Whole genome sequencing reveals extensive community-level transmission of group A *Streptococcus* in remote communities

A. C. BOWEN¹,²,³, T. HARRIS¹, D. C. HOLT¹, P. M. GIFFARD¹, J. R. CARAPETIS²,³, P. T. CAMPBELL⁴, J. McVERNON⁴ AND S. Y. C. TONG¹,⁵*

¹Menzies School of Health Research, Charles Darwin University, Darwin, NT, Australia
²Princess Margaret Hospital for Children, Perth, WA, Australia
³Telethon Kids Institute, University of Western Australia, Perth, WA, Australia
⁴University of Melbourne, Melbourne, VIC, Australia
⁵Royal Darwin Hospital, Darwin, NT, Australia

Received 19 September 2015; Final revision 11 December 2015; Accepted 16 December 2015; first published online 1 February 2016

SUMMARY

Impetigo is common in remote Indigenous children of northern Australia, with the primary driver in this context being *Streptococcus pyogenes* [or group A *Streptococcus* (GAS)]. To reduce the high burden of impetigo, the transmission dynamics of GAS must be more clearly elucidated. We performed whole genome sequencing on 31 GAS isolates collected in a single community from children in 11 households with ≥2 GAS-infected children. We aimed to determine whether transmission was occurring principally within households or across the community. The 31 isolates were represented by nine multilocus sequence types and isolates within each sequence type differed from one another by only 0–3 single nucleotide polymorphisms. There was evidence of extensive transmission both within households and across the community. Our findings suggest that strategies to reduce the burden of impetigo in this setting will need to extend beyond individual households, and incorporate multi-faceted, community-wide approaches.

Key words: Impetigo, infectious disease epidemiology, *Streptococcus pyogenes*, transmission.

INTRODUCTION

Impetigo is a common childhood infection [1] with an estimated 160 million prevalent cases globally [2]. Children living in Oceania have the highest documented prevalence [2], with the most severely affected group being Indigenous Australian children living in remote communities [3, 4], where the median childhood prevalence is 43% (95% CI 40.2–45.7) [2]. Impetigo is a non-benign disease that drives outbreaks of acute post-streptococcal glomerulonephritis [5] with consequent chronic kidney disease [6, 7] and probably contributes to the highest reported rates of rheumatic heart disease in the world found in these communities [8, 9]. In a large randomized controlled trial, *Streptococcus pyogenes* [or group A *Streptococcus* (GAS)] was confirmed to be the primary driver of impetigo in an endemic context [10]. Therefore a clear understanding of the transmission dynamics of GAS is required to inform the design of interventions to reduce the prevalence of impetigo.

It has long been thought that housing conditions are a major contributor to the high prevalence of infectious and parasitic disease in children in remote Indigenous communities. For example, households are crowded in remote Indigenous communities with...
a median of 3–7 persons per bedroom [10, 11] and a correlation exists between the level of crowding and prevalence of impetigo [11]. Interventions that improve housing quality have had only limited success in reducing the burdens of infectious disease [12], although interventions to reduce household crowding have not been tested. Part of the reason for the lack of impact of improved housing quality may be extensive transmission of infectious agents outside the household. Community interactions within Indigenous communities are complex, with considerable mobility of children between households, and much unstructured mixing opportunities in school and other community settings that may be more or less influential to infection risk than the home environment [13].

To better understand the relative contributions of household-level and community-level transmission, we obtained whole genome sequences (WGS) of 31 GAS isolates from household clusters of impetigo in a single community over a 3-day period. By assessing the relatedness of GAS strains associated with skin infections in multiple members of individual households, we sought to assess whether the transmission from other household members, or acquisition outside the household, was the likely source of infection, and thus whether individual households would likely be the most useful target for interventions to reduce acquisition of GAS and the subsequent burden of impetigo.

**METHODS**

Isolates were collected from children with impetigo who were participants in a randomized controlled trial of oral trimethoprim-sulfamethoxazole vs. intramuscular benzathine benzylpenicillin G for the treatment of skin sores [10]. The trial recruited children aged 3 months to 13 years from seven remote communities in the Northern Territory (NT), Australia where the prevalence of impetigo is high year-round [14, 15]. Screening for eligibility in the trial occurred predominantly in schools. Following this, research nurses visited households to discuss the study with caregivers. Overall 65% of recruited children with impetigo were identified through school screening. During these household visits, siblings or relatives were also screened for participation (an additional 25% of recruitment). The remaining participants (10%) were referred directly from the clinic [10]. No attempt was made to screen children who did not attend school or were below school age, outside of the recruitment described above. Impetigo was graded as ‘mild’ if there was one purulent or crusted sore and <5 sores in total, or ‘severe’ if there were ≥2 purulent or crusted sores or ≥5 sores in total. All children with impetigo in the trial had at least one microbiological swab collected [16]. To ascertain household crowding and number of other children with impetigo in the household a survey of the primary caregiver was conducted. Apart from children with impetigo, no swabs or clinical assessments of other household members were made.

For this WGS substudy, we concentrated on a single community of ~1500 population, at a single recruitment visit conducted between 8 and 11 November 2011. This community was chosen because it had the highest number of GAS isolates available from different sized household clusters. We included for sequencing isolates recovered from children residing in households where ≥2 children from the same household had culture-confirmed GAS impetigo. This sampling frame was selected to determine whether there was evidence of children acquiring GAS within the household or elsewhere. All isolates were collected prior to the commencement of antibiotic therapy.

Cotton swabs (Copan, Italy) were used to collect pus from skin sores for microbiological culture according to standard methods [26]. Where GAS was recovered, a single isolate from the agar plate was stored at −80 °C. DNA was extracted from the stored isolate using a QIAamp DNA Mini kit (Qiagen, Germany) according to the manufacturer’s instructions.

We obtained genome sequences with paired end libraries on the Illumina HiSeq (Illumina, USA) platform through Macrogen Inc. (South Korea). The reads have been deposited in NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) with accession numbers SAMN03988118–SAMN03988148. Reads were mapped against the M1 GAS reference sequence of *S. pyogenes* isolates SF370 [28] (accession no. AE004092-2) using SPANDx [19] to define orthologous core genome single nucleotide polymorphisms (SNPs). To restrict the analysis to core genome SNPs, reads mapping to repetitive regions, regions with less than half the average depth, and regions with more than three times the average coverage of the entire genome were filtered out. We built a maximum likelihood tree using RaXML [20] with default settings and visualized with the Interactive Tree of Life [21]. The multilocus sequence type (MLST) for each isolate was determined using SRST2 [22]. Short read data for each isolate was assembled using Velvet [23] with reference to SF370 (accession no. AE004092-2). Draft assemblies have been deposited at DDBJ/EMBL/GenBank.
under the accessions LRGF00000000–LRHJ00000000. The versions described in this paper are versions LRGF01000000–LRHJ01000000, and assembly metrics are provided in Supplementary Table S1. The best assembly for each sequence type (ST) (based on the smallest number of contigs and largest N50) was then used as a reference to map reads from the other strains of that ST to identify orthologous SNPs using SPANDx [19]. Reads from the assembled ’reference’ strain were also mapped back to itself as a quality control procedure. SNPs were manually visualized in Artemis [24]. Two sets of short reads (SST2097-1, SST2090-1) did not assemble well (>300 contigs rather than 10–32 for other assemblies). However, the short reads from SST2097-1 and SST2090-1 mapped well against the reference assembly and alignments were manually checked. Epidemiological and genome sequence data were visualized using Circos [25].

Ethical standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

RESULTS

Of 69 children who were screened at the school for participation in the trial during the November 2011 recruitment, 45 (65%) with crusted or purulent impetigo were enrolled in the study. These 45 children had a median age of 7.4 years [interquartile range (IQR) 4.4–9.7 years] and 27 (60%) were female. Fourteen (31%) of these 45 children were the only member of their household with impetigo recruited in the trial and isolates recovered from these participants were not included for WGS. The remaining 31 children (69%) resided in 11 households, with a median of 3 (IQR 2–3) infected individuals in each household. Twenty-five (81%) of the 31 children in households where multiple infected children were identified had severe impetigo, compared to nine (64%) of the 14 children with impetigo from households without other cases (P = 0.4). (See Fig. 1 for the study profile.) Of the 31 isolates, there were nine STs that were clearly delineated following alignment against the reference GAS strain SF370 (Fig. 2). In order to provide improved resolution and finer mapping within each ST, the best de novo assembly within each ST was chosen to be the reference for that ST (median size of reference assemblies 1 801 299 bp, range 1 745 425–1 878 702 bp). Mapping of short reads from the reference isolate back to the reference assembly (i.e. itself) demonstrated no SNPs. Overall, the median depth of read coverage for all 31 sequences aligning against the reference assemblies was 133×. Within each ST, the number of SNPs for any one isolate compared to the reference assembly ranged from 0 to 3 (Table 1). Thus, isolates within each ST were essentially identical when considering orthologous SNPs at a
whole genome level. Functional details of these SNPs are provided in Supplementary Table S2.

There was evidence of acquisition of GAS both within and outside of the household unit (Fig. 3). Five of the 11 households had a single circulating ST, consistent with transmission within the household. There were only two instances where a household was found to have an ST that was not found in any other household (i.e. the household was uniquely identified with an ST and vice versa – household I with ST182, and household K with ST641). There was also abundant evidence of transmission of STs outside of the household unit. The six ST10 isolates were identical (i.e. 0 SNPs were identified) and spread across three different households. The two ST176 isolates were identical and from two different households. Of the four ST304 isolates, two variants (differing by 3 SNPs from each other) were identified. Three of the ST304 isolates were from a single household, where both variants were present. Within the eight ST332 isolates, four SNP positions resulted in five variants, with a maximum of two SNPs between any two variants. No household had more than one isolate of any ST332 variant.

**DISCUSSION**

Our study has revealed evidence of extensive within-household and community-level transmission of GAS in an endemic setting. There was considerable diversity of GAS clones as represented by nine STs in 11 households, and also a lack of variation within each ST, indicating very recent transmission of each of the STs. This lack of variation, together with finding the same ST shared across households, infers the transmission of multiple strains outside of the household. It is of note that the majority of index study participants were recruited from the community school, a setting likely to be important for cross-household transmission.
This study was conducted in households with a heavy burden of disease. The households had a median of three children with impetigo at the time of screening and the impetigo was typically graded as severe. Previous studies in this regional endemic setting have demonstrated a considerable diversity of GAS when longitudinal sampling was performed in a small number of communities [26, 27], or with broader sampling for diversity across the entire region [28]. We show here that extensive diversity of STs exists even at a single time point (sampling over a 4-day period) and within a single community.

The previous studies in our region [26–28] were able to determine diversity at a MLST or *emm*-type level, but lacked the resolution now available with WGS to understand the relationship of strains within these STs or *emm* types. We found that strains within each ST differed by at most three SNPs despite being in multiple households in some cases, representing clear evidence of recent transmission between several households (Fig. 3).

Most of the clones presented in this study are uncommon in the MLST database (http://pubmlst.org/spyogenes/). Specifically, *emm*166·1/ST332, *emm*219/ST641, *emm*164·3/ST330 (SLV) are all novel combinations not present in the MLST database; there are four *emm*70/ST10 strains – all from the NT; five *emm*53/ST11 strains – four from the NT, one from Brazil; two *emm*108/ST304 strains – both from the NT; one *emm*44/ST641 strain from Guyana; one *emm*205/ST641 strain from Brazil; and one *emm*44/ST641 strain from South Africa.

Table 1. Details of the 31 group A Streptococcus (GAS) strains

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Household</th>
<th>ST</th>
<th><em>emm</em> type</th>
<th>SNP variant</th>
<th>SRA number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2090-1</td>
<td>A</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988125</td>
</tr>
<tr>
<td>2097-1</td>
<td>A</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988127</td>
</tr>
<tr>
<td>2096-1</td>
<td>A</td>
<td>332</td>
<td>166·1</td>
<td>CCGA</td>
<td>SAMN03988126</td>
</tr>
<tr>
<td>1023-1</td>
<td>B</td>
<td>332</td>
<td>166·1</td>
<td>CAGG</td>
<td>SAMN03988121</td>
</tr>
<tr>
<td>2245-1</td>
<td>B</td>
<td>332</td>
<td>166·1</td>
<td>TAGG‡</td>
<td>SAMN03988122</td>
</tr>
<tr>
<td>1024-1*</td>
<td>C</td>
<td>11</td>
<td>53</td>
<td></td>
<td>SAMN03988123</td>
</tr>
<tr>
<td>2232-1</td>
<td>C</td>
<td>176</td>
<td>58</td>
<td>No SNPs</td>
<td>SAMN03988124</td>
</tr>
<tr>
<td>1028-1</td>
<td>D</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988118</td>
</tr>
<tr>
<td>2237-1</td>
<td>D</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988119</td>
</tr>
<tr>
<td>2243-1</td>
<td>D</td>
<td>332</td>
<td>166·1</td>
<td>CCTG</td>
<td>SAMN03988120</td>
</tr>
<tr>
<td>2241-1*</td>
<td>E</td>
<td>205</td>
<td>230</td>
<td></td>
<td>SAMN03988130</td>
</tr>
<tr>
<td>2240-1</td>
<td>E</td>
<td>332</td>
<td>166·1</td>
<td>CCGA</td>
<td>SAMN03988129</td>
</tr>
<tr>
<td>2229-1</td>
<td>E</td>
<td>330</td>
<td>164·3</td>
<td>No SNPs</td>
<td>SAMN03988128</td>
</tr>
<tr>
<td>2102-1*</td>
<td>F</td>
<td>304</td>
<td>108</td>
<td>CTA</td>
<td>SAMN03988132</td>
</tr>
<tr>
<td>2099-1</td>
<td>F</td>
<td>332</td>
<td>166·1</td>
<td>CCTG</td>
<td>SAMN03988131</td>
</tr>
<tr>
<td>2103-1*</td>
<td>F</td>
<td>332</td>
<td>166·1</td>
<td>CCGG</td>
<td>SAMN03988133</td>
</tr>
<tr>
<td>2095-1*</td>
<td>G</td>
<td>176</td>
<td>58</td>
<td>No SNPs</td>
<td>SAMN03988138</td>
</tr>
<tr>
<td>2091-1</td>
<td>G</td>
<td>304</td>
<td>108</td>
<td>TGG</td>
<td>SAMN03988134</td>
</tr>
<tr>
<td>2093-1</td>
<td>G</td>
<td>304</td>
<td>108</td>
<td>TGG</td>
<td>SAMN03988136</td>
</tr>
<tr>
<td>2094-1</td>
<td>G</td>
<td>304</td>
<td>108</td>
<td>CTA</td>
<td>SAMN03988137</td>
</tr>
<tr>
<td>2092-1</td>
<td>G</td>
<td>332</td>
<td>166·1</td>
<td>CCGG</td>
<td>SAMN03988135</td>
</tr>
<tr>
<td>1020-1</td>
<td>H</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988139</td>
</tr>
<tr>
<td>1021-1*</td>
<td>H</td>
<td>10</td>
<td>70</td>
<td>No SNPs</td>
<td>SAMN03988140</td>
</tr>
<tr>
<td>2098-1*</td>
<td>I</td>
<td>182</td>
<td>205</td>
<td>G</td>
<td>SAMN03988141</td>
</tr>
<tr>
<td>2242-1</td>
<td>I</td>
<td>182</td>
<td>205</td>
<td>A</td>
<td>SAMN03988142</td>
</tr>
<tr>
<td>1019-1</td>
<td>J</td>
<td>330</td>
<td>164·3</td>
<td>No SNPs</td>
<td>SAMN03988143</td>
</tr>
<tr>
<td>2088-1*</td>
<td>J</td>
<td>330</td>
<td>164·3</td>
<td>No SNPs</td>
<td>SAMN03988144</td>
</tr>
<tr>
<td>2233-1*</td>
<td>K</td>
<td>641</td>
<td>219</td>
<td>No SNPs</td>
<td>SAMN03988145</td>
</tr>
<tr>
<td>2235-1</td>
<td>K</td>
<td>641</td>
<td>44</td>
<td>No SNPs</td>
<td>SAMN03988146</td>
</tr>
<tr>
<td>2236-1</td>
<td>K</td>
<td>641</td>
<td>219</td>
<td>No SNPs</td>
<td>SAMN03988147</td>
</tr>
<tr>
<td>2238-1</td>
<td>K</td>
<td>641</td>
<td>219</td>
<td>No SNPs</td>
<td>SAMN03988148</td>
</tr>
</tbody>
</table>

SNP, Single nucleotide polymorphism; SRA, short read archive.
* Isolate used as the ‘reference’ assembly for that sequence type (ST).
† Single locus variant of ST330.

The SNP (single nucleotide variant) column refers to the SNP variants within each of the STs. (See Supplementary Table S1 for details of these SNPs.) Where there is only one representative of a ST, the cell is left blank.
ST182 strain from the United States; and one emm230/ST205 strain from Australia. emm58/ST176 appears to be a globally distributed clone with 36 strains in the MLST database from Australia (15), Europe (7), UK (3), the Americas (8) and Asia (3).

Although this report is the first to utilize WGS to examine transmission of GAS in impetigo lesions in an endemic region, the similarity of the clinical epidemiology and diversity of emm types from our region with other countries such as India [29, 30], Nepal [31], the Pacific Islands [32], Brazil [33] and Africa [34], suggests that our findings are likely to be applicable to other areas with a high burden of impetigo [2, 35, 36]. With regards to GAS population biology, despite the small size of this study and previous sampling in our region, we found novel combinations of emm types and STs. For example, emm166·1/ST332, emm219/ST641, and emm164·3/ST330 (SLV) are all novel combinations not present in the MLST database. In addition, we found direct evidence of recombination at the emm locus in the four ST641 strains, with three being emm219 and one emm44. Notably, ST182 was associated with emm205 in this study, but ST182 has also been associated with emm98 and emm101 in strains from Australia. Taken together with the findings of multiple STs, each with little intra-lineage variation, it appears that in endemic, high-transmission settings such as in tropical northern Australia, not only is there frequent strain replacement but also present are the conditions that may facilitate recombination and horizontal gene transfer.

Limitations of this study include the use of a single time-point and a single community to assess for household clustering. However, even with this small sample size, it is clear that community-level
transmission of GAS is a key component of the dynamic epidemiology of GAS in an endemic setting. In addition, an inference of multiple importations of GAS into a household may be incorrect if individuals harboured multiple co-infecting strains, making attribution of a source other than the household spurious on the basis of a single isolate. Studies involving WGS of multiple *S. aureus* colonies from a single swab have demonstrated carriage of a cloud of variation [37, 38]; however, all colonies were of the same ST. Such studies have not yet been reported for GAS. Finally, our sampling in households was not complete and included only symptomatic children recruited into the study. Swabbing of all household members for either active infection or asymptomatic colonization over time would be required to understand underlying patterns of transmission in the household that may be influential, other than just observed disease.

Nonetheless, our findings suggest that interventions purely targeted at individual households (e.g. improved living conditions, reduced household crowding) may not be effective at reducing the prevalence of impetigo unless the majority of residents in a community benefit from the intervention and/or there are other community-wide intervention strategies. Therefore, it is not surprising that small-scale interventions targeting housing quality alone did not result in an appreciable reduction in infectious diseases [12].

**CONCLUSIONS**

We report the first use of WGS to define the household transmission of GAS in a remote Indigenous community. It is likely that strategies aimed at reducing the burden of impetigo in remote Indigenous Australian children, and possibly in other high-burden settings, will need to include community-wide interventions.

**SUPPLEMENTARY MATERIAL**

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S095026881500326X.

**ACKNOWLEDGEMENTS**

We thank the participants, their families and study staff from Menzies School of Health Research for their contributions to this study.

An Australian National Health and Medical Research Council (NHMRC) project grant (545 346) and a Menzies School of Health Research Small Grant supported this work. A.B. is an NHMRC Early Career Fellow (1 088 735) and S.T. (1 065 736) and J.McV. (1 061 321) are NHMRC Career Development Fellows.

**DECLARATION OF INTEREST**

None.

**REFERENCES**


