Genome sequencing of an historic *Staphylococcus aureus* collection reveals new enterotoxin genes and sheds light on the evolution and genomic organisation of this key virulence gene family

Jo Dicks,[a] Jake Turnbull,[a] Julie Russell,[a] Julian Parkhill,[b] Sarah Alexander,[a,*]
Culture Collections, Public Health England, London, United Kingdom[a]; Department of Veterinary Medicine, University of Cambridge, Cambridge, CB3 0ES, United Kingdom[b]

Compiled February 4, 2021
This is a draft manuscript, pre-submission

Address correspondence to Sarah Alexander, Sarah.Alexander@phe.gov.uk.

ABSTRACT We take advantage of an historic collection of 133 *Staphylococcus aureus* strains accessioned between 1924 and 2016, whose genomes have been long-read sequenced as part of a major National Collection of Type Cultures (NCTC) initiative, to conduct a gene family-wide computational analysis of enterotoxin genes. We identify two novel Staphylococcal enterotoxin (pseudo)genes (*sel29p* and *sel30*), the former of which has not been observed in any contemporary strain to date. We provide further information on five additional enterotoxin genes or gene variants that have either recently entered the literature or for which the nomenclature or description is currently unclear (*selz*, *sel26*, *sel27*, *sel28* and *ses-2p*). An examination of over 11,000 RefSeq genomes in search of wider support for these seven (pseudo)genes led to the identification of an additional three novel enterotoxin gene family members (*sel31*, *sel32* and *sel33*) plus two new variants (*seh-2p* and *ses-3p*). We cast light on the genomic distribution of the enterotoxin genes, further defining their arrangement in gene clusters. Finally, we show that co-occurrence of enterotoxin genes is prevalent, with individual NCTC strains possessing as many as eighteen enterotoxin genes and pseudogenes, and that Clonal Complex membership rather than time of isolation is the key factor in determining enterotoxin load.

IMPORTANCE *Staphylococcus aureus* strains pose a significant health risk to both human and animal populations. Key amongst this species' virulence factors are the Staphylococcal enterotoxin gene family. Certain enterotoxin forms can induce a potentially life-threatening immune response, while others are implicated in less fatal though often severe conditions such as food poisoning. Genetic characterisation of Staphylococcal enterotoxin gene family members has steadily accumulated over recent decades, with over 20 genes now established in the literature. Despite the current wealth of knowledge on this important gene family, questions remain about the presence of additional enterotoxin genes and the genomic composition of family members. This study further expands knowledge of the Staphylococcal enterotoxins while shedding light on their evolution over the last century.

KEYWORDS: *Staphylococcus aureus*, enterotoxin gene family, genome analysis, National Collection of Type Cultures.

INTRODUCTION

Staphylococcus aureus is a Gram-positive, coccoid bacterium belonging to the Firmicutes phylum of mainly low G+C bacteria. *S. aureus* is a common member of the human microbiota, with studies estimating approximately 20-30% of the population to be long-term carriers of *S. aureus* strains in the skin, nostrils or female lower reproductive tract (1). In addition to its prevalence as a commensal organism of humans and animals, *S. aureus* is an important opportunistic pathogen. Strains can produce a variety of exotoxins, key amongst which are the staphylococcal enterotoxins (SEs), emetic toxins widely implicated in food poisoning. Gene family members are also associated with more severe, life-threatening conditions. For example, SEB is classified as a potential bioterrorism threat given its rapid and acute stimulation of the immune system, and it is also potentially implicated in the inducement of auto-immunity (2). Toxic shock syndrome (TSS) is a serious, and potentially fatal, condition with roughly half of cases denoted as menstrual-associated and the remainder as non menstrual-associated. TSST-1, a protein very closely related to the SEs, gives rise to the majority of menstrual-associated TSS cases and approximately half of the non menstrual cases, with the remainder - ~25% in total - associated with SEB and SEC (3, 4).

The SE and TSST-1 proteins are superantigens (SAgs), immunomodulatory toxins that have the ability to stimulate large populations of T cells by interacting with the Variable region of the β-chain (Vβ) of the T-cell receptor. Structurally, SAgs are two-domain proteins characterised by a β-grasp domain and an OB-fold domain. The SE proteins are encoded by a family of genes related by their DNA sequence. The recent literature on the Staphylococcal enterotoxin gene family encompasses 24 genes: *sea*, *seb*, *sec*, *sed*, *see*, *seg*, *seh*, *sei*, *selj*, *sek*, *sel*, *sem*, *sen*, *seo*, *sep*, *seq*, *ser*, *ses*, *set*, *selu*, *selv*, *selw* (formerly *selu2*), *selx* and *sely*. The protein encoded by the closely related *tsst-1* gene was initially discovered independently by two groups as PEC (staphylococcal pyrogenic exotoxin C; 5) and SEF (staphylococcal enterotoxin F; 6) and later renamed TSST-1 upon agreement, hence the absence of the *sef* nomenclature within the SE gene list. Only genes whose proteins have demonstrated emetic activity are given the "se" prefix, with others designated as "sel" for "stapylococcal enterotoxin-like". The nomenclature used here is largely taken from Fisher et al. (7).

In phylogenetic terms, the majority of the SE genes group closely to one another and with a number of *Streptococcus pyogenes* genes (8). The *selx* and *tsst-1* gene sequences group more distantly, with those of approximately 26 Staphylococcal Superantigen-like exoproteins (SSLs), which unlike the SAgs are immune invasion molecules and will not be considered here further. Despite its phylogenetic placement amongst the SSLs, *selx* is functionally similar to the SAg genes and is therefore referred to as an "SSL-like SAg" (8). Uniquely, SElX is a single-domain SAg, lacking the OB-fold domain seen in all other staphylococcal SAgs to date.

The majority of the SAg genes are located on mobile genetic elements such as pathogenicity islands, prophages and plasmids (7, 9). Clustering of the enterotoxin genes is also observed, most notably the *egc* cluster, which in a given strain can comprise up to seven genes and pseudogenes from a repertoire of nine (pseudo)gene forms (10). Consequently, while there is considerable variability with regard to the enterotoxin gene content between strains, the co-occurrence of the individual genes is highly non-random.

The National Collection of Type Cultures (NCTC) was founded in 1920 to address a recognised need for accumulating and disseminating information on human, animal, fungal and plant pathogens. It is one of four culture collections operated by Public Health England as part of a globally recognised biological resource centre, providing

many thousands of historical and emerging strains to researchers and biomedical scientists worldwide. Recently, a Wellcome-funded initiative to sequence the genomes of $\sim 3,000$ NCTC strains was completed. Amongst the datasets developed were Pacific Bioscience (PacBio) long-read sequences and associated genome assemblies for 133 *Staphylococcus aureus* strains accessioned between 1924 and 2016, with at least one strain isolated prior to June 1924. NCTC strains are accessioned either proactively, based upon scientific requests from one of its nine past and current curators or passively, deposited by members of the research community. While we think it unlikely that such a moderately-sized dataset would be representative - either geographically or temporally - of globally circulating *S. aureus* strains over the last century, the dataset is nonetheless diverse, with 43 distinct Sequence Types each represented by one or more strains.

Here, we analysed this new dataset on an historic strain collection to answer questions on the number of staphylococcal enterotoxin genes and their genomic organisation. We showed that each examined strain possessed between two and eighteen SE genes. We identified seven putative SE genes outside our search list, four of which were not seen in NCTC strains accessioned after 1951, and one of which is the most prevalent enterotoxin-like sequence identified to date. We also examined the genomes of over 11,000 *Staphylococcus aureus* strains in the RefSeq database in order to gain support for this expanded SE gene repertoire. The RefSeq dataset offered significant support for the newly identified genes while additionally presenting a further three SE genes and two gene variants. Collectively the two datasets shed light on the genomic distribution of SE genes, further delineating six gene clusters and introducing a new one. Crucially, the NCTC dataset enabled the examination of temporal patterns of enterotoxin birth and death to be made over a period of a century, showing a remarkable stability of gene content over this time, with all but one gene well represented in global sequence datasets. Finally, in accordance with this observed stability, analysing the inter-relationships between the NCTC strains showed that their Clonal Complex origins were more important than their time of isolation in determining their enterotoxin gene load.

RESULTS

The NCTC strains revealed novel enterotoxin-like sequences. Using a profile Hidden Markov Model approach to hunt for DNA sequences within a set of genomes provides the opportunity to find novel SE genes that have yet to be formally characterised. We identified 825 SE- and 20 *tsst-1*-like sequences within the 133 *S. aureus* strains, with genomic co-ordinates and annotation details provided within Data Set S1. The 845 sequences clustered into 29 easily distinguishable gene-specific groups. In addition to finding 22 of the 25 expected SE/*tsst-1* gene-specific groups (reference sequences for each group are shown in Table S1; no copies of *see*, *ses* or *set* were identified in any of the 133 strains), seven additional putative SE genes were identified, which we initially termed Gr1-Gr7. Strikingly, 133 copies of the Gr1 sequence were found, one in every strain analysed, spanning 42 distinct Sequence Types (including three not found in the *S. aureus* BIGSdb). While almost half (62 out of 133 copies) were likely to be pseudogenes due to premature stop codons or frameshift-inducing indels, seemingly intact versions of the coding sequence were seen across the accessioning period, with the most recent confirmed copy seen in strain NCTC 13434, isolated and accessioned in 2008.

All other groups (denoted Gr2 to Gr7) consisted of between two and sixteen members. The sixteen copies of Gr2 were found in strains isolated between 1932 and 2008.

Gr7 was less prevalent, with only five copies, but was observed over only a slightly reduced timespan, between 1938 and 1997. Despite its wide timespan, all five copies of Gr7 are likely to be pseudogenes. The other four genes were limited to the earlier strains, with Gr3/Gr4 (co-located), Gr5 and Gr6 most recently seen in strains accessioned in 1951, 1949 and 1948 respectively. Furthermore, both copies of Gr5 were presumed pseudogenes, likely the result of two small deletions. However, the prevalence of pseudogeny outside the cases of Gr1, Gr5 and Gr7 was generally less common, with fewer potential or likely pseudogenes in all other gene groups (see Data Set S1 for details).

Origins of the putative enterotoxin-like sequences. We searched for close sequence matches to each of the seven initially unidentified enterotoxin-like sequences in order to establish whether they had been observed previously. Overall, we found 882 high-scoring *megablast* hits to the GenBank nr/nt database (on 12/02/2020), with sequence-specific frequencies shown in Table S2a.

- **Gr1 is the near-ubiquitous chromosomal gene *sel26*.** The majority of the *megablast* hits, numbering 735 (83%, excluding 32 hits to NCTC genome sequences), were to the Gr1 nucleotide sequence at 93.24-100% sequence identity, including two full-length copies in a single strain (*S. aureus* strain ch22 chromosome; CP017807.1) and all but 17 of which were to complete genomes or chromosomes. The 17 gene hits were all annotated as enterotoxin-like W genes (e.g. 98.14% sequence identity to a purported *selw* gene identified in strain TD101; KX655716.1), though this gene was significantly different from the egc gene cluster *selw* gene used in our HMM search. It would appear that two distinct genes have been using the *selw* nomenclature: the egc gene formerly known as *selu2* and the seemingly ubiquitous (or at least highly prevalent) chromosomal gene previously seen in studies of human (11) and bovine (12) SEs. We will henceforth refer to this chromosomal gene as *sel26*, generally in line with the recommended nomenclature for SE genes (13) and similar to that used within (12), though currently lacking experimental confirmation to the best of our knowledge.
- **Gr2 is the *orfX*-associated gene *selz*.** We found 40 hits to the Gr2 nucleotide sequence, all at 96.54-99.49% sequence identity. Fourteen hits were to gene sequences, the majority to a gene recently described as a staphylococcal enterotoxin-like Z gene (*selZ*) in strains of *Staphylococcus argenteus* (14). Interestingly, four hits (e.g. KT316803.1) were annotated as a Staphylococcal Cassette Chromosome *mec* (SCC*mec*) element, a mobile genetic element implicated in broad-spectrum beta-lactam resistance via the mecA gene (15). An additional hit (U10927.2) appears to lie adjacent to an SCC*cap1* element, an SCC element with structural similarities to SCC*mec* but which instead harbours a type 1 capsular polysaccharide biosynthesis gene cluster (16). Further investigation of the annotation of U10927.2 shows that *selz* is the enterotoxin gene observed but unnamed in 2002 by Luong et al. (16). We searched for proximity of *selz* to *orfX* (which encodes an RlmH-type ribosomal methyltransferase) within each of the 16 strains with copies of Gr2/*selz*, as SCC elements are known to insert within the C terminus of this locus (17). We found that all copies of Gr2 were indeed in close proximity to *orfX* (between 2.5kb and 38.6kb), with two strains (NCTC 10399 and NCTC 10649) possessing an adjacent SCCcap element but none an SCCmec element. Six Gr2/*selz* copies were found in strains with Sequence Type 121, with others belonging to ST123, ST151, ST351, ST395, ST705, ST707, ST1254 and a novel Sequence Type. None of these strains belong to the six major identified

Clonal Complex groups (CC1, CC5, CC8, CC22, CC30 and CC97) in this study.

- **Gr3 and Gr4 are the clustered genes *sel27* and *sel28*.** The clustered Gr3 and Gr4 nucleotide sequences found 39 hits each, up to 98.41% and 99.12% sequence identity respectively. Of particular note were strong hits for each gene to a gene cluster pathogenicity island in strain 364P and to two recently identified enterotoxin genes annotated as *Sel27* and *Sel28* in strains SJTU F20365 (MF370878.1; 18), 86, 72, 50, SG19, SG16, SG13, SG11, SG09, SG05-2, SG05-1, SG04 and SG01. The oldest confirmed NCTC strain identified as carrying these two genes, NCTC 5664, was isolated in 1936 and the putative youngest, NCTC 8765, during or prior to 1951. Five of the seven NCTC strains possessing the two genes belong to ST9 (CC1) and the remainder to ST350.
- **Gr5 is the (pseudo)gene *sel29p*.** We failed to find any highly similar hits to the Gr5 nucleotide sequence. The two strains possessing this gene, NCTC 6966 and NCTC 7856, were accessioned in 1945 and 1949 respectively, both belong to ST890, and the two sequences appear to be pseudogenes. We propose that this gene is referred to henceforth as *sel29p*.
- **Gr6 is the plasmid gene *sel30*.** The Gr6 nucleotide sequence produced just eight high scoring hits, all to complete plasmid sequences and differing at most by one nucleotide substitution. The absence of any hits to annotated coding sequences meant it was difficult to form any further conclusions about the origins of this gene, other than its clear plasmid location. The two NCTC strains possessing the Gr6 gene were both accessioned in the 1940s and belonged to ST5 and ST1021. We propose that this gene is referred to henceforth as *sel30*.
- **Gr7 is the *orfX*-associated pseudogene *ses-2p*, a variant of *ses* clustered with *seh*.** Finally, the Gr7 nucleotide sequence, present in all five NCTC copies in close proximity to an *seh* gene sequence, found 21 hits ranging between 93.51% and 100% sequence identity. While the majority were to complete genome or chromosome sequences, two hits (EU272079.1 and KX690110.1) were - similar to Gr2 - to insertion sites of SCCmec elements. Further investigation of these hits identified previous reports of a partial enterotoxin gene with sequence similarity to *seo*, in close proximity to an *seh* gene and associated with SCCmec Type IV element insertion (19, 20). Similar to our analysis of Gr2, we searched for *orfX* genes within the five strains with copies of both Gr7 and *seh*, plus the single strain (NCTC 13435) possessing a presumed pseudogenised version of *seh* but no copy of Gr7. We found that all five copies of Gr7 and six copies of *seh*/*sehp* were close to an *orfX* gene (between 17.5kb and 43.1kb). However, only one copy of Gr7 (in NCTC 13297) was adjacent to an SCC element (in this case likely an SCC*fus* element) while the *sehp* copy in NCTC 13435 was adjacent to an SCC*mec* Type IV element. Further sequence analysis showed that while the 3' region of the Gr7 gene (~300bp) was highly similar to the corresponding region of *ses* (rather than *seo*), the 5' region of approximately 30bp was ~40bp shorter and dissimilar at the sequence level, with the intervening region showing a moderate level of sequence similarity interrupted by several presumed mutations. All copies of Gr7 in NCTC strains therefore look to be (only partially) truncated pseudogenes, containing several premature stop codons, though it is uncertain whether Gr7 was ever a functional gene. Four of the NCTC strains with the Gr7 sequence are ST10, with the remaining strain ST1. We propose that this pseudogene is referred to henceforth as *ses-2p*.

Most novel enterotoxin-like gene and pseudogene sequences are also observed within RefSeq genomes. We investigated the numbers of high-scoring hits

for each of the seven novel or recently identified (pseudo)genes within 11,351 *Staphylococcus aureus* genome sequences within the RefSeq database (on 11/05/2020). Of the 64,281 total hits to staphylococcal enterotoxin-like sequences, 12,505 were to six of these seven genes. Only Gr5/*sel29p* failed to find hits in strains other than the two NCTC strains (which were present within the RefSeq database). Table S2a shows that the relative frequencies of the other six genes within the NCTC dataset are mirrored within the *nr/nt* and *RefSeq* databases. This suggests to a certain extent that the frequencies of SE genes within the NCTC dataset are indicative of their frequencies within larger datasets, notwithstanding the absence of three SE genes from our dataset. Although we did find instances of *see*, *ses* and *set* within the sizeable RefSeq dataset (2, 39 and 36 copies, respectively), their low frequencies suggest these may be relatively rare genes, or certainly within strains whose genomes have been sequenced thus far.

The RefSeq database harbours a further cache of novel enterotoxin-like sequences. We classified 11,026 of the 11,351 RefSeq genomes into 468 distinct Sequence Types (see Data Set S2), 39 of them putative new STs. Analysis of all genomes led to the identification of a further three putative Staphylococcal enterotoxin genes plus two additional gene variants. In total, 258 sequences grouped into five distinct sets which we initially termed Gr8-Gr12. Similar to the NCTC-derived sequences, we searched the GenBank nr/nt database using BLAST to determine further information on the origins of these groups (see Table S2b).

- **Gr8 and Gr9 are the clustered genes *sel31* and *sel32*.** Fifteen copies of the Gr8 sequence in the RefSeq genomes were to a likely functional (based on its amino acid translation) gene which we term *sel31*. Of the eighteen Gr9 sequences, fifteen directly neighboured *sel31*. We term this new gene *sel32*, and hence delineate a new enterotoxin gene cluster. Fifteen of the eighteen RefSeq genomes could be classified as four Sequence Types (ST1, ST121, ST97, ST508) from three Clonal Complexes (CC1, CC45, CC97), see Data Set S3 for details of these strains. Each gene found three identical BLAST hits to the GenBank nr/nt database, from the same genomic sources, two of which were to plasmid genomes, highlighting the likely origin of the gene pair.
- **Gr10 is the egc cluster gene *sel33*, a recombinant of *selw* and *sen*.** A single copy of a new recombinant derivative of egc gene cluster genes *selw* and *sen* was identified in strain BSAC1477 from the BSAC Resistance Surveillance Project (GenBank accession NZ_FGMI01000018.1; BSAC), which we term *sel33*. This strain was isolated in or after 2001 and derives from CC22. Of the five sequences investigated here, only *sel33* failed to find any similar sequences within the *nr/nt* database, suggesting this recombination to be a rare occurrence.
- **Gr11 is the variant of *seh*, *seh-2p*, strongly associated with SCCmec Type IV elements.** The RefSeq dataset contained thirty copies of Gr11, a pseudogenised or truncated form of *seh* which we will refer to henceforth as *seh-2p*. The sequence of *seh-2p* likely possesses two single nucleotide deletions relative to the canonical form of *seh*. All but two classified genomes derive from ST80, with one each of ST4563 and a novel Sequence Type, both differing in a only single MLST allele from ST80 (both within the glpF gene). Notably, one of the ST80 strains is NCTC 13435, which unlike the five copies of *seh* in the NCTC strains, lacked a neighbouring *ses-2p* sequence. Similarly, none of the 30 copies of *seh-2p* within the RefSeq dataset have a neighbouring enterotoxin-like sequence, and 29 show significant evidence of a neighbouring SCCmecIVc(2B) element, with 27 genomes showing all SCCmec genes spread over one to four contigs and two additional genomes showing partial SCCmec matches (both including

mecA presence). In contrast, 219 of the 224 copies of *seh* possessed an adjacent *ses-2p* or *ses-3p* (see below) sequence, with 29 and 190 copies respectively.

- **Gr12 is the pseudogene *ses-3p*, a further variant of *ses* linked to *seh*.** We saw above that five copies of *ses-2p* were found within the NCTC genomes adjacent to copies of *seh*, and 29 such gene pairs were also observed within the RefSeq genomes. The Gr12 group of 194 sequences was found to constitute a second, distinct, variant of *ses* which we term *ses-3p* given its likely pseudogene status. This new variant is highly similar to *ses-2p* except for a divergent 5′ end. As noted above, most copies of *ses-3p* were found adjacent to *seh*, with only four of 194 instances lacking a neighbouring enterotoxin sequence. All but four classified genomes were found to be members of CC1 (ST1, ST81, ST474, ST1207, ST2764, ST3248, ST3497 and a novel ST), with the remainder from ST182 and ST944, Sequence Types highly distinct from CC1 but differing from one another in a single allele.

The novel enterotoxin gene sequences are spread across the SE phylogeny. A phylogenetic tree (Fig 1) was estimated from the amino acid sequences of eleven of the twelve putative novel or recently identified SE genes (or 'repaired' amino acid sequences in the cases of *sel29p*, *ses-2p* and *ses-3p*, and a 'short' sequence truncated by a premature stop codon in the case of *seh-2p*), alongside sequences of the established SE genes that were used in the pHMM search process. The tree shows the eleven sequences to group across the SE gene tree: *sel28*, *sel29p* and *sel32* with *sei*, *sek*, *sel*, *sem*, *seq* and *selv*; *sel26* and *sel30* with *sea*, *sed*, *see*, *selj* and *sep*; *seh-2p* with *seh*; *ses-2p*, *ses-3p* and *sel31* with *sen*, *seo* and *ses*; *selz* and *sel27* with *seb*, *sec*, *seg*, *ser*, *selu* and *selw*. The tree groupings of the established SE genes remained largely consistent with earlier analyses (e.g. 8) following the addition of the new gene family members. Note however that the amino acid sequence of *sel33* was omitted from the tree. As it is a recombinant gene derived from genes in two distinct clades (*selw* in the yellow clade and *sen* in the cyan clade in Fig 1), its inclusion distorts the topology of the resulting tree. This contrasts with *selv*, which derives from two genes within the same clade (*sem* and *sei*, purple clade in Fig 1).

There are at least seven Staphylococcal enterotoxin gene clusters. SE genes are known to sometime co-locate with others, often on plasmids or pathogenicity islands. The most striking example of this phenomenon is the egc gene cluster. The full characterisation of this operon has taken place in a stepwise fashion. The initial discovery of the *seg* and *sei* genes (21) was followed by identification of *sem*, *sen*, and *seo* in the neighbouring genomic regions, along with evidence of their co-transcription, while two pseudogenes (ϕent1 and ϕent2) were found between *sei* and *sen* (22). A sixth gene, *selu*, thought to be the product of deletions within ϕent1 and ϕent2) was identified later (23). Most recently, *selw* (formerly *selu2*) and *selv* were identified (10). While the nucleotide sequence of *selw* was highly similar to that of *selu*, the main difference being a 15bp deletion in the former compared to the latter, and thought to be the result of a different mutation of the ϕent1 and ϕent2 sequences from that hypothesised in *selu*, *selv* was found to be the product of a recombination of *sem* and *sei*.

The egc gene cluster appears to be highly prevalent within *S. aureus* genomes (14, 24). In this study, 59 egc gene clusters were found in 58 of the 133 *S. aureus* strains (43.6%), all but one seemingly complete. One of these 58 strains (NCTC 7972) possessed two gene clusters, with one of them the only case observed in this strain set of a large gap between any two egc genes, 13.6kb between *seo* and *sem*. In NCTC 11963, the only clear case of an incomplete egc gene cluster, we identified genes *seg* and *seo* on two separate genomic contigs, such that the omission of intervening genes

may be the result of an incomplete genome assembly rather than their absence from the genome. Further investigation of the raw sequencing reads overlapping this region showed the central region of the egc gene cluster to suffer from a very low read coverage but nonetheless to offer support for the existence of a full gene cluster most likely of type OMIUNG (read 27628 runs from the middle of *seo* to the middle of *seg* and is highly similar to the corresponding sequence of NCTC 2669, albeit with a single large run of T's breaking the alignment, presumably a sequencing artefact).

The NCTC dataset suggests the ϕent1 and ϕent2 pseudogenes should no longer be considered as entities distinct from *selu* and *selw*. While all four pseudogenised copies of *selw* and one of four pseudogenised copies of *selu* possess a single nucleotide frameshift (a run of 6 A's increased to 7 A's at positions 365 in *selw* and 380 in *selu*) that would lead to a two-ORF prediction similar to ϕent1 and ϕent2, the underlying nucleotide sequences are clearly merely a minor change to *selu* or *selw*. Furthermore, none of these sequences possess the 69bp deletion (relative to *selw*) observed in strain A900322, from which ϕent1 and ϕent2 were first defined (22), nor could we infer this deletion from any other strain sequence within the GenBank database. We feel that it is therefore more appropriate going forward to refer to *selu* or *selw* and their pseudogenes only. Given that the earliest copies of these genes within the NCTC collection (NCTC 2669 from 1928 for *selu* and NCTC 6134 from 1941 for *selw*) appear to be full-length, functional copies, the historical data would also support this view.

Frequencies of the distinct egc gene cluster arrangements identified in the NCTC strains are given in Table 1, showing that OMIUNG (i.e. the gene order *seo-sem-sei-selu-sen-seg*) and its close variant OMIWNG (including 'minor' pseudogenes of all six genes) are the predominant gene cluster variants, seen in this strain set in a ratio of 2.6:1. As well as three strains isolated and accessioned in the 1930s or 1940s possessing an OMIUN variant (i.e. apparent absence of the *seg* gene), we see a recent strain (NCTC 13373, accessioned in 2005 and equivalent to ATCC 43300, a clinical isolate from Kansas) with the OVUNG form, the potentially rare gene *selv* the result of a recombination between *sem* and *sei*. A comparison of the sequence of *selv* in NCTC 13373 to the canonical form in strain A900624 (10) from the French National Reference Center for Staphylococci (see Fig S1) shows evidence for the recombination between *sem* and *sei* in NCTC 13373 having occurred slightly closer to the 5' end of the sequence, though both events clearly took place within a central sequence region highly similar between the two progenitor genes. That observation, together with a high number of single nucleotide differences between the NCTC 13373 and *selv* reference sequences plus the alternative OVWNG form of the A900624 gene cluster, indicates that the two genes likely arose from two distinct recombination events. Note that the OVWNG form and the OMI33G form (i.e. containing the novel *sel33* recombinant of *selw* and *sen*), which we discovered in strain BSAC1477, were not observed within the NCTC dataset.

Five additional gene clusters were observed within the NCTC strains, as shown in Table 2. All but one instance of the 122 gene clusters, the broken egc cluster mentioned above, were found in intact gene cluster form. Interestingly, the *sel27-sel28* gene cluster, found within 7 strains accessioned in the 1930s-50s, was seen to be located close to the egc gene cluster (2 of the 3 OMIUN strains and 5 of the 14 OMIWNG strains). The distance of the egc cluster to the *sel27-sel28* gene cluster ranged between 17,465 and 39,464bp, with an average of approximately 27.6 kb.

While the *sek-seq* and *seh-ses-2p* gene clusters are comprised of genes within the same clade in Fig 1 (the cyan and orange groups are often referred to as a single clade elsewhere), the remaining four clusters contain genes spanning two or even three clades. In particular, the two most prevalent egc gene cluster arrangements (OMI-

UNG and OMIWNG) which account for 54 of its 59 copies, possess two genes from three main SE clades (shaded yellow, cyan and purple in Fig 1). It has been speculated that this divergence of the egc cluster genes may indicate the cluster's role as the progenitor of the majority of SE genes in *Staphylococcus aureus* (22). The seventh gene cluster, *sel31-sel32*, identified in fifteen RefSeq genome sequences from three Clonal Complexes, was not observed within the NCTC strains.

***Staphylococcal aureus* strains can possess many SE genes.** The frequencies of the 37 SE/*tsst-1* gene groups, including the twelve additional enterotoxin-like sequences, within the 133 NCTC strains are shown in Fig 2. Data Set S1 further shows the numbers of each gene identified within each strain. We see that the chromosomal genes *sel26* and *selx* are most common, with lesser frequencies of genes present on mobile genetic elements. Further work would be necessary to determine whether these frequencies were representative of the population as a whole or whether they were biased by sampling and temporal effects.

We found ten cases of a strain possessing two copies of the same gene and a single case of three gene copies (see Data Set S1). Notable examples include NCTC 7415, in which we found both three copies of *tsst-1* and two copies of the *sec-sel* gene cluster, and NCTC 7972, which harboured two intact egc gene clusters. In the former case, the likelihood of two of the three *tsst-1* copies and one of the *sec* copies being pseudogenes within a 20kbp region of a single contig may contribute to this finding.

Consistent with other studies such as Varshney et al. (24), individual strains were found to possess numerous SE/*tsst-1* genes, with a range of 2 to 18 genes per strain and a mean of 6.35 (median of 6). While this value is slightly higher than the average 5 SE genes per strain seen in (24), that prior study had looked at fewer genes, 19 of the 37 genes examined here, which may have led to the lower gene counts.

Associations between unclustered SE genes. Unlike the 19 SE genes involved in gene clusters, and which we discussed above, the *sea*, *seb*, *sep*, *selx*, *sely* and *tsst-1* genes are not clustered within this dataset in a conventional form, as neither are the newly identified *selz*, *sel26* and *sel30* genes nor the *sel29p* and *seh-2p* pseudogenes. Nevertheless, both positive and negative associations between these and other SE genes may still exist, likely the result of enterotoxin gene co-presence on plasmids, prophages, pathogenicity islands and other mobile genomic islands (9). Fig S2a shows a heatmap of Pearson correlation coefficients of gene presence/absence for all gene pairs (with the exception of *sel26*, which is always present), with genes arranged so that they are close to other genes with which they show the greatest associations.

The six gene clusters identified in the previous section are easily apparent as either single or sets of large red circles. Additional positive and negative associations are also apparent. Notable positive associations include *tsst-1* with the *sec-sel* gene cluster, *seb* with the *sek-seq* gene cluster, *sely* with *selz*, and *selw* with the *sel27-sel28* gene cluster. The former two associations are previously noted and likely the products of co-presence on SaPIm1/n1 (or SaPIbov1) and SaPI3 pathogenicity islands respectively. Examination of the relevant gene co-ordinates shows that the associations are without exception underpinned by co-presence on the same contig, though not by co-location. The most compact examples are the 11 cases of *seb*/*sek-seq*, which are approximately 11kb apart in all strains. However, to the best of our knowledge the latter two associations have not been observed prior to this study. *sely* has previously been seen only on the chromosome, so its association with the orfX-associated *selz* gene could be down to chance alone. The *selw*/*sel27-sel28* association is likely to involve particular forms of the *ν*Sa*β* genomic island, given the known presence of the egc cluster on this mobile element (25).

Notable negative associations include *selx* with the OMIUNG form of the egc gene cluster and the prophage-encoded *sea/sek-seq* gene cluster combination with OMIUNG and *sely*. Although Varshney et al. (24) note an absence of egc gene clusters within *seb*$^+$ strains, we observe a more complex pattern in this dataset. Examining the 17 *seb*$^+$ strains identified here, we find that 7 of the 8 strains isolated in or prior to 1949 carried the egc gene cluster whereas none of the 9 strains believed to derive from the 1950s onwards were found to possess it. Furthermore, no strains with *sel27* and *sel28* harboured a *seb* gene. Taken together, these observations suggest different SE gene combinations have circulated within the *S. aureus* population, some of which were restricted to particular timeframes, or perhaps to different parts of the *S. aureus* population. Fig S2b shows a depiction of the assocations observed in this study and/or described in Argudin et al. (9).

Phylogenetic and temporal patterns of Staphylococcal enterotoxin genes. We estimated a phylogenetic tree of the 133 NCTC *S. aureus* strains using HarvestTools (26) and IQ-TREE (27), based on 96,541 core genome SNPs, and annotated it with SE/*tsst-1* gene content and Clonal Complex group. It is immediately clear from Fig 3 that certain SE gene combinations are restricted to particular clades within the tree. For example, only CC1, CC5 and CC22 strains harbour the OMIWNG form of the egc gene cluster in this dataset. Indeed, the observed patterns of gene content explain many of the associations between genes described in the previous sections. For example, we see that *selx* is completely absent from CC30 (purple strip), a monophyletic group in which the OMIUNG form of the egc gene cluster is highly prevalent, thereby explaining the strong negative association between *selx* and *selu*.

Most clades cover a broad timespan (e.g. confirmed isolation periods of at least 1941-1997 for CC1; 1948-1988 for CC5; 1932-2003 for CC8; 1928-2003 for CC30) with a potential span of 1933-1985 for CC97 (though only presently confirmed up to 1954) and only CC22 represented here by a narrow group of strains (1990-2005). While Fig 3 suggests gene content to be highly correlated with Clonal Complex, the mean number of enterotoxin genes per strain was found to be higher in strains isolated within the 1920s-1940s (7.60) than in the 1950s-2010s (5.65), irrespective of Clonal Complex membership. We examined the relationship between the number of enterotoxin genes per strain with Clonal Complex and year of isolation by fitting a generalised linear model to the data, collapsing gene clusters to single observations as described in the Materials and Methods section. We found that for the 90 NCTC strains with Clonal Complex designations (see Fig S4 for plots of the data), the number of enterotoxin genes/gene clusters was strongly associated with Clonal Complex identity ($p < 0.05$ for three of the five factors when compared to CC1) but neither with year of isolation ($p = 0.274$) nor the interaction between Clonal Complex and time ($p > 0.425$ for all factor interactions). This suggests that enterotoxin gene content within Clonal Complexes has remained stable across the century of strain isolation and that the putative temporal difference in gene content described above may be due to sampling effects, with a higher frequency of strains harbouring the egc gene cluster isolated during the earlier period (60% vs. 34%).

HarvestTools Gingr plots of the core genome SNP alleles alongside the phylogenetic tree (see Fig S3 for an example, with NCTC 1803 as a reference genome) also indicate horizontal transmission between disparate *S. aureus* clades has taken place. Consequently, Staphylococcal enterotoxin gene content may also be influenced by horizontal processes in addition to clonal expansion. In future it would be interesting to analyse whether, for example, the two cases of the *sel27-sel28* gene cluster outside CC1 (uppermost green circles in Fig 3) were due to horizontal transfer of the

pathogenicity island on which they are located.

DISCUSSION

We analysed a strain set of 133 *Staphylococcus aureus* strains from the UK National Collection of Type Cultures, with a particular goal of understanding the complement of enterotoxin genes captured within, thereby further enhancing the utility of the strains for the benefit of the research community. While we did not initially anticipate uncovering any potential novel genes, particularly given the size of the dataset and the lack of an enterotoxin-focussed strategy for its collection, the use of a pHMM-profile approach allowed us to identify new sequences that we hope will be investigated further by researchers in this area. Given the relative ease with which we found putative enterotoxins first within the NCTC dataset, and subsequently within the RefSeq database, this leads us to speculate as to whether there might yet be other enterotoxin genes left for others to uncover. While the NCTC and RefSeq datasets encompass a sizeable proportion of the global diversity of *S. aureus* strains, with 43 and 468 distinct Sequence Types represented respectively, they will not have captured the full range. Consequently, as yet unidentified enterotoxin genes may still be present in strains whose genomes are currently outside the reach of strain and sequence collections.

Our study has also added to the understanding of the genomic organisation of the enterotoxin genes, particularly via gene clusters carried by mobile genetic elements. Gene families harboured by bacterial genomes are presented with an array of strategies that enable them to thrive and mobilise. The Staphylococcal enterotoxin gene family has indeed shown it is capable of exploiting many of these routes, from use of plasmids, prophages, pathogenicity islands and genomic islands in addition to stable chromosomal inheritance. Despite the consequent stability of the SE genes over the past century, in general large gene families appear to be rare in many bacteria. A study of the sequenced genomes of species including *Escherichia coli*, *Streptomyces pyogenes* and *Chlamydophila pneumoniae* showed limited numbers of gene families of size 20 or over, with only ~10 such gene families in *S. aureus* strains Mu50, MW2 and N315 (28). A more recent study found greater variation in the number of gene families between strains but again the number of duplicated genes was relatively limited, with a maximum of 190 duplicates for 84 gene families across 473 strains (29). Furthermore, the majority of duplicated genes in the latter study were thought to have a phage origin. The findings here are consistent with this observation.

The egc gene cluster has clearly been a key component in the expansion of the SE gene family. Interestingly, the genomic island harbouring the egc gene cluster was recently found, similarly to pathogenicity islands, to be capable of mobilisation due to a temporate bacteriophage (30). The egc is the only SE gene cluster to date that has been shown to have produced novel recombinant SE genes and from the growing number of components on offer, six distinct combinations were observed in the strains analysed here. As mentioned earlier, the egc gene cluster has been mooted as a putative SE nursery, whereby the observed genetic diversity has been generated by the processes of tandem duplication and subsequent divergence (22). The level of variation observed in this study would seem consistent with that view. Looking at the seven gene clusters in Table 2, all but one (*sek - seq*) has members from two or more of the shaded clades in Figure 1. In future, it might be illuminating to carry out a dating analysis of SE gene sequences to see if the results can help us to understand how these structures might have evolved. For example, the constituents of the *sec - sel* and *se27 - se28* clusters derive from the same two clades, and further these two clades are two of the three clades from which the egc genes all derive. It would be interesting to

determine whether these common features are due to a (partially) shared inheritance or whether they are coincident to independent origins.

The cases of *seh*(/*ses-2p*) and *selz* genes also indicate that transposition of genes to insertion sites otherwise used by elements such as Staphylococcal Cassette Chromosomes may be an additional strategy for gene survival and proliferation. Interestingly, Luong et al. (16) and Noto et al. (19) have suggested that enterotoxin insertion at this genomic location may have been implicated in the loss of ccrAB-mediated SCC element excision.

Alongside this general picture of gene family stability and proliferation, however, individual genes may also be lost. One interesting case is that of *sel29p*, observed in two ST890 strains isolated in or prior to the 1940's but not seen subsequently in any public sequence dataset. Could this gene have become extinct during the last 70 years? That it was seen only in a pseudogenised form could lend weight to such a hypothesis, as pseudogenisation may lead to gene excision or deterioration, particularly during host adaptation (31). To put the absence of copies of *sel29*/*sel29p* within the GenBank nr/nt and RefSeq databases into sharper view, we attempted to determine the Sequence Types of all 11,351 *S. aureus* genomes downloaded from RefSeq, and were able to easily achieve unambigious predictions without manual intervention in 11,026 cases (97%, see Data Set S2 for all predictions). We failed to find any further ST890 strains within this dataset. Consequently, an alternative hypothesis of restriction of this gene to the ST890 lineage cannot be ruled out. As well as genes that are widespread across *S. aureus* strains, we have seen cases of genes restricted to a narrow range of lineages (e.g. *sel33*), so this scenario would not be unprecedented. We further note that the lineage itself has not become extinct, with recent reports of ST890 strains derived from small mammals (32, 33). However, it does not appear that any of these strains have yet been subjected to whole genome sequencing. It will be interesting to discover patterns of *sel29*/*sel29p* presence and absence within these strains should sequencing data become available, particularly as at least one of the two NCTC ST890 strains were isolated from a different host (human; see Data Set S1).

Our analyses indicated an association between Clonal Complex and the number of SE genes/gene clusters in the strains investigated. It would be interesting to further investigate possible associations between CCs and SE gene profiles (e.g. the particular pattern of SE genes a strain possesses), though the larger RefSeq dataset may be required to achieve statistical significance. Many studies have indicated associations between SE gene profiles and disease type, for example between the egc gene cluster and both cystic fibrosis (34) and Toxic Shock Syndrome (35). Other studies have shown links between disease and both SE profile and CC, such as those between CC30, infective endocarditis and the genes *tsst-1*, *sea*, *sed*, *see* and *sei* (36). Differences in regulatory system (7) may contribute to disease/SE gene associations, with some SE genes more likely to occur in chronic rather than acute infections. The plasticity of the mobile genetic elements carrying the majority of SE genes, with the different variant combinations circulating (e.g. see Figure S2b), and the potential for their rapid loss and gain, mean that selection could act swiftly on SE gene profiles. Considering such a system, it would seem plausible that strains with particular SE gene profiles would be selected for their roles in specific diseases and that clonal expansion would subsequently drive (at least some) CCs specialised for certain diseases. Limited within-CC recombination (37) would preserve these associations, establishing the patterns we see among extant and preserved strains. However, SE gene co-location might also mean some observed associations are indirect. A substantial meta-analysis of sequenced strains with high quality disease status would undoubtedly be illuminating. It would be interesting to

investigate, for example, whether the absence of *selx* on CC30 strains and resulting negative associations between this gene and other genes prevalent (e.g. *tsst-1*, *selu*) in this CC are due to simple gene loss and subsequent clonal expansion or due to strong selection for gene content.

Possessing multiple SEs may also be an advantage in itself. Distinct SE constituents of the egc gene cluster have been shown to exhibit different Vβ specificities, and are therefore likely to have complementary effects on a host's immune system (22). Strains possessing multiple SEs could therefore possess a selective advantage regarding host colonisation/invasion. Additionally, should two or more SEs be genetically linked, such as in the egc gene cluster, there are further opportunities for production of novel SE forms through processes such as recombination, such as we have seen here with the single observed case to date of *sel33*. Indeed, the high prevalence of distinct, non-trivial SE combinations presents a problem to researchers attempting to produce SE toxin-based vaccines for *S. aureus* infections. Despite promising results concerning strains with simple SE profiles (38), producing such a vaccine against strains with multiple SE genes remains a challenge (39). Consequently, disease-specific SE-based vaccines may be required, with a tailored combination of anti-toxins.

Of the 845 SE and *tsst-1* gene sequences identified in this study, 123 (14.6%) were potential or likely pseudogenes. Approximately half of these cases were the *sel26* gene. The high rate of pseudogenisation of this gene is potentially a consequence of its chromosomal location, as excision is not so easily possible as it is for the majority of SE genes residing on mobile elements. That said, we see a much lower rate of pseudogenisation for *selx*, another chromosomal gene with only 4.6% cases present in strains with distinct Sequence Types. The different rates of the two genes may be due to the age of the genes or their importance in certain environmental (e.g. disease) niches, which further research might uncover. In general, rates of pseudogenisation vary both between genes and strains, indicating that selection may have played a key role in this process. Eleven strains possess three or more pseudogenes. For example, NCTC 6133 has seven cases, including four of its six egc genes and the *selx* gene. Within the egc gene cluster, *seg* has the greatest number of pseudogenes. The observation of OMIUN egc arrangements in other strains may indicate that *seg* is not always essential to the success of the gene cluster. Most cases of pseudogenisation do appear to be minor sequence changes to an established gene, which are likely to render it dysfunctional. However, the *ses-2p* gene adjacent to the *orfX* locus is more intriguing and has potential to be an emerging SE gene that has yet to be functional. Further sequencing of past and future *S. aureus* strains may shed light on the evolutionary trajectory of this sequence.

Here we have shown how the analysis of even a medium-sized strain set can provide valuable information to the study of an important bacterial gene family. An added dimension to our analysis is that the strain set was collected over a period of almost a century, thereby granting access to biological material that can no longer be collected today. However, while unique in terms of the specific strains involved, there are other collections worldwide that will now be contemplating or even carrying out sequence programs such as that which enabled this study. It will be interesting to see what information emerges.

MATERIALS AND METHODS

Dataset preparation. The genome assemblies of 133 *Staphylococcus aureus* strains (see Data Set S1 for strain identities) derived from PacBio raw reads at the Wellcome Sanger Institute (WTSI) were downloaded in FASTA format from GenBank. The UniProt

database was searched for protein and nucleotide sequences attributed to each of the 25 target genes (24 SEs plus *tsst-1*) from other *S. aureus* strains. In the two cases where no SE matches were found (*selv* and *selw*), sequences were instead obtained from GenBank. Sequences were divided into two sets, where set 1 consisted of *selx* and *tsst-1* and set 2 consisted of the remaining 23 SEs.

Enterotoxin gene hunting. The software tool HMMER (40) was used to build profile hidden Markov models (pHMMs) for each gene set. The two pHMMs were then used to search the 133 genome assemblies for target gene matches. Searches were made using stringent parameters to guarantee full length, or close to full length, matches (for set 1, $E < 1 \times e^{-10}$, $a < 5$ and $b > 600$; for set 2, $E < 1 \times e^{-10}$, $a < 88$ and $b > 590$; where a is the starting co-ordinate of the match relative to the pHMM and b is the end co-ordinate of the match) and more relaxed parameters ($E < 1 \times e^{-10}$ for sets 1 and 2). Co-ordinates were chosen following visual inspection of the initial HMMER output for a non-trivial subset of strains.

The HMMER accessory tool Easel was subsequently used to extract the nucleotide sequences of all predicted target genes in all strains, keeping the two gene sets distinct. Gene-specific pHMMs were also built for each of the 25 target genes and HMMER was again used to search for and extract predicted sequences, this time separated by gene identity. The gene-specific datasets were compared to the set-specific results for all strains to confirm the gene family wide-approach was consistent with the gene-specific approach. No inconsistencies were identified. Extracted nucleotide sequences were aligned, along with 25 reference sequences (one for each target gene - see Table S1 for the GenBank accession numbers of all SE gene references), using MUSCLE (41) and were manually divided into gene-specific groups within BioEdit (42). Using the reference sequences as guides, gene co-ordinates within the HMMER output were manually adjusted to ensure all sequences were full length. Using the modified co-ordinates, full length target gene matches were extracted from the 133 *S. aureus* strains with Easel and re-aligned into gene-specific groups.

We also attempted both to gain support for the existence of the putative novel enterotoxin genes identified via this approach and to glean information on their origins by analysing additional *S. aureus* genomes. Briefly, all 11,351 genomes within the RefSeq database (43) available on 11/05/2020 were downloaded and were subjected to an almost identical HMM-searching procedure as the 133 NCTC genome sequences. The only difference in the two procedures was the use of MAFFT (44) for gene sequence alignment, in place of MUSCLE, due to its ability to align tens of thousands of gene sequences within a few hours (using parameters *–retree 1 –maxiterate 0 –reorder*).

Phylogenetic analysis. Translated amino acid sequences of novel or recently identified gene-specific groups not included in the gene hunting process, one from each group, were aligned using MUSCLE along with reference sequences for previously known groups and the alignment input to the IQ-TREE phylogenetic software (27) with amino acid substitution model selection requested. During this process, the translated reference sequences for four groups (Gr5, Gr7, Gr11 and Gr12) were “repaired" with minor manual editing as the group members appeared to be pseudogenes with various mutations such as indel-causing frameshifts and premature stop codons. The repairs, which effectively estimated the amino acid states of the sequences before their putative pseudogenisation but after their divergence from the other enterotoxin sequences, were made to maximise the phylogenetic signal in the dataset and hence the reliability of the resulting tree. The resulting sequences are shown in Table S3. We also compared the nucleotide reference sequences of all established and putative gene family groups with the GenBank nr/nt database using BLAST (45).

Strain typing. Multilocus Sequence Typing (MLST) was conducted for each NCTC strain using the established seven gene set for *S. aureus* (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi* and *yqiL*) (46). For each gene, all *S.aureus* sequences in the relevant BIGSdb database (47) were downloaded and a pHMM calculated using HMMER. For each strain the MLST gene sequences were extracted and concatenated into a single file, with the file subsequently input to BIGSdb for MLST characterisation and identification of Clonal Complex. We also carried out spa-typing for each strain, whereby the combination of differing repeat sequence types within the *SpA* gene was established. This process was carried out using the get_spa_type.py software, which compares repeats found between pairs of primer sequences against the Ridom and eGenomics typing nomenclature. For 12 strains, manual editing of their genome sequence was required to achieve a spa type, as sequence mutations within their *SpA* genes meant that 100% matches to primer sequences were no longer achievable, preventing the software from extracting the repeats. All Sequence Type, Clonal Complex and Spa Type predictions are shown in Data Set S1. The Sequence Type/Clonal Complex designation process was repeated for the 11,351 *S.aureus* genome assemblies downloaded from the RefSeq database. Predictions for the 11,026 strains (97%) for whom manual intervention was not required to achieve a result are shown in Data Set S2.

SNP analysis. Each NCTC genome assembly was compared to that of NCTC 1803 (one full-length chromosome only, represented by a single contig) with Parsnp from the HarvestTools suite (26). The MFA file output from the suite's Gingr tool, which consisted of core genome single nucleotide polymorphisms across the 133 strain set, was used as input to the IQ-TREE phylogenetic analysis tool. Gingr was also used to visualise patterns of recombination between the strains.

Statistical analysis. The gene contents of strains for whom a Clonal Complex origin was determined were analysed along with year of isolation (set to the most recent year possible given the strain metadata shown in Data Set 1) within the R statistical environment (version 3.6.1) (48). A generalised linear model with a logarithmic link function and Poisson error distribution was fitted to the remaining data, with the number of genes/gene clusters (gene clusters were counted as if they were a single gene to account for the dependence of gene number counts on gene cluster presence and absence) as the independent variable and Clonal Complex (a factor with six levels) and year (after 1924) as dependent variables.

Data availability. All NCTC genomes have been deposited in the NCBI Sequence Read Archive under NCBI BioProject accession number PRJEB6403. For individual accession numbers, please see Data Set S1, worksheet 1, in the supplemental material.

ACKNOWLEDGMENTS

The authors are indebted to the work of all past NCTC staff and collaborators for their efforts in establishing, curating and maintaining this collection, which at time of writing is celebrating its 101st year of operation. We also express our gratitude to all who deposited strains into the NCTC and which were examined in this manuscript. We would also like to thank three anonymous reviewers for their supportive and insightful comments. The PacBio sequencing and genome assembly of the 133 *S. aureus* strains analysed within this article were funded as part of the Wellcome Trust Grant no: 101503/Z/13/Z "Creation of an e-resource centre to underpin the provision and use of Type and reference strains of human pathogens".

REFERENCES

1. **Graham 3rd PL, Lin SX, Larson EL**. 2006. A U.S. population based survey of Staphylococcus aureus colonization. Ann Intern Med 144:318–325.
2. **Chowdhary VR, Tilahun AY, Clark CR, Grande JP, Rajagopalan G**. 2012. Chronic exposure to staphylococcal superantigen elicits a systemic inflammatory disease mimicking lupus. J Immunol 189:2054–2062.
3. **McCormick JK, Yarwood JM, Schlievert PM**. 2001. Toxic Shock Syndrome and Bacterial Superantigens: An Update. Annu Rev Microbiol 55:77–104.
4. **McCormick JK, Tripp TJ, Llera AS, Sundberg EJ, Dinges MM, Mariuzza RA, Schlievert PM**. 2003. Functional Analysis of the TCR Binding Domain of Toxic Shock Syndrome Toxin-1 Predicts Further Diversity in MHC Class II/Superantigen/TCR Ternary Complexes. The J Immunol 171 (3):1385–1392.
5. **Schlievert PM, Shands KN, Dan BB, Schmid GP, Nishimura RD**. 1981. Identification and characterization of an exotoxin from *Staphylococcus aureus* associated with toxic-shock syndrome. J Infect Dis 143 (4):509–516.
6. **Bergdoll MS, Crass BA, Reiser RF, Robbins RN, Davis JP**. 1981. A new staphylococcal enterotoxin, enterotoxin F, associated with toxic-shock-syndrome *Staphylococcus aureus* isolates. Lancet 1(8228):1017–1021.
7. **Fisher EL, Otto M, Cheung GYC**. 2018. Basis of Virulence in Enterotoxin-Mediated Staphylococcal Food Poisoning. Front Microbiol 9:436.
8. **Langley RJ, Ting YT, Clow F, Young PG, Radcliff FJ, Choi JM, Sequeira RP, Holtfreter S, Baker H, Fraser JD**. 2017. Staphylococcal enterotoxin-like X (SElX) is a unique superantigen with functional features of two major families of staphylococcal virulence factors. PLoS Pathog 13 (9):e1006549.
9. **Argudin MA, Mendoza MC, Rodicio MR**. 2010. Food Poisoning and *Staphylococcus aureus* Enterotoxins. Toxins 2:1751–1773.
10. **Thomas DY, Jarraud S, Lemercier B, Cozon G, Echasserieau K, Etienne J, Gougeon ML, Lina G, Vandenesch F**. 2006. Staphylococcal Enterotoxin-Like Toxins U2 and V, Two New Staphylococcal Superantigens Arising from Recombination within the Enterotoxin Gene Cluster. Infect Immun 74 (8):4724–4734.
11. **Okumura K, Shimomura Y, Murayama SY, Yagi J, Ubukata K, Kirikae T, Miyoshi-Akiyama T**. 2012. Evolutionary paths of streptococcal and staphylococcal superantigens. BMC Genom 13:404.
12. **Wilson GJ, Tuffs SW, Wee BA, Seo KS, Park N, Connelley T, Guinane CM, Morrison WI, Fitzgerald JR**. 2018. Bovine Staphylococcus aureus Superantigens Stimulate the Entire T Cell Repertoire of Cattle. Infect Immun 86 (11):e00505–18.
13. **Lina G, Bohach GA, Nair SP, Hiramatsu K, Jouvin-Marche E, Mariuzza R**. 2004. Standard Nomenclature for the Superantigens Expressed by Staphylococcus. The J Infect Dis 189 (12):2334–2336.
14. **Aung MS, Urushibara N, Kawaguchiya M, Sumi A, Takahashi S, Ike M, Ito M, Habadera S, Kobayashi N**. 2019. Molecular Epidemiological Characterization of *Staphylococcus argenteus* Clinical Isolates in Japan: Identification of Three Clones (ST1223, ST2198, and ST2550) and a Novel Staphylocoagulase Genotype XV. Microorganisms 7 (10):E389.
15. **IWG-SCC**. 2009. Classification of Staphylococcal Cassette Chromosome *mec* (SCC*mec*): Guidelines for Reporting Novel SCC*mec* Elements. Antimicrob Agents Chemother 53 (12):4961–4967.
16. **Luong TT, Ouyang S, Bush K, Lee CY**. 2002. Type 1 capsule genes of Staphylococcus aureus are carried in a staphylococcal cassette chromosome genetic element. J. Bacteriol. 184:3623–3629.
17. **Boundy S, Safo MK, Wang L, Musayev FN, O'Farrell HC, Rife JP, Archer GL**. 2013. Characterization of the Staphylococcus aureus rRNA Methyltransferase Encoded by orfX, the Gene Containing the Staphylococcal Chromosome Cassette mec (SCCmec) Insertion Site. The J Biol Chem 288 (1):132–140.
18. **Zhang DF, Yang XY, Zhang J, Qin X, Huang X, Cui Y, Zhou M, Shi C, French NP, Shia X**. 2018. Identification and characterization of two novel superantigens among *Staphylococcus aureus* complex. Int J Med Microbiol 308 (4):438–446.
19. **Noto MJ, Kreiswirth BN, Monk AB, Archer GL**. 2008. Gene acquisition at the insertion site for SCCmec, the genomic island conferring methicillin resistance in *Staphylococcus aureus*. J Bacteriol 190 (4):1276–1283.
20. **Tunsjø HS, Kalyanasundaram S, Worren MM, Leegaard TM, Moen AEF**. 2017. High frequency of occupied attB regions in Norwegian *Staphylococcus aureus* isolates supports a two-step MRSA screening algorithm. Eur J Clin Microbiol Infect Dis 36 (1):65–74.
21. **Munson SH, Tremaine MT, Betley MJ, Welch RA**. 1998. Identification and characterization of staphylococcal enterotoxin types G and I from Staphylococcus aureus. Infect. Immun. 66:3337—-3348.
22. **Jarraud S, Peyrat MA, Lim A, Tristan A, Bes M, Mougel C, Etienne J, Vandenesch F, Bonneville M, Lina G**. 2001. egc, A Highly Prevalent Operon of Enterotoxin Gene, Forms a Putative Nursery of Superantigens in Staphylococcus aureus. The J Immunol 166 (1):669–677.
23. **Letertre C, Perelle S, Dilasser F, Fach P**. 2003. Identification of a new putative enterotoxin SEU encoded by the egc cluster of Staphylococcus aureus. J Appl Microbiol 95 (1):38–43.
24. **Varshney AK, Mediavilla JR, Robiou N, Guh A, Wang X, Gialanella P, Levi MH, Kreiswirth BN, Fries BC**. 2009. Diverse Enterotoxin Gene Profiles among Clonal Complexes of *Staphylococcus aureus* Isolates from the Bronx, New York. Appl Environ Microbiol 75 (21):6839–6849.
25. **Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K**. 2008. Genome sequence of Staphylococcus aureus strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. J Bacteriol 190:300—-310.
26. **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. 15:524.
27. **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):268–274.
28. **Pushker R, Mira A, Rodríguez-Valera F**. 2004. Comparative genomics of gene-family size in closely related bacteria. Genome Biol 5:R27.
29. **Sanchez-Herrero JF, Bernabeu M, Prieto A, Hüttener M, Juárez A**. 2020. Gene Duplications in the Genomes of Staphylococci and Enterococci. Front molecular biosciences 7:160.
30. **Moon BY, Park JY, Hwang SY, Robinson DA, Thomas JC, Fitzgerald JR, Park YH, Seo KS**. 2015. Phage-mediated horizontal transfer of a *Staphylococcus aureus* virulence-associated genomic island. Sci Reports 5:9784.
31. **Goodhead I, Darby AC**. 2015. Taking the pseudo out of pseudogenes. Curr Opin Microbiol 23:102—-109.
32. **Monecke S, Gavier-Widén D, Hotzel H, Peters M, Guenther S, Lazaris A, Loncaric I, Müller E, Reissig A, Ruppelt-Lorz A, Shore AC, Walter B, Coleman DC, Ehricht R**. 2016. Diversity of *Staphylococcus aureus* Isolates in European Wildlife. PLoS ONE 11 (12):e0168433.
33. **Mrochen DM, Schulz D, Fischer S, Jeske K, El Gohary H, Reil D, Imholt C, Trübe P, Suchomel J, Tricaud T, Jacob J, Heroldová M,**

Bröker BM, Strommenger B, Walther B, Ulrich RG, Holtfreter S. 2018. Wild rodents and shrews are natural hosts of *Staphylococcus aureus*. Int J Med Microbiol 308:590–597.

34. **Fischer AJ, Kilgore SH, Singh SB, Allen PD, Hansen AR, Limoli DH, Schlievert PM**. 2019. High Prevalence of *Staphylococcus aureus* Enterotoxin Gene Cluster Superantigens in Cystic Fibrosis Clinical Isolates. Genes 10:1036.
35. **Jarraud S, Cozon G, Vandenesch F, Bes M, Etienne J, Lina G**. 1999. Involvement of enterotoxins G and I in staphylococcal toxic shock syndrome and staphylococcal scarlet fever. J Clin Microbiol 37 (8):2446–2449.
36. **Nienaber JJ, Sharma Kuinkel BK, Clarke-Pearson M, Lamlertthon S, Park L, Rude TH, Barriere S, Woods CW, Chu VH, Marín M, Bukovski S, Garcia P, Corey GR, Korman T, Doco-Lecompte T, Murdoch DR, Reller LB, Fowler Jr VG, on Endocarditis-Microbiology Investigators IC**. 2011. Methicillin-susceptible *Staphylococcus aureus* endocarditis isolates are associated with clonal complex 30 genotype and a distinct repertoire of enterotoxins and adhesins. J Infect Dis 204 (5):704–713.
37. **Driebe EM, Sahl JW, Roe C, Bowers JR, Schupp JM, Gillece JD, Kelley E, Price LB, Pearson TR, Hepp CM, Brzoska PM, Cummings CA, Furtado MR, Andersen PS, Stegger M, Engelthaler DM, Keim PS**. 2015. Using Whole Genome Analysis to Examine Recombination across Diverse Sequence Types of Staphylococcus aureus. PLOS ONE 10 (7):1–19.
38. **Aman MJ**. 2017. Superantigens of a superbug: Major culprits of *Staphylococcus aureus* disease? Virulence 8 (6):607–610.
39. **Aguilar JL, Varshney AK, Pechuan X, Dutta K, Nosanchuk JD, Fries BC**. 2017. Monoclonal antibodies protect from Staphylococcal Enterotoxin K (SEK) induced toxic shock and sepsis by USA300 Staphylococcus aureus. Virulence 8 (6):741–750.
40. **Wheeler TJ, Eddy SR**. 2013. nhmmer: DNA Homology Search With Profile HMMs. Bioinformatics 29:2487–2489.
41. **Edgar RC**. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (5):1792–1797.
42. **Hall TA**. 1999. BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98.
43. **O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD**. 2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44 (D1):D733–D745.
44. **Katoh K, Standley DM**. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30 (4):772–780.
45. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
46. **Enright MC, Day NPJ, Davies CE, Peacock SJ, Spratt BG**. 2000. Multilocus Sequence Typing for Characterization of Methicillin-Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. J Clin. Microbiol. 38 (3):1008–1015.
47. **Jolley KA, Maiden MCJ**. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinform 11:595.
48. **R Core Team**. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
49. **Letunic I, Bork P**. 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23 (1):127–128.
50. **Bohach GA, Schlievert PM**. 1989. Conservation of the Biologically Active Portions of Staphylococcal Enterotoxins Cl and C2. Infect Immun 57 (7):2249–2252.

733 ## TABLES

TABLE 1 egc gene cluster forms within 133 NCTC *Staphylococcus aureus* genomes

Cluster composition	No. of copies
OMIUNG	39
OMIWNG	15
OMIUN	3
OVUNG	1
Unconfirmed (O and G only)	1
OVWNG	0
OMI33G	0
Total	59

TABLE 2 Staphylococcus enterotoxin gene clusters within 133 NCTC *Staphylococcus aureus* genomes

Cluster composition	No. of copies
egc	59
sek - seq	23
sec - sel	19
sed - selj - ser	9
sel27 - sel28	7
seh - ses-2p	5
sel31 - sel32	0
Total	122

734 ## FIGURES

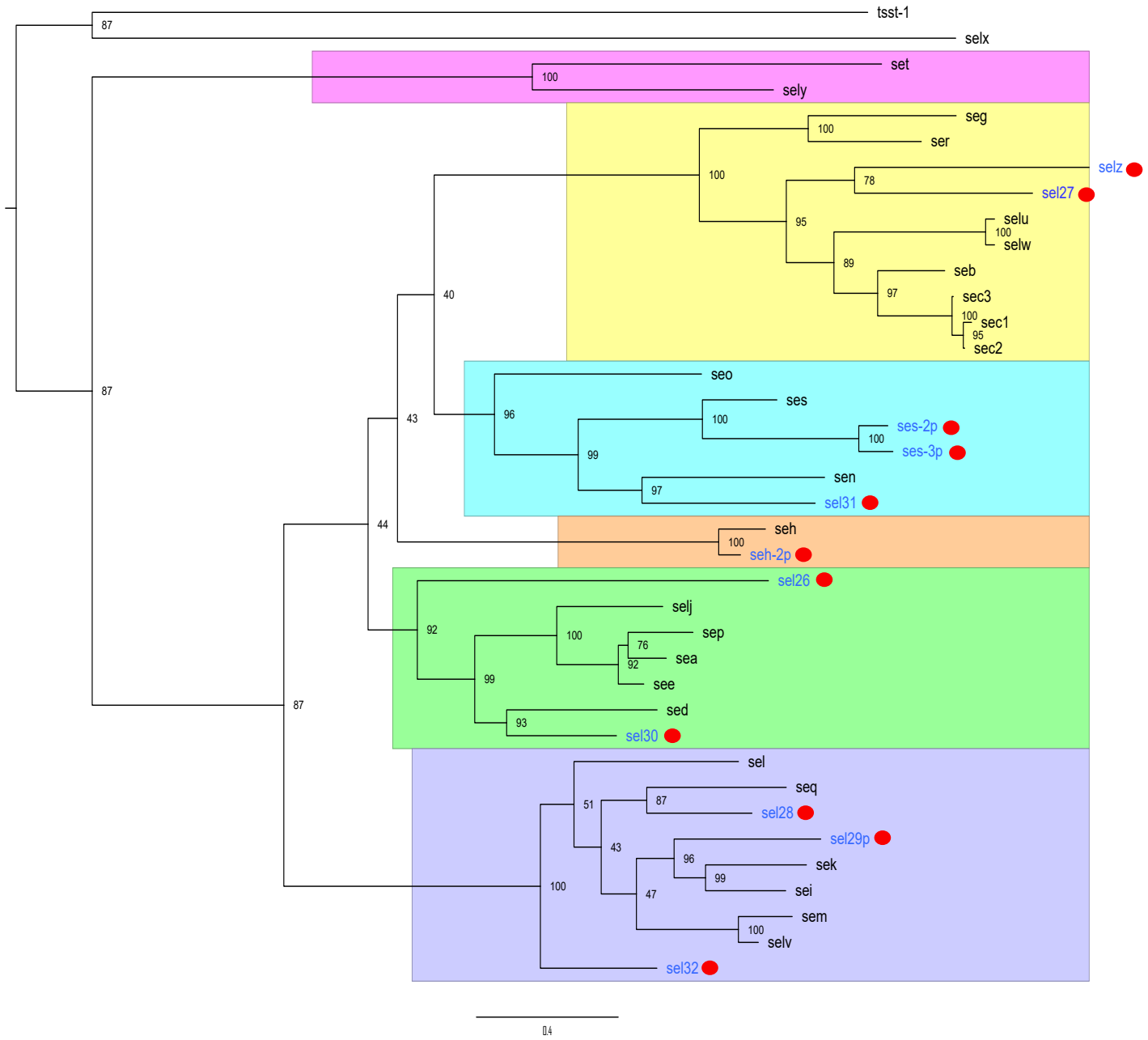

FIG 1 Phylogeny of the Staphylococcal enterotoxin genes.
Maximum likelihood (ML) phylogenetic tree of eleven of the twelve new Staphylococcal enterotoxin gene family members (gene names shown in blue text and with adjacent red circle; *sel33* is not shown as its between-clade recombinant origin distorts the tree topology) identified in the NCTC and RefSeq strain sets, along with reference sequences for 24 previously identified SE genes (including three variants of *sec*) plus *tsst-1*. Compact gene groups (clades) are highlighted as coloured blocks. The tree was estimated with IQ-TREE (27) using the VT+F+R4 amino acid substitution model, maximum log-likelihood = -14507.9726, and with the *tsst-1*/*selx* clade used as an outgroup. 1000 ultrafast bootstraps were performed, with percentages of bootstrapped trees supporting the ML tree shown at each internal node. The tree was further annotated by clade with FigTree.

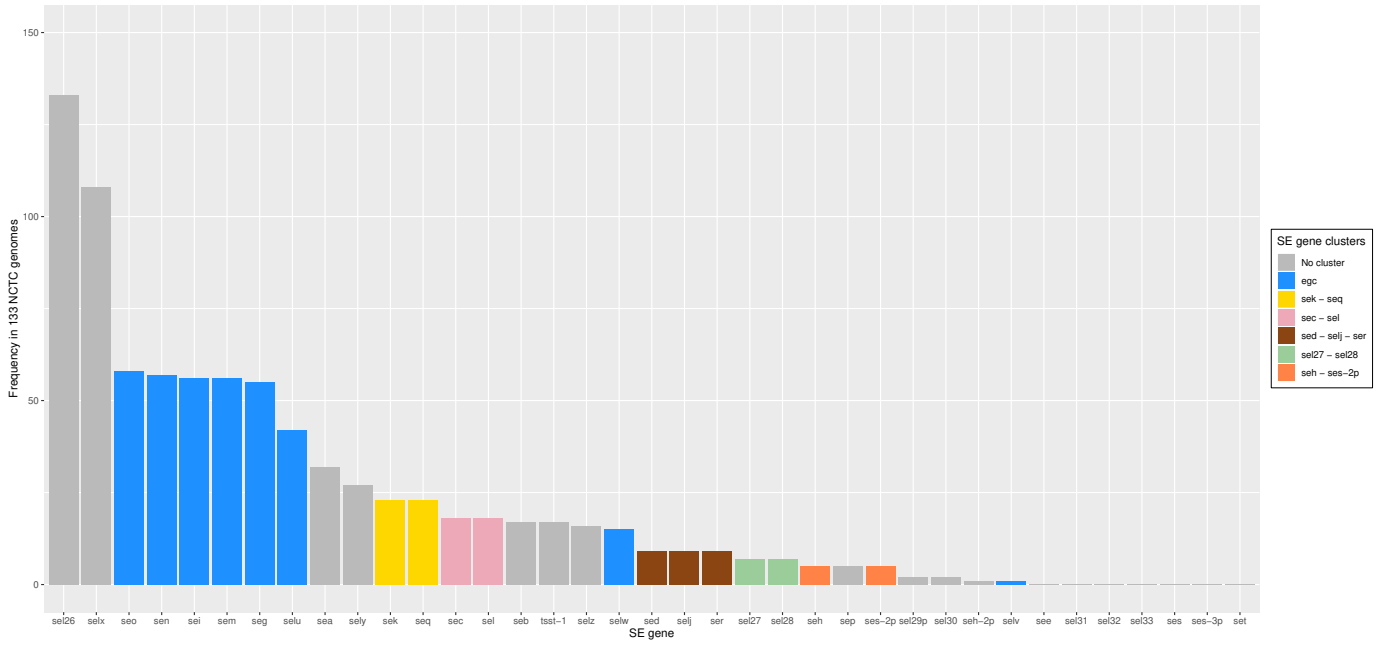

FIG 2 Staphylococcal enterotoxin gene presence. Frequencies of 36 Staphylococcal enterotoxin genes (or putative genes/pseudogenes) plus *tsst-1* in 133 NCTC *S. aureus* strains. Membership of one of the six gene clusters present in this dataset is indicated by a colour code.

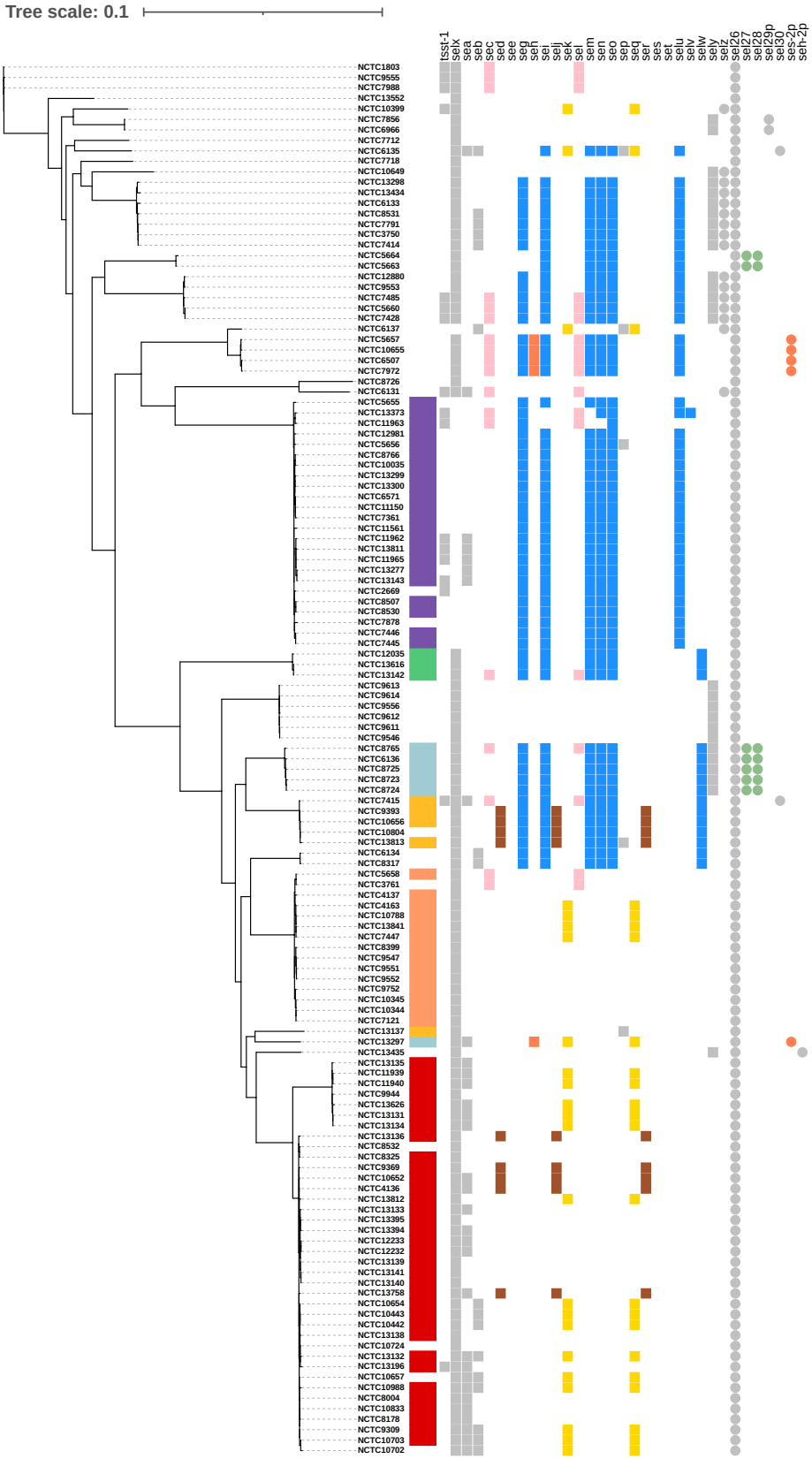

FIG 3 Staphylococcal enterotoxin gene content of 133 NCTC *Staphylococcus aureus* strains. An unrooted maximum likelihood phylogenetic tree based on 96,541 core genome SNPs is annotated with the Clonal Complex (colour strip adjacent to strain names: CC1 blue; CC5 gold; CC8 red; CC22 green; CC30 purple; CC97 orange) and SE/*tsst-1* gene content (established genes as squares and recent/novel genes as circles). Gene presence is coloured according to the scheme in Fig 2, so that membership of a common gene cluster can be identified easily. SNPs were called using HarvestTools. The tree was estimated with IQ-TREE using the SYM+ASC+R3 nucleotide substitution model, with a maximum log-likelihood = -1022309.6472. The figure was generated using the iTOL web server (49).