ESGEM-AMR

ESCMID

AMRrules

Agenda

1. **Qualibact**
   - Motivation
   - Tool
   - Method
   - Examples
2. **Resources**
   - SpecCheck
3. **Future Planning**
   - Next meeting
4. **Other questions**

Defining AMRrules via geno/pheno analysis

Need to establish standards for quality control (QC) of

- Genome data
- Phenotype data

Need an approach that can be applied consistently across organisms

QualiBact provides a potential solution to this by establishing a systematic approach to defining QC thresholds for WGS data from multiple organisms

AMRrules

ESCMID | ESGEM-AMR

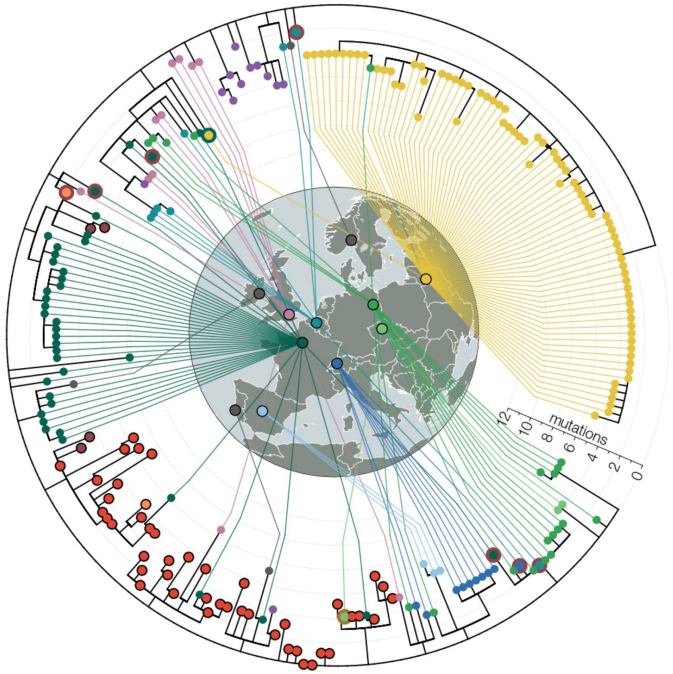
QualiBact

Nabil-Fareed Alikhan

https://happykhan.github.io/qualibact/species/

Motivation

- Many laboratories across the globe are implementing pathogen surveillance using genomics.
- However, **the expertise in each individual pathogen is lacking** and laboratories face challenges to interpret the sequencing results, which may have direct impact in downstream analysis (e.g. AMR detection).



AMRrules

ESCMID | ESGEM-AMR

Motivation

Gap:

- Current QC tools are species-agnostic, but bacterial genomes differ widely → species-specific thresholds are needed.

Our contribution – *QualiBact*:

- QC metrics + thresholds for **304 bacterial species** from large-scale Illumina data (AllTheBacteria).
- Curated with machine learning. Procedural. Automated.

Objective

- To develop a new, generalised and automated approach — scalable across species and adaptable beyond AMR priorities.
- An approach that ensures consistent, reliable assemblies for genotyping, AMR/virulence detection, and population structure analyses.

What is Qualibact?

Scale:

- 2.3 M genomes, 304 species (98 genera).
- 10 species dominate dataset (78%).

Filtering:

- Avg 5.7% genomes removed; <10% filtered in 283/304 species.
- “Passed” / “Failed” genome lists provided.

Variation:

- Genome size 0.8–8.35 Mb; GC content 27–73%.
- Wide inter-species diversity.

Dataset skew:

- Highly uneven taxa: *Salmonella* 625 k, *Escherichia* 390 k, *Streptococcus* 211 k.

Differences:

- RefSeq vs SRA assemblies differ in size, CDS count, contamination.

Limitations:

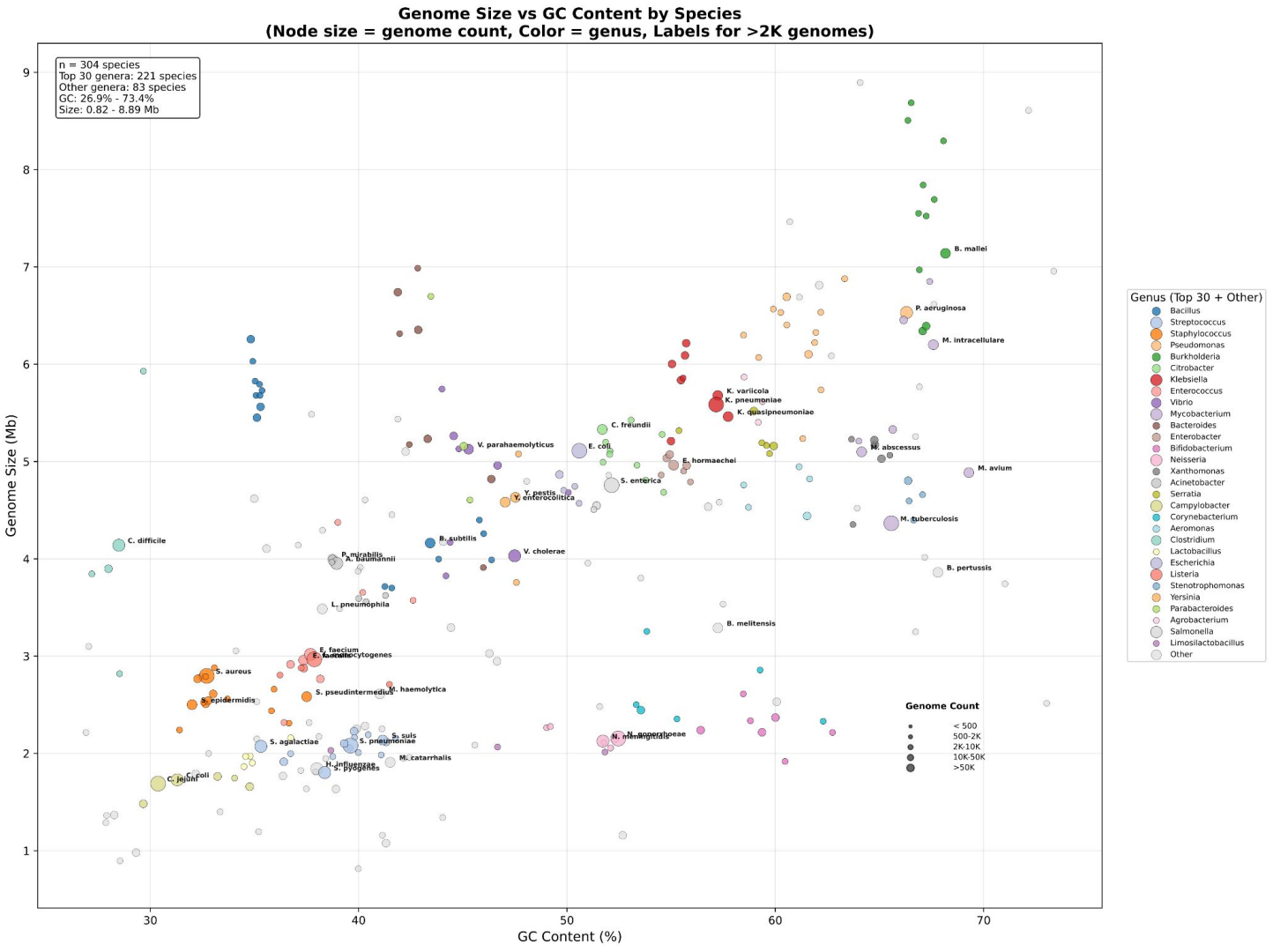
- Species definitions via GTDB; thresholds tuned for SPAdes/Shovill.

Reproducibility:

- Automated, scalable workflow; thresholds update with new data.

An aside

- Broad diversity: 0.8–8.9 Mb, 27–73% GC.
- 10 species dominate (78% of data).
- Node size = genome count;
- Colour = genus.
- Clear genus-level clustering by GC % / size.
- Highlights uneven taxon representation.

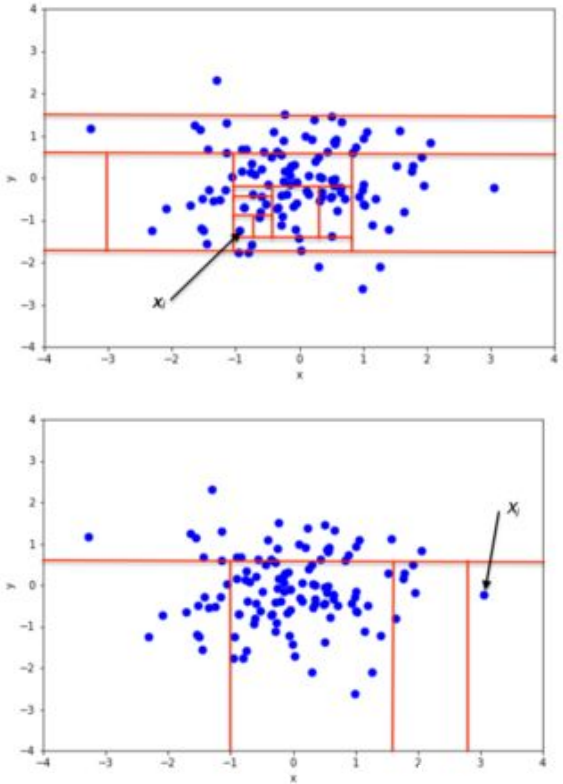


Method for calculating thresholds

What didn't work: mean +/- std.dev, Z-score, Winsorizing to handle outliers, Gaussian mixture models, Kernel density estimation, Clustering e.g. DBSCAN to find majority, Math.random()

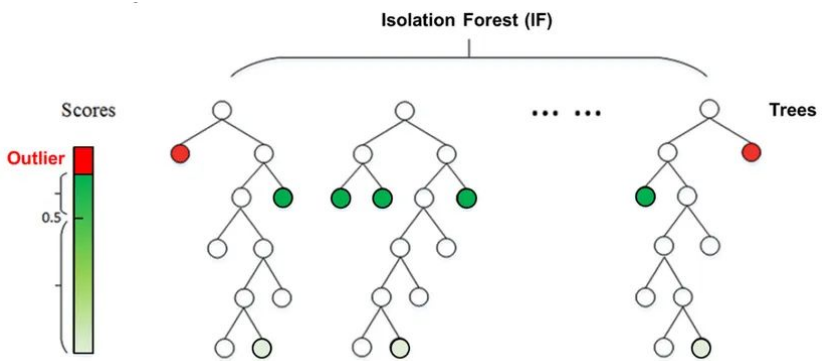
Isolation forests

- Unsupervised anomaly detection algorithm
  Anomalies are easier to isolate than normal points
- Builds multiple **Isolation Trees (iTrees)** using random splits
- Each tree recursively partitions data using:
  - Randomly selected feature
  - Random split value between min & max
- **Path length** (root → leaf) indicates how isolated a point is
- Anomalies have **shorter average path lengths** across trees
- Fast, memory-efficient – good for high-dimensional data
- Outputs an **anomaly score** based on path length



Anomaly scores

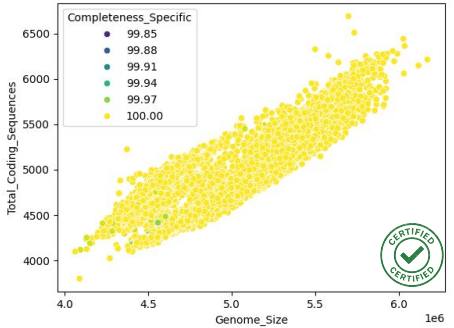
- Unsupervised score that quantifies *isolation* of a point.
- Built from many random isolation trees (iTrees).
- **Path length** (root → leaf): shorter ⇒ more isolated ⇒ more anomalous.
- Score = transforms average path length into 0–1 range (higher → more anomalous).
- Typical interpretation: **score ≈ 1 → anomaly; score ≈ 0 → normal**.
- Thresholding: score > cutoff ⇒ flag as anomaly (choose cutoff by validation or expected contamination).
- Fast and memory-efficient; works well in high-D and with mixed features.
- Practical tips: use many trees + small subsample for robustness; tune cutoff using labelled data or expected anomaly rate.

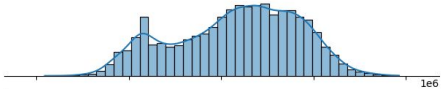


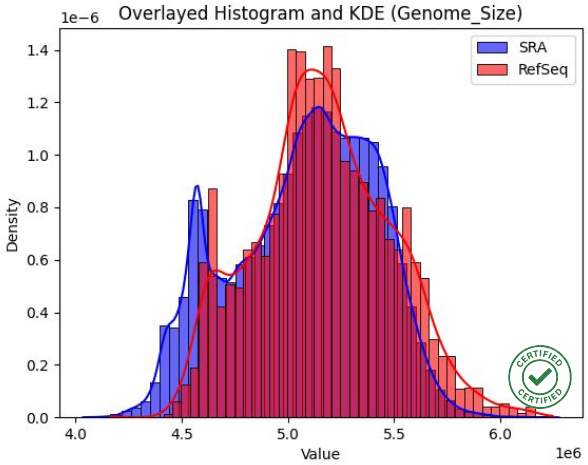
Escherichia coli: a successful example

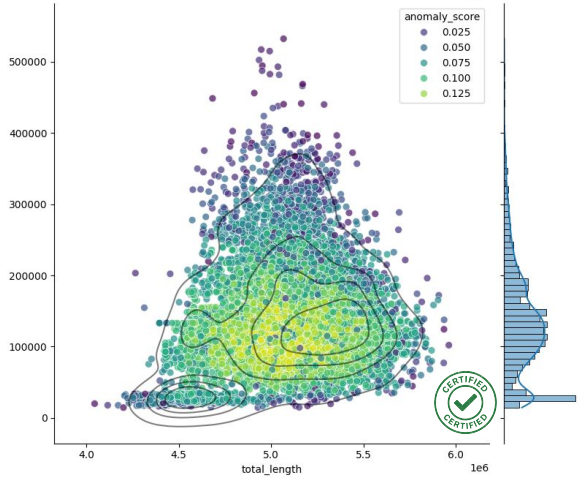
Based on **4,033** RefSeq & **399,884** ATB/SRA genomes.

metric	lower_bounds	upper_bounds
N50	20000.0	
no_of_contigs		700.0
GC_Content	50.0	52.0
Completeness	95.0	
Contamination		16.0
Total_Coding_Sequences	3900.0	6500.0
Genome_Size	4100000.0	6300000.0







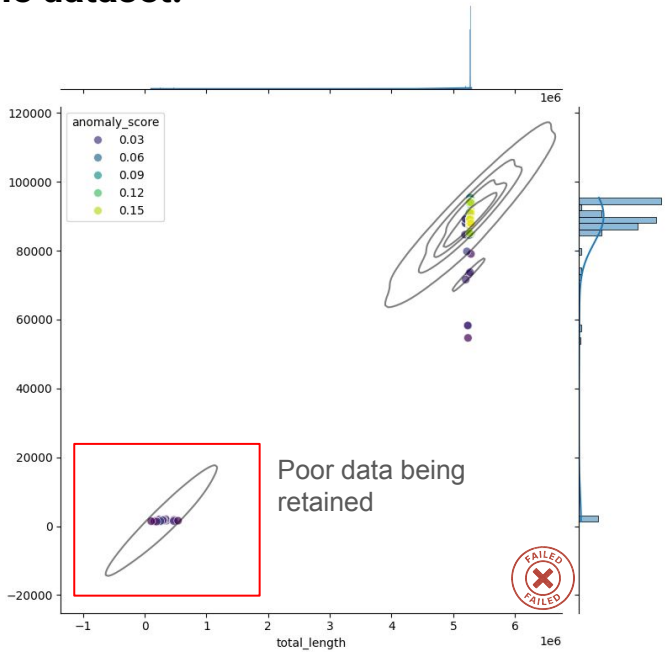


Citrobacter rodentium - cause for concern

- These thresholds are based on 353 genomes
- Thresholds constrained by available data
- Strange cutoffs
- Can be addressed manually (but that defeats the aim)

metric	lower_bounds	upper_bounds
N50	1000.0	
no_of_contigs		310.0
GC_Content	51.0	55.0
Completeness	5.0	
Contamination		1.0
Total_Coding_Sequences	200.0	5300.0
Genome_Size	100000.0	5500000.0

N50 vs total length, post filtering on the dataset.



Main reasons for rejecting assemblies

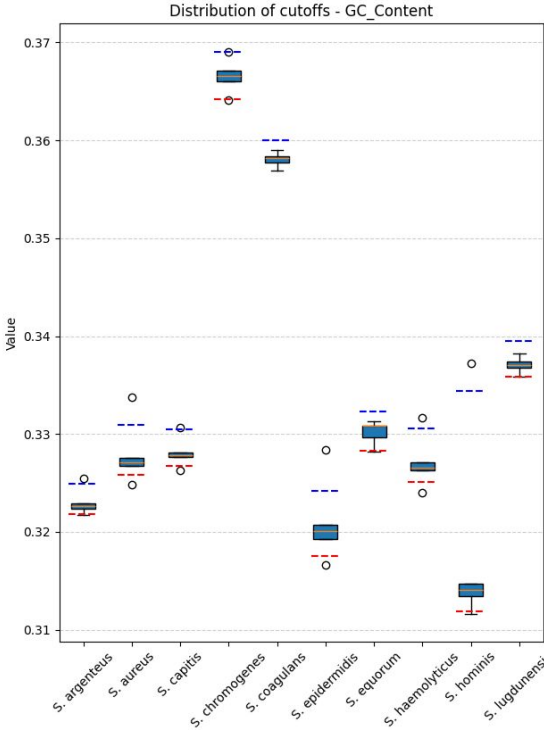
Reason for rejection	Count	Percentage
N50	33755	69.1%
No. of contigs	7208	14.7%
Various reasons*	4948	10.1%
Contamination (CheckM2)	1460	3.0%
Total Coding Sequences	820	1.7%
Genome size	486	1.0%
Completeness (CheckM2)	130	0.3%
GC content	70	0.1%

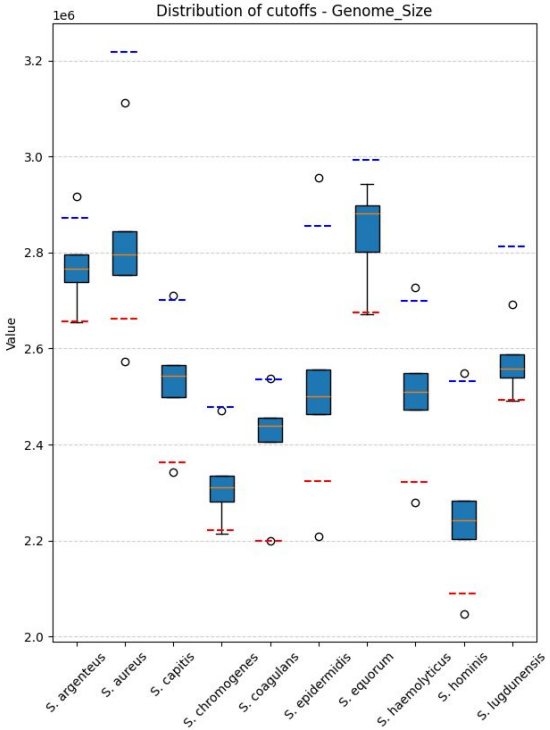
*Most common being a combination of N50+No. of contigs

An example from *Staphylococcus* spp. (*S. aureus*)

Staphylococcus included (> 100 genomes)

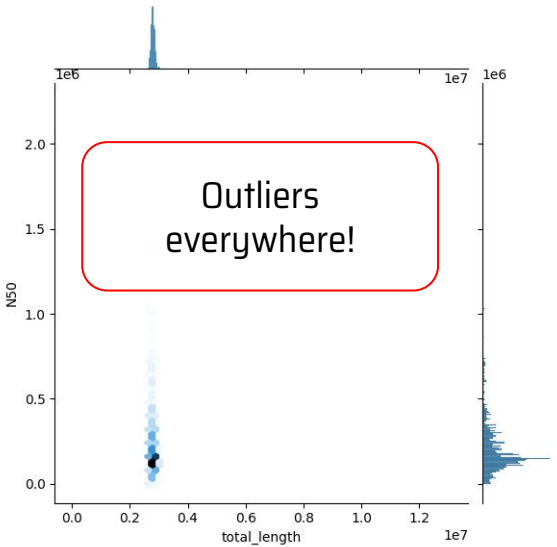
- *Staphylococcus argenteus*
- *Staphylococcus aureus*
- *Staphylococcus capitis*
- *Staphylococcus chromogenes*
- *Staphylococcus coagulans*
- *Staphylococcus epidermidis*
- *Staphylococcus equorum*
- *Staphylococcus haemolyticus*
- *Staphylococcus hominis*
- *Staphylococcus lugdunensis*



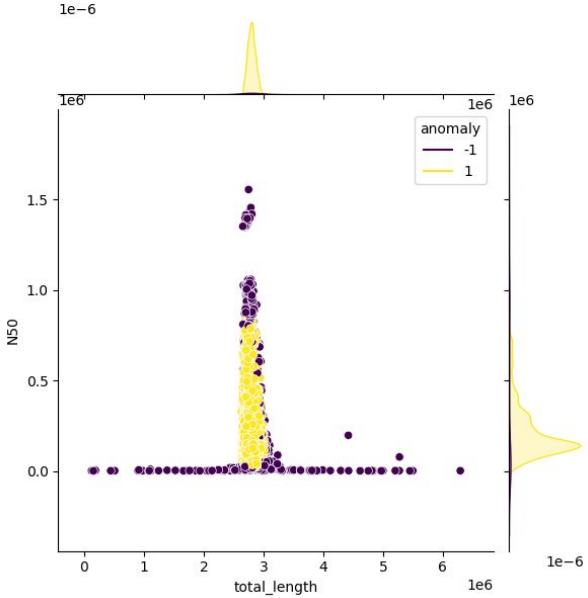


S. aureus N50 and total length

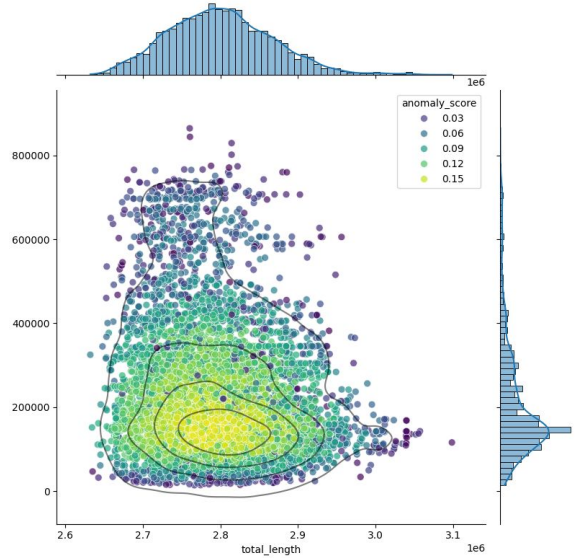
All the data: #nofilter



Purple filtered as anomalies



Pick ranges from this distribution

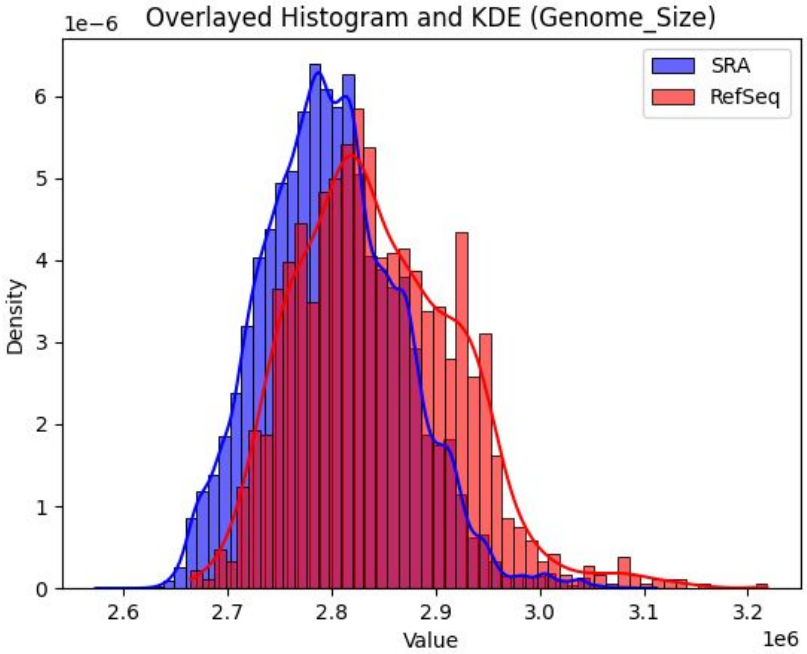


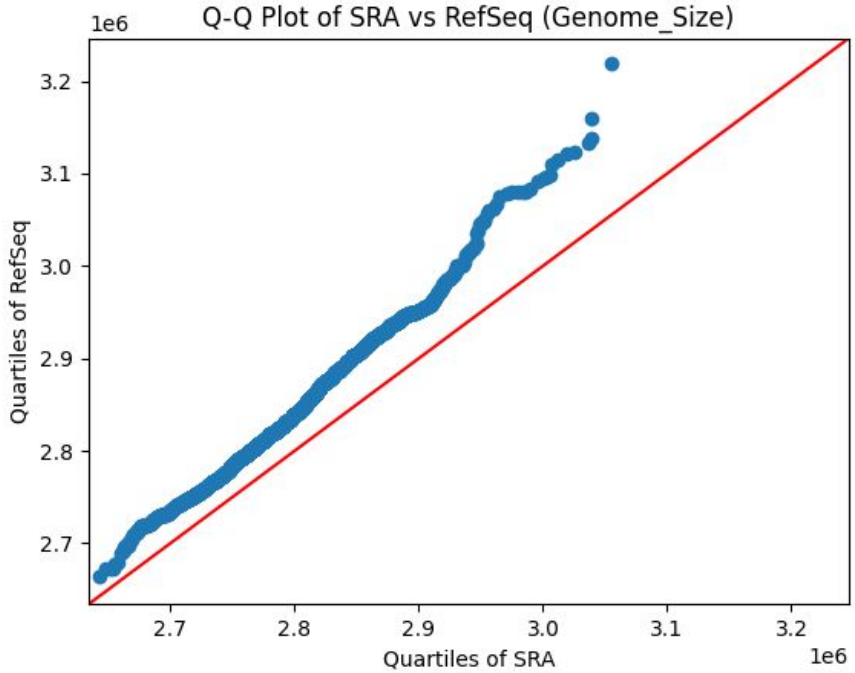
What does this look like? - *S. aureus*

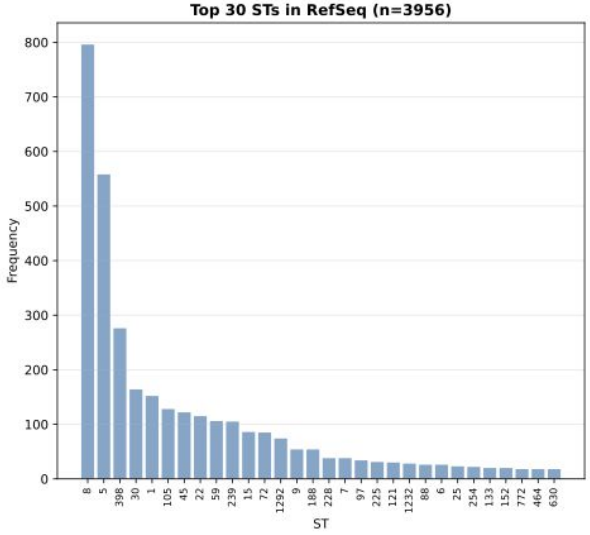
Metric	Lower bounds	Upper bounds
N50	26000	
No. of contigs		280
GC content	32	34
Completeness (CheckM2)	95	
Contamination (CheckM2)		4
Total coding sequences	2400	3400
Assembly size (Shovill)	2600000	3300000

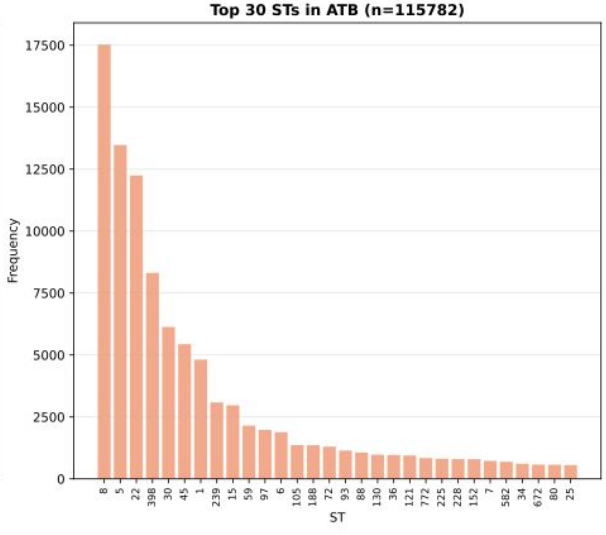
S. aureus example

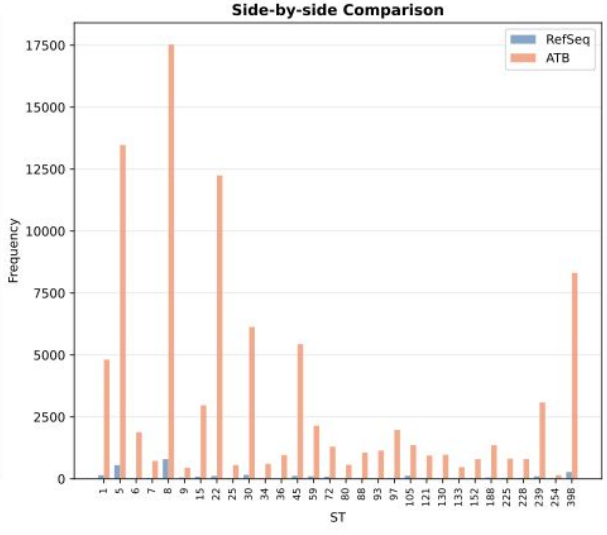
A Q-Q plot (quantile-quantile plot) visually compares the distribution of a dataset to a theoretical distribution, highlighting deviations from normality or other expected distributions.

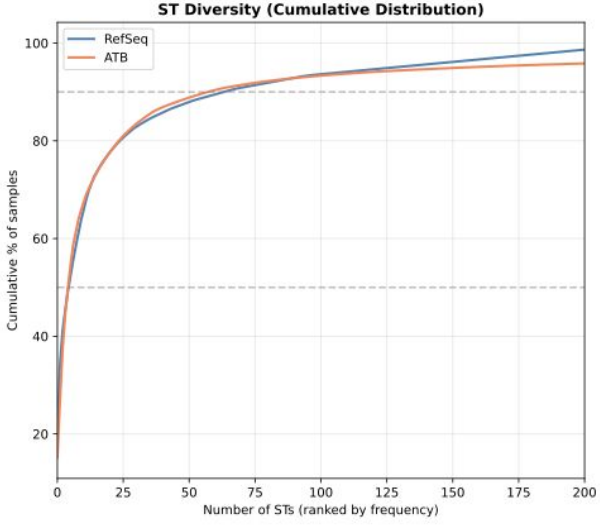


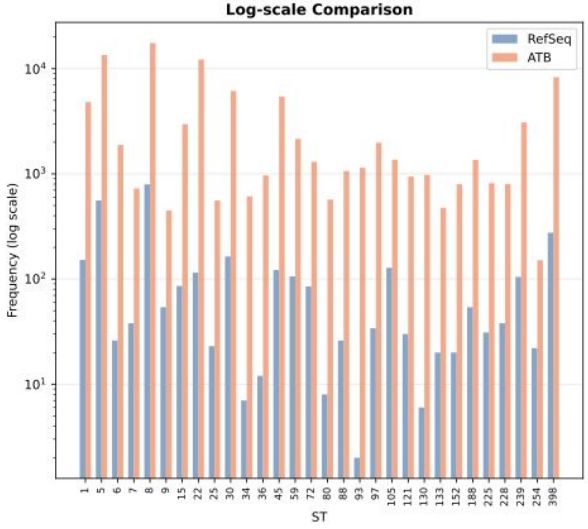






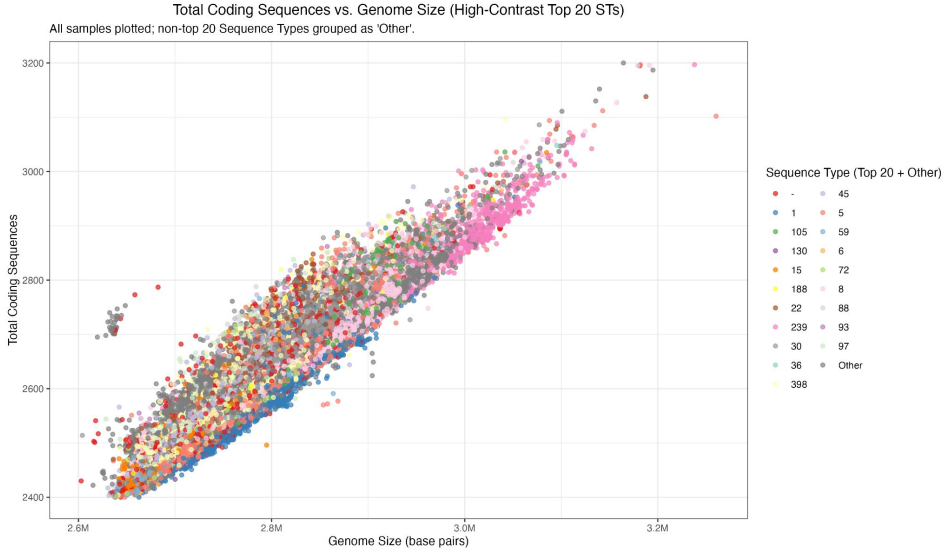


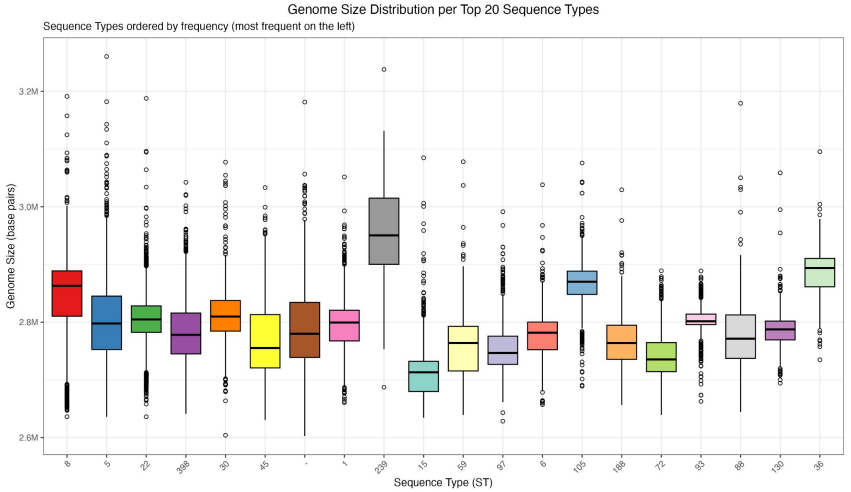




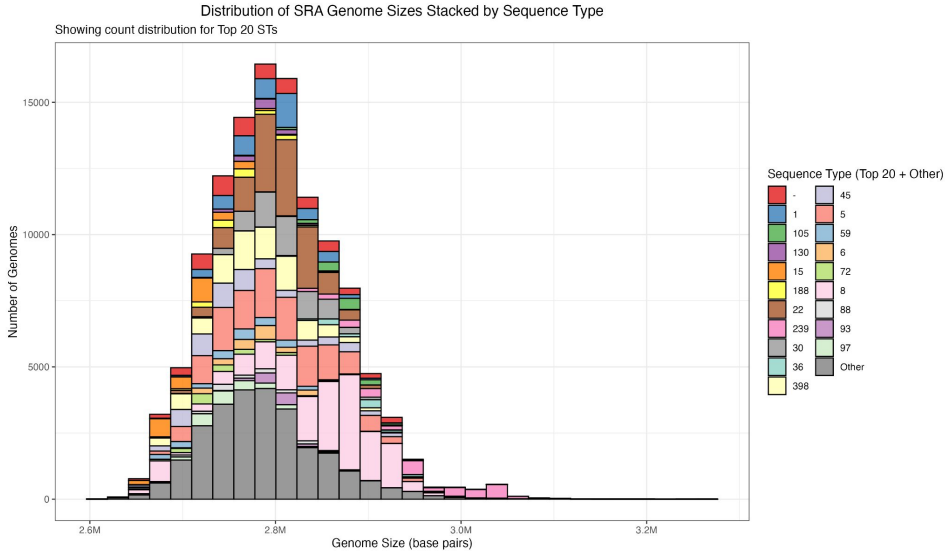

```
SUMMARY STATISTICS

RefSeq Database:
• Total samples: 3,956
• Unique STs: 239
• Samples in shared STs: 3,808 (96.3%)
• Most common ST: 8 (n=796)

ATB Database:
• Total samples: 115,782
• Unique STs: 2,225
• Samples in shared STs: 106,404 (91.9%)
• Most common ST: 8 (n=17529)

Overlap:
• Shared STs: 182
• RefSeq-only STs: 57
• ATB-only STs: 2,043

Top 5 Shared STs (by combined frequency):

1. ST8: RefSeq=796, ATB=17529
2. ST5: RefSeq=558, ATB=13476
3. ST22: RefSeq=115, ATB=12244
4. ST398: RefSeq=276, ATB=8315
5. ST30: RefSeq=164, ATB=6132
```

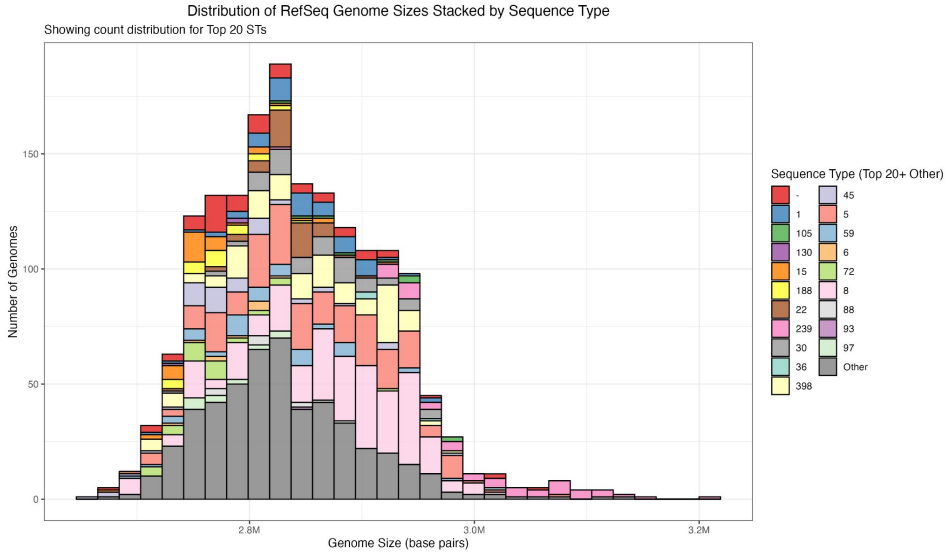
Genome size per ST



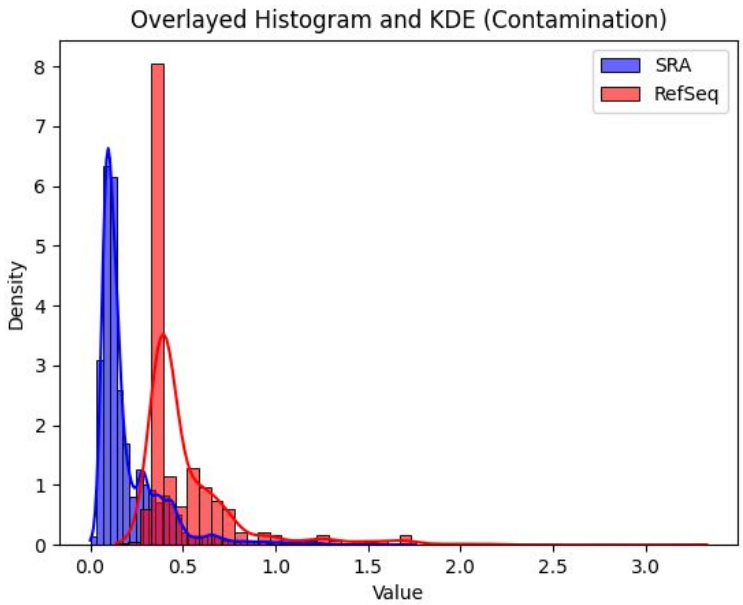


Genome size per ST





Contamination



Contamination in the RefSeq dataset is slightly higher than in the SRA dataset:

- Reference, complete assemblies with true contaminants?
- Regions in the assemblies that are not in the databases (or are but coded as something else) and are therefore reported as “contamination”.

Conclusions and future directions

- Automated, generic QC thresholds from Qualibact.
- Serve as guidelines, not replacements for expert judgment.

- **Goal: evolve into a regularly updated, versioned, consistent framework.**
- **Need to automatically flag weakly supported thresholds.**
- **Separate read-based QC guidance in development.**
- **Speciation - Speciator? Separate benchmark to come.**

ESGEM-AMR contributions

Revise species-specific thresholds and provide feedback through a form

Why?

- We would like the thresholds to be validated by experts on those species.
- We would like to use the QualiBact thresholds to assess the quality of the genomes shared and included in the ESGEM-AMR project.

ESGEM-AMR contributions

Feedback requested from subgroups:

- Visit https://happykhan.github.io/qualibact/species/ and check the thresholds that have been set for your organism.
- Fill in the form: https://forms.gle/S4wGGvbwvZkSjCUr9.
- **Submit one form per species until November 21.**

All contributing subgroups will be listed as co-authors of the Qualibact manuscript and given an option to review it before submission.

Agenda

1. **Qualibact**
   - Motivation
   - Tool
   - Method
   - Examples
2. **Resources**
   - SpecCheck
3. **Future Planning**
   - Next meeting
4. **Other questions**

AMRrules

ESCMID | ESGEM-AMR

SpecCheck

SpecCheck is a modular command-line tool for collecting, validating, and summarizing quality control (QC) metrics from genomic analysis pipelines. It automatically detects and processes outputs from multiple bioinformatics tools, validates them against customizable criteria, and generates comprehensive reports with optional interactive visualizations.

Supported Modules

The collect command automatically detects outputs from:

- **CheckM2: Completeness, contamination, genome metrics**
- **QUAST: Assembly statistics (N50, contigs, GC content)**
- Speciator: Species identification and confidence
- ARIBA: Marker gene detection
- Sylph: Profiling and ANI values

https://github.com/happykhan/speccheck

ReadQC coming soon!

https://github.com/ghruproject/bactscout

Should we adopt QualitBact as our standard approach to genome data QC?

metric	Lower bounds	Upper bounds
N50	54000.0	
no_of_contigs		390.0
GC_Content	56.0	59.0
Total_Coding_Sequences	4700.0	6500.0
Genome_Size	4900000.0	6400000.0
Completeness	95.0	
Contamination		5.0

QUAST

CheckM2

Example thresholds for *Klebsiella pneumoniae:* https://happykhan.github.io/qualibact/Klebsiella/Klebsiella_pneumoniae/

Agenda

1. Qualibact

- Motivation
- Tool
- Method
- Examples

2. Resources

- SpecCheck

3. Future Planning

- Next meeting

4. Other questions

Future planning

November 2025

- Updates from the subgroups
- Updates from Kat re AMRrulemakeR package
- Updates from Jane re AMRrules Python package

December 2025

- Manuscript proposals and planning

AMRrules

ESCMID | ESGEM-AMR

Any other questions?