of Antimicrobial Chemotherapy*.
119. **Stosor V, Peterson LR, Postelnick M, Noskin GA.** 1998. *Enterococcus faecium* bacteremia: does vancomycin resistance make a difference? *Archives of internal medicine* **158**:522-527.
120. **Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, van Embden JD, Willems RJ.** 2002. Multilocus sequence typing scheme for *Enterococcus faecium*. *Journal of clinical microbiology* **40**:1963-1971.
121. **Hall L, Duke B, Guiney M, Williams R.** 1992. Typing of *Enterococcus* species by DNA restriction fragment analysis. *Journal of clinical microbiology* **30**:915-919.
122. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95**:3140-3145.
123. **Lam MM, Seemann T, Bulach DM, Gladman SL, Chen H, Haring V, Moore RJ, Ballard S, Grayson ML, Johnson PD.** 2012. Comparative analysis of the first complete *Enterococcus faecium* genome. *Journal of bacteriology* **194**:2334-2341.

124. **Lam MM, Seemann T, Tobias NJ, Chen H, Haring V, Moore RJ, Ballard S, Grayson LM, Johnson PD, Howden BP.** 2013. Comparative analysis of the complete genome of an epidemic hospital sequence type 203 clone of vancomycin-resistant *Enterococcus faecium*. BMC Genomics **14**:595.
125. **Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S.** 2013. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. mBio **4**:e00534-00513.
126. **Top J, Willems R, Bonten M.** 2008. Emergence of CC17 *Enterococcus faecium*: from commensal to hospital-adapted pathogen. FEMS Immunology & Medical Microbiology **52**:297-308.
127. **Top J, Willems R, van der Velden S, Asbroek M, Bonten M.** 2008. Emergence of clonal complex 17 *Enterococcus faecium* in The Netherlands. Journal of clinical microbiology **46**:214-219.
128. **Johnson PD, Ballard SA, Grabsch EA, Stinear TP, Seemann T, Young HL, Grayson ML, Howden BP.** 2010. A sustained hospital outbreak of vancomycin-resistant *Enterococcus faecium* bacteremia due to emergence of vanB E. faecium sequence type 203. Journal of Infectious Diseases **202**:1278-1286.
129. **Mahony AA, Buultjens AH, Ballard SA, Grabsch EA, Xie S, Seemann T, Stuart RL, Kotsanas D, Cheng A, Heffernan H.** 2018. Vancomycin-resistant *Enterococcus faecium* sequence type 796-rapid international dissemination of a new epidemic clone. Antimicrobial Resistance & Infection Control **7**:44.
130. **Lister DM, Kotsanas D, Ballard SA, Howden BP, Carse E, Tan K, Scott C, Gillespie EE, Mahony AA, Doherty R.** 2015. Outbreak of vanB vancomycin-resistant *Enterococcus faecium* colonization in a neonatal service. American journal of infection control **43**:1061-1065.
131. **Leong KW, Cooley LA, Anderson TL, Gautam SS, McEwan B, Wells A, Wilson F, Hughson L, O'Toole RF.** 2018. Emergence of Vancomycin-Resistant *Enterococcus faecium* at an Australian Hospital: A Whole Genome Sequencing Analysis. Sci Rep **8**:6274.
132. **Coombs G, Daley D.** 2017. Australian Enterococcal Sepsis Outcome Program (AESOP) 2016: Final Report. AESOP-Australian Enterococcal Sepsis Outcome Program.

133. **Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE.** 1977. Legionnaires' disease: description of an epidemic of pneumonia. *New England Journal of Medicine* **297**:1189-1197.
134. **McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA, Dowdle WR, Team* LI.** 1977. Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *New England Journal of Medicine* **297**:1197-1203.
135. **Fliermans C, Cherry W, Orrison L, Smith S, Tison D, Pope D.** 1981. Ecological distribution of *Legionella pneumophila*. *Applied and environmental microbiology* **41**:9-16.
136. **Fliermans C, Cherry W, Orrison L, Thacker L.** 1979. Isolation of *Legionella pneumophila* from nonepidemic-related aquatic habitats. *Applied and environmental microbiology* **37**:1239-1242.
137. **Pastoris MC, Passi C, Maroli M.** 1989. Evidence of *Legionella pneumophila* in some arthropods and related natural aquatic habitats. *FEMS Microbiol Lett* **62**:259-263.
138. **Marrão G, Veríssimo A, Bowker RG, da Costa MS.** 1993. Biofilms as major sources of *Legionella* spp. in hydrothermal areas and their dispersion into stream water. *FEMS Microbiol Ecol* **12**:25-33.
139. **Hay J, Seal D, Billcliffe B, Freer J.** 1995. Non - culturable *Legionella pneumophila* associated with *Acanthamoeba castellanii*: detection of the bacterium using DNA amplification and hybridization. *Journal of applied bacteriology* **78**:61-65.
140. **Fields BS.** 1996. The molecular ecology of legionellae. *Trends Microbiol* **4**:286-290.
141. **Atlas RM.** 1999. *Legionella*: from environmental habitats to disease pathology, detection and control. *Environmental microbiology* **1**:283-293.
142. **Desai R, Welsh C, Summy M, Farone M, Newsome AL.** 1999. The potential of in situ hybridization and an immunogold assay to identify *Legionella* associations with other microorganisms. *J Microbiol Methods* **37**:155-164.
143. **Kwaik YA, Gao L-Y, Stone BJ, Venkataraman C, Harb OS.** 1998. Invasion of protozoa by *Legionella pneumophila* and its role in bacterial ecology and pathogenesis. *Applied and environmental microbiology* **64**:3127-3133.
144. **Rowbotham TJ.** 1998. Isolation of *Legionella pneumophila* serogroup 1 from human feces with use of amebic cocultures. *Clinical Infectious Diseases* **26**:502-503.

145. **Newton HJ, Ang DK, van Driel IR, Hartland EL.** 2010. Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clin Microbiol Rev* **23**:274-298.
146. **Muder RR, Victor LY, Woo AH.** 1986. Mode of transmission of *Legionella pneumophila*: a critical review. *Archives of internal medicine* **146**:1607-1612.
147. **Fields BS, Benson RF, Besser RE.** 2002. *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev* **15**:506-526.
148. **Stout J, Yu V, Best M.** 1985. Ecology of *Legionella pneumophila* within water distribution systems. *Applied and environmental microbiology* **49**:221-228.
149. **Mouchtouri VA, Goutziana G, Kremastinou J, Hadjichristodoulou C.** 2010. *Legionella* species colonization in cooling towers: risk factors and assessment of control measures. *American journal of infection control* **38**:50-55.
150. **Isberg RR, O'connor TJ, Heidtman M.** 2009. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. *Nature Reviews Microbiology* **7**:13.
151. **Xu L, Luo Z-Q.** 2013. Cell biology of infection by *Legionella pneumophila*. *Microbes and infection* **15**:157-167.
152. **Gomez-Valero L, Rusniok C, Rolando M, Neou M, Dervins-Ravault D, Demirtas J, Rouy Z, Moore RJ, Chen H, Petty NK.** 2014. Comparative analyses of *Legionella* species identifies genetic features of strains causing Legionnaires' disease. *Genome biology* **15**:505.
153. **Dominguez A, Alvarez J, Sabria M, Carmona G, Torner N, Oviedo M, Cayla J, Minguell S, Barrabeig I, Sala M.** 2009. Factors influencing the case-fatality rate of Legionnaires' disease. *The International Journal of Tuberculosis and Lung Disease* **13**:407-412.
154. **Haley CE, Cohen ML, Halter J, Meyer RD.** 1979. Nosocomial Legionnaires' disease: a continuing common-source epidemic at Wadsworth Medical Center. *Annals of internal medicine* **90**:583-586.
155. **Keller DW, Hajjeh R, DeMaria Jr A, Fields BS, Pruckler JM, Klein C, Benson RS, Kludt PE, Lett SM, Mermel LA.** 1996. Community outbreak of Legionnaires' disease: an investigation confirming the potential for cooling towers to transmit *Legionella* species. *Clinical Infectious Diseases* **22**:257-261.
156. **Den Boer JW, Yzerman P, Schellekens J, Lettinga KD, Boshuizen HC, Van Steenbergen JE, Bosman A, Van den Hof S, Van Vliet HA, Peeters MF.** 2002. A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. *Emerging Infectious Diseases* **8**:37.

157. **Ott M, Bender L, Marre R, Hacker J.** 1991. Pulsed field electrophoresis of genomic restriction fragments for the detection of nosocomial *Legionella pneumophila* in hospital water supplies. *Journal of clinical microbiology* **29**:813-815.
158. **Lück P, Köhler J, Maiwald M, Helbig JH.** 1995. DNA polymorphisms in strains of *Legionella pneumophila* serogroups 3 and 4 detected by macrorestriction analysis and their use for epidemiological investigation of nosocomial legionellosis. *Applied and environmental microbiology* **61**:2000-2003.
159. **Gaia V, Fry NK, Harrison TG, Peduzzi R.** 2003. Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *Journal of clinical microbiology* **41**:2932-2939.
160. **Gaia V, Fry NK, Afshar B, Lück PC, Meugnier H, Etienne J, Peduzzi R, Harrison TG.** 2005. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *Journal of clinical microbiology* **43**:2047-2052.
161. **Ratzow S, Gaia V, Helbig JH, Fry NK, Lück PC.** 2007. Addition of *neuA*, the gene encoding N-acetylneuraminyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing *Legionella pneumophila* serogroup 1 strains. *Journal of clinical microbiology* **45**:1965-1968.
162. **Mentasti M, Underwood A, Lück C, Kozak - Muiznieks N, Harrison T, Fry N.** 2014. Extension of the *Legionella pneumophila* sequence - based typing scheme to include strains carrying a variant of the N - acetylneuraminyl transferase gene. *Clinical Microbiology and Infection* **20**:0435-0441.
163. **Harrison T, Afshar B, Doshi N, Fry N, Lee J.** 2009. Distribution of *Legionella pneumophila* serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000–2008). *European journal of clinical microbiology & infectious diseases* **28**:781.
164. **Sánchez-Busó L, Guiral S, Crespi S, Moya V, Camaró ML, Olmos MP, Adrián F, Morera V, González-Morán F, Vanaclocha H.** 2016. Genomic investigation of a legionellosis outbreak in a persistently colonized hotel. *Frontiers in Microbiology* **6**:1556.

165. **Graham R, Doyle C, Jennison A.** 2014. Real-time investigation of a *Legionella pneumophila* outbreak using whole genome sequencing. *Epidemiology & Infection* **142**:2347-2351.
166. **Lévesque S, Plante P-L, Mendis N, Cantin P, Marchand G, Charest H, Raymond F, Huot C, Goupil-Sormany I, Desbiens F.** 2014. Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. *PLoS One* **9**:e103852.
167. **McAdam P, van der Broek C, Lindsay D, Ward M, Hanson M, Gillies M, Watson M, Stevens J, Edwards G, Fitzgerald R.** 2014. Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol* **15**:504.
168. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents.
169. **Greig JE, Carnie JA, Tallis GF, Ryan NJ, Tan AG, Gordon IR, Zwolak B, Leydon JA, Guest CS, Hart WG.** 2004. An outbreak of Legionnaires' disease at the Melbourne Aquarium, April 2000: investigation and case-control studies. *Medical Journal of Australia* **180**:566-572.
170. **Mac CP, Tolhurst JC, et al.** 1948. A new *mycobacterial* infection in man. *J Pathol Bacteriol* **60**:93-122.
171. **Johnson PD, Stinear T, Small PL, Pluschke G, Merritt RW, Portaels F, Huygen K, Hayman JA, Asiedu K.** 2005. Buruli ulcer (*M. ulcerans* infection): new insights, new hope for disease control. *PLoS Med* **2**:e108.
172. **Organization WH.** 2009. Buruli ulcer: first programme review meeting for west Africa-summary report. *Wkly Epidemiol Rec* **84**:45.
173. **Amofah G, Bonsu F, Tetteh C, Okrah J, Asamoah K, Asiedu K, Addy J.** 2002. Buruli ulcer in Ghana: results of a national case search. *Emerg Infect Dis* **8**:167-170.
174. **Debacker M, Aguiar J, Steunou C, Zinsou C, Meyers WM, Guedenon A, Scott JT, Dramaix M, Portaels F.** 2004. *Mycobacterium ulcerans* disease (Buruli ulcer) in rural hospital, Southern Benin, 1997-2001. *Emerg Infect Dis* **10**:1391-1398.
175. **Marston BJ, Diallo MO, Horsburgh Jr CR, Diomande I, Saki MZ, Kanga J-M, Patrice G, Lipman HB, Ostroff SM, Good RC.** 1995. Emergence of Buruli ulcer disease in the Daloa region of Cote d'Ivoire. *Am J Trop Med Hyg* **52**:219-224.

176. **Hayman J.** 1985. Clinical features of *Mycobacterium ulcerans* infection. *Australas J Dermatol* **26**:67-73.
177. **Francis G, Whitby M, Woods M.** 2006. *Mycobacterium ulcerans* infection: a rediscovered focus in the Capricorn Coast region of central Queensland. *Med J Aust* **185**:179-180.
178. **Johnson PD, Veitch MG, Leslie DE, Flood PE, Hayman JA.** 1996. The emergence of *Mycobacterium ulcerans* infection near Melbourne. *Med J Aust* **164**:76-78.
179. **Hayman J, McQueen A.** 1985. The pathology of *Mycobacterium ulcerans* infection. *J. Pathol.* **17**:594-600.
180. **George KM, Chatterjee D, Gunawardana G, Welty D, Hayman J, Lee R, Small PL.** 1999. Mycolactone: a polyketide toxin from *Mycobacterium ulcerans* required for virulence. *Science* **283**:854-857.
181. **Stinear TP, Mve-Obiang A, Small PL, Frigui W, Pryor MJ, Brosch R, Jenkin GA, Johnson PD, Davies JK, Lee RE, Adusumilli S, Garnier T, Haydock SF, Leadlay PF, Cole ST.** 2004. Giant plasmid-encoded polyketide synthases produce the macrolide toxin of *Mycobacterium ulcerans*. *Proc Natl Acad Sci U S A* **101**:1345-1349.
182. **Van der Werf TS, Van der Graaf WT, Tappero JW, Asiedu K.** 1999. *Mycobacterium ulcerans* infection. *Lancet* **354**:1013-1018.
183. **Portaels F, Aguiar J, Debacker M, Guedenon A, Steunou C, Zinsou C, Meyers WM.** 2004. *Mycobacterium bovis* BCG Vaccination as Prophylaxis against *Mycobacterium ulcerans* Osteomyelitis in Buruli Ulcer Disease. *Infect Immun* **72**:62-65.
184. **Chauty A, Ardant M-F, Marsollier L, Pluschke G, Landier J, Adeye A, Goundoté A, Cottin J, Ladikpo T, Ruf T, Ji B.** 2011. Oral Treatment for *Mycobacterium ulcerans* Infection: Results From a Pilot Study in Benin. *Clin Infect Dis* **52**:94-96.
185. **Nienhuis WA, Stienstra Y, Thompson WA, Awuah PC, Abass KM, Tuah W, Awua-Boateng NY, Ampadu EO, Siegmund V, Schouten JP, Adjei O, Bretzel G, van der Werf TS.** 2010. Antimicrobial treatment for early, limited *Mycobacterium ulcerans* infection: a randomised controlled trial. *Lancet* **375**:664-672.
186. **O'Brien DP, Hughes AJ, Cheng AC, Henry MJ, Callan P, McDonald A, Holten I, Birrell M, Sowerby JM, Johnson PD, Athan E.** 2007. Outcomes for *Mycobacterium ulcerans* infection with combined surgery and antibiotic therapy: findings from a south-eastern Australian case series. *Med J Aust* **186**:58-61.

187. **Sizaïre V, Nackers F, Comte E, Portaels F.** 2006. *Mycobacterium ulcerans* infection: control, diagnosis, and treatment. *Lancet Infect Dis* **6**:288-296.
188. **Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, Savioli L.** 2007. Control of neglected tropical diseases. *N. Engl. J. Med.* **357**:1018-1027.
189. **Asiedu K, Etuaful S.** 1998. Socioeconomic implications of Buruli ulcer in Ghana: a three-year review. *Am J Trop Med Hyg* **59**:1015-1022.
190. **Raghunathan PL, Whitney EA, Asamoah K, Stienstra Y, Taylor TH, Jr., Amofah GK, Ofori-Adjei D, Dobos K, Guarner J, Martin S, Pathak S, Klutse E, Etuaful S, van der Graaf WT, van der Werf TS, King CH, Tappero JW, Ashford DA.** 2005. Risk factors for Buruli ulcer disease (*Mycobacterium ulcerans* Infection): results from a case-control study in Ghana. *Clin Infect Dis* **40**:1445-1453.
191. **Aiga H, Amano T, Cairncross S, Adomako J, Nanas OK, Coleman S.** 2004. Assessing water-related risk factors for Buruli ulcer: a case-control study in Ghana. *Am J Trop Med Hyg* **71**:387-392.
192. **Stinear T, Ross BC, Davies JK, Marino L, Robins-Browne RM, Oppedisano F, Sievers A, Johnson PD.** 1999. Identification and characterization of IS2404 and IS2606: two distinct repeated sequences for detection of *Mycobacterium ulcerans* by PCR. *J Clin Microbiol* **37**:1018-1023.
193. **Ross BC, Marino L, Oppedisano F, Edwards R, Robins-Browne RM, Johnson PD.** 1997. Development of a PCR assay for rapid diagnosis of *Mycobacterium ulcerans* infection. *J Clin Microbiol* **35**:1696-1700.
194. **Portaels F, Elsen P, Guimaraes-Peres A, Fonteyne P-A, Meyers WM.** 1999. Insects in the transmission of *Mycobacterium ulcerans* infection. *Lancet* **353**:986.
195. **Duker AA, Portaels F, Hale M.** 2006. Pathways of *Mycobacterium ulcerans* infection: a review. *Environ Int* **32**:567-573.
196. **Merritt RW, Benbow ME, Small PLC.** 2005. Unraveling an emerging disease associated with disturbed aquatic environments: the case of Buruli ulcer. *Front Ecol Environ* **3**:323-331.
197. **Marsollier L, Robert R, Aubry J, Saint Andre JP, Kouakou H, Legras P, Manceau AL, Mahaza C, Carbonnelle B.** 2002. Aquatic insects as a vector for *Mycobacterium ulcerans*. *Appl Environ Microbiol* **68**:4623-4628.

198. **van Zyl A, Daniel J, Wayne J, McCowan C, Malik R, Jelfs P, Lavender CJ, Fyfe JA.** 2010. *Mycobacterium ulcerans* infections in two horses in south-eastern Australia. *Aust Vet J* **88**:101-106.
199. **Elsner L, Wayne J, O'Brien CR, McCowan C, Malik R, Hayman JA, Globan M, Lavender CJ, Fyfe JA.** 2008. Localised *Mycobacterium ulcerans* infection in a cat in Australia. *J Feline Med Surg* **10**:407-412.
200. **Mitchell PJ, Jerrett IV, Slee KJ.** 1984. Skin ulcers caused by *Mycobacterium ulcerans* in koalas near Bairnsdale, Australia. *Pathology* **16**:256-260.
201. **Portaels F, Chemlal K, Elsen P, Johnson PD, Hayman JA, Hibble J, Kirkwood R, Meyers WM.** 2001. *Mycobacterium ulcerans* in wild animals. *Rev Sci Tech* **20**:252-264.
202. **Fyfe JA, Lavender CJ, Handasyde KA, Legione AR, O'Brien CR, Stinear TP, Pidot SJ, Seemann T, Benbow ME, Wallace JR, McCowan C, Johnson PD.** 2010. A major role for mammals in the ecology of *Mycobacterium ulcerans*. *PLoS Negl Trop Dis* **4**:e791.
203. **Stinear TP, Jenkin GA, Johnson PD, Davies JK.** 2000. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. *J Bacteriol* **182**:6322-6330.
204. **Portaels F, Fonteyne PA, de Beenhouwer H, de Rijk P, Guedenon A, Hayman J, Meyers MW.** 1996. Variability in 3' end of 16S rRNA sequence of *Mycobacterium ulcerans* is related to geographic origin of isolates. *J Clin Microbiol* **34**:962-965.
205. **Hilty M, Yeboah-Manu D, Boakye D, Mensah-Quainoo E, Rondini S, Schelling E, Ofori-Adjei D, Portaels F, Zinsstag J, Pluschke G.** 2006. Genetic diversity in *Mycobacterium ulcerans* isolates from Ghana revealed by a newly identified locus containing a variable number of tandem repeats. *J Bacteriol* **188**:1462-1465.
206. **Chemlal K, De Ridder K, Fonteyne PA, Meyers WM, Swings J, Portaels F.** 2001. The use of IS2404 restriction fragment length polymorphisms suggests the diversity of *Mycobacterium ulcerans* from different geographical areas. *Am J Trop Med Hyg* **64**:270-273.
207. **Chemlal K, Huys G, Fonteyne PA, Vincent V, Lopez AG, Rigouts L, Swings J, Meyers WM, Portaels F.** 2001. Evaluation of PCR-restriction profile analysis and IS2404 restriction fragment length polymorphism and amplified fragment length polymorphism fingerprinting for identification and typing of *Mycobacterium ulcerans* and *M. marinum*. *J Clin Microbiol* **39**:3272-3278.

208. **Ablordey A, Swings J, Hubans C, Chemlal K, Locht C, Portaels F, Supply P.** 2005. Multilocus variable-number tandem repeat typing of *Mycobacterium ulcerans*. J Clin Microbiol **43**:1546-1551.
209. **Stragier P, Ablordey A, Meyers WM, Portaels F.** 2005. Genotyping *Mycobacterium ulcerans* and *Mycobacterium marinum* by using mycobacterial interspersed repetitive units. J Bacteriol **187**:1639-1647.
210. **Huys G, Rigouts L, Chemlal K, Portaels F, Swings J.** 2000. Evaluation of Amplified Fragment Length Polymorphism Analysis for Inter-and Intraspecific Differentiation of *Mycobacterium bovis*, *M. tuberculosis*, and *M. ulcerans*. J Clin Microbiol **38**:3675-3680.
211. **Ablordey A, Kotlowski R, Swings J, Portaels F.** 2005. PCR amplification with primers based on IS2404 and GC-rich repeated sequence reveals polymorphism in *Mycobacterium ulcerans*. J Clin Microbiol **43**:448-451.
212. **Stinear T, Davies JK, Jenkin GA, Portaels F, Ross BC, Oppedisano F, Purcell M, Hayman JA, Johnson PD.** 2000. A simple PCR method for rapid genotype analysis of *Mycobacterium ulcerans*. J Clin Microbiol **38**:1482-1487.
213. **Kaser M, Gutmann O, Hauser J, Stinear T, Cole S, Yeboah-Manu D, Dernick G, Certa U, Pluschke G.** 2009. Lack of insertional-deletional polymorphism in a collection of *Mycobacterium ulcerans* isolates from Ghanaian Buruli ulcer patients. J Clin Microbiol **47**:3640-3646.
214. **Stinear TP, Seemann T, Pidot S, Frigui W, Reyssset G, Garnier T, Meurice G, Simon D, Bouchier C, Ma L, Tichit M, Porter JL, Ryan J, Johnson PD, Davies JK, Jenkin GA, Small PL, Jones LM, Tekai F, Laval F, Daffe M, Parkhill J, Cole ST.** 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res **17**:192-200.
215. **Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG.** 2001. Massive gene decay in the *leprosy bacillus*. Nature **409**:1007-1011.
216. **Tonjum T, Welty DB, Jantzen E, Small PL.** 1998. Differentiation of *Mycobacterium ulcerans*, *M. marinum*, and *M. haemophilum*: mapping of their relationships to *M.*

- tuberculosis* by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequence analysis. J Clin Microbiol **36**:918-925.
217. **Yip MJ, Porter JL, Fyfe JA, Lavender CJ, Portaels F, Rhodes M, Kator H, Colorni A, Jenkin GA, Stinear T.** 2007. Evolution of *Mycobacterium ulcerans* and other mycolactone-producing *mycobacteria* from a common *Mycobacterium marinum* progenitor. J Bacteriol **189**:2021-2029.
 218. **Doig KD, Holt KE, Fyfe JA, Lavender CJ, Eddyani M, Portaels F, Yeboah-Manu D, Pluschke G, Seemann T, Stinear TP.** 2012. On the origin of *Mycobacterium ulcerans*, the causative agent of buruli ulcer. BMC Genomics **13**:258.
 219. **Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsche-Gerdes S, Supply P, Kalinowski J, Niemann S.** 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS Med **10**:e1001387.
 220. **Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K.** 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A **99**:3684-3689.
 221. **Ablordey AS, Vandelannoote K, Frimpong IA, Ahoritor EK, Amissah NA, Eddyani M, Durnez L, Portaels F, de Jong BC, Leirs H.** 2015. Whole genome comparisons suggest random distribution of *Mycobacterium ulcerans* genotypes in a Buruli ulcer endemic region of Ghana. PLoS Negl Trop Dis **9**:e0003681.
 222. **Vandelannoote K, Meehan CJ, Eddyani M, Affolabi D, Phanzu DM, Eyangoh S, Jordaens K, Portaels F, Mangas K, Seemann T.** 2017. Multiple Introductions and Recent Spread of the Emerging Human Pathogen *Mycobacterium ulcerans* across Africa. Genome Biol Evol **9**:414-426.
 223. **Jackson K, Edwards R, Leslie DE, Hayman J.** 1995. Molecular method for typing *Mycobacterium ulcerans*. J Clin Microbiol **33**:2250-2253.
 224. **Veitch MG, Johnson PD, Flood PE, Leslie DE, Street AC, Hayman JA.** 1997. A large localized outbreak of *Mycobacterium ulcerans* infection on a temperate southern Australian island. Epidemiol Infect **119**:313-318.
 225. **Boyd SC, Athan E, Friedman ND, Hughes A, Walton A, Callan P, McDonald A, O'Brien DP.** 2012. Epidemiology, clinical features and diagnosis of *Mycobacterium ulcerans* in an Australian population. Med J Aust **196**:341-344.

226. **Tiong A.** 2005. The epidemiology of a cluster of *Mycobacterium ulcerans* infections in Point Lonsdale. Vic Infect Dis Bull **8**:2-4.
227. **van Ravensway J, Benbow ME, Tsonis AA, Pierce SJ, Campbell LP, Fyfe JA, Hayman JA, Johnson PD, Wallace JR, Qi J.** 2012. Climate and landscape factors associated with Buruli ulcer incidence in Victoria, Australia. PLoS One **7**:e51074.
228. **Hayman J.** 1991. Postulated Epidemiology of *Mycobacterium ulcerans* Infection. Int J Epidemiol **20**:1093-1098.
229. **Williamson HR, Benbow ME, Nguyen KD, Beachboard DC, Kimbirauskas RK, McIntosh MD, Quaye C, Ampadu EO, Boakye D, Merritt RW.** 2008. Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana. PLoS Negl Trop Dis **2**:e205.
230. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP.** 2015. Prospective whole genome sequencing enhances national surveillance of *Listeria monocytogenes*. Journal of clinical microbiology:JCM. 02344-02315.
231. **Wassilew N, Seth-Smith HM, Rolli E, Fietze Y, Casanova C, Führer U, Egli A, Marschall J, Buetti N.** 2018. Outbreak of vancomycin-resistant *Enterococcus faecium* clone ST796, Switzerland, December 2017 to April 2018. Eurosurveillance **23**:1800351.
232. **Pidot SJ, Gao W, Buultjens AH, Monk IR, Guerillot R, Carter GP, Lee JYH, Lam MMC, Grayson ML, Ballard SA, Mahony AA, Grabsch EA, Kotsanas D, Korman TM, Coombs GW, Robinson JO, Gonçalves da Silva A, Seemann T, Howden BP, Johnson PDR, Stinear TP.** 2018. Increasing tolerance of hospital *Enterococcus faecium* to handwash alcohols. Sci Transl Med **10**.