Title Page

Investigation of transmission of human respiratory pathogens using whole genome sequencing

Volume 1

Patrick Stapleton

Supervisor: Prof Martin Cormican, School of Medicine, Nursing & Health Sciences; Bacteriology; University of Galway

Submitted March 2023

Table of Contents

Summary

This doctoral thesis explores the role of whole genome sequencing in the investigation of transmission of human respiratory pathogens. It is a "PhD thesis by publication" which consists primarily of three published first-author journal articles that describe investigations of the transmission of Mumps virus, Adenovirus and *Pseudomonas aeruginosa* respectively. The work was undertaken in Toronto, Canada between 2016 and 2019, when the author was a Clinical and Research Fellow in Medical Microbiology at the University of Toronto, the Hospital for Sick Children and the Public Health Ontario Laboratory, as well as a PhD student with University of Galway supervised by Prof. Martin Cormican. The topics of the research were chosen for (1) their relevance to public health or hospital outbreak prevention and infection control and (2) as contexts in which best practices had yet to be established for genomic epidemiology investigations using whole genome sequencing. The majority of infections described occurred in the greater Toronto area.

The introduction outlines the development of early generation pathogen typing techniques and the advent of low-cost and widely distributed whole genome sequencing technologies. The strengths and limitations of sequencing technologies that have seen widespread adoption in the field of clinical microbiology are described. An outline of the epidemiological context for each of the three articles is provided, along with rationales for using sequencing to explore transmission. The focus of the publications consists of (1) investigating the community spread of a Mumps virus outbreak in Ontario, (2) investigating a prolonged nosocomial outbreak of human Adenovirus-A31 affecting a paediatric bone marrow transplantation unit over 3 years and (3) the retrospective interrogation of a large collection of *P. aeruginosa* isolates from children with Cystic Fibrosis to identify cryptic nosocomial transmission.

The discussion section that follows the main chapters describes how the findings from all investigations illustrate the need for clinicians to employ an approach to genomic epidemiology that can define thresholds for relatedness between pathogens where this has not already been clearly established, in order to "rule in" or "rule out" transmission as the core finding. The novelty and significance of specific findings from each study are identified and commonalities between them discussed. Finally, a post-script section explores how the lessons learned can be applied to the integration of whole genome sequencing into the routine work of clinical microbiology in hospitals laboratories in Ireland to support outbreak investigation. It briefly describes how initial steps towards this integration taken during the course of the Covid-19 pandemic through the setup of a national laboratory network for whole genome sequencing of SARS-CoV-2.

Acknowledgements

I am deeply grateful to the National Doctors in Training Program for awarding me support from the Dr Richard Steevens Scholarship to take up a clinical fellowship in medical microbiology in the Hospital for Sick Children (SickKids) in 2016, which started me on my journey to specialise in genomic epidemiology. Credit for this is also due to Ronan Leahy from the Temple Street Paediatric Infectious Disease (ID) service for putting me in touch with former colleagues in SickKids Microbiology and ID who helped setup my fellowships (clinical first, followed by research) and became my supervisors. To them, Yvonne Yau and Valerie Waters, I owe a huge debt of gratitude for supporting me through three years of fellowship in clinical and research duties but also for helping make the transition of a young and growing family to a new country and a large unfamiliar city so easy and welcoming. They are valued mentors whose insights on balancing work and family commitments and on maintaining varied and stimulating interests throughout ones career stay with me.

I could not have completed the research fellowship without the support of Cystic Fibrosis Canada who awarded me the Jennifer and Robert Sturgess Research Fellowship to study cross transmission of *Pseudomonas aeruginosa*; I particularly want to thank their review board for giving an early stage researcher without an extensive track record a chance to work on a large multi-year project. The final piece of the puzzle for funding support was arranged by Susan Richardson as head of department of Medical Microbiology and later as head of Laboratory Medicine. She assisted with fellowship extension and managed to find some crucial salary for occasional clinical on-call commitments at SickKids (because as Val says about research fellowships “you gotta eat”). Susan became another inspiring mentor during the course of our long collaboration on the adenovirus project; at her retirement party in 2019 I was struck by just how much she had contributed to her hospital and her field of paediatric microbiology over the course of a long and distinguished career. I am particularly thankful for her continuing support in her Emeritus position for the adenovirus project for several years after my fellowship ended to secure publication of the manuscript. Together with co-first author Ramzi Fattouh, my predecessor as SickKids clinical fellow we formed a triumvirate of dedicated senior and first authors who worked on submissions, resubmissions and a multitude of requested revisions in late night zoom meetings all the way to acceptance at the end of 2022.

From Val’s research lab Ana Blanchard and Trevor Beaudoin were wise and kind compatriots who enlivened many days spent working up stacks of Pseudomonas culture plates at the Peter Gilgan Research Institute; I wish them well in their promising new careers. At Public Health Ontario Samir

Patel was always supportive of my desire to gain experience on genomic epidemiology projects by linking with as many teams as possible; he was a supportive supervisor for both the mumps and the adenovirus projects. Critical to the success of both was the technical expertise of AliReza Eshagi who taught me all I know about sanger and whole genome sequencing protocols, multichannel pipettes and checking amplification products on agarose gels; he was very patient with an inexperienced medical doctor getting to grips with molecular microbiology techniques and also taught me some of the rudiments of bioinformatics. Other colleagues at Public Health Ontario especially Jonathan Gubbay and Sarah Wilson were very helpful with a trialling a sequencing approach to investigate a community mumps outbreak that would really come into its own in the Covid-19 pandemic. Aaron Campigotto at SickKids was a valuable role model as a young consultant microbiologist starting in a new post and eager to integrate sequencing into research and clinical activities.Our trips to the genome diagnostics department to beg the use of their extensive sequencing equipment for the adenovirus project were the first step on an important transition for the SickKids microbiology department.

David Guttman was kind enough to take me on, sight unseen, to his lab at University of Toronto as part of his ongoing collaborations with Val and Yvonne. I am particularly thankful to him for formally taking on the role of co-supervisor of my PhD. His weekly lab meeting where I and the other post-grad researchers such as Marcus Dillon and Renan Almeida presented project updates brought a welcome structure to the my project. It was a helpful forum for open and sometimes robust feedback which imposed a new level of rigour on my analyses. David's door was always open when I needed a steer on how to proceed and his manuscript and figure polishing skills are fearsome. Thanks also to fellow postgrads Shawn Clarke for sequencing over a thousand bacterial isolates for our project, to Julio Diaz Caballero for sharing his bioinformatics expertise without which I would have been utterly lost in the early years. Special thanks to Conrad Izydorczyk for working with me on the bacterial isolate collection. We analysed for different projects but developed a joint approach to ensure robust bioinformatics quality control, troubleshoot sequencing contamination issues and to keep track of the many isolates and sequences in the collection. He also helped out by creating bespoke R and python scripts which were beyond my informatics ability. Also from the Guttman lab Pauline Wang was always kind and supportive as supervisor of the genome sequencing lab and facilitated all requests for additional sequencing runs.

Martin Cormican's role as lead PhD supervisor is only the latest way in which he has supported me from the earliest days of my medical career. From inspiring me with a love of the discipline as an undergraduate, to providing clear eyed advice on the joys and challenges of the career when I was

making decisions on higher specialist training, he has always provided wise counsel and supervision that continued through the complex process of co-ordinating my research activity between Canada and Ireland. Although it was contemporaneous with the PhD rather than part of it, I cannot let his term as Antimicrobial Resistance and Infection Control (AMRIC) lead before and especially during the Covid-19 pandemic go unmentioned. His support not only to me but to all my colleagues nationwide felt like a crucial lifeline in a time of crisis. The weekly calls, impromptu check-ins and his availability at all hours for advice provided not just practical help and responsive guidance but also crucial psychological and emotional support during a professional experience that might otherwise have been completely overwhelming. I also wish to acknowledge his support while on the National SARS-CoV-2 Whole Genome Sequencing and Surveillance steering group, to which we were both appointed on its establishment in 2021. The vision set out by this group of a hub and spoke model was subsequently realised in 2022. The network now enables weekly whole genome sequencing runs in seven hospital clinical laboratories (the majority of which had no previous local sequencing capacity) plus the National Virus Reference Laboratory for real time surveillance and outbreak investigation. In a certain way it represents the culmination and validation of the work set out in this thesis.

Thanks also to members of my Graduate Review Committee (GRC) committee for their counsel and advice and to the infection prevention and control team in SickKids for their work to investigate and terminate the adenovirus outbreak and subsequent assistance with manuscript preparation over many years. Members of reference laboratories in Dublin and Winnipeg who contributed samples or expertise for the adenovirus project have my thanks also.

Last of all I want to thank my friends and family for their support over the past five years: my parents for continuing to encourage my academic endeavours (not forgetting the help with childcare for three young children). Lee, Rob and Antoine, medical and clinical microbiology trainees and friends from Toronto, for the game nights that kept me sane, the stories of PhD completion that kept me encouraged, and for opening my eyes to the benefits and possibilities of medically and scientifically trained consultant microbiologists working together in teams. Finally but most importantly my ever supportive, loving and patient wife Aoife; the rock of stability for our family and the source of all of our strength - thank you for moving to Canada and traipsing to work through snowdrifts while pregnant to support me in this endeavour; without you none of this would be possible.

Introduction

Purpose

This work explores the use of whole genome sequencing (WGS) of microbes to investigate person-to-person cross-transmission of three distinct respiratory pathogens and in both community and hospital settings. WGS has the highest resolution of available microbiological techniques for distinguishing respiratory pathogens obtained from different human hosts [1]. If pathogens from different hosts are indistinguishable or very closely related by WGS, then a recent common ancestor is likely and recent cross-transmission may have occurred, either by direct spread or indirectly via the environment [2]. Understanding transmission pathways using WGS aids development of effective infection prevention and control and informs mitigation measures, but crucially any analysis must acknowledge a key limitation; whilst WGS analysis can in essence conclusively refute that recent transmission of the organism sequenced has occurred, it cannot demonstrate that cross-transmission occurred without strong supporting epidemiological evidence. Sequencing technology has matured to the extent that sequencers now commercially available permit routine sequencing at scale of diverse pathogens in well-resourced academic and reference microbiology laboratory environments[3], but much of this work focuses on routine surveillance of pathogens of public health importance, of which enteric bacteria feature prominently [4,5]. In this work WGS led investigations of outbreaks of Mumps and Adenovirus are detailed that were performed at a time when neither pathogen had been previously interrogated using WGS. In addition, an investigation of transmission of *P. aeruginosa* is the first to use very large scale WGS to examine in detail the question of acquisition of early infection (including mixed strain infection) in children with Cystic Fibrosis (CF).

Background

1. Whole genome sequencing for transmission analysis

The relatively recent arrival of cheap, widely accessible genomic sequencing for a multitude of scientific and clinical applications is probably the single most important development in biology this century[6]. Reading an organisms nucleic acid makeup, whether RNA or DNA can in theory provide us with the blueprint for replication of that organism and it's biological functions. Pioneering advancements in the underlying technology were achieved via massive public and private

investment in the last quarter of the 20th century, primarily through multiple parallel efforts to sequence the first human genome[7]. When this aspiration was met, the goal expanded to sequencing thousands and then millions of human genomes[8]. This demand catalysed rapid growth of a market for sequencing technology, with a commensurate dramatic fall in price from 300 million US dollars (USD) to 1000 USD to sequence a human genome. The fall in costs has plateaued since 2015 and only incorporates consumable costs[9]. The steady decrease in the cost of sequencing is commonly compared with the predictable decreased cost and increased processing speed of microchips embodied by "Moore's Law". These technological developments were accompanied by an expanded suite of applications beyond sequencing human chromosomal DNA. WGS now encompasses animals, plants and microbes. Modern examples of the utility of genome sequencing include the more recent field of genome editing facilitated by techniques such as CRISPR-Cas9[10]. Specific applications include the design of new disease resistant crops to offset the impact of climate change[11] and deepening our understanding of evolutionary processes such as the complex origins of *Homo* sapiens and the evolution of plants and animals generally[12–14]. However, for the purposes of this work the WGS development of interest is with respect to clinical microbiology.

WGS is a tool that allows for comparison of pathogens at the nucleotide level, allowing far higher resolution analysis of transmission networks than was possible with legacy methods, detailed below. Transmission analysis is not the only microbial WGS application relevant to clinical microbiology. Since genome sequence predicts phenotypes, it can be used to predict susceptibility or resistance to antimicrobials, which is relevant for both viral and bacterial pathogens[15,16]. It can also be used to identify the presence or absence virulence factors such as genes encoding toxins or effector systems (or alternations in existing genes such as mutations in promoter regions, leading to overexpression) which may result in altered disease severity. As parasites and fungi are Eukaryotes they have more complex genomes compared to Prokaryotic bacteria or to viruses which may have only a single stranded RNA or DNA molecule. Eukaryotic pathogen genomics therefore remains relatively underdeveloped for clinical microbiology applications[17] and is not considered further here.

The relative simplicity of viral genomes makes them tractable from the technical process of sequencing, as outlined in Mumps and Adenovirus chapters methodology, which describe both enrichment and enrichment free sequencing approaches. Viruses are also ideal subjects from the perspective of learning the rudiments of genomic analysis. Virus genomes contain in some cases only a dozen genes which occur in predictable order. This relative lack of complexity is due to a small genome size[18], on the order of a few thousand to several hundred thousand nucleotides for the clinically relevant viruses that infect humans. From a bioinformatics perspective, this facilitates a

straightforward bioinformatic inter-genome comparison at the nucleotide level. The purpose of this comparison is to determine single nucleotide polymorphisms (SNP), i.e. individual nucleotide differences in the base sequences. Such nucleotide comparison may be between a set of genomes all sequenced for the same project or it may commonly also incorporate external comparison to the genome set e.g. using a well characterized reference genome. External comparison to databases containing sequences associated with particular phenotypic traits such as antimicrobial resistance[19] underpin many WGS applications relevant to clinical microbiology. The term single nucleotide variant (SNV) also appears in the literature and is used in the first article presented in this work. For all practical purposes related to transmission analysis the terms SNP and SNV are interchangeable, and SNP appears to be favoured in recent years.

Nucleotide comparison involving bacteria is considerably more complex, as discussed in more depth in the article involving bacterial genome (*P. aeruginosa*) WGS based transmission analysis. This greater genomic complexity results not only from larger genome size of several million bases, but also from the potential presence of mobile genetic elements. Such elements may contain entire gene complexes. Large-scale genome rearrangements of nucleotide order within a bacterial chromosome are also possible. Finally, repetitive regions are always present in a bacterial genome. These contain long stretches of repetitive nucleotide sequence and are distributed at intervals around the bacterial chromosome. These regions pose particular bioinformatic challenges when using certain sequencing approaches as discussed in further detail in the section on WGS technology[20].

The WGS application in this work is directly concerned with the comparison of respiratory pathogen genomes to identify if they are so closely related at the nucleotide level that the likely explanation is recent transmission from one source to another. It involves identifying potential cross-transmission from the similarity of WGS data, investigating for other sources of evidence to support the WGS analysis and discussing the infection control considerations arising from results. For the pathogens considered in this work, it is rare for genomes recovered from different hosts to be absolutely identical at the nucleotide level. This is not surprising as accumulation of genomic mutations is inevitable due to errors that occur during nucleic acid replication. The rate of mutation is pathogen and context dependant but it occurs at a higher rate in viruses than bacteria [21,22]. Since identifying identical pathogen genomes from different hosts rarely occurs, even in the presence of recent cross-transmission, it is necessary to define the number of sequence changes between genomes that is still plausibly consistent with recent transmission. Stated another way, it is necessary to set a maximum number of changes (a SNP difference cut-off) allowed between pairs of pathogen

genomes when conducting transmission analysis. If SNP differences above this threshold are identified then the possibility of recent transmission is effectively ruled out. A key challenge encountered in the main chapters of this work is that consensus SNP difference thresholds had not been previously established for the three pathogens considered. Indeed, very few mumps or adenovirus A31 whole genome sequences had ever been published prior to commencement of our studies. It was thus necessary to establish and justify SNP thresholds in the manuscripts when using SNP distances between isolates to argue that transmission had or had not occurred in a particular instance.

For establishing strain relatedness, the mutations of interest are those that have been inherited from the pathogen's common ancestor through the process of random nucleotide substitution during genome replication i.e. by vertical inheritance. Over short timescales, for example an outbreak of respiratory pathogen in a hospital ward, the exact location and functional effect of the mutation is less important than the absolute number of distinct nucleotide mutations accrued. This is for two main reasons; because the process of natural selection requires time to purify deleterious mutations or positively select for beneficial ones, and because most random mutations are "silent" or synonymous[23]. Unlike non-synonymous mutations, these do not change the instruction encoded by a codon resulting in change in the amino acid content determined by a gene. They also do not alter the function of an important genomic region such as a gene promotor. In the short timescales involved in many outbreaks of respiratory pathogens, the purifying effects of natural selection (the removal of random deleterious mutations from the pathogen population) does not occur. There is therefore no need to model how many deleterious mutations might have occurred during vertical inheritance and subsequently been removed by purifying selection from the population before the genomic analysis on the descendants is performed. Accounting for such backwards mutation is a pertinent issue in long scale evolutionary analyses but it is valid when describing differences between pairs of genomes from an outbreak to use a straightforward measure such as SNP distance (the absolute number of mutations identified between genomes) without need to account for the effects of natural selection[24].

Genome mutations by random substitution occur at a relatively constant rate that is often well characterised for a particular pathogen. This rate (termed the molecular clock) is primarily dependant on how error prone the genome replication process is. Even if the molecular clock is not known with confidence for a particular organism it can be estimated based on the genome replication error rate and presence or absence of nucleotide repair mechanisms – RNA viruses typically have mutation rates orders of magnitude higher than DNA viruses or bacteria[25,26] .

Bioinformatic analysis can identify the number of nucleotide differences or SNPs between genomes and these usually result from rate dependant accumulation of mutations in a predictable fashion over time (exceptions are discussed below). It is therefore possible to infer when pathogens last shared a common ancestor and thus to determine if recent transmission occurred. If pathogens are indistinguishable at the nucleotide level it is usually possible to infer that transmission has occurred recently, although this can be a chance finding, as "*low viral sequence diversity limits the power of genomics to infer transmission clusters…it is possible some genetically similar viruses are from unconnected introduction events*"[27].

As mentioned earlier, over evolutionary timescales the effects of natural selection must be incorporated into any analysis, for example by using nucleotide substitution models when performing maximum likelihood analysis to generate a phylogenetic tree of genomes collected years or decades apart that graphically displays isolate relationships[28,29]. Such analyses are discussed in the three articles when comparing genomes from outbreaks to ancestral strains and non-outbreak strains isolated many years or decades prior to the outbreak. Genome reorganisation can further complicate the picture, for example mobile elements for bacteria[30], and recombination for both bacteria and viruses [31,32]. At a minimum, it is necessary to establish if the presence of mobile genetic elements is a consideration when analysing cross-transmission. The basic assumption when discussing SNP differences is that they result from random substitution in genome sequences that are descended from a common ancestor, unless specifically stated otherwise. However not all nucleotide differences between pathogens result from mutation following descent from common ancestor. During recombination, which can occur when two pathogens of the same or similar species occupy the same host or environment, they can exchange entire segments of genome encoding more than one gene. A long stretch of exogenous genome therefore replaces the prior nucleotide sequence. Such events can artificially inflate the apparent time to last common shared ancestor (can increase the SNP distance) through the introduction of SNP dense regions. The introduction of multiple SNPs in this context results from a single event, rather than multiple independent events, as might be naively assumed if treating each SNP as independently arising from point mutations. It is therefore usually advisable in transmission studies is to identify these regions if relevant, and to exclude them from further consideration in the analysis[33]., However, this approach has been challenged more recently as it has the potential to artificially inflate the number of apparently closely related isolates through exclusion of hundreds or thousands of SNPs in the presence of multiple large recombinant sites[24]

2. Alternative Methods for transmission analysis

Several methods of pathogen genome comparison to understand transmission pre-date WGS . Early methods such as Pulse Field Gel Electrophoresis (PGFE) and Restriction Fragment Length Polymorphism (RFLP) analysis could be considered an early prototype of the "whole genome" characterisation approach for bacteria, albeit ones with much lower resolution[34,35]. These techniques depended on the isolation and manipulation of entire bacterial genomes, using restriction endonucleases to cut at specific restriction sites, numbering perhaps a few hundred scattered across a genome potentially containing millions of bases. Comparing the patterns of the resulting fragmented genomes is technically challenging, for instance requiring gel electrophoresis, visualisation of the resulting bands and construction of a distance matrix quantifying the inferred differences in the underlying genomes. The core principle is similar to WGS; if sufficient difference between final analysis outputs were observed (ultimately a reflection of the starting genome sequences) then isolates were so different that cross-transmission could effectively be ruled out. As the technique did not examine potential nucleotide changes in the majority of the sites of the genome significant genomic differences could still be present despite indistinguishable PFGE profiles. Scalability to large datasets was also a challenge that was met by setup of co-ordinating bodies such PulseNet International to standardise genotyping approaches and streamline isolate comparison between countries [36]. Other variations on the low-resolution genomic analysis include techniques such as microarrays that looked for several hundred well-characterised mutations in a bacterial sequence. While PFGE and microarrays have disappeared from clinical microbiology in recent years, another technique relevant to cross-transmission analysis is still extant: Sanger sequencing. This involves sequencing several hundred to a maximum of a few thousand nucleotide bases using primers targeting a defined region of interest in a bacterial or viral isolate allowing for comparison of nucleotide sequence of the targeted region in the same manner as discussed earlier for WGS. The targeted genomic region may be chosen for its relatively high mutability, if the primary interest is difference between pathogens such as determining the genotype of mumps virus by targeting the small hydrophobic (SH) gene. Conversely, a target with highly conserved nature, such as the 16S ribosomal RNA encoding region, may be preferred when identifying a bacterial species. In this case most mutations arising are highly deleterious and thus removed relatively quickly by natural selection.

Sanger sequencing is fundamentally limited by the underlying technology to delineating a fraction of any given pathogen genome, due to deterioration in obtained chromatogram sequence and therefore base calls after approximately 1000 nucleotide bases[37]. This limitation can be overcome to

an extent by sequencing multiple non-adjacent gene targets. For bacteria, this is the approach underlying the still widely used typing scheme called multi locus sequence typing (MLST)[38], as discussed in the article on *P. aeruginosa*. MLST leverages several regions to provide a higher level of detail then Sanger sequencing one target alone. The typing scheme allows comparison between different research groups. Yet it still cannot provide high enough resolution for an in depth transmission network analysis as it examines a maximum of several thousand bases for difference. Large *P. aeruginosa* strain collections from people in Ontario with CF have been extensively characterised by MLST. Fortunately, MLST is "backwards compatible" with whole genome sequencing technology – it is straightforward to obtain MLST sequence from a bacterial whole genome analysis with bioinformatic tools and assign an ST to the isolate using standardised nomenclature[39,40]. This approach, implemented in *P. aeruginosa* analysis, permitted some limited comparison of WGS sequences to isolates typed using historical methods.

An approach conceptually similar to MLST has been proposed for Adenovirus isolates to allow cross transmission analysis using Sanger sequencing of three gene targets[41]. This approach can enhance precision over traditional sequencing of single gene targets, when WGS is not available. A disadvantage of such multiple target Sanger sequencing is the significant effort required to identify suitable regions for sequencing, design primers optimised to these regions and need to account for ongoing genetic drift which may require primer updating. The substantial work required to develop such Sanger based typing approaches is quite pathogen dependant and not readily transferrable to different pathogens. In contrast, I demonstrate in later chapters WGS methodology (quality control of sequence data and the generation of phylogenetic trees) that I applied to multiple pathogens.

Regardless of whether a Sanger sequencing approach (which in practice is usually limited to a handful of gene targets) or a WGS approach is taken to pathogen typing there is a requirement for an actively maintained database and curated nomenclature to facilitate inter-laboratory comparison. Such a database permits allele calls to be generated for each region sequenced and is relatively agnostic to the technology used, though whole genome MLST (wg-MLST) permits the characterisation of orders of magnitude numbers of alleles per pathogen more than Sanger. The combination of a particular set of allele calls can then be assigned an ST.

3. WGS technology

There are two main approaches to WGS, short read and long read sequencing. Both are used in this work. The former was the "workhorse" technology used in whole genome sequencing in public health and reference laboratories in Ireland and in Canada during the period of this research. Short

read sequencing was used to sequence all the viral and bacterial genomes investigated in this work in academic and reference laboratories, specifically the genome sequencers produced by the Illumina company. Illumina markets a range of instruments that differ primarily in their throughput in terms of data generation from a single run, with more data acquisition associated with longer run times. They also differ in physical footprint and capital cost of the instruments. This flexibility in instrument form and function, along with the fact it is a long established technology with a mature ecosystem of associated analysis tools, means it has established a dominant market share in the field of laboratory medicine[42].

The main alternative to the dominant approach is long-read sequencing. The previous major vendor in this field was Pacific Biosciences, which was eclipsed in recent years by technology produced by Oxford Nanopore. Instruments produced by the latter have been commercially available since 2014 with widespread adoption in academic and reference laboratories more recently following improvements in base calling accuracy and bioinformatic tractibility[42] . The final article in this work describes use of Nanopore sequencing for three adenovirus genomes sequenced in the SickKids clinical microbiology lab. Like Illumina, its chief competitor in the microbial sequencing space, Oxford Nanopore offers instruments in a range of form factors. Some Nanopore offerings have a very small instrument footprint - the “MinION” and MinION Mk1C devices approximately the size of a mobile phone and tablet respectively. They are also significantly cheaper than the alternative devices (circa 1000 to 4000 euro for the Oxford Nanopore devices discussed above versus approximately 50,0000 to 100,000 for the low and medium throughput Illumina MiniSeq and MiSeq. Both instruments have relatively expensive consumables, but by multiplexing dozens of samples on a single flow cell and employing low cost protocols a total consumable cost in the range of 50 to 100 euro per bacterial genome is achievable[43,44] . The main drawback of long read sequencing as implemented by Oxford Nanopore is relative inaccuracy during base calling (the determination of nucleotide call at each position sequenced, which is expressed in terms of a quality score indicating likelihood of error for each), discussed in detail below.

Both approaches work on the basis of massively parallel sequencing which involves shearing the target genomic material of interest prior to sequencing, using for example physical methods (bead beating) or enzymatic action. The starting material may be DNA or RNA extracted from the pathogen itself after some form of enrichment step. This may involve simple culture of bacteria in selective solid or liquid media or cell culture for viruses followed by extraction of supernatant from cells showing cytopathic effect. As mentioned previously, it is possible to extract genomic material directly from primary clinical samples to obtain a “metagenome” consisting of nucleic acid from

human cells, commensal bacteria and pathogenic bacteria or viruses, in which case two different sequencing library preparations are required prior to loading of the instrument flow cell, one each for DNA and RNA. Since human DNA and (ribosomal) RNA are usually present in quantities orders of magnitude higher than micro-organisms from clinical sample, some sort of host nucleic acid depletion is advisable to avoid wasteful sequencing of off-target genome.

The sheared genomic fragments, numbering in the hundreds or thousands or millions, are then sequenced in parallel on the Illumina or Nanopore instrument. The sequencing chemistry differs between Illumina and Nanopore. Illumina operates using sequencing by synthesis, whereby labelled nucleotides are added in a complementary and stepwise fashion to single stranded DNA or complementary DNA (cDNA) fixed on a flow cell and amplified into clusters to aid identification in the next step. Then in successive rounds of sequencing by synthesis an in-built camera images each cluster as fluorescently tagged nucleotides are added to the complementary strands. Knowing the nucleotide added at each position allows the instrument to determine the original base present in the length of genome sequenced. This is a highly accurate technique with error rates on the order of <0.1% per base call using modern instruments[45]. It is however, limited by the underlying chemistry underpinning the additive reactions to sequencing stretches of genome (reads) ranging in size from 100 base pair (bp) to 600bp in length, hence the term “short read” sequencing. Through a process known as paired end sequencing, it is possible to sequence the extremities of a genomic fragment that may be several thousand bases long. The resulting “paired end reads” are a few hundred bps in length and are separated by hundreds or thousands of bps whose sequence is unknown. Despite the lack of sequencing of the middle of the fragment, the approach preserves data about the relative position and orientation of the paired reads with respect to each other. This is valuable information for downstream bioinformatic analysis as it can be used to overcome some limitations imposed by short read lengths that are particularly relevant to bacterial DNA. The limitations of a short read approach to bacterial sequencing arise from the biological structure of bacterial chromosomes and mobile genetic elements. In order to average out potential errors even in highly accurate short reads, it is desirable to have “overlap” of nucleotide calls at least 10 reads covering any position, as random errors in any position or read can be discounted by favouring majority consensus calls. With Illumina sequencing, sufficient overlap to eliminate the effect of errors in individual reads “adequate coverage” can be achieved for 90% of the genome of a sequenced bacteria. The reason the genome sequence for the remaining 10% cannot be established with accuracy with this approach is that this proportion of bacterial genomes contains structure called “repeat regions” which are each several thousands of bps in length. They are comprised mostly of multiple copies of genes encoding ribosomal RNA[46,47]. Repeat regions are very similar at the nucleotide level though are not necessarily

identical and they can contain SNPs introduced by random mutation that are informative for purposes of establishing pathogen nucleotide similarity. However, many repeat regions are too long to be spanned by individual reads, even with the paired-end approach discussed earlier. Therefore, reads from repeat region cannot be assigned with confidence to the parent repeat region; the lack of overlap at read ends with non-repeat region leads to low confidence in assignment. As well as representing a significant loss of potentially informative phylogenetic SNPs, this shortcoming means that we cannot be sure how different stretches of genome in the 90% of the accurately sequenced genome relate structurally to each other, since they are separated by repeat regions which we essentially cannot visualise. Bacterial chromosome draft assemblies reassembled from short read data alone thus consist of non-contiguous fragments thousands or hundreds of thousands of nucleotides long and of uncertain position with respect to the other fragments, unlike the contiguous, circularised input material. I describe in the article on *P. aeruginosa* methods for comparing hundreds of bacterial genomes despite these limitations using what was then a newly released bioinformatic tool called "mashtree"[65]. This software compared the dozens or hundreds of non-contiguous nucleotide fragments that each bacterial draft genome assembly consisted of using a minhash algorithm that was agnostic to the position of any SNPs in the genome. Indeed, the underlying computational approach could be used to quickly estimate the similarity of the information compared in any set of large text files, whether the text was nucleotide sequences or baking recipes. This allowed rapid screening of over a thousand assemblies for clusters of bacterial genomes that appeared to be closely related. These candidates were then subjected to more computationally intensive analysis that could efficiently analyse dozens of genomes to accurately identify SNP differences that were due to vertical inheritance (i.e. excluding recombinant regions).

Long read sequencing generates much longer fragments of genomic material. Single reads in excess of one million bp length are possible[48] although read length of several thousand bps is the norm and sufficient for most purposes in clinical microbiology. The input is typically either DNA or else complementary DNA (cDNA) which is generated as an initial step in workflows investigating RNA pathogens. However, in contrast to short read sequencing with Nanopore direct sequencing of RNA is possible[49] . An Oxford Nanopore instrument such as a MinION uses a flow cell containing several thousand pore forming proteins. Double stranded DNA fragments are pulled through the resulting hollow pores by enzymatic action. As they are fed through the pore the change in nucleotide sequence within results in changes in electrical current being passed along a conductive membrane in which thousands of pores are embedded which can be sensed and allows for the nucleotide sequence to be inferred. The primary limitation of this approach is a significantly higher base call error rate compared to Illumina, on the order of 6%[50]. Such errors in base calling are non-random,

being more likely to occur in stretches of homologous nucleotides. This is because several homologous nucleotides transit a pore together so that the accuracy of base calling of the number of nucleotides and thus the genomic sequence drops. This has implications for downstream bioinformatic reconstruction of such regions, particularly when assembling individual reads into draft bacterial genomes. It can introduce a characteristic error pattern in which downstream annotation of gene function artificially introduces hundreds of frameshift mutations, many of which if present in the underlying DNA would encode premature stop codons that block protein synthesis. Whilst nanopore sequencing alone cannot therefore produce a highly accurate and complete "closed" genomes consisting of the bacterial chromosome and accompanying mobile genetic elements like plasmids, it can provide data of sufficient quality to type bacterial organisms using traditional schemes such as MLST[51] .

Given the contrasting shortcomings of short-read and long-read sequencing for bacterial sequencing, a resource intense approach to overcome these is to sequencing a bacterial genome using both methods. This approach is called hybrid assembly. The long reads act as a "scaffold" that provide useful information on the large scale structure of the bacterial genome (by bridging repeat regions) and a first draft of the consensus genome sequence, which can then be "polished" by the more accurate short reads which can now be more confidently assigned to the correct repeat region. This approach is costlier in terms of laboratory workload and resources but does yield a complete "closed" circularised bacterial genome[52]. Also, if the bacterium carries extra-chromosomal DNA in the form of plasmids, a hybrid assembly approach is a suitable method for simultaneously closing the plasmid genome assemblies. Fully characterising any plasmids present in a bacterium via hybrid assembly may be of the first importance for transmission studies in cases where plasmid mobilised genes are relevant to infection prevention and control (e.g. some beta-lactamases). In these cases, the plasmids themselves may be transmitted to other bacteria and thus to new hosts rather than clonal spread of the pathogen from one patient to another. Tracking cross-transmission of virulence factors or antimicrobial resistance genes via detailed plasmid characterisation, in addition to evaluating transmission of an individual pathogen species or clone may be necessary. Plasmid borne genes are not a major consideration for *P. aeruginosa*. It has such a large and diverse genome that in-situ chromosomal mutation of genes or their promotor regions is sufficient to allow it adapt to most niches or to antimicrobial selection pressure[20], hence hybrid assembly method was not used for that study.

In contrast to bacterial sequencing limitations, reconstruction of the input genome for viral pathogens is possible using either the Illumina or Nanopore approach proceeding from clinical

samples either by direct metagenomic or enrichment approach. It is possible to leverage the high depth of coverage achieved to overcome limitations imposed by errors in individual reads. Since viral genomes are orders of magnitude smaller than most bacterial genomes, it is possible to obtain for each nucleotide position hundreds of reads that overlap this position and therefore provide information about it. This can be done even if sequencing up to 96 viral genomes on a flow cell[53], where allocating similar flow cell capacity for data (base call) generation for larger bacterial genomes would provide dozens of reads at any given position and reduce confidence in the accuracy of final consensus genome. Thus, the error rates for genome sequencing of viruses can be averaged out more easily than for bacteria, leading to higher confidence in the final reconstructed consensus genome and permitting direct and accurate comparison of viral genomes at the nucleotide level regardless of approach.

The three peer-reviewed published journal articles which make up the body of this thesis constitute analyses undertaken while I was a Clinical and Research Microbiology Fellow in Toronto, Canada from 2016 to 2019. While the use of WGS to investigate pathogen transmission per se is well established [54–58], each article focuses on a specific pathogen/epidemiological context transmission dyad that had not previously been explored in detail using WGS as the key analysis tool. In consequence, basic parameters such as what constitutes "closely related" or a case "cluster" for each pathogen had to be established and justified using a combination of *a priori* reasoning and deductions from the sequencing data, rather than by reference to existing published WGS studies as would be possible for certain key public health pathogens where these parameters are well described e.g. *Staphylococcoccus aureus* and *Clostridioides difficile* [59,60]. The articles are presented in chronological order and concern:

- The first published description of a community Mumps virus outbreak interrogated using WGS
- The first description of *Pseudomonas aeruginosa* cross transmission causing early infection of children with Cystic Fibrosis using WGS methodology to explore the important role played by intra-host pathogen diversity via multiple colony sequencing
- The first multi-institution investigation of human Adenovirus A31 (hAdv-A31) transmission in children undergoing haematopoeitic stem cell transplantation using WGS.

The initial WGS methodology was developed from first principles as described above, since no previous publications on this topic were available at the time of study inception. A separate group in the UK published the first WGS based analyses of hAdv-A31 transmission based on their own single

centre experience and methods while analysis for this final article was in progress. Their results and published genome sequences profoundly altered the conclusions of our own outbreak investigation. In this work, in addition to investigating the application of WGS varied infection control contexts, I also detail in the methods (within the main text of each article and in the more detailed accompanying supplementary material) transferable approaches and concepts required to ensure reliable, repeatable analysis that can be performed for a range of respiratory pathogens. When I undertook this research, such use of WGS in Ireland was primarily undertaken in individual reference laboratories with remits for specific pathogens where "laboratory functions were appointed or established based on personal initiatives…without a formal selection process…inhibiting long-term planning" and with capacity gaps for some healthcare-associated pathogens[61]. The research below presents an approach that can be applied to a range of pathogens and could help address capacity gaps.

The research question

Does use of WGS in investigating suspected cross-transmission of varied respiratory pathogens in different public health and hospital epidemiology contexts add value in terms of understanding how transmission occurs, aid development of control and mitigation strategies and inform future research efforts?

Research objectives

1. For a large community outbreak of Mumps virus that occurred in South East Ontario (Toronto and surrounding regions), to use WGS to determine its origin and its utility in aiding the response of Public Health Ontario (PHO). I aimed to use WGS to investigate a Mumps outbreak and to determine if WGS is superior for elucidating transmission networks compared to existing methods of typing based on Sanger sequencing of a portion of the SH gene. Secondary objectives were:
    a. Develop a "wet lab" method for performing WGS directly from the background human genomic material present in much greater abundance in clinical specimens. The intention was to accomplish this by iteratively refining a "tiling amplicon" based approach developed in PHO Laboratory. The tiling amplicon method involves PCR

amplification of many short fragments of the Mumps virus genome, with the ends of each amplicon overlapping slightly with its neighbours like tiles arranged in an offset pattern. This ensures no gaps in coverage when the subsequently sequenced products are bioinformatically reassembled to make a contiguous genome. Previously this approach had been deployed for individual Mumps genomes to enable Sanger sequencing from virus culture supernatant. Our aim was to develop a WGS method that would be independent of expensive and slow virus culture.

b. Integrate WGS data with epidemiological information from outbreak investigation and infection transmission dynamics (incubation period, infectious period and generation time). The latter was achieved using two tools, the first of which was GenGIS[66]. This allowed for a phylogenetic tree of the outbreak to be divided into several clades (each containing closely related sequences) and each sequence within the clade mapped to the case location in Ontario. The purpose of this was to provide a readily accessible visualisation for public health specialists who did not have a background in genomics and did not have experience interpreting traditional phylogenetic tree layouts. The second tool used was Transphylo, which produced an output in the style of transmission map that provides a visual estimate of the direction and timing of disease transmission and the presence of likely unsampled intermediary cases. This again represented a visualisation for the benefit of public health specialists.

c. Incorporate contemporaneous Mumps whole genome sequences from outbreaks in the USA to understand role of international transmission in outbreak generation. This was a post-hoc analysis initiated after our project commenced, prompted by the publication of genome sequences from USA in online repositories such as THE National Centre for Biotechnology Information (NCBI) Genbank, academic laboratory websites and as appendixes to pre-print manuscripts. It included analysis using Nextstrain, a tool which permitted analysis of when the Ontario outbreak clade sequences coalesced with USA sequences. This incorporated both sequence and geospatial data (place and date of sample collection) to provide an estimate (with confidence intervals) of when the clades last shared a recent common ancestor. This is known as time to most recent common ancestor (tMRCA) and calculating it provides important information in outbreaks of infectious disease as it indicates the earliest time that the Ontario virus clades were potentially imported.

2. For early *P. aeruginosa* infection in children with CF, to determine if cryptic person to person transmission occurs in early infection in an non-outbreak setting by sequencing at large scale (>1000 isolates) bacterial isolates cultured from early CF infections. The overarching context in CF care at the time was that patient to patient transmission was considered primarily a feature of late infection and mostly related to particular "epidemic strains"[63]. In contrast, early infection was thought due to acquisition of very diverse strains from the general environment. The background to the study was the availability of a comprehensive sputum biobank from a cohort of children attending SickKids for CF care and incomplete biobanking of some bacterial isolates from early infections. Inter-patient transmission was not highly suspected on clinical or microbiological grounds (e.g. shared phenotypes such as distinctive patterns of antimicrobial susceptibility). Bacterial isolates from early infection growing in solid agar culture did however exhibit a diverse range of morphological characteristics. Some "morphotypes" isolated from children with early infection had distinctive appearances such as mucoidy which are typically seen in late infection as a result of adaptation of the CF airways[64]. This prompted the question could the appearance of such morphotypes in early infection represent acquisition of strains from persons with late infection. Early *P. aeruginosa* infection is a significant event for children with CF, as it can progress to chronic infection which is associated with reduced lung function and poor outcomes. To better understand the role of transmission in early infection it was necessary to sequence not only early infection bacterial isolates, but also late (chronic CF infection) isolates from older children, isolates from non-CF patients e.g. from bloodstream infections, and any bacteria recovered from the hospital environment. My secondary objectives were:
    a. culture bacterial isolates from the biobanked respiratory samples and sequence up to a dozen isolates per sample
    b. identify where cross-transmission may be occurring by linking sequencing analysis to epidemiological investigations focused on overlapping inpatient stays, clinic attendance and community links
    c. develop a bespoke bioinformatic workflows to efficiently analyse > 1000 bacterial sequences. A traditional SNP based phylogenetic analysis involving whole genome data was not practical. The final stage of this workflow therefore used the tool mashtree, which allowed for the draft genome assemblies created by standard bioinformatic tools in the early pipeline steps to be compared to each other and reference strains in a computationally efficient manner.

d. detect the presence of hypermutator genotype sequences, where random mutation of genes involved in DNA repair produce strains with significantly higher mutation rate. Development of hypermutator genotype is relatively common in chronic *P. aeruginosa* CF infection and the SNP distance of the hypermutator from recent ancestors can become rapidly inflated. Consequently, instead of using SNP distance as a metric of relatedness it is necessary to examine the phylogenetic tree topology for evidence of *paraphyletic* relationships; a long branch leading to a hypermutator genotype isolate, but arising within a clade of isolates obtained from other host indicates a likely transmission event.
e. to identify if there were genotypic differences in isolates of the same morphotype from a single sample and to identify how such intra-patient genotypic diversity correlates with the inter-patient transmission
f. determine if cross-transmission was associated with adverse outcomes for patients

3. For Adenovirus infections of children receiving haematopoietic stem cell transplantation (HSCT) at SicKids, the objective was to investigate a hospital outbreak of adenovirus type A31 (hAdV-A31) persisting over several years. Adenovirus infection in this population may occur due to either reactivation of endogenous infection or cross-transmission, and those due to hAdV-A31 often become disseminated. This has important clinical consequences for both stem cell transplant recipients and for the HSCT unit. hAdV-A31 infection necessitates strict patient isolation for extended periods and may require treatment with expensive and potentially toxic antiviral agents such as cidofovir or brincidofovir, with associated mortality risk. For the HSCT unit, evidence of sustained cross-transmission may prompt consideration of unit closure for a period to new transplant recipients, if the risk of virus acquisition and mortality is deemed too high.

   This outbreak investigation was accomplished with the aid of WGS of virus culture isolates and also metagenomic sequencing i.e. direct sequencing of primary clinical specimen, including viral and human genome sequences, without prior enrichment methods such as culture or tiling amplicon. These post-transplant infections imposed a high burden of morbidity on critically ill children and were disruptive to the functioning of the transplant unit. A high incidence of post-transplant infections with hAdV-A31 was identified and an outbreak declared in 2015, but infections occurred intermittently for years despite the introduction of control measures. WGS of the virus genomes was arranged to explore

potential cross-transmission. At the time we commenced the investigation there were no existing studies of infections that reported using WGS for investigating hAdV-A31 outbreaks and there were no contemporaneous whole genome sequences available for comparison. Additional virus isolates were obtained from four other paediatric HSCT centres for sequencing as comparison to external sequences would assist with interpreting the sequencing diversity found within a single centre. As mentioned in the introduction, a group in the UK published single centre experience of paediatric HSCT adenovirus transmission during the later outbreak period. The newly available UK genome sequences were incorporated into the final multi-centre analysis and significantly informed the results. My secondary objectives were:

a. describe the clinical and laboratory features of the outbreak
b. describe the use of whole-genome sequencing in delineating local outbreak cases
c. explore the relationship of SickKids virus hAdV-A31 genomes to international strains
d. propose strategies for optimal monitoring of such patients in paediatric facilities including the role of WGS in suspected outbreaks

REFERENCES

1. Kwong JC, Mccallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. doi:10.1097/PAT.0000000000000235
2. Stimson J, Gardy JL, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *bioRxiv*. Published online May 10, 2018:319707. doi:10.1101/319707
3. Perez-Sepulveda BM, Heavens D, Pulford C v., et al. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biol*. 2021;22(1). doi:10.1186/s13059-021-02536-3
4. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: Entering the Age of Next-Generation Sequencing. *Foodborne Pathog Dis*. 2019;16(7):451-456. doi:10.1089/fpd.2019.2634
5. Kubota KA, Wolfgang WJ, Baker DJ, et al. PulseNet and the Changing Paradigm of Laboratory-Based Surveillance for Foodborne Diseases. *Public Health Reports*. 2019;134(2_suppl):22S-28S. doi:10.1177/0033354919881650
6. Gibbs RA. The Human Genome Project changed everything. *Nature Reviews Genetics 2020 21:10*. 2020;21(10):575-576. doi:10.1038/s41576-020-0275-3
7. Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. *Genome Med*. 2013;5(9):1-8. doi:10.1186/GM483/METRICS
8. Saunders G, Baudis M, Becker R, et al. European infrastructures to access one million human genomes by 2022. *Nat Rev Genet*. 2019;20(11):693. doi:10.1038/S41576-019-0156-9

9. The Cost of Sequencing a Human Genome. Accessed January 13, 2023. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

10. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols 2013 8:11*. 2013;8(11):2281-2308. doi:10.1038/nprot.2013.143

11. Karavolias NG, Horner W, Abugu MN, Evanega SN. Application of Gene Editing for Climate Change in Agriculture. *Front Sustain Food Syst*. 2021;5:296. doi:10.3389/FSUFS.2021.685801/BIBTEX

12. Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: Where are we now? *Proc Natl Acad Sci U S A*. 2021;118(52):e2109019118. doi:10.1073/PNAS.2109019118/SUPPL_FILE/PNAS.2109019118.SD01.XLSX

13. Soltis PS, Soltis DE. Plant genomes: Markers of evolutionary history and drivers of evolutionary change. *Plants, People, Planet*. 2021;3(1):74-82. doi:10.1002/PPP3.10159

14. Prüfer K, Posth C, Yu H, et al. A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. *Nat Ecol Evol*. 2021;5(6):820-825. doi:10.1038/S41559-021-01443-X

15. Suárez NM, Blyth E, Li K, et al. Whole-Genome Approach to Assessing Human Cytomegalovirus Dynamics in Transplant Patients Undergoing Antiviral Therapy. *Front Cell Infect Microbiol*. 2020;10. doi:10.3389/FCIMB.2020.00267

16. Ferreira I, Beisken S, Lueftinger L, et al. Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: An analysis of isolates from the unyvero lower respiratory tract infection trial. *J Clin Microbiol*. 2020;58(7). doi:10.1128/JCM.00273-20/SUPPL_FILE/JCM.00273-20-S0001.PDF

17. Talavera-López C, Andersson B. Parasite genomics—Time to think bigger. *PLoS Negl Trop Dis*. 2017;11(4). doi:10.1371/JOURNAL.PNTD.0005463

18. Dimaio D. Viruses, Masters at Downsizing. *Cell Host Microbe*. 2012;11(6):560-561. doi:10.1016/J.CHOM.2012.05.004

19. Alcock BP, Huynh W, Chalil R, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res*. 2023;51(D1). doi:10.1093/NAR/GKAC920

20. Valot B, Guyeux C, Rolland JY, Mazouzi K, Bertrand X, Hocquet D. What It Takes to Be a Pseudomonas aeruginosa? The Core Genome of the Opportunistic Pathogen Updated. *PLoS One*. 2015;10(5):e0126468. doi:10.1371/journal.pone.0126468

21. Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. *J Mol Evol*. 2002;54(2):156-165. doi:10.1007/s00239-001-0064-3

22. Chevallereau A, Meaden S, van Houte S, Westra ER, Rollie C. The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. *Philos Trans R Soc Lond B Biol Sci*. 2019;374(1772). doi:10.1098/RSTB.2018.0094

23. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. *PLoS Genet*. 2008;4(12):1000304. doi:10.1371/JOURNAL.PGEN.1000304

24. Gorrie CL, da Silva AG, Ingle DJ, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. *Lancet Microbe*. 2021;2(11):e575-e583. doi:10.1016/S2666-5247(21)00149-X

25. Beaty SM, Lee B. Constraints on the Genetic and Antigenic Variability of Measles Virus. *Viruses*. 2016;8(4). doi:10.3390/V8040109

26. Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of Mycobacterium tuberculosis. *PLoS Pathog*. 2019;15(9). doi:10.1371/JOURNAL.PPAT.1008067

27. Hamilton WL, Tonkin-Hill G, Smith ER, et al. Genomic epidemiology of COVID-19 in care homes in the east of England. *Elife*. 2021;10. doi:10.7554/ELIFE.64618

28. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res*. 2018;3:33. doi:10.12688/wellcomeopenres.14265.1

29. Zhou X, Shen XX, Hittinger CT, Rokas A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol Biol Evol*. Published online November 21, 2017. doi:10.1093/molbev/msx302

30. Freschi L, Bertelli C, Jeukens J, et al. Genomic characterisation of an international Pseudomonas aeruginosa reference panel indicates that the two major groups draw upon distinct mobile gene pools. *FEMS Microbiol Lett*. Published online June 12, 2018. doi:10.1093/femsle/fny120

31. Turakhia Y, Thornlow B, Hinrichs A, et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature 2022 609:7929*. 2022;609(7929):994-997. doi:10.1038/s41586-022-05189-9

32. Darch SE, Mcnally A, Harrison F, et al. Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. doi:10.1038/srep07649

33. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015;11(2). doi:10.1371/JOURNAL.PCBI.1004041

34. de Boer AS, Kremer K, Borgdorff MW, de Haas PEW, Heersma HF, van Soolingen D. Genetic Heterogeneity in Mycobacterium tuberculosis Isolates Reflected in IS6110 Restriction Fragment Length Polymorphism Patterns as Low-Intensity Bands. *J Clin Microbiol*. 2000;38(12):4478. doi:10.1128/JCM.38.12.4478-4484.2000

35. Simar SR, Hanson BM, Arias CA. Techniques in bacterial strain typing: past, present, and future. *Curr Opin Infect Dis*. 2021;34(4):339. doi:10.1097/QCO.0000000000000743

36. Nadon C, van Walle I, Gerner-Smidt P, et al. Pulsenet international: Vision for the implementation of whole genome sequencing (WGS) for global foodborne disease surveillance. *Eurosurveillance*. 2017;22(23). doi:10.2807/1560-7917.ES.2017.22.23.30544

37. Manyana S, Gounder L, Pillay M, Manasa J, Naidoo K, Chimukangara B. HIV-1 Drug Resistance Genotyping in Resource Limited Settings: Current and Future Perspectives in Sequencing Technologies. *Viruses*. 2021;13(6). doi:10.3390/V13061125

38. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3. doi:10.12688/WELLCOMEOPENRES.14826.1

39. Feijao P, Yao HT, Fornika D, et al. MentaLiST – A fast MLST caller for large MLST schemes. *Microb Genom*. 2018;4(2). doi:10.1099/MGEN.0.000146

40. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom*. 2017;3(10). doi:10.1099/mgen.0.000131

41. Kaján GL, Lipiec A, Bartha D, Allard A, Arnberg N. A multigene typing system for human adenoviruses reveals a new genotype in a collection of Swedish clinical isolates. *PLoS One*. 2018;13(12). doi:10.1371/JOURNAL.PONE.0209038

42. Zhong Y, Xu F, Wu J, Schubert J, Li MM. Application of Next Generation Sequencing in Laboratory Medicine. *Ann Lab Med*. 2021;41(1):25. doi:10.3343/ALM.2021.41.1.25

43. Butt SL, Taylor TL, Volkening JD, et al. Rapid virulence prediction and identification of Newcastle disease virus genotypes using third-generation sequencing. *Virol J*. 2018;15(1). doi:10.1186/S12985-018-1077-5

44. Alvarez Narvaez S, Shen Z, Yan L, et al. Optimized conditions for Listeria, Salmonella and Escherichia whole genome sequencing using the Illumina iSeq100 platform with point-and-click bioinformatic analysis. *PLoS One*. 2022;17(11):e0277659. doi:10.1371/JOURNAL.PONE.0277659

45. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform*. 2021;3(1). doi:10.1093/NARGAB/LQAB019

46. Mavromatis K, Land ML, Brettin TS, et al. The Fast Changing Landscape of Sequencing Technologies and Their Impact on Microbial Genome Assemblies and Annotation. *PLoS One*. 2012;7(12). doi:10.1371/JOURNAL.PONE.0048837

47. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2011;8(1):61. doi:10.1038/NMETH.1527

48. Ma Z (Sam), Li L, Ye C, Peng M, Zhang YP. Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. *Genomics*. 2019;111(6):1896-1901. doi:10.1016/J.YGENO.2018.12.013

49. Jain M, Abu-Shumays R, Olsen HE, Akeson M. Advances in nanopore direct RNA sequencing. *Nat Methods*. 2022;19(10):1160-1164. doi:10.1038/S41592-022-01633-W

50. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. *PLoS One*. 2021;16(10). doi:10.1371/JOURNAL.PONE.0257521

51. Liou CH, Wu HC, Liao YC, Lauderdale TLY, Huang IW, Chen FJ. nanoMLST: accurate multilocus sequence typing using Oxford Nanopore Technologies MinION with a dual-barcode approach to multiplex large numbers of samples. *Microb Genom*. 2020;6(3):1-8. doi:10.1099/MGEN.0.000336

52. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics*. 2020;21(1). doi:10.1186/S12864-020-07041-8

53. Liu H, Li J, Lin Y, et al. Assessment of two-pool multiplex long-amplicon nanopore sequencing of SARS-CoV-2. *J Med Virol*. 2022;94(1):327. doi:10.1002/JMV.27336

54. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. *Curr Opin Microbiol*. Published online 2015. doi:10.1016/j.mib.2014.11.004

55. Hall MD, Woolhouse MEJ, Rambaut A. Using genomics data to reconstruct transmission trees during disease outbreaks. *Rev Sci Tech Off Int Epiz*. 2016;35(1):287-296. doi:10.20506/rst.35.1.2433

56. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (1979)*. 2014;345(6202):1369-1372. doi:10.1126/SCIENCE.1259657

57. Török ME, Reuter S, Bryant J, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *J Clin Microbiol*. 2013;51(2):611-614. doi:10.1128/JCM.02279-12

58. Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLoS Currents Outbreaks*. 2014;1. doi:10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d

59. Golubchik T, Batty EM, Miller RR, et al. Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. *PLoS One*. 2013;8(5). doi:10.1371/JOURNAL.PONE.0061319

60. Eyre DW, Cule ML, Wilson DJ, et al. Diverse Sources of C. difficile Infection Identified on Whole-Genome Sequencing. *N Engl J Med*. 2013;369(13):1195-1205. doi:10.1056/NEJMOA1216064

61. *Country Visit to Ireland to Discuss Policies Relating to Antimicrobial Resistance*.; 2019. Accessed January 13, 2023. https://www.ecdc.europa.eu/en/publications-data/country-visit-ireland-discuss-policies-relating-antimicrobial-resistance

62. ECDC supports EU/EEA Member States in rapid detection of SARS-CoV-2 variants. Accessed January 13, 2023. https://www.ecdc.europa.eu/en/news-events/ecdc-supports-eueea-member-states-rapid-detection-sars-cov-2-variants

63. Panagea S, Winstanley C, Walshaw MJ, Ledson MJ, Hart CA. Environmental contamination with an epidemic strain of Pseudomonas aeruginosa in a Liverpool cystic fibrosis centre, and study of its survival on dry surfaces. *Journal of Hospital Infection*. Published online 2005. doi:10.1016/j.jhin.2004.09.018

64. Clark ST, Caballero JD, Cheang M, et al. Phenotypic diversity within a Pseudomonas aeruginosa population infecting an adult with cystic fibrosis. *Sci Rep*. 2015;5:10932. doi:10.1038/SREP10932

65. Katz L, Griswold, T, Carleton HA. Generating WGS Trees with Mashtree. Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines. Washington, DC. ASM, 2017.

66. Parks, D. H. et al. GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. PLoS One 8, e69885 (2013).

Chapter 1

Evaluating the use of whole genome sequencing for the investigation of a large mumps outbreak in Ontario, Canada.

Sci Rep. 2019 Aug 30;9(1):12615. doi: 10.1038/s41598-019-47740-1. PMID: 31471545; PMCID: PMC6717193. Impact Factor of Scientific Reports in 2019: 4.16

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6717193/

Contribution of author to manuscript: design of study, amplicon generation and sequencing library preparation, bioinformatics analysis, drafting of manuscript, data sharing via Genbank.

AUTHORS AND AFFILIATIONS

Patrick J Stapleton [1 2], AliReza Eshaghi [2], Chi Yon Seo [3], Sarah Wilson [3 4], Tara Harris [3], Shelley L Deeks [3 4], Shelly Bolotin [1 2 4 5], Lee W Goneau [1 2], Jonathan B Gubbay [1 2], Samir N Patel [6 7] *

[1] Laboratory Medicine & Pathobiology, University of Toronto, Toronto, ON, Canada.

[2] Public Health Ontario Laboratory, Public Health Ontario, Toronto, ON, Canada.

[3] Communicable Diseases, Emergency Preparedness and Response, Public Health Ontario, Toronto, ON, Canada.

[4] Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada.

[5] Applied Immunisation Research and Evaluation, Public Health Ontario, Toronto, ON, Canada.

[6] Laboratory Medicine & Pathobiology, University of Toronto, Toronto, ON, Canada. samir.patel@oahpp.ca.

[7] Public Health Ontario Laboratory, Public Health Ontario, Toronto, ON, Canada. samir.patel@oahpp.ca.

*Corresponding author

ABSTRACT

In 2017 Ontario experienced the largest mumps outbreak in the province in 8 years, at a time when multiple outbreaks were occurring across North America. Of 259 reported cases, 143 occurred in Toronto, primarily among young adults. Routine genotyping of the small hydrophobic gene indicated that the outbreak was due to mumps virus genotype G. We performed a retrospective study of whole genome sequencing of 26 mumps virus isolates from early in the outbreak, using a tiling amplicon method. Results indicated that two of the cases were genetically divergent, with the remaining 24 cases belonging to two major clades and one minor clade. Phylogeographic analysis confirmed circulation of virus from each clade between Toronto and other regions in Ontario. Comparison with other genotype G strains from North America suggested that the presence of co-circulating major clades may have been due to separate importation events from outbreaks in the United States. A transmission network analysis performed with the software program *TransPhylo* was compared with previously collected epidemiological data. The transmission tree correlated with known epidemiological links between nine patients and identified new potential clusters with no known epidemiological links.

INTRODUCTION

Mumps virus is a single-stranded negative-sense RNA virus in the Rubulavirus genus of the *Paramyxoviridae* family, with a 15.3 kilobase (kb) genome that encodes 8 proteins. It is highly contagious and causes outbreaks of respiratory illness. Disease is normally self-limiting, but can be complicated by meningitis, encephalitis, orchitis or oopheritis[1]. While the incidence of mumps in high income countries has declined dramatically from the 1970s and onwards following the introduction of effective live-attenuated vaccines (https://www.canada.ca/en/public-health/services/immunization/vaccine-preventable-diseases/mumps/health-professionals.html), the past decade has seen a relative resurgence of mumps activity in some high income countries[2,3,4]. This increase has been attributed to waning of immunity in young adults who were immunized with one or two doses of measles, mumps and rubella (MMR) vaccine[5,6,7].

In Ontario, mumps is a reportable disease. All cases are investigated, and the detection of a cluster of cases prompts outbreak investigation and control measures by public health authorities. Mumps outbreak investigations are labour intensive and require public health professionals to interview cases and identify transmission networks and potential common exposures in public settings;

commonly these include post-secondary education settings, social gatherings or sporting events. Identifying links between individual cases is often challenging, as up to 40% of individuals with mumps are either asymptomatic, or present with primarily respiratory symptoms, and therefore lack the classic clinical presentation of parotitis[1].

Epidemiological investigations can be complemented by molecular genotyping studies, which can help confirm or refute potential transmission events by comparing strain relatedness. The most widely used genotyping method for mumps virus involves sequencing a 316 nucleotide region of the small hydrophobic (SH) gene. This is usually the most variable region of the mumps genome and encodes a membrane associated protein whose function is incompletely understood. There are 12 distinct mumps genotypes, which are distributed globally. Most outbreaks in North America in recent years have been caused by Genotype G[8]. Genotyping using the SH gene is of limited utility during genotype G outbreaks, as the most variable region of this genotype is not the SH gene, but rather has been reported to be in non-coding regions of the genome[8,9].

Whole genome sequencing (WGS) provides the ultimate resolution for genomic epidemiological investigations, by identifying single nucleotide variants (SNVs) between isolates. Even over the relatively short timeframe of a typical mumps outbreak (i.e. a few months), genome substitutions in RNA viruses are likely to arise frequently enough to allow sufficient discrimination of distinct lineages within the outbreak, and even individual transmission events. This approach has been described for outbreaks of other paramyxoviruses[10]. However, WGS of mumps was until recently performed infrequently, with only 110 full genome sequences available in the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/index.html) in July 2017 when this study commenced, compared with approximately 500 Zika virus genomes and over 2000 Zaire ebolavirus genomes.

In 2017, Ontario experienced a mumps virus outbreak which was the largest in the province since 2008. There were 259 cases reported (https://www.publichealthontario.ca/en/DataAndAnalytics/Pages/RDTO2016.aspx), 143 of which (55%) occurred in Toronto. Immunisation status was known for 155 Ontario cases (60%); of these 67 (43%) had received 2 or more doses of MMR vaccine. It was unclear from initial epidemiological investigations if the cases outside Toronto were part of the same outbreak, or represented a separate provincial cluster, potentially due to importation of cases from simultaneous outbreaks occurring elsewhere in North America. The most frequent common exposure for the Toronto cases was attendance at downtown bars (n = 70, 49%), and only 22 (15%) were linked with education

settings[11]. This is in contrast to prior Ontario outbreaks, which were associated with secondary or post-secondary education settings[12], and it therefore presented unique challenges for outbreak control efforts. Routine case finding and outbreak control measures were expanded to include targeted bar inspections to reinforce good infection prevention and control practices, and a social media campaign targeting young adults[11].

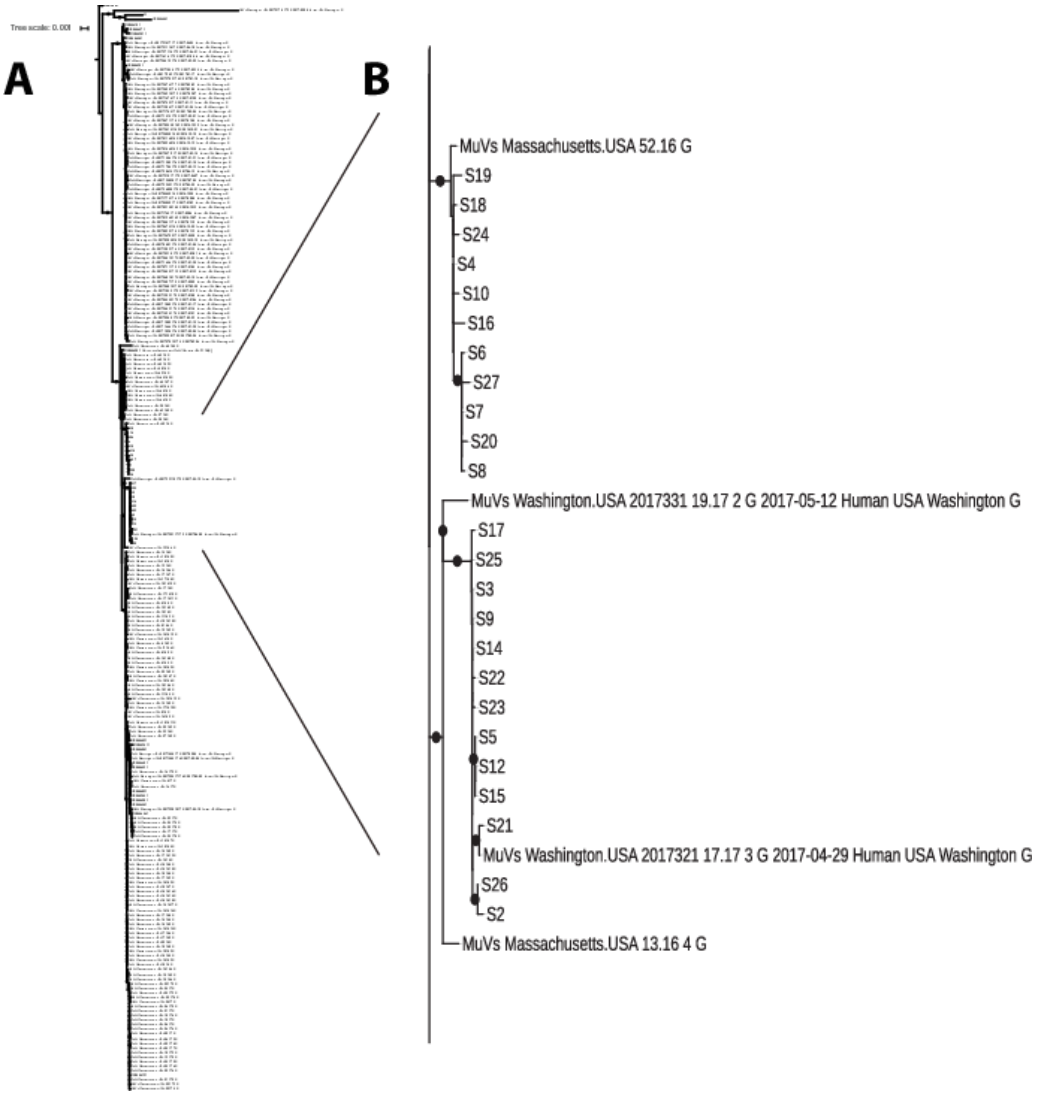
Routine SH genotyping of isolates from 203 PCR confirmed cases indicated that 194 (96%) were genotype G. The limited resolution of SH genotyping could not help resolve transmission networks. We performed a retrospective study using WGS of a convenience sample of virus isolates from 27 cases from the first three months of the outbreak, 17 of them (63%) from Toronto. Our aims were to determine if the results of WGS and transmission network analysis correlated with epidemiological data, and to evaluate the feasibility and desirability of using this approach prospectively in future outbreaks.

RESULTS

Phylogenetic analysis

Amplicon based WGS was successful in 26 cases (96%). All 26 samples had average nucleotide coverage per site of at least 500X. Initial phylogenetic analysis revealed that isolates from patient 1 (S1) and patient 11 (S11) were genomically distinct from the main outbreak clade (Fig. 1).

FIGURE 1



Phylogenetic tree created using whole genome alignment of 26 outbreak isolates and selected reference strains from NCBI GenBank. Text outside the tree indicates major phylogenetic groups. The main 2017 Ontario outbreak clade of 24 isolates is collapsed into a pyramid at the base of the tree. Two genomically distinct outliers from the outbreak, S1 (genotype G) and S11 (genotype C) are marked with red dots. The tree was created using the maximum likelihood method with iqtree v1.6.2 using the GTR + G model and 1000 ultrafast bootstrap approximations Black circles indicate nodes with >90% ultrafast bootstrap approximation support.

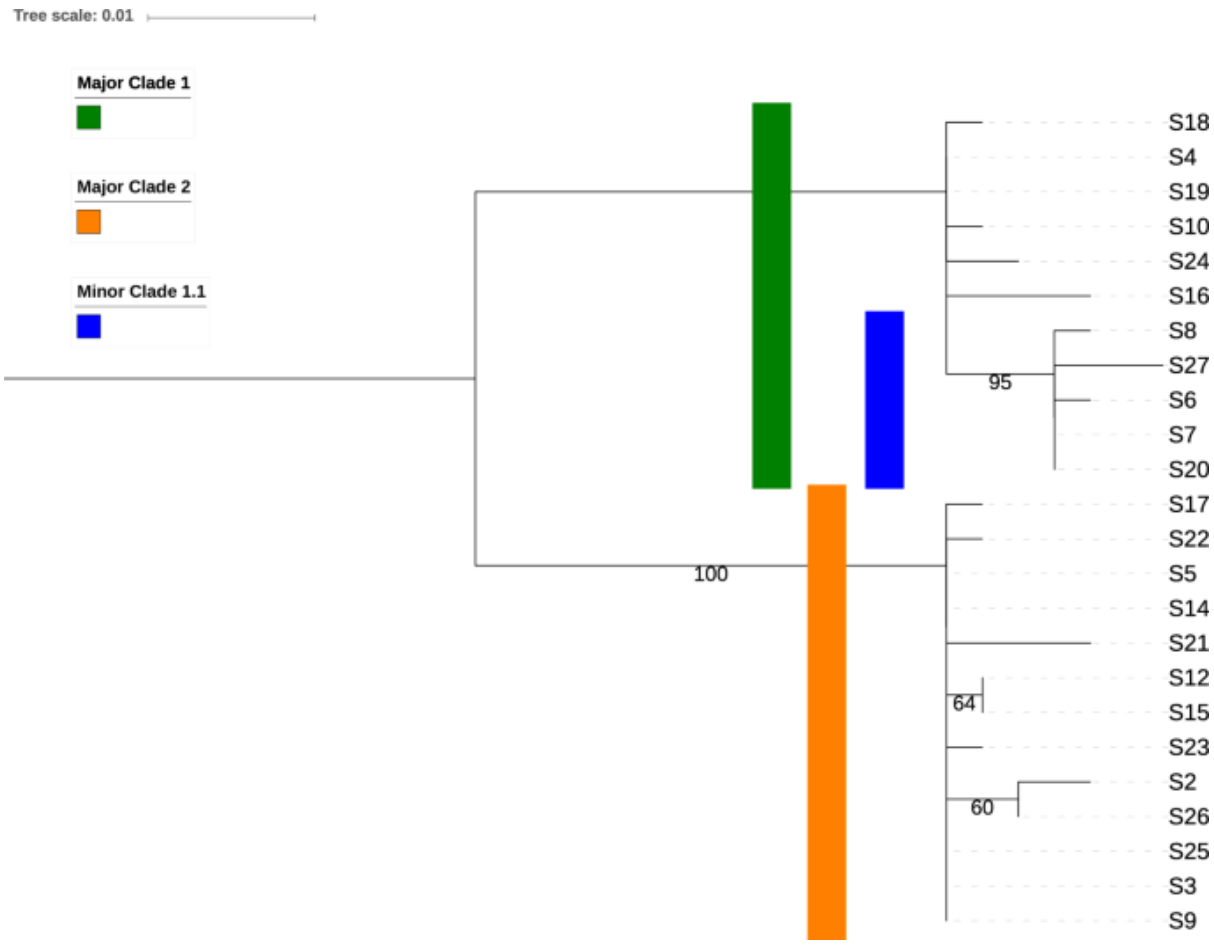
Traditional SH genotyping differentiated S11, which is genotype C, from the main outbreak, but did not identify that case S1 was distinct from the other Genotype G strains and was therefore likely the result of an independent introduction of mumps virus. From the WGS phylogenetic tree it is apparent that S1 is more closely related to a sample we previously sequenced from a 2010 Ontario outbreak than it is to the other outbreak strains[13].

Outbreak specific SNVs were identified by mapping reads of 23 isolates from the main outbreak clade against an in-group reference (S7). All samples had >99.9% site coverage compared to the

reference assembly length of 15285 nucleotides. We identified 51 variable sites specific to the outbreak (Supplementary Dataset 1). All outbreak mutations were due to SNVs; there were no complex variants or insertions/deletions. The RNA-directed RNA polymerase L gene had the most variable sites (n = 23, 45%). There were only four variable sites in intergenic regions. A minority of SNVs (n = 14, 27%) were missense mutations that resulted in amino acid substitutions.

Maximum likelihood (ML) tree analysis of the outbreak strains revealed two major clade and one minor clade (or cluster) with bootstrap support values >0.9 (Fig. 2).

FIGURE 2

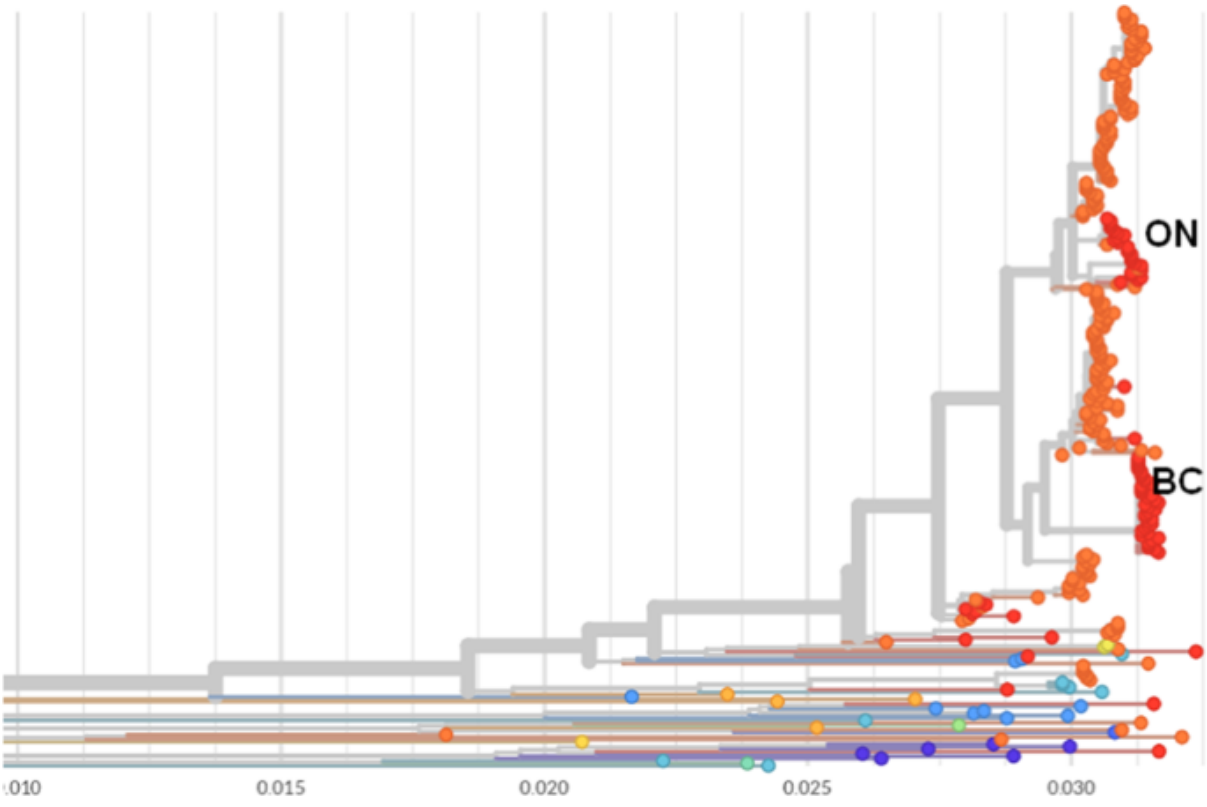


Phylogenetic tree of the outbreak clades. The two major and one minor clade are indicated with colored bars. The maximum likelihood tree was created by mapping reads to in-group reference (an S7 de-novo assembly). The tree is midpoint rooted and was constructed with iqtree using the GTR model with correction for ascertainment bias and 1000 traditional bootstrap replicates. Nodes with >50% bootstrap support are annotated with the support value.

Interestingly, the minor clade consisted of five isolates that shared a common missense mutation in the SH gene (c.158C > T p.Ser53Phe), meaning that traditional SH genotyping can discriminate the minor clade from other samples, but does not identify the 2 major clades. The only other SNV in the SH gene (c.41T > C p.Ile14Thr, S2) was a phylogenetically uninformative singleton mutation. Overall, ML analysis helped us to identify clusters, but did not allow us to identify probable transmission events.

Comparison of our strains with genotype G mumps virus circulating in North America was facilitated by the Nextstrain project (https://nextstrain.org/mumps/global), which includes sequences that are not publically available in GenBank. This analysis indicated our outbreak strains were more closely related to strains in the United States (US) than to an outbreak in British Columbia that occurred in summer 2016 (Fig. 3).

FIGURE 3

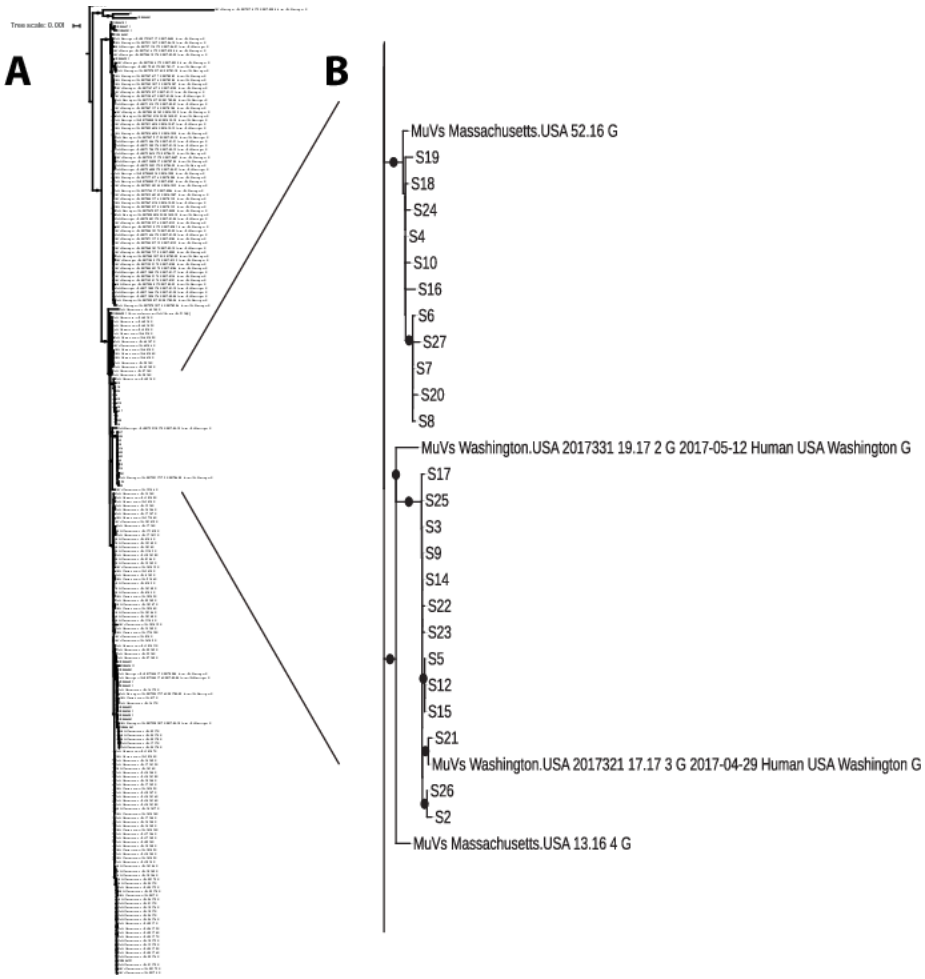


Phylogenetic tree of mumps isolates reproduced from nextstrain.org. Thick branches connect genotype G isolates. Orange circles indicate isolates from the USA and red circles isolates from Canada. The Y axis indicates percentage nucleotide diversity. ON, Ontario 2017 outbreak isolates; BC, British Columbia 2016 outbreak isolates.

Nextstrain analysis estimated the date of the most recent common ancestor for each of our major clades, using an ML discrete traits model. This indicated that sequences within each major Ontario clade coalesced in late 2016 or early 2017; the confidence interval (CI) for major clade 1 is 19th September to 7th January, and for major clade 2 is 1st October to 12th January. Both Ontario clades coalesced with clades from the US before they coalesced with each other in spring 2016.

When we examined the US association in greater detail, by ML comparison of our outbreak with 211 mumps genotype G strains from 2016/2017 outbreaks in the US, we identified 4 isolates which were closely related to our strains (Supplementary Fig. S1).

SUPPLEMENTARY FIGURE S1



(A) Maximum Likelihood tree of 211 mumps Genotype G samples from US in 2016 and 2017 and 25 Ontario 2017 isolates. The tree is rooted on JX287389.1, an ancestral strain from New York in 2012. (B) Enlargement of the Ontario outbreak clades, showing closely related Massachusetts and

Washington strains. Nodes with ultrafast bootstrap support values >90 are indicated with black circles.

Within Ontario major clade 1 is MuVs Massachusetts.USA 52.16 (GenBank accession MG986382), sequenced from a case with symptom onset on 27th of December 2016, but no history of travel according to accompanying epidemiological metadata. Ontario major clade 2 is associated with MuVS Massachusetts.USA 13.16 (GenBank accession MF965213), with symptom onset on 31st of March 2016 and a history of travel out of the country, but the specific destination of travel was not mentioned in the Massachusetts outbreak report[14]. Major clade 2 isolates are also closely related to 2 viruses from a Washington mumps virus outbreak in May 2017. This suggests that major clade 1 and 2 originated independently, from strains circulating in the US in 2016, with subsequent onward transmission from major clade 2 to Washington isolate MuVs Washington.USA 17.17. Only US 2016/2017 genotype G outbreak sequences were available for local comparison, so links to outbreaks in other countries cannot be definitively excluded.

Phylogeographic analysis

We examined the data for evidence to support an early hypothesis that distinct transmission networks existed in Toronto and other regions. We superimposed the main outbreak phylogenetic tree on a map of Southern Ontario to illustrate the geographic structure of the outbreak (Fig. 4).

FIGURE 4

The outbreak phylogenetic tree superimposed on a map of Southern Ontario using the software program GenGIS v2.5.3. Red circles indicate the public health unit regions (PHUR) and are connected by colored dashed lines to the tree tips. The two major and one minor clades contain isolates from Toronto and from outlying PHURs, but major clade 2 is predominately associated with Toronto (11

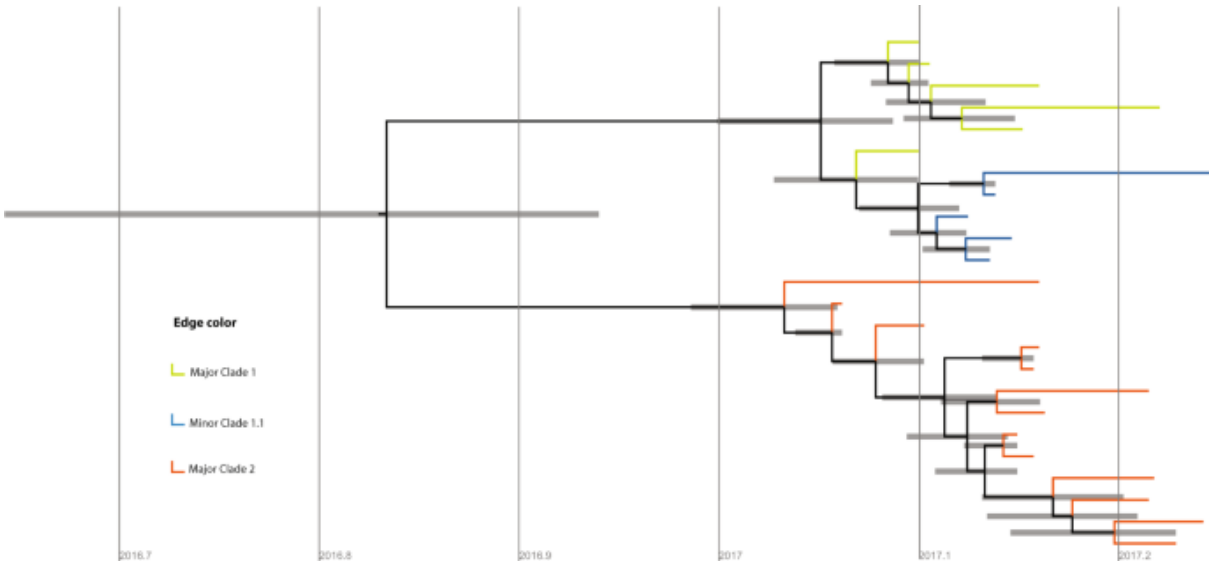
cases, 85%). Tree tips are colored as follows: Green, isolates in major clade 1 (excluding those that are also contained in minor clade 1.1); Blue, isolates in minor clade 1.1; Orange; isolates in major clade 2. Map image is the intellectual property of Esri and is used herin under license. Copyright © 2018 Esri and its licensors. All rights reserved. Map sources: National Geographic, Esri, Garmin, HERE, UNEP-WCMC, USGS, NASA, ESA, Increment P Corp.

Most cases in major clade 2 (n = 11, 85%) are from the city of Toronto. However cases from major clade 1 and minor clade 1.1 come from both Toronto and surrounding regions. This indicates that the geographical structure of the outbreak was more complex than assumed, with transmission networks extending across the province, rather than forming distinct Toronto and outlying area outbreaks.

Bayesian phylogenetic and transmission analysis

Bayesian phylogenetic analysis indicated that the time to Most Recent Common Ancestor (tMRCA) of our strains, and therefore the most likely date for origin of the outbreak, was October 25th 2016, but with a 95% high probability distribution (HPD) for the date of August 23rd to December 10th 2016 (Fig. 5).

FIGURE 5



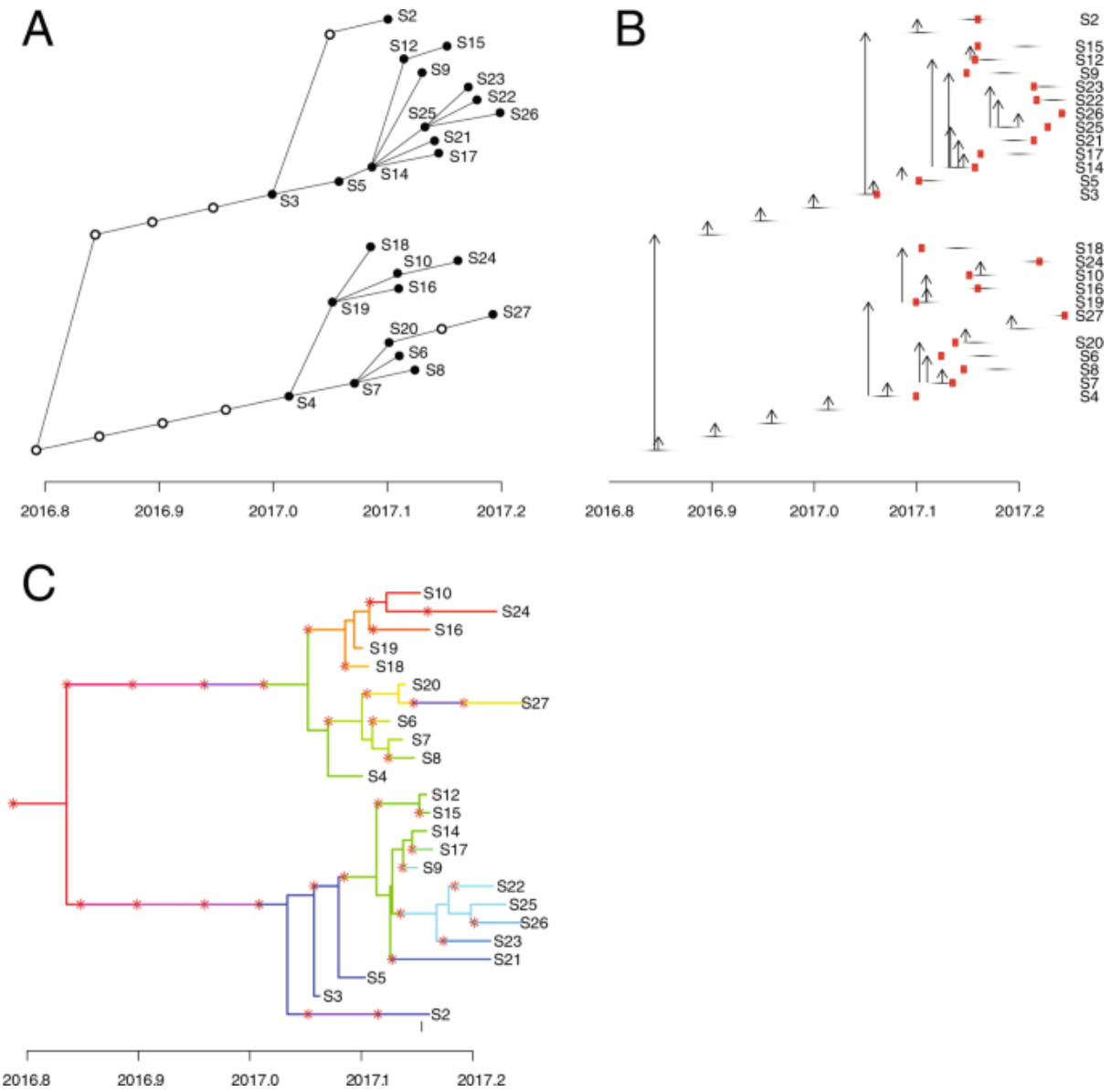
Timed tree from BEAST2 analysis. Time in decimal format is plotted on the Y axis. Nodes represent the mean estimate of time to most recent common ancestor (tMRCA) of descendant tips. Horizontal thick bars indicate 95% high probability distribution for the node height estimate. Tree tips are colored as follows: Green, isolates in major clade 1 (excluding those that are also contained in minor clade 1.1); Blue, isolates in minor clade 1.1; Orange isolates in major clade 2.

This is earlier than the date of the first case detected public health units, January 9th. The differences in estimates for tMRCA from Nextstrain (spring 2016) and Bayesian analysis are likely due to the different models used, and the greater diversity of samples in the Nextstrain dataset. Although the Bayesian model is a better fit than the Nextstrain model for rapid outbreak growth dynamics, the Nextstrain model has the advantage of using other strains to refine the date estimates. Results of Nextstrain, ML and Bayesian analyses support the hypothesis that the two major clades were introduced by separate importation events in late 2016 or early 2017. The tMRCA estimates allow for the possibility that there was some degree of silent virus transmission after each importation (e.g. due to mumps infections that were asymptomatic or caused only non-specific respiratory symptoms) for several weeks, but are equally consistent with rapid case detection.

The mean molecular clock estimate for the mumps outbreak from Bayesian analysis was 2.24×10^{-3} substitutions/site/year (95% HPD 1.39, 3.1). This is higher than earlier mean clock estimates for mumps based on analysis of F-SH-HN gene sequence of 0.25 (0, 0.43), or 0.65 (0, 1.4) for synonymous sites only[15]. In contrast to the dataset on which the earlier estimate is based, our data incorporated numerous substitutions in the most variable region of our strains, the L polymerase gene, and consists of samples collected over weeks rather than decades. Viral molecular clock rates will be overestimated by up to several orders of magnitude when derived from samples collected over a short period, compared to samples taken over an evolutionary timescale where purifying selection plays a significant role[16]. As the transmission tree software *Transphylo* requires a timed tree containing only outbreak sequences, we accepted the higher clock estimate imposed by these limitations.

The timed tree generated by Bayesian phylogenetic analysis was input to *TransPhylo* to generate consensus transmission trees (Fig. 6).

FIGURE 6


A
S2
S12
S15
S9
S23
S22
S25
S26
S21
S17
S14
S5
S3
S18
S10
S24
S16
S19
S27
S20
S6
S8
S7
S4
2016.8
2016.9
2017.0
2017.1
2017.2
B
S2
S15
S12
S9
S23
S22
S26
S25
S21
S17
S14
S5
S3
S18
S24
S10
S16
S19
S27
S20
S6
S8
S7
S4
2016.8
2016.9
2017.0
2017.1
2017.2
C
S10
S24
S16
S19
S18
S20
S27
S6
S7
S8
S4
S12
S15
S14
S17
S9
S22
S25
S26
S23
S21
S5
S3
S2
2016.8
2016.9
2017.0
2017.1
2017.2

(A) Consensus TransPhylo transmission tree, plotted in 'economical format'. The horizontal axis represents time in years, in decimal format. Filled circles represented sequenced cases. Empty circles represent putative unsampled cases, which the program has determined are required to reconcile the input phylogenetic tree with the pre-specified parameters for incubation period and infectious period, which must be provided as ranges for generation time and sampling time. If TransPhylo has determined that a direct transmission event is likely, then filled circles (sampled cases) are directly connected by a line. If sampled cases are separated by an unfilled circle (e.g. between S20 and S27), the program has determined that although the genomic sequences are similar, direct transmission is unlikely given known incubation and infectious period for mumps, and an unsampled intermediate case is required for transmission. (**B**) The transmission tree plotted in 'standard' format. Horizontal lines represent the infectious period for each sampled and unsampled case. Vertical arrows represent transmission from one case to another. Red squares indicate the time of patient sampling. Some of the purported transmission events are seen to reflect extreme assumptions e.g. S19 transmits to S18 at the earliest limit of the infectious period, and S18 is sampled before onset of symptoms. (**C**) The 'colored phylogenetic tree' view, consisting primarily of the timed tree also displayed in Fig. 5. The tree topology is unaltered, but Transphylo has colored the branches so that each case is represented by a unique color, and has inserted red asterisks to indicate each transmission event. This information provided by this view is the same as in panels A and B, and of the 3 potential Transphylo visual output formats, it is the least intuitive. Transphylo will maintain the clustering of cases seen on the input timed tree in its final outputs. Where 2 genomically similar cases share the same branch of the timed tree, but are separated by significant horizontal distance (S20 and S27, sampled weeks apart), it has colored a segment of the line purple, with red asterisks at the borders, to represent the unsampled intermediate case.

We found the 'economical' and 'standard' transmission tree viewing formats (Figs. 6A,B) easiest to interpret than the 'colored phylogenetic tree' format (Fig. 6C) when correlating the transmission tree with known epidemiological links. Epidemiologic data collected during routine public health investigation pertaining to 25 of the cases in our dataset were analysed separately to the genomic analysis, and identified three clusters of patients who shared common exposures; these clusters were then correlated with the transmission tree (Table 1).

Table 1 Correlation of epidemiological data for three clusters with *Transphylo* transmission tree.

Cluster	Cases	Epidemiological link	Transmission tree	Comments
A	S14, S17	Cases attended the same workplace and had collection of oral swab samples within one week of each other. No other case was identified at the workplace.	S14 transmits infection to S17	As the minimum mumps incubation period is 15 days, it is unlikely that S14 could have transmitted infection directly to S17
B	S4, S19, S10	S4 and S19 acquired their infection on the same date from a common source at a private residence. S10 lived nearby. Samples were collected from S4 and S19 on the same date and from S10 two weeks later	S4 transmits infection to S19 who transmits to S10 (and others)	*TransPhylo* model inferred that S4 was sampled after the infectious period, and S19 sampled at the beginning of the infectious period, to reconcile the fact that the collection date was the same for these samples but the

Cluster	Cases	Epidemiological link	Transmission tree	Comments
				timed tree suggested direct transmission from one to other
C	S12, S22, S23, S26 and S27	Cases linked by common exposure "Alcohol Serving Establishment (Bar/Tavern/Other)" and deemed to represent a potential cluster of cases associated with this general exposure setting.	S25[a] transmitted to S22, S23 and S26. S14[b] transmitted to S12 and S25. S27 is not linked to this cluster	S27 mumps virus strain is distinct from this cluster (it is part of a different major clade on phylogenetic analysis), so it can be confidently excluded from cluster C, despite the common exposure.

1. [a]No epidemiological data is available for this case. [b]No recorded exposure to a Toronto bar.

Although most cases were not part of a cluster (no common exposures, n = 14, 58%), we identified three clusters (A - C) comprised of two, three and five cases respectively. The transmission tree proposes close links between the cases concerned. The only exception is case S27 from cluster C, where there is strong phylogenetic evidence to refute the proposed epidemiological link. In clusters A and B we observed discordance between the epidemiology and the transmission tree regarding the likely direction of transmission. In these clusters, *TransPhylo* analysis proposed direct transmission between patients, when the epidemiological data suggests they are more likely to have acquired the infection from a common source. The transmission tree identifies new purported links between cases that were not part of an epidemiological cluster previously (S7 transmitting to S6, S8 and S20), or that are now implicated as additional cases in an existing cluster (S16 and S18 are linked with the cases in cluster B). Due to the retrospective nature of this study, we were unable to perform further epidemiological investigations to support or refute these new links.

DISCUSSION

Genomic epidemiology based on sequencing an entire microbial genome is now routinely used for outbreaks of high profile pathogens such as Ebola virus and Zika virus[17,18] and surveillance of pathogens such as *Mycobacterium tuberculosis*, where public health agencies expend enormous effort in identifying contacts who require chemoprophylaxis or treatment[19]. There are few studies published to date on the application of genomic epidemiology to mumps virus outbreaks; at the time the Ontario outbreak commenced we could find no studies correlating WGS results with outbreak investigations and to date only one study has been reported[14].

In this retrospective study we developed protocols for amplicon sequencing from virus cultures and for identifying SNVs and clusters of closely related strains. Meaningful interpretation of clusters then requires close collaboration with public health professionals, who have specialist expertise in outbreak control, but who may not be accustomed to interpreting traditional visualisations of comparative genomic analyses, such as phylogenetic trees. We particularly focused on generating intuitive visual summaries of the phylogenomic data, such as maps illustrating the geographic structure of the outbreak, and transmission trees showing patient to patient spread.

We refined our previously published method for amplifying mumps virus RNA from virus culture samples[13]. At PHO, virus culture is routinely performed on all specimens that are reactive by RT-PCR, to assist with SH genotyping, and we have previously found the sensitivity of culture and RT-PCR to be broadly equivalent. Both virus culture and amplicon sequencing enrich for reads from the target virus rather than from the host or commensal bacteria. This allowed us to sequence our samples in a multiplex fashion on the Illumina MiSeq at PHO laboratories as part of runs where two thirds of sequencing capacity was allocated to bacterial pathogens for routine surveillance activities, and still achieve high depth of coverage of the target virus.

In future outbreaks, it may be desirable to sequence mumps virus directly from primary clinical specimens. This would reduce turnaround time by eliminating the culture step, which at our institution typically takes 7–10 days, but for samples weakly positive by RT-PCR may take up to 17 days or may fail to grow. Direct specimen sequencing would eliminate the possibility that SNVs could arise during cell culture passage. The impact of this potential confounder is unknown, as to our knowledge comparison of mumps virus sequence from before and after cell culture has not been performed. We investigated if it was possible to use our 9 amplicon protocol to sequence virus directly from 7 oral swabs positive for mumps virus, all with Ct values less than 33. However we could not obtain amplicon adequate product for all fragments under standard thermocycler conditions. Quick *et al*. recently described modified primer design and RT-PCR protocols for amplicon sequencing of Zika virus and chikungunya virus directly from clinical specimens[20]. Alternative methods for direct specimen sequencing include metagenomic approaches, either unbiased, or enriched with viral hybrid capture[21]. Hybrid capture has recently been shown to reliably recover sequence from buccal swabs positive by RT-PCR for mumps virus if the Ct value is under 30[14]. However in our experience the bioinformatics analysis was the time-limiting step, particularly optimisation of the models for generating timed trees and transmission trees.

Standard phylogenetic analysis based on SNV identification and construction of the ML tree was able to differentiate 24 outbreak strains from 2 non-outbreak strains. Only one Genotype C strain would have been differentiated using SH genotyping alone. The higher resolution provided by WGS allowed us to identify 2 major clade and 1 minor subclade within the outbreak. Identifying subclades is a starting point for identifying transmission events, since each clade is presumed to share a common ancestor, and therefore an epidemiological link. The phylogenetic tree was also used as input for a GenGIS phylogeographic analysis, which showed that strains from Toronto and surrounding regions were closely related. This was a question early in the outbreak, when the extent of strain sharing between public health units was unclear. Identifying strains shared between multiple health units may in the future help with multi-jurisdictional co-ordination of outbreak control efforts.

The transmission tree generated using *TransPhylo* was, in our opinion, superior to other means of visualising genomic data such as ML trees. Unlike ML trees it was able to incorporate available data about the timing of sample collection, which is a key factor to consider when using genomic data to identify transmission clusters[22]. *TransPhylo* used partial sampling data to generate a representation of case to case transmission, with clear illustrations of the assumptions the model made for each case to reconcile input data (a timed tree and dates of case sampling) with the constraints we imposed regarding the ranges for the infectious period and from virus acquisition to sampling. When we compared cases from 3 clusters with known epidemiological associations to the transmission tree, we found that the transmission tree independently identified close links between the cases. In the only exception, the results strongly support the genomic data over the epidemiologic data, since the case was infected with a strain from a different clade to other cases with the same category of epidemiological exposure.

Clearly *TransPhylo* has limitations; in 2 clusters it postulated direct patient to patient transmission when it is much more likely that cases acquired their infection from a common source. When we explored the inferences made by the model leading to errors for cluster B, we found the model made extreme assumptions about the time from case acquisition to sampling for a pair of cases (very early and very late in the course of illness respectively). This appears to have been done in order to reconcile an input timed tree showing considerable genomic distance between 2 strains, with the fact that the sampling date was the same for both cases. Although the transmission tree appears to be only a rough approximation of the true transmission network, we believe it may prove useful in future outbreaks. A key factor in future outbreak investigations will be the ability to generate accurate ML and timed trees from epidemiologically targeted or comprehensive case sequencing, rather than retrospective convenience sampling as in this study. The accuracy of

phylogeographic and transmission models is constrained by the accuracy of the phylogenetic tree used as input, and the transmission model presented here would benefit from further refinement with independent datasets.

Whether mumps virus outbreaks have sufficient public health impact to warrant the expenditure of time and resources required to perform WGS based transmission network analysis is a subject for debate. Over the course of our outbreak investigation only a minority of cases had clear epidemiological links to other cases, so there is a need for genomic analyses to generate hypotheses with respect to transmission networks, and to inform additional case tracking measures. In most mumps outbreaks, interventions are limited to vaccination clinics for at risk populations, and to case isolation, as well as general messaging aimed at limiting transmission (i.e. advice to avoid sharing utensils and water bottles). These interventions were applied during the period of increased mumps activity in Ontario, but advice on immunisation was disseminated primarily through both traditional and social media, asking individuals to speak with their personal healthcare professional about immunisation. Prospective phylogenomic transmission network analysis could play an important role, by helping to identify hotspots for transmission and to define more precisely and vaccinate the population at risk. Public health agencies interested in prospectively applying these novel techniques should consider undertaking preparatory work to develop the necessary sequencing, bioinformatics and data visualisation methods. We are not aware of any study that has demonstrated a real, rather than potential, public health benefit from mumps WGS analysis, and to do so will require methods optimised in advance to deliver rapid results. Our study demonstrates that WGS of mumps virus is readily performed from virus culture and that traditional phylogenomic analyses are complemented by phylogeographic and transmission network analyses. Comparison of outbreak strains with sequences from traditional and novel data repositories helps identify potential international transmission events, which can then be correlated with results of epidemiological investigations. Transmission network analysis based on sequences from a small fraction of total cases generated results that were partly supported by known epidemiological associations. Limitations of our study included the small number of isolates sequenced, that our model inferred direct patient to patient transmission when acquisition from common sources was more likely, and that we were unable to further investigate potential new transmission links to confirm or refute them with epidemiological data. We believe that prospective phylogenomic analyses are needed to determine if WGS can be used to identify cryptic transmission chains in real-time and define the at-risk populations who would benefit from mumps containing vaccine.

MATERIALS AND METHODS

Strain collection

Throat and buccal samples from all potential cases, identified either by primary clinicians or as a result of public health investigations of mumps, are routinely sent to the Public Health Ontario (PHO) Laboratories for analysis. All swabs were tested at PHO by reverse-transcription polymerase chain reaction (RT-PCR) targeting both SH and Fusion (F) genes, using an in-house assay adapted from protocols developed by Canada's National Microbiology Laboratory (NML) and the US Centers for Disease Control and Prevention (CDC)[23,24]. Samples that were reactive in this assay were cultured in rhesus monkey kidney cell primary cell lines to assist with genotyping (Quidel Corporation, San Diego, CA). Total culture time is 17 days, including one passage at 10 days, although in our center most samples with Ct values < 30 are usually culture positive by 7 to 10 days. As this was a pilot study, we selected a convenience sample of 27 positive cultures, 17 of which were from Toronto, for sequencing. The first 20 cultures were selected randomly from samples collected in the first 2 months of 2017, at the beginning of the outbreak, and were sequenced in March. The remaining 7 cultures were selected from samples collected in the final 2 weeks of March and sequenced in June. We chose samples over a 3 month window to ensure adequate temporal signal in our dataset to enable us to perform phylogenetic molecular clock analysis.

RNA extraction and sequencing protocol

Nucleic acid extraction and sequencing was carried out using a modified version of a tiling amplicon-based method to enrich the culture supernatant for viral RNA, which we previously used to sequence mumps virus from a 2010 Ontario outbreak[13]. We extracted RNA using either the QIAamp Viral RNA Mini Kit (Qiagen, Mississauga, ON) or the NucliSENS easyMAG instrument. For the initial eight samples we performed amplification of 18 overlapping amplicons, of mean length 977 bp. We optimised the protocol to reduce the number of amplicons, so for the last 19 samples we sequenced 9 amplicons of mean length 1958 bp (Supplementary Dataset 2 for primers). Amplification of the fragments in 96 well plates was performed on a SimpliAmp thermal cycler using the superscript III One Step RT-PCR system (Invitrogen,Thermo Fisher Scientific).

Amplicon fragments from individual samples were pooled together in equal amounts and cDNA concentration checked using a Qubit fluorometer. Mumps cDNA libraries were prepared with the Nextera XT kit. We checked the quality of the indexed libraries by Bioanalyzer. Sequencing on the

Illumina MiSeq instrument was performed with V2 reagent kit (2 × 150 bp, Illumina Inc. San Diego, California, USA), according to the manufacturer's instructions.

Phylogenetic analysis

We removed adapters, primer sequences and low quality reads with Trimmomatic[25]. We created de-novo assemblies for each isolate using Spades v3.12.0[26] as implemented in shovill v.0.9.0 (https: https://github.com/tseemann/shovill). Assembly errors were corrected by mapping trimmed reads back to each assembly with snippy v3.2-dev (https://github.com/tseemann/snippy).

We used MEGA7 to manually align our assemblies with a reference genome (accession JX287389) and with representative sequences of various mumps genotypes obtained from NCBI Genbank. We created a maximum likelihood (ML) tree from the full alignment using iqtree v1.6. To identify outbreak specific SNVs, we used snippy with default parameters to map sequencing reads of outbreak strains against an in-group reference (annotated de-novo assembly S7) which had >99% average nucleotide identify to other genotype G sequences in GenBank.

In order to compare our genotype G strains with as many whole genome sequences as possible, in addition to searching NCBI GenBank, we also conducted internet searches for sequences located outside of traditional data repositories. We retrieved 121 relevant Massachusetts outbreak sequences and associated clinical metadata from the bioRxiv preprint server for biology; an alignment of "clade-II" sequences was published as a supplement[14]. From github.com we retrieved 72 sequences from a research laboratory repository of sequences relating to a Washington outbreak, after obtaining permission from the researchers. Ultimately we created an ML tree from an alignment of 25 Ontario and 211 USA genotype G complete sequences from outbreaks occurring in 2016 and 2017, but we did not find sequences any from other countries in this period. Trees were visualised and prepared for publication using iTOL (http://itol.embl.de). We uploaded our outbreak clade genomes to the Nextstrain project (https://nextstrain.org).

Phylogeographic analysis

To illustrate the phylogeographic structure of the outbreak we used the program GenGIS v2.5.3[27] to combine the outbreak clade ML tree with a digital map with a location of the health unit where the sample was collected.

Bayesian phylogenetic analysis

We required a time labelled phylogenetic tree as a starting point for the *TransPhylo* analysis, so we performed a Bayesian phylogenetic analysis of the outbreak strains using BEAST2 v2.4.7[28]. The complete consensus genome alignment of the outbreak clade was labelled with the collection date for each specimen. We assessed regression of root-to-tip distance in TempEst v1.5 and confirmed adequate temporal signal to proceed to Bayesian analysis. We used the birth-death skyline serial model as implemented in the BDSKY package v1.3.3, as an appropriate model for a RNA virus outbreak with changing dynamics due to the presence of resistant individuals and the depletion of the susceptible individuals[29]. We used the following parameters: HKY Model of evolution with empirical frequencies, gamma category count 4, proportion invariant sites 0.98 and a strict molecular clock. A strict clock was chosen as appropriate to a single outbreak in one location and was supported by the root-to-tip regression. When we attempted to run the analysis with a relaxed molecular clock we did not achieve convergence of the chain. We chose diffuse priors for the virus evolution rate, proportion of outbreak sampled, the rate at which patients become uninfectious and the reproductive number. The analysis was run for 40 million Markov Chain Monte Carlo (MCMC) simulations, with sampling from the posterior distribution every 4000 steps. Evaluation of the posterior probability of the parameters with Tracer v1.6 indicated adequate mixing of the chain, and all parameters achieved an effective sampling size (ESS) >200. The posterior sample of phylogenetic timed trees was summarised with TreeAnnotator v2.4.7, with the first 10% discarded as burn-in and an output tree of maximum sum of clade credibility with median node heights, which was visualised with Icytree[30].

***TransPhylo* analysis**

For input to *TransPhylo* we used the time labelled phylogenetic tree along with initial estimates for the following parameters: sampling proportion Pi 0.1, date sampling of the outbreak stopped (last specimen collection, 2017.246) and a gamma distribution specifying the generation time, or the time between an individual's primary infection and a secondary infection that they give rise to. Authorities such as WHO, CDC and the public health agency of Canada (http://www.health.gov.on.ca/en/pro/programs/publichealth/oph_standards/docs/mumps_chapter.pdf) give slightly different intervals and ranges for the incubation period and the period of infectiousness[31]. We specified a gamma distribution for the generation time with shape 64 and scale 0.000856, resulting in a mean of 20 days and 95% distribution of 14–30 days, in an attempt to incorporate WHO guidance on the incubation period (14–28 days, mean interval 16–18 days), and infectious period (−2 to +7 days from symptom onset), in a single distribution. We also specified a similar gamma distribution for sampling time (incubation time plus time from symptom onset to

collection date) with a mean of 23 days (ws.shape 7, ws.scale 0.000856), specifying that our samples were most likely collected between 1 and 7 days from the onset of symptoms. This was an empiric estimate of time from symptoms to sampling, but subsequently our data linkage revealed that 23 of 25 cases with known onset and collection dates were within this time window (median 2 days, outliers were 0 days and 13 days).

TransPhylo uses MCMC simulation to analyse many thousands of possible transmission trees. Our simulation was run for 100000 MCMC simulations, with sampling of a tree every 1000 steps. We generated a consensus transmission tree from the output, with burn-in proportion of 0.5 and a minimum probability for inclusion of a partition in the consensus of 0.5.

Epidemiological data

We compared the resulting transmission tree with previously collected epidemiological data recorded in the integrated Public Health Information System (*iPHIS*), which is Ontario's electronic reporting system for reportable diseases. Of the 26 cases that were sequenced, 25 were matched to cases in iPHIS; one case could not be linked as the individual resided outside Ontario. Epidemiological data were extracted from iPHIS on April 20, 2018. PHO identified possible transmission clusters from the epidemiological data before reviewing the results of the genomic analysis. Clusters were defined as cases that had close contact with each other or that shared common exposures, as recorded in iPHIS.

ETHICAL APPROVAL

The study protocol was approved by the PHO Ethics Review Board (ERB, File number 2017-053.01) and Privacy Office (Privacy assessment RRB-18-010). The ERB waived requirement for informed consent as the study satisfied the conditions of article 5.5A of the Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS2).

DATA AVAILABILITY

Complete genome sequences for the 26 strains have been deposited in NCBI GenBank with accession numbers MK033747 to MK033772. GenBank accession numbers and WHO names for each sequence are provided in Supplementary Dataset 3.

REFERENCES

1. Hviid, A., Rubin, S. & Mühlemann, K. Mumps. *Lancet* **371**, 932–944 (2008).
2. Dayan, G. H. *et al*. Recent Resurgence of Mumps in the United States. *N. Engl. J. Med.* **358**, 1580–1589 (2008).
3. Carr, M. J. *et al*. Molecular epidemiological evaluation of the recent resurgence in mumps virus infections in Ireland. *J. Clin. Microbiol.* **48**, 3288–94 (2010).
4. Willocks, L. J. *et al*. An outbreak of mumps with genetic strain variation in a highly vaccinated student population in Scotland. *Epidemiol. Infect.* **145**, 3219–3225 (2017).
5. Deeks, S. L. *et al*. An assessment of mumps vaccine effectiveness by dose during an outbreak in Canada. *CMAJ* **183**, 1014–20 (2011).
6. Donahue, M. *et al*. Notes from the Field: Complications of Mumps During a University Outbreak Among Students Who Had Received 2 Doses of Measles-Mumps-Rubella Vaccine — Iowa, July 2015–May 2016. *MMWR. Morb. Mortal. Wkly. Rep.* **66**, 390–391 (2017).
7. Patel, L. N. *et al*. Mumps Outbreak Among a Highly Vaccinated University Community—New York City, January–April 2014. *Clin. Infect. Dis.* **64**, ciw762 (2016).
8. Jin, L. *et al*. Genomic diversity of mumps virus and global distribution of the 12 genotypes. *Reviews in Medical Virology*, https://doi.org/10.1002/rmv.1819 (2015).
9. Gavilán, A. M. *et al*. Genomic non-coding regions reveal hidden patterns of mumps virus circulation in Spain, 2005 to 2015. *Eurosurveillance* **23**, 17–00349 (2018).
10. Gardy, J. L. *et al*. Whole-Genome Sequencing of Measles Virus Genotypes H1 and D8 During Outbreaks of Infection Following the 2010 Olympic Winter Games Reveals Viral Transmission Routes. *J. Infect. Dis.* **212**, 1574–1578 (2015).
11. Dubey, V. *et al*. Investigation and management of a large community mumps outbreak among young adults. *Can Commun Dis Rep* **44**, 309–16 (2018).
12. Trotz-Williams, L. A. *et al*. Challenges in Interpretation of Diagnostic Test Results in a Mumps Outbreak in a Highly Vaccinated Population. *Clin. Vaccine Immunol.* **24** (2017).
13. L'Huillier, A. G. *et al*. Laboratory testing and phylogenetic analysis during a mumps outbreak in Ontario, Canada. *Virol. J.* **15**, 98 (2018).
14. Wohl, S. *et al*. Co-circulating mumps lineages at multiple geographic scales. *bioRxiv* 343897, https://doi.org/10.1101/343897 (2018).
15. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. *J. Mol. Evol.* **54**, 156–165 (2002).
16. Duchene, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. B Biol. Sci.* **281**, 20140732–20140732 (2014).
17. Gire, S. K. *et al*. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (80-.).* **345**, 1369–1372 (2014).
18. Thézé, J. *et al*. Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. *Cell Host Microbe* **23**, 855–864.e7 (2018).
19. Lee, R. S. *et al*. Reemergence and Amplification of Tuberculosis in the Canadian Arctic. *J. Infect. Dis.* **211**, 1905–1914 (2015).
20. Quick, J. *et al*. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
21. Briese, T. *et al*. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6**, e01491–15 (2015).
22. Stimson, J. *et al*. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *bioRxiv* 319707, https://doi.org/10.1101/319707 (2018).

23. Uchida, K. *et al*. Rapid and sensitive detection of mumps virus RNA directly from clinical samples by real-time PCR. *J. Med. Virol.* **75**, 470–474 (2005).
24. Boddicker, J. D. *et al*. Real-time reverse transcription-PCR assay for detection of mumps virus RNA in clinical specimens. *J. Clin. Microbiol.* **45**, 2902–8 (2007).
25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
26. Bankevich, A. *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).
27. Parks, D. H. *et al*. GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. *PLoS One* **8**, e69885 (2013).
28. Bouckaert, R. *et al*. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
29. Barido-Sottani, J. *et al*. Taming the BEAST – A community teaching material resource for BEAST 2. *Syst Biol* **67**, 170–174 (2017).
30. Vaughan, T. G. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* **33**, 2392–2394 (2017).
31. Czumbel, I., Quinten, C., Lopalco, P., Semenza, J. C. & ECDC expert panel working group, T. E. expert panel working. Management and control of communicable diseases in schools and other child care settings: systematic review on the incubation period and period of infectiousness. *BMC Infect. Dis.* **18**, 199 (2018)

ACKNOWLEDGEMENTS

We thank Louise Moncla for assistance with nextstrain.org analysis and permission to use unpublished Washington strains in the ML comparison, and Jennifer Gardy for permission to use unpublished British Columbia strains in the nextstrain.org comparison. Eddie Chongking, Mark Cardona and Steve Masney for provision of viral culture isolates; Gursimran Bahia for assistance with amplicon generation; Aimin Li for assistance with Illumina sequencing; Venkata Duvvuri for assistance with Bayesian analysis; Ilone Harrison, Catherine Oyiliagu, Gillian Lim and Alex Marchand-Austin for supporting linkage of genomic and epidemiological data. Finally we thank all individuals with mumps infection included in this analysis and the staff across the many public health units in Ontario involved in the public health response to this outbreak.

CONTRIBUTIONS

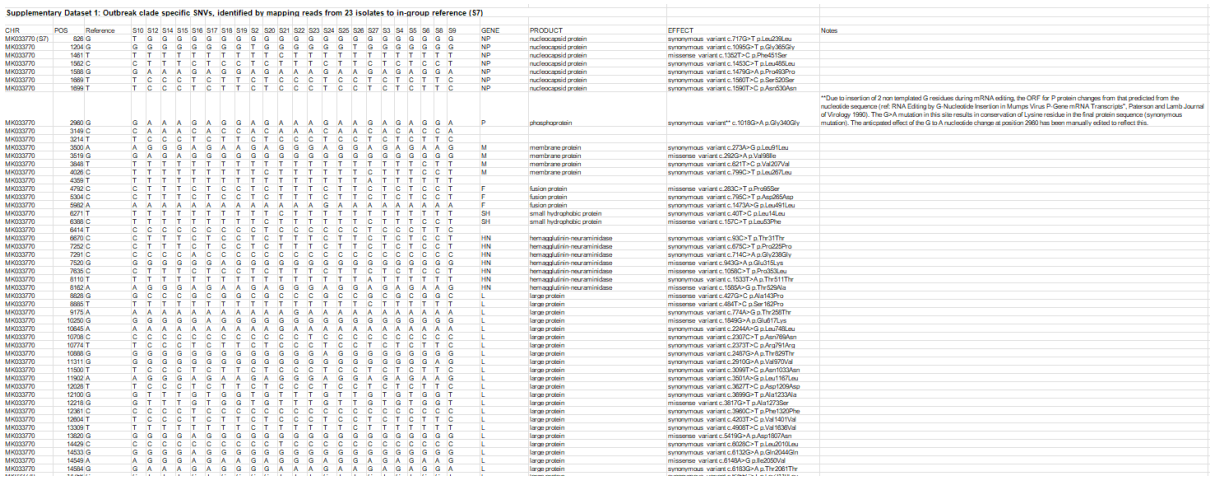
P.J.S., A.E., S.B. J.B.G. and S.N.P. designed the study. A.E. and P.J.S. performed the amplicon generation and sequencing library preparation. P.J.S. performed the bioinformatics analysis and drafted the manuscript. C.Y.S, S.W., T.H. and S.L.D. provided support to Ontario public health units who led the epidemiological investigations, extracted and linked epidemiologic data to the genomic data and interpreted the linked dataset. L.W.G. and J.G. provided valuable insights on study design and interpretation of results. All authors reviewed and approved the manuscript.

ETHICS DECLARATIONS

The authors declare no competing interests.

SUPPLEMENTARY INFORMATION

Supplementary Dataset 1



Supplementary Dataset 1: Outbreak clade specific SNVs, identified by mapping reads from 23 isolates to in-group reference (S7)

CHR	POS	Reference	S10	S12	S14	S15	S16	S17	S18	S19	S2	S20	S21	S22	S23	S24	S25	S26	S27	S3	S4	S5	S6	S8	S9	GENE	PRODUCT	EFFECT	Notes
MK033770 (S7)	626	G	T	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	NP	nucleocapsid protein	synonymous variant c.717G>T p.Leu239Leu	
MK033770	1204	G	G	G	G	G	G	G	G	T	G	G	G	G	G	G	T	G	G	G	G	G	G	G	G	NP	nucleocapsid protein	synonymous variant c.1096G>T p.Gly365Gly	
MK033770	1461	T	T	T	T	T	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	T	T	NP	nucleocapsid protein	missense variant c.1352T>C p.Phe451Ser	
MK033770	1552	C	C	T	T	T	C	T	C	C	T	C	T	T	T	C	C	T	T	C	T	C	T	C	C	T	NP	nucleocapsid protein	synonymous variant c.1453C>T p.Leu485Leu	
MK033770	1588	G	G	A	A	G	A	G	G	A	G	A	A	A	G	A	A	G	A	G	A	G	G	A	A	NP	nucleocapsid protein	synonymous variant c.1479G>A p.Pro493Pro	
MK033770	1869	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	NP	nucleocapsid protein	missense variant c.1560T>C p.Ser509Ser	
MK033770	1899	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	NP	nucleocapsid protein	synonymous variant c.1590T>C p.Asn530Asn	
MK033770	2960	G	G	A	A	A	G	A	G	G	A	G	A	A	A	G	C	A	A	G	A	G	A	G	A	P	phosphoprotein	synonymous variant** c.1019G>A p.Gly340Gly	**Due to insertion of 2 non templated G residues during mRNA editing, the ORF for P protein changes from that predicted from the nucleotide sequence (ref. RNA Editing by G-Nucleotide Insertion in Mumps Virus P-Gene mRNA Transcripts", Paterson and Lamb Journal of Virology 1990). The G>A mutation in this site results in conservation of Lysine residue in the final protein sequence (synonymous mutation). The anticipated effect of the G to A nucleotide change at position 2960 has been manually edited to reflect this.
MK033770	3149	C	C	A	A	C	A	C	C	A	C	A	A	A	C	C	A	A	C	A	C	A	C	C	A				
MK033770	3214	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C				
MK033770	3500	A	A	G	G	G	A	G	A	A	G	A	G	G	G	A	A	G	G	A	G	A	G	A	A	M	membrane protein	synonymous variant c.273A>G p.Leu91Leu	
MK033770	3519	G	G	A	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	M	membrane protein	missense variant c.292G>A p.Val98Ile	
MK033770	3848	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	C	T	T	M	membrane protein	synonymous variant c.621T>C p.Val207Val	
MK033770	4026	C	T	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	C	T	T	T	C	C	T	M	membrane protein	synonymous variant c.799C>T p.Leu267Leu	
MK033770	4399	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	A	T	T	T	T	T	T				
MK033770	4792	C	C	T	T	T	C	T	C	C	T	C	T	T	T	C	T	T	C	T	C	T	C	C	T	F	fusion protein	missense variant c.263C>T p.Pro663se	
MK033770	5304	C	C	T	T	T	C	T	C	C	T	C	T	T	T	C	T	T	C	T	C	T	C	C	T	F	fusion protein	synonymous variant c.765C>T p.Asp365Asp	
MK033770	5662	A	A	A	A	A	A	A	A	A	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	F	fusion protein	synonymous variant c.1473A>G p.Leu491Leu	
MK033770	6271	T	T	T	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	SH	small hydrophobic protein	synonymous variant c.42T>C p.Leu14Leu	
MK033770	6398	C	T	T	T	T	T	T	T	T	C	T	T	T	T	T	T	T	C	T	T	T	C	T	T	SH	small hydrophobic protein	missense variant c.157C>T p.Leu53Phe	
MK033770	6414	T	C	C	C	C	C	C	C	C	T	C	C	C	C	C	C	C	T	C	C	C	T	T	C				
MK033770	6670	C	C	T	T	C	T	C	C	T	C	T	T	T	C	T	T	C	C	T	C	T	C	C	T	HN	hemagglutinin-neuraminidase	synonymous variant c.93C>T p.Thr31Thr	
MK033770	7252	C	C	T	T	T	C	T	C	C	T	C	T	T	T	C	T	T	C	T	C	T	C	C	T	HN	hemagglutinin-neuraminidase	synonymous variant c.675C>T p.Pro225Pro	
MK033770	7291	C	C	C	C	A	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	HN	hemagglutinin-neuraminidase	synonymous variant c.714C>A p.Gly238Gly	
MK033770	7520	G	G	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	HN	hemagglutinin-neuraminidase	missense variant c.943G>A p.Glu315Lys	
MK033770	7635	C	C	T	T	C	T	C	C	T	C	T	T	T	C	T	T	C	C	T	C	T	C	C	T	HN	hemagglutinin-neuraminidase	missense variant c.1058C>T p.Pro353Leu	
MK033770	8110	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	A	T	T	T	T	T	HN	hemagglutinin-neuraminidase	synonymous variant c.1533T>A p.Thr511Thr	
MK033770	8162	A	A	G	G	G	A	G	A	A	G	A	G	G	G	A	A	G	G	A	G	A	A	G	G	HN	hemagglutinin-neuraminidase	missense variant c.1585A>G p.Thr529Ala	
MK033770	8628	G	G	C	C	C	G	C	G	G	C	G	C	C	C	G	C	C	G	C	G	C	G	G	C	L	large protein	missense variant c.427G>C p.Ala143Pro	
MK033770	8895	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	C	T	T	T	T	T	L	large protein	missense variant c.464T>C p.Ser162Pro	
MK033770	9175	A	A	A	A	A	A	A	A	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	A	L	large protein	synonymous variant c.774A>G p.Thr258Thr	
MK033770	10250	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	L	large protein	missense variant c.1949G>A p.Gly617Lys	
MK033770	10645	A	A	A	A	A	A	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	A	A	A	L	large protein	synonymous variant c.2244A>G p.Leu748Leu	
MK033770	10798	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	L	large protein	synonymous variant c.2397C>T p.Asn799Asn	
MK033770	10774	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	L	large protein	synonymous variant c.2373T>C p.Arg791Arg	
MK033770	10999	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	L	large protein	missense variant c.2467G>A p.Thr597Thr	
MK033770	11311	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	L	large protein	synonymous variant c.2910G>A p.Val970Val	
MK033770	11500	T	C	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	L	large protein	synonymous variant c.3099T>C p.Asn1033Asn	
MK033770	11902	A	A	G	G	A	G	A	A	G	G	G	A	G	G	A	G	G	A	G	A	G	A	A	G	L	large protein	synonymous variant c.3501A>G p.Leu1167Leu	
MK033770	12028	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	L	large protein	synonymous variant c.3627T>C p.Asp1209Asp	
MK033770	12100	G	G	T	T	T	G	T	G	G	T	T	T	T	T	G	T	T	G	T	G	T	G	G	T	L	large protein	missense variant c.3699G>T p.Ala1233Ala	
MK033770	12218	G	G	T	T	T	G	T	G	G	T	G	T	T	T	G	T	T	G	T	G	T	G	G	T	L	large protein	missense variant c.3817G>T p.Ala1273Ser	
MK033770	12361	C	C	C	C	C	T	C	C	C	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	L	large protein	synonymous variant c.3960C>T p.Phe1320Phe	
MK033770	12604	T	T	C	C	C	T	C	T	T	C	T	C	C	C	T	C	C	T	C	T	C	T	T	C	L	large protein	synonymous variant c.4203T>C p.Val1401Val	
MK033770	13099	T	T	T	T	T	T	T	T	T	C	T	T	T	T	T	C	T	T	T	T	T	T	T	T	L	large protein	synonymous variant c.4698T>C p.Val1636Val	
MK033770	13820	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	L	large protein	missense variant c.5419G>A p.Asp1807Asn	
MK033770	14429	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	L	large protein	synonymous variant c.6028C>T p.Leu2010Leu	
MK033770	14533	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	L	large protein	synonymous variant c.6132G>A p.Gln2044Gln	
MK033770	14549	A	A	G	G	G	A	G	A	A	G	A	G	G	G	A	G	G	A	G	A	G	A	A	G	L	large protein	missense variant c.6148A>G p.Ile2050Val	
MK033770	14584	G	G	A	A	A	G	A	G	G	G	G	A	A	A	G	A	A	G	A	G	A	G	G	A	L	large protein	synonymous variant c.6183G>A p.Thr2061Thr	

Supplementary Dataset 2

Primer scheme for amplicon sequencing

Mumps whole genome sequence 18 fragment protocol primer scheme

Fragment	Primer name	Sequence 5'- 3'	Position	Product size
1	forward	ACCAAGGGGAAAATGAAGATGGGA	1-24	1049
	reverse	AACATCATCAAGGATCGGAGCTTA	1027-1050	
2	forward	AATGGTGGGTGACATTGGAAAGTA	901-924	1188
	reverse	ACTTGCTCAACGAGTTGGTTCCCT	2067-2089	
3	forward	TGATCGTGTTTCTTTCCGGGCAAG	1953-1976	1047
	reverse	AATCCTTTACTGGTTGGGGTTTTG	2977-3000	
4	forward	GACCAGGAGATGTGTCATTCAGCT	2881-2904	1088
	reverse	CGGATGCAATGCACCCTTCTCCAT	3946-3969	
5	forward	ATCTAGACTAGTAAGAGCGGTTCA	3851-3874	1098
	reverse	ACGTGCATTTGTCTGTGCTTGAAC	4927-4950	
6	forward	CTTCCAATTGCAGAAAACATAAAC	4786-4809	1149
	reverse	CGGCATTTTGGAGGGATGCATTAA	5912-5935	
7	forward	CATCACTTATGCTGAGAACCTTAC	5826-5849	974
	reverse	CTTGATCATTGATCATCCTTACAA	6777-6800	

8	forward	AGATGGAGCCCTCGAAGCTCTTCA	6612-6635	788
	reverse	GTTGAGTTGAGACGTAACAGTACA	7377-7400	
9	forward	GGTATCCTATGTTCAAAACCTTAA	7281-7304	1087
	reverse	ACTATGACTCAAGTGATAGTCAAT	8347-8370	
10	forward	TACACCACAACCACCTGCTTTCAA	8231-8254	1135
	reverse	ACAACAGGATCACCATACTGTAAT	9343-9366	
11	forward	ATTCTCCTAACATTGGTTGATGAC	9251-9274	853
	reverse	TCCTTCTCCTTGAGAGAGTAAGAT	10081-10104	
12	forward	GAATTTTCTGGAGGATGACAGATT	9991-10014	957
	reverse	TATGAAGAAATCAGAACTAAGGAT	10925-10948	
13	forward	ATTCCGAGAAGAAGGAGCAAGCTT	10821-10844	979
	reverse	CAAATTGTAGGACAGATACAATAA	11777-11800	
14	forward	GATTCAAGGTTACTTGGATTCCAC	11677-11700	1013
	reverse	ATCTCTAGCAAACTGAGTGGTTCA	12670-12693	
15	forward	CACTTGCACACAGGCTCGTCTTGC	12551-12574	949
	reverse	AGATAAAACAAATTACTAGGGTAG	13477-13500	
16	forward	CAGGCTTTAACTACACATCTACTC	13373-13396	917
	reverse	GATTAGTAAGCCACCTGGCTTTAG	14267-14290	
17	forward	GTTCATGTGGATTTGGAGGGTGTg	14177-14200	910
	reverse	TGATAGGCTCGATTTAACAATATG	15067-15090	
18	forward	AGATCCTTAAACTATCCCCCAACAG	14979-15003	405
	reverse	ACCAAGGGGAGAAAGTAAAATCAA	15361-15384	

Mumps whole genome sequence 9 fragment protocol primer scheme

Fragment	Primer name	Sequence 5'- 3'	Position	Product size
1	forward	ACCAAGGGGAAAATGAAGATGGGA	1-24	2090
	reverse	ACTTGCTCAACGAGTTGGTTCCCT	2067-2089	
2	forward	AGTCGGTACAGTCCTAGATGTCCA	1159-11182	2810
	reverse	CGGATGCAATGCACCCTTCTCCAT	3946-3969	
3	forward	ATCTAGATTAGTGAGAGCAGTTCA	3851-3874	1841
	reverse	CGGCATTTTGGAGGGATGCATTAA	5912-5935	
4	forward	CATCACTTATGCTGAGAACCTTAC	5826-5849	1575
	reverse	GTTGAGTTGAGACGTAACAGTACA	7377-7400	
5	forward	GGTATCCTATGTTCAAAACCTTAA	7281-7304	2086
	reverse	ACAACAGGATCACCATACTGTAAT	9343-9366	
6	forward	ATTCTCCTAACATTGGTTGATGAC	9251-9274	1698
	reverse	TATGAAGAAATCAGAACTAAGGAT	10925-10948	
7	forward	ATTCCGAGAAGAAGGAGCAAGCTT	10821-10844	1873
	reverse	ATCTCTAGCAAACTGAGTGGTTCA	12670-12693	
8	forward	CACTTGCACACAGGCTCGTCTTGC	12551-12574	1740
	reverse	GATTAGTAAGCCACCTGGCTTTAG	14267-14290	
9	forward	CCCAAGTTTGTTGATGACGGCTGA	13708-13731	1679
	reverse	ACCAAGGGGAGAAAGTAAAATCAA	15361-15384	

Supplementary Dataset 3

List of Genbank Identifiers and WHO names for Ontario cases

Case ID	Genotype	Genbank ID	WHO sample name	Isolation source
S1	G	MK033756	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/08.17/1[G], genomic sequence	buccal swab

S2	G	MK033765	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/1[G], genomic sequence	buccal swab
S3	G	MK033766	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/04.17/1[G], genomic sequence	buccal swab
S4	G	MK033767	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/06.17/1[G], genomic sequence	buccal swab
S5	G	MK033768	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/06.17/2[G], genomic sequence	buccal swab
S6	G	MK033769	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/07.17/1[G], genomic sequence	buccal swab
S7	G	MK033770	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/08.17/2[G], genomic sequence	buccal swab
S8	G	MK033771	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/2[G], genomic sequence	buccal swab
S9	G	MK033772	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/3[G], genomic sequence	buccal swab
S10	G	MK033747	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/4[G], genomic sequence	buccal swab
S11	C	MK033748	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/5[C], genomic sequence	buccal swab
S12	G	MK033749	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/6[G], genomic sequence	buccal swab
S14	G	MK033750	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/7[G], genomic sequence	buccal swab
S15	G	MK033751	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/8[G], genomic sequence	buccal swab
S16	G	MK033752	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/9[G], genomic sequence	buccal swab
S17	G	MK033753	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/09.17/10[G], genomic sequence	buccal swab
S18	G	MK033754	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/06.17/3[G], genomic sequence	buccal swab
S19	G	MK033755	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/06.17/4[G], genomic sequence	buccal swab
S20	G	MK033757	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/08.17/3[G], genomic sequence	buccal swab
S21	G	MK033758	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/12.17/1[G], genomic sequence	buccal swab
S22	G	MK033759	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/12.17/2[G], genomic sequence	buccal swab
S23	G	MK033760	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/12.17/3[G], genomic sequence	buccal swab
S24	G	MK033761	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/12.17/4[G], genomic sequence	buccal swab
S25	G	MK033762	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/12.17/5[G], genomic sequence	buccal swab
S26	G	MK033763	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/13.17/1[G], genomic sequence	buccal swab
S27	G	MK033764	[organism=Mumps rubulavirus] strain MuVi/Ontario.CAN/13.17/2[G], genomic sequence	buccal swab

Chapter 2

Pseudomonas aeruginosa Strain-sharing in Early Infection Among Children With Cystic Fibrosis.

Clin Infect Dis. 2021 Nov 2;73(9):e2521-e2528. doi: 10.1093/cid/ciaa788. PMID: 32544950; PMCID: PMC8563227. Impact Factor of Clinical Infectious Diseases in 2021: 20.99

https://pubmed.ncbi.nlm.nih.gov/32544950/

Contribution of author PS to manuscript: Design, plan and direct study, environmental sampling and bacterial cultures, sequencing, bioinformatics analysis, epidemiological investigations, main analyses and drafting manuscript, data sharing via NCBI SRA.

AUTHORS AND AFFILIATIONS

Patrick J Stapleton1,2,4, Conrad Izydorcyzk2, Shawn Clark2, Ana Blanchard3, Pauline W Wang5, Yvonne Yau1,4*, Valerie Waters1,3*, David S Guttman2,5,* #

1. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada.

2. Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada.

3. Division of Infectious Diseases, The Hospital for Sick Children, Toronto, Ontario, Canada.

4. Division of Microbiology, The Hospital for Sick Children, Toronto, Ontario, Canada.

5. Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada

*shared senior authorship

Corresponding author:

SUMMARY

In this study, 41% of CF children shared *Pseudomonas aeruginosa* strains with other children. In approximately a third of patients with shared strains, epidemiologic links were identified suggesting that patient to patient transmission of *P. aeruginosa* strains may have occurred.

ABSTRACT

Background: We previously identified *Pseudomonas aeruginosa* isolates with characteristics typical of chronic infection in some early infections in children with Cystic Fibrosis (CF), suggesting these isolates may have been acquired from other patients. Our objective was to define the extent of P. aeruginosa strain sharing in early CF infections and its impact on antibiotic eradication treatment failure rates.

Methods: We performed whole genome sequencing on isolates from early pediatric CF pulmonary infections and from comparator groups in the same hospital: chronic CF infection, sink drains, sterile site infections and asymptomatic carriage. Univariate logistic regression was used to assess factors associated with treatment failure.

Results: In this retrospective observational study, 1,029 isolates were sequenced. The CF clones Strain B and Clone C were present. In 70 CF patients with early infections, 14 shared strains infected 29 (41%) patients over five years; 16% (n=14) of infections had mixed-strains. In the 70 children, approximately one third of shared strain infections were likely due to patient-to-patient transmission. Mixed-strain infections were associated with strain sharing (odds ratio 8.50; 95% confidence interval 2.2 - 33.4, P = 0.002). Strain sharing was not associated with antibiotic eradication treatment failure; however, nosocomial strain transmission was associated with establishment of chronic infection in a CF sibling pair.

Conclusions: Although early P. aeruginosa CF infection is thought to reflect acquisition of diverse strains from community reservoirs, we identified frequent early CF strain sharing which was associated with the presence of mixed-strains and instances of possible patient-to-patient transmission.

Key Words: Pseudomonas aeruginosa, cystic fibrosis, respiratory tract infections, whole genome sequencing, cross infection

INTRODUCTION

Early *Pseudomonas aeruginosa* (Pa) infection in individuals with cystic fibrosis (CF) is usually characterised by acquisition of diverse Pa strains from the environment [1]. These Pa strains undergo rapid diversification into a highly heterogeneous population that is adapted to the CF airways (e.g. acquisition of mucoid phenotype) [2]. We previously identified early Pa infection with mucoid isolates, suggesting patient to patient transmission of Pa, which prompted us to question the degree of genetic diversity and strain sharing within a population of CF children with initial Pa infection [3].

Although strain sharing and transmission of Pa between individuals has been described in adult and pediatric patients with chronic Pa infection, it is thought to be a very rare occurrence in early Pa infection [4]. Previous studies demonstrating shared Pa strains between CF children, which may be due to either patient to patient transmission or transmission from a common environmental source, have been limited by the use of lower resolution typing techniques such as Pulse Field Gel Electrophoresis (PFGE) or Multi Locus Sequence Typing (MLST) [5-8]. The advent of newer molecular typing techniques, in particular whole genome sequencing (WGS), has improved our ability to distinguish between nearly identical strains. The one study that has used WGS to examine the genetic relatedness of early Pa isolates from children with CF identified strain sharing in approximately 15% of patients, with epidemiologic data supporting patient to patient transmission

[9]. This study was limited by a small sample size (35 patients) and limited depth of sampling (one isolate per clinical specimen). Analysis of multiple bacterial isolates per clinical sample is critical when investigating strain relatedness from infections with intra-patient pathogen subpopulations (i.e. that have significant intra-host bacterial diversity), such as Pa infections in CF patients. Furthermore, comparator collections of Pa isolates, such as from the environment and non-CF populations, are also helpful when evaluating the degree of genetic relatedness between strains and are missing in earlier studies.

To address this gap in knowledge, we performed a 5 year retrospective observational cohort study of all children with CF and new-onset Pa infection followed at the Hospital for Sick Children (Toronto, Canada) to determine the extent of strain sharing, the potential for patient-to-patient transmission and the impact on antibiotic eradication treatment. Using WGS of multiple Pa isolates from each sputum sample, we aimed to characterise the level of Pa strain sharing in our pediatric CF cohort, the largest such studied cohort to date. Additionally, we compared isolates from early Pa infection in children with CF to isolates from chronic Pa infection in CF patients, isolates from the hospital environment, and clinical isolates from invasive infection or asymptomatic intestinal carriage in non-CF patients. From these analyses, we identified that a surprising proportion of CF children (41%) share Pa strains with other children, challenging previous beliefs that early Pa infection occurs due to acquisition of genetically diverse strains from the environment.

MATERIALS AND METHODS

Early Pa Cohort

This was a retrospective observational cohort study. The primary study population consisted of the Early Cohort: all children with CF from SickKids (Toronto, Canada) with at least one new-onset Pa infection between 2011 and 2015, and who underwent Antibiotic Eradication Therapy (AET) [10]. Pa isolates from the time of initial infection were recovered from the CF Sputum Biobank, which has been prospectively storing frozen sputum samples from SickKids CF patients since 2011 [3].

Comparator Cohorts

Comparator cohorts included: [1] The Chronic Cohort consisted of a CF patient population chronically infected with Pa, and included 24 children who were enrolled in a randomized control trial of CF Pa biofilm antimicrobial susceptibility testing at SickKids between January 2009 and September 2013 [11]. One Pa isolate from up to three morphotypes was prospectively isolated and stored from each sputum culture. We required children in the Chronic Cohort to have two or more

positive cultures between 2011 and 2013. Two patients in the Early Cohort, who subsequently developed chronic infection, were also included in the Chronic Cohort. [2] An Environmental Cohort consisted of isolates cultured from sink drain sampling performed in CF clinical areas in 2018 (See Supplementary Methods in the online data supplement for details). [3] The Carriage Cohort consisted of isolates cultured from stool or rectal swabs during an inpatient point prevalence screen for carbapenemase producing Enterobacteriaceae performed in 2017 (no CF patients). And [4] a Sterile Site Cohort consisted of Pa cultured from children without CF who had invasive Pa infection (includes one isolate of each morphotype of Pa from each blood or sterile body fluid sample) at the hospital between 2000 and 2017. We have included all the isolate numbers for all cohorts with STs, collection dates and shared strain number in Supplementary Table 1.

Whole Genome Sequencing and Analysis

All Pa isolates were sequenced and analysed as described in Supplemental Methods.

Definitions of Strain Sharing

We defined a shared strain as a set of identical isolates found to infect multiple individuals with CF [4] . We chose a cut-off of four or fewer SNP differences between isolates to define a strain based on our observations that intra-patient sequence diversity from new-onset infections was up to four SNPs (excluding outliers >50 SNPs), and that three to four SNPs per year were accumulated in patients who experienced recurrent new-onset infection years apart (Supplementary Methods) [12]. An exception to the SNP cut-off was made for two strains with complex phylogenetic relationships and hypermutator genotype Pa, which substantially increased the SNP distance between isolates.

Mixed-strain infection was defined as the presence of two or more Pa strains from different clonal complexes (i.e. differing in at least three of seven MLST alleles) [13] in a new-onset sputum sample. Superinfection occurred when a chronically Pa infected individual was co-infected with a different strain of Pa, at a later point in time, that may or may not have supplanted the original strain.

Statistical Analysis

Univariate logistic regressions were used to assess associations between mixed-strain infection and AET failure, strain sharing and AET failure, and mixed-strain infection and strain sharing. All statistical analyses were done using SAS 9.04.01 (SAS Institute, USA). The study was approved by the Research Ethics Board of the Hospital for Sick Children (#1000061322).

RESULTS

Pa Sequencing Results

A total of 435 Early Cohort isolates were cultured and sequenced from 87 new-onset episodes, in 70 patients (in those with repeated new-onset episodes, 35% (5 of 14 patients) had re-infection with a Pa strain of the same ST). A median of four isolates (IQR 3-8) and two morphotypes (IQR 1-2) were sequenced per episode. Pa could not be recovered from frozen sputum in 41 eligible episodes (32%), thus they were excluded from the analysis. The clinical characteristics of patients included in the Early Cohort are shown in Table 1, and were similar to patients excluded from the Early Cohort (Supplementary Table 2).

Table 1: Clinical Characteristics of Patients in the Early Cohort (N = 70)

Condition	Value
Age at Commencement of study, mean (SD)	9.7 (3.5)
Female, n (%)	36 (51%)
Mutation Class, n (%)	
Class I-III	68 (97%)
Class IV-V	2 (3%)
Complications	
CF Related Diabetes, n (%)	3 (4%)
Pancreatic insufficiency, n (%)	68 (97%)
Baseline Forced Expiratory Volume in 1 second	
L, mean (SD)	1.72 (0.77)
% pred, mean (SD)	89.7 (19.8)
Body Mass Index, mean (SD)	39.0 (29.8)

Supplementary Table 2. Demographics of excluded patients

		New Onset (sequenced) N=70	New Onset (not sequenced) N=23
Age at Commencement of study, mean (SD) (range)		9.7 (3.5) (0.1, 17.1)	8.1 (4.3) (0.1, 15.6)
Female, n (%)		36 (51%)	10 (43%)
Homozygous ΔF508, n (%)		35 (50%)	10 (43%)
Mutation Class, n (%)			
	Class I-III	68 (97%)	21 (91%)
	Class IV-V	2 (3%)	2 (9%)
Complications			
	CFRD, n (%)	3 (4%)	1 (4%)
	Pancreatic insufficiency, n (%)	68 (97%)	20 (87%)
FEV1 at commencement of study			
	L, mean (SD) (range)	1.72 (0.77) (0.41, 4.28)	1.56 (0.60) (0.96, 3.16)
	% pred, mean (SD) (range)	89.7 (19.8) (38.2, 130.4)	95.0 (25.9) (42.0, 137.3)
BMI			
	kg/m2, mean (SD) (range)	16.7 (2.8) (11.0, 25.0)	16.6 (2.1) (12.3, 20.3)
	z-score, mean (SD) (range)	-0.42 (1.06) (-3.56, 1.57)	-0.13 (0.96) (-1.64, 1.70)
	centile, mean (SD) (range)	39.0 (29.8) (0.1, 94.2)	46.1 (29.9) (5.0, 95.6)

FEV1: forced expiratory volume in 1 second
BMI: body mass index
CFRD: cystic fibrosis related diabetes

In the Chronic Cohort, 331 isolates were sequenced from 24 patients enrolled in the biofilm antimicrobial susceptibility trial (including two siblings from the Early Cohort), with a mean of 14 isolates sequenced per patient. 20 Environmental Cohort isolates were collected and sequenced from five sinks installed in separate patient rooms on the CF inpatient ward (2-6 isolates per sink); no Pa was recovered from sink drain cultures taken from the CF clinic or PFT laboratory. 22 Carriage Cohort isolates were sequenced (one per patient), as well as 221 Sterile Site Cohort isolates from 201 patients. In total, 1,029 SickKids Pa sequences and 81 reference genomes were included in the first pass analysis (Figure 1).

FIGURE 1

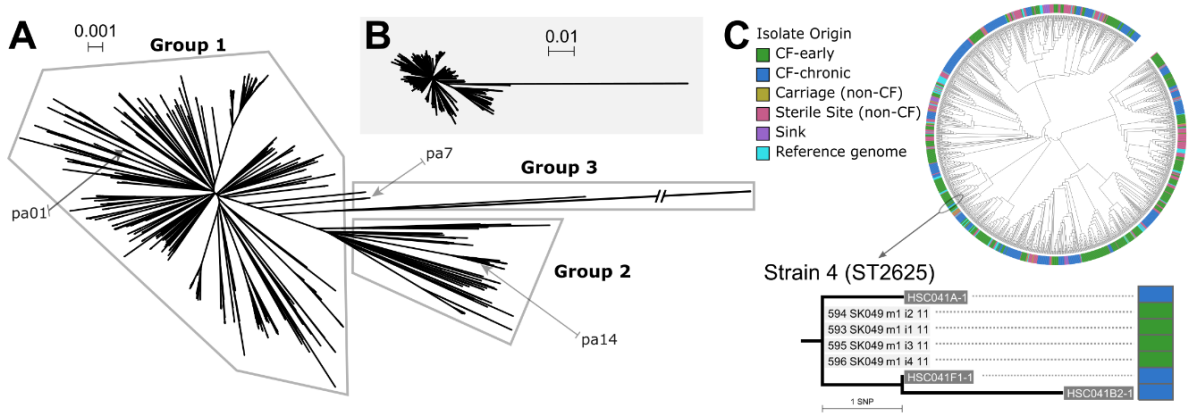


Figure 1. (A) Unrooted Neighbour-Joining mashtree of 1,029 de-novo assemblies from SickKids Pa isolates and 81 complete Pa reference genomes. Most sequences cluster in either Group 1, which contains reference strain PA01, or Group 2, which contains reference strain PA14. This population structure is similar to previous reports of phylogenetic analyses of diverse Pa isolates (39, 40). The remaining outlier sequences cluster in group 3 together with the Pa7 reference genome. Pa7 is a phylogenetic outlier which diverged early in evolutionary history from other Pa lineages. Note that a branch leading to a sterile site isolate (STS031) in group 3 is truncated. (B) Neighbour-joining tree from panel A with the group 3 branch shown to scale. (C) A circular mashtree cladogram (branch lengths ignored). Isolates from the different cohorts (CF and non-CF patients and hospital sinks) are dispersed across the tree. A group of closely related sequences from the Early and Chronic Cohorts is circled and shown in an enlarged phylogram (branch lengths proportional to evolutionary distance) on the right. Sequences from Early Cohort case 49 (SK049) and Chronic Cohort case 14 (HSC014) appear highly related. In fact, some sequences from SK049 appear as closely related to HSC014 sequences as they are to other SK049 sequences. This group was therefore subjected to further analysis by mapping sequencing reads to the most closely related reference genome, in this case PAER_119, to generate a pairwise SNP distance matrix and ML tree, and thereby determine if these sequences represent a shared strain. **Isolate coding system**: **Early Cohort**: 594 (isolate number) SK (Early Cohort) 049 (patient number) m1(morphotype number) i2 (isolate number) 11 (collection year 20**11**); **Chronic Cohort**: HSC (Chronic Cohort) 041 (patient number) B2 (visit type and number; A: enrolment, B: baseline, E: exacerbation, F: follow up) -1(isolate number); **Environmental Cohort**: ENV (Environmental Cohort) 64 (room number where isolate collected)-3 (isolate number).

Strain B (ST439), a well-recognized clone accounting for 7% of isolates from adult CF patients in Ontario (19), was identified in one Early Cohort patient. Clone C (ST17), which is widely disseminated in the general environment, was detected in eight CF patients (five Early Cohort and three Chronic

Cohort patients) however, using WGS, only 4 patients were classified as harbouring a shared strain (see Supplementary Table 1 for more detail).

Mixed-strains (up to four strains) were present in 14 new-onset infections (16%), from 13 Early Cohort patients and could not be reliably predicated from the morphotypic appearance of isolates alone. However, we found an association between the presence of multiple morphotypes in a sample and mixed-strain infection (OR 2.18, 95%CI 1.14-4.20, p=0.02).

Shared Strains in the Early Cohort

The majority of shared strains were found in the Early Cohort. 78 different Pa strains from 60 STs were identified in the Early Cohort (8 of these STs have previously been reported in other Early CF cohorts [5]). Of these, 64 strains (82%) were recovered only once, while 14 were shared among patients. These 14 shared strains were found among 29 (41%) patients, with shared strains found in 33 (38%) of the new-onset episodes. 11 of the 14 shared strains were shared only among Early Cohort patients. Of these 11 strains, eight (strains 14, 41, 43, 53, 55, 59, 60, 61) were found in patients pairs, one (strain 45) was shared among three patients, and two (strains 2, 51) were shared among four patients. One strain (strain 51) included a patient with isolates that had a hypermutator genotype (Figure 2).

FIGURE 2

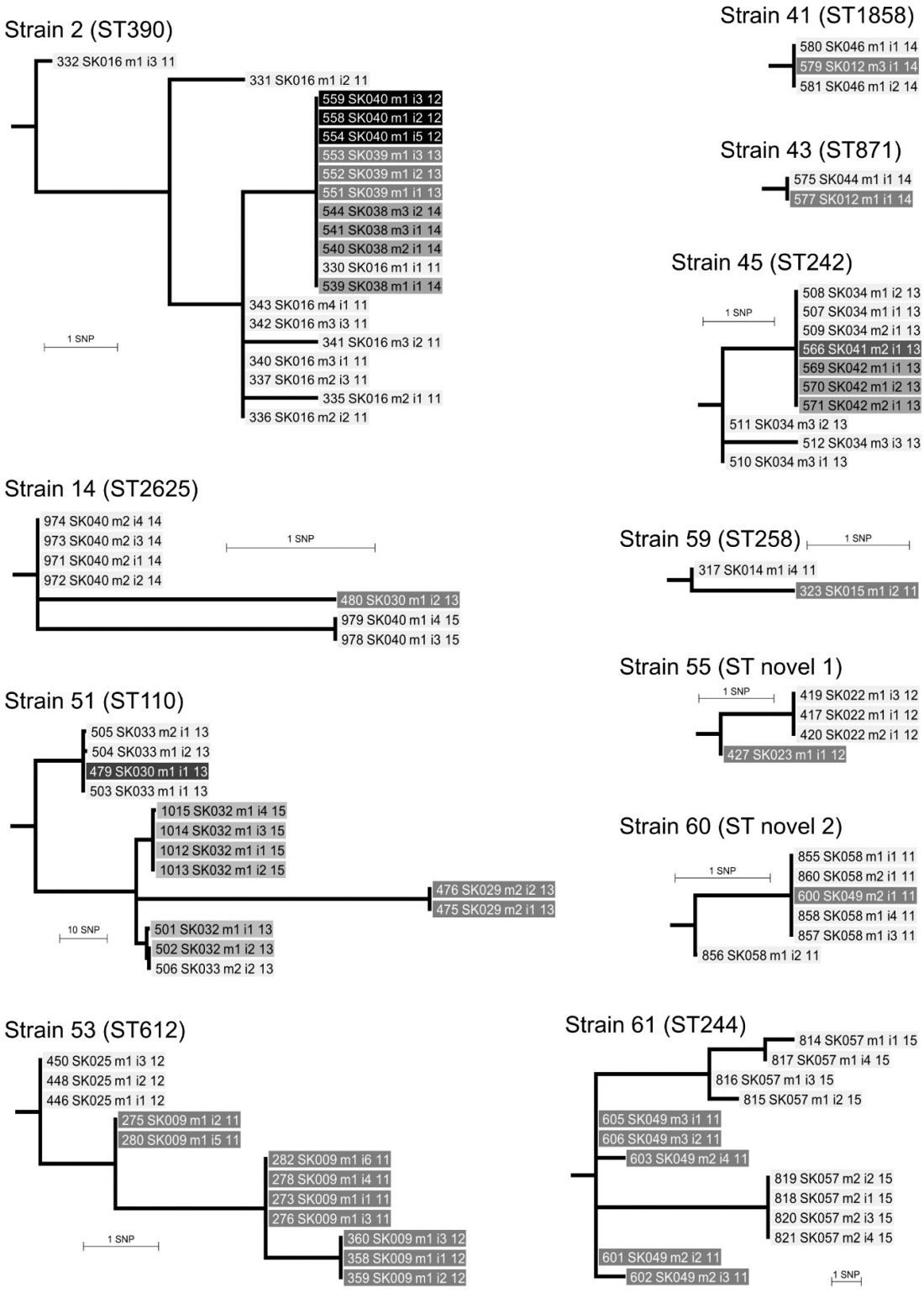


Figure 2. Maximum likelihood trees for 11 shared strains in Early Cohort patients rooted on reference genome (not shown) used for SNP calling as detailed in **Supplementary Table 1**. Strain 51 contains a

long branch to two hypermutator isolates from case SK029 (475 and 476). These isolates are 89 to 90 SNPs different from isolates from infections which occurred in one case (SK032) in 2013 and 2015. We identified a mutS DNA mismatch repair gene frameshift mutation at codon 333 (of total 2568), resulting in non-functional MutS in isolates 475 and 476 only, and so consider them part of the shared strain. The SK032 isolates are paraphyletic with respect to 475 and 476, which suggests the direction of Pa transmission was from SK032 to SK029.

Overall, we found few epidemiological links for the 11 Early Cohort shared strains. No siblings were affected, no social links outside hospital were identified, and cases were not co-infected with similar pathogens other than Pa. One shared strain may have been transmitted between two patients at a clinic visit in 2011. Another was potentially transmitted between two patients during a same-day PFT laboratory visit in 2012; however two other patients who acquired the same strain a year later had no epidemiological links. Nine shared strains had no epidemiological link between any patients: the gap between detection of infection ranged from two days to 14 months (under six months for eight strains). All infections were detected between late March and mid-November and only one shared strain was newly identified in a patient after 2014.

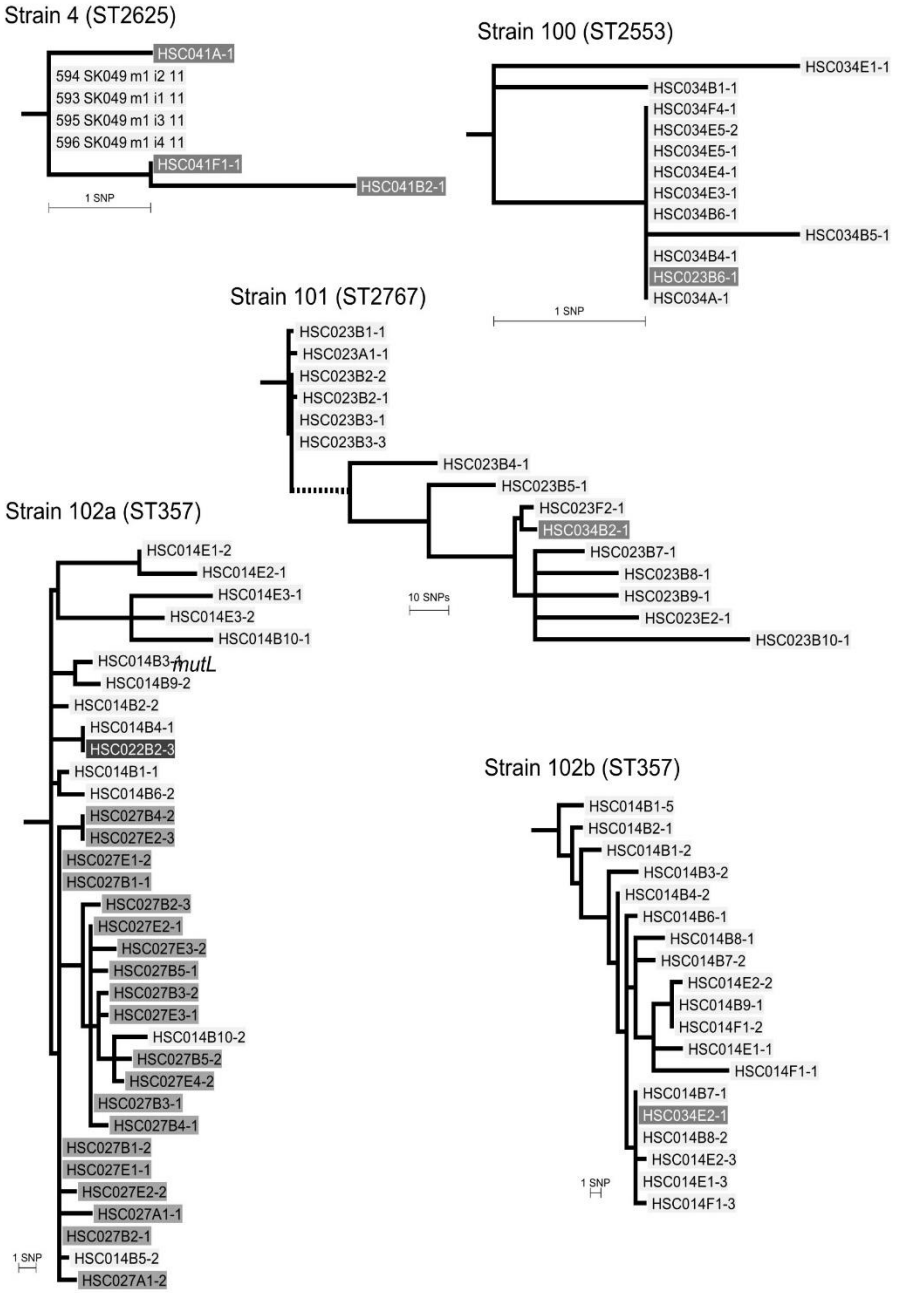
Mixed-strain infection was associated with strain sharing (OR 8.50, 95% CI 2.2 - 33.4, p = 0.002). There was no association between shared strain infections and AET failure, or mixed-strain infection and AET failure, using Pa sputum culture positivity at five weeks or three months after time of initial Pa infection to define AET failure, or using development of chronic infection [14] after 18 months.

Shared Strains in the Chronic Cohort

Overall, 24 Pa strains were identified in the Chronic Cohort; 17 (71%) strains recovered from 15 patients were unique, while four strains recovered from four patients were shared with other cohorts and are discussed separately.

The remaining three strains were shared by five patients (Supplementary Figure E1). Strain 102 was found in a sibling pair consistently throughout their longitudinal sampling. It was also briefly present in two other chronically infected non-sibling patients, in whom it caused superinfection. One non-sibling patient was at baseline status at the time (isolate HSC022B2-3) and the other non-sibling was experiencing an exacerbation (isolate HSC034E2-1). Strain 102 did not supplant the original chronic strain for the non-sibling patient in whom follow up samples were obtained. Two other shared strains (strains 100, 101) involved pairs of non-sibling patients who attended clinic on the same day and superinfected each other; the superinfecting strains did not supplant the original patient strain.

SUPPLEMENTARY FIGURE E1



Supplementary Figure E1. Maximum likelihood trees for shared stains in the chronic cohort. Strain 04 is shared between one patient in the Early Cohort and one in the Chronic Cohort. Strain 100, 101 and 102 are shared between chronic patients only. Note that strain 101 has eight isolates from case HSC023 and one isolate from case HSC034 that have a hypermutator genotype. The hypermutator clade is separated by a dashed lined from six non-hypermutator HSC023 isolates collected earlier in the study. All hypermutator isolates have a frameshift mutation in the mutL DNA mismatch repair gene (codon 593/633). Although the distance between isolates HSC023F2-1 and HSC034B2-1 is seven SNPs, the two patients are still considered to have shared a strain.

Shared Strains between Cohorts

Three strains were shared between patients in the Early and Chronic Cohorts. For two of these strains, there was evidence of possible patient-to-patient transmission. Genomic and epidemiological analyses demonstrate that a strain (strain 01) was shared between a chronic patient and three patients in the Early Cohort on the inpatient CF ward (Figure 3A, B).

FIGURE 3

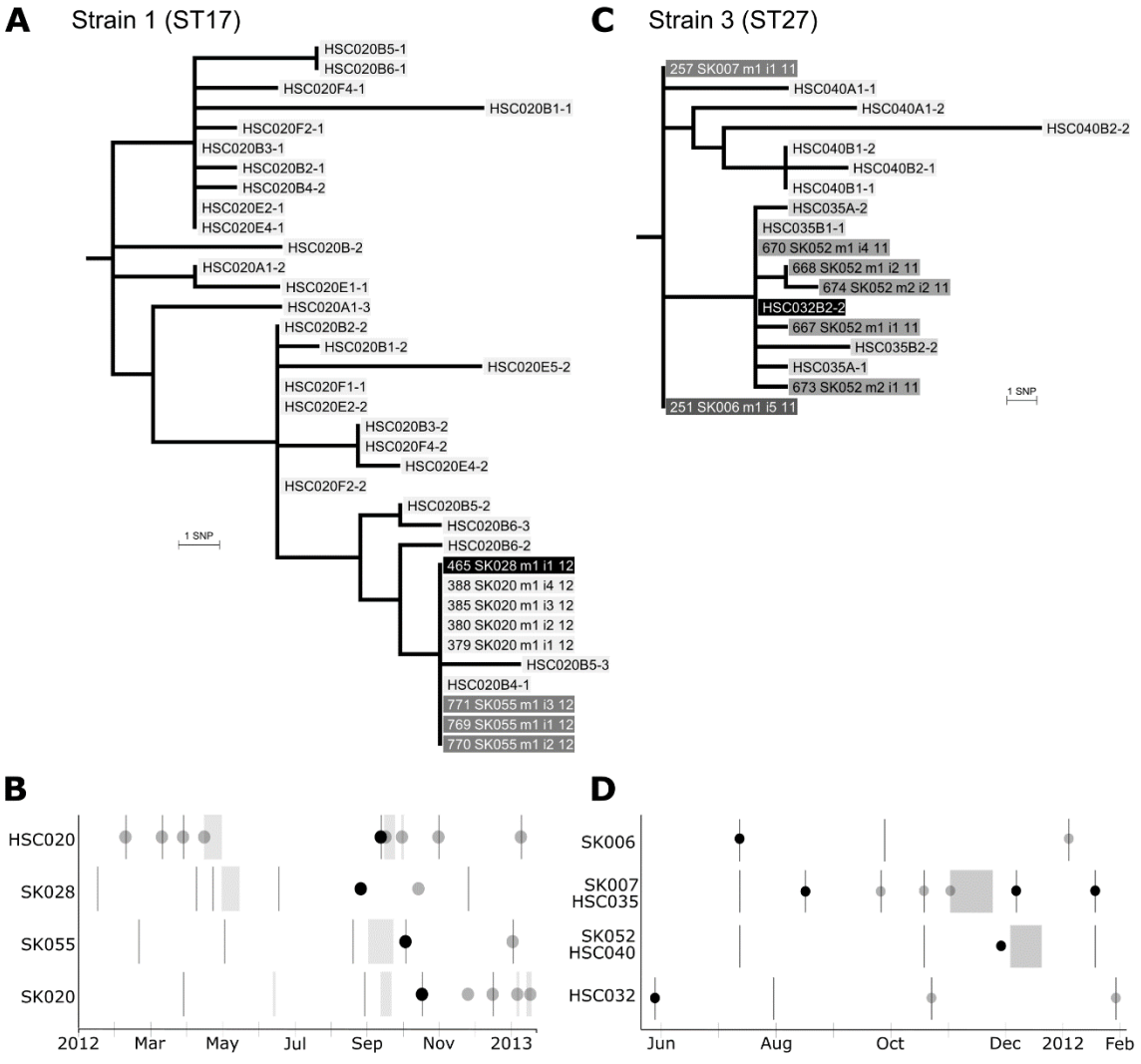


Figure 3. (A) Maximum likelihood tree of strain 1 isolates rooted on reference genome (not shown) used for SNP calling as detailed in **Supplementary Table 1.** Isolates of ST17 Pa from three Early Cohort cases (SK020, 028 and 055), are identical to each other, and to one isolate from a Chronic Cohort case (HSC020). The chronic case exhibits significant intra-patient Pa sequence diversity from the time of enrolment (sample A), through baseline pulmonary status assessments (B), pulmonary exacerbations (E), and follow-up visits post-exacerbation (F). However, only one HSC020 isolate, B4-1, is identical to the Early patient isolates, which allow us to date the approximate period that strain transmission

occurred to when this isolate was collected, in September 2012. (B) Timeline of Pa infections and hospital visits. Each row represents one patient. Dark vertical lines: Clinic or PFT laboratory visits. Light grey boxes: Inpatient Admissions. Dark circles: Sputum was positive for Pa and isolates were sequenced, with presence of shared strain confirmed. Light grey circles: Sputum positive for Pa but isolates were not sequenced. Sibling cases SK020 and 055 were admitted to the CF ward in September 2012, and overlapped for 6 days with the stay of the Chronic Cohort patient. Shortly thereafter, SK020 and 055 had new-onset infection with the chronic patient's strain. Another Early Cohort patient, SK028, was infected in August 2012, several months after an admission to the ward that commenced on the day after the chronic patient had been discharged. (C) Maximum likelihood tree of strain 3 isolates rooted on reference genome. Isolates from Early Cohort cases SK006 and 007 are identical, and separated by four SNPs from an isolate from SK052 (the sibling of SK007) and an isolate from Chronic Cohort case HSC032. The siblings developed chronic infection with this Pa strain and were subsequently enrolled in the biofilm trial. All subsequent chronic isolates from the siblings were with the same strain, and were sequenced under the study code numbers HSC040 (for SK0007) and HSC035 (for SK052). (D) Timeline of infections and hospital visits as in panel B. Timeline of Pa infections and hospital visits. Each row represents one patient (SK007 and SK052 also have chronic cohort study codes HSC035 and HSC040). SK006 attended clinic in July 2011 and had a new-onset infection detected. On the same day, the siblings also attended clinic; SK007 developed infection with the shared strain a month later, while SK052 developed infection with the shared strain in December. SK052 could have acquired the strain from SK006 at the clinic visit, or from their sibling subsequently. In May 2012, HSC032 was superinfected with the shared strain, but had no overlapping visits with the other patients in the prior 6 months. Afterwards, HSC032 reverted to their pre-existing Pa strain.

Another shared strain (strain 03) may have been transmitted in CF clinic from an Early Cohort patient to a pair of Early Cohort siblings, with the siblings then becoming chronically infected. Additionally, a patient in the Chronic Cohort became superinfected with this strain around the same time but had no clear epidemiological link (Figure 3C, D).

The third strain (strain 04) was shared between an Early Cohort case and a Chronic Cohort case with no epidemiological link.

The final strain shared between cohorts (strain 200) consisted of four isolates collected from the hand wash sink of a room on the CF ward in 2018 and isolates from one patient each in the Chronic and Sterile Site Cohorts (Figure 4).

FIGURE 4

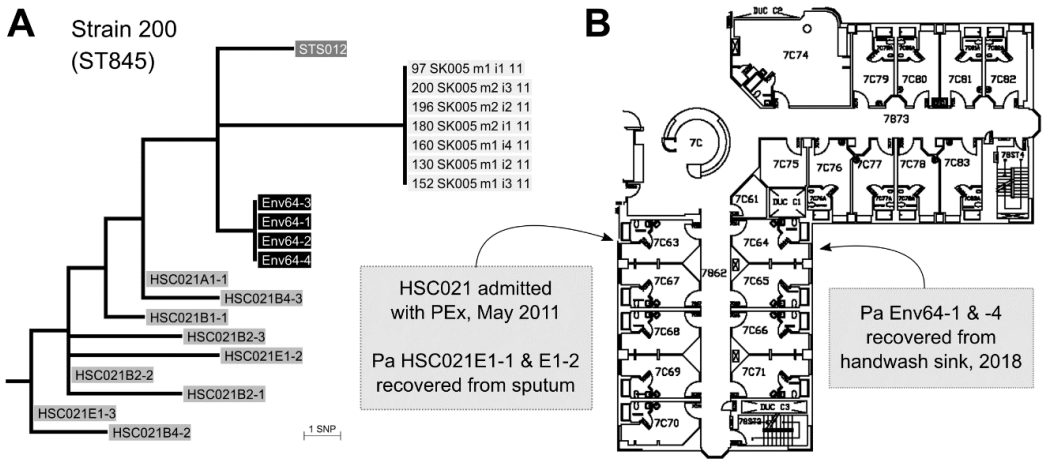


Figure 4. (A) Maximum likelihood tree of isolates rooted on reference genome (not shown) used for SNP calling as detailed in **Supplementary Table 1.** One isolate from chronic case HSC021 (HSC021-A1-1, collected at trial enrolment in 2010) is three SNPs from isolates from a sink in CF ward room 64 (Env64 1-4). The other nine isolates from HSC020 (collected 2010-2012) are five to nine SNPs distant. An isolate from a sterile site infection (STS012) is also three SNPs from the sink isolates, and four SNPs from HSC021-A1-1. Isolates from Early Cohort case SK005 were included as they appeared closely related to the other isolates on the mashtree, but we found at least six SNPs differences from all other isolates, so we deemed the SK005 isolates to be unique, i.e. they are not part of the shared strain. (B) Floor map of the inpatient CF ward with locations of sink samples and inpatient admissions. Case HSC021 was admitted in May 2011 to a room opposite room 64, where the sink isolates were retrieved seven years later. This suggests isolates from HSC021 were introduced to the ward environment during the 2011 admission, and persisted in the ward sinks. The STS012 isolate came from a post-operative blood culture from a 2001 patient (non-CF) with no epidemiological links to this ward.

The CF patient was hospitalized in the room across the hall in 2011. However, the Sterile Site isolate came from a blood culture taken from a non-CF patient hospitalized 16 years earlier on a different ward.

A summary of the degree of strain sharing between cohorts is shown in Supplementary Figure E2.

SUPPLEMENTARY FIGURE E2

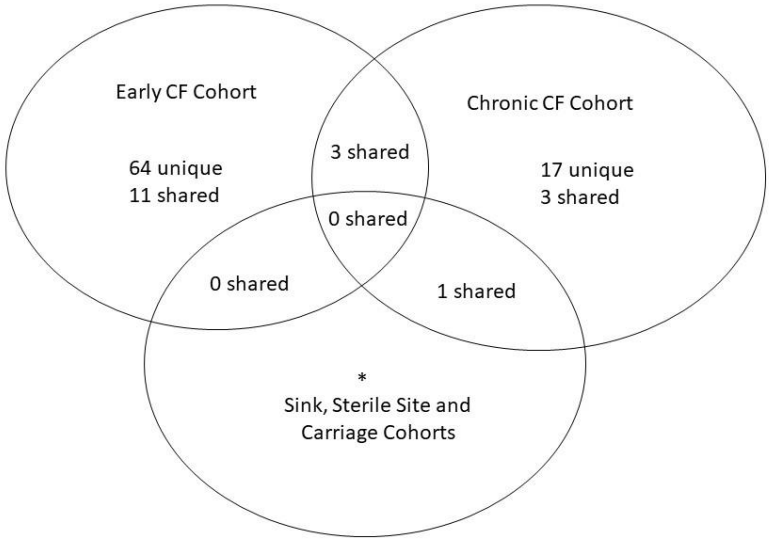


Figure E2. Sharing of strains within and between different cohorts. The majority of shared strains were shared between patients in the Early Cohort, or between Early and Chronic Cohort patients. The strain shared between CF and non-CF cohorts consisted of isolates from a Chronic Cohort patient, a CF ward sink and a sterile site infection. * Based on review of the mashtree, all 24 carriage isolates from individual patients were unique, as were isolates from 169 (84%) patients with sterile site infection. Confirmation of instances of strain sharing amongst the remaining 32 patients with sterile site infection (non-CF) was beyond the scope of this study. Pa isolates from the five CF ward sinks were subject to in-depth investigation of strain relatedness: [1] a shared Pa strain was found in sinks from two inpatient rooms located on different wings of the ward, [2] one sink strain (mentioned above) was shared with a chronic and sterile site patient, and [3] the remaining two inpatient sinks harboured unique strains.

DISCUSSION

Our study demonstrated that strain sharing occurred in 41% of patients with early Pa infection. Mixed-strain infection was relatively frequent in the Early Cohort (16% of episodes) and was strongly associated with strain sharing. Epidemiological links between patients were found for four of 29 shared strain (14%) infected patients; these links were comprised of overlapping ward, clinic or PFT lab visits, and occasionally sibling relationships. Strain sharing was not associated with antibiotic eradication treatment failure; however, potential nosocomial strain transmission was associated with establishment of chronic infection in a CF sibling pair.

Previous studies have reported that between 0-62% of CF patients may share strains of Pa, although few studies have focused specifically on early Pa acquisition in children with CF [4]. Work by Kidd et al. demonstrated that the majority of Pa strains infecting CF children under the age of 5 years were unique and commonly found in different ecological settings, however, there was limited within patient sampling of Pa colonies [1, 5]. With low resolution typing techniques of single Pa colonies from respiratory tract specimens and the absence of proven epidemiological links, it is reasonable to conclude that genotypically similar strains were acquired independently from the general

environment. In a smaller, retrospective study performed at the Copenhagen CF Centre, 474 isolates of Pa sampled from the airways of 34 children and young individuals with CF were genotyped by WGS; in only a few cases (n=5) were strains closely related, differing by only a few SNPs, suggesting patient to patient transmission, supported by epidemiologic links [9]. Although this degree of patient to patient transmission (approximately 15%) is similar to that in our study, we found a significantly higher percentage of strain sharing overall (41%).

Our shared strain infections occurred in a pediatric CF centre which adheres to national CF IPAC recommendations [15]: no designated waiting room, environmental cleaning of pulmonary function testing lab (not under negative pressure) after every patient, standard precautions and single room isolation in all CF care areas from 2011 to 2014, and additional contact precautions (gloving and gowning for staff) from 2014 onwards. Given our study occurred from 2011-2015, we cannot judge the impact of such a change in infection control practices on strain sharing due to the unequal follow up times pre and post 2014. Due to the retrospective nature of the study, we could not establish how strains were shared during overlapping hospital visits, although cough aerosol generation and fomites have been implicated in previous studies [16, 17]. Hospital water distribution networks, including sink outlets, may be the source of outbreaks of waterborne bacteria. Although we did not sample water distribution networks, we did not find evidence of widespread dissemination of a single Pa strain, which would be typical of point source transmission. Of note, we did not find Pa in CF clinic sinks. A shared Pa strain was present in a CF ward sink, but it is possible this represents unidirectional contamination of sink drains by patients. Shared strains may have come from other unsampled reservoirs in CF clinical areas. If a community reservoir was the main source, we would expect considerable sharing between cohorts, however, only one Pa strain was shared between CF and non-CF patients. In fact, there was little to no strain sharing amongst the Environmental, Carriage and Sterile Site Cohorts, suggesting that inclusion in the CF cohort played a role in strain sharing. Movement of patients between CF centres has previously been shown to be a risk factor for the acquisition of shared Pa strain infections [18].

This study had several limitations as well as strengths. We were only able to recover Pa from two thirds of the frozen patient samples, limiting our study population. Additionally, comparator cohort isolates were obtained at different time periods to the Early Cohort, and fewer isolates per sample were sequenced. Hospital environmental sampling was performed three years after the last infections occurred, and therefore was restricted to potential long-term reservoirs, not surfaces or fomites that might be transiently contaminated. Although one strain was found shared between sink drain, CF and non-CF patients, the samples were collected 16 years apart, and there were insufficient environmental samples to put the SNP differences in context (e.g. to determine if sink

isolates have lower diversity and or mutation rates than clinical isolates). The main strength of this study was the depth of sampling in the early CF Cohort that allowed us to identify the frequent presence of nearly identical isolates shared between patients. Furthermore, we were able to set a clear definition of “shared strain” with a SNP threshold derived from intra-host isolate diversity observed with deep sampling. Although the term strain is frequently used to describe nearly identical CF Pa isolates, these shared strains may be more accurately referred to as clones, evolved from a common bacterial ancestor.

In conclusion, our study demonstrated that 41% of CF patients with early Pa infection had shared strains of which approximately a third were potentially associated with patient-to-patient transmission. Although strain sharing was not associated with failure of antibiotic eradication therapy, nosocomial transmission of a shared strain was associated with the establishment of chronic infection in a pair of CF siblings. Further studies are warranted to determine whether silent Pa strain sharing is common in other pediatric CF centres and how this may be prevented with enhanced infection prevention and control policies.

AUTHOR CONTRIBUTIONS

PJS, VW, YY and DSG conceived, planned and directed the study. PJS performed environmental sampling and bacterial cultures. SC and PWW sequenced the isolates. CI and PJS developed the methods for bioinformatics analyses. AB performed the statistical analyses and assisted PJS with epidemiological investigations. PJS performed the main analyses and wrote the manuscript. All authors discussed the results and provided critical feedback on the manuscript.

ACKNOWLEDGEMENTS

We thank the following for their expertise and assistance: Julio Diaz Caballero and Yunchen Gong for assistance with bioinformatics analyses, Michelle Klingel for Toronto CF Database support, Alvin Li for help with the sputum biobank cultures. We are indebted to the SickKids clinical microbiology laboratory, infection control and CF clinical care teams for help with Pa cultures and epidemiological investigations. Finally, we wish to express our sincere gratitude to the patients with CF, and their families, for their participation in the Toronto CF database and this research study.

FUNDING

This work was supported by a Collaborative Health Research Project grant awarded to DSG with joint funding provided by the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada (CP-151952). PJS and AC are in receipt of Cystic Fibrosis Canada post-doctoral research fellowships.

POTENTIAL CONFLICTS OF INTEREST

VW reports grants from Gilead and consultancy fees from Astra Zeneca, outside the submitted work. All other authors have no potential conflicts.

REFERENCES

1. Ranganathan SC, Skoric B, Ramsay KA, et al. Geographical differences in first acquisition of Pseudomonas aeruginosa in cystic fibrosis. Ann Am Thorac Soc 2013; 10(2): 108-14.

2. Hogardt M, Heesemann J. Adaptation of Pseudomonas aeruginosa during persistence in the cystic fibrosis lung. Int J Med Microbiol 2010; 300(8): 557-62.

3. Vidya P, Smith L, Beaudoin T, et al. Chronic infection phenotypes of Pseudomonas aeruginosa are associated with failure of eradication in children with cystic fibrosis. Eur J Clin Microbiol Infect Dis 2016; 35(1): 67-74.

4. Parkins MD, Somayaji R, Waters VJ. Epidemiology, Biology, and Impact of Clonal Pseudomonas aeruginosa Infections in Cystic Fibrosis. Clin Microbiol Rev 2018; 31(4).

5. Kidd TJ, Ramsay KA, Vidmar S, et al. Pseudomonas aeruginosa genotypes acquired by children with cystic fibrosis by age 5-years. J Cyst Fibros 2015; 14(3): 361-9.

6. Johansson E, Welinder-Olsson C, Gilljam M, Pourcel C, Lindblad A. Genotyping of Pseudomonas aeruginosa reveals high diversity, stability over time and good outcome of eradication. J Cyst Fibros 2015; 14(3): 353-60.

7. Hall AJ, Fothergill JL, McNamara PS, Southern KW, Winstanley C. Turnover of strains and intraclonal variation amongst Pseudomonas aeruginosa isolates from paediatric CF patients. Diagnostic microbiology and infectious disease 2014; 80(4): 324-6.

8. Speert DP, Campbell ME, Henry DA, et al. Epidemiology of Pseudomonas aeruginosa in cystic fibrosis in British Columbia, Canada. Am J Respir Crit Care Med 2002; 166(7): 988-93.

9. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet 2015; 47(1): 57-64.

10. Blanchard AC, Horton E, Stanojevic S, Taylor L, Waters V, Ratjen F. Effectiveness of a stepwise Pseudomonas aeruginosa eradication protocol in children with cystic fibrosis. J Cyst Fibros 2017; 16(3): 395-400.

11. Yau YC, Ratjen F, Tullis E, et al. Randomized controlled trial of biofilm antimicrobial susceptibility testing in cystic fibrosis patients. J Cyst Fibros 2015; 14(2): 262-6.

12. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome analysis of a transmissible lineage of pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS genetics 2013; 9(9): e1003741.

13. Middleton MA, Layeghifard M, Klingel M, et al. Epidemiology of Clonal Pseudomonas aeruginosa Infection in a Canadian Cystic Fibrosis Population. Annals of the American Thoracic Society 2018; 15(7): 827-36.

14. Lee TW, Brownlee KG, Conway SP, Denton M, Littlewood JM. Evaluation of a new definition for chronic Pseudomonas aeruginosa infection in cystic fibrosis patients. J Cyst Fibros 2003; 2(1): 29-34.

15. Saiman L, Siegel JD, LiPuma JJ, et al. Infection prevention and control guideline for cystic fibrosis: 2013 update. Infection control and hospital epidemiology 2014; 35 Suppl 1: S1-S67.

16. Wainwright CE, France MW, O'Rourke P, et al. Cough-generated aerosols of Pseudomonas aeruginosa and other Gram-negative bacteria from patients with cystic fibrosis. Thorax 2009; 64(11): 926-31.

17. Panagea S, Winstanley C, Walshaw MJ, Ledson MJ, Hart CA. Environmental contamination with an epidemic strain of Pseudomonas aeruginosa in a Liverpool cystic fibrosis centre, and study of its survival on dry surfaces. J Hosp Infect 2005; 59(2): 102-7.

18. Kidd TJ, Soares Magalhaes RJ, Paynter S, Bell SC. The social network of cystic fibrosis centre care and shared Pseudomonas aeruginosa strain infection: a cross-sectional analysis. Lancet Respir Med 2015; 3(8): 640-50.

SUPPLEMENTARY METHODS

Recovery of Isolates from the Sputum Biobank

Pseudomonas aeruginosa (Pa) positive sputum samples from pediatric cystic fibrosis (CF) patients with new-onset infections at the Hospital for Sick Children, Toronto, Canada were collected and frozen in a 1:1 ratio of sputum and sputolysin (Calbiochem, La Jolla, California). To recover isolates from the initial infection, we plated both undiluted and diluted (1:10 sputum/sputolysin to broth) frozen samples to MacConkey agar (Oxoid, Nepean, Ontario). Plates were incubated at 42°C to select for Pseudomonas aeruginosa for 72 hours. If no growth was observed, a second and (if required) third undiluted frozen sputum sample(s) (1:1 sputum to sputolysin) were cultured as described above. Putative Pa colonies were identified based on appearance and morphotypes and degree of growth recorded. The identity of one colony per morphotype was confirmed with MALDI-TOF (Bruker Biotyper) and four to eight colonies per morphotype were sub-cultured on Columbia blood agar with 5% sheep blood (Oxoid, Nepean, Ontario) overnight at 37°C. The isolate was stored at -80°C in 1 ml of 10% glycerol in water for later sequencing.

Environmental Sampling

We sampled clinical handwashing station and patient washroom sink trap biofilms in an effort to obtain hospital environment Pa isolates for comparison with clinical strains. Sampling was carried out on the CF inpatient ward (late 2018), outpatient clinic and pulmonary function test (PFT) labs (early 2019). We limited our environmental sampling to potential long-term Pa reservoirs only, because it was undertaken three years after the end of the clinical study. Sink traps were sampled by inserting sterile cotton swabs 10-15 cm into the sink tailpiece and using brush strokes and swab rotation to obtain biofilm. Swabs were placed in sterile saline and were vortexed within one hour. The re-suspended biofilm were plated in multiple dilutions to MacConkey agar and cetrimide agar,

and incubated at 42°C for 72 hours. If confluent growth of non-target bacteria was present, then further serial dilutions were performed in order to visualise isolated colonies. Presumptive Pa colonies were identified by MALDI-TOF and four to six Pa isolates from each positive sink were selected for sequencing. Due to poor recovery of Pa from clinic and PFT laboratory sinks, refrigerated biofilm samples from these sites were also cultured on minimal media that were incubated at 28°C for one week. This method, however, did not increase recovery of Pa, although there was heavy growth of heterotrophs from all samples.

Whole Genome Sequencing and Analysis

All sequencing was performed on the Illumina NextSeq instrument at the University of Toronto Centre for the Analysis of Genome Evolution and Function (CAGEF), as previously described [1]. Sequencing reads have been deposited in the NCBI short read archive with BioProject accession number PRJNA556419.

Sequencing read quality assessment and adaptor trimming was followed by de-novo assembly using SPAdes v3.12.0 [2]. The sequence type (ST) was determined from assemblies using MLST v2.15.1.

First-pass phylogenetic analysis used Mashtree v0.3 [3] to generate a neighbour-joining tree (the mashtree) from our assemblies and from 81 complete reference Pa genomes obtained from the Pseudomonas Genome Database [4] on 21 January 2018 [5]. If first-pass analysis indicated that some sequences from multiple individuals appeared closely related, we further investigated each group of sequences independently with a SNP based approach [6].

Bioinformatics Analysis

Quality control: Read count and quality was assessed with fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed and adapter sequences removed with Trimmomatic v0.33[7]. Read contamination with non-Pa genomes was screened for using Kraken [8]. De-novo assembly of each isolate was performed using SPAdesv 3.12.0 [2], as implemented in the wrapper program shovill V0.9.0 (https://github.com/tseemann/shovill), which improves the speed and accuracy of assemblies. Only assemblies with at least 25X average depth of coverage were included in the analysis.

Contamination of Illumina sequences with non-target genomes is common and can result in poor assembly quality [9]; it can occur due contamination of reagents, operator errors during DNA preparation, or during sequencing itself through processes such as index hopping or sample bleed. We screened draft assemblies for the presence of contaminating contigs based on methods recommended by Douglass et al [10]: we used custom scripts to create plots of read coverage depth

versus contig length using the contig header information in each assembly, we identified each contig using blastn, and then used an R script to plot coverage vs length and contig identity. The scripts used in this study are available at https://github.com/cizydorczyk/PAEarlyInfCF

To screen out low level non-Pa genome contamination in some assemblies, which we believe most likely came from DNA present in sequencing reagents [11], we included in our final assemblies only those contigs with a minimum length of 1000 bases (using the shovill flag --minlen 1000) and default minimum sequencing kmer depth of 2X. If plots indicated an assembly had significant contamination, e.g. due to cross-contamination of samples during DNA preparation, then the assembly was excluded, and the isolate was sequenced again if required.

Multi-Locus Sequence Typing

We used results of analysis with MLST v2.15.1 (https://github.com/tseemann/mlst) to identify the presence of CF epidemic Pa clones listed in a recent review [12]. Detection of an epidemic clone did not automatically equate to detection of a shared strain. For instance, the Strain B isolates found in one CF patient were not found in any other patient, so are considered unique. Similarly, although eight patients were infected with Sequence Type(ST) 17, a putative CF epidemic clone, only four of these patients had isolates of ST17 that were so closely related as to constitute a shared strain, so the remaining patients are each considered to have a unique strains of ST17.

Mixed-strain infection was defined as the presence of two or more Pa strains from different clonal complexes (i.e. differing in at least three of seven MLST alleles) in a new-onset sputum sample. We required that the strains of different STs were in different clonal complexes, because there are examples of closely related strains being classified as different ST, e.g. the CF-associated Liverpool Epidemic Strain (LES) is both ST148 and ST683, which share six identical alleles. All of our mixed-strain infections, by definition, differed by at least three alleles. This population diversity may be due to the simultaneous colonization by multiple strains, or a secondary infection that fails to displace an original infection. It does not encompass the long-term evolution within a patient of multiple lineages derived from a single strain, since these are highly unlikely to have diversified over short timescales to such an extent that they no longer constitute a clonal complex.

Mashtree

The epidemic population structure of Pa means that a broad diversity of Pa STs will be found in adequately sampled environments. This favours an alignment free approach for the first pass analysis, since the “gold standard” alternative requires a single reference genome for SNP calling; this can be prohibitively computationally expensive, and additionally has reduced accuracy for large

collections of diverse sequences, since a bacterial reference sequence closely related to the sequences of interest is required for accurate determination of SNPs [13].

We used the program Mashtree [3] (https://github.com/lskatz/mashtree) to create a neighbour joining tree from approximately 1000 Pa assemblies. All trees were visualised using iTOL [14]. The mashtree analysis took less than five minutes using 40 cpu cores and 160 Gb of memory on a single node of the Niagara high performance computing cluster at University of Toronto. A de-novo assembly can be generated in as little as 15 minutes from a new sequence, so the Mashtree software allowed us to quickly update the tree with new sequences as they became available. In contrast, when we evaluated the "gold standard" method on a subset of 340 Pa isolates (see below) the analysis took several days.

We used reads from 340 new-onset infection isolates to a single reference genome (PA01) with an in-house pipeline [1] to identify high quality, non-recombinant single nucleotide polymorphisms (SNPs), and used the SNP alignment to create to create a maximum likelihood (ML) tree with iqtree v1.6.3 [15]. We manually compared the clusters of isolates recovered by this tree with the clusters recovered by mashtree using phylo.io [16] and found exactly the same clusters were recovered by the mashtree. The mashtree topology was also always congruent with the ST determined from each de-novo assembly. Isolates of the same ST, but which ML analysis indicated were hundreds of SNPs apart, were always clearly distinguishable as separate clusters using the mashtree; differences on the order of 100 or fewer SNPs could not be reliably determined using mashtree alone, hence these were further investigated as potentially shared strains. Identifying potentially shared strains from the mashtree for further analysis was greatly aided by the inclusion of multiple isolates per sample: inspection of such a cluster on the mashtree reveals that the inter-patient genomic distance for isolates, based on the mash distance formulation that uses Jaccard distance between kmer sets [17], was similar to the intra-patient isolate genomic distance.

SNVPhyl

Potentially shared strains on the mashtree were investigated with SNVPhly (Single Nucleotide Variant Phylogeny v1.0.1 [6], an end-to-end pipeline designed to map reads to a reference genome and generate an alignment of high quality SNPs for the purposes of determining strain relatedness: outputs include a pairwise core-genome SNP distance matrix, and ML trees with branch support determined using the approximate likelihood-ratio test as implemented in PhyML [18]. We used the command line instance of SNVPhyl (implemented with Docker) which allowed for an automated workflow that could be run with a single command for each potential shared strain.

SNVPhyl identifies and excludes from the SNP alignment regions where ambiguous read mapping to the reference may occur (long repeats in the reference sequence); we left this setting at default parameters. It also excludes SNPs that occur in recombinant regions of each sequenced genome, which it identifies using SNP density analysis. This leverages the fact that if several SNPs are located close together on the genome, creating a SNP dense region, this is likely to reflect a single horizontal gene transfer event rather than multiple independent mutations occurring after organisms diverged from a common ancestor. SNPs that are due to horizontal gene transfer should be excluded when estimating the core genome SNP distance between isolates to avoid inflating the estimate of core genome distance. Alternative tools such as Gubbins identify recombinant SNPs after an alignment has been created: SNVPhyl uses a more computationally efficient approach to filter out SNP dense regions from each sequence independently, while achieving similar results to other approaches used to handle recombination [6, 19].

The default settings for SNPhyl identify a SNP dense region if two SNPs occur in a 500bp window. We empirically chose a more conservative SNP density filter of two SNPs in a 200bp window (using the flag --filter-density-window 200). Our initial tests indicated that the default settings, which were developed specifically for Streptococcus pneumoniae, would filter out too many true positive SNPs, thus underestimating the SNP distance between isolates. This would lead to us incorrectly identifying isolates as a shared strain when in fact they had diverged from each other several years previously. We tested different filter-density windows on several shared strains, including one (Group 51, ST110) with two new-onset infections in the same patient occurring two years apart. Using the 200 bp filter window, the median distance between the two infections was nine SNPs. Varying the parameters used changed the inter-years difference from three SNPs (--filter-density-window 500) to 29 SNPs (recombination filter off). With all --filter-density-window settings, tree topologies remained the same, and isolates that were four or fewer SNPs different with recombination filter off remained less than 4 SNPs different with any filter applied. Thus, isolates that have few overall SNP differences (including core and recombinant SNPs) are relatively insensitive to changes in SNP density filter settings, whereas reducing the --filter-density-window from 500 to 200 reduces the likelihood that isolates that diverged from each other several years ago would be deemed to be less than four SNPs apart.

Strain sharing cut-off

We chose a four SNP threshold to define strain sharing for the following reasons:

[1] We identified an intra-patient SNP distance between isolates from new-onset infections (excluding outliers) ranged from zero to four. Although most patients had isolates that were

identical, isolates showing intra-patient diversity (excluding polyclonal differences) were distributed as follows amongst shared strains (Patient, year, intra-patient SNP differences): (SK016, 2011, 1-4), (SK034, 2013, 0-2), (SK009, 2012, 0-2), (SK058, 2011, 0-1), (SK052, 2011, 0-3), (SK032, 2013, 2) (SK032, 2015, 1) . In Group 51, one patient (SK033 in 2013) had an outlier isolate 50 SNPs from the other three isolates. [2] We identified that approximately four SNPs per year accumulated within patients with multiple new-onset infections with the same strain. We hypothesise these isolates persisted in a patient reservoir such as the nasal passages, since by definition sputum was Pa culture negative between new-onset infections. Multiple year SNP differences, between patients repeatedly infected with shared strains, were as follows (Patient, years infections occurred, inter-year SNP differences): (SK040, 2014 – 2015, 4), (SK032, 2013-2015, 9), (SK009, 2012-2013, 1-3). Three other patients had repeat infections with the same, non-shared strain identified from the mashtree alone: SK010 2010 and 2014, SK036 2012 and 2015, and SK060 2014 and 2015. [3] The isolate difference within chronically infected patients with shared strains seemed to increase by approximately three to seven SNPs per year (Patient, years sampled, maximum intra-patient isolate difference): (HSC014, 2010-2013, 21), (HSC034, 2012-2013, 4), (HSC020, 2010-2012, 8), (HSC020, 2010-2013, 17), (HSC035, 2013, 3), (HSC040, 2012, 7), (HSC041, 2013, 3).

Hypermutator genotype confirmation

Potential hypermutator genotype isolates, identified by reviewing SNVPhyl ML trees for long branches, were investigated by mapping reads from these and related isolates to the PA01 reference genome using snippy v4.3.6 (https://github.com/tseemann/snippy). This pipeline returns annotated SNPs, unlike SNVPhyl, which does not provide functional annotation. We extracted annotations for non-synonymous SNPs in genes associated with Pa hypermutation [20, 21] (mutS, mutL, mutM, mutD, mutT, mutY, uvrD) and manually reviewed them for frameshift mutations.

Visualization

Epidemiological data was visualized together with ML trees using the Healthcare Associated Infections Visualization Tool (http://haiviz.beatsonlab.com/).

References (for Supplemental Methods)

1. Diaz Caballero J, Clark ST, Coburn B, et al. Selective Sweeps and Parallel Pathoadaptation Drive Pseudomonas aeruginosa Evolution in the Cystic Fibrosis Lung. MBio 2015; 6(5).

2. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology 2012; 19(5): 455-77.

3. Katz L, Griswold, T, Carleton HA. Generating WGS Trees with Mashtree. In: Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines. Washington, DC: ASM, 2017.

4. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic acids research 2016; 44(D1): D646-53.

5. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome open research 2018; 3: 33.

6. Petkau A, Mabon P, Sieffert C, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microbial genomics 2017; 3(6): e000116.

7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30(15): 2114-20.

8. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology 2014; 15(3): R46.

9. Jeong H PJ-G, Park S-H. . Contamination as a major factor in poor Illumina assembly of microbial isolate genomes. . bioRxiv, 2016.

10. Douglass AP OBC, Offei B, Coughlan AY, Ortiz-Merino RA, Butler G, Byrne KP, Wolfe KH. . Coverage-versus-Length plots, a simple quality control step for de novo yeast genome sequence assemblies. . bioRxiv 421347.doi:10.1101/421347., 2018.

11. Asplund M, Kjartansdottir KR, Mollerup S, et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 2019; 25(10): 1277-85.

12. Parkins MD, Somayaji R, Waters VJ. Epidemiology, Biology, and Impact of Clonal Pseudomonas aeruginosa Infections in Cystic Fibrosis. Clin Microbiol Rev 2018; 31(4).

13. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PloS one 2014; 9(8): e104579.

14. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research 2019; 47(W1): W256-W9.

15. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015; 32(1): 268-74.

16. Robinson O, Dylus D, Dessimoz C. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. Mol Biol Evol 2016; 33(8): 2163-6.

17. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology 2016; 17(1): 132.

18. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010; 59(3): 307-21.

19. Katz LS, Griswold T, Williams-Newkirk AJ, et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. Frontiers in microbiology 2017; 8: 375.

20. Oliver A, Baquero F, Blazquez J. The mismatch repair system (mutS, mutL and uvrD genes) in Pseudomonas aeruginosa: molecular characterization of naturally occurring mutants. Molecular microbiology 2002; 43(6): 1641-50.

21. Oliver A, Mena A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 2010; 16(7): 798-808.

Chapter 3

A Prolonged Outbreak of Human Adenovirus A31 (HAdV-A31) Infection on a Pediatric Hematopoietic Stem Cell Transplantation Ward with Whole Genome Sequencing Evidence of International Linkages

J Clin Microbiol. 2022 Nov 16;60(11):e0066522. doi: 10.1128/jcm.00665-22. Epub 2022 Oct 12. PMID: 36222515; PMCID: PMC9667772. Impact Factor Journal of Clinical Microbiology 2022: 7.1

https://pubmed.ncbi.nlm.nih.gov/36222515/

Contribution of author PS to manuscript: Design of study, bioinformatics analysis, drafting of manuscript, data sharing of sequences, manuscript submission, data sharing via Genbank.

AUTHORS AND AFFILIATIONS

Ramzi Fattouh[*a,b], Patrick J. Stapleton[#*a,b], AliReza Eshaghi[b,c], Angela D. Thomas[d], Michelle E. Science[e,f], Tal Schechter[g,f], Laurie Streitenberger[d], Petr Hubacek[h,i], Yvonne C.W. Yau[a,b], Martha Brown[b], Morag R. Graham[j,k], Jonathan B. Gubbay[b,c], Aaron J. Campigotto[a,b], Samir N. Patel[b,c], Susan E. Richardson[a,b]

* Contributed equally to this work. Author order was determined in chronological order of when the author joined the project.

[#] Submitting author and corresponding author

a. Division of Microbiology, The Hospital for Sick Children (SickKids), Toronto, Ontario, Canada
b. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
c. Public Health Ontario Laboratories, Public Health Ontario, Toronto, Ontario, Canada
d. Infection Prevention and Control, The Hospital for Sick Children, Toronto, Ontario, Canada
e. Division of Infectious Diseases, The Hospital for Sick Children, Toronto, Ontario, Canada
f. Department of Paediatrics, University of Toronto, Toronto, Ontario, Canada
g. Division of Hematology/Oncology, The Hospital for Sick Children, Toronto, Ontario, Canada
h. Department of Medical Microbiology, University Hospital Motol, Prague, Czech Republic
i. Charles University, Prague, Czech Republic
j. National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada
k. University of Manitoba, Winnipeg, Manitoba, Canada

ABSTRACT

A surge in hematopoietic stem cell transplantation (HSCT) adenovirus A31 (HAdV-A31) infections was initially observed in late 2014/2015 at SickKids (SK) Hospital, Toronto, Canada. In response, enhanced laboratory monitoring for all adenovirus infections was undertaken. Positive samples underwent genotyping, viral culture and in selected cases, whole genome sequencing (WGS). HAdV-A31 specimens/DNA obtained from four international pediatric HSCT centers also underwent WGS. During the SK outbreak period (October 27, 2014 – October 31, 2018), 17/20 HAdV-A31 isolates formed a distinct clade with 0-8 mutations between closest neighbors. Surveillance before and after the outbreak detected six additional HAdV-A31 HSCT cases; three of four sequenced cases clustered within the outbreak clade. Two SK outbreak isolates were identical to sequences from two patients in an outbreak in England. Three SK non-outbreak sequences also had high sequence similarity to strains from three international centers. Environmental PCR testing of the HSCT ward showed significant adenovirus contamination. Despite intense infection control efforts, we observed re-occurrence of infection with the outbreak strain. Severe but non-fatal infection was observed more commonly with HAdV-A31 compared to other genotypes, except HAdV-C1. Our findings strongly implicate nosocomial spread of HAdV-A31 over 10 years on a HSCT unit and demonstrate the value of WGS in defining and mapping the outbreak. Close linkages among strains in different countries suggest international dissemination, though the mechanism is undetermined. This large, extended, outbreak emphasizes the pre-eminent role of HAdV-A31 in causing intractable pediatric HSCT outbreaks of severe illness worldwide.

INTRODUCTION

Adenovirus infections are usually acquired in the first decade of childhood, often followed by latency in the oropharynx and gastrointestinal tracts(1). This sets the stage for reactivation of infection in the setting of severe immune compromise, such as hematopoietic stem cell transplantation (HSCT)(2, 3). However, as 60% of children undergoing HSCT are less than 10 years of age(4), primary infection also occurs, and may be nosocomially acquired(5-8). Adenoviral infections following pediatric HSCT are more common (6-42% viremia) than in adults (3-15% viremia)(1), and are more likely to be severe, disseminated and life-threatening(1, 9).

In late 2014/2015, the Infection Prevention and Control (IPAC) team noted a cluster of human adenovirus genotype A31 (HAdV-A31) infections on the HSCT ward at SickKids Hospital, Toronto,

Canada. Since HAdV-A31 was previously uncommon among our adenovirus-infected HSCT patients, and there was a precedent for nosocomial spread of this genotype in pediatric HSCT centers (5, 7), we proceeded to determine whether the HAdV-A31 strains were genetically related.

The objective of this report is to describe the clinical and laboratory features of the outbreak, describe the use of whole genome sequencing in delineating local outbreak cases and their relationship to international strains, and to propose strategies for optimal monitoring of such patients in pediatric facilities.

MATERIALS AND METHODS

SickKids Study Population

Outbreak period

All children with a positive adenovirus PCR test from any body site after undergoing HSCT at SickKids between October 27, 2014, and October 31, 2018, were enrolled (outbreak period). Patients were identified by search of the Microbiology database. Demographic, clinical and laboratory information were obtained by chart review. The study was approved by the SickKids Research Ethics Board (1000068500).

Pre-outbreak period

All HSCT children with a positive adenovirus PCR test from any site between January 1, 2012 and October 26, 2014.

Post-outbreak period

All HSCT children with a positive adenovirus PCR test from any site between November 1, 2018 and December 31, 2021.

HAdV Detection

Adenovirus PCR was performed using the NUCLISENS easyMAG system for nucleic acid extraction (bioMerieux Canada, Inc., St. Laurent, Canada) for all sample types, followed by the quantitative RealStar Adenovirus PCR Kit (Altona Diagnostics, Hamburg, Germany), shown to detect all adenovirus species (A-G). Amplification and detection were performed using the Rotor-Gene Q (Qiagen, Hilden. Germany). For quantification one positive control provided by the manufacturer

(QS3, 100,000 copies/ml) was included in each run and used to adjust an imported standard curve generated using four positive controls (QS1-QS4, range 10^7- 10^4 copies/ml).

HAdV Genotyping

HAdV genotyping was performed using a previously described method targeting the hypervariable region 7 of the hexon gene(10) and determined on the basis of BLAST analysis by comparing the sequence to the GenBank nr/nt database using Blastn. HAdV specimens with low viral loads (usually $\leq 10^5$ genome copies/ml) were not able to be reliably genotyped.

HAdV Culture

Viral culture from stool, urine and nasopharyngeal (NP) samples was performed to enrich for virus from primary specimens prior to whole genome sequencing at Public Health Ontario Laboratories (PHOL). Stool samples (pea-sized if solid; 0.5 ml if liquid) were transferred to 4.5 ml of penicillin-streptomycin-amphotericin B (PSF solution), to yield a final volume of 5ml (10% suspension). Stool was crudely homogenized by manual grinding, and then allowed to stand at 4^{o}C for 30 minutes. The suspension was centrifuged for 30 minutes at 1,000g, then transferred to a sterile tube containing 0.2 ml of PSF. 0.2 ml of urine was added to 0.2 ml PSF and allowed to stand at 4^{o}C for 30 minutes. Nasopharyngeal swabs were vortexed briefly, then processed as per urine samples. Prepared samples were then transferred (0.2 ml) to each of two cell lines, rhesus monkey kidney and MRC5 cells (Diagnostic Hybrids Inc., Ohio, US) and incubated at 37^{o}C on a roller drum. Inoculated cell culture tubes were examined for cytopathic effect every other day for up to 10 days. Adenovirus positivity was confirmed in cultures demonstrating evidence of cytopathic effect using an immunofluorescence-based assay (Diagnostic Hybrids Inc.). Cell culture supernatant was than harvested (no serial passage performed), aliquoted and frozen at -80^{o}C until time of sequencing.

Adenovirus Screening (see Online Supplement for comprehensive IPAC approach)

Pre-Outbreak: Routine weekly adenovirus monitoring from plasma was performed in all SickKids patients from day 0-100 post-HSCT. An adenovirus positive plasma result prompted adenovirus PCR testing in the urine, and also in the respiratory tract, if the patient had respiratory symptoms. The stool was tested when a patient had diarrhea, using routine bacterial/viral diagnostics in place at the time [pre-July, 2013: electron microscopy (EM), beginning July, 2013: multiplex PCR xTAG Gastrointestinal Pathogen Panel, Luminex Corp. Toronto, Canada) was added to EM]. Stool was not tested by the Altona adenovirus PCR test pre-outbreak. Genotyping was only performed on request by the IPAC team.

Outbreak onset and post-outbreak: Following recognition of the outbreak in 2014/15, weekly plasma screening was continued, and both routine adenovirus PCR (Altona) and multiplex GI PCR (Luminex) testing of stools in all HSCT patients with diarrhea was introduced, while stool EM was discontinued. An initial adenovirus positive PCR test from any site (usually plasma) prompted automatic adenovirus screening of urine and stool of symptomatic or asymptomatic patients. The nasopharynx was sampled if the patient was symptomatic. Additional IPAC precautions were continued until stool/urine/NP were <1000 genome copies/ml, whether inpatient or outpatient (during clinic visits). The frequency of stool monitoring during admission was weekly or every two weeks, and monthly after discharge until negative. At onset of the outbreak, genotyping became routine and was continued thereafter for every new adenovirus infection on the HSCT ward. Genotyping was only repeated if a patient presented with new symptoms of infection after complete clinical recovery from a preceding adenovirus infection, and negative adenovirus PCR results at all body sites.

Other HSCT-associated HAdV-A31 SickKids infections outside the outbreak period (pre- and post-outbreak periods)

We searched for missed HSCT HAdV-A31 cases in the pre-outbreak period, in addition to maintaining surveillance and genotyping for all new adenovirus infections in HSCT after the outbreak (post-outbreak). One pre-outbreak case (July, 2012) and five post-outbreak HAdV-A31 cases were identified, one in November, 2019 and four cases in August - October 2021. We also retrieved a remote HAdV-A31 isolate from a non-HSCT patient with gastrointestinal infection (1982) for comparison to outbreak strains.

Environmental Sampling

Environmental sampling was performed in selected rooms/areas in 2015 on and off the outbreak ward from high touch surfaces. Rooms were chosen at a point in time to include i) one room on the HSCT ward occupied by a patient with active A31 infection, ii) three rooms on the HSCT ward with resolved A31 infections, after discharge terminal cleaning, and iii) the HSCT outpatient clinic, which sees discharged HSCT patients who may have persistent or recurrent positive PCR results for adenovirus. In addition, the HSCT ward common workstation, and common family areas were screened. Control groups included i) one non-HSCT general infectious diseases outpatient clinic room, and ii) two non-HSCT non-HAdV-infected patient rooms elsewhere in the hospital.

Approximately 10 high-touch surfaces per patient room (up to a 100 cm^2 surface area per site), such as door handles, blood pressure cuff, side table, sink faucet handles, mattress, and curtains were

sampled using a flocked swab (Copan Diagnostics, Murietta, USA) pre-moistened with Universal Transport Medium (UTM, Copan Diagnostics). The swabs were tested for HAdV using the same qPCR method as for clinical samples. Five surfaces with the highest viral loads were subjected to viral culture, direct genotyping and metagenomic sequencing.(5, 7)

Samples from collaborators

We obtained frozen stool specimens associated with a pediatric HSCT outbreak in Czech Republic, with sequences obtained from 4 cases following culture enrichment. Viral culture supernatants from pediatric HSCT HAdV-A31 infections in Ireland (n=3 samples from 2 patients) were successfully sequenced. Purified DNA from culture enrichment was sequenced from US cases on a pediatric HSCT unit (n=4 samples from 3 patients); 3 from stool specimens and one from a nasopharyngeal sample. We also obtained consensus HAdV-A31 sequences, both published and unpublished, from a UK pediatric hospital study(8, 11).

Whole genome sequencing (WGS)

WGS was performed on all SK HAdV-A31 isolates obtained at any time during the outbreak or pre-/post-outbreak periods, in addition to as many HAdV-C1 and HAdV-C2 isolates from the outbreak period as possible. Viral culture supernatant was sequenced (PHOL) with Illumina MiSeq using the V2 (2x150bp) or V3 (2x300bp) reagent kit as previously described(12). For those cases where culture failed, shotgun metagenomic sequencing was attempted directly on available clinical specimens. Complete genome sequences could not be obtained from one Toronto 2017 HAdV-A31 case, despite repeated attempts at viral culture and shotgun metagenomic sequencing.

For the Toronto sequences, the phylogenetic analysis includes one sequenced specimen per case, with the exception of one HSCT patient (Case 07) who had two samples sequenced from distinct infections in 2015 and 2017. Several cases from collaborator sites had two samples collected at different times from a single patient; if the final HAdV sequences from a patient at a collaborator site showed sequence variation over time, then both sequences were retained in the phylogenetic analysis, indicated with suffix -1 and -2 on phylogenetic trees.

Sequencing reads were mapped to suitable HAdV reference genomes (KF268119.1 for HAdV-A31), and the resulting BAM files were investigated with bamstats v1.25 to determine the average depth of coverage per site. Reads accepted for further analysis had over 35-fold mean coverage depth and over 99% reference coverage, and underwent adaptor and quality trimming with Trimmomatic v0.33(13). We used snippy v4.3.6 (https://github.com/tseemann/snippy) with default parameters to perform mapping of reads to reference genomes, identification and annotation of single nucleotide

polymorphisms (SNPs) and indels, and generation of consensus genome sequences. Consensus sequences of each HAdV type were aligned using MAFFT v7.4(14) and SNP distance matrices were generated from the alignment with snp-dists v0.6.3 [https://github.com/tseemann/snp-dists]. SNPs at alignment ends (positions 1-50 and 33792-33802) were masked. Maximum likelihood (ML) phylogenetic trees were generated from sequence alignments using iqtree v1.6.9 (15) with the following parameters: nucleotide substitution model (-m) GTR, 1000 ultrafast bootstrap support replicates (-bb 1000) and keep identical isolates (-keep-ident). The resulting trees were visualised in iTOL (16). Nucleotide variants and alignment were visualised together with phylogenies using the gingr tool from the harvest suite (17).

We investigated monophyletic groups comprising 2 or more patient samples, with ultrafast bootstrap support values of at least 90% as indicating possible clusters of cross-transmission. We analysed pairwise SNP differences within clusters in light of previously identified differences in one centre of approximately 3 to 4 SNPs for within cluster/within host HAdV nucletotide variation and >30 SNPs between many apparently unrelated HAdV sequences (8).

WGS was carried out on three of the post-outbreak HAdV-A31 isolates (November 1, 2018 - December 31, 2021), using Nanopore technology at SickKids. Culture supernatant was sequenced using a GridION (Oxford Nanopore Technologies, UK) utilizing an R.9.4.1 flow cell and LSK 109 barcoding kit. Onboard, high accuracy basecalling was performed and consensus genome assembly performed by mapping to KF268119.1 using MiniMap2 (version 2.0) and samtools. All assemblies had average read depth of at least 100X. Consensus sequences were then aligned to existing sequences and analysed as described above. The underlying nanopore read support for phylogenetically informative SNPs was confirmed by manual inspection.

Statistical analysis

All statistical analysis was performed in R version 4.0.0. Significance levels were $\alpha < 0.05$. Y axis data were $\log_{10}$ transformed before statistical analysis to reduce skewness and the influence of outliers. Comparison of genome copies per ml and days among HAdV-A31, HAdV-C1 and all other genotypes were conducted using Welch's analysis of variance (ANOVA) and Games-Howell post-hoc tests (does not assume equal variance or sample size between groups).

Data availability statement

The adenovirus sequences generated for this study were deposited in NCBI GenBank and are associated with BioProjects PRJNA598822 and PRJNA790639 (https://www.ncbi.nlm.nih.gov/nuccore/MN901805…MN901840 and OM112280…OM112294).

RESULTS

Epidemiology

A cluster of HSCT HAdV-A31 cases in late 2014/2015 triggered genotyping of all new cases, heightened case ascertainment, a search for pre-outbreak cases over the preceding three years, line listing of common exposures and implementation of enhanced infection control measures (Figure 1A).

FIGURE 1

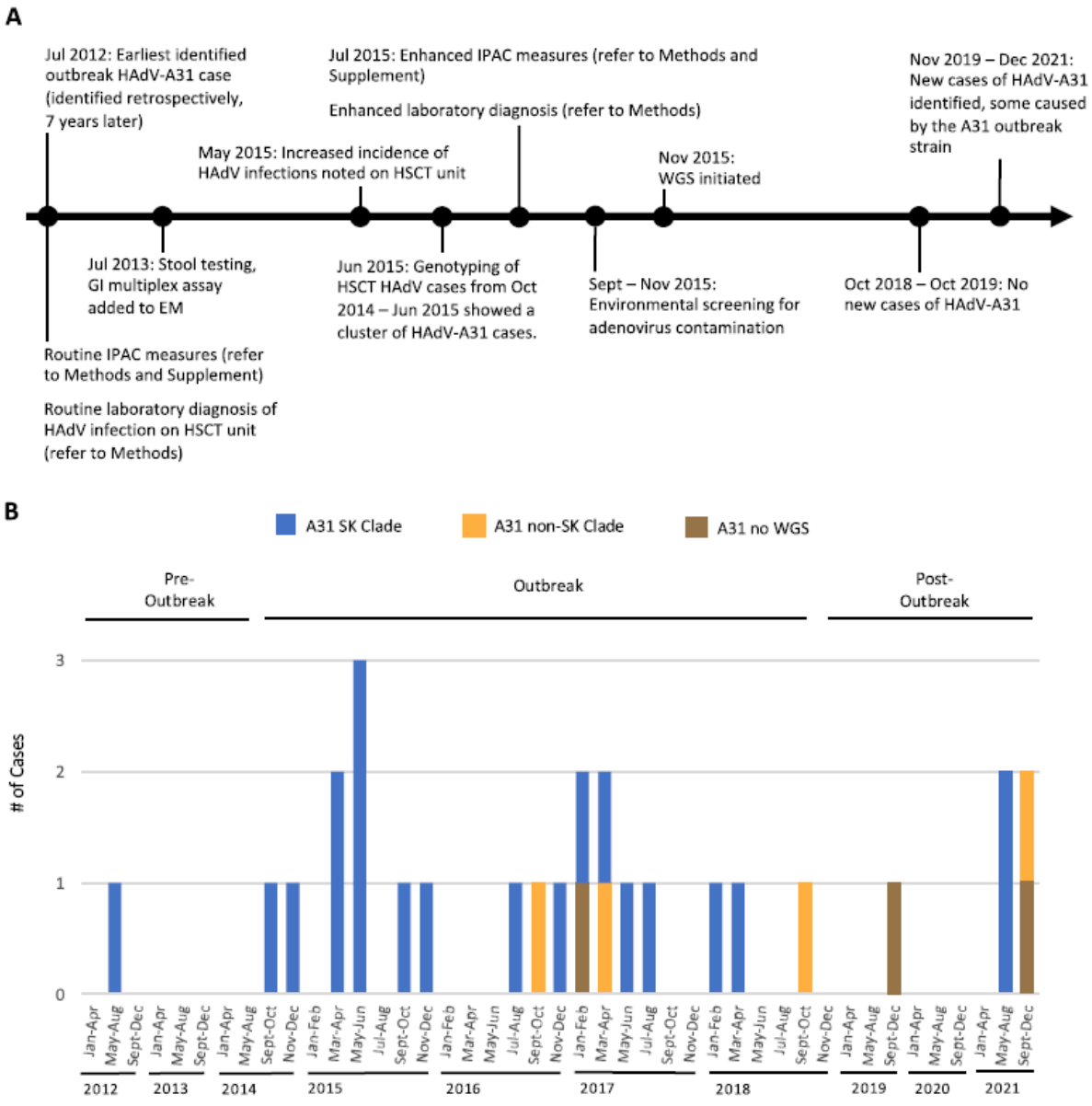


Figure 1: Time course of investigations and epidemic curve of HAdV-A31 cases pre-outbreak, outbreak and post-outbreak, 2012-2021

A) Flow chart outlining the course of investigations. B) Epidemic curve of HAdV-A31 cases. Three of the 21 cases in the outbreak period (October 27, 2014 – October 31, 2018) and one case in the post-outbreak period (April 2018 – August 2021) were subsequently revealed to be non-outbreak strains (orange) by whole genome sequencing. Adequate WGS results could not be obtained from three cases (brown) and could not be included in the phylogenetic analyses. HAV - human adenovirus; GI – gastrointestinal; EM - electron microscopv: IPAC - infection prevention and control

Outbreak period (October 27, 2014 to October 31, 2018)

Seventy-one adenovirus episodes of infection (mean 17.8 cases per year) were detected, from which 21 isolates were genotyped as HAdV-A31 (29.6% of all adenovirus infections). The HAdV-A31 epidemic curve showed two peaks of infection, one in 2014-15 and one in 2016-17 (Figure 1B). Other genotypes detected in the outbreak period included C1 (n=9), C2 (n=21), A12 (n=4), C6 (n=3), A61 (n=1) and D45 (n=1). Eleven isolates were not typeable due to low viral loads in stool. In general, non-typeable strains often showed low, transient viremia and were seldom detectable in multiple body sites (data not shown).

Pre- and post-outbreak strains

During the pre-outbreak period (January 1, 2012 to October 26, 2014), fewer adenovirus infections in HSCT patients were documented (30 total episodes, mean 10 cases/year), although screening was not as vigilant at this time. Consistent with clinical practice at the time, stools were not routinely tested for adenovirus, and routine genotyping was not performed. However, one additional HSCT HAdV-A31 case was detected retrospectively from July, 2012 on a stool that had been tested by electron microscopy and banked. This 2012 patient did not have HAdV-A31 detected again at any time during subsequent admissions to SK, which spanned the outbreak period, despite being tested on multiple occasions. Five new HAdV-A31 cases were detected in the post-outbreak period, one in September 2019 and four during August to October 2021.

Clinical-laboratory correlation with genotype of HAdV-infected patients during the outbreak period

Mean maximum plasma viral load (genome copies/ml) was compared for each genotype. Preliminary results indicated significant differences within A31- and C1-infected patients, compared to C2 and all other genotypes (A31: 5.9×10^5, C1: 1.2×10^6, C2: 2.7×10^4, all others: 2.5×10^4). As such, further data analysis regarding clinical and laboratory parameters was divided into three groups: A31, C1 and all other genotypes (including C2, representing 70% of other genotypes).

There were 20 individual patients (21 episodes) with HAdV-A31 infections. HAdV-A31 patients had a lower median age at HSCT (3.3 years) and were more commonly female (55%) than in infections with other genotypes. Solid tumors were the most common underlying illness in A31 patients, compared to patients with infections due to other adenovirus genotypes, in whom leukemias and lymphomas predominated (Table 1).

TABLE 1 Demographic and clinical data for HAdv infections occurring during the outbreak period[a]

Data	All genotypes	A31	C1	Other genotypes
Demographic data				
Total patients (*n*)[b]	55	20	7	28
Total adenovirus episodes (*n*)[c]	57	21	8	28
Age (yrs) at HSCT, median (range)	4.4 (0.4–18.7)	3.3 (0.4–14.7)	3.8 (1.4–13.5)	8.1 (0.8–18.7)
Female, *n* (%)	24 (44)	11 (55)	3 (43)	10 (36)
Underlying disease (by individual patient), *n* (%)				
Leukemia/lymphoma	24 (43)	5 (25)	5 (71)	14 (50)
Solid tumor	13 (24)	7 (35)	0	5 (18)
Immunodeficiency disorder	9 (16)	4 (20)	1 (14.5)	4 (14)
Hematologic disorder	7 (13)	2 (10)	1 (14.5)	5 (18)
Hurler syndrome	2 (4)	2 (10)	0	0
Type of transplant, *n* (%)				
Allogeneic	42 (76)	13 (65)	7 (100)	22 (79)
Autologous	13 (24)	7 (35)	0	6 (21)
Clinical data				
Time (d) from transplant to adenovirus infection, median (range)	20 (1–335)	18 (6–141)	15 (4–86)	36 (1–335)
Blood PCR positive (%)	49/57 (86)	21/21 (100)	7/8 (88)	21/28 (75)
First positive sample = blood, *n* (%)	37/57 (65)	17/21 (81)	4/8 (50)	16/28 (57)
NP PCR positive, *n* (%)	20/25 (80)	11/12 (92)	2/2 (100)	7/11 (64)
Urine PCR positive, *n* (%)	32/46 (70)	19/21 (90)	5/7 (71)	8/18 (44)
Stool PCR positive, *n* (%)	53/53 (100)	19/19 (100)	8/8 (100)	26/26 (100)
HAdV infection episodes treated, *n* (%)	29/57 (51)	14/21 (67)	5/8 (63)	10/28 (36)
Cidofovir alone	15/29 (52)	8/14 (57)	1/5 (20)	6/10 (60)
Brincidofovir alone	7/29 (24)	5/14 (36)	1/5 (20)	1/10 (10)
Cidofovir and brincidofovir	7/29 (24)	1/14 (7)	3/5 (60)	3/10 (30)
Deaths from any cause occurring within 90 days onset adenovirus infection, *n* (%)	5/55 (9)	0/20	1/7 (14)	4/28 (14)
Time (d) from adenovirus infection onset to death from any cause, median (range)	220.5 (3–1,008)	294 (207–781)	78 (3–530)	259 (22–1,008)

[a]HAdV, human adenovirus.
[b]Includes only data from cases where isolate could be genotyped and clinical information was complete.
[c]One patient had two episodes of A31 infection two years apart and one patient had separate infections with A31 and C1 three months apart.

Allogenic transplants were more common in HAdV-A31-infected patients. HAdV-A31 and C1 infections were diagnosed earlier post-transplantation than other genotypes (median 18 d and 15 d, respectively, vs. 36 d). HAdV-A31 infections were detected earlier in blood and were more commonly detected in multiple body sites, compared to infections due to C1 or other genotypes (Figure 2).

FIGURE 2

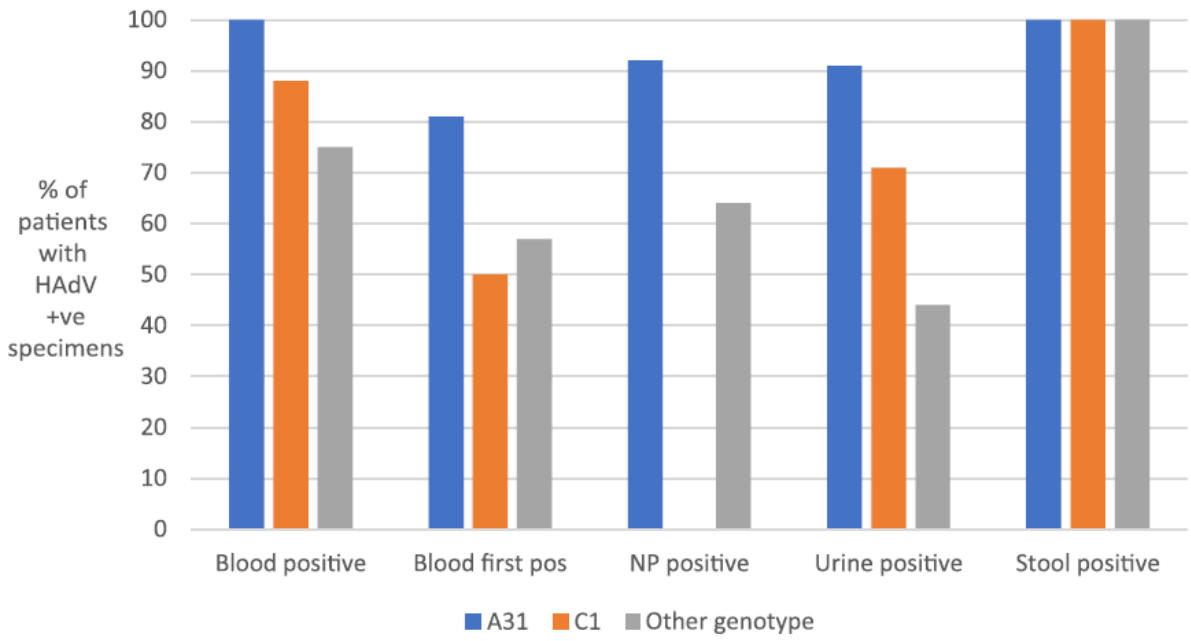


Figure 2: PCR positivity body site vs. Genotype during outbreak period: HAdV-A31 presented earlier in the blood and was more frequently disseminated to multiple body sites than other genotypes. Includes only data from cases where isolate could be genotyped and clinical information was complete. A31 (n=21), C1 (n=8), other genotype (n=28)

Laboratory correlates of clinical severity indicated highest peak viral load and greatest length of time at a plasma load >1000 copies/ml for HAdV-A31 and HAdV-C1 infections compared to other genotypes (C2, A12, C6, A61, D45 combined), although the differences were not statistically significant (Figure 3).

FIGURE 3

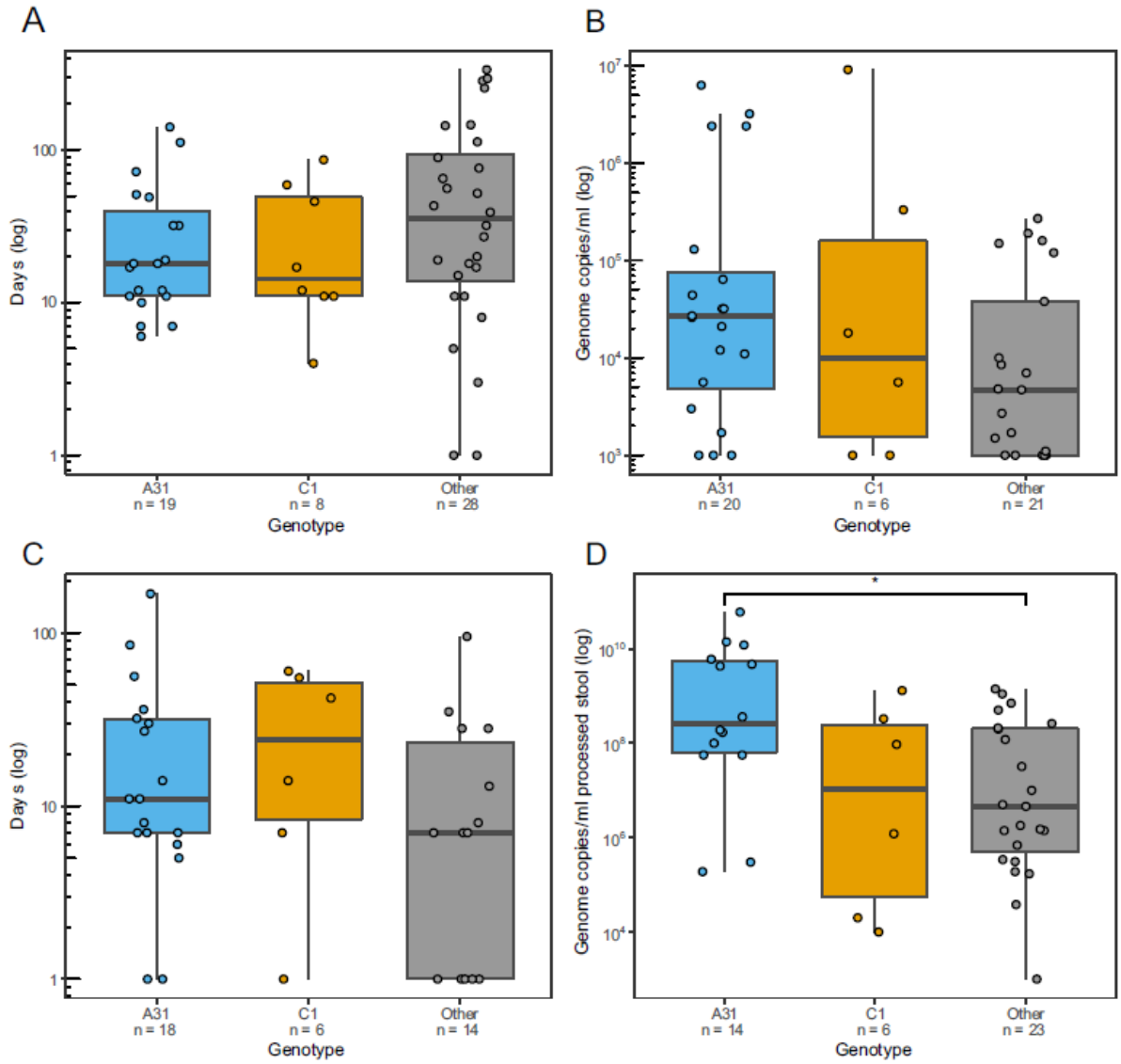


Figure 3: Clinical-laboratory correlation with respect to genotype during outbreak period: A) Transplant to first positive PCR (days): HAdV-A31 and HAdV-C1 infection presented earlier than other genotypes. B) Maximum plasma viral load: Median peak viral load in plasma was highest for A31, lower for C1 and lowest for other genotypes (although not statistically significant, Welch's ANOVA F=0.68, p>0.05). Peak viral load in plasma was greater than 1000 genome copies/ml in 86% of A31, 75% of C1 and only 50% of infections caused by other genotypes (Data not shown). C) Plasma load >1000 genome copies/ml (days): The mean number of days that the viral load in the plasma remained greater than 1000 genome copies/ml was highest for the C1 infected patients, lower in the A31 group and lowest for all other genotypes (not statistically significant, Welch's ANOVA, F=1.33, p>0.05). D) Median stool load: Stool viral load was highest for A31.

Stool specimens were universally positive by PCR for infectious episodes due to all genotypes, but the highest median viral loads (1.8 x 10^8 genome copies/ml of processed stool) occurred in A31 infections (Welch's ANOVA F=4.98, p=0.012; Figure 3).

About half of all infections were treated with antivirals (cidofovir, brincidofovir), more commonly for A31 and C1 infections (Table 1). All-cause mortality within the entire cohort of adenovirus-infected patients in the first 90 days following onset of infection was 9%; none of these were due to HAdV-A31.

Environmental Sampling

Environmental sampling showed 85% of high-touch surfaces in a room and anteroom of an actively infected patient were positive for HAdV by qPCR positive (Supplementary Figure S1).

FIGURE S1

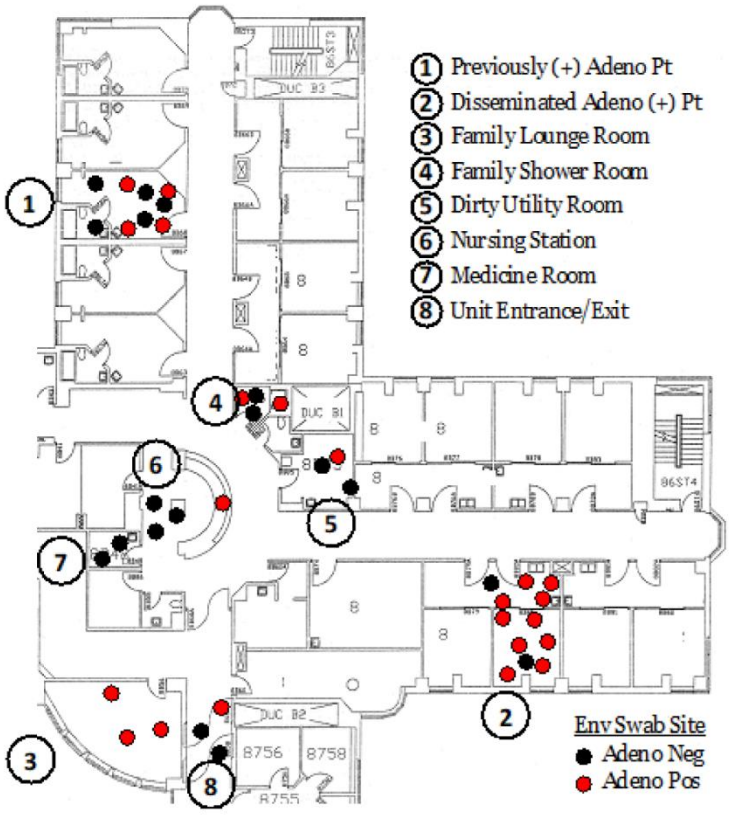


Figure S1: Floor plan of hematopoietic stem cell transplantation (HSCT) unit with results of environmental screening for HAdV

44% of surfaces were positive for HAdV by qPCR in a room where a patient with resolved infection had resided (1). 85% of surfaces tested positive in a room with a patient who had disseminated HAdV infection, including samples from an anteroom shared with another patient (2). The heavily contaminated patient room had the highest viral load detected from the entire screening exercise (2 x 10^4 genome copies/ml), on the interior door-opening button 6-8 hours post routine cleaning. 63% of the common family areas were positive by PCR including the couch, chair, refrigerator door, and TV remote control (3 and 4). One positive sample was detected in the dirty utility room (5). The ward workstation and medication room showed just one positive counter top (6 and 7). One of three entry/exit doors to the ward was PCR positive (8).

Surface positivity and viral load were reduced by each successive intensive clean, but eradication was not possible. Compared to rooms with actively infected patients, lower surface positivity rates and viral loads in other HSCT ward and clinic rooms were observed. Broader screening of the outbreak ward (Table 2) showed that three of six anterooms were PCR positive. Clean care areas (e.g. nursing workstation) showed little positivity, whereas 63% of the common family areas (e.g. lounge couch, refrigerator) were PCR positive.

TABLE 2 Environmental screening of HSCT ward and non-HSCT areas during outbreak period[a]

Area screened	qPCR-positive sites/ total tested (*n*, %)	Highest viral load (genome copies/mL)	Site of highest load
HSCT room/anteroom of patient with active disseminated HAdV-A31 infection ($n = 1$)	11/13 (85)	2×10^4	Interior door opening button
Post-clean number 1 HSCT room/anteroom of patient with disseminated infection	7/11 (64)	1.6×10^4	Chair
Post-clean number 2 HSCT room/anteroom of patient with disseminated infection	8/12 (67)	10^4	Interior door opening button
Post-clean number 3 HSCT room/anteroom of patient with disseminated infection	4/8 (50)	<1,000	Chair
HSCT ward and HSCT clinic rooms of patients with previous HAdV-A31 infection, now negative ($n = 3$)	10/30 (33)	1.2×10^3	Interior door handle
HSCT room of patient without adenovirus infection ($n = 1$)	3/10 (33)	<1,000	Computer mouse/IV pump
Non-HSCT ID ward and ID outpatient clinic rooms of patients without HAdV infection ($n = 2$)	0/19 (0)	NA	NA
Ward common workstation/medication room	1/6 (17)	<1,000	Counter
Dirty utility room	1/3 (33)	6.8×10^3	Commode in dirty utility room
Family lounge/shower	5/8 (63)	6.2×10^3	Couch
Unit entrance/exit doors	1/3 (33)	<1,000	Door

[a]HSCT, hematopoietic stem cell transplantation; qPCR, quantitative PCR; HAdV, human adenovirus; NA, not applicable; ID, infectious diseases.

None of the five surfaces subjected to viral culture, direct genotyping (i.e. PCR-based) and metagenomic sequencing yielded genotype/strain related information. The general infectious diseases ward and clinic did not show any positive environmental surfaces by qPCR.

WGS Analysis

SickKids HAdV-A31 strains

Of 21 HAdV-A31 episodes during the outbreak period, 20 high quality sequences were included in the final phylogenetic analysis, using WGS of one sample per infection: stool culture isolates (n=15), urine culture (n=2), direct WGS of urine (n=2) and stool (n=1). In all cases (n=4 patients) in whom genomes were sequenced from both a culture isolate and direct sample, no sequence mismatches were observed; only one sequence from each pair was included in the final analysis.

Of the 20 sequenced HAdV-A31 viruses in the outbreak period, 17 sequences (from 16 patients) clustered into a distinct clade based on phylogenetic analysis (0-8 SNPs between closest neighbors; maximum pairwise SNP distance between most distantly related strains in the cluster was 11 SNPs). In comparison, the two complete HAdV-A31 genomes available at the time from NCBI GenBank and a 1982 non-HSCT SickKids isolate differed by >100 SNPs from the outbreak isolates (Figure 4A).

FIGURE 4

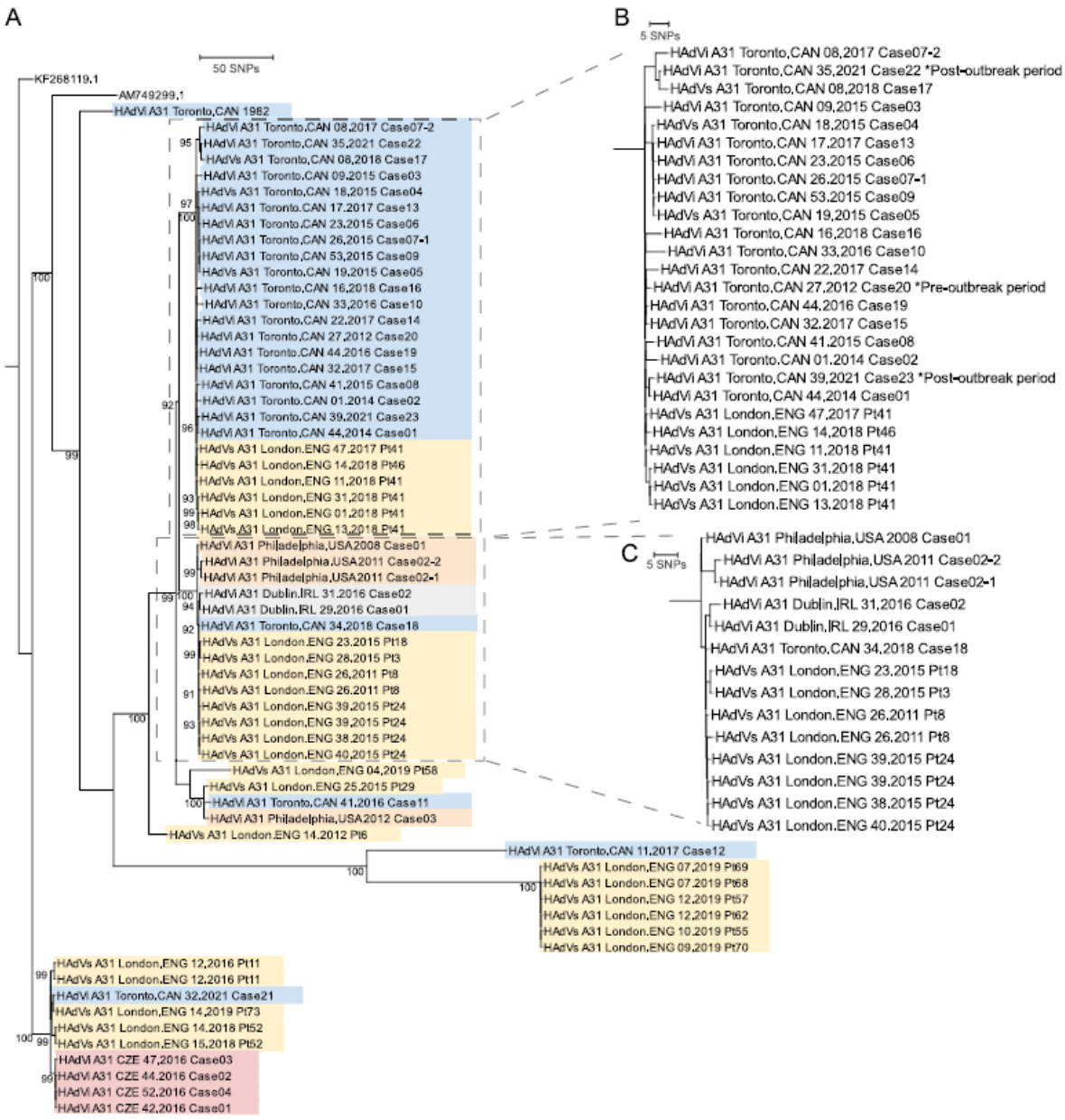


Figure 4: Phylogenetic tree of all available HAdV-A31 sequences: (A) Phylogenetic tree of all sequenced HAdV-A31 strains from Toronto (blue) (pre-outbreak, outbreak and post-outbreak) and comparator strains from pediatric patients in HSCT units from USA (beige), Ireland (grey) and Czech republic (pink) sequenced for this study and from a UK HSCT unit obtained from GenBank (cream). Dashed lines surround two clades of interest shown in greater detail in the insets to the right. Several unique sequences are present including a historical (1982) HAdV-A31 stool isolate from Toronto and two GenBank reference strains. The tree was created from a whole genome alignment (first and last 20bp masked) using the maximum likelihood method with GTR substitution model. It is rooted on reference strain KF268119.1. Nodes with ultrafast bootstrap support values of at least 90% are annotated. Sequence naming convention for this study: HAdVi (Human Adenovirus isolate) or HAdVs (Human Adenovirus sequence from clinical sample)/City.Country/Week(ISO format). Year/Case

number (with -1 or -2 suffix if more than one sequence per patient included). Further details on the origin of each sequence are included in the Supplementary material.

(B) Inset showing that the majority of Toronto HSCT sequences form a distinct clade (20 sequences from 19 patients) together with six sequences from two patients who were part of a UK HSCT outbreak. Within this clade, six identical Toronto isolates form a subclade (cases 04, 05, 06, 07-1, 09, 13); these were isolated over 2 years and each case had an overlapping inpatient admission with another member of the subclade

(C) Inset showing Toronto Case 18 differing by zero or one SNP from sequences from another UK HSCT outbreak, and by three to five SNPs from sequences from Ireland and the USA.

The remaining three HAdV-A31 strains from the outbreak period appeared distinct from both the major cluster (SNP difference range 30 to 376) and each other and thus were suspected to be 'unique' and inconsistent with recent transmission.

Within the major cluster, 6 isolates were identical (Figure 4B), all of which had overlapping admissions with a known positive case (100%). In contrast, only 20% of the ten other cases within the outbreak cluster (1-8 SNPs different from the six identical cases above), had an overlapping admission with a known positive case. Surprisingly, pre-outbreak genotyping revealed an HSCT case from 2012 (Case 20), 2.3 years before the first case in the study period, that located within the outbreak cluster and differed by only 1 SNP from identical Cases 15 (2017) and 19 (2016).

WGS analysis of the five new post-outbreak strains was undertaken, but 2/5 isolates did not grow in culture (one each from 2019 and 2021). Of the 3 sequenced isolates, two 2021 isolates were positioned within the outbreak cluster; one (Case 22) was 4 SNPs different from Cases 17 and 07-2, and one (Case 23) was 2 SNPs from Case 01 and 5 SNPs from Case 08 (Figure 4B).

Eight SNPs were identified between two viral sequences obtained from the only known patient with two discrete episodes of A31 infection, separated by two years, with interim clinical and virologic recovery (Case 07). The 2017 sequence (Case 07-2) had acquired 5 new SNPs and lost 3 SNPs relative to the 2015 isolate.

SickKids HAdV-C1 and HAdV-C2 strains

Five of nine (56%) HAdV-C1 strains (3 culture isolate, 2 direct specimen) and 7/21 (33%) HAdV-C2 (5 culture isolate, 2 direct specimen) were sequenced by WGS. Sequencing results did not support recent transmission, even after inclusion of C1 and C2 strains from London (8) and GenBank reference strains (Figure 5).

FIGURE 5

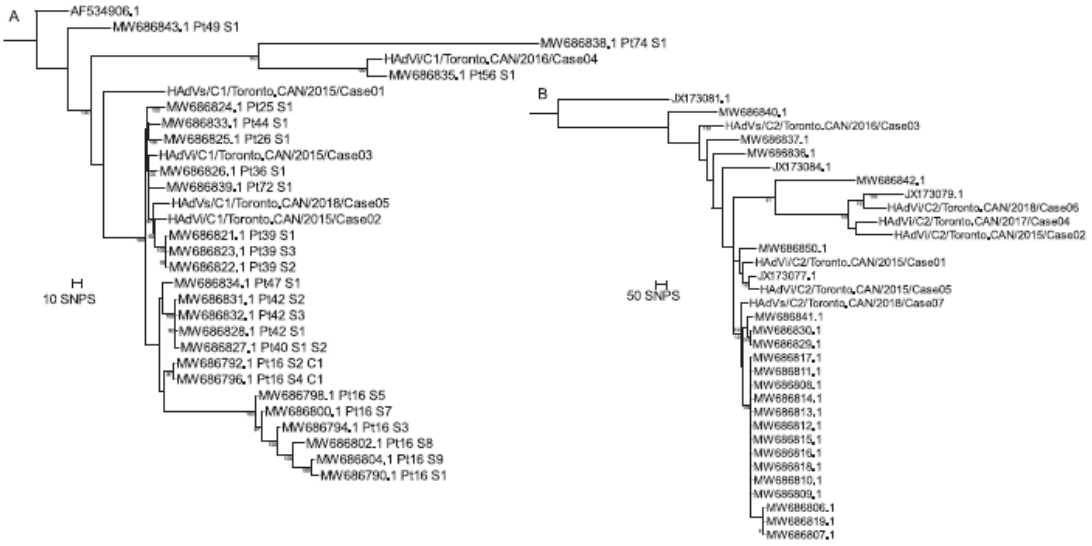


Figure 5: Phylogenetic tree of (A) HAdV type C1 and (B) C2 whole genome sequences from HSCT patients in Toronto during the outbreak period, and from the UK, with selected GenBank reference strains: The closest relationship involving a C1 Toronto sequence is 23 SNPs (Toronto C1 Case 03 and UK sequence MW686833.1). For type C2 sequences the closest relationship involving a Toronto sequence is 22 SNPs (Toronto C2 Case 07 and UK sequence MW686808.1). The SNP distances are not consistent with recent cross-transmission involving Toronto isolates.

In contrast to a previous study, we did not identify any mixed strain infections (8) or recombination.

International strains

Analysis of strains from four international pediatric HSCT centers revealed several instances where very closely related HAdV-A31 sequences were found in different centers, strongly suggesting inter-center transmission of outbreak strains (Figure 4A). One compelling illustration of this involves two strains from our outbreak clade. Our Cases 15 (August 2017) and 19 (November 2016) harbored sequences identical (0 SNPs) to two patients from the “Cluster 3” outbreak in England, comprising Pt41 (November 2017) and Pt46, (April 2018). Pt41 went on to develop intra-patient diversity: of 4 later isolates, 3 have a unique SNP not present in Toronto sequences.

We also observed unexpectedly close relationships of three non-outbreak clade HAdV-A31 sequences from Toronto (cases 11, 18, and 21, Figure 4A) to sequences from other countries. Case 11 exhibited only 8 SNPs from UK and USA strains. Case 18 from Toronto demonstrated minimal pairwise differences (range 0 to 5 SNPs) from sequences from London (UK), Philadelphia (USA) and Dublin (Ireland) (Figure 4C). Case 21 showed one SNP difference from one UK patient (Pt73) and 5 SNPs from another (Pt52) (11).

In terms of diversity found within different countries (one pediatric HSCT unit per country); two Irish isolates differed by 2 SNPs. USA isolates fell into two distinct clades with within-clade differences of 4 SNPs. Sequences from Czech Republic were practically identical (1 SNP among 4 cases). Sequences from a London HSCT outbreak exhibited a previously described mix of unique and outbreak strains, with some limited within-patient diversity (8).

DISCUSSION

We describe a distinct strain of HAdV-A31 causing the largest described outbreak of adenovirus infection on a pediatric HSCT unit, involving at least 19 patients over 10 years. Whole genome sequencing confirmed and defined the initial outbreak of 2014-2018, which included a cluster of 16 cases (17 sequences) with genomic differences of 0 to 8 SNPs between closest neighbors. WGS then provided evidence for three additional pre- and post-outbreak strains existing within the same cluster, extending the outbreak of genetically-related strains from 2012 to 2021. Periods of heightened activity, followed by quiescent periods, some as long as 3.3 years, occurred over the 10-year period. A number of factors likely contribute to the lack of easily identifiable epidemiologic links in this setting and with this virus, including prolonged subclinical shedding of the virus from the GI tract of patients and perhaps staff, persisting environmental contamination, and failure to identify HAdV-A31 in stools with low viral loads and within mixed adenovirus infections in the GI tract.

In 2014/2015, when an outbreak was first suspected, only two complete HAdV-A31 sequences were available in GenBank and there were no established criteria to appraise relatedness of adenovirus sequences within specific genotypes. To provide context for our results, we sequenced material (DNA or isolates) from 11 HAdV-A31 isolates from pediatric HSCT units in three other countries and obtained other recent sequences from NCBI GenBank (11). Sequences from Czech Republic, Ireland and the majority of sequences from England had strong epidemiological linkages primarily with their own pediatric HSCT populations and showed few nucleotide within-clade differences (<5), but demonstrated 29 to 150+ SNPs relative to the main Toronto clade of 17 sequences. Interestingly, two of three cases from one American HSCT center, which showed >29 SNPs relative to our cluster, showed high similarity to each other (<5 SNPs), despite being collected more than 3 years apart. Considered collectively, these findings afforded compelling evidence of healthcare associated transmission of HAdV-A31 over time on our and other HSCT wards.

A surprising finding was the presence of six English outbreak sequences which clustered within the main Toronto outbreak clade. In fact, there were 4 identical contemporaneous HAdV-A31 isolates

identified from Toronto and England. The chronology, along with long-term persistence of the outbreak strain in Toronto but not England, suggests transmission originating in Toronto and arriving in England by an unknown mechanism in 2017. Subsequently, the opposite directionality was observed, when a 2019 English strain differed by 1 SNP from a strain appearing in Toronto in 2021. In addition, we observed a clade of viral genomes from different HSCT units (USA, Ireland, England, and non-outbreak genomes from SK) separated by a range of 1-9 SNPs.

It is possible these results reflect international spread of a particular strain of HAdV-A31. However, limitations of this analysis include the absence of known epidemiological links between units and relatively low total numbers of HAdV-A31 sequences available. Our analysis included both culture enriched and direct sample sequences; viral mutation in culture may lead to “false positive” SNPs and distort the underlying phylogenetic relationships, although our cultures were not subjected to serial passage, thus minimizing mutational events. Reassuringly, we saw no sequence differences for four pairs of direct and cultured sequences from SickKids patients.

Interestingly, we isolated two slightly different strains (=8 SNP) of HAdV-A31 from the same patient at an interval of two years. This suggests possible re-infection with an altered “outbreak” strain that had acquired different mutations by passing through other patients. Alternatively, this profoundly immunosuppressed patient could have harbored distinct subpopulations of HAdV-A31 in the GI tract over 2 years, resulting in the reactivation of a previously undetected strain.

Previous studies have described limited genetic diversity within HAdV-A31, using RFLP of 79 strains from immunocompromised and immunocompetent patients (18) and sequencing of 5 coding regions from 7 strains associated with disseminated disease (19). However, reported genetic differences among strains in both studies are still much greater than the 1-9 SNP differences across entire genomes described above. Thus, it appears unlikely that our observations are an artifact of under-sampling of non-HSCT strains or constrained sequence diversity within HAdV-A31. Rather, this very low “between-country” inter-patient diversity is typical of “within-unit” and “within-outbreak" inter-patient diversity. We are unaware of epidemiological links that could explain the Toronto-London chronology detailed above, but suspect that the findings presented here implicate cryptic international spread of outbreak-causing strains of HAdV-A31, analogous to rare intercontinental transmission events involving *Mycobacterium abscessus* in cystic fibrosis patients resulting in continuous spread of dominant clones in that patient population (20).

Environmental screening by PCR revealed contamination of multiple surfaces in patient rooms and common anterooms. While staff working areas showed little environmental contamination, two thirds of common family areas were contaminated. Coupled with findings of extremely high HAdV-

A31 viral loads in stool and persistence in stool for weeks to months after acute infection in patients who were frequently re-admitted, these findings suggest that environmental contamination and fomites play a significant role in transmission. Closing communal family areas at the peak of the outbreak was thought to help reduce transmission among families, and therefore to patients. We suggest that all spaces, non-clinical or clinical (both in-patient and out-patient), involving interactions between HSCT patients and/or family, should be managed with rigorous infection prevention practices.

Enhancements were made to IPAC protocols including placing all adenovirus positive patients on contact precautions, increased cleaning of adenovirus positive patient rooms and anterooms, and ongoing education of staff and families, emphasizing the role of exporting adenovirus from infected patient's rooms to other areas of the ward or clinic. In addition, the microbiology laboratory implemented routine genotyping of all new adenovirus cases on the HSCT unit and enhanced adenovirus testing of stool, urine and nasopharyngeal specimens. A portable disinfection system using aerosolized H_2O_2 in silver nitrate (Nocospray^{TM}) was introduced in July 2018. Collectively, IPAC measures appeared to achieve cessation of HAdV-A31 infections on several occasions, for periods between 8 months and more than 2 years duration, but despite this, the outbreak has continued until very recently.

We observed a more severe clinical phenotype with HAdV-A31 and HAdV-C1 infections on our unit. This was supported by more frequent viremia/systemic infection (21) in HAdV-A31 infections, more frequent positivity at other body sites, higher maximum plasma viral loads, and longer duration of viremia. Correspondingly, two thirds of HAdV-A31 and HAdV-C1 infections received antiviral therapy with cidofovir and/or brincidofovir, compared to 36% of other genotypes. Despite the clinical severity of HAdV-A31 infections in our cohort, none of these patients died in the first 90 days post-transplant.

Significantly higher median viral loads were found in stool in HAdV-A31 patients than with other genotypes, supporting the fecal-oral route as a primary means of transmission in our outbreak. Putative virulence factors for HAdV-A31 include unique motifs in the E1, E3 and protein IX regions of 7 HAdV-A31 isolates from disseminated infections, which may mediate enhanced escape from immune surveillance, viral persistence and promotion of promiscuous tropism for various tissues and enhanced dissemination(19). These may allow HAdV-A31 to establish a niche in pediatric HSCT patients, where primary and reactivation adenovirus infections are common.

Several pediatric HSCT adenovirus outbreaks have been associated with HAdV-A31(5, 7, 8). Interestingly, in both our study and Myer's(11), retrospective sequencing identified closely related

HAdV-A31 isolates more than two and four years, respectively, prior to suspicion of an outbreak. Nosocomial transmission may therefore persist undetected over long periods of time. In a survey of 12 North American pediatric HSCT centers, none routinely performed adenovirus genotyping (personal communication R. Fattouh, Toronto, 2017). Genotyping was available on request in only five centers. Notably, four centers had suspected but not proven the presence of an outbreak. We contend that nosocomial spread of adenovirus within pediatric HSCT wards may be commonplace and advise implementation of routine adenovirus genotyping in pediatric HSCT centers so that ongoing and future outbreaks can be identified rapidly. To complement this, we suggest, in addition to weekly blood screening by PCR, prompt screening of stool and urine at first positive blood PCR, regardless of symptoms, to reveal the extent of dissemination, potential sources of environmental contamination, and to aid genotyping. Weekly stool monitoring by quantitative PCR may detect infection earlier (22).

Finally, we strongly recommend routine whole genome sequencing for clusters of adenoviral infections of a single genotype, as a clinical tool in adenovirus outbreaks. As experience with WGS increases and more sequences are available in public databases, it will be much easier to make a critical clinical call of nosocomial transmission. This is particularly important for adenovirus, as conventional epidemiological evidence for chains of transmission may be circumstantial or not demonstrable, and yet the virus may persist through multiple patients, across continents, and over a number of years.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the following people who provided invaluable assistance to the investigation and management of this outbreak, and support of this research project: Farhad Gharabaghi (technical support, SickKids), Nursrin Dewsi (technical support, SickKids), David Venhuis (technical support, SickKids), Cillian De Gascun (HAdV-A31 strains, Dublin, Ireland), Jonathan Dean (HAdV-A31 strains, Dublin, Ireland), Adriana Kajon (HAdV-A31 sequences, Lovelace Biomedical Research Institute, New Mexico, USA), Brian Fisher (HAdV-A31 sequences, Children's Hospital of Philadelphia Research Institute, Pennsylvania, USA), Eddie Chong-King (technical support, PHOL), Kathleen Magee (Nursing leadership support, SickKids), Jennifer LaRosa (Nursing leadership support, SickKids), Donna Wall (manuscript review, SickKids), the Bone Marrow Transplant Team (patient management and outbreak control, SickKids), Travis Murphy (Sanger sequencing, NML), Shaun Tyler (Sanger sequencing, NML), Gwyneth MacMillan (statistical analysis, McGill University). We thank

Charlotte Houldcroft and Judith Breuer (Great Ormond Street Hospital, London UK) for providing sequence files prior to upload to NCBI GenBank.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

DISCLOSURES

The authors of this manuscript have no conflicts of interest to disclose

REFERENCES

1. Lion T. 2014. Adenovirus infections in immunocompetent and immunocompromised patients. Clin Microbiol Rev 27(3):441-62. doi: 10.1128/CMR.00116-13. PubMed PMID: 24982316; PubMed Central PMCID: PMC4135893.
2. Seidemann K, Heim A, Pfister ED, Köditz H, Beilken A, Sander A, Melter M, Sykora KW, Sasse M, Wessel A. 2004. Monitoring of adenovirus infection in pediatric transplant recipients by quantitative PCR: report of six cases and review of the literature. Am J Transplant 4(12):2102-8. doi: 10.1111/j.1600-6143.2004.00631.x. PubMed PMID: 15575915.
3. Kampmann B, Cubitt D, Walls T, Naik P, Depala M, Samarasinghe S, Robson D, Hassan A, Rao K, Gaspar H, Davies G, Jones A, Cale C, Gilmour K, Real M, Foo M, Bennett-Rees N, Hewitt A, Amrolia P, Veys P. 2005. Improved outcome for children with disseminated adenoviral infection following allogeneic stem cell transplantation. Br J Haematol 130(4):595-603. doi: 10.1111/j.1365-2141.2005.05649.x. PubMed PMID: 16098075.
4. Khandelwal P, Millard HR, Thiel E, Abdel-Azim H, Abraham AA, Auletta JJ, Boulad F, Brown VI, Camitta BM, Chan KW, Chaudhury S, Cowan MJ, Angel-Diaz M, Gadalla SM, Gale RP, Hale G, Kasow KA, Keating AK, Kitko CL, MacMillan ML, Olsson RF, Page KM, Seber A, Smith AR, Warwick AB, Wirk B, Mehta PA. 2017. Hematopoietic Stem Cell Transplantation Activity in Pediatric Cancer between 2008 and 2014 in the United States: A Center for International Blood and Marrow Transplant Research Report. Biol Blood Marrow Transplant 23(8):1342-9. Epub 2017/04/24. doi: 10.1016/j.bbmt.2017.04.018. PubMed PMID: 28450183; PubMed Central PMCID: PMC5669065.
5. Leruez-Ville M, Chardin-Ouachée M, Neven B, Picard C, Le Guinche I, Fischer A, Rouzioux C, Blanche S. 2006. Description of an adenovirus A31 outbreak in a paediatric haematology unit. Bone Marrow Transplant 38(1):23-8. Epub 2006/05/15. doi: 10.1038/sj.bmt.1705389. PubMed PMID: 16699529.
6. Mattner F, Sykora KW, Meissner B, Heim A. 2008. An adenovirus type F41 outbreak in a pediatric bone marrow transplant unit: analysis of clinical impact and preventive strategies. Pediatr Infect Dis J 27(5):419-24. doi: 10.1097/INF.0b013e3181658c46. PubMed PMID: 18382384.
7. Swartling L, Allard A, Törlen J, Ljungman P, Mattsson J, Sparrelid E. 2015. Prolonged outbreak of adenovirus A31 in allogeneic stem cell transplant recipients. Transpl Infect Dis 17(6):785-94. Epub 2015/10/07. doi: 10.1111/tid.12443. PubMed PMID: 26284461.
8. Houldcroft CJ, Roy S, Morfopoulou S, Margetts BK, Depledge DP, Cudini J, Shah D, Brown JR, Romero EY, Williams R, Cloutman-Green E, Rao K, Standing JF, Hartley JC, Breuer J. 2018. Use of Whole-Genome Sequencing of Adenovirus in Immunocompromised Pediatric Patients to Identify Nosocomial Transmission and Mixed-Genotype Infection. J Infect Dis 218(8):1261-71. doi: 10.1093/infdis/jiy323. PubMed PMID: 29917114.

9. Ali S, Krueger J, Richardson SE, Sung L, Waespe N, Renzi S, Chiang K, Allen U, Ali M, Schechter T. 2019. The yield of monitoring adenovirus in pediatric hematopoietic stem cell transplant patients. Pediatr Hematol Oncol 36(3):161-72. Epub 2019/04/30. doi: 10.1080/08880018.2019.1607961. PubMed PMID: 31037986.
10. Sarantis H, Johnson G, Brown M, Petric M, Tellier R. 2004. Comprehensive detection and serotyping of human adenoviruses by PCR and sequencing. J Clin Microbiol 42(9):3963-9. doi: 10.1128/JCM.42.9.3963-3969.2004. PubMed PMID: 15364976; PubMed Central PMCID: PMC516336.
11. Myers CE, Houldcroft CJ, Roy S, Margetts BK, Best T, Venturini C, Guerra-Assunção JA, Williams CA, Williams R, Dunn H, Hartley JC, Rao K, Rolfe KJ, Breuer, J. 2021. Using Whole Genome Sequences to Investigate Adenovirus Outbreaks in a Hematopoietic Stem Cell Transplant Unit. Front Microbiol 12:667790. Epub 20210702. doi: 10.3389/fmicb.2021.667790. PubMed PMID: 34276599; PubMed Central PMCID: PMC8284422.

12. Stapleton PJ, Eshaghi A, Seo CY, Wilson S, Harris T, Deeks SL, Bolotin S, Goneau LW, Gubbay JB, Patel SN. 2019. Evaluating the use of whole genome sequencing for the investigation of a large mumps outbreak in Ontario, Canada. Sci Rep 9(1):12615. Epub 2019/08/30. doi: 10.1038/s41598-019-47740-1. PubMed PMID: 31471545; PubMed Central PMCID: PMC6717193.
13. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114-20. Epub 2014/04/01. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID: PMC4103590.
14. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772-80. Epub 2013/01/16. doi: 10.1093/molbev/mst010. PubMed PMID: 23329690; PubMed Central PMCID: PMC3603318.
15. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268-74. Epub 2014/11/03. doi: 10.1093/molbev/msu300. PubMed PMID: 25371430; PubMed Central PMCID: PMC4271533.
16. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47(W1):W256-W9. doi: 10.1093/nar/gkz239. PubMed PMID: 30931475; PubMed Central PMCID: PMC6602468.
17. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 15(11):524. doi: 10.1186/s13059-014-0524-x. PubMed PMID: 25410596; PubMed Central PMCID: PMC4262987.
18. Johansson ME, Brown M, Hierholzer JC, Thörner A, Ushijima H, Wadell G. 1991. Genome analysis of adenovirus type 31 strains from immunocompromised and immunocompetent patients. J Infect Dis 163(2):293-9. doi: 10.1093/infdis/163.2.293. PubMed PMID: 1988511.
19. Hofmayer S, Madisch I, Darr S, Rehren F, Heim A. 2009. Unique sequence features of the Human adenovirus 31 complete genomic sequence are conserved in clinical isolates. BMC Genomics 10:557. Epub 2009/11/25. doi: 10.1186/1471-2164-10-557. PubMed PMID: 19939241; PubMed Central PMCID: PMC2794291.
20. Ruis C, Bryant JM, Bell SC, Thomson R, Davidson RM, Hasan NA, van Ingen J, Strong M, Floto RA, Parkhill J. 2021. Dissemination of Mycobacterium abscessus via global transmission networks. Nat Microbiol 6(10):1279-88. Epub 20210920. doi: 10.1038/s41564-021-00963-3. PubMed PMID: 34545208; PubMed Central PMCID: PMC8478660.
21. Matthes-Martin S, Feuchtinger T, Shaw PJ, Engelhard D, Hirsch HH, Cordonnier C, Ljungman P, Fourth European Congress on Infections in Leukemia. 2012. European guidelines for diagnosis and treatment of adenovirus infection in leukemia and stem cell transplantation: summary of ECIL-4 (2011). Transpl Infect Dis 14(6):555-63. Epub 2012/11/12. doi: 10.1111/tid.12022. PubMed PMID: 23146063.
22. Jeulin H, Salmon A, Bordigoni P, Venard V. Diagnostic value of quantitative PCR for adenovirus detection in stool samples as compared with antigen detection and cell culture in haematopoietic stem cell transplant recipients. Clin Microbiol Infect. 2011;17(11):1674-80. Epub 2011/04/11. doi: 10.1111/j.1469-0691.2011.03488.x. PubMed PMID: 21481083.

SUPPLEMENTARY MATERIAL

Infection Prevention and Control Management /Epidemiology

Pre-outbreak: Routine pre-outbreak measures included screening for communicable diseases on admission and additional precautions as indicated. Routine daily cleaning of patient rooms with 0.5% w/w accelerated hydrogen peroxide wipes (Accel or Oxivir TB, Diversey, Mississauga, Canada) included high touch surfaces, floors, waste bins, bathrooms, and the unit common areas. In addition to targeted application of transmission-based precautions, each patient room received a pre-transplant "scrub", meaning the room was empty and underwent deep cleaning, including walls, before a new patient was admitted.

Outbreak onset and post-outbreak:
A line list for outbreak management was created including documentation of common exposures among infected patients, bed allotment, sharing of anterooms, admission and discharge dates, exposure to common procedures, and potential contact in the outpatient setting. An epidemiologic link was defined for a new case of HAdV-A31 infection as infection which was acquired after admission to the HSCT ward, while an active case of HAdV-A31 was present on the ward.

Enhanced IPAC measures included Contact Precautions for all adenovirus-positive patients, enhanced cleaning and laboratory screening. Cleaning included a one-time full-unit purge and deep clean of all patient rooms, anterooms, bathrooms, common spaces, staff/physician spaces, and equipment. 4.5% accelerated hydrogen peroxide wipes (Rescue, Diversey, Mississauga, Canada) were introduced for deep cleaning of high-touch surfaces in all areas of the unit by dedicated housekeeping staff, twice per day cleaning of all adenovirus positive patient rooms and anterooms (continued to this day), reduction of patient personal belongings to a minimum so that all horizontal surfaces in rooms could be easily accessed by cleaners, and closure of common areas to patients and family. Once those spaces reopened, the frequency of cleaning was increased, as housekeeping resources allowed. Deep cleaning was repeated intermittently when there was a high burden of adenovirus on the inpatient unit.

Accelerated hydrogen peroxide 4.25% (Virox Technologies Inc, Oakville, Canada) using hydrogen peroxide via Nocospray^{TM} (AMG Medical, Montreal, Canada), was added to the disinfection regimen in July, 2018. All patient rooms (adenovirus positive or negative) are treated with Nocospray during the pre-BMT room scrub/deep clean and on patient discharge or transfer. Common areas (e.g. family kitchen, patient shower room) receive a Nocospray treatment when the burden of adenovirus-infected patients on the unit is elevated.

In the HSCT outpatient follow-up clinic, similar infection control procedures were put in place, i.e. Contact Precautions for all adenovirus-positive patients, isolation in a single room upon arrival, and terminal clean with 4.5% accelerated H_2O_2 wipes of the examination room at the end of their appointment. Patients with diarrhea were provided a dedicated bathroom when possible.

Important components of outbreak control included daily Environmental Services rounds, regular multidisciplinary meetings and continuing education of staff and families, especially with respect to emphasizing the role of exporting adenovirus from infected patient's rooms on hands and shared equipment to other areas of the ward or clinic.

Table S1: Demographics summary of all HAdV cases included in the phylogenetic analysis

Country of origin	HSCT patient	Patient #	Year_Month_Day first isolate	Year_Month_Day WGS sample	# days 1st pos to WGS sample	Sample code: HAdVi or HAdVs (i=isolate, s=sequence) genotype city.country ISO week.year Case#	WGS successful	WGS sample type - direct and/or culture	Viral load genome copies/ml	GenBank accession #	Chronology (pre-outbreak, outbreak, post-outbreak)	Within Toronto clade by WGS
Canada	yes	20	2012/07/03	2012/07/03	0	HAdVi A31 Toronto.CAN 27.2012 Case20	yes	stool culture	Unknown (EM pos)	OM112280	Pre-outbreak	yes
Canada	yes	1	2014/10/27	31/10/2014	4	HAdVi A31 Toronto.CAN 44.2014 Case01	yes	stool culture	Unknown (EM pos)	MN901838	Outbreak	yes
Canada	yes	2	2014/12/29	31/12/2014	2	HAdVi A31 Toronto.CAN 01.2014 Case02	yes	stool culture	Unknown (EM pos)	MN901828	Outbreak	yes
Canada	yes	3	2015/03/30	10/04/2015	11	HAdVi A31 Toronto.CAN 09.2015 Case03	yes	stool culture	4.3 x 10^9	MN901816	Outbreak	yes
Canada	yes	4	2015/04/27	30/04/2015	3	HAdVs A31 Toronto.CAN 18.2015 Case04	yes	urine - direct and culture	9.3 x 10^7	MN901810	Outbreak	yes
Canada	yes	5	08/05/2015	12/05/2015	4	HAdVs A31 Toronto.CAN 19.2015 Case05	yes	urine - direct and culture	5.6 x 10^9	MN901825	Outbreak	yes
Canada	yes	6	05/06/2015	12/06/2015	7	HAdVi A31 Toronto.CAN 23.2015 Case06	yes	stool culture	Unknown (EM pos)	MN901818	Outbreak	yes
Canada	yes	7_1	2015/06/22	23/06/2015	1	HAdVi A31 Toronto.CAN 26.2015 Case07-1	yes	urine - culture and direct	2.2 x 10^7	MN901815	Outbreak	yes
Canada	yes	8	2015/10/05	13/10/2015	8	HAdVi A31 Toronto.CAN 41.2015 Case08	yes	urine - culture and direct	2.3 x 10^6	MN901824	Outbreak	yes
Canada	yes	9	2015/12/28	29/12/2015	1	HAdVi A31 Toronto.CAN 53.2015 Case09	yes	stool culture	1.7 x 10^8	MN901807	Outbreak	yes
Canada	yes	10	2016/08/16	16/08/2016	0	HAdVi A31 Toronto.CAN 33.2016 Case10	yes	stool culture	3 x 10^5	MN901830	Outbreak	yes
Canada	yes	11	10/10/2016	11/10/2016	1	HAdVi A31 Toronto.CAN 41.2016 Case11	yes	stool culture	5.6 x 10^7	MN901820	Outbreak	no
Canada	yes	19	2016/11/01	10/11/2016	9	HAdVi A31 Toronto.CAN 44.2016 Case19	yes	stool culture	5.6 x 10^7	MN901813	Outbreak	yes
Canada	yes	24	07/02/2017	N/A		N/A	no	stool - culture and direct	3.6 x 10^8	N/A	Outbreak	N/A
Canada	yes	7_2	2017/02/21	21/02/2017	0	HAdVi A31 Toronto.CAN 08.2017 Case07-2	yes	stool culture	6 x 10^10	MN901819	Outbreak	yes
Canada	yes	12	13/03/2017	14/03/2017	1	HAdVi A31 Toronto.CAN 11.2017 Case12	yes	stool culture	1.9 x 10^8	MN901840	Outbreak	no
Canada	yes	13	2017/04/27	27/04/2017	0	HAdVi A31 Toronto.CAN 17.2017 Case13	yes	stool culture	6 x 10^9	MN901821	Outbreak	yes
Canada	yes	14	2017/05/29	29/05/2017	0	HAdVi A31 Toronto.CAN 22.2017 Case14	yes	stool culture	4.7 x 10^9	MN901832	Outbreak	yes
Canada	yes	15	2017/08/07	08/08/2017	1	HAdVi A31 Toronto.CAN 32.2017 Case15	yes	stool culture	1.2 x 10^10	MN901833	Outbreak	yes
Canada	yes	17	2018/02/21	21/02/2018	0	HAdVs A31 Toronto.CAN 08.2018 Case17	yes	stool direct	10^8	MN901809	Outbreak	yes
Canada	yes	16	2018/04/04	04/04/2018	0	HAdVi A31 Toronto.CAN 16.2018 Case16	yes	stool culture	1.9 x 10^5	MN901839	Outbreak	yes
Canada	yes	18	17/09/2018	23/09/2018	6	HAdVi A31 Toronto.CAN 34.2018 Case18	yes	stool culture	1.4 x 10^10	MN901814	Outbreak	no
Canada	yes	25	23/11/2019	N/A		N/A	no	stool culture	2.2 x 10^4	N/A	Post-outbreak	N/A
Canada	yes	21	01/08/2021	01/08/2021	0	HAdVi A31 Toronto.CAN 32.2021 Case21	yes	stool culture	7.6 x 10^9	OM112281	Post-outbreak	yes
Canada	yes	22	23/08/2021	23/08/2021	0	HAdVi A31 Toronto.CAN 35.2021 Case22	yes	stool culture	1.0 x 10^10	OM112283	Post-outbreak	yes
Canada	yes	26	06/09/2021	N/A		N/A	no	stool culture	9.3 x 10^9	N/A	Post-outbreak	N/A
Canada	yes	23	20/09/2021	29/09/2021	9	HAdVi A31 Toronto.CAN 39.2021 Case23	yes	stool culture	3.4 x 10^6	OM112282	Post-outbreak	no
Canada	no	27	1982-00-00	1982-00-00		HAdVi A31Toronto.CAN 1982	yes	stool culture	Unknown	MN901835	Historical isolate	no
Czech Republic		1	N/A	18/10/2016		HAdVi A31 CZE 42.2016 Case01	yes	stool culture	Unknown	MN901829		
Czech Republic		2	N/A	31/10/2016		HAdVi A31 CZE 44.2016 Case02	yes	stool culture	Unknown	MN901836		
Czech Republic		3	N/A	21/11/2016		HAdVi A31 CZE 47.2016 Case03	yes	stool culture	Unknown	MN901811		
Czech Republic		4	N/A	28/12/2016		HAdVi A31 CZE 52.2016 Case04	yes	stool culture	Unknown	MN901822		
USA		1	N/A	2008-00-00		HAdVi A31 Philadelphia.USA 2008 Case01	yes	DNA from stool culture	Unknown	MN901812		
USA		2_1	N/A	2011-00-00		HAdVi A31 Philadelphia.USA 2011 Case02-1	yes	DNA from stool culture	Unknown	MN901834		
USA		2_2	N/A	2011-00-00		HAdVi A31 Philadelphia.USA 2011 Case02-2	yes	DNA from nasopharyngeal aspirate culture	Unknown	MN901823		
USA		3	N/A	2012-00-00		HAdVi A31 Philadelphia.USA 2012 Case03	yes	DNA from stool culture	Unknown	MN901808		
Ireland		1	N/A	2016-07-00		HAdVi A31 Dublin.IRL 29.2016 Case01	yes	stool culture	Unknown	MN901837		
Ireland		2	N/A	2016-08-00		HAdVi A31 Dublin.IRL 31.2016 Case02	yes	stool culture	Unknown	MN901831		
Data below was extracted from NCBI Genbank to allow renaming of the English sequences published in Front Microbiol 12, 667790 (2021) using the sample code in this study												
England		Pt18		01/06/2015		HAdVs A31 London.ENG 23.2015 Pt18				MW686757		
England		Pt3		06/07/2015		HAdVs A31 London.ENG 28.2015 Pt3				MW686758		
England		Pt6		05/04/2012		HAdVs A31 London.ENG 14.2012 Pt6				MW686759		
England		Pt8		30/06/2011		HAdVs A31 London.ENG 26.2011 Pt8				MW686760		
England		Pt8		27/06/2011		HAdVs A31 London.ENG 26.2011 Pt8				MW686761		
England		Pt11		23/03/2016		HAdVs A31 London.ENG 12.2016 Pt11				MW686762		
England		Pt11		25/03/2016		HAdVs A31 London.ENG 12.2016 Pt11				MW686763		
England		Pt24		17/09/2015		HAdVs A31 London.ENG 38.2015 Pt24				MW686764		
England		Pt24		21/09/2015		HAdVs A31 London.ENG 39.2015 Pt24				MW686765		
England		Pt24		24/09/2015		HAdVs A31 London.ENG 39.2015 Pt24				MW686766		
England		Pt24		28/09/2015		HAdVs A31 London.ENG 40.2015 Pt24				MW686767		
England		Pt55		04/03/2019		HAdVs A31 London.ENG 10.2019 Pt55				MW686768		
England		Pt52		09/04/2018		HAdVs A31 London.ENG 15.2018 Pt52				MW686769		
England		Pt52		05/04/2018		HAdVs A31 London.ENG 14.2018 Pt52				MW686770		
England		Pt57		18/03/2019		HAdVs A31 London.ENG 12.2019 Pt57				MW686771		
England		Pt29		16/06/2015		HAdVs A31 London.ENG 25.2015 Pt29				MW686772		
England		Pt62		22/03/2019		HAdVs A31 London.ENG 12.2019 Pt62				MW686773		
England		Pt68		11/02/2019		HAdVs A31 London.ENG 07.2019 Pt68				MW686774		
England		Pt58		21/01/2019		HAdVs A31 London.ENG 04.2019 Pt58				MW686775		
England		Pt69		14/02/2019		HAdVs A31 London.ENG 07.2019 Pt69				MW686776		
England		Pt41		24/11/2017		HAdVs A31 London.ENG 47.2017 Pt41				MW686777		yes
England		Pt41		01/01/2018		HAdVs A31 London.ENG 01.2018 Pt41				MW686778		yes
England		Pt41		16/03/2018		HAdVs A31 London.ENG 11.2018 Pt41				MW686779		yes
England		Pt41		26/03/2018		HAdVs A31 London.ENG 13.2018 Pt41				MW686780		yes
England		Pt41		01/08/2018		HAdVs A31 London.ENG 31.2018 Pt41				MW686781		yes
England		Pt73		02/04/2019		HAdVs A31 London.ENG 14.2019 Pt73				MW686782		
England		Pt46		02/04/2018		HAdVs A31 London.ENG 14.2018 Pt46				MW686783		yes
England		Pt70		25/02/2019		HAdVs A31 London.ENG 09.2019 Pt70				MW686784		

Discussion

This discussion explores how the three articles above describe work that was performed at a time of rapid technological change in the field of WGS. In the case of mumps virus and adenovirus**,** the articles are among the first efforts to use WGS to aid investigations of outbreaks of the respective pathogens and indeed to generate any significant numbers of genome sequences for publication in open source databases. With respect to *P. aeruginosa,* where the use of WGS to characterise the pathogen was well established, the analysis leveraged cutting edge bioinformatic workflows for quality control, preliminary identification of potential clusters and fine scale SNP typing required to prove transmission.

The importance of demonstrating flexibility in adapting WGS based outbreak analysis to "new" pathogens and the ability to integrate the latest bioinformatic tools was shown during the Covid-19 pandemic which occurred subsequent to the work described in this thesis. During the pandemic, the main challenges of establishing the basic procedures for sequencing SARS-CoV-2 and identifying instances of human to human transmission could be met by adapting the methods described herein, namely amplicon based sequencing from clinical samples and use of new bioinformatic tools to generate phylogenetic trees and establish SNP distances.

The discussion therefore considers the research objectives of each article in turn and explains how WGS was adapted in the specific circumstances, exploring not only the novelty and significance of the specific findings but also the differences and commonalities in approach across the articles.

Finally, a post-script considers the implications of the broader implementation of WGS for transmission based analysis and the benefits of expansion of its role in clinical microbiology in the years ahead.

How the research objectives were met and the strengths and limitations of each article.

Mumps

The objective to develop a novel protocol for successful WGS amplicon sequencing of 26 mumps virus genomes to investigate a large community Mumps outbreak in Ontario was met and demonstrated to be superior to the typing techniques in general use at that time.

- WGS data analysis was able to distinguish outbreak from non-outbreak cases within genotype G, whereas SH gene analysis using traditional Sanger sequencing approach for mumps virus typing had insufficient resolution for this.

- Three distinct outbreak clades (two major and one minor) were distinguishable through SNP analysis and generation of ML phylogenetic trees.

- I performed Bayesian phylogenetic analysis to create a timed tree; this indicated that the two major clades diverged from each other prior to the epidemiological start date of the outbreak. Comparison with international sequences identified links with cases in the USA, suggesting multiple importation events causing what were in fact separate outbreaks rather than 1 outbreak. This supported the conclusions of the Bayesian analysis and was an unexpected finding.

- Details from the epidemiological investigations were integrated with WGS data to visualise the different geospatial distribution of the various clades. A transmission tree modelled how person-to-person transmission might occur. This linked the timed tree with a set of assumptions about the proportion of outbreak cases sequenced, the incubation time and the infectious period. When this transmission network was linked to epidemiological information (e.g. suspected common exposures related to work, private residences and indoor hospitality) it supported some suspected transmission events and refuted others. Limitations were also apparent i.e. the transmission tree displayed some visualisations of inferred transmission events from one person to another that could not be supported by epidemiological information and indeed were considered very implausible when subjected to expert review.

- A secondary objective was to develop a “wet lab” method for performing Mumps amplicon sequencing direct from oral swabs to reduce turnaround time by eliminating the viral culture enrichment step which took between 7 and 17 days. This was not successful because of “drop out” out of significant numbers of amplicons (failed amplification) leading to incomplete sequence generation in test cases. This was not due to inadequate viral template, as all samples had cycle threshold (Ct) values under 33. It did prove possible to successfully optimise the primers to generate long amplicons of 2kb from viral culture supernatant, reducing by half the number of amplicon reactions required to generate a genome compared to previous small scale sequencing efforts, and therefore reducing the number of pipetting steps and overall preparatory time. We also demonstrated the practicality of sequencing viral outbreaks by amplicon method using spare capacity on Illumina flow cells being used to sequence bacteria for typing purposes (the core function of

the sequencing service in PHO Laboratory at the time). Ultimately, though all the sequences in our final report still relied on an initial culture step. As I outlined in the discussion, a seminal paper published by Jonathan Quick around the time the work was completed outlined a modified approach for successfully generating 2kb amplicons direct from clinical samples for a variety of viruses, using optimised thermocycler conditions and automated primer design scheme via a web application for more efficient amplification. This would be a better approach for future Mumps WGS efforts. This prediction was borne out when the method saw extremely widespread uptake during the Covid pandemic to sequence SARS-CoV-2 from clinical samples[1,2]. All laboratories in the Irish Covid Hub and Spoke methods use amplicon based sequencing from clinical samples for Illumina or Nanopore sequencing.

As discussed above, another secondary objective was integrating WGS data with epidemiological information from outbreak investigation and infection transmission dynamics. For the Mumps study this integration did not consider strain comparison in terms of SNP distance, a metric which is an extremely useful way to discuss relatedness of genomes with multiple stakeholders in fields such as microbiology, infectious diseases, infection control, public health and academic research. I did identify SNPs in the results, which I termed SNVs in this article, but as indicated in the thesis introduction the two terms may be considered synonymous for all practical purposes; while SNP has become favoured over time. Mumps SNPs were discussed in terms of distribution across the genome and functional impact, however the actual SNP distances between individual sequences and clades was not used as a core metric for describing transmission. Additionally, the scale bars in the figures displaying maximum likelihood trees of mumps virus sequences were not altered to display tree length differences as (intuitive) SNP distances. Instead, the figures displayed a non-intuitive metric of substitutions per site in the scale bar. By contrast, the subsequent articles in this work that present evidence for or against cross-transmission contain explicit statements of SNP distances. The subsequent articles also present detailed arguments for particular “cut-offs”, whereby SNP distances above the cut-off effectively rule out cross-transmission and below it requires further investigation. Phylogenetic trees in the later articles also display figures where the scale bar is expressed in SNP distance, each such figure requiring manual calculation by correlation of SNP distance matrices from bioinformatic analysis with the default output of ML tree generating software.

Instead of a focus on SNP distances in this article, a complex Bayesian analysis was performed using the program BEAST2 and a transmission tree generated using the R package “Transphylo”. These are labour intensive analyses -optimising the run parameters to produce the final analyses took several

weeks. This limitation, it could be argued, rules out such types of analysis as a practical tool to describe outbreaks in real time using WGS for outbreaks and provide rapid results to inform control measures. Consequently, these techniques were not used in the subsequent articles. Bayesian analysis is however used on an ongoing basis by web hosted real-time surveillance and outbreak investigation software such as Nextstrain, which is described in this paper and is used for a several other pathogens of public health importance. It is therefore desirable to understand how such analyses are constructed and how the results should be interpreted if performing transmission analyses.

The final secondary objective, which was added post hoc, was to incorporate contemporaneous sequences from mumps virus outbreaks in the USA. This led to the unexpected finding of multiple international importation events, and that the Ontario outbreak was distinct from another outbreak in Canada occurring in British Columbia. It necessitated searching for sequences from diverse sources such as laboratory websites and supplementary appendixes to pre-print publications. These are genome sequence equivalents to "grey literature" – they are published outside of mainstream public databases like the Global Initiative on Sharing All Influenza Data (GISAID) or Genbank. Finding and incorporating these sequences added important context to the understanding of the Ontario outbreak. The availability of individual patient level metadata (travel histories and symptom onset dates) from one preprint provided evidence to support a potential USA to Canada importation event. This highlights the importance of open sharing of both sequence data and of associated metadata.

The articles in this work adhere to good research practices for genomics as they all include data availability sections describing how consensus genomes or read sets were updated to open source databases. The main and supplementary results provide links to the accession numbers from such open source databases. Best practice in such cases is to upload consensus genomes or read sets to public databases and linking to the accession numbers from such databases in final publications. It is also preferable to include in supplementary material metadata for each individual accession/sequence, where this is permitted under terms of REB study approval. The need for data sharing and the imperative to protect patient privacy are sometimes seen to be in opposition to one another, which can make the members of an REB concerned about breach of privacy if metadata is included in publication, even where this has been de-identified. For example, the metadata (or even the name of the sequence as it appears in database or publication) may have geographic and temporal information (date and location of case) that in theory could allow for linkage to individual patients. It is necessary however to share carefully de-identified metadata to facilitate genomic

epidemiology by the scientific community, as other scientists and infection control teams are likely to need spatial and temporal information for analyses that include published sequence data as comparators to their own sequences.

In the three main chapters of this thesis, supplementary materials for each describe the metadata for each individual accession, where this was permitted under terms of the relevant REB study approval (which was not the case for the mumps study). Achieving REB approval for the mumps virus WGS project was a protracted process. The REB members expressed concerns about the suitability of including even de-identified patient level infection data such as location where infection occurred (to the level of suburb not street or house/business), and suspected epidemiological links. Their concern was the absence of express patient written consent to use viral genome sequence data for outbreak analysis and for sharing in public databases. Due to delays in study initiation after outbreak declaration, the outbreak had actually terminated by the time REB approval was sought, and so it was initially deemed that the public health interest of this study did not override concerns about express consent for use of healthcare records for research purposes. After support for the use of such data was provided by the local public health department the REB agreed that such use of patient level data was required for the mumps study to have value, and that it was not feasible to obtain individual consent retrospectively. Conversely, for the Pseudomonas study, all the CF participants had previously enrolled in a biobank and no difficulties were encountered obtaining ethical approval once approval from the biobank regulators was granted to access stored material and data. Similarly, in the case of adenovirus there was clearly a pressing need for higher resolution analysis with new techniques as part of a research project to help terminate an ongoing outbreak with significant clinical consequences. It is essential to make the case to members of an REB at the outset that identifying cryptic transmission is important, even retrospectively as this informs future control measures. Indeed, in my experience this argument is rarely disputed, however it must always be paired with the argument that it is imperative to share sequence data and de-identified metadata in the wider public interest. The studies presented here, in particular Mumps and Adenovirus, show the importance of inclusion of such data in local analyses to provide essential context, without which interpretation of local sequence data may be challenging or impossible. Measures to limiting metadata to a level of resolution that provides context but cannot be used to identify an individual should be specified at the outset, for instance specifying the location as a local municipality area rather than street and the date as a week of the year rather than a day.

In summary the stated research objectives were achieved by refining methods for use in further research and clinical applications, by identifying specific types of analyses that would be more or less valuable in future outbreaks and detailing various factors relevant to successful WGS based outbreak

investigation. These factors comprised wet lab approaches, bioinformatics analysis techniques, sourcing comparator sequences, overcoming challenges of metadata linkage and data sharing and how to present the results to collaborators to draw actionable conclusions. A major limitation was the ability to sequence only a fraction of identified infections; reagents and staff time were limited as no specific funding for the project was available. The approach presented remains relevant however, since capacity issues and case ascertainment difficulties will be encountered in many other outbreak scenarios and bioinformatics approaches that take account of this are necessary; some tools not used in this analysis such as outbreaker required that essentially all infections in an outbreak to be sequenced to generate a transmission network analysis.

Finally, as stated in the last paragraph of the article conclusions, it was not possible to state that the benefits of performing WGS analysis routinely for community Mumps virus outbreaks justify the outlays that we identified in terms of time and resources required (even assuming real-time provision of results). This is because limited public health interventions are available and infections are hardly ever fatal, though sometimes result in serious sequelae. It was possible to specify the preparatory work advisable for public health laboratories to ensure capacity to respond with WGS investigation where there was a pressing clinical need. The advent of the Covid pandemic demonstrated that public health systems with experience of this type of WGS outbreak response were well positioned to respond rapidly to a new threat.

P. aeruginosa

Here WGS was used to confirm the hypothesis that cryptic person to person transmission of *P. aeruginosa* occurs in the setting of early infection of children with CF. This transmission was due not to established epidemic clones (lineages that transmit primarily between people with CF who are chronically infected) but rather due to limited strain sharing between small clusters of children. These strains were presumably initially acquired from the general environment and then transmitted from person to person. These findings were made possible by the novel approach of combining WGS for typing with decision to type multiple bacterial colonies (up to 12) from each “new-onset” CF infection in an unbiased manner (not predicated on suspicion of an outbreak due to e.g. abnormal incidence or unusual resistance pattern). We identified the occurrence of what we termed “mixed-strain” infection, which could not be identified by phenotypic examination of colonies. We demonstrated that this finding was associated with strain sharing, at least in one centre (SickKids), although a causal relationship between the mixed-strain infection and strain sharing could not be identified. Given the novel nature of the findings, further studies are required to see if these findings

are reproduced in similar patient populations. It is notable however that a subsequent multicentre study of *Staphylococcus aureus* infections in a similar CF patient population, using similar methodological concepts, also identified a hitherto underappreciated role for early mixed-strain infection and strain sharing[3]. Though a multiple colony WGS approach is indeed resource intensive and challenging to implement for cost reasons, the findings clearly demonstrate that cross-transmission in this context is likely to be under-appreciated or missed entirely unless an unbiased, WGS based approach accounting for intra-host diversity is used.

Regarding secondary objectives, over a thousand bacterial isolates were recovered and frozen from biobanked samples using standard sputum culture methods and several hundred picked as appropriate colony representatives in proportion to underlying morphotypic diversity. This can be considered "deep sampling" in the sense that the true intra-host strain diversity can be identified by this method, compared with the converse of "superficial sampling" where a single colony or single colony of each morphotype is picked as a representative for sequencing. This approach was aided by the fact that sputum for children with early onset infections (3 x 2ml tubes) had been frozen as standard by the clinical microbiology laboratory since 2011 since the commencement of clinical studies examining why chronic infection phenotypes of *P. aeruginosa* are associated with failure of eradication in children with CF. In Toronto, at least 99% of children with CF are registered with the Toronto Cystic Fibrosis registry, which prospectively captures clinical data in a secure redcap database. All participants have agreed to have their data collected through this registry and have agreed for any potential use of data for research purposes. Access to the sputum biobank was governed by the primary investigators on the original studies Dr Waters and Dr Yau and was primarily for studies originating from the local research group. The CF registry by contrast held data that was open to all researchers providing that the access was approved by Registry Review Panel, comprised of Canadian CF clinicians and researchers, and supported by local REB approval letter with safeguards around confidentiality and data retention. The combination of a bank of clinical samples (3 x 2ml tubes of residual sputum from each sample collected from a child with CF and stored at -80 degrees centigrade in the research institute adjacent to the hospital) and a complete dataset of clinical data permitted easy linkage of sequence data to metadata.

Several hundred other isolates previously frozen from infections in other clinical contexts (chronic CF and non-CF infections) formed the comparator collections of chronic CF infection. We hypothesized environmental reservoirs in CF ward and clinic might be relevant to cross-transmission and developed a method to sample biofilms from sink drains in the clinical areas and culture them for

Pseudomonas. We then identified *P. aeruginosa* strains and stored them for sequencing. The resulting isolate collection was sequenced using standard Illumina bacterial sequencing protocols on multiple runs of a high throughput Nextseq instrument. Simultaneously the bioinformatic methods were developed to deal with a large isolate collection.

The secondary objective of identifying where cross-transmission may be occurring through the linkage of WGS data to epidemiological investigations was less successful. A minority of identified cross-transmission clusters did have supporting epidemiological evidence to support the WGS evidence of transmission. Some interesting vignettes are reported in the article. This included transmission from a child chronically infected with *P. aeruginosa* to other patients, likely occurring on the inpatient ward. There was evidence that another transmission event was responsible for the chronic colonisation of the recipient, a potentially life altering event,. Finally intermittent detection of one shared strain in different patient cohorts and the hospital environment over several years was observed. It was also possible in the context of certain hypermutator strains to prove the directionality of transmission from one patient to another. However, for the majority of shared strain clusters no epidemiological link was identified beyond the fact that they all received their ongoing clinical care in SickKids. There was no evidence of community links. The failure to identify the main sources of cross-transmission limited the ability to recommend mitigating interventions. The article did not demonstrate statistically significant evidence of adverse clinical outcomes for patients who have acquired a shared strain, in term of success of eradication therapy. Nonetheless, the findings have clinical significance as they demonstrate the origin of some cases of chronic *P. aeruginosa* infection is in childhood cross-transmission. Also, the demonstration of relatively frequent cryptic cross-transmission for this pathogen by using high resolution techniques generates hypotheses as to whether this occurs for other pathogens also and methods by which that could be identified. A few instances of superinfection were also identified, where a child who was chronically Pa infected became co-infected with a different strain of Pa, at a later point in time, and this did not always supplant the original strain.

This article used improved methods compared with chapter two (mumps), specifically the figures illustrated the SNP distances between isolates using information conveyed by scale markers on phylogenetic trees. This article also displayed the relevant epidemiological information in the form of Gannt charts showing overlapping admissions, or diagrams of the ward showing positive environmental reservoirs.

The results of the environmental sampling, another secondary objective, failed to provide robust support for the hypothesis that sinks served as a reservoir for early infection strains. Relatively little

P. aeruginosa was recovered from biofilms obtained from the sink drains (of clinical handwash basins and general sinks) located on the inpatient ward, and none was recovered from the CF clinic and pulmonary function test area. This environmental sampling was undertaken many years after the early CF infections occurred, so there was potential for disappearance of bacterial reservoirs over time confounding the results. Findings from inpatient ward sink cultures were relevant to cross-transmission in one instance.

A bespoke bioinformatic workflow was developed to efficiently construct and search phylogenetic trees with >1000 bacterial genome assemblies. The specific methods are detailed in the article, but I note analysis at this scale would have been impossible without the use of high performance computing (HPC) cluster at University of Toronto to generate bacterial genome assemblies, to group them using mashtree, then identify SNPs in clusters using SNVPhyl. This necessitated using Unix command line (the required interface with all high performance computing clusters) for installing and running software, submitting analysis jobs to compute clusters and iteratively refining the methods over time. The first step in the bioinformatics pipeline development was to design rigorous quality control analysis to exclude genomes contaminated before or during sequencing from the analysis. This necessitated resequencing isolates in some instances. This Quality Control (QC) approach was detailed in extensive supplementary methods. The other bioinformatic methods were described with the specific purpose that the entire analysis is reproducible using freely available, open source software by another research group on the dataset, or on their own dataset. This was to avoid a situation where the software used was tied to one particular laboratory computing environment and would thus not be repeatable by other investigators. The sequencing reads from the project were uploaded to the short read archive (SRA), this permits other researchers to repeat the analysis with the same or different methods to see if they arrive at the same conclusions.

The article reports use of multiple strands of evidence to support the assertion that a SNP distance of four or fewer SNPs was appropriate in the local context as a cut-off for defining a shared strain. The evidence detailed in supplementary methods included directly observed within host diversity and longitudinal diversity from serial sampling of patients, but also justified exceptions to the rule. Exceptions were situations where the cut-off was not applicable because of specific circumstances which could be determined from the WGS data (presence of hypermutator strains). The result was the first attempt to codify a WGS based SNP cut-off for identifying shared *P. aeruginosa* strains in CF infection.

For the final objective, determining if cross-transmission was associated with adverse outcomes for patients, no association was found for an impact of either mixed-strain or shared-strain infection on

clinical outcomes (failure of antibiotic eradication therapy), though progression to chronic infection as a result of transmission events was observed. Nonetheless, the findings were sufficiently concerning to warrant a renewed focus on adherence to recommended infection control practices in CF units including in the outpatient clinic[4] .

In conclusion, this study represented a progression from the first article in multiple domains: bacterial rather than viral pathogens, a greater wet lab commitment (though focused on basic bacteriology techniques rather than DNA extraction and sequencing), hundreds more genomes, more complex quality control and analysis steps, more detailed metadata with respect to possible cross-transmission links and the generation of more detailed figures that combined SNP distances and epidemiological details. It demonstrated again that the most challenging aspect of sequencing in clinical microbiology labs is the post sequencing workflow, especially defining clusters of genomes that might represent cross-transmission.

Adenovirus

For the final article the objective was to investigate a hospital outbreak of hAdV-A31, one which initially was detected in 2015, although ultimately, we determined that the first case had occurred in 2012 and the outbreak strains we identified persisted until 2021. This is the longest duration hAdV-A31 outbreak in a paediatric HSCT unit yet described and amongst the first study to use WGS for this pathogen (none had been reported when we commenced our analysis). I demonstrated conclusively the presence of an outbreak strain persisting over years, complicated by interspersed non-outbreak strain hAdV-A31 infections at times. It was possible to resolve transmission details within hAdV-A31 such as subclades of sequences differing by 0 to 1 SNPs, correlating with bursts of outbreak activity that were suspicious for point source dissemination. Transmission persisted despite all patients residing in positive pressure isolation rooms throughout admission with application of strict cross-transmission precautions, and while ward staff and parents were observed by infection control practitioners to adhere closely to recommended practices. The increase in hAdV-A31 above baseline over such a prolonged period initially prompted consideration that the cases burden might reflect independent reactivation of endogenous hAdV-A31, or parental introduction of diverse hAdV-A31 strains, in the context of immunosuppression from transplant and occasional waves of community adenovirus infection of various genotypes. WGS analysis conclusively refuted this by going beyond the genotype level to demonstrate that within hAdV-A31 the majority of sequences formed a clade of closely related sequences (pairwise minimum distance 0 to 8 SNPs) with occasional diverse “non-outbreak” infections occurring that were dozens or hundreds of SNPs different. Due to the absence

of any criteria for defining closely related strains, it was necessary to define how an outbreak strain could be identified and discriminated from sporadic cases. It was challenging to establishing SNP thresholds for this genotype as on commencing the analysis only two complete hAdV-A31 genomes were available in Genbank for comparison. The paper details how incorporation of external comparator genomes into the analysis was required to contextualise the SNP distances found in our single centre population. To achieve this, AdV-A31 isolates from paediatric HSCT patients in Czech Republic and USA, as well as clinical material from colleagues in Ireland were analysed. Genome sequences were obtained from researchers in the UK who kindly shared their non-annotated genomes on request. They had recently described WGS of cases in a similar setting (though with limited evidence for transmission) but encountered technical difficulties making their annotated genomes publicly available[5]. The final manuscript incorporates the fully annotated UK hAdV-A31 genomes that were eventually successfully published in Genbank.

A major unexpected finding that resulted from the international comparison was the identification of two hAdV-A31 clades containing closely related isolates from paediatric HSCT units in different countries, raising the possibility of international dissemination of certain strains. There was some circumstantial epidemiological evidence that suggested export of our outbreak strain from SickKids to the paediatric HSCT unit in London (Great Ormond Street Hospital) with subsequent onward transmission there to another patient. It was not possible to establish how this may have occurred (contaminated bone marrow transplant material from registries was considered and excluded by the transplant team and no direct patient transfers unit to unit). A strain present in London, Dublin, Philadelphia was also identified associated with a single infection in Toronto. It was not possible to explain how this finding occurred in the absence of much larger scale hAdV-A31 genome sequencing availability to determine the true underlying genomic diversity of this pathogen. Overall, the paper demonstrated hAdV-A31 is uniquely disposed towards causing outbreaks in this patient population. No inter- or intra-hospital outbreak strains of hAdV-C1 or hAdV-C2 genotypes were found Chronic transmission within and between centres appears to be playing a very substantial role.

A key limitation of this study was that the mechanism of transmission of adenovirus in this outbreak remained unknown. Heavy environmental contamination of common areas for patient families was demonstrated during the initial outbreak stages. Initially this was with non-WGS genotyping.

Environmental contamination of patient rooms was unsurprisingly also present and persisted (though at declining assessed viral load) despite repeated cleaning efforts, suggesting a role for cross-transmission from fomites. We also did not consider at the time sampling staff or family members for adenovirus for the possibility of episodic asymptomatic or minimally symptomatic

infection e.g. using nasopharyngeal and throat swabs (staff and parents were ordered to leave the unit if even minimal symptoms developed and compliance deemed high by IPC staff).

Subsequently, the substantial role played by asymptomatic and pre-symptomatic viral transmission by individuals in the community was established for SARS-CoV-2 [6]. Extrapolating from this evidence, testing of asymptomatic individuals in healthcare (patients and staff) for outbreak prevention purposes was sometimes practiced during the early stages of the Covid-19 pandemic, though more recently this has been discouraged by healthcare associated infection prevention societies in the context of the changing epidemiology of Covid-19[7]. In future hAdV-A31 outbreaks, testing and sequencing in this context should be considered if no other source of infection is apparent.

Secondary objectives are discussed next. The first was to describe the clinical and laboratory features of the outbreak. The clinical findings were described above. This was a multi-system infection with multiple body sites positive for hAdV-A31 on testing, usually detected first in blood asymptomatically as part of weekly surveillance efforts. In some cases this proceeded to multi-system infection including respiratory, gastrointestinal and urinary tract PCR positivity. Respiratory or other localising symptoms were uncommon. Virus was detected in some nasopharyngeal aspirate specimens although virus loads were higher in stool samples, which proved to be better samples for metagenomics or viral culture enriched sequencing discussed further below. We demonstrated that hAdV-A31 is an adenovirus genotype with particular propensity to cause disseminated infection requiring treatment. We sequenced genomes from six other genotypes and we described hAdV-C1 or hAdV-C2 clinical features in detail. We also demonstrated there was no inter- or intra-hospital outbreak strains detectable by WGS for other genotypes, notably the other two most common infections, hAdV-C1 or hAdV-C2. We described how hAdV-A31 had more concerning laboratory (viral load) and clinical features than other genotypes.

Another secondary objective was to describe the use of whole-genome sequencing in delineating local outbreak cases. A viral culture enrichment step was used for most cases and when this failed a direct metagenomics approach was used (shotgun sequencing of all DNA present in the clinical sample). This often failed to generate adequate viral sequence for the reasons outlined in the discussion of the mumps paper, namely competition from human DNA. We did not opt to develop a culture free method of enrichment from clinical sample, such as the amplicon method developed for Mumps virus, but such a method (oligonucleotide bead based capture enrichment) was described by others after the bulk of our sequencing was complete[8]. Though the bulk of the sequencing was performed on Illumina MiSeq for three 2021 hAdV-A31 cases the project utilised a new hospital laboratory microbiology pathogen WGS service in SickKids. Our group had attempted a test

sequencing run of five hAdV-A31 positive environmental swab samples in SickKids in 2019, using a Illumina instrument sited in the human genome diagnostic lab. Though this experiment failed (and description of method and negative results removed from the manuscript during the revision process at peer reviewer suggestion), it was a first effort to perform WGS at SickKids for clinical microbiology purposes. In 2021 it was possible to use a local Oxford Nanopore MinION device to sequence hAdV-A31 cases from the HSCT unit. An enrichment step was still needed, and this was attained by using culture supernatant provided by PHO Laboratory. This further sequencing effort pushed back the date of final detection of an outbreak case out from 2018 to 2021. It also informed the recommendations in the article that in addition to performing active surveillance for asymptomatic infection on paediatric HSCT patients and genotyping of cases, WGS should be performed on all identified hAdV-A31 cases to identify cryptic outbreaks.

We did not attempt to identify specific genomic markers that could account for the apparent increased virulence and transmissibility of hAdV-A31 versus other genotype. By publishing dozens of new annotated hAdV genome assemblies on Genbank, accompanied by de-identified metadata for individual cases in the supplement (with REB approval), it is possible for other specialised groups specifically investigating adenovirus function to leverage the sequences generated for such purposes. We also sequenced virus from most patients at a single time point only i.e. we did not include many longitudinally collected isolates. We had to prioritise limited WGS capacity and chose breadth rather than depth at an early stage of study design. As with the previous two studies, completing analysis required a deep familiarity with UNIX command line bioinformatic tools to create alignments of nearly a hundred genomes, perform QC, identify SNPs, visualise the SNPs as to where they were occurring in the genome, and annotate the consensus genomes for eventual Genbank upload and release. The final secondary objectives were already addressed above; namely exploring the relationship of SickKids virus hAdV-A31 genomes to international strains and proposing strategies for optimal monitoring of patients and role of WGS.

The final article represents the most complex project of this work and the refinement of approaches developed earlier in the thesis. The clinical scenario combined the virus outbreak dynamics of the first article with the detailed protracted hospital outbreak epidemiological investigation of the second. A variety of sequencing approaches (culture enrichment and metagenomics, Illumina short read and Nanopore long read, public health and hospital laboratory based sequencers) were employed. Samples from more than one hospital and region were sequenced and the accompanying clinical metadata was richer than that in the two previous articles. The bioinformatics analysis and results presentation combined elements of two previous articles by incorporating whole genome assemblies from other outbreaks obtained from a variety of sources (as with mumps virus) and

generating SNP cut-offs and phylogenetic trees with scale bars specified in SNP distance as with *P. aeruginosa*. The findings and recommendations were the most novel of any of the articles presented, specifically that cryptic, high impact hAdV-A31 can persist far longer than suspected and that international transmission of outbreak strains appears to occur by mechanisms unknown, and that routine surveillance and WGS should be carried out. The total body of work contained in this thesis shows the potential benefits of a routine WGS based approach to transmission investigations and details the necessary methods to implement it, potentially within a hospital microbiology laboratory, for this and other respiratory pathogens.

REFERENCES

1. Lucey M, Macori G, Mullane N, et al. Whole-genome Sequencing to Track Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Transmission in Nosocomial Outbreaks. *Clin Infect Dis*. 2021;72(11):E727-E735. doi:10.1093/CID/CIAA1433
2. Mallon PWG, Crispie F, Gonzalez G, et al. Whole-genome sequencing of SARS-CoV-2 in the Republic of Ireland during waves 1 and 2 of the pandemic. *medRxiv*. Published online February 10, 2021:2021.02.09.21251402. doi:10.1101/2021.02.09.21251402
3. Long DR, Wolter DJ, Lee M, et al. Polyclonality, Shared Strains, and Convergent Evolution in Chronic Cystic Fibrosis Staphylococcus aureus Airway Infection. *Am J Respir Crit Care Med*. 2021;203(9):1127-1137. doi:10.1164/RCCM.202003-0735OC
4. Sala MA, Jain M. Pseudomonas aeruginosa and Children With Cystic Fibrosis. *Clinical Infectious Diseases*. 2021;73(9):e2529-e2530. doi:10.1093/CID/CIAA765
5. Houldcroft CJ, Roy S, Morfopoulou S, et al. Use of Whole-Genome Sequencing of Adenovirus in Immunocompromised Pediatric Patients to Identify Nosocomial Transmission and Mixed-Genotype Infection. *J Infect Dis*. 2018;218(8):1261-1271. doi:10.1093/INFDIS/JIY323
6. Muller CP. Do asymptomatic carriers of SARS-COV-2 transmit the virus? *The Lancet Regional Health - Europe*. 2021;4. doi:10.1016/j.lanepe.2021.100082
7. Talbot TR, Hayden MK, Yokoe DS, et al. Asymptomatic screening for severe acute respiratory coronavirus virus 2 (SARS-CoV-2) as an infection prevention measure in healthcare facilities: Challenges and considerations. *Infect Control Hosp Epidemiol*. 2023;44(1):1-6. doi:10.1017/ICE.2022.295
8. Wohl S, Metsky HC, Schaffner SF, et al. Combining genomics and epidemiology to track mumps virus transmission in the United States. *PLoS Biol*. 2020;18(2):e3000611. doi:10.1371/JOURNAL.PBIO.3000611

Post-Script

Key challenges to implementing WGS in hospital laboratories

This work explored how WGS can fill an unmet need for high resolution pathogen typing. There are also other highly promising future applications in hospital clinical microbiology laboratories such as identification of virulence factors[1] or the use of sequence data to predict antimicrobial resistance[2]

WGS is now implemented widely in Irish hospital clinical labs, partly as a result of the experience of the Covid pandemic and the establishment of a national SARS-CoV-2 sequencing program, discussed further below. To fully realise its potential, certain key challenges remain. These do not stem from a technical difficulty with producing sequencing data. Rather, as presented in the thesis, the challenges are:

1. A specific skillset and careful selection of methods is required to generate a valid analysis once pathogen WGS data has been generated, in order to have confidence in the conclusions generated from data analysis
2. The analysis must be integrated with an overarching clinical or public health context to test hypotheses and generate new ones for a particular cross-transmission challenge.

Complex bioinformatic analysis and integration with epidemiology can be undertaken by a clinical microbiologist with domain specific experiences in outbreak investigation and clinical microbiology but lacking prior formal training in sequencing, molecular biology or bioinformatics techniques. This requires “upskilling” personnel who already work in clinical microbiology labs. The ability to bring analyses in-house without reliance on hiring highly trained and specialised personnel such as bioinformaticians is key to promoting the uptake of WGS in clinical laboratories. Once the technology and expertise are in place WGS is a supremely flexible tool that can be used to explore transmission of a range of respiratory and other pathogens. In future it is likely that additional clinical applications such as metagenomic detection of pathogens will mature to the point they are ready for patient use. Yet even for transmission analysis as the sole applications there are advantages to performing local analysis rather than reliance exclusively on academic partners or reference laboratories: avoidanceof sample transport can reduce turnaround times, high priority samples can be expedited and local expertise in sequencing in medical or scientific staff can help with presentation of results to non-specialist stakeholders such as infection control teams, public health teams and management. Even if some steps such as sequencing or culture enrichment must

be carried out off site, a microbiology scientist or clinician trained in bioinformatics could analyse reads of interest to the local service regardless of where they are generated, as demonstrated in the adenovirus article.

A flexible approach to training that is not restricted to any particular pathogen is required, which is why upskilling existing personnel in "wet lab" technique and bioinformatic techniques should emphasise transferable skills. The potential array of respiratory pathogens in healthcare or community is vast. In all three articles presented, no bespoke bioinformatic workflow or WGS criteria for strain relatedness existed and the wet laboratory methods usually had to be developed also. There is a need for personnel who can adapt to outbreaks of pathogens that have not been extensively studied by WGS and can generate new analysis pipelines as required.

Such analysis pipelines will by necessity focus on creating genome assemblies and then constructing SNP distance matrices and phylogenetic trees, and for bacterial pathogens the use of typing schemes such as MLST, cg-MLST and wg-MLST. All these approaches require to a great extent the use of open-source bioinformatics software tools, which themselves are primarily designed to be run on UNIX. However, the potential array of pathogens that can cause outbreaks at the level of individual institutions, healthcare delivery networks or the wider public context is vast. Where no bespoke pre-existing complete workflow exists it is necessary to rapidly generate reliable sequence data, interpret it with new workflows and define criteria for strain relatedness de-novo. This addresses the clinical question of interest, generates sequence data that others can use to add context to their own investigations and provides examples of workflows that others can emulate to establish an investigation in a shorter timeframe.

Sequencing analyses have to a large extent been normalised and even automated for key public health pathogens, through the use of commercially available software packages or open source web portals that allow strain typing and quantification of nucleotide difference by the non-specialist. In addition to standardised methodology, a near consensus has emerged in clinical academic literature for a narrow range of pathogens on results interpretation with respect to definitions of what counts as closely related for the purposes of indicating likely recent transmission from person to person as discussed in chapter 1 of this thesis for *S. aureus* and *C. difficile*.

As mentioned earlier, UNIX Operating System (OS) expertise is required to operate High Performance Computing (HPC) clusters, which are needed to provide the computing power to tackle very large scale sequencing analyses. The use of these UNIX tools and clusters was described in the methods sections of the articles.

Most UNIX bioinformatics tools are operated using the "command line" (CL), a purely text based interface for installing and running software and reviewing some analysis outputs. This will be unfamiliar to many users who are mainly familiar with graphical user interface employed in Windows or Mac OS. There is a steep initial learning curve to achieve sufficient competency to perform even basic computer tasks like retrieving and editing a text document using the CL. Once this hurdle is overcome the flexibility offered by CL approach to bioinformatics analysis is that it opens up access to a massive ecosystem of bioinformatics software tools that are nearly all offered free of charge and developed and maintained by researchers and bioinformaticians to aid the WGS research and clinical applications community. If a desired analysis is too complex to be performed with a single tool, multiple tools can be relatively easily integrated in a modular fashion into "workflows" where outputs of one tool provide inputs to the next until all desired analyses are complete. There is no cost barrier to testing these tools and nowadays installation on UNIX system using methods such as conda or nextflow takes only minutes, permitting rapid experimentation with many tools to find one that best suits the requirements of a particular analysis. For high profile pathogens like SARS-CoV-2, a CL tool such as nextflow artic pipeline will take either Illumina or Nanopore reads and run an entire analysis on 96 samples (QC, assembly, SNP identification, mutation and variant analysis and visualisation in html format) in under an hour on a suitable powerful workstation by running a single command.

Where the modular approach is required, the short text scripts required to run it can be easily communicated to other researchers for their own use. In the Pseudomonas articles the scripts were made available on the github website to aid study replication.

Today installing and using CL tools is relatively standardised, with most tools hosted on github and usually adhering to a common set of installation commands. This is a significant advance on the state of the field when this PhD thesis commenced, at that time there were no defacto standards for hosting and distributing CL tools. Illustrating the significant trend towards standardisation of CL hosting and distribution is the experience that most CL tools do not adhere to a code of best practice described by Torsten Seemans, a notable bio-informatician, termed the *"10 commandments for bioinformaticians".* A condensed version of this best practice guide is that the code for command line tools is made available on major websites such as gituhub.com, rather than offered exclusively through various academic laboratory websites that do not have robust version tracking and download support, that feedback and improvement suggestion from users are offered via long term support, that installation of the tool and any essential dependencies is simplified through making the tool available via "wrapper" programs such as Bioconda or Docker that permit installation with by typing a single command[3,4]. It is also recommended to avoid creating (usually accidentally)

“vapourware”, a software tool which users download and spend significant time and resources attempting to use before discovering that it has fundamental flaws that have not been fixed, or which is no longer being actively supported by the developer(s) - for instance a doctoral student who has moved to a new lab. Through experience it is usually possible to identify such CL tools, for instance by checking the number of open and closed issues on the tools github page for evidence of ongoing engagement by the developer.

Additionally, though not an absolute requirement, CL tools that launch complex workflows often include links in the description to manuscripts (preprints or peer-reviewed) detailing the principles underpinning the analysis the tool performs and often an objective assessment of the performance of a new tool relative to peer tools for model datasets for speed, accuracy and compute resource intensiveness. They can be supplemented with other windows OS or web based tools for effective visualisation of results to enable communication of findings to clinical and public health colleagues.

The disadvantages of UNIX are the need to rote memorise a specific directory and file structure as well as a large vocabulary and unique syntax for structuring the written commands. These commands can be used both interactively and though the use of submission of “shell” scripts written in BASH scripting language for job submission to HPCs. HPC’s may not offer an interactive computing environment, outside of limited instances for debug work. Any error in a submitted HPC script may cause a job to fail and queueing it the job up for repeat analysis may take hours or days. There is also an element of judgement required to pick CL tools, aided by keeping abreast of what peers are using via reading methods sections of current publications or attending conferences like ASM[5].

Even though learning to use UNIX and BASH is akin to learning a new language, it is a simplified, formalised and highly logical one. Importantly, it is not required of the new user to learn how to write basic computer code to use UNIX and CL programs. One interacts with prewritten programs, but using text interface rather than a mouse pointer on a screen. To write entirely new software programs from scratch to manipulate and analyse genomic sequence data may be considered firmly in the realm of specialised bioinformaticians and beyond the scope of what can be expected of clinical or scientific microbiology lab staff who are upskilling to take advantage of the unique flexibility offered by CL tools. Indeed, efforts by inexperienced non-bioinformaticians to develop new software, written in computing languages like Python or C++ (and later operated using UNIX CL) are unlikely to result in useful tools that adhere to best practice standards. In this work, the focus in the bioinformatics methods was firmly on use of existing CL tools rather than the generation of new ones. This is a valid approach to research, especially for clinical microbiologists who are familiar at sitting at the interface between medical science and laboratory science.

Combined with ever improving WGS benchtop technology, training hospital laboratory staff to use command line tools could user in an era of "democratised" access to sequencing where a wide range of non-reference clinical microbiology labs have the tools available to allow them to investigate pathogens as needed. This is likely to help with uncovering cryptic person to person transmission and devising effective prevention strategies in conjunction with local infection control team.

Applications of WGS based transmission analysis in Ireland during the Covid pandemic

SARS-CoV-2 sequencing was established in University Hospital Limerick (UHL) during the pandemic. The bioinformatic analysis is performed on a local UNIX workstation and access to the national HPC cluster running UNIX in Ireland called the Irish Centre for High End Computing (ICHEC)[6]. Access to the latter was provided not only to one clinical lab to but to the NVRL and other laboratories (clinical, academic and private industry) in the Irish Coronavirus Sequencing Consortium (ICSC)[7]. This project was approved by the National Research Ethics Council for Covid-19. The Health Research Consent Declaration Committee granted approval to link limited de-identified Covid metadata (age range, sex, location expressed as county and date of infection) with SARS-CoV-2 genome sequences without express individual patient consent[7,8]. The experience detailed in the thesis discussion on the challenges obtaining ethical approval for the Mumps paper article was salutary. There is a high bar to clear to obtain a derogation from obtaining individual patient consent for pathogen genome sequencing and data sharing but it is possible to demonstrate an over-rising public health interest.

With respect to the bioinformatics pipeline that enabled the function of the ICSC, access to one HPC, containing a single instance of a UNIX based artic analysis pipeline[9] (artic-ncov) allowed all participating laboratories that generated sequence data to have access to adequate compute resources and complete standardisation of post-sequencing analysis steps across the consortium. More importantly, when the ICSC was established, the primary way to generate consensus genomes from the available wet lab methods was using UNIX tools; commercial options utilising website uploads of sequencing files to identify variants e.g. EPISEQ became available from late 2021 onwards[10].

The knowledge gained undertaking research for the ICSC was used to advocate for and to guide the setup of a "Hub and Spoke" model for an Irish laboratory network to sequence SARS-CoV-2[15]. This model relies to a very great extent on hospital clinical microbiology laboratories. The majority of these laboratories had no WGS experience prior to 2020, including UHL clinical laboratory. Both Illumina and Oxford Nanopore instruments are used by the participating laboratories for generating

sequences. This reflects a conscious decision to ensure resilience in terms of access to reagents and consumables in an era when supply chains for a single company can be fragile in times of crises. The national model delivers near real-time SARS-CoV-2 sequencing results, a significant improvement on the older centralised model which depended primarily upon reference laboratory. Results from the network inform local and national infection control responses to the evolving pandemic[15]. It aims to develop a resilient, flexible resource to ensure adequate WGS capacity to respond to urgent public health threats and ensure expertise is available for sequencing other pathogens as circumstances dictate.

Variant identification is the key metric used in the program surveillance reports. SNP distance analysis is used in specific instances to assist local outbreak identification and control and to inform interventions in individual hospitals[11]. Bayesian analysis of the type described in the Mumps article does not form part of the standard bioinformatics analyses. There is a role for such complex analyses outside the program: when SARS-CoV-2 first emerged as a "novel coronavirus nCoV" re-print articles in January 2020 analysed coronavirus genomes using the same Bayesian approach described for Mumps (Birth Death Skyline Serial model). The authors used it to determine the tMRCA – the time to most recent common ancestor of SARS-CoV-2 genomes isolated from the initial Wuhan outbreak and therefore how long had the virus had been circulating in humans[12]. Conclusions from these analyses, borne out by subsequent analyses of much larger datasets, pointed unambiguously towards very recent origin of all genomes isolated from humans sometime around late 2019. This effectively ruled out one hypothesis at the time on its origin: that SARS-CoV-2 had been potentially circulating unnoticed in humans for long periods without causing significant outbreaks or mortality. Training individuals who work in clinical microbiology and public health in Ireland to perform these types of analyses is therefore valuable as a form of pandemic preparedness.

Some of funding available to the network is to be used to provide training in WGS data analysis to medical scientists and to trainees in clinical microbiology and public health. One approach is though funding development of new Irish courses, fellowships and academic qualifications including MSc and PhD, in addition to subsiding attendance at existing training courses in Ireland and abroad. The types of training activities supported are informed by the various bioinformatic and data visualisation challenges encountered in the three articles and discussed earlier. A model that could be used is an example at UHL where a medical scientist with no prior sequencing experience performed small scale WGS investigation of bacterial isolates as a MSc thesis using equipment initially procured for viral sequencing. The thesis focused on optimisation of wet lab sequencing protocols but also the gave the scientist the opportunity to learn the rudiments of bioinformatics analysis through mentoring from scientists and microbiologists, and ultimately via attendance at a

welcome trust training course in pathogen genomics at the Sanger Institute, UK. The expansion of this model of "in-service" training in practical genomics skills, including bioinformatics skills, to more individuals within the Irish health service is highly desirable. Repeatedly demonstrating the benefit of WGS based analysis by publication of new insights into transmission networks is one means to support and encourage this model.

In order to demonstrate such benefit, the following scenarios and enablers are proposed for use of WGS in routine hospital settings:

- Sequencing of bacterial isolates where cross-transmission is suspected and where there is no existing reference laboratory function for the pathogen in Ireland. In UHL this has been undertaken for an outbreak of *Burkholderia cepaciae complex* in intensive care and for suspected transmission of *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* in the neonatal intensive care unit. The ability to identify or rule out transmission on site allows for finite infection control resources to be more effectively allocated.
- Identification of unusual or rare bacteria. As an example, in UHL a case report is currently in preparation of the first described human infection with a rare non-fermenting gram negative bacillus, *Chryseobacterium shandogenes***.** The species identification was performed with routine laboratory methods (protein mass spectrometry) but on-site WGS allows for confirmation of the identity. No reference laboratory service is known in Europe for this species or genus of organism. Transfer out to services such as the rare and imported pathogens laboratory in UK would result in missed opportunity to expand local expertise in WGS and demonstrate its utility to local clinicians for patient management and to support research.
- Identification of bacterial virulence factors: several reference laboratories identify virulence factors routinely in Ireland using WGS such as PVL toxin in the *Staphylococcus aureus* reference lab. Where no reference lab for the pathogen exists, whole genome assemblies or the sequencing reads could be checked against the virulence factor database. It must be noted the clinical utility of this is unknown and such use would not be validated. Similarly, the clinical utility of WGS in hospital laboratories for antimicrobial resistance gene identification is not well established compared with the reproducibility, clinical utility and relatively low cost and fast turnaround time of existing phenotypic antimicrobial susceptibility testing methods.
- Target enriched metagenomic assay identification of respiratory pathogens (viruses and bacteria) from NP swabs as part of formal surveillance programs, rather than as

a diagnostic assay. UHL is a site in the expanding national program for Severe Acute Respiratory Illness surveillance. This requires identifying at minimum SARS-CoV-2, influenza and respiratory syncytial virus in all patients admitted with severe respiratory illness. A metagenomics assay approach would detect these pathogens and potentially dozens or hundreds of other pathogens that could result in similar clinical presentation. It would also allow for the assembly of SARS-CoV-2 whole genome and variant identification with the same assays, such as the "Illumina Respiratory Pathogen ID/AMR Panel". Once laboratory and medical staff became familiar with the operational aspects of such assays and the quality assurance system for the assay was robust, a business case could be developed to use the assay for diagnostic purposes including reporting of results to clinicians for action, following an extensive validation (required if the assay is marked as "research use only").

If the above applications are to be made a reality in a number of hospital labs, then funding, training and institutional support are required. The national SARS-CoV-2 hub and spoke sequencing program has secured approximately 2.6 million euro of funding from 2023 to 2027 from the European Union and the Irish health service to expand the sequencing applications of current sites and embed pathogens sequencing into routine lab workflows. Such funding can be used to support the projects outlined above whether by consumable purchase or training courses in Ireland or abroad for medical scientists. Senior medical scientist posts specifically for WGS are to be funded in each spoke lab and sequencing equipment and bioinformatics workstations have already been purchased for Covid sequencing applications. The equipment and scientist time it is hoped can be partially repurposed to the above applications, as the number of Covid patients requiring WGS falls with the end of the pandemic and the lack of emergence of dangerous new variants in 2022 and 2023 that would threaten the success of the vaccination program.

REFERENCES

1. Brennan C, DeLappe N, Cormican M, et al. A geographic cluster of healthcare-associated carbapenemase-producing hypervirulent Klebsiella pneumoniae sequence type 23. *Eur J Clin Microbiol Infect Dis*. Published online 2022. doi:10.1007/S10096-022-04535-Z
2. Chung The H, Boinett C, Pham Thanh D, et al. Dissecting the molecular evolution of fluoroquinolone-resistant Shigella sonnei. *Nat Commun*. 2019;10(1). doi:10.1038/S41467-019-12823-0
3. Dale R, Grüning B, Sjödin A, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*. 2018;15(7):475-476. doi:10.1038/S41592-018-0046-7
4. Merkel D. Docker. *Linux Journal*. 2014;2014(239). doi:10.5555/2600239.2600241

5. ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines | Overview | ASM.org. Accessed January 24, 2023. https://asm.org/Events/ASM-NGS/Home

6. Introduction to the Irish Centre for High-End Computing | ICHEC. Accessed January 23, 2023. https://www.ichec.ie/about

7. Irish Coronavirus Sequencing Consortium - Teagasc | Agriculture and Food Development Authority. Accessed January 23, 2023. https://www.teagasc.ie/food/research-and-innovation/research-areas/food-bioscience/irish-coronavirus-sequencing-consortium/

8. Funk T, Pharris A, Spiteri G, et al. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Euro Surveill*. 2021;26(16). doi:10.2807/1560-7917.ES.2021.26.16.2100348

9. Quick J, Grubaugh ND, Pullan ST, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc*. 2017;12(6):1261-1276. doi:10.1038/nprot.2017.066

10. Mugnier N, Griffon A, Simon B, et al. Evaluation of EPISEQ SARS-CoV-2 and a Fully Integrated Application to Identify SARS-CoV-2 Variants from Several Next-Generation Sequencing Approaches. *Viruses*. 2022;14(8). doi:10.3390/V14081674

11. Hare D, Meaney C, Powell J, et al. Repeated transmission of SARS-CoV-2 in an overcrowded Irish emergency department elucidated by whole-genome sequencing. *J Hosp Infect*. 2022;126:1-9. doi:10.1016/J.JHIN.2022.04.015

12. Zhang C, Wang M. MRCA time and epidemic dynamics of the 2019 novel coronavirus. *bioRxiv*. Published online January 31, 2020:2020.01.25.919688. doi:10.1101/2020.01.25.919688

List of Abbreviations

AET	Antibiotic Eradication Therapy
AMRIC	Antimicrobial Resistance and Infection Control
ANOVA	Analysis of Variance
ASM	American Society of Microbiology
BMI	Body Mass Index
BMT	Bone Marrow Transplant
CAGEF	Centre for the analysis of Genome Evolution and Function
CDC	Centres for Disease Control and Prevention
CF	Cystic Fibrosis
CFRD	Cystic Fibrosis Related Diabetes
cg-MLST	core genome Multi Locus Sequence Typing
CI	Confidence Interval
CID	Clinical Infectious Diseases
CL	Command Line
DNA	Deoxyribonucleic acid
EM	Electron Microscopy
ERB	Ethics Research Board
ESS	Effective Sample Size
GI	Gastrointestinal
GISAID	Global Initiative on Sharing All Influenza Data
GRC	Graduate Review Committee
HAV	Human Adenovirus
HAdV-A31	Human Adenovirus-A31

HPC	High Performance Computing
HPD	High Probability Distribution
HSCT	Hematopoietic Stem Cell Transplant
ICHEC	Irish Centre for High End Computing
ICSC	Irish Coronavirus Sequencing Consortium
ID	Infectious Diseases
IPAC	Infection Prevention And Control (Canada)
IPC	Infection Prevention and Control (Ireland)
LES	Liverpool Epidemic Strain
MCMC	Markov chain Monte Carlo
ML	Maximum Likelihood
MLST	Multi Locus Sequence Typing
MMR	Measles Mumps Rubella
NCBI	National Centre for Biotechnology Information
NP	Nasopharyngeal
OS	Operating System
Pa	*Pseudomonas aeruginosa*
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Electrophoresis
PFT	Pulmonary Function Test
PHO	Public Health Ontario
PHUR	Public Health Unit Regions
PSF	Penicillin-streptomycin-amphotericin B
QC	Quality Control

RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SH define in Intro	Small Hydrophobic
SK	SickKids
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SRA	Short Read Archive
ST	Sequence Type
tMRCA	time to Most Recent Common Ancestor
UHL	University Hospital Limerick
UK	United Kingdom
US	United States
USD	United States Dollar
wg-MLST	whole genome Multi Locus Sequence Typing
WGS	Whole Genome Sequencing
WHO	World Health Organisation