A peer-reviewed version of this preprint was published in PeerJ on 12 July 2018.

Petit RA III, Read TD. 2018. *Staphylococcus aureus* viewed from the perspective of 40,000+ genomes. PeerJ 6:e5261
https://doi.org/10.7717/peerj.5261

Staphylococcus aureus viewed from the perspective of 40,000+ genomes

Robert A Petit III [1] , **Timothy D Read** Corresp. [1]

[1] Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, Georgia, United States

Corresponding Author: Timothy D Read
Email address: tread@emory.edu

Low-cost Illumina sequencing of clinically-important bacterial pathogens has generated thousands of publicly available genomic datasets. Analyzing these genomes and extracting relevant information for each pathogen and the associated clinical phenotypes requires not only resources and bioinformatic skills but organism-specific knowledge. In light of these issues, we created Staphopia, an analysis pipeline, database and Application Programming Interface, focused on *Staphylococcus aureus*, a common colonizer of humans and a major antibiotic-resistant pathogen responsible for a wide spectrum of hospital and community-associated infections.

Written in Python, Staphopia's analysis pipeline consists of submodules running open-source tools. It accepts raw FASTQ reads as an input, which undergo quality control filtration, error correction and reduction to a maximum of approximately 100x chromosome coverage. This reduction significantly reduces total runtime without detrimentally affecting the results. The pipeline performs *de novo* assembly-based and mapping-based analysis. Automated gene calling and annotation is performed on the assembled contigs. Read-mapping is used to call variants (single nucleotide polymorphisms and insertion/deletions) against a reference *S. aureus* chromosome (Type strain, N315).

We ran the analysis pipeline on more than 43,000 *S. aureus* shotgun Illumina genome projects in the public ENA database in November 2017. We found that only a quarter of known multi-locus sequence types (STs) were represented but the top ten STs made up 70% of all genomes. MRSA (methicillin resistant *S. aureus*) were 64% of all genomes. Using the Staphopia database we selected 380 high quality genomes deposited with good metadata, each from a different multi-locus sequence type, as a non-redundant diversity set for studying *S. aureus* evolution. In addition to answering basic science questions, Staphopia could serve as a potential platform for rapid clinical diagnostics of *S. aureus* isolates in the future. The system could also be adapted as a template for other organism-specific databases.

Staphylococcus aureus viewed from the perspective of 40,000+ genomes

Robert A. Petit III[1], Timothy D. Read[1]

[1]Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine.

Corresponding Author:
Timothy Read[1]

Email address: tread@emory.edu

Abstract

Low-cost Illumina sequencing of clinically-important bacterial pathogens has generated thousands of publicly available genomic datasets. Analyzing these genomes and extracting relevant information for each pathogen and the associated clinical phenotypes requires not only resources and bioinformatic skills but organism-specific knowledge. In light of these issues, we created Staphopia, an analysis pipeline, database and Application Programming Interface, focused on *Staphylococcus aureus*, a common colonizer of humans and a major antibiotic-resistant pathogen responsible for a wide spectrum of hospital and community-associated infections.

Written in Python, Staphopia's analysis pipeline consists of submodules running open-source tools. It accepts raw FASTQ reads as an input, which undergo quality control filtration, error correction and reduction to a maximum of approximately 100x chromosome coverage. This reduction significantly reduces total runtime without detrimentally affecting the results. The pipeline performs *de novo* assembly-based and mapping-based analysis. Automated gene calling and annotation is performed on the assembled contigs. Read-mapping is used to call variants (single nucleotide polymorphisms and insertion/deletions) against a reference *S. aureus* chromosome (Type strain, N315).

We ran the analysis pipeline on more than 43,000 *S. aureus* shotgun Illumina genome projects in the public ENA database in November 2017. We found that only a quarter of known multi-locus sequence types (STs) were represented but the top ten STs made up 70% of all genomes. MRSA (methicillin resistant *S. aureus*) were 64% of all genomes. Using the Staphopia database we selected 380 high quality genomes deposited with good metadata, each from a different multi-locus sequence type, as a non-redundant diversity set for studying *S. aureus* evolution.
In addition to answering basic science questions, Staphopia could serve as a potential platform for rapid clinical diagnostics of *S. aureus* isolates in the future. The system could also be adapted as a template for other organism-specific databases.

Introduction

Staphylococcus aureus is a common and deadly bacterial pathogen that has been frequently investigated by whole genome sequencing over the last decade. It was the subject of arguably the first large scale bacterial genomic epidemiology study using Illumina sequencing technology (Harris et al., 2010). The cumulative number of Illumina shotgun genome projects deposited in public repositories [the National Center for Biotechnology Information Short Read Archive (NCBI SRA) and the European Nucleotide Archive (ENA)] had grown to almost 50,000 by March 2018 (**Figure 1**). *S. aureus* is therefore on the front edge of a cohort of bacterial species that are acquiring broad whole genome shotgun coverage, offering possibilities of new types of large scale analysis.

S. aureus is a Gram-positive bacterium with a chromosome of ~2.8 Mbp. Plasmid content varies between strains. A multi-locus sequence typing (MLST) scheme that assigns each strain a 'sequence type' (ST) based on seven core genes alleles has proven a robust way of describing individual strain genotypes and membership of larger 'clonal complexes' (CCs) (Planet et al., 2016). The accumulated public *S. aureus* genome datasets present an opportunity for investigating basic questions about how genetic variations that cause antibiotic resistance evolve within populations and how long genes traded by horizontal gene transfer persist in populations. However, there has been a problem of access, as few public tools fill the niche of providing fine scale access to very large datasets from a pathogen species. For example, PATRIC (Wattam et al., 2014) and BIGSdb (Jolley & Maiden, 2010) web based analysis sites focus on high quality annotation and complete genome MLST (cgMLST), respectively, while Aureowiki (Fuchs et al., 2017) and PanX (Ding, Baumdicker & Neher, 2018) provide very detailed information on a smaller number of strains. In this study we describe the creation of Staphopia, an integrated analysis pipeline, database and Application Programming Interface (API) to analyze *S. aureus* genomes.

Materials & Methods

Staphopia Analysis Pipeline

The Staphopia Analysis Pipeline (StAP) processed FASTQ files from a single genome through quality control steps and bioinformatic analysis software. StAP (https://github.com/staphopia/staphopia-ap/) consisted of custom Python3 scripts and open source software organized by the the Nextflow (Di Tommaso et al., 2017) (v0.28.2) workflow management platform (**Figure 2**). When available we used BioConda (Grüning et al., 2017) to install the open source software. Summary statistics of the original input and subsequent downstream results files were collected at each step of the pipeline. For portability, StAP was wrapped in a Docker container. The

version of the pipeline used in this work was Docker Image Tag: 112017 (https://hub.docker.com/r/rpetit3/staphopia/).

The input to StAP was either single or paired end FASTQ file (or files). StAP contained an option that allowed FASTQ data to be pulled from the ENA based on the experiment accession number (ena-dl v0.1, https://github.com/rpetit3/ena-dl). A MD5 hash (md5sum) was generated from the input FASTQ data and cross-referenced against a list generated from processed genomes to prevent reanalysis of the same input. BBduk (Bushnell, 2016) (v37.66) was used to filter out adapters associated with Illumina sequencing and trim reads based on quality. Read errors were corrected using SPAdes (Bankevich et al., 2012) (v3.11.1). Based on the corrected reads, low quality reads were filtered out and the total dataset was subsampled to a maximum of 281 Mbases (100x coverage of the N315 reference chromosome (Kuroda et al., 2001)) with illumina-cleanup (v0.3, https://github.com/rpetit3/illumina-cleanup/). This file (or files, if paired end) we termed “processed FASTQ” or “pFASTQ”.

pFASTQ reads were assembled *de novo* using SPAdes (Bankevich et al., 2012) (v3.11.1). SPAdes also marked assembles as putative plasmids based on evidence such as relative read coverage (Antipov et al., 2016). Summary statistics of the assembly are created using the assembly-summary script (https://github.com/rpetit3/assembly-summary). A BLAST nucleotide database was created from the assembled contigs to be used subsequently for sequence query matching. Open reading frames and their putative functions were predicted and annotated using PROKKA (Seemann, 2014) (v1.12) and its default database.

The *S. aureus* type strain N315 (Kuroda et al., 2001) chromosome (ST5 MRSA; accession NC_002745.2; length 2,814,816 bp) was used as a reference for calling consensus SNPs and indels in the pFASTQ reads using the GATK (McKenna et al., 2010) (v3.8.0) pipeline. GATK pipeline also incorporated BWA (Li & Durbin, 2009) (v0.7.17), SamTools (Li et al., 2009) (v1.6) and PicardTools (v2.14.1, http://broadinstitute.github.io/picard/) software. Identified variants were annotated using the vcf-annotator script (v0.4, https://github.com/rpetit3/vcf-annotator). Jellyfish (Marçais & Kingsford, 2011) (v2.2.6) was used to count k-mers of length 31 base pairs (31-mers) in the pFASTQ file. If the pFASTQ was paired-end, Ariba (Hunt et al., 2017) (v2.10.2) was used to make antibiotic resistance and virulence predictions. Resistance phenotypes were predicted using the MegaRes reference database (Lakin et al., 2017) and virulence using the Virulence Factor Database (Chen et al., 2016) core dataset.

MLST was determined by two or three methods depending on the whether the pFASTQ was paired end. All methods used the *S. aureus* MLST allele sequence database

downloaded from http://saureus.mlst.net/ (November 2017). Alleles for each for each of the seven loci were aligned against the assembled genome using BLAST (Altschul et al., 1990, 1997) (v2.7.1+). Alleles and sequence type (ST) were determined based on perfect matches (100% nucleotide identity with no indels). We also used the MentaLiST (Feijao et al., 2018) (v0.1.3) software to call MLST and cgMLST (complete genome MLST) based on k-mer matching of the alleles to the pFASTQ file. Unlike the BLAST-based MLST method, MentaLiST did not require exact matches to alleles to predict a ST. If the pFASTQ was paired-end, Ariba (Hunt et al., 2017) (v2.10.2) also determined MLST alleles and ST. The default ST call for each genome was determined in the following order: agreement between each method, agreement between MentaLiST and Ariba, agreement between MentaLiST and BLAST, agreement between Ariba and BLAST, Ariba alone without a novel or uncertainty call, MentaLiST alone, and finally BLAST alone.

Evidence for SCC*mec* predictions were based on multiple approaches. The primary approach was to align the primers developed for PCR-based SCCmec typing against the assembled genome using BLAST (Altschul et al., 1990, 1997; Zhang et al., 2005; Chongtrakool et al., 2006; Milheiriço, Oliveira & de Lencastre, 2007; Kondo et al., 2007) (v.2.7.1+). Based on both primer pairs for a given amplicon having a perfect match, an SCCmec type was assigned following the Kondo et. al. algorithm (Kondo et al., 2007). We labelled a genome “MRSA” only if there was at least one match to *mecA* specific primer but no conclusive SCC*mec* assignment. We also aligned proteins associated with SCCmec are also aligned against the assembled genome using TBLASTN and mapped the pFASTQ BWA (Li & Durbin, 2009) (v0.7.17) to to each SCCmec cassette using BWA. The overall cassette and *mec* region coverage statistics were determined as well as the per-base coverage determined for each cassette using genomeCoverageBed (Quinlan & Hall, 2010) (v2.26.0). The methods described above were based on on the 11 SCC*mec* types currently listed in the http://www.sccmec.org (I - XI) and hence did not include recently described types XII and XIII (Wu et al., 2015; Kaya et al., 2018).

Web Application, Relational Database and Application Programming Interface

We used Django (v2.0), a Python web framework, to develop a PostgreSQL (v10.1) backed relational database for storing the results from the analysis pipeline (**Supplemental Figure 1**). A Django application was created for each module of the pipeline, automating the creation of database tables for the results. Python scripts building off Django were developed for insertion of results from each StAP module or the StAP as a whole. A web front-end was developed (staphopia.emory.edu) using the Bootstrap (v4.0) and jQuery (v3.2.1) web frameworks. We used the Django REST framework to develop an extensive application programming interface (API) that allowed

users to create queries accessing multiple samples. We also developed an R package, Staphopia-R (https://github.com/staphopia/staphopia-r), to programmatically access the API. The API and its endpoints were documented to allow users to further develop their own packages in a language of their choice. The source code for our web application was made available at https://github.com/staphopia/staphopia-web/.

Processing Public Data

We used the Cancer Genomics Cloud (CGC) Platform, powered by Seven Bridges (http://www.cancergenomicscloud.org/), to process *S. aureus* genomes through StAP in November 2017. CGC allows users to create custom workflows based on Docker containers, then execute these workflows on the Amazon Web Services (AWS) cloud platform. We obtained a list of publicly available *S. aureus* sequencing projects from the ENA web API using the following search term:

> *"tax_tree(1280) AND library_source=GENOMIC AND (library_strategy=OTHER OR library_strategy=WGS OR library_strategy=WGA) AND (library_selection=MNase OR library_selection=RANDOM OR library_selection=unspecified OR library_selection="size fractionation")".*

CGC opened AWS r3.xlarge instances (30.5GB RAM, 4 processors) that downloaded FASTQ files from the ENA using ena-dl for each genome and ran the StAP pipeline. Results files were returned to the CGC, then uploaded into the Staphopia database server.

Metadata Collection

We used the ENA API to download and store any information linked to the 'Experiment', 'Study', 'Run' and 'BioSample' accessions into the database for each genome. We also determined each sample's publication status using three approaches. #1 using NCBI's Entrez Programming Utilities web API(*Entrez Programming Utilities Help*, 2010), we created a script to identify existing links between SRA, a mirror of ENA, and PubMed. For any links identified, we used the corresponding PubMed ID to extract information corresponding to the publication and stored them in the database. #2 for datasets not linked to a publication in SRA we searched for links in the text of scientific articles. We searched PubMed using the term, "*Staphylococcus aureus"*, limited to the years between and including 2010 (the date of the first publicly available Illumina data upload), and 2017. The saved results, stored as XML, were then loaded into Paperpile, a subscription-based reference management tool, and the corresponding main-text PDFs were automatically downloaded. This process did not include supplementary information files, which required a manual operation. For those articles in which a PDF could not be automatically downloaded, attempts to manually acquire the PDF were

made. Using the text search program 'mdfind', available on Apple OS X, each accession (BioSample, Experiment, Study and Run) in the Staphopia database was used as a separate query to search all the PDF files. Experiment accessions with a corresponding PubMed ID were then stored in the database. In cases where a Study, BioSample or Run accession was identified in PDF text, each associated Experiment accession was linked to the corresponding PubMed ID. #3 a collection of PubMed articles with primary descriptions of *S. aureus* genome sequencing studies was manually curated (https://gist.github.com/plasmid02/48d1fb293c0d394ae650922cdaa62302). For these studies, the PDF and all available supplementary information were downloaded. The process of text-mining the articles and linking Experiment information to PubMed ID was repeated as described for approach #2.

Creating non-redundant *S. aureus* diversity set

Using available metadata, we selected a non-redundant diversity (NRD) set of genomes that were gold quality, linked to a publication and each had a unique ST. When more than one strain from a ST was available, we randomly selected one individual giving priority to samples with collection date, site of isolation and location of isolation fields filled.

Using predicted variants against N315, we extracted a list of genes that had complete sequence coverage (ie "core" genes) but no predicted indels. We extracted the reference gene sequence and created an alternative gene sequence with SNPs predicted in each sample. The alternative gene sequences were split into 31-mers. Presence on these 31-mers in the pFASTQ file were cross-validated using the Jellyfish (Marçais & Kingsford, 2011) tool. These reconstructed gene sequences or all genomes were stored in the database and made available through the API for rapid phylogenetic comparisons.

A set of 31-mer validated genes in which no more than 3 samples contained unvalidated 31-mers were selected for phylogenetic analysis. The set of validated genes were extracted and concatenated into a single sequence for each sample and saved in multi-FASTA and PHYLIP formats. A guide tree was generated with IQ-Tree (Nguyen et al., 2015) (v8.2.11, -fast option) for identification recombination events with ClonalFrameML (Didelot & Wilson, 2015)(v1.11). A recombination free alignment was created with maskrc-svg (https://github.com/kwongj/maskrc-svg). We used IQ-Tree to generate the final maximum likelihood tree with the GTR model and bootstrap support. Bootstrap support was generated from 1000 UFBoot2 (Hoang et al., 2018) (ultrafast bootstrap) replicates. We annotated the tree using iTOL (Letunic & Bork, 2016).

Results

Design of the Staphopia Analysis Pipeline and processing 43,000+ genomes

The Staphopia analysis pipeline (StAP; **Figure 2**) was written to automate processing of individual *S. aureus* genomes from Illumina shotgun data. The pipeline was designed as a series of modules running individual software packages, organized by the Nextflow (Di Tommaso et al., 2017) workflow language, which made it possible to run the entire pipeline or individual components as needed. The first step of the pipeline was to import single- or paired-end FASTQ files either as local files, or from the ENA database. We selected ENA over SRA due to ENA offering direct FASTQ downloads. Following quality-based trimming and down selection of the FASTQ to 281 Mbases (~100x coverage of the N315 reference chromosome (Kuroda et al., 2001), NC_002745.2), analyses were run on the raw processed FASTQ (pFASTQ) files directly, or on *de novo* genome assemblies constructed by the SPAdes program (see Methods for more details). We decided to down sample the input FASTQ files for two reasons: to manage the computational burden when running thousands of genome projects and also to achieve genome datasets with consistently sized pFASTQ input files. The threshold of ~100x coverage was chosen after preliminary studies showed that there was either small or no improvements in outcome for downstream assembly and remapping steps for input files > 100x but large increases in processing time and memory requirement. We created a Postgres database to store results from the StAP analysis and a web front end and a web API for mining the data. An R package (Staphopia-R) was written for interacting with the API and was used for most analysis presented in the results.

In November 2017 there were 44,012 publicly-available shotgun sequencing projects with FASTQ files in ENA. Illumina technology was the dominant platform, accounting for 99% of samples (N=43,972). Eighty-one percent (N=35,580) of them had at least 281 Mbases sequence data. We processed all Illumina genomes through the StAP using cloud servers (please see Methods section). On r3.xlarge instances with 30.5 Gb RAM and 4 processors, the mean time to process a genome was 52 minutes with an interquartile range of 47 to 56 minutes (**Figure 3**).

Sequence and assembly quality trends

We identified samples that were likely not *S. aureus* whole genome shotgun projects and/or were of low technical quality and marked them to not be included in subsequent analysis. We removed genomes that did not have a match to any known allele of the seven MLST loci (232 genomes), had a total assembly size that differed by more than 1Mb from a typical *S. aureus* chromosome (<1.8Mb or >3.8Mb; 764 genomes), or had a GC content differing more than 5% (<28% or > 38%; 467 genomes) of the expected 33% GC content. Failure to complete the StAP pipeline due to poor data quality, and

coverages less than 20x were flagged in 101 and 142 genomes, respectively. In total, we removed 1,023 genome projects, leaving 42,949 for further analysis.

We placed genomes into an arbitrary ranking of 1-3 ("Bronze", "Silver" and "Gold") based on the pFASTQ coverage and average sequencing quality. Paired-end genomes that had read lengths exceeding 100bp, a coverage of 100x and an average per base quality score of at least 30 were given a Gold rank. The purpose of the Gold rank was to group together high-quality samples with near-identical coverage. Paired-end genomes with similar read length and quality cutoffs but a lower sequence coverage (between 50x and 100x) were classified as Silver. The remaining samples were given a rank of Bronze. Single-end reads were classified as Bronze no matter the read length, quality or coverage. More than 70% of the samples were of rank Gold (N=31,014). There were 5,931 Silver and 6,004 Bronze rank samples. Each year since 2012, the number of Gold ranked genomes have exceeded Silver and Bronze (**Figure 4**).

Changes in sequence quality and *de novo* genome assembly metrics over time reflected the development of Illumina technology. Mean per based quality scores increased from ~ 32 in 2010 to > 35 in 2012 and have stayed at that level since. The mean sequence read length rose in steps from < 50 in 2010 to ~ 150 bp in 2017. Assembly metrics such as N50 (Earl et al., 2011), and mean and maximum contig length have gradually increased since 2010. Bronze ranked genome projects had similar (or sometimes even higher) mean per read quality scores than Gold and Silver since 2011. However, Silver and Gold assembly metrics such as N50 and mean contig size were generally quite similar and higher than Bronze.

Genetic diversity measured by MLST

We obtained a view into the genetic diversity of the sequenced *S. aureus* genomes by *in silico* MLST using Ariba (Hunt et al., 2017), MentaLiST (Feijao et al., 2018) (both taking pFASTQ as input, but using different algorithms) and BLASTN against assembled contigs. A sequence type (ST) was assigned to 42,337 (98.6%) genomes. Of these, 41,226 (97.7%) calls were in agreement between MentaLiST, BLAST and (if paired-end) Ariba methods; 828 had agreement between two methods and a no-call on the other, and 189 were supported by one program with no-calls from the other two. Of the remaining 612 genomes not assigned to a known ST, 306 were predicted to be in a novel ST based on matches to known alleles of each of the 7 loci. The remaining 306 genomes had 1-6 known *S. aureus* MLST alleles.

The 42,337 genomes assigned to existing STs represented only 1,090 STs of 4,466 in the saureus.mlst.net database (November 2017). The abundance distribution was

weighted toward common strains, with the top ten sequence types (STs 22, 8, 5, 239, 398, 30, 45, 15, 36, and 105) representing 70% (N=29,851) of the genomes (**Figure 5**).

The cgMLST (complete genome MLST) set of 1861 loci (November 2017) were assigned to the genome set using MentaLiST. There were 38,677 distinct patterns, with only 1,850 patterns found in more than one sample, the remaining 36,827 patterns were represented by a single genome.

Antibiotic resistance genes

Treatment of *S. aureus* infections has been complicated by the evolution of strains resistant to many commonly used antibiotics (Foster, 2017). In particular, methicillin-resistant *S. aureus* (MRSA), carrying the *mecA* gene encoding the PBP2a protein that confers resistance to beta-lactam antibiotics, has become a global problem. We designated a genome as MRSA if each *mecA* typing primer (Kondo et al., 2007) had a perfect BLASTN match on the *de novo* assemblies (26,743 strains), a predicted *mecA* gene ortholog had a BLASTN score ratio of at least 95% (26,430 strains), or the Ariba (Hunt et al., 2017) algorithm predicted reads in the paired-end pFASTQ file matching a *mecA* target in the MegaRes (Lakin et al., 2017) database (27,120 strains). The number of genomes having at least one of these criteria (27,548) was 64% of the total number. Of these, 95% (26,076) of the samples had agreement between each of the three criteria. The top five most common STs had a large portion of MRSA strains (**Figure 6**), which reflects the selection bias of the research community in investigating these significant hospital and community pathogen strains over other *S. aureus*.

The *mecA* gene is usually horizontally acquired as part of a mobile genetic element called “Staphylococcal Cassette Chromosome mec” (SCC*mec*) (Katayama, Ito & Hiramatsu, 2000). SCCmec elements have been classified into at least eleven classes that vary in composition of *mec* genes, *ccr* cassette recombinase genes and spacer regions (http://www.sccmec.org). Knowledge of the SCCmec type can be useful for high-level characterization of MRSA strain types (Kaya et al., 2018). We showed that ten of the eleven cassettes in the current schema map to at least one genome with highest coverage (an approximate method for assigning SCCmec type) (**Table 1**). Of the 26,462 (26,185 paired-end) genomes with at least 50% cassette coverage, 96%, 96% and 99% are MRSA based on primer BLASTN, protein BLASTN or MegaRes, respectively. All type XI cassettes were *mecA* negative by primer BLASTN because these contained the *mecC* allele (García-Álvarez et al., 2011; Shore et al., 2011), which was sufficiently different to be outside the normal distance for a positive match. We found 53 genomes which matched to at least 50% of a SCCmec cassette but were not MRSA and had no reads mapping to the *mec* region of the cassette.

In addition to *mecA*, we found numerous other classes of non-core genes using the MegaRes (Lakin et al., 2017) class designations (**Table 2**). We did not consider SNPs/indels in core genes associated with resistance for this analysis. The most common class of resistance genes were beta-lactamases found in 37,758 genomes. Following this, the most common were the genes putatively conferring fosfomycin, macrolide-lincosamide-streptogramin (MLS), and aminoglycosides resistance (24,205, 22,322, 17,968 genomes respectively). As with MRSA, the other common resistance genes were not distributed evenly among the top ST groups (**Figure 7**), reflecting sampling ascertainment bias and also possibly differences in geographic distribution and prevalence of healthcare-isolated strains in the most common genotypes.

Publication, metadata and strain geographic distribution

One challenge to using publicly available datasets through ENA or SRA is determining whether there is a published article describing the sequenced genome. We found through NCBI's Entrez Tools (eLink) that 6,712 genomes were linked to 48 publications in PubMed (March 2018). We attempted to add to the number by using text-mining methods to find *S. aureus* accession numbers in PDFs of *S. aureus* genome publications, ascertaining an additional 5,209 genomes in 30 publications. Therefore, of the 42,949 samples deposited between 2010 and 2017, only 28% (N=11,921) could be linked to a publication (**Figure 8**). Since many genomes have been deposited in the last 1-3 years, this reflected the often significant time lag between depositing sequence data and final publication.

We noted that collection of metadata from public sequencing projects was another challenge. When submitting genome sequences to databases only a limited number of metadata fields are required, leading to the bulk of the information needing to be extracted manually from a publication, if it can be found. Only 40% (N=17,034) genomes had a collection date, 35% (N=14,983) had a geographic location and 35% (N=14,768) had isolate source metadata. Using the available geographic data to geocode the sites of collection, we found that strains were from five continents and at least 40 countries. There was a strong bias toward strains from Europe (N=7,314) and North America (N=5,882), reflecting where the funding for most of the early sequencing studies had originated.

A non-redundant *S. aureus* diversity set

The number of SNPs compared to the N315 reference strain varied from 6 to 141,893 within our collection of 42,949 genomes. The stepped pattern of the distribution (**Figure 9**) reflected the organization of *S. aureus* into clonal complexes. Apart from CC5 strains closely related to N315, the majority of *S. aureus* had ~50-50,000 SNPs and ~500-1500

indels called by the GATK pipeline (McKenna et al., 2010). There were a group of 240 most distant strains with > 55,000 SNP (**Figure 9**) that were found to be closer to the sister species, *S. argenteus* (Holt et al., 2011) based on ANI imputed by mash (Ondov et al., 2016), although 230 of these were assigned a *S. aureus* ST.

Of the 6,904 *S. aureus* genomes of Gold rank linked to a publication we selected a group of 380 each having a distinct ST as a non-redundant diversity (NRD) set of genomes. Of the 2,756 annotated N315 genes (excluding RNAs), 1,113 genes had no indels when reads from each genome in the NRD dataset were mapped. Of these, 838 were “core” genes found in every genome. We reconstructed these genes for each of the NRD genomes starting with the N315 sequence and substituting predicted SNPs. These predicted sequences were then validated by decomposing into 31-mers and cross-checking whether each k-mer was present in pFASTQ files processed by Jellyfish (Marçais & Kingsford, 2011). We concatenated the 838 genes for each member of the NRD set and created a tree based on the 60,191 variant SNP positions (**Figure 10**). The structure of the unrooted species tree resembles previous *S. aureus* phylogenies (Planet et al., 2016).

Discussion

The huge public library of genome sequence projects of *S. aureus* and other pathogens are a resource for microbiologists for testing genetic hypotheses in silico. Unfortunately, this has been a library of blank covers: most projects cannot be browsed to identify features such as ST, key SNPs and non-core genes. Staphopia makes the library searchable for a number of important attributes, and we have described example workflows in the results section.

We used three strategies for analysis of raw sequence data: mapping reads to a reference chromosome to identify variants; *de novo* genome assembly, and direct analysis of the reads. Each has its strengths and weaknesses. Reference mapping retains quality information about variant calls but is limited to regions of the core genome and accuracy is reduced as genetic distance increases between the query and the reference. *De novo* assembly allows for discovery of novel accessory genes and is reference independent but could be affected by genomic contamination and with Illumina short read data, and small portions of the sequence could be lost in gaps between contigs. Direct analysis of reads based on k-mer decomposition approaches allows examination of sequence independent of mapping and assembly algorithms but are susceptible to false results arising from contamination and random sequence error. Using different approaches to cross-validate wherever possible builds confidence and we showed that MLST and MRSA/MSSA identification were robust with different underlying data types collected.

There are many possible avenues for future extensions of the project. New tools for efficient direct querying of raw reads have recently become available (e.g BigSI (Bradley et al., 2017), and mash (Ondov et al., 2016)) and we plan to incorporate them in future iterations of the pipeline. Some of the principal improvements need to be in protein functional annotation. For speed and simplicity, we elected to map genes called from *de novo* assemblies against the included PROKKA (Seemann, 2014) RefSeq database. This has the advantage of giving consistent proteins naming that can be linked to many functional annotation databases through UniProt cross-references. However, for fine resolution studies of sets of genomes from Staphopia, we recommend reprocessing with ROARY (Page et al., 2015) to incorporate paralog detection and to use more extensive databases for homology matching. Even then, specific modules would need to be incorporated to improve naming of intrinsically hard to annotate protein families (e.g MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) (Foster et al., 2014)).

A key problem highlighted in this study is the difficulty in tracing publications linked to public genome data and finding typical metadata on strains (date and place of isolation, body site). We were able here to link thousands of records to publications through searching text in PDFs. For this reason, we urge researchers publishing microbial genomes to quote the project id (ie the PRJN ID) of publically submitted data in the full text of the publication. Extracting metadata from publications to link from strains was much more manual. We believe that journals need to start to enforce machine readable standards for metadata associated with deposited strains. The routine usage of BioSample id (https://www.ncbi.nlm.nih.gov/books/NBK169436/), which links strains to genomic information, would be a major step forward.

Staphopia was designed with Illumina shotgun data in mind but increased use of alternative sequencing technologies in the future may necessitate new development. “Long read” technologies (e.g. PacBio, Oxford Nanopore) tend to have assemblies with fewer gaps, higher per base errors and lower coverage. A “gold standard” PacBio assembly will have a different quality profile to Illumina technology data (which itself is also evolving). Another challenge for automated assembly of public data will be to identify projects sequenced with multiple technologies and assembled as hybrids (e.g. as demonstrated by the Unicycler tool (Wick et al., 2017)). To do this would mean altering the pipeline to perform hybrid assembly when experiments with multiple technologies are associated with a strain. Currently, within ENA (and SRA) a BioSample can be associated with multiple Experiments, but an Experiment can only be associated with a single BioSample. When a BioSample was linked to more than one Experiment, it was difficult to determine in an automated way if it is actually the same genomic DNA

input to multiple experiments or, in rare cases, a mistaken assignment of a set of genetically non-identical isolates with the BioSample (e.g. all isolates from a study given the common strain name “USA300”). Because of this, Staphopia treated each ENA Experiment as a unique sample, rather than the BioSample.

It is unclear at this time whether the approach of processing of every public dataset will be sustainable as sequencing data production grows in the future. It would only be possible if storage and processing costs fall faster than the accumulation of new data, and multi-genome database queries may still be prohibitively slow. An alternative strategy to processing all strains, would be to filter the isolates for redundancy, by removing isolates that are less than *n* SNPs from any member of a canonical genome set. However, there is still information in deep sequencing studies that can be captured from distributions of reads and kmer distribution, even if the consensus sequences of the strains are identical. Plasmid copy number may differ between clones grown under different conditions and the distribution of reads across the genome can itself be used to infer relative growth rate (Brown et al., 2016). No two shotgun genome sequencing projects are identical, and all have some potential value, especially if they have strong supporting metadata.

Conclusions

- We analyzed 43,972 *S. aureus* public Illumina genome projects using the newly developed “Staphopia” analysis pipeline and database. 42,949 genomes were retained for subsequent analysis after filtering against low quality
- The data quality was high overall: 36,945 (86%) were from paired end projects with greater than 50-fold coverage and 35 average quality (“Gold” and “Silver” quality)
- There has been a great concentration of effort on a sequencing a small number of sequence types: only 1,090 STs of 4,466 previously collected STs were recovered and 10 STs make up 70% of all genomes.
- 26,050 to 27,548 genomes were predicted MRSA depending on the criteria used for classification.
- We could link only 28% of the genomes to a PubMed referenced publication.
- We identified 380 non-redundant highly quality published genomes as a reference subset for diversity within the species.
- We identified 838 cores genes that can be reliably used for rapid tree building based on SNPs compared to the reference N315 genome.

Funding

Funding was from Emory University, Amazon AWS in Education Grant Program, and NIH grants AI091827 and AI121860. The Seven Bridges NCI Cancer Genomics Cloud

pilot was supported in part by the funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.

Acknowledgements

We would like to thank Tauqeer Alam, Jim Hogan, Santiago Castillo-Ramírez. Michelle Su, Michael Frisch and Erik Lehnert for their helpful suggestions. We would also like to acknowledge our gratitude to the many scientists and their funders who provided genome sequences to the public domain, ENA and SRA for storing and organizing the data, and the authors of the open source software tools and databases used in this work.

Links

Code for most analysis described in the results section - https://github.com/staphopia/staphopia-paper
R Package - https://github.com/staphopia/staphopia-r
StAP - https://github.com/staphopia/staphopia-ap
Web Package - https://github.com/staphopia/staphopia-web
Docker Image - https://hub.docker.com/r/rpetit3/staphopia/

References

Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410.

Altschul SF., Madden TL., Schäffer AA., Zhang J., Zhang Z., Miller W., Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25:3389–3402.

Antipov D., Hartwick N., Shen M., Raiko M., Lapidus A., Pevzner P. 2016. plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data. *bioRxiv*:048942. DOI: 10.1101/048942.

Bankevich A., Nurk S., Antipov D., Gurevich AA., Dvorkin M., Kulikov AS., Lesin VM., Nikolenko SI., Pham S., Prjibelski AD., Pyshkin AV., Sirotkin AV., Vyahhi N., Tesler G., Alekseyev MA., Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology: a journal of computational molecular cell biology* 19:455–477.

Bradley P., den Bakker H., Rocha E., McVean G., Iqbal Z. 2017. Real-time search of all bacterial and viral genomic data. *bioRxiv*:234955. DOI: 10.1101/234955.

Brown CT., Olm MR., Thomas BC., Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nature biotechnology* 34:1256–1263.

Bushnell B. 2016. BBMap short read aligner. *University of California, Berkeley, California. URL http://sourceforge. net/projects/bbmap*.

Chen L., Zheng D., Liu B., Yang J., Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic acids research* 44:D694–7.

Chongtrakool P., Ito T., Ma XX., Kondo Y., Trakulsomboon S., Tiensasitorn C.,

562 Jamklang M., Chavalit T., Song J-H., Hiramatsu K. 2006. Staphylococcal cassette
563 chromosome mec (SCCmec) typing of methicillin-resistant Staphylococcus aureus
564 strains isolated in 11 Asian countries: a proposal for a new nomenclature for
565 SCCmec elements. *Antimicrobial agents and chemotherapy* 50:1001–1012.
566 Didelot X., Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in
567 whole bacterial genomes. *PLoS computational biology* 11:e1004041.
568 Ding W., Baumdicker F., Neher RA. 2018. panX: pan-genome analysis and exploration.
569 *Nucleic acids research* 46:e5.
570 Di Tommaso P., Chatzou M., Floden EW., Barja PP., Palumbo E., Notredame C. 2017.
571 Nextflow enables reproducible computational workflows. *Nature biotechnology*
572 35:316–319.
573 Earl D., Bradnam K., St John J., Darling A., Lin D., Fass J., Yu HOK., Buffalo V.,
574 Zerbino DR., Diekhans M., Nguyen N., Ariyaratne PN., Sung W-K., Ning Z., Haimel
575 M., Simpson JT., Fonseca NA., Birol İ., Docking TR., Ho IY., Rokhsar DS., Chikhi
576 R., Lavenier D., Chapuis G., Naquin D., Maillet N., Schatz MC., Kelley DR.,
577 Phillippy AM., Koren S., Yang S-P., Wu W., Chou W-C., Srivastava A., Shaw TI.,
578 Ruby JG., Skewes-Cox P., Betegon M., Dimon MT., Solovyev V., Seledtsov I.,
579 Kosarev P., Vorobyev D., Ramirez-Gonzalez R., Leggett R., MacLean D., Xia F.,
580 Luo R., Li Z., Xie Y., Liu B., Gnerre S., MacCallum I., Przybylski D., Ribeiro FJ., Yin
581 S., Sharpe T., Hall G., Kersey PJ., Durbin R., Jackman SD., Chapman JA., Huang
582 X., DeRisi JL., Caccamo M., Li Y., Jaffe DB., Green RE., Haussler D., Korf I., Paten
583 B. 2011. Assemblathon 1: a competitive assessment of de novo short read
584 assembly methods. *Genome research* 21:2224–2241.

585 *Entrez Programming Utilities Help* 2010. National Center for Biotechnology Information
586 (US).
587 Feijao P., Yao H-T., Fornika D., Gardy J., Hsiao W., Chauve C., Chindelevitch L. 2018.
588 MentaLiST - A fast MLST caller for large MLST schemes. *Microbial genomics*. DOI:
589 10.1099/mgen.0.000146.
590 Foster TJ. 2017. Antibiotic resistance in Staphylococcus aureus. Current status and
591 future prospects. *FEMS microbiology reviews*. DOI: 10.1093/femsre/fux007.
592 Foster TJ., Geoghegan JA., Ganesh VK., Höök M. 2014. Adhesion, invasion and
593 evasion: the many functions of the surface proteins of Staphylococcus aureus.
594 *Nature reviews. Microbiology* 12:49–62.
595 Fuchs S., Mehlan H., Bernhardt J., Hennig A., Michalik S., Surmann K., Pané-Farré J.,
596 Giese A., Weiss S., Backert L., Herbig A., Nieselt K., Hecker M., Völker U., Mäder
597 U. 2017. AureoWiki The repository of the Staphylococcus aureus research and
598 annotation community. *International journal of medical microbiology: IJMM*. DOI:
599 10.1016/j.ijmm.2017.11.011.
600 García-Álvarez L., Holden MTG., Lindsay H., Webb CR., Brown DFJ., Curran MD.,
601 Walpole E., Brooks K., Pickard DJ., Teale C., Parkhill J., Bentley SD., Edwards
602 GF., Girvan EK., Kearns AM., Pichon B., Hill RLR., Larsen AR., Skov RL., Peacock
603 SJ., Maskell DJ., Holmes MA. 2011. Meticillin-resistant Staphylococcus aureus with
604 a novel mecA homologue in human and bovine populations in the UK and
605 Denmark: a descriptive study. *The Lancet infectious diseases* 11:595–603.
606 Grüning B., Dale R., Sjödin A., Rowe J., Chapman BA., Tomkins-Tinch CH., Valieris R.,
607 The Bioconda Team., Köster J. 2017. Bioconda: A sustainable and comprehensive

software distribution for the life sciences. *bioRxiv*:207092. DOI: 10.1101/207092.

Harris SR., Feil EJ., Holden MTG., Quail MA., Nickerson EK., Chantratita N., Gardete S., Tavares A., Day N., Lindsay JA., Edgeworth JD., de Lencastre H., Parkhill J., Peacock SJ., Bentley SD. 2010. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* 327:469–474.

Hoang DT., Chernomor O., von Haeseler A., Minh BQ., Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution* 35:518–522.

Holt DC., Holden MTG., Tong SYC., Castillo-Ramirez S., Clarke L., Quail MA., Currie BJ., Parkhill J., Bentley SD., Feil EJ., Giffard PM. 2011. A Very Early-Branching Staphylococcus aureus Lineage Lacking the Carotenoid Pigment Staphyloxanthin. *Genome biology and evolution* 3:881–895.

Hunt M., Mather AE., Sánchez-Busó L., Page AJ., Parkhill J., Keane JA., Harris SR. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics* 3:e000131.

Jolley KA., Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics* 11:595.

Katayama Y., Ito T., Hiramatsu K. 2000. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in Staphylococcus aureus. *Antimicrobial agents and chemotherapy* 44:1549–1555.

Kaya H., Hasman H., Larsen J., Stegger M., Johannesen TB., Allesøe RL., Lemvigh CK., Aarestrup FM., Lund O., Larsen AR. 2018. SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette ChromosomemecinStaphylococcus

aureusUsing Whole-Genome Sequence Data. *mSphere* 3. DOI: 10.1128/mSphere.00612-17.

Kondo Y., Ito T., Ma XX., Watanabe S., Kreiswirth BN., Etienne J., Hiramatsu K. 2007. Combination of multiplex PCRs for staphylococcal cassette chromosome mec type assignment: rapid identification system for mec, ccr, and major differences in junkyard regions. *Antimicrobial agents and chemotherapy* 51:264–274.

Kuroda M., Ohta T., Uchiyama I., Baba T., Yuzawa H., Kobayashi I., Cui L., Oguchi A., Aoki K., Nagai Y., Lian J., Ito T., Kanamori M., Matsumaru H., Maruyama A., Murakami H., Hosoyama A., Mizutani-Ui Y., Takahashi NK., Sawano T., Inoue R., Kaito C., Sekimizu K., Hirakawa H., Kuhara S., Goto S., Yabuzaki J., Kanehisa M., Yamashita A., Oshima K., Furuya K., Yoshino C., Shiba T., Hattori M., Ogasawara N., Hayashi H., Hiramatsu K. 2001. Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *The Lancet* 357:1225–1240.

Lakin SM., Dean C., Noyes NR., Dettenwanger A., Ross AS., Doster E., Rovira P., Abdo Z., Jones KL., Ruiz J., Belk KE., Morley PS., Boucher C. 2017. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research* 45:D574–D580.

Letunic I., Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* 44:W242–5.

Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The Sequence

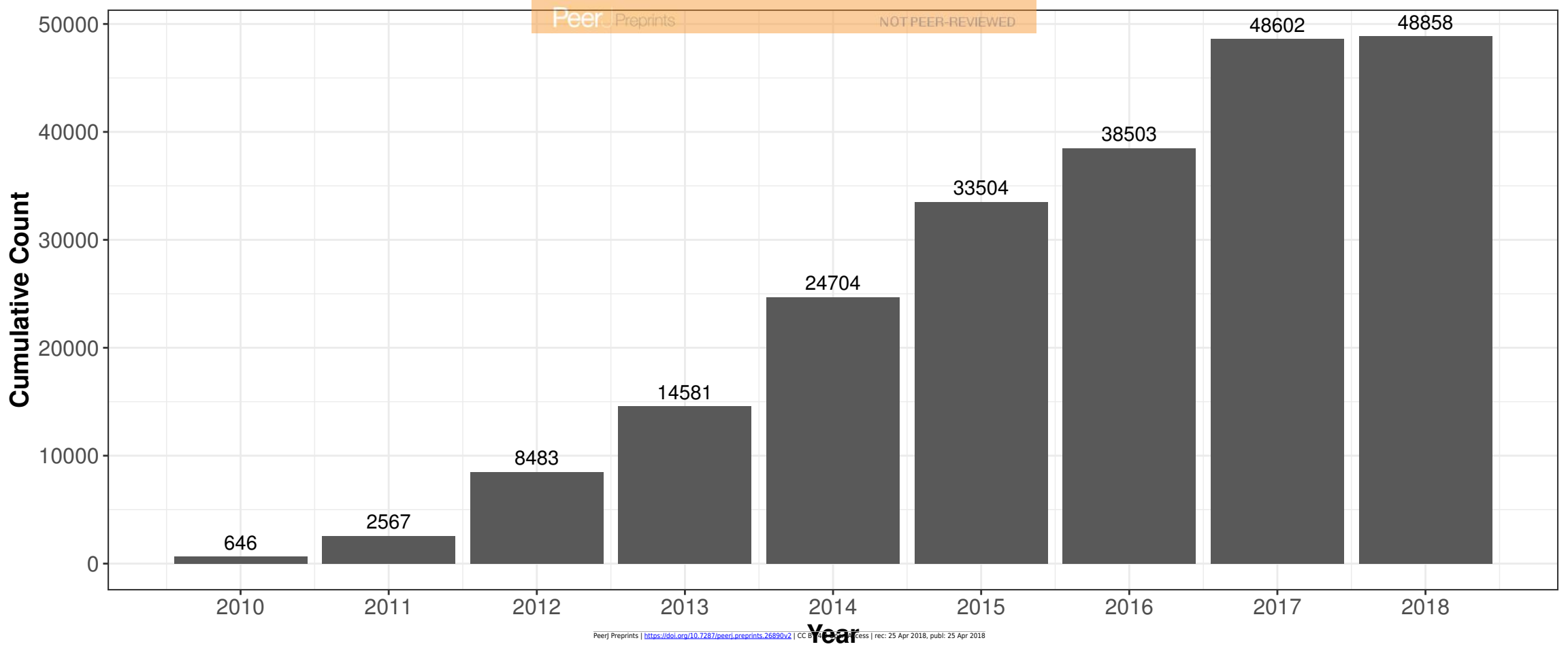
654 Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
655 Marçais G., Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting
656 of occurrences of k-mers. *Bioinformatics* 27:764–770.
657 McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella
658 K., Altshuler D., Gabriel S., Daly M., DePristo MA. 2010. The Genome Analysis
659 Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
660 data. *Genome research* 20:1297–1303.
661 Milheiriço C., Oliveira DC., de Lencastre H. 2007. Multiplex PCR strategy for subtyping
662 the staphylococcal cassette chromosome mec type IV in methicillin-resistant
663 Staphylococcus aureus:“SCC mec IV multiplex.” *The Journal of antimicrobial*
664 *chemotherapy* 60:42–48.
665 Nguyen L-T., Schmidt HA., von Haeseler A., Minh BQ. 2015. IQ-TREE: a fast and
666 effective stochastic algorithm for estimating maximum-likelihood phylogenies.
667 *Molecular biology and evolution* 32:268–274.
668 Ondov BD., Treangen TJ., Melsted P., Mallonee AB., Bergman NH., Koren S., Phillippy
669 AM. 2016. Mash: fast genome and metagenome distance estimation using
670 MinHash. *Genome biology* 17:132.
671 Page AJ., Cummins CA., Hunt M., Wong VK., Reuter S., Holden MTG., Fookes M.,
672 Falush D., Keane JA., Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan
673 genome analysis. *Bioinformatics* . DOI: 10.1093/bioinformatics/btv421.
674 Planet PJ., Narechania A., Chen L., Mathema B., Boundy S., Archer G., Kreiswirth B.
675 2016. Architecture of a Species: Phylogenomics of Staphylococcus aureus. *Trends*
676 *in microbiology*. DOI: 10.1016/j.tim.2016.09.009.

677 Quinlan AR., Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
678 features. *Bioinformatics* 26:841–842.
679 Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*
680 30:2068–2069.
681 Shore AC., Deasy EC., Slickers P., Brennan G., O'Connell B., Monecke S., Ehricht R.,
682 Coleman DC. 2011. Detection of staphylococcal cassette chromosome mec type XI
683 carrying highly divergent mecA, mecI, mecR1, blaZ, and ccr genes in human
684 clinical isolates of clonal complex 130 methicillin-resistant Staphylococcus aureus.
685 *Antimicrobial agents and chemotherapy* 55:3765–3773.
686 Wattam AR., Abraham D., Dalay O., Disz TL., Driscoll T., Gabbard JL., Gillespie JJ.,
687 Gough R., Hix D., Kenyon R., Machi D., Mao C., Nordberg EK., Olson R.,
688 Overbeek R., Pusch GD., Shukla M., Schulman J., Stevens RL., Sullivan DE.,
689 Vonstein V., Warren A., Will R., Wilson MJC., Yoo HS., Zhang C., Zhang Y., Sobral
690 BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource.
691 *Nucleic acids research* 42:D581–91.
692 Wick RR., Judd LM., Gorrie CL., Holt KE. 2017. Unicycler: Resolving bacterial genome
693 assemblies from short and long sequencing reads. *PLoS computational biology*
694 13:e1005595.
695 Wu Z., Li F., Liu D., Xue H., Zhao X. 2015. Novel Type XII Staphylococcal Cassette
696 Chromosome mec Harboring a New Cassette Chromosome Recombinase, CcrC2.
697 *Antimicrobial agents and chemotherapy* 59:7597–7601.
698 Zhang K., McClure J-A., Elsayed S., Louie T., Conly JM. 2005. Novel multiplex PCR
699 assay for characterization and concomitant subtyping of staphylococcal cassette

chromosome mec types I to V in methicillin-resistant Staphylococcus aureus. *Journal of clinical microbiology* 43:5026–5033.

Figure 1(on next page)

Figure 1. Cumulative submissions of *Staphylococcus aureus* genome projects 2010 - March 2018.

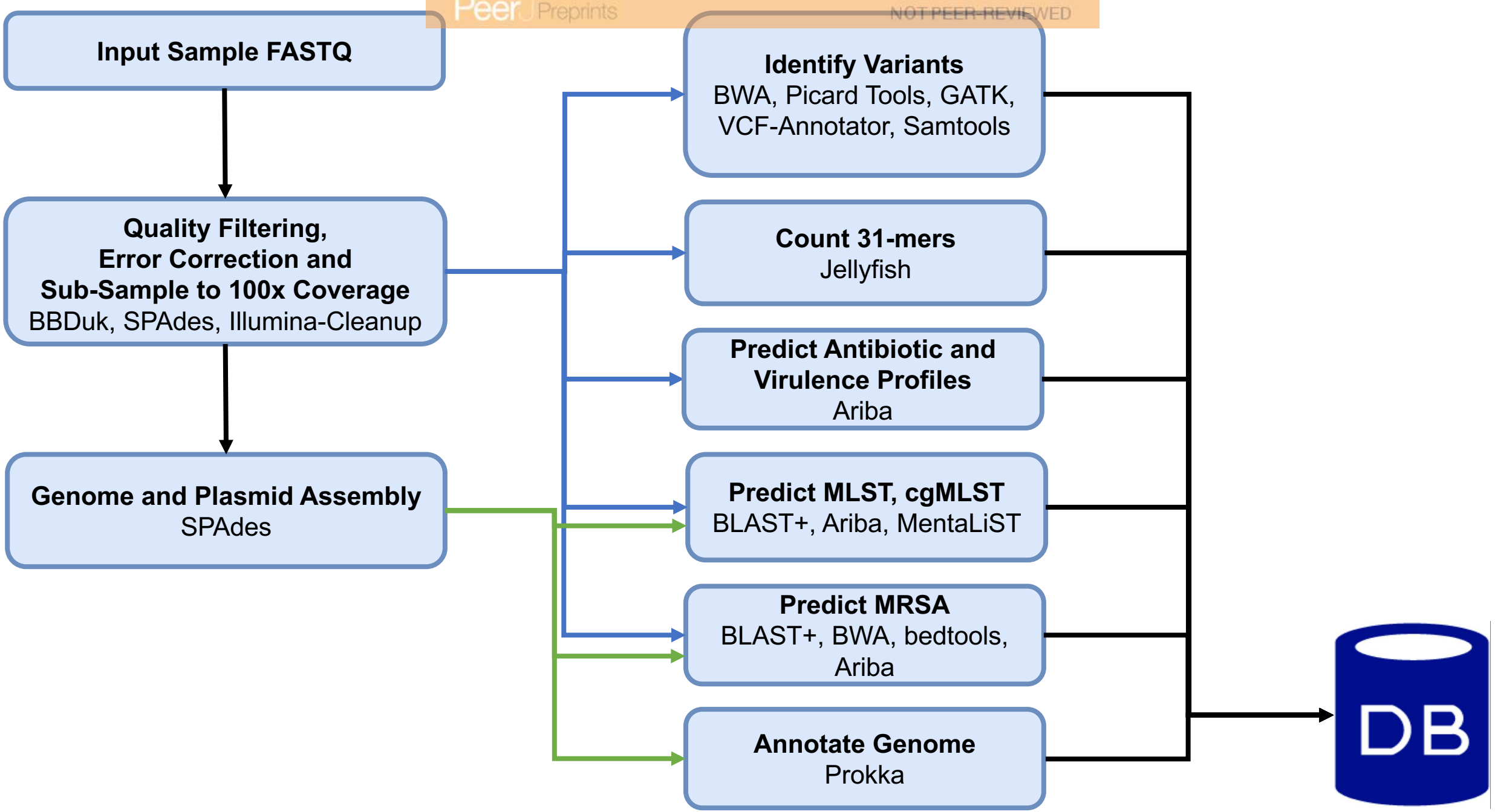


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26890v2 | CC BY 4.0 Open Access | rec: 25 Apr 2018, publ: 25 Apr 2018

Figure 2(on next page)

Figure 2. Staphopia Analysis Pipeline (StAP) Workflow.

The diagram describes basic operations of the pipeline on a single genome input (FASTQ file) before uploading into the Postgres relational database. Details of the programs used are in the methods and https://github.com/staphopia/staphopia-ap . Green arrows indicate input from *de novo* assembled contigs, blue arrows were operations performed on pFASTQ files.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26890v2 | CC BY 4.0 Open Access | rec: 25 Apr 2018, publ: 25 Apr 2018

Figure 3(on next page)

Figure 3. StAP run time using Cancer Genomics Cloud (CGC) platform.

Overall run time statistics were available for 31,587 of the completed CGC jobs. Mean run time was 51 minutes (median 52 minutes). There were 983 jobs that took more than 80 minutes to complete.

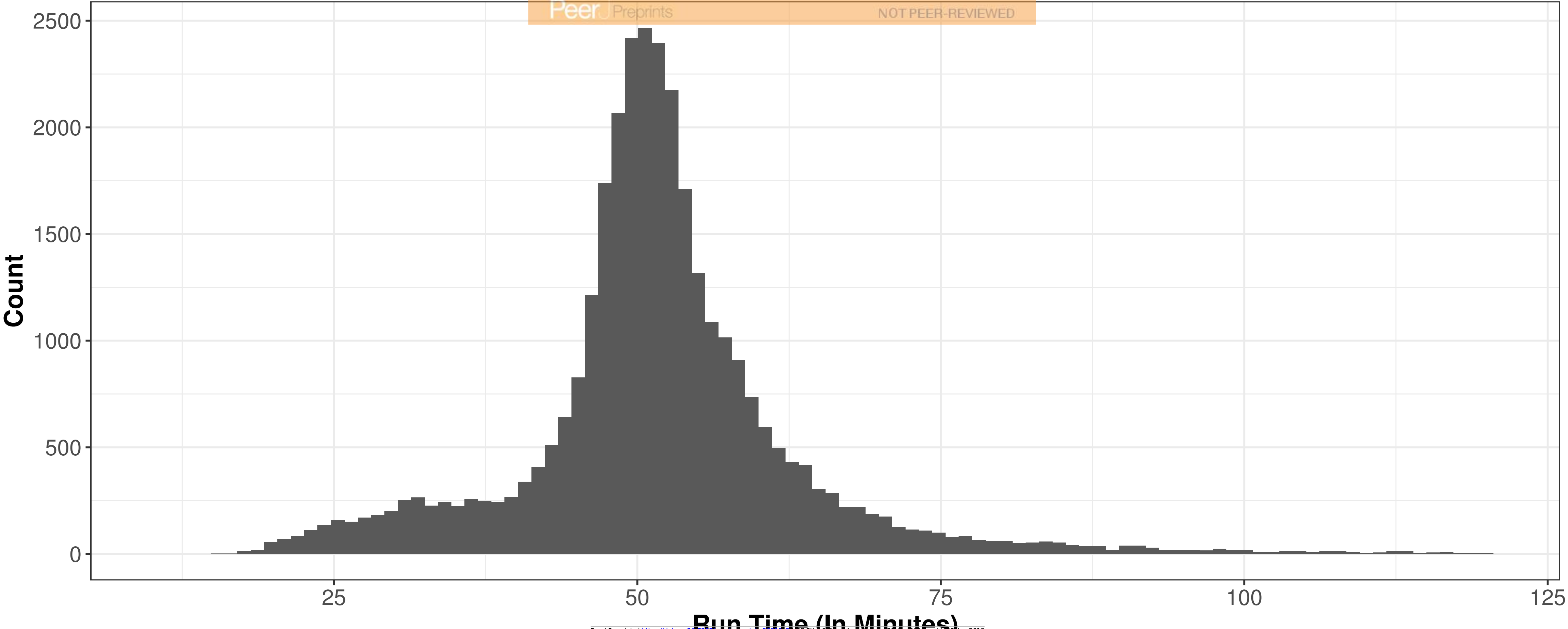
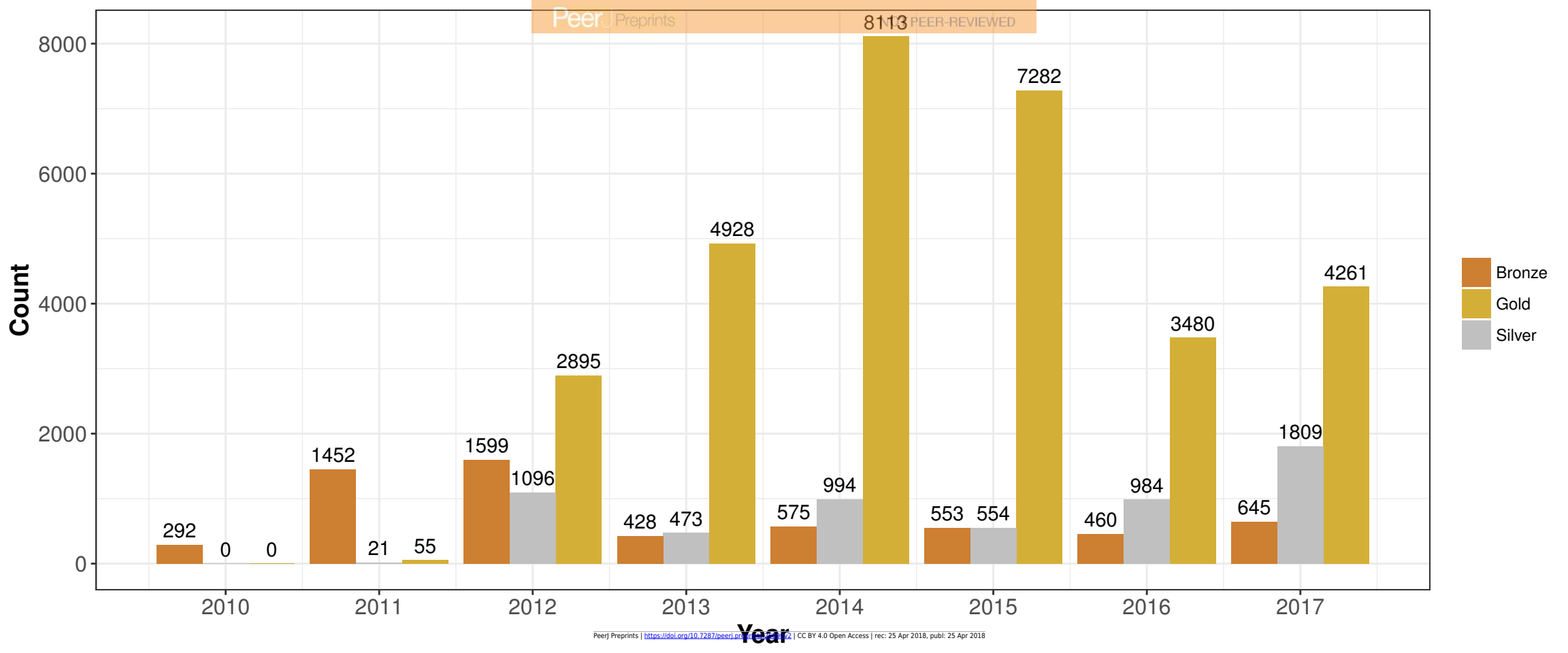


Figure 4(on next page)

Figure 4. Sequencing quality ranks per year 2010-2017.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26872v2 | CC BY 4.0 Open Access | rec: 25 Apr 2018, publ: 25 Apr 2018

Figure 5(on next page)

Figure 5. Top ten STs.

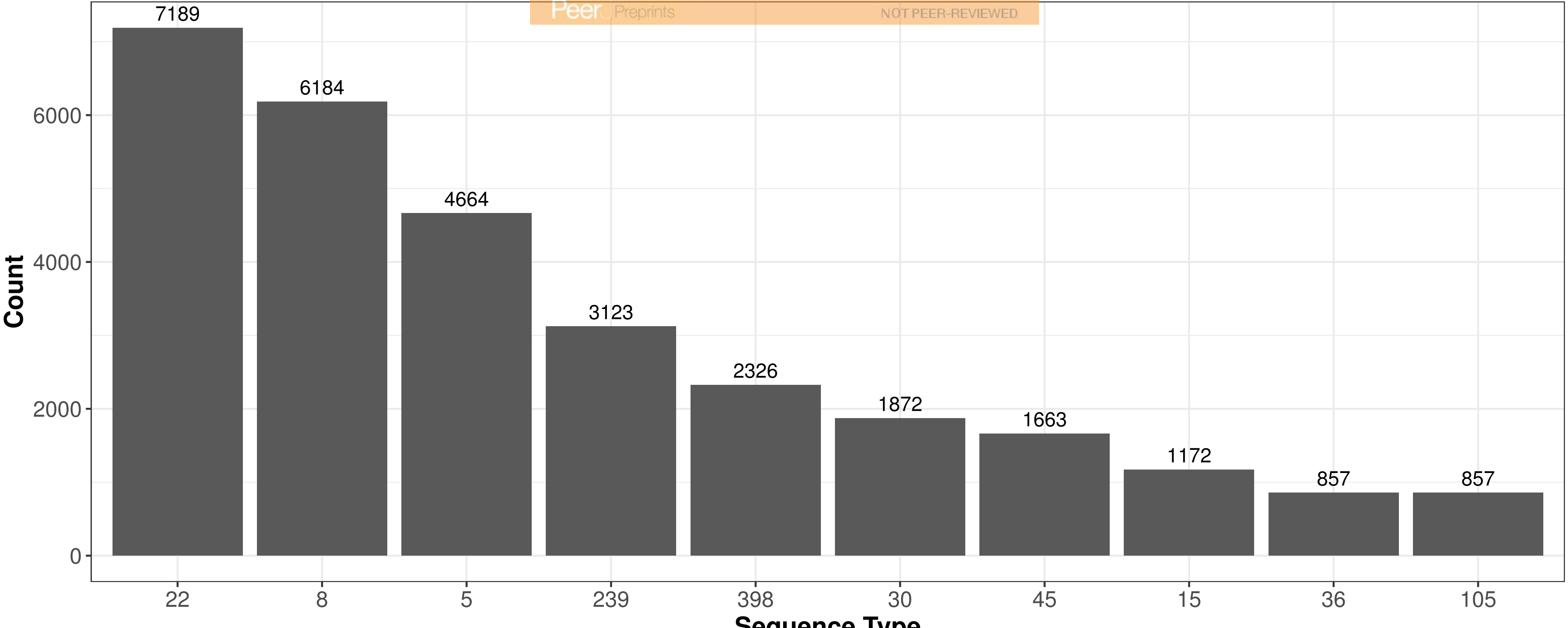


Figure 6(on next page)

Figure 6. Breakdown of predicted MRSA and MSSA genomes in the top ten STs

MRSA was predicted based with Ariba (Hunt et al., 2017) using the MegaRes (Lakin et al., 2017) database.

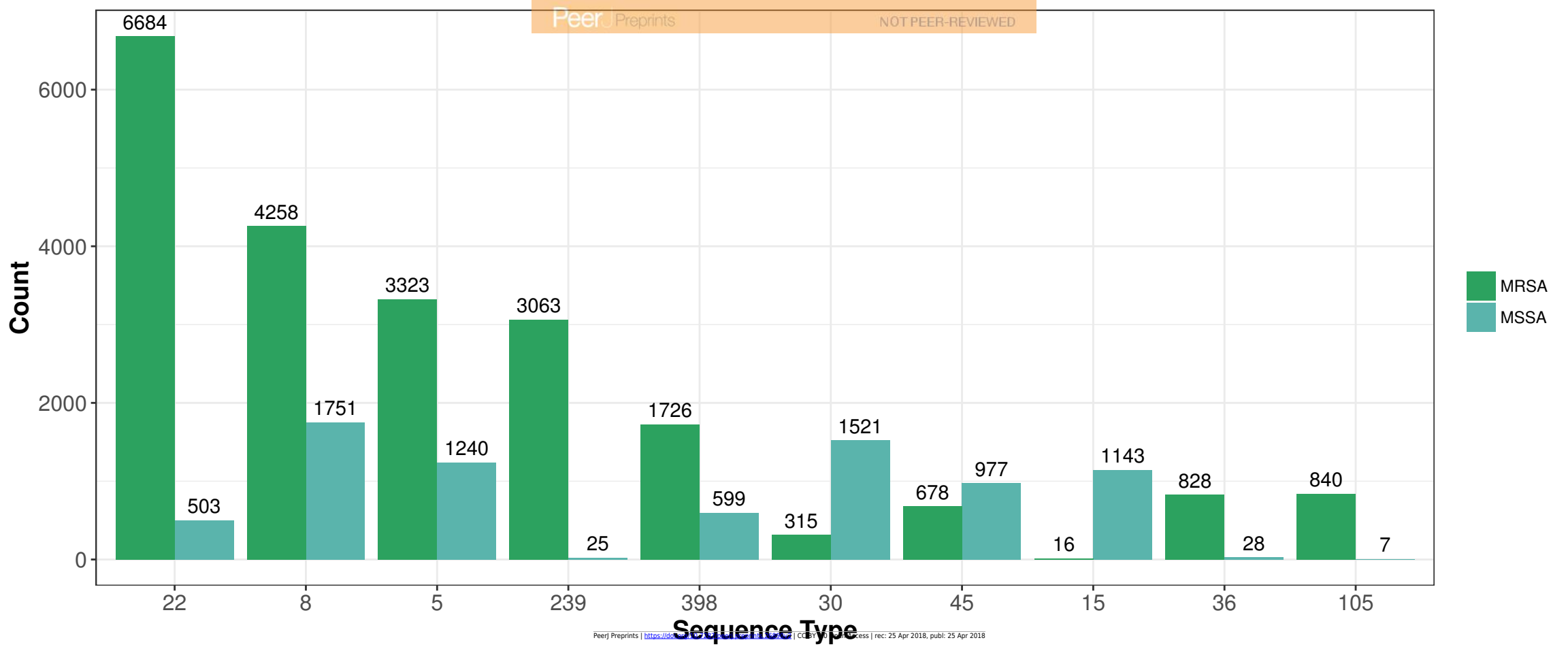
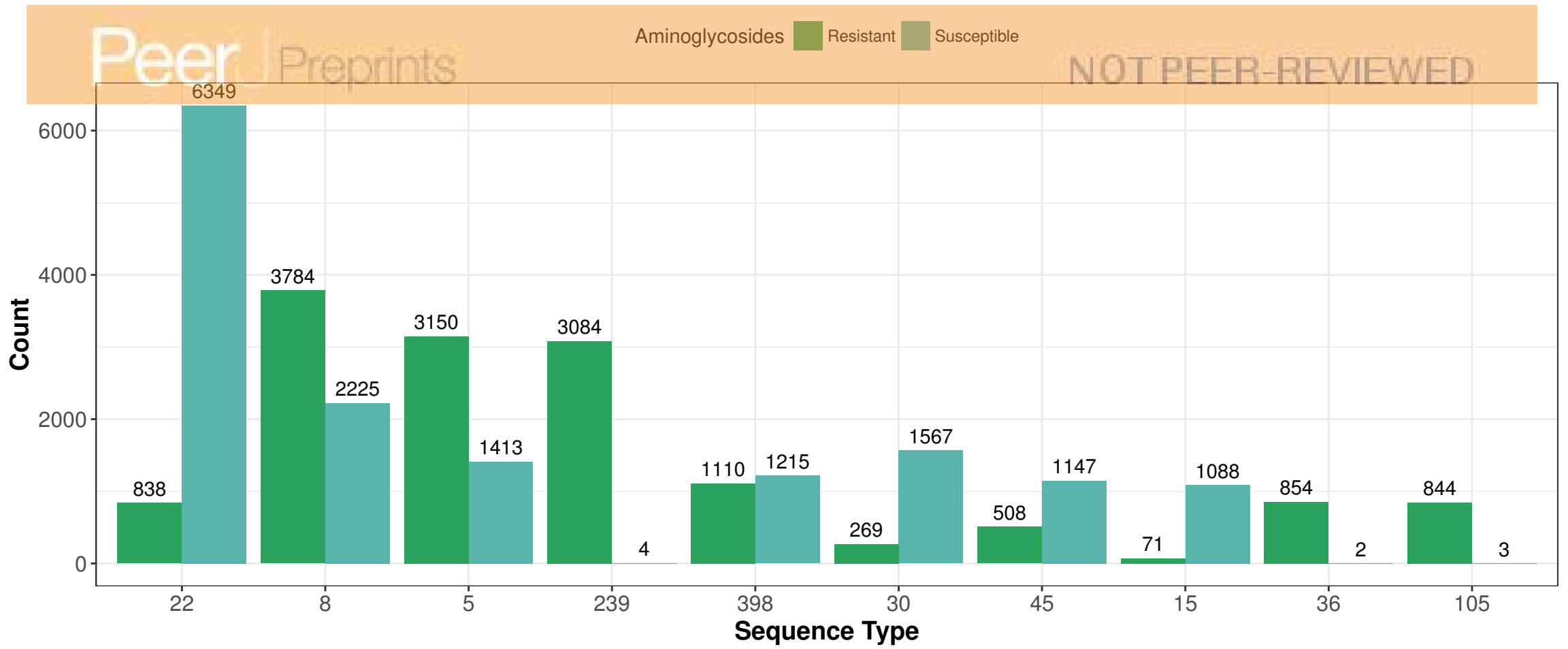
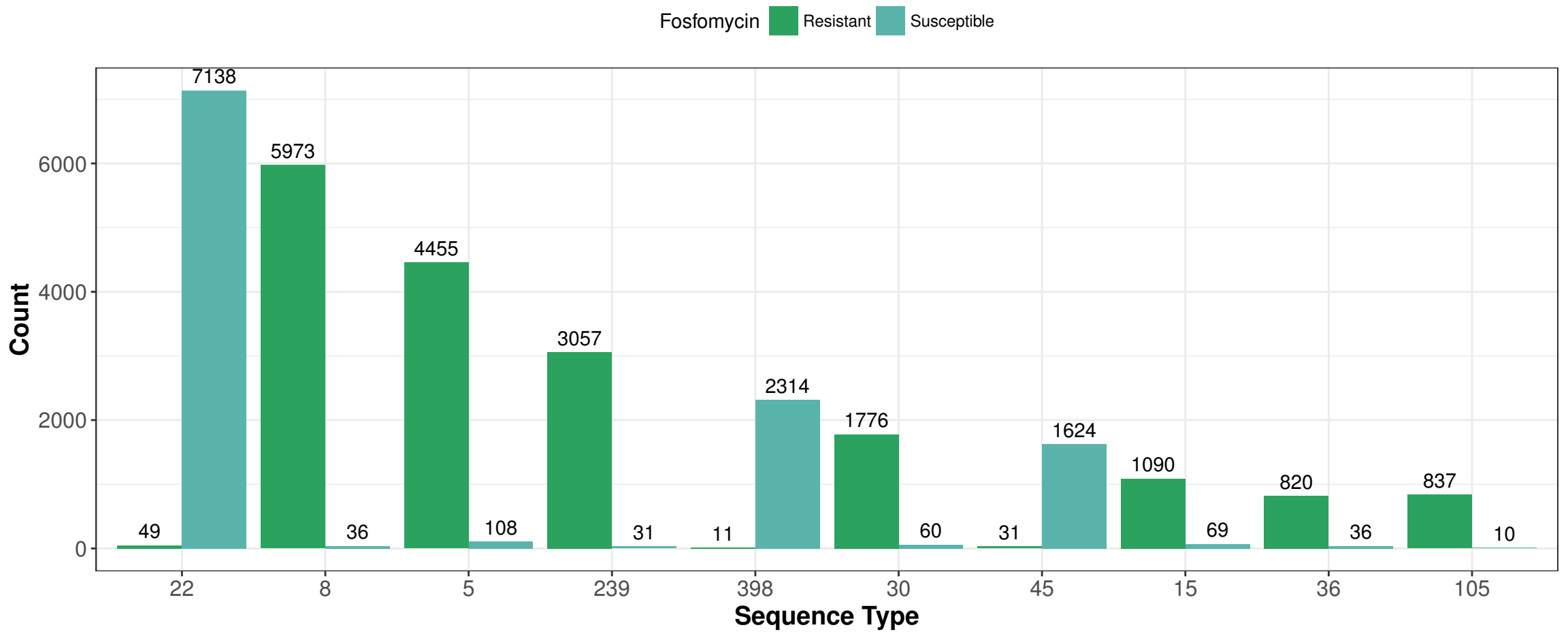


Figure 7 (on next page)

Figure 7. Resistance genes to aminoglycoside, fosfomycin, and macrolide-lincosamide-streptogramin (MLS) antibiotic in the top 10 STs.

The presence of resistance genes was predicted by Ariba (Hunt et al., 2017) using the reference MegaRes (Lakin et al., 2017) database. Calls were based on MegaRes classes. Core genes (found in > 41,000 genomes were excluded).





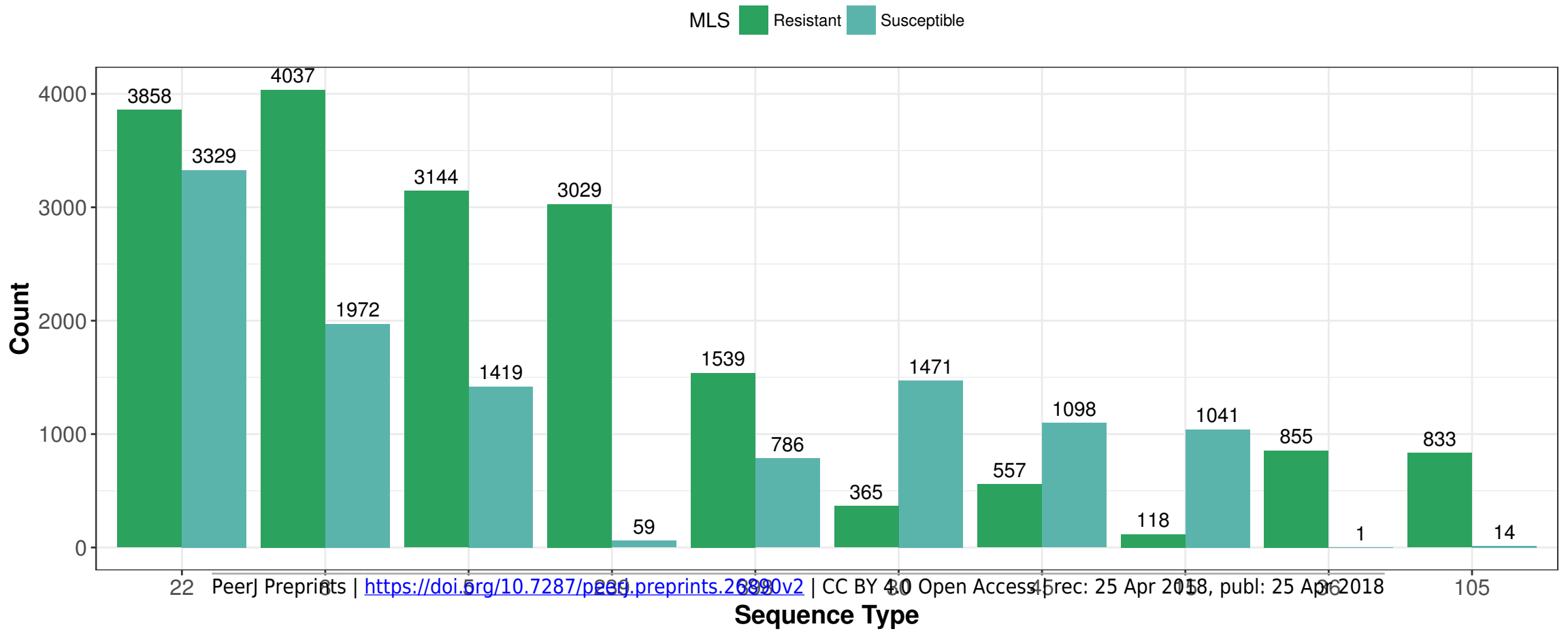
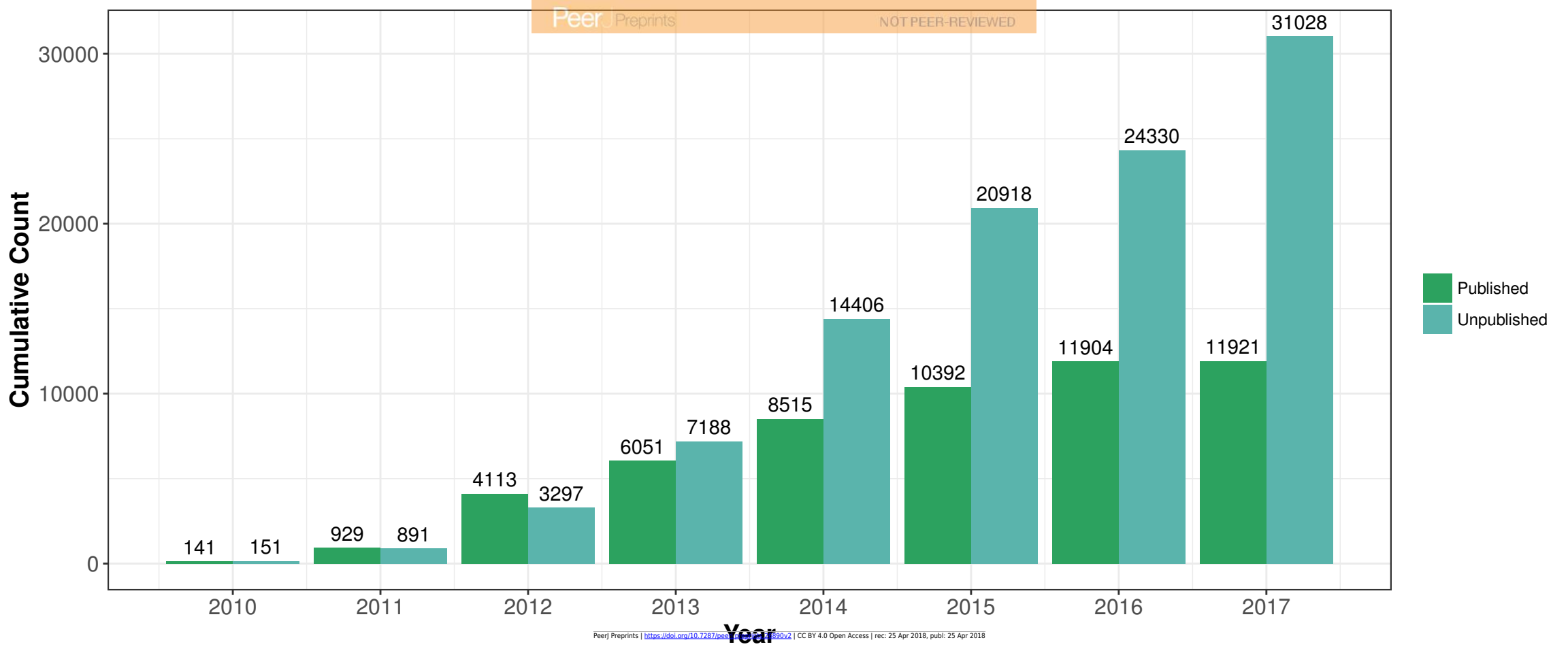


Figure 8(on next page)

Figure 8. Cumulative genomes linked to publications 2010-2017.

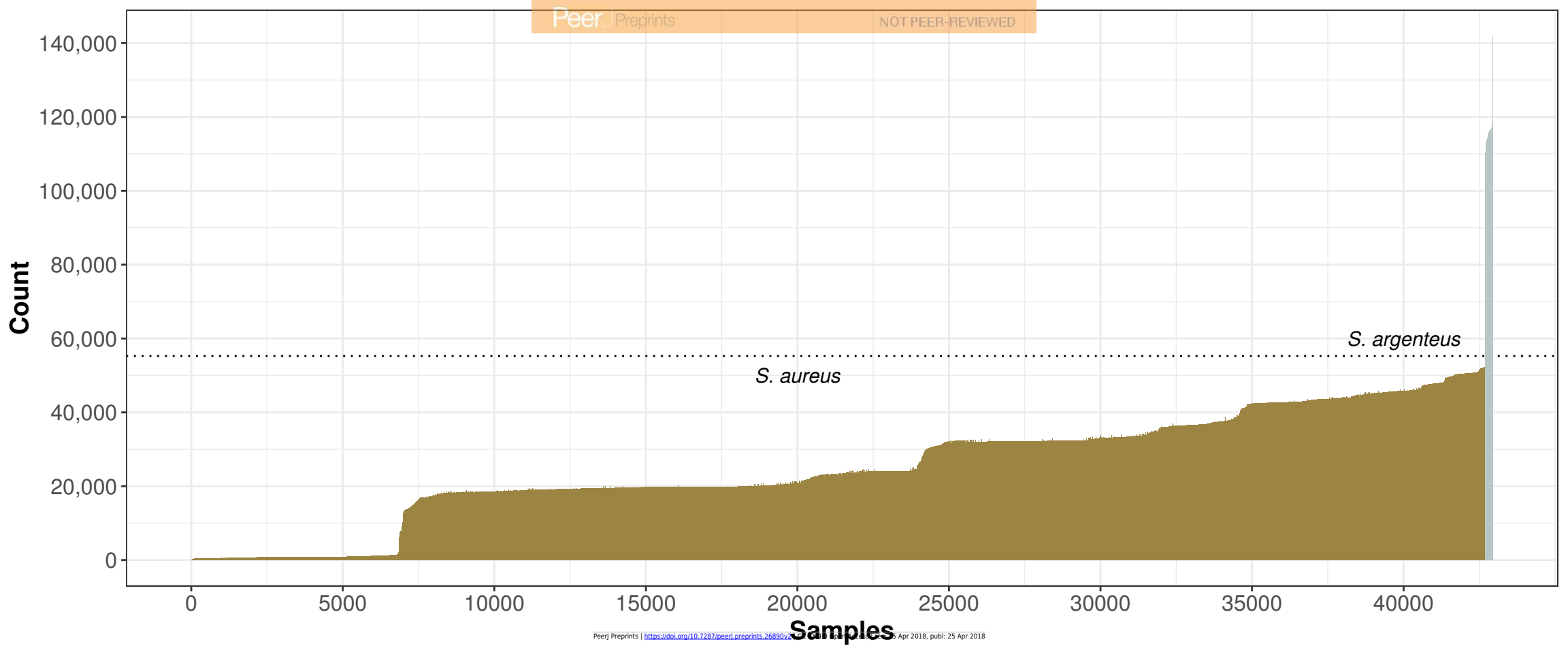


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3890v2 | CC BY 4.0 Open Access | rec: 25 Apr 2018, publ: 25 Apr 2018

Figure 9(on next page)

Figure 9. *S. aureus* SNP distance from reference S. aureus N315.

For each genome, the number of SNPs found by mapping reads to the N315 reference using GATK (McKenna et al., 2010) was plotted, with genomes ordered from least to most SNPs. 240 genomes with > 55,000 SNPs (dotted line) that had best matches to *S. argenteus* using mash (Ondov et al., 2016) were indicated by silver bars, the rest were *S. aureus* (gold).

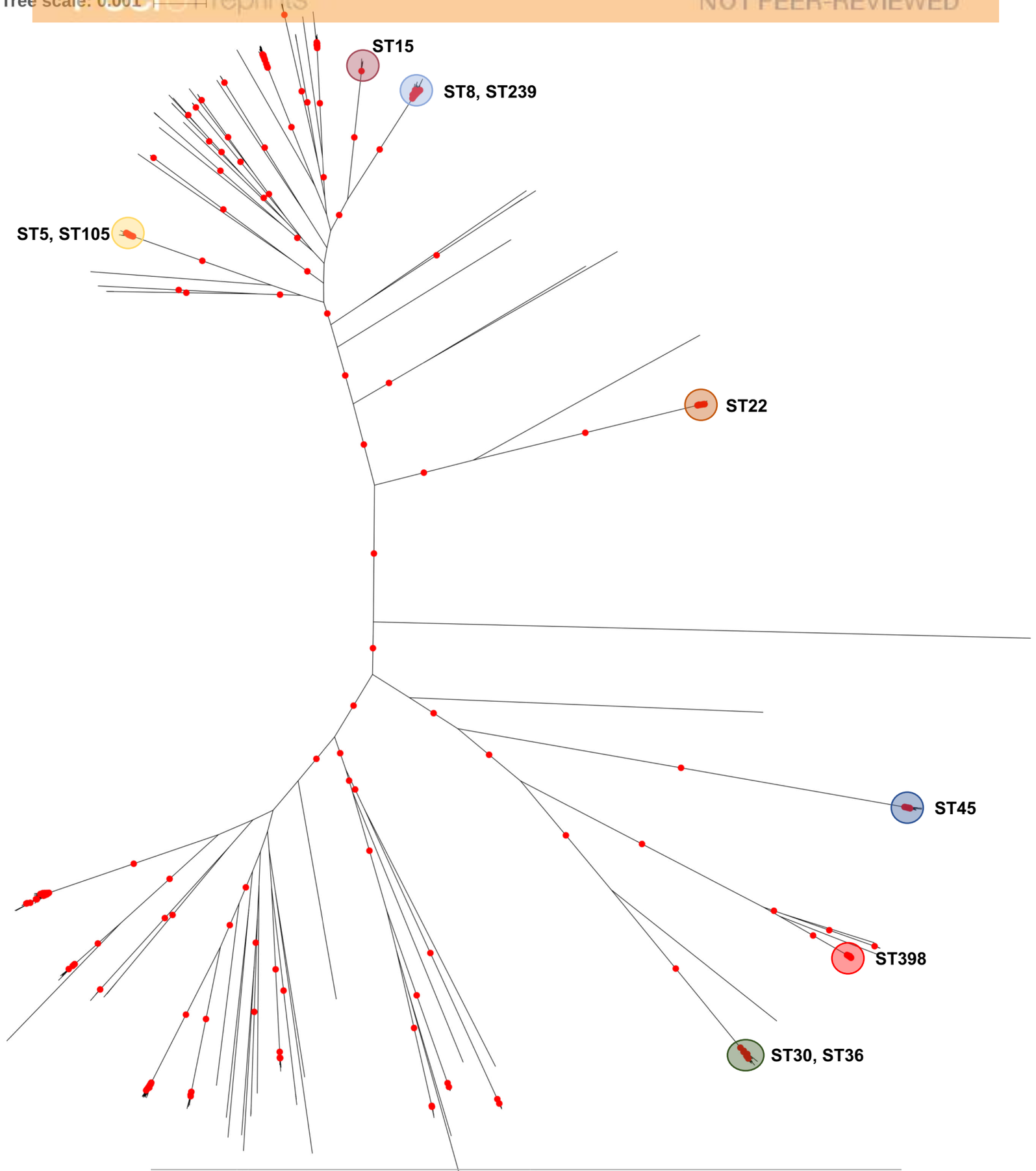


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26890v2 ... Apr 2018, publ: 25 Apr 2018

Figure 10(on next page)

Figure 10. Unrooted phylogeny of the *S. aureus* Non-Redundant Diversity (NRD) dataset.

An unrooted phylogenetic representation of the 380 genome non-redundant set (one representative per ST, all published and gold rank) using IQ-Tree (Nguyen et al., 2015) . Recombination sites were identified with ClonalFrameML (Didelot & Wilson, 2015) were filtered from the alignment. Clonal complexes containing the top ten most common STs are indicated with colored circles. The tree was built from 838 reconstructed core genes (please see Methods section) with 44,377 sites. Branches supported with probability > 0.9 are marked by red dots. The likelihood score for the tree was -1,890,510.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26890v2 | CC BY 4.0 Open Access | rec: 25 Apr 2018, publ: 25 Apr 2018

Table 1(on next page)

Table 1. Predicted SCCmec cassette type representation.

There were 26,462 samples with reads mapped to at least 50% of a SCCmec cassette. The table is a breakdown of the SCCmec cassettes with the highest percent match for each sample.

SCCmec Type	Count
I	689
II	5,183
III	2,807
IV	14,526
V	1,684
VI	171
VII	19
VIII	468
IX	0
X	20
XI	895

1

Table 2(on next page)

Table 2. Antibiotic resistance classes predicted by non-core genes.

Number of genomes with genes of resistance classes predicted by Ariba using the reference MegaRes database naming scheme.

Antibiotic Resistance Class	Count
Aminocoumarin	46
Aminoglycoside	17,968
Beta-lactam	37,758
Fluoroquinolone	69
Fosfomycin	24,205
Fusidic Acid	346
Glycopeptide	5,777
Lipopeptide	44
Macrolide-Lincosamide-Streptogramin (MLS)	22,322
Multi-Drug Resistance	13,653
Phenicol	852
Rifampin	46
Sulfonamide	36
Tetracycline	8,638
Trimethoprim	6,605

1