Genome sequence analysis of emm89 Streptococcus pyogenes strains causing infections in Scotland, 2010–2016

Stephen B. Beres,1,4 Randall J. Olsen,1,2 Matthew Ojeda Saavedra,1 Roisin Ure,3 Arlene Reynolds,3 Diane S. J. Lindsay,3 Andrew J. Smith3,4 and James M. Musser1,2

Abstract

Purpose. Strains of type emm89 Streptococcus pyogenes have recently increased in frequency as a cause of human infections in several countries in Europe and North America. This increase has been molecular epidemiologically linked with the emergence of a new genetically distinct clone, designated clade 3. We sought to extend our understanding of this epidemic behavior by the genetic characterization of type emm89 strains responsible in recent years for an increased frequency of infections in Scotland.

Methodology. We sequenced the genomes of a retrospective cohort of 122 emm89 strains recovered from patients with invasive and noninvasive infections throughout Scotland during 2010 to 2016.

Results. All but one of the 122 emm89 infection isolates are of the recently emerged epidemic clade 3 clonal lineage. The Scotland isolates are closely related to and not genetically distinct from recent emm89 strains from England, they constitute a single genetic population.

Conclusions. The clade 3 clone causes virtually all-contemporary emm89 infections in Scotland. These findings add Scotland to a growing list of countries of Europe and North America where, by whole genome sequencing, emm89 clade 3 strains have been demonstrated to be the cause of an ongoing epidemic of invasive infections and to be genetically related due to descent from a recent common progenitor.

INTRODUCTION

Streptococcus pyogenes, a strict human pathogen, causes highly diverse infections that range in severity from common self-limiting skin and throat infections (impetigo and pharyngitis), to less common invasive infections that have high morbidity and mortality such as scarlet fever, sepsis and necrotizing fasciitis [1]. For decades, S. pyogenes strains were subtyped based on serologic diversity in the M protein, a highly polymorphic surface-displayed molecule that is an important virulence factor involved in multiple host-pathogen interactions, including adherence and resistance to phagocytosis [2]. Over the last 20 years, a shift away from M protein serotyping to emm gene sequence typing has occurred. Greater than 220 distinct emm types have been described based on allelic variation in the region of the emm gene that encodes the highly polymorphic amino terminal 50 amino acids of the mature M protein [3].

Considerable new information has been obtained by emm gene sequence typing. However, whole genome sequence analysis of approximately 10 000 S. pyogenes strains [4–15] has revealed that polymorphisms exist among strains classified by emm type and found that the genome of nearly every strain is genetically distinct. The largest contributors to this genome-to-genome variation are mobile genetic elements (MGE), such as prophages and integrative conjugative elements (ICE), and small sequence differences such as single
nucleotide polymorphisms (SNPs) and short insertions and deletions (indels). The genome sequence data provide a rich source of information for epidemiologic and pathogenesis studies framed within an evolutionary population genomic context.

In recent years, strains of emm89 S. pyogenes have been reported as a cause of increasing disease frequency in many countries worldwide, and have been documented to be a cause of local disease outbreaks [6, 15-43]. Genome sequence analysis of almost 1500 emm89 strains from the USA, Finland, Iceland, Canada (Ontario), the UK (the great majority from England) and Portugal has identified three major genetically distinct strain clusters, designated clades 1, 2 and 3 [6, 14, 15, 19, 24, 41]. Clade 1 strains have been identified only in the USA and Canada. Importantly, concurrent with a significant increase in the number and frequency of invasive emm89 infections, clade 3 strains emerged in the early 2000s and around 2007–2009 expanded rapidly in the USA, Finland and Iceland, displacing their predecessor clade 1 and 2 strains in these countries [6, 41]. Similarly, clade 3 organisms have become abundant in England, where they also displaced clade 2 organisms [15]. The emm89 clades differ predominantly due to multiple horizontal genetic transfer (HGT) events, several of which involve known virulence factors [6, 15]. Moreover, whole genome RNAseq expression analysis found that these regions of recombination were significantly non-randomly associated with altered gene expression [6].

The goal of the present study was to analyse the genome sequences of all available emm89 strains causing infections in Scotland between 2010 and 2016. We also sought to determine if geographic and temporal variation occurred in the genomic types of organisms causing disease in the country. Here we report that with a single exception, recently emergent clade 3 strains caused all invasive emm89 infections in Scotland during the study period. There is restricted evidence of genetic substructure of emm89 clade 3 subclones within Scotland and other regions of the UK.

METHODS

Bacterial strains

The 122 emm89 strains newly sequenced and studied herein were cultured from individuals throughout Scotland between 2010 and 2016, including 98 from patients with symptomatic infections at normally sterile sites (designated invasive infection isolates). Strains were identified as emm89 by emm gene sequencing as described [44], and confirmed to be emm89 based on whole genome sequence data (see below). The strain sample includes virtually all emm89 organisms causing invasive infections in Scotland from 2013 onward. The strain samples from 2010 through 2012 were largely from Glasgow and Lanarkshire, that is, the west central belt of Scotland (Fig. S1, available in the online version of this article). The sample also included 24 strains cultured from throats of individuals with pharyngitis or asymptomatic colonization (designated non-invasive infection isolates). The strain, year, infection type, geographic location and phage-profile are presented in Table S1.

For the purpose of epidemiologic and genetic comparisons, genome data from 133 previously sequenced emm89 S. pyogenes strains [6, 15, 41] were also used in this investigation, and are referenced at the appropriate places below.

Multiplexed genome sequencing

Bacterial cells were lysed with mutanolysin and lysozyme (Sigma, St. Louis, MO) and chromosomal DNA was extracted with the DNeasy 96 Blood and Tissue Kit (Qiagen, Germantown, MD). The extracted DNA was quantified with a Qubit Fluorometer using a Qubit dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA). Sequencing libraries were generated by tagmentation from the extracted DNA with the Nextera XT DNA Sample Preparation Kit and Index Kit V2 (Illumina, San Diego, CA). The DNA concentration of the resultant individual index tagged libraries was measured with a Qubit HS DNA Kit (Agilent, Santa Clara, CA). Libraries in sets of 96 samples were pooled at equivalent concentrations, and paired-end 150 nt sequence reads were obtained from the pooled libraries using a NextSeq 500/550 High Output Reagent v2 300 Cycle Cartridge and an Illumina NextSeq instrument. Data for the 122 newly sequenced emm89 genomes were deposited in the NCBI Sequence Read Archive under accession number PRNJNA387243.

Data preprocessing and initial genetic characterization

Before the sequence data were used in subsequent genetic analyses, reads were preprocessed to remove low quality sequences and library generation artifacts (e.g. adaptor contamination) with Trimmomatic [45], and base call errors were corrected with Musket [46]. Reads were then assembled de novo with SPAdes [47]. The S. pyogenes genome is approximately 2 Mbp in length. Strains for which the de novo assembly was larger than 2.2 Mbp were considered potentially contaminated and were reisolated and resequenced (n=9, 5 from the throat a non-sterile site of isolation) or excluded (n=2, viable pure cultures were not obtained) from further analysis. The resultant high-quality and contaminant-free reads were used in subsequent genetic analyses. The multi-locus sequence type (https://pubmlst.org/spyogenes/), emm type (ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/), presence of hyaluronate capsule synthesis genes (hasABC) and presence of potential antibiotic resistance genes (http://en.mediterranea-infection.com/article.php?laref=283%26titre=arg-annot) were assessed using SRST2 [48] with the indicated databases. This initial genetic typing confirmed the species and emm type of the strains studied.
Polymorphism discovery
Preprocessed reads were aligned to the genome of \textit{emm}89 clade 2 reference strain MGAS23530 (GenBank accession number CP013839) using SMALT (http://sourceforge.net/projects/smalt/). Sequence differences between the aligned reads and the reference genome were identified with Free-Bayes [49]. These polymorphism data were filtered on the basis of alternate base call consensus (\(\geq 70\%\)), mapped quality (\(\geq Q30\)) and coverage depth (\(\geq 10\)-fold) using VCFfilter (https://github.com/ekg/vcflib#vcflib). Since the genome of strain MGAS23530 lacks prophages and integrative and conjugative elements, the resultant polymorphism calls were made relative to only the core chromosome, thereby obviating the need to exclude phylogenetic inference potentially distorting polymorphisms within MGEs.

Phylogenetic inference and population structure
SNPs (i.e. variant site nucleotides) identified relative to the 1709394 genome of \textit{emm}89 clade 2 reference strain MGAS23530 along with all invariant site nucleotides were concatenated in sequential order to generate entire core genome aligned sequences for each strain studied using Prephix, Phrecon and SNPswapper (https://github.com/coding-hedgehog). Based on the 1709394 nucleotide full-length core genome aligned sequences, regions of putative horizontal transfer were predicted and polymorphisms within these regions were removed using Gubbins [50]. To constrain the inferences to primarily vertically inherited nucleotide differences, all phylogenetic reconstructions presented are based on Gubbins-filtered core genome SNPs. Phylogeny among strains was inferred by the neighbour-joining method with SplitsTree [51], and tree figures were generated with Dendroscope [52]. Genetic distances between strains (pairwise and overall mean SNP differences) were determined with MEGA [53]. Root-to-tip genetic distances were determined using TempEst [54].

Gene content and MGE analysis
Total gene and MGE content were assessed for the strains studied relative to the non-redundant pangenome gene content defined for 30 complete \textit{S. pyogenes} genomes representing 18 \textit{emm} types, as previously described, with minor modifications [6]. Briefly, reads were mapped to the pangenome (denoted GAS-30) with SMALT and the mean depth of coverage was determined with BEDTools2 [55] for all 2835 gene clusters represented. To compensate for differences in the depth of sequencing, depth-of-coverage for reads mapped relative to genome of reference strain MGAS23530 was normalized to 100-fold. Based on mapping of the sequencing reads from the clade 3 reference genome MGAS27061 to the GAS-30 pangenome, a normalized mean DOC value of >10 corresponded with gene presence and <10 with gene absence. The resultant gene content data were used to infer phage presence or absence, which was expressed as prophage content profiles or prophage genotypes as previously defined for 1193 \textit{emm}89 strains [6]. A prophage was called present if a minimum of 75 percent of its gene content represented in the GAS-30 pangenome was determined to be present. Reads not mapping to the GAS-30 pangenome were assembled \textit{de novo} with SPAdes. Resultant contigs greater than 100 nucleotides were queried against the NCBI non-redundant database using BLAST to determine their nature.

RESULTS

Genome sequencing and initial analysis
The genomes of all 122 strains from individuals from Scotland (\(n=98\) invasive disease isolates and \(n=24\) non-invasive isolates) were sequenced to a mean 231-fold depth-of-coverage (range, 108-fold–519-fold) using a paired-end 150 nt read strategy, and polymorphisms were identified. Inference of genetic relationships using core chromosomal SNPs found that these \textit{emm}89 strains from Scotland have a major population of 121 strains, and a single genetically divergent outlier strain (Fig. 1). Phylogenetic analysis showed the major \textit{emm}89 population of 121 strains corresponded to previously identified epidemic clade 3, whereas the single outlier strain (MGAS31744) was a prepandemic clade 2 organism (Fig. 1).

Variation in gene content and prophage genotypes
We, and others, have shown that \textit{emm}89 \textit{S. pyogenes} strains can vary in gene content as a consequence of HGT and differences in MGE content [6, 15, 41]. The gene content (i.e. gene presence or absence) of the 122 Scotland \textit{emm}89 strains was determined relative to the pangenome constructed from the combined gene content (>53 000 genes) of 30 \textit{S. pyogenes} genomes of 18 \textit{emm} types [6]. No ICEs were detected in the 122 strains. Virtually all differences in gene content are due to strain-to-strain differences in the five prophages detected among the 122 strains (Fig. 2a). Our analysis identified nine different prophage content profiles or prophage genotypes (Fig. 2a). The three more prevalent prophage genotypes are each present in 13 or more strains and account for 109 (89.3 %) of all 122 isolates (Fig. 2b). The remaining six prophage genotypes are each present in four or fewer strains and account for only 13 (10.7 %) of the isolates. The three most prevalent prophage genotypes (PG01, PG02 and PG04) among these Scotland isolates are those previously identified to be prevalent among 756 \textit{emm}89 clade 3 strains in the USA and Finland [6]. The most prevalent prophage genotype (PG01) was found in 60 (49.1 %) of the 122 isolates. Clade 3 reference strain MGAS27061 has prophage genotype PG01. PG01 is characterized by the presence of one prophage (\textit{φ}27061.1) encoding two virulence factors, superantigen pyrogenic exotoxin Spec and secreted DNase Spd1. The second most prevalent prophage genotype (PG02), was found in 36 (29.5 %) isolates, and is characterized by a lack of prophages in the genome. Clade 2 reference strain MGAS23530 (GenBank accession number CP013839) has prophage genotype PG02. The third most prevalent type (PG04), was found in 13 (10.7 %) isolates, and consists of prophage \textit{φ}27061.1 and a second prophage integrated into the genome at the two-component system regulator \textit{yesMN}.
This prophage shares substantial gene content with prophage 10270.3, and both encode the virulence factors' superantigen pyrogenic exotoxin SpeK and secreted phospholipase Sla (Fig. 3). The intermingling of these three prophage genotypes on multiple branches of the tree is consistent with these prophages being dynamically gained and lost in the population (that is, the prophage genotypes are not uniquely associated with a given genetic background). Additionally, as expected, all 121 of the Scotland clade 3 strains lacked the hyaluronic acid capsule synthesis genes (hasABC) consistent with previous findings [6, 15, 41, 42].
Genome analysis of strains cultured from throats

Twenty-four organisms cultured from the throat were studied. We found that the throat isolates were genetically intermixed with the invasive isolates. That is, the throat and invasive strains were not separate genetically distinct groups (Fig. 4).

Genetic structure with respect to year of isolation

We examined if a distinct temporal relationship existed between the year of strain isolation and the topology of the phylogenetic tree. In general, isolates from the early years of the Scotland sample tend to be located closer to the root of the tree, whereas those recovered in the later years (2015 and 2016) were located farther toward the tips of the tree branches (Fig. 5a). This impression is supported by analysis of root-to-tip genetic distances for the best-fit rooted tree, which found an overall trend for increasing genetic distance with increasing time since a point in the past (Fig. 5b).

Comparison of the genetic structure of strains from Scotland and other regions of the UK

We next tested the hypothesis that emm89 strains recovered in Scotland were a distinct genetic subpopulation compared to all other emm89 strains from elsewhere in the UK. For this analysis, we included all emm89 genome data reported in the Turner et al. manuscript [15] and deposited in publically available databases. Four strains from Scotland were analysed in the Turner et al. study, including one each from Dundee, Glasgow, Edinburgh and Kilmarnock (described in Table S1...
of the Turner et al. manuscript). The great majority of the strains \((n=122, 93.1\%\) were from England localities. Among the 121 clade 3 organisms from our Scotland sources studied herein, several genetically distinct phylogenetic branches were identified (Fig. 1, right). That is, the strains from Scotland sources do not form a single phylogenetic branch sharing a large number of identical polymorphisms relative to a most recent common ancestor. Rather, the clade 3 portion of the rooted-tree has multiple independent branches of short length radiating away from a central hub, suggesting that these strains have evolved fairly randomly away from a relatively recent common ancestor (Fig. 1, right). In combining all 122 strains from Scotland with all 122 strains from England in phylogenetic reconstruction (Fig. 6a), strains were intermixed

Fig. 5. Temporal distribution of the Scotland emm89 strains. (a) Reconstructed phylogeny of the 122 Scotland strains inferred by neighbour-joining based on 1410 core SNPs is shown. Strains are coloured by the year of isolation as indicated in the index. In general, strains isolated in earlier years are more proximal and in later years are more distal on the tree branches. (b) A scatter plot of root-to-tip genetic distances for the adjacent 122 Scotland strain tree is shown. The weighted fit shows an overall upward trend and a consistent yearly increase in the root-to-tip distances over 2010–2015.

Fig. 6. Genetic relationships among Scotland and England emm89 strains. (a) Reconstructed phylogeny of 255 emm89 strains inferred by neighbour-joining based on 2766 Gubbins-filtered core SNPs is shown. Strains are coloured by country, Scotland in blue and England in red. Consistent with the earlier period of collection of the England strain cohort (ca. 2003–2013) relative to the Scotland strain cohort (ca. 2010–2016), one-third (41/122) of the England strains are of preepidemic clade 2, whereas only 3 of the 126 Scotland strains are clade 2. Moreover, in the epidemic clade 3 portion of the tree (i.e. the right side), by in large the England strains are more proximal on the branches than the Scotland strains. (b) Reconstructed phylogeny of 89 emm89 strains inferred by neighbour-joining based on 1087 Gubbins-filtered core SNPs is shown. The 44 Scotland and 43 England strains are temporally matched sets of all strains collected between 2010–2013.
independent of country on virtually all branches of the clade 3 side of the tree. This is consistent with the Scotland and England strains being a single genetic population. Constraining the reconstruction to only those isolates collected in Scotland (n=44) and England (n=43) over the same time period 2010–2013 (Fig. 6b), similarly results in the strains being intermixed independent of country of origin.

DISCUSSION
Our genome sequence data show that with only one exception, all emm89 isolates recovered in Scotland since 2010 are members of the recently emerged genetic clade 3. These findings add Scotland to the list of countries in Europe (Finland, Iceland and England) and North America (USA and Canada) where clade 3 emm89 strains have been documented by full genome sequencing to be important causes of recent invasive infections and to be genetically related due to descent from a common progenitor.

Clade 3 organisms are noteworthy because they have an evolutionary history involving multiple recent HGT events, which have contributed to their epidemic emergence [6, 15]. Due to one HGT, emm89 clade 3 strains lack the hasABC gene region required for synthesis of hyaluronic acid capsule, as previously reported for GAS emm4 and emm22 strains [56]. Capsule was long thought to be a critical S. pyogenes virulence factor by virtue of its antiphagocytic properties [57, 58]. Lack of capsule may also affect virulence by altering GAS adherence interactions [59, 60]. Additionally, emm89 strains have undergone at least two recent HGT events involving the nga-ifs-slo gene region, encoding the secreted toxins NADase and streptolysin O that are documented virulence factors [6, 15]. As a consequence, relative to predecessor clade 1 and clade 2 organisms, clade 3 strains produce an increased amount of these two important toxins [6, 15, 41]. Upregulation of the nga-ifs-slo operon and increased production of these two toxins has been shown to enhance virulence of S. pyogenes in mouse and non-human primate models of upper respiratory tract infection and severe invasive infection [41, 42]. We recently reported that production of both NADase and SLO are required for full virulence in a mouse infection model [61]. The loss of capsule production and the increased production of NADase and SLO may be closely linked in the emergence of the epidemic clade 3 strains, as it has been shown that when NGA and SLO are highly produced, capsule is no longer required for virulence in a mouse invasive infection model [42].

A further evolution of the emm89 clade 3 clone, via an additional HGT event involving the spyBA virulence factors [62, 63] characterized an emergent variant, designated subclade 3D, in bacteremic case isolates from Finland [6, 24]. The subclade 3D strains strongly increased in 2014 as a cause of bacteremic infections, and have persisted in Finland. Relative to the five other subclades identified among the bacteremic cases, subclade 3D strains had a significantly higher 30 day case fatality rate. In assessment of phylogeny and recombination among the Scotland emm89 clade 3 strains, no subclade 3D strains were found, nor were any new HGT events involving other virulence factors. However, the clear propensity for the GAS emm89 strains to rapidly evolve through HGT recombinational replacement events generating novel variants of enhanced virulence warrants continued vigilance in monitoring of the epidemic clade 3 clone.

Genomic sequencing has revealed differences in MGE content, phages and ICEs, as the most common source of gene content differences between GAS strains [64]. GAS phages commonly encode one or more secreted virulence factors, such as streptococcal pyrogenic exotoxins (spe gene superantigens) and streptococcal phage DNases (spd gene exonuclease). GAS ICEs commonly encode one or more antibiotic resistance factors, usually for macrolides and tetracyclines. Small and large outbreaks of GAS infections have been associated with the acquisition of phages and/or ICEs. Although no specific ICE was detected in the Scotland cohort, the tetracycline resistance gene for TetM was detected in strain MGAS31641. An association with antibiotic resistance has not been reported to our knowledge in any investigation of the current geographically wide spread increase in emm89 infections. All differences in gene content between the Scotland strains were attributable to differences in phage content. Five prophages in nine combinations were found among the 122 Scotland strains (Fig. 2). The phage content, of the 121 Scotland clade 3 strains, mirrored that found for 756 emm89 clade 3 strains of the USA, Finland and Iceland [6]. The same three prophage genotypes (PG01, PG02, PG04) in the same ranking were found most prevalent. Moreover, the same spec and spd1 encoding phage was found most prevalent among the Scotland and England emm89 clade 3 strains [15]. The paucity of antibiotic resistance and the diversity of prophage profiles found among the emm89 clade 3 strains in multiple investigations indicates that no single MGE is conferring a strong selective advantage in the spread and persistence of the epidemic.

Considerable human flux occurs annually across the relatively short (174 km) border between Scotland and England. Thus, we anticipated considerable sharing of common genetic pool of emm89 strains between the two countries. To test this idea, we compared the SNP-based phylogeny obtained for this Scotland strain sample with the genome data published previously by Tuner et al. for the UK [15] available in public databases. The resultant combined phylogeny demonstrated extensive intermixing of the emm89 strains causing infections in these two areas (Fig. 6).

Given the clear recent numerical success of clade 3 organisms in multiple geographically dispersed countries, it is reasonable to speculate that clade 3 strains are present in other countries that have reported an increase in emm89 disease activity in recent years. However, full-genome sequence data are required to confirm this suspicion, since neither emm gene sequencing nor multi-locus sequence typing
provide sufficient genetic information to permit clade identification.

**Funding information**
This project was supported in part by the Fondren Foundation, Houston Methodist Research Institute.

**Acknowledgements**
We thank Drs Kathryn Stockbauer and Helen Chifotides for critical comment and editorial assistance, and Dr Shiranee Srisukanan for providing metadata for previously studied UK emm89 strains.

**Conflicts of interest**
The authors declare that there are no conflicts of interest.

**Ethical statement**
Patient data were anonymized. Thus, individual patient consent was not required.

**References**