Integrating Staphopia into *Staphylococcus aureus* genomic investigations

•••

Robert A. Petit III, PhD
Emory University

Virtual *S. aureus* Seminar Series - Early Career Scientist Symposium
February 4th, 2021

Genomic Epidemiology

- Using whole-genome sequencing (WGS) for surveillance, outbreak investigations, and retrospective studies

- Case in point: SARS-CoV-2
  - 435k genomes sequenced
  - standardized workflows for data collection and analysis

- *S. aureus* has 70k+ WGS samples publicly available

Wouldn’t it be nice to use them?

It’s not that simple

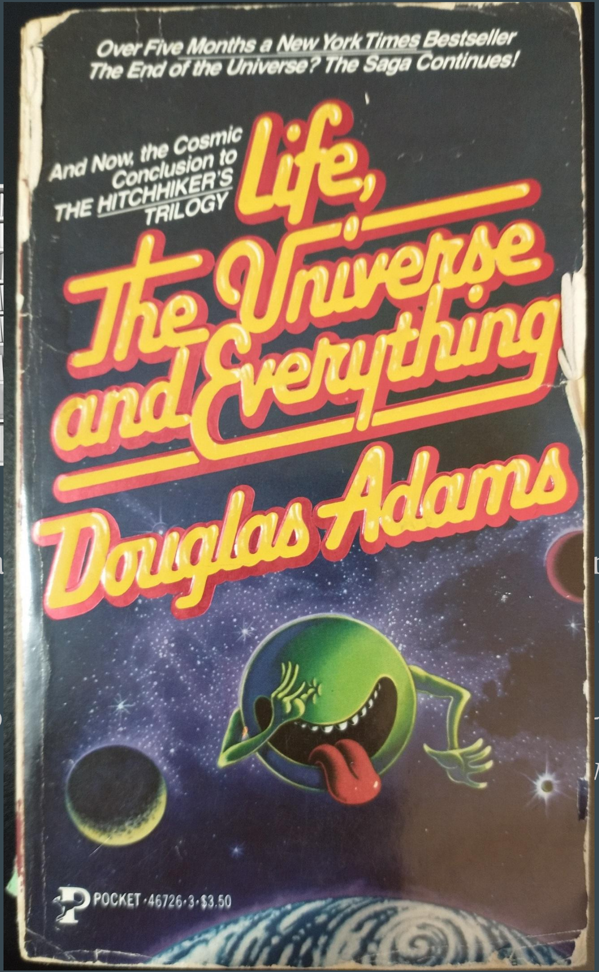
Let me explain why

Imagine, you’re at the *only* library in town looking for a sci-fi book

Where’s the Sci-Fi section? Here you go, DON’T PANIC!

What if your Sci-Fi book club curated the section?

What is the differen


Over Five Months a New York Times Bestseller
The End of the Universe? The Saga Continues!
And Now, the Cosmic Conclusion to THE HITCHHIKER'S TRILOGY
Life, The Universe and Everything
Douglas Adams
POCKET·46726·3·$3.50

- Blank covers tell you nothing a
  contents

- To find specific information yo
  open and read every book!

n adds descriptive book covers that
about the contents

-Fi section at the *only* library in town
/ be better utilized

Imagine, you want to use *S. aureus* genomes from the Sequence Read Archive

You were probably hoping for *"Descriptive Book Covers"*

Processed sequences

Collected metadata

Known genomic context

Possible to screen

Instead, the *only* library in town has *"Blank Book Covers"*

Raw sequences

Minimal metadata

No genomic context

Impossible to screen

Why is this the case? *They solve different problems*

Sequence Read Archive

- Hosts WGS for thousands of species
- Critical resource for genomics

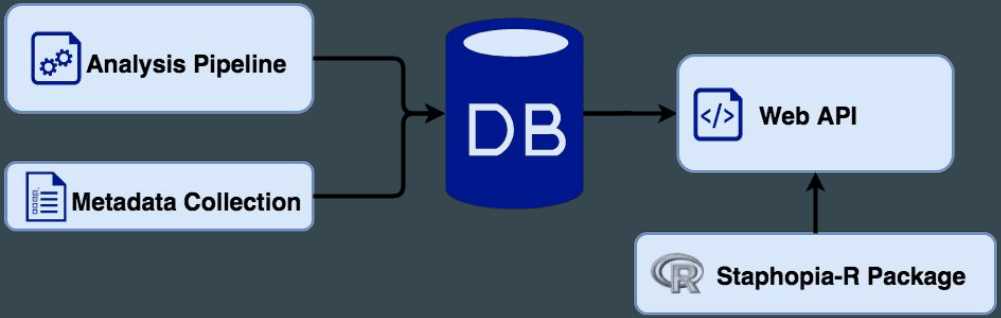
Curated Section of SRA

- Applies species-specific knowledge
- Can now better use the *only* library in town

How do we get "descriptive book covers" for *S. aureus?*

“*S. aureus club*” has to curate it, so we built Staphopia

- Analysis pipeline
  - Clean and standardized raw sequences
  - Assembly and annotation
  - Multi-locus sequence type (MLST)
  - Predicted AMR and virulence
  - Variants against *S. aureus N315*



- Metadata mined from literature

- Results for 40k+ genomes, free and publicly available
  - >75% high quality, 65% predicted MRSA, 1000+ STs represented
  - Programmatic access to all the data

Staphopia Case Study: *qacA* mediated chlorhexidine resistance

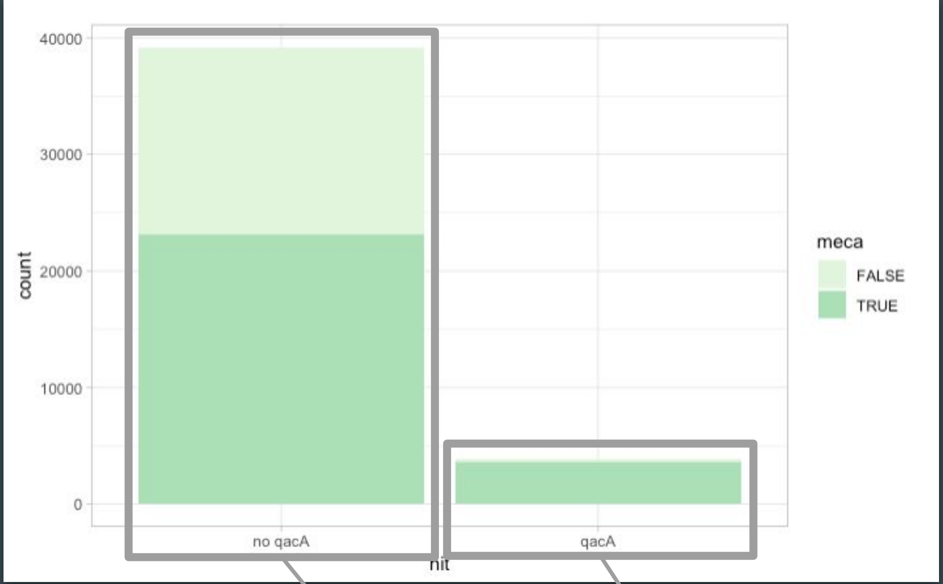
- Chlorhexidine is a biocide used for *S. aureus* decolonization

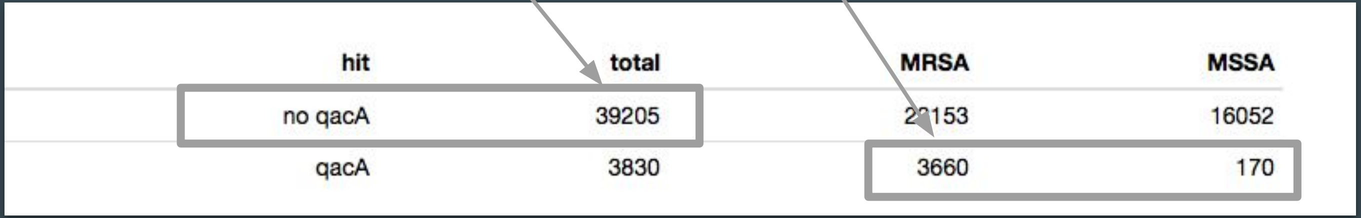
- *qacA* gene
  - Usually plasmid-borne
  - Encodes a muli-drug efflux pump associated with chlorhexidine resistance

- *Is qacA enriched in methicillin-resistant (MRSA) strains?*

- If present, how long has *qacA* been in the *S. aureus* population?

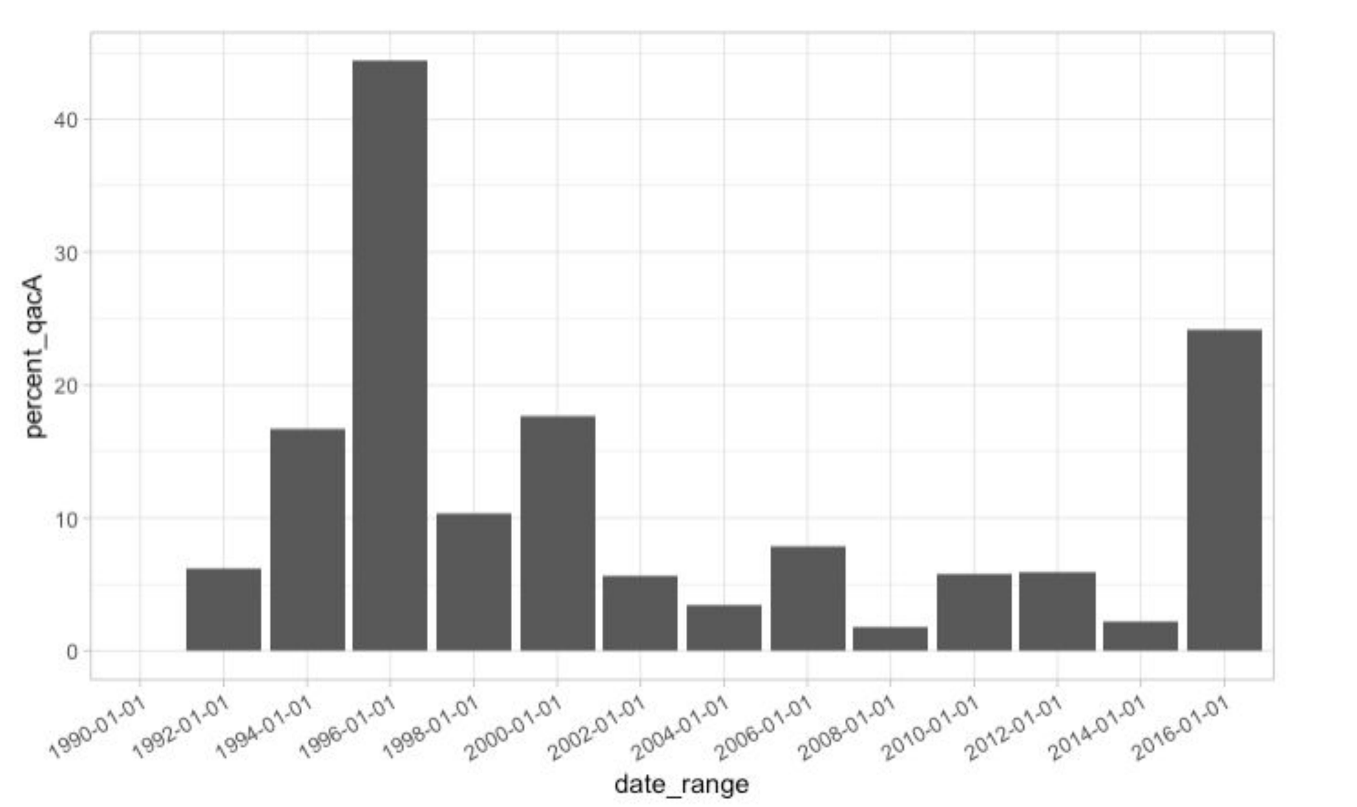
qacA is enriched in MRSA strains





hit	total	MRSA	MSSA
no qacA	39205	23153	16052
qacA	3830	3660	170

qacA has been present for many years



How can you use Staphopia?

Hey Staphopia, can I *please* have genomes that look like this?

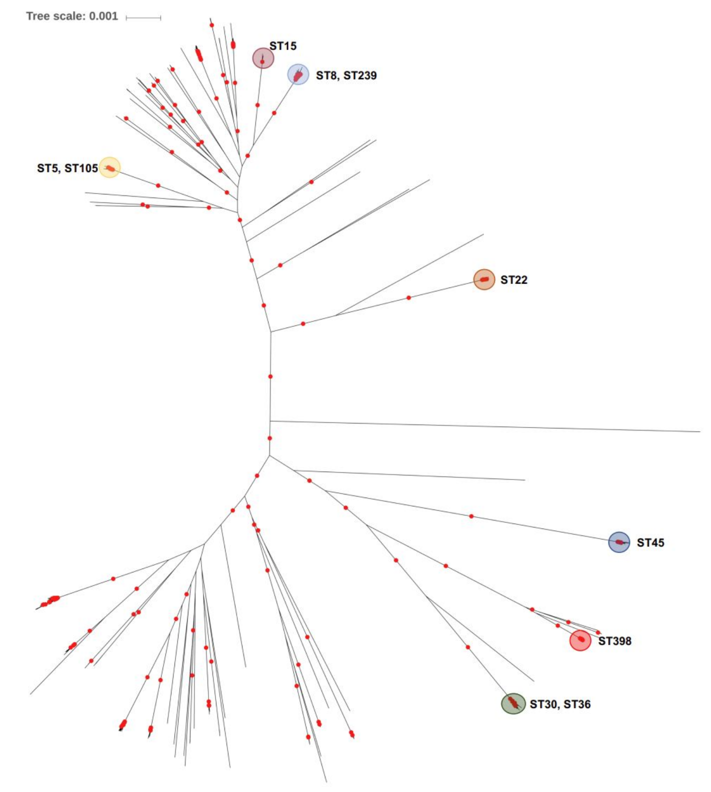
Staphopia: … searching …

Staphopia: *“Here you go!”*

Do you have to use all 40k+ genomes? Nope!

- Non-Redundant Diversity (NRD) Dataset
  - 380 genomes
  - High consistent quality
  - Linked to a publication
  - Each represents a unique sequence type

- Overcomes sample bias in Staphopia
  - 60% of the genomes are represented by only 10 sequence types



How have other studies used Staphopia?

- Rationally sub-sample public genomes (e.g. by sequence type)
    - (*Vrieling et. al, mBio, 2020*), (*Gill et. al., bioRxiv, 2021*)

- Search a specific signature against all public genomes
    - (Bennett et. al., J. Infect Dis. 2019)

- Uses for secondary tools
    - (Moustafa and Planet, Genome Biology, 2020) (*e.g. WhatsGNU*)

What’s next for Staphopia? *Version 2!*

- Reprocess all *S. aureus* public genomes with Bactopia
  - It’s done! *Using AWS Batch to process 67,000 genomes with Bactopia*
- Repackage as a wrapper around Bactopia to include *S. aureus* specific analyses
  - *S. aureus* specific datasets (e.g. *spa*, SCCmec, etc...)
- Improve Staphopia API useability

Let’s wrap it up

- The SRA provides genomes with a *“Blank Cover”*
- Staphopia creates *“Descriptive Book Covers”* for *S. aureus* section of SRA
- Staphopia can be used in many ways
- Version 2 will be available soon!

Thank you very much! Questions?

The many scientists and their funders who provided WGS to the public domain, DDBJ, ENA, and SRA for storing and organizing the data, and the authors of the open source software tools and databases used in this work.

EMERGENT Group

- Tim Read (PI)
- Ashley Alexander
- Jon Moller
- Michelle Su
- Brooke Talbot
- Vishnu Raghuram
- Emily Wissel

Links

- Staphopia
- Bactopia
- Twitter @rpetit3
- GitHub @rpetit3

Publications

- Staphopia PeerJ
- Bactopia mSystems

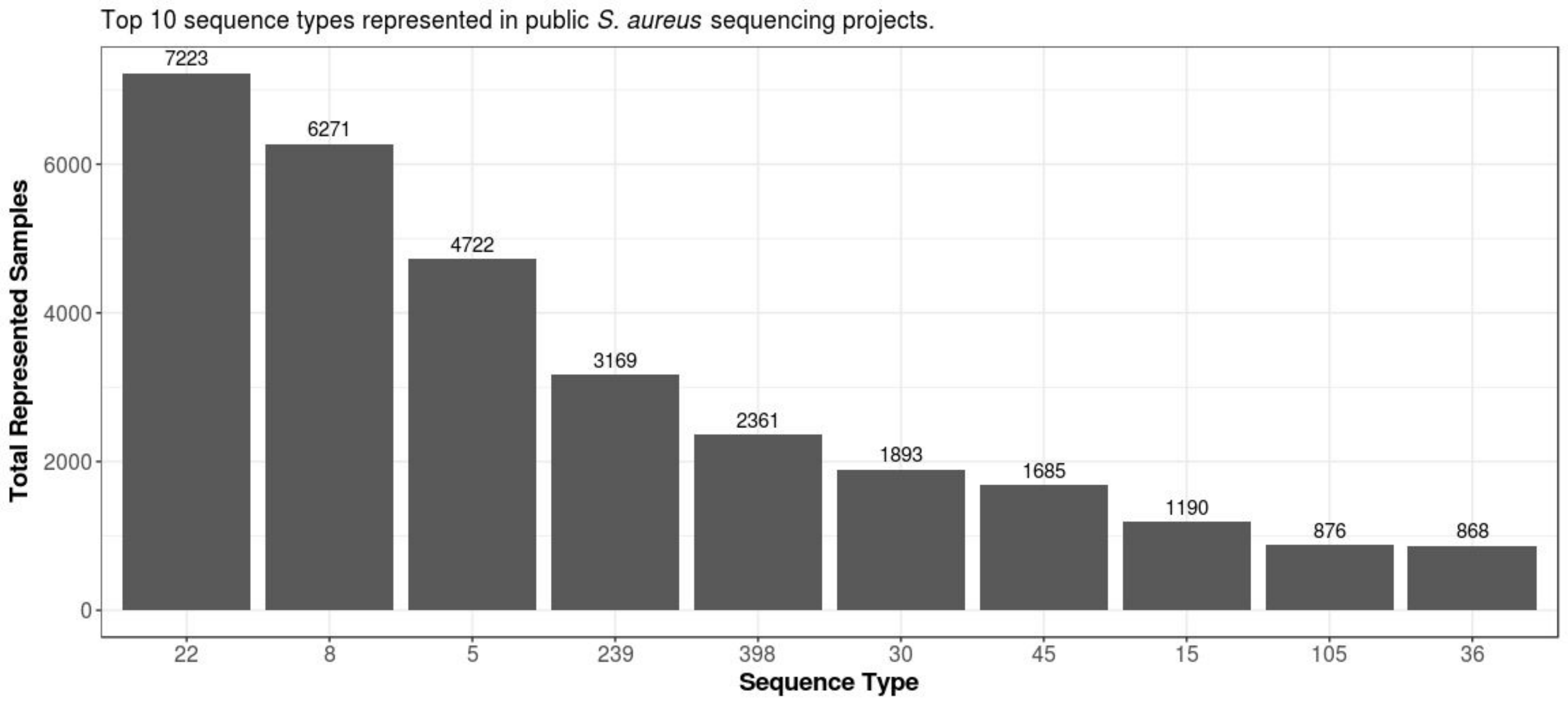
EMORY UNIVERSITY SCHOOL OF MEDICINE

NIH National Institute of Allergy and Infectious Diseases

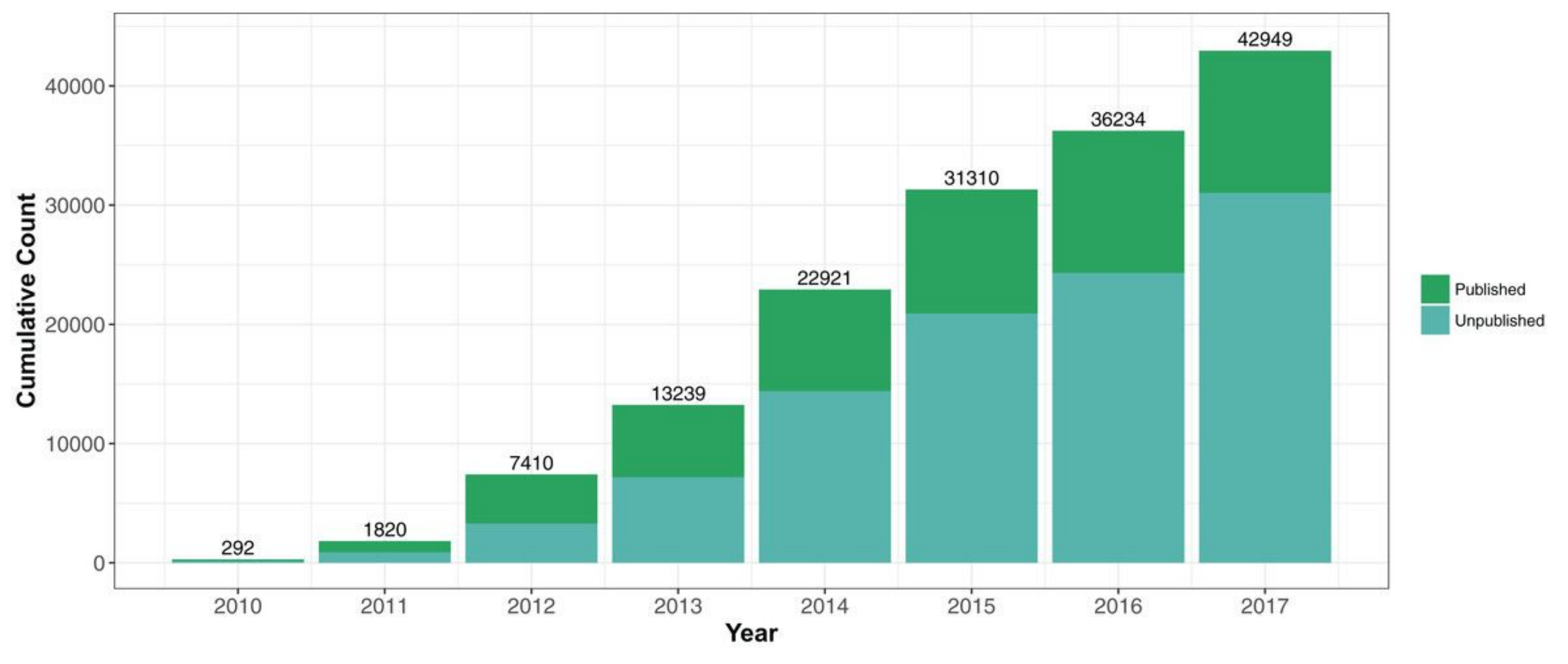
GEORGIA EIP emerging infections program

amazon web services | Cloud Credits for Research

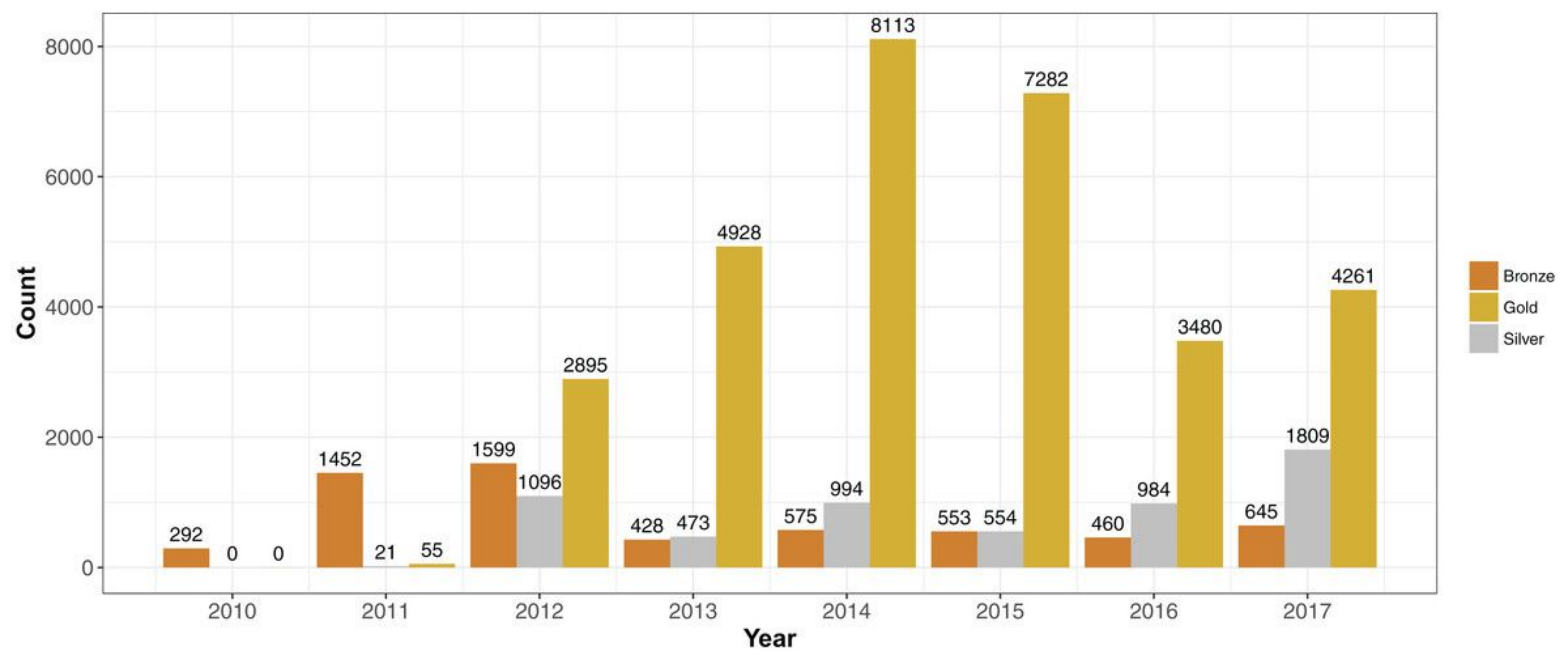
Top 10 STs are 60% of all *S. aureus* genomes



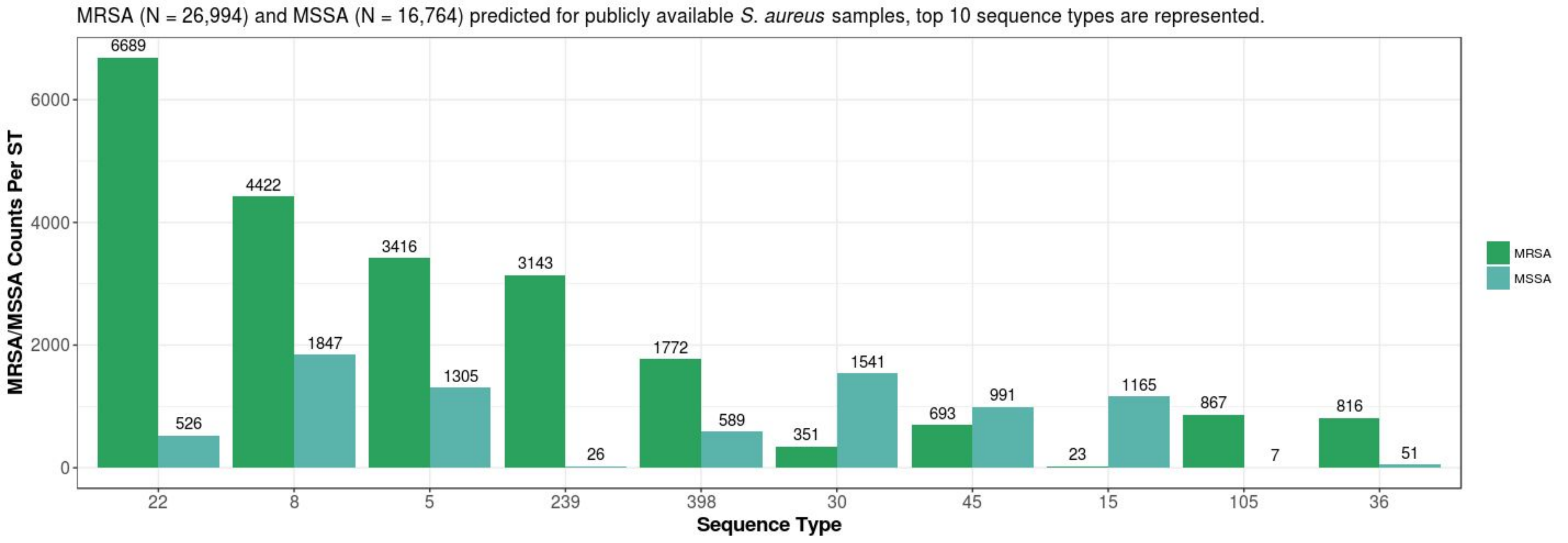
Published vs Unpublished in *S. aureus* SRA genomes



S. aureus quality grades by year



Methicillin-resistance by Sequence Type


MRSA (N = 26,994) and MSSA (N = 16,764) predicted for publicly available *S. aureus* samples, top 10 sequence types are represented.