Development of a Rapid, High-Resolution Microbial Identification Platform

Zachary Pimentel[1]; Andrea Watson[1]; Giel Göertz[1]; Marcel Hillebrand[1]; Bernice Westrek-Esselink[1]; Julia Maritz[1]; James Cass[1]

[1]Merck & Co., Inc., Rahway, NJ, USA

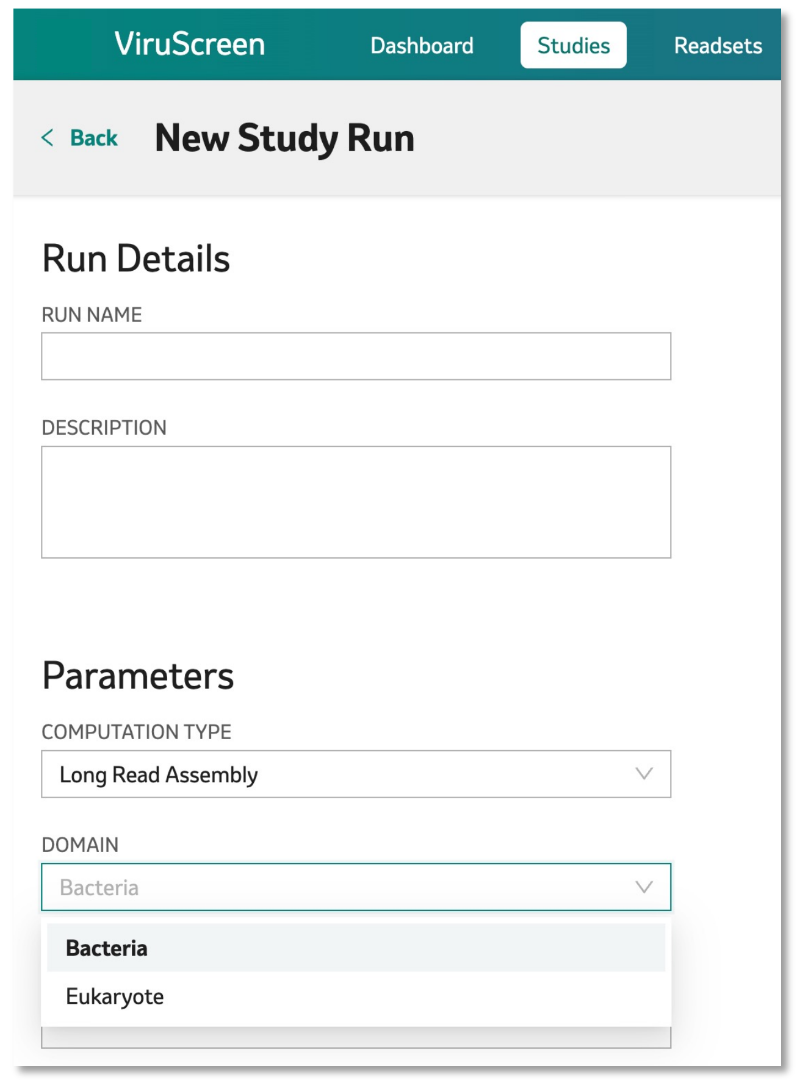
Background

- Producing safe medicinal products is dependent on pharmacopeial and conventional methods for identification of microbial contaminants, also known as adventitious agents.
- Genotypic methods for microbial identification use only selective loci covering only small fractions of the genome, limiting taxonomic resolution and missing whole-genome insights.
- Advancements in DNA sequencing technology have reduced the time and cost needed to sequence whole genomes, enabling their use for microbial identification.
- **We leveraged Oxford Nanopore whole-genome sequencing (WGS) to develop workflows for rapid, high-resolution, in-house identification of bacteria and yeast.**
  - Our bacterial workflow can differentiate closely related strains of *S. flexneri* and *E. coli*.
  - Our yeast workflow can identify *S. cerevisiae* and *K. phaffii* species.
- We are automating these pipelines as part of Merck's patent-pending ViruScreen platform which enables multi-'omic analyses through an easy-to-use web portal.

ViruScreen Platform

- ViruScreen is a patent-pending GMP bioinformatics enablement platform.
- Supports multiple high-throughput sequencing-based analysis pipelines for detection of adventitious agents in vaccine and biologics samples.
- Designed for use by non-bioinformaticians (**Figure 1**):
  - Web-based
  - User-friendly graphical interface
  - Suggested pre-set parameters tested for efficiency
- End-to-end pipeline execution from simple data upload from the cloud or a local computer, to detailed tabular and graphical summaries of analysis results (**Figure 2**).
- New features added regularly through continuous development and ongoing testing to verify pipeline functionality.

Figure 1. ViruScreen assembly workflow new run page



New Study Run

Run Details

RUN NAME

DESCRIPTION

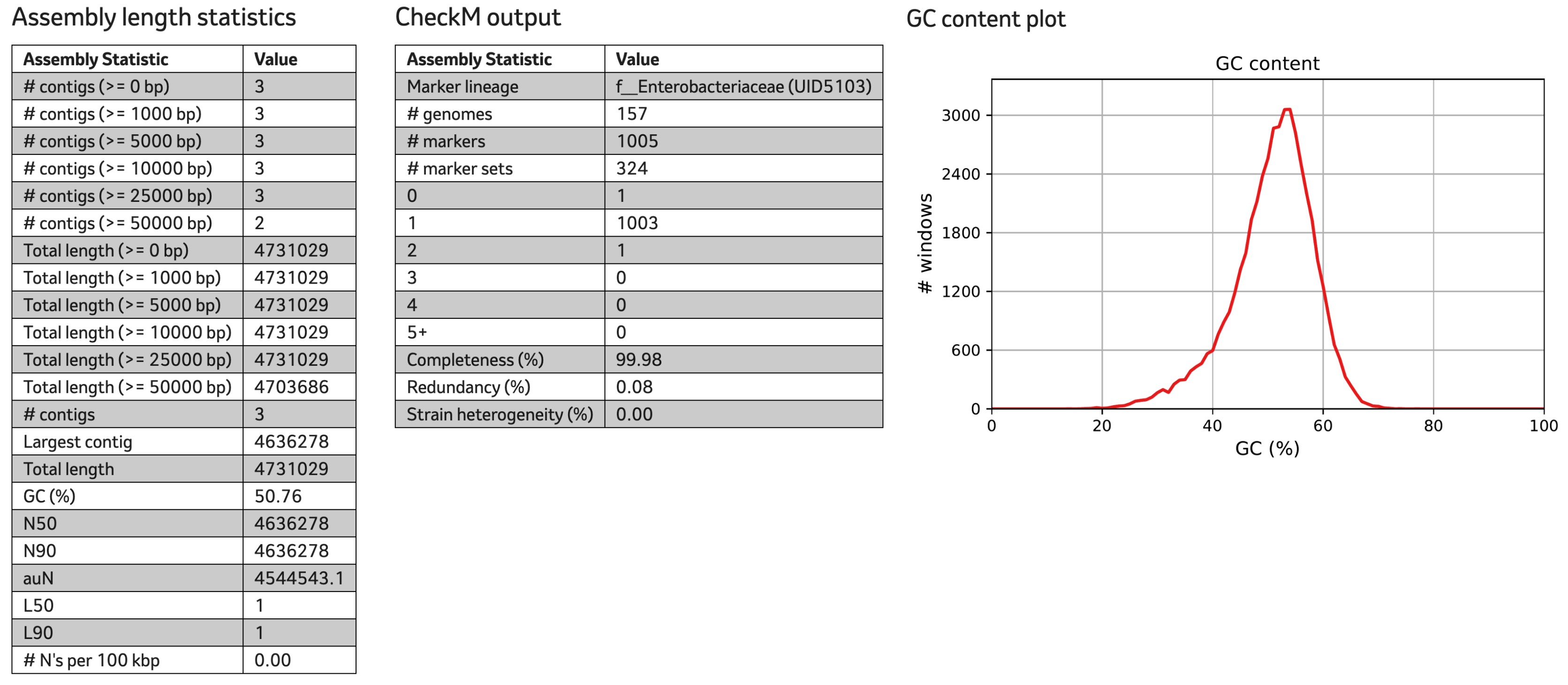
Parameters

COMPUTATION TYPE: Long Read Assembly

DOMAIN: Bacteria

- Bacteria
- Eukaryote

Figure 2. ViruScreen assembly workflow results summary



Assembly length statistics

Assembly Statistic	Value
# contigs (>= 0 bp)	3
# contigs (>= 1000 bp)	3
# contigs (>= 5000 bp)	3
# contigs (>= 10000 bp)	3
# contigs (>= 25000 bp)	3
# contigs (>= 50000 bp)	2
Total length (>= 0 bp)	4731029
Total length (>= 1000 bp)	4731029
Total length (>= 5000 bp)	4731029
Total length (>= 10000 bp)	4731029
Total length (>= 25000 bp)	4731029
Total length (>= 50000 bp)	4703686
# contigs	3
Largest contig	4636278
Total length	4731029
GC (%)	50.76
N50	4636278
N90	4636278
auN	4544543.1
L50	1
L90	1
# N's per 100 kbp	0.00

CheckM output

Assembly Statistic	Value
Marker lineage	f__Enterobacteriaceae (UID5103)
# genomes	157
# markers	1005
# marker sets	324
0	1
1	1003
2	1
3	0
4	0
5+	0
Completeness (%)	99.98
Redundancy (%)	0.08
Strain heterogeneity (%)	0.00

GC content plot

Long-Read Whole-Genome Assembly in ViruScreen

- The ViruScreen long-read genome assembly workflow leverages Oxford Nanopore Technologies (ONT) long-read sequencing.
- ONT long-reads span thousands to millions of bases, enabling single-read coverage of large genomic regions in both bacteria and yeast.
- Long reads allow reconstruction of complete or near-complete bacterial and yeast genomes with minimal fragmentation.
- The ViruScreen long-read genome assembly workflow includes quality filtering, trimming and sub-sampling of long-reads, followed by de novo genome assembly, and genome quality and completeness assessment (**Figure 3**).
- Our bacterial assemblies have near-identical (>99.8%) average nucleotide identity (ANI) to their reference genomes, high completeness, and low redundancy (**Table 1, Figure 4**).

Figure 3. ViruScreen microbial genome assembly workflow

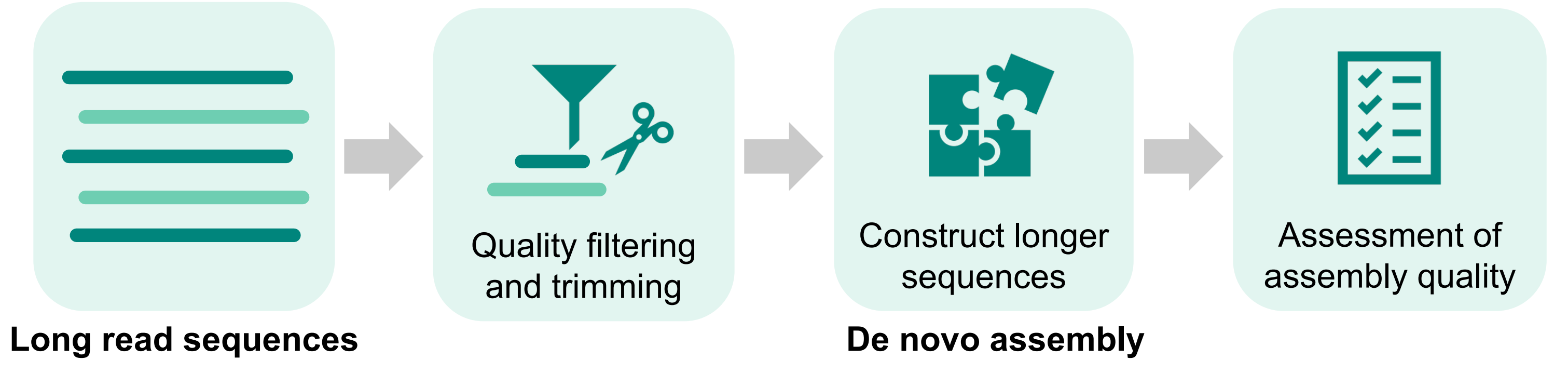

Quality filtering and trimming
Construct longer sequences
Assessment of assembly quality
Long read sequences
De novo assembly

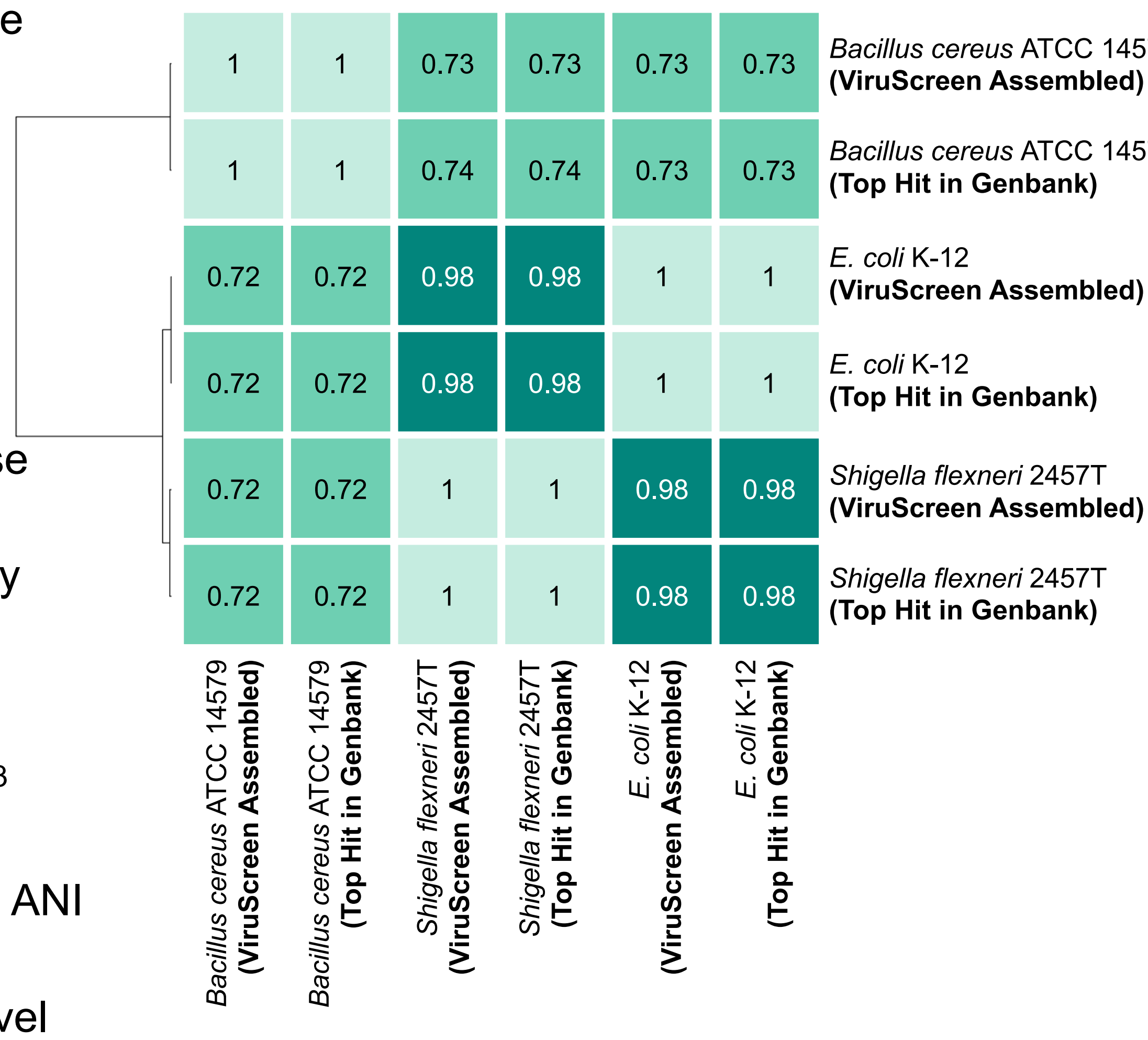
Table 1. ViruScreen bacterial genome assembly quality evaluation

Strain	Identity to Reference	Completeness	Redundancy
E. coli K-12	99.99%	99.98%	0.08%

Bacterial Identification: *E. coli* and *S. flexneri*

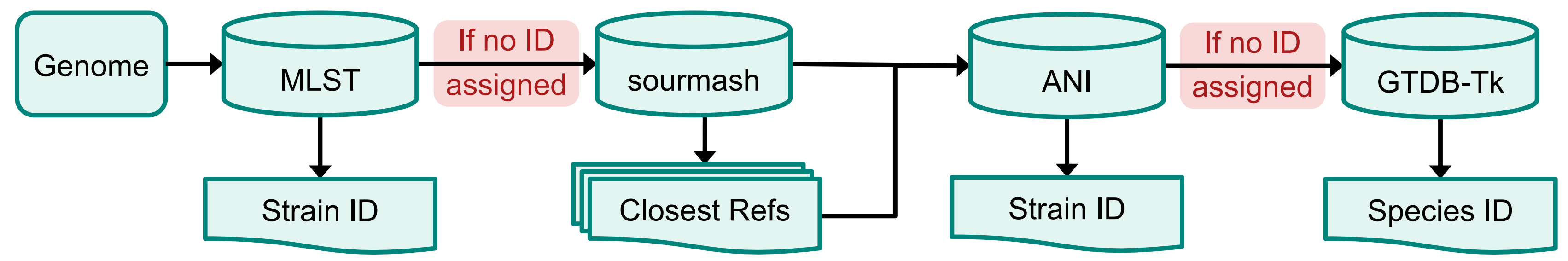
- Our workflow assigns taxonomy to bacterial genomes using three possible methods (**Figure 5**):
  1. Multi-locus sequence typing (MLST)
  2. Rapid comparison to database of genome sourmash[1] sketches and subsequent ANI
  3. Genome Taxonomy Database Toolkit (GTDB-Tk)[2]
- *E. coli* and *S. flexneri* are closely related with highly conserved genomic content; it has been suggested that they should be classified as the same species.[3]
- Our workflow differentiates *S. flexneri* and *E. coli* despite 98% ANI (**Figure 4**).
- Investigations into sub-strain-level differentiation using ANI are ongoing.

Figure 4. ANI of *E. coli* and *S. flexneri* genomes



	Bacillus cereus ATCC 14579 **(ViruScreen Assembled)**	*Bacillus cereus* ATCC 14579 **(Top Hit in Genbank)**	*Shigella flexneri* 2457T **(ViruScreen Assembled)**	*Shigella flexneri* 2457T **(Top Hit in Genbank)**	*E. coli* K-12 **(ViruScreen Assembled)**	*E. coli* K-12 **(Top Hit in Genbank)**
Bacillus cereus ATCC 14579 **(ViruScreen Assembled)**	1	1	0.73	0.73	0.73	0.73
Bacillus cereus ATCC 14579 **(Top Hit in Genbank)**	1	1	0.74	0.74	0.73	0.73
E. coli K-12 **(ViruScreen Assembled)**	0.72	0.72	0.98	0.98	1	1
E. coli K-12 **(Top Hit in Genbank)**	0.72	0.72	0.98	0.98	1	1
Shigella flexneri 2457T **(ViruScreen Assembled)**	0.72	0.72	1	1	0.98	0.98
Shigella flexneri 2457T **(Top Hit in Genbank)**	0.72	0.72	1	1	0.98	0.98

Figure 5. Bacterial whole-genome identification workflow



Yeast Identification: *S. cerevisiae* and *K. phaffii*

- Yeast whole-genome assembly, an ongoing addition to ViruScreen, is complicated by their relatively larger genome size, repetitive low-complexity sequences, and potential for more than one complete set of chromosomes in a cell.
- Multiple copies of the genome complicate whole-genome assembly, as first-pass assemblies are haploid and therefore represent a composite, or consensus, of different haplotypes.
- We evaluated our yeast assemblies using completeness and redundancy of single-copy gene sets shared by most fungi (**Table 2**).

Table 2. Yeast assembly quality evaluation

Strain	Completeness	Redundancy
Saccharomyces cerevisiae	75.1%	6.1%
Komagataella phaffii	82.3%	1.2%

- Despite low completeness scores we were able to identify *S. cerevisiae* and *K. phaffii* species using our genome assemblies, and unassembled long-reads in the case of *K. phaffii*, using the sourmash genome identification tool (**Table 3**).

Table 3. Yeast long-read and assembly taxonomic identification

	Long-Reads Only	Genome Assembly
Strain	**Taxonomic ID**	
Saccharomyces cerevisiae	None	*Saccharomyces cerevisiae*
Komagataella phaffii	*Komagataella phaffii*	*Komagataella phaffii*

Conclusions

- ViruScreen is a powerful, GMP-ready platform that makes high-resolution microbial identification of adventitious agents accessible to non-bioinformaticians.
- Our workflows leverage long-read sequencing to rapidly distinguish closely related *E. coli* and *S. flexneri* strains for precise bacterial identification, and taxonomically identify yeasts *S. cerevisiae* and *K. phaffii*.
- Together, these capabilities reduce time and cost of adventitious agent identification in vaccines and biologics samples.

References

1. Irber, Luiz, et al. "Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers." *BioRxiv* (2022): 2022-01.
2. Chaumeil, Pierre-Alain, et al. "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." *Bioinformatics* (2020): 1925-1927.
3. Parks, Donovan H., et al. "Reclassification of Shigella species as later heterotypic synonyms of Escherichia coli in the Genome Taxonomy Database." *BioRxiv* (2021): 2021-09.

Presented at PDA Pharmaceutical Microbiology Conference 2025; Washington, DC, USA and Online; October 27-29, 2025.

Disclosures & Acknowledgments

Funding for this research was provided by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

Contact information

Contact the authors at zachary.pimentel@merck.com or andrea.watson@merck.com