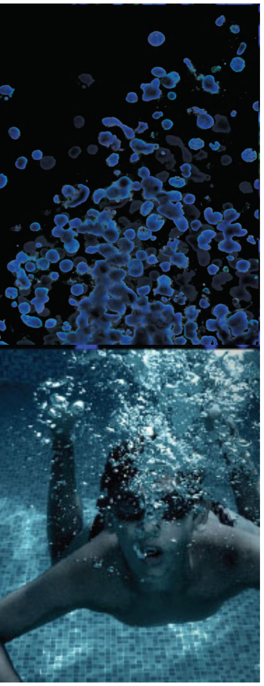


Prevent Analysis Variability by Using NGS-authenticated Microbial Genomes — Shift from Consensus to Discernible

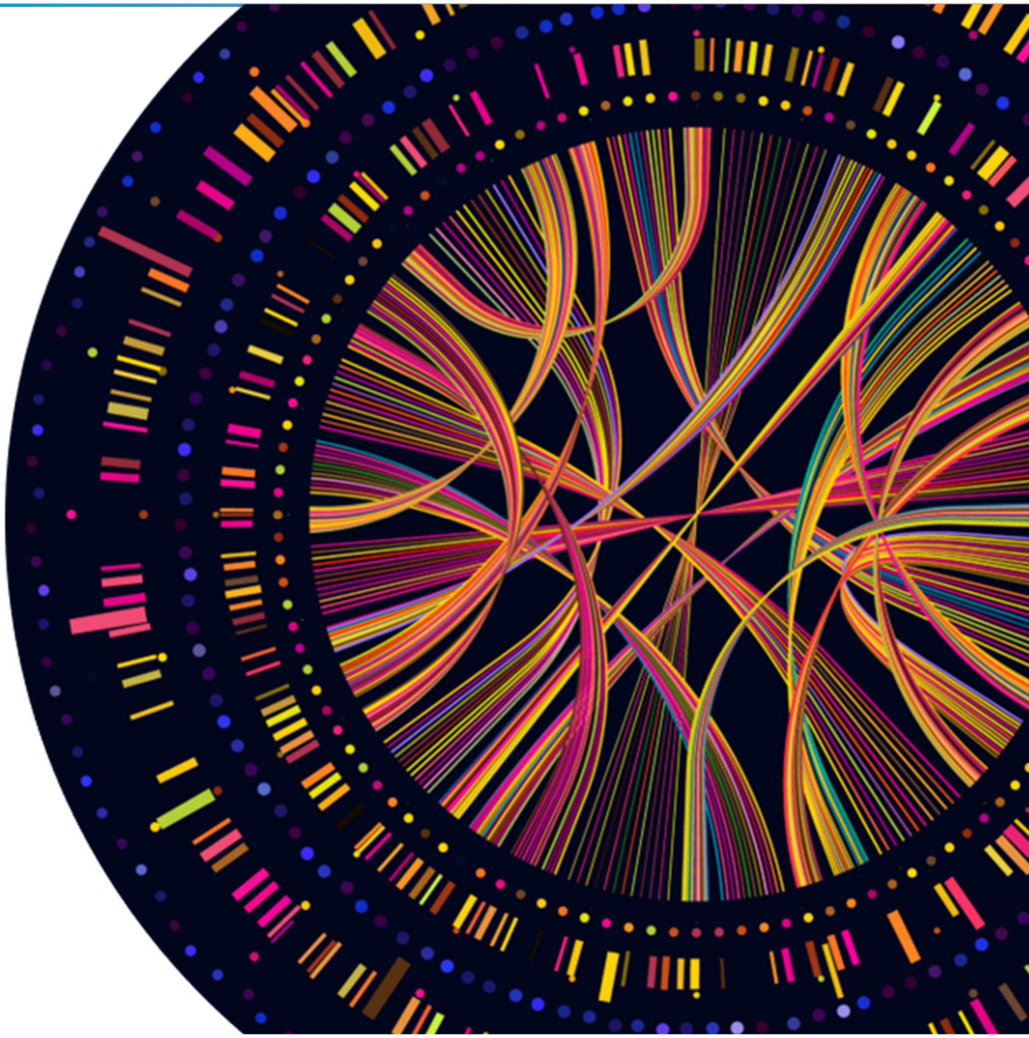
Andrew Frank, Scientist
Bioinformatician, ATCC

Credible Leads to Incredible™

About ATCC

- Founded in 1925, ATCC is a not-for-profit organization with HQ in Manassas, VA, and an R&D and Services center in Gaithersburg, MD
- World's largest, most diverse biological materials and information resource for microbes – the "*gold standard*"
- Innovative R&D company featuring gene editing, microbiome, NGS, and advanced models
- cGMP biorepository
- Partner with government, industry, and academia
- Leading global supplier of authenticated cell lines, microorganisms, and molecular standards
- Sales and distribution in 150 countries, 19 international distributors
- Talented team of 450+ employees, over one-third with advanced degrees

Overview



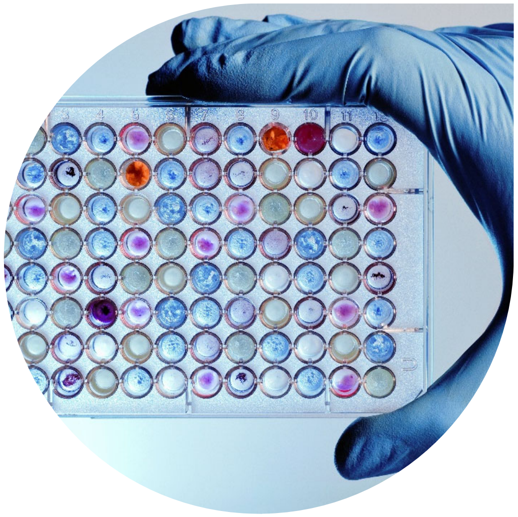
Enhanced Authentication Initiative

Enriching the characterization of ATCC's biological collections through whole-genome sequencing

ATCC Genome Portal

Providing reference-grade genomes matched to authenticated ATCC biological materials

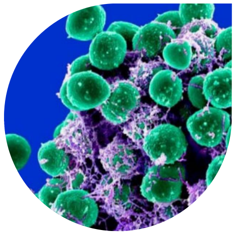
Science is Changing



The Reproducibility Crisis

The Genomics Revolution

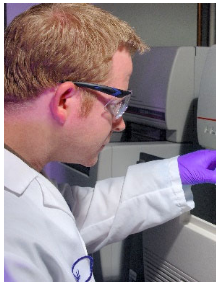
Modern Bacterial Strain Authentication Approaches



ATCC® 12228™
Staphylococcus epidermidis

CERTIFICATE OF ANALYSIS

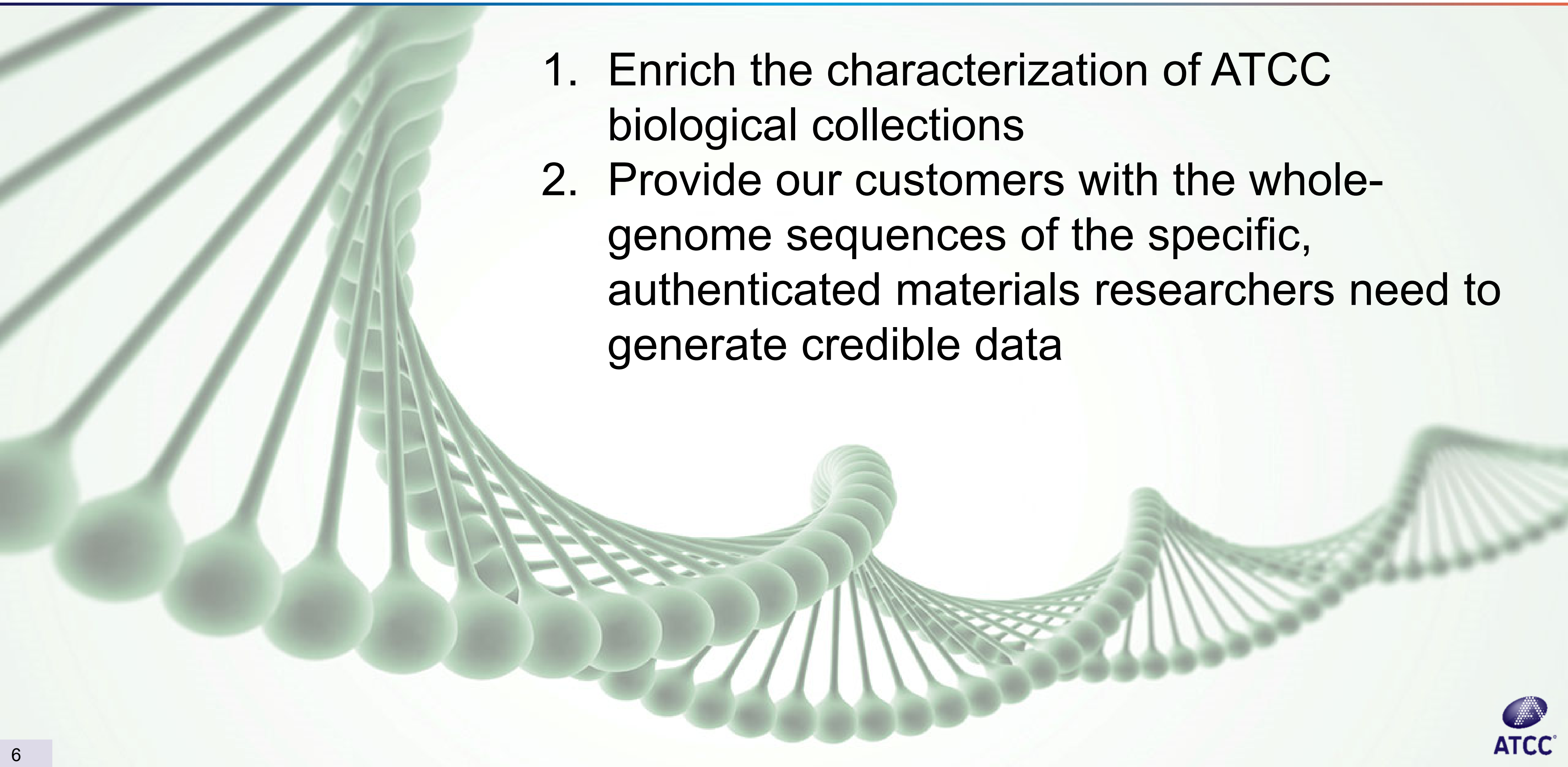
1. ATCC polyphasic testing
    - Phenotypic testing
    - Genotypic testing
    - Functional analyses



2. Sequence the genome yourself
    - Expensive but highly accurate
    - Must be done as close to receipt of the item as possible, or else not reproducible

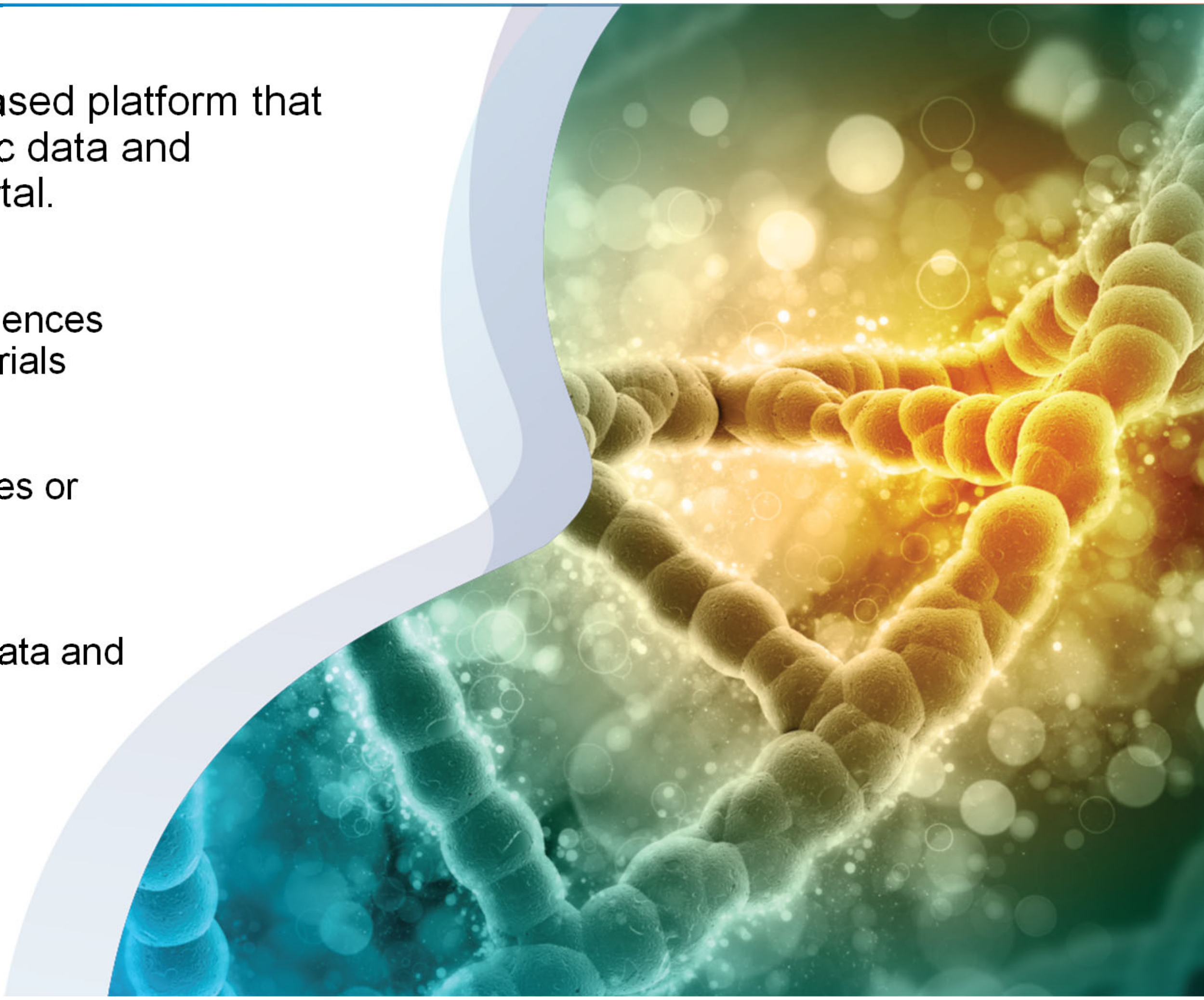
3. "Crowd-source" the genome
    - Free but risky
    - Requires bioinformatics expertise
    - Issues with quality, traceability, etc.

The Enhanced Authentication Initiative



1. Enrich the characterization of ATCC biological collections
2. Provide our customers with the whole-genome sequences of the specific, authenticated materials researchers need to generate credible data

ATCC Genome Portal



The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal.

- Download whole-genome sequences and annotations of ATCC materials
- Search for nucleotide sequences or genes within genomes
- View genome assembly metadata and quality metrics

genomes.atcc.org

Differentiating ATCC by Material and Practices

Material

Practices

Extract DNA from authenticated material

Sequence high-quality DNA

Assemble with short and long reads

Annotate the resultant genome

Validate strain designation

ATCC Bioinformatics Best Practices

Extract DNA with optimized, proprietary protocols to get the best quality input material
Avoid “garbage in, garbage out” data and results

Sequence on both Illumina® and Oxford Nanopore Technologies®
Impose strict quality control thresholds to save only highest quality reads per instrument

Combine data from both technologies to achieve high-quality, complete genomes
We perform hybrid genome assembly, leveraging the strengths of both platforms while avoiding biases and weaknesses of them individually

Annotate genomes to enable gene-level analyses
Provide users with a reliable annotation, so they can identify genes of interest rapidly to enable gene-specific research

Confirm strain identity against highly curated databases
We’ve partnered with One Codex, an industry leader in bioinformatic bacterial identification, to verify our strain designations by using *k*-mer and genomic distance approaches

ATCC Extraction Quality Control

ATCC® no.	Species	PicoGreen® (ng/µL)	$A_{260/}A_{280}$	DNA fragment size (range)**
8739DX™*	*Escherichia coli*	101.9	1.92	49.5 kb (1.5 – >60 kb)
13048DX™*	*Klebsiella aerogenes*	98.1	1.86	49.5 kb (1.6 – >60 kb)
11828DX™*	*Cutibacterium acnes*	197.7	1.84	29.8 kb (0.8 – >60 kb)
6538DX™*	*Staphylococcus aureus*	97.8	1.85	32.9 kb (2.7 – >60 kb)
BAA-2797DX™*	*Pseudomonas aeruginosa*	153.3	1.99	44.1 kb (1.1 – >60 kb)
824D-5™	*Clostridium acetobutylicum*	73.8	2.05	12.5 kb (4.6 – 57.8 kb)
6538D-5™	*Staphylococcus aureus*	37.1	2.00	26.2 kb (6.9 – >60 kb)
27774D-5™	*Desulfovibrio desulfuricans*	69.2	1.99	58.5 kb (13.3 – >60 kb)
11842D-5™	*Lactobacillus delbrueckii*	64.8	2.02	41.9 kb (6.1 – >60 kb)
15697D-5™	*Bifidobacterium longum*	76.2	1.95	51.3 kb (10.5 – >60 kb)

* New NGS-ready DNA product
** Main peak reported

NGS-Ready DNA Products Now Available

NGS-Ready DNA products (-DX) now available to customers

- ≥ 10 µg of DNA
- DNA fragment sizes ≥ 20 kb
- Launched with 50 of the most popular ATCC bacterial strains
- Check them out at **atcc.org/NGS-ready**

ATCC Bioinformatics Best Practices

Extract DNA with optimized, proprietary protocols to get the best quality input material
Avoid “garbage in, garbage out” data and results

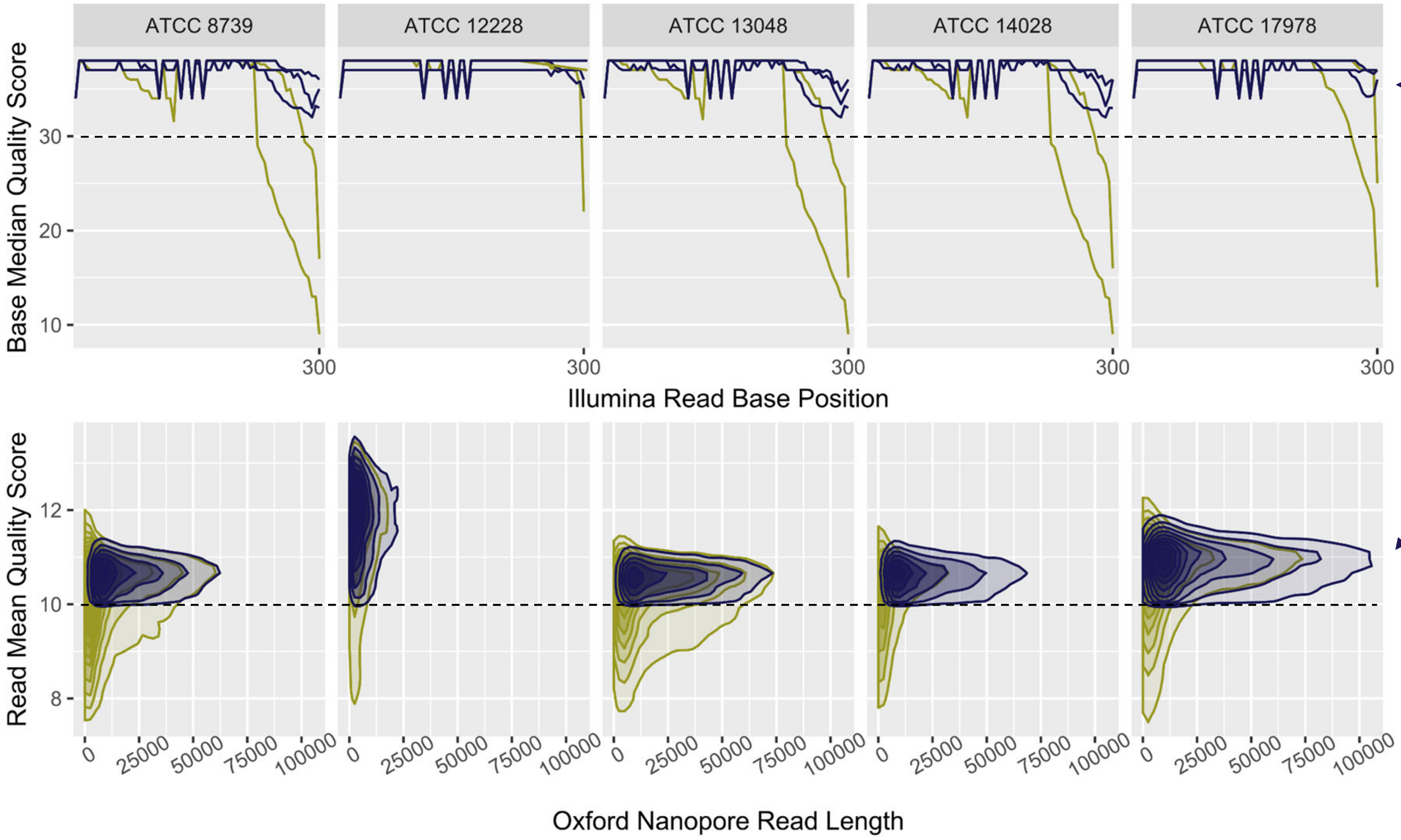
Sequence on both Illumina® and Oxford Nanopore Technologies®
Impose strict quality control thresholds to save only highest quality reads per instrument

Combine data from both technologies to achieve high-quality, complete genomes
We perform hybrid genome assembly, leveraging the strengths of both platforms while avoiding biases and weaknesses of them individually

Annotate genomes to enable gene-level analyses
Provide users with a reliable annotation, so they can identify genes of interest rapidly to enable gene-specific research

Confirm strain identity against highly curated databases
We’ve partnered with One Codex, an industry leader in bioinformatic bacterial identification, to verify our strain designations by using *k*-mer and genomic distance approaches

Sequencing QC – Read Trimming/Filtering



Only keep high-quality base calls

Only keep long, good quality reads

■ Quality controlled data

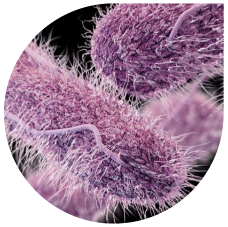
■ Raw data

Sequencing QC – Changing Conditions

Test	Species	ATCC® no.	Reference Genome	Analysis	# of SNPs	# of indels	Number of variants	Variant coverage
A	*Mycoplasma hominis*	23114™	GCA_000085865.1	Preparation 1	14	10	24	1042.1
				Preparation 2	14	10	24	900.0
	Cutibacterium acnes	11828™	GCA_000231215.1	Preparation 1	28	37	65	121.1
				Preparation 2	28	39	67	128.6
B	*Clostridium acetobutylicum*	824D-5™	GCA_000008765.1	Kit 1	171	55	226	95.9
				Kit 2	170	55	225	202.0
	Aeromonas hydrophila	7966D-5™	GCA_000014805.1	Kit 1	1	1	2	216.8
				Kit 2	1	1	2	203.0
C	*Escherichia coli*	700926™	GCA_000005845.2	Extraction 1	0	1	1	137.0
				Extraction 2	0	1	1	186.3
	Streptococcus pyogenes	19615™	GCA_000743015.1	Extraction 1	2	44	46	314.2
				Extraction 2	2	41	43	460.2

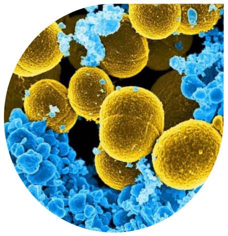
Test A: Same DNA sequenced using two different DNA preparations
Test B: Same DNA sequenced with two different library kits
Test C: Same strain extracted with two different methods

Application – NGS Proficiency Testing

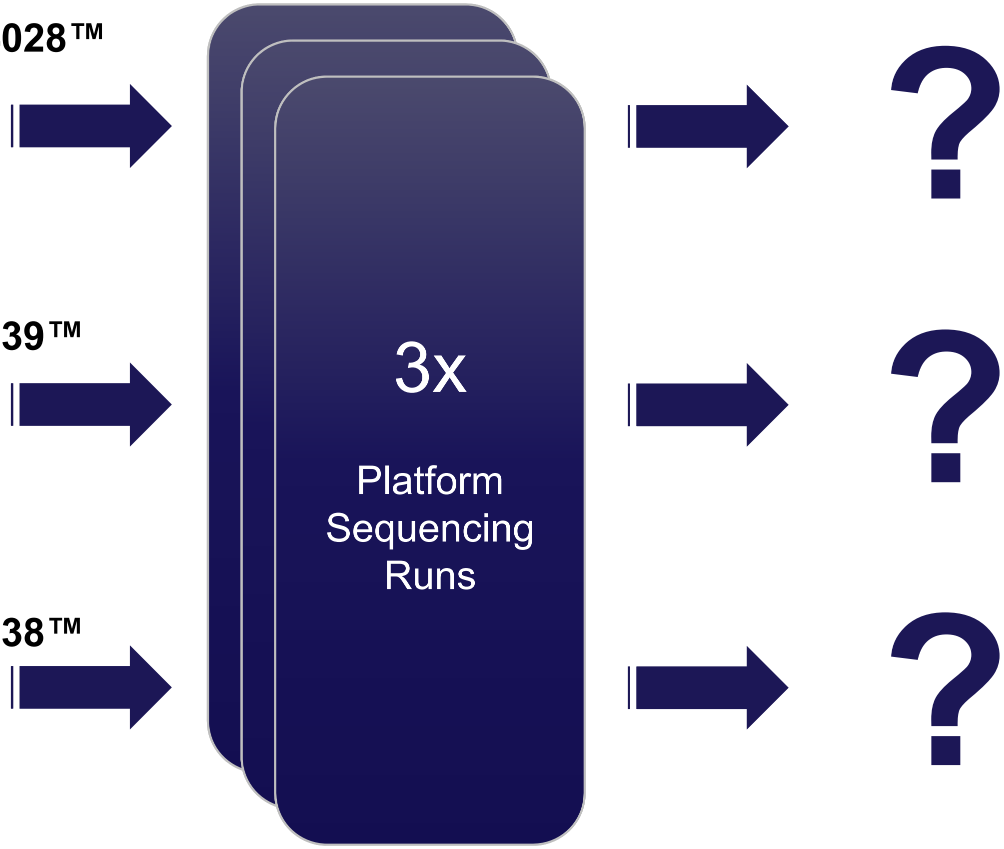


ATCC® CRM-14028™
Salmonella enterica

ATCC® CRM-8739™
Escherichia coli



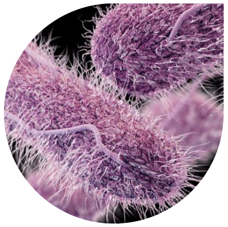
ATCC® CRM-6538™
Staphylococcus aureus



3x

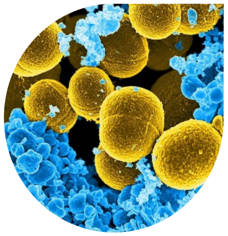
Platform Sequencing Runs

Application – NGS Proficiency Testing

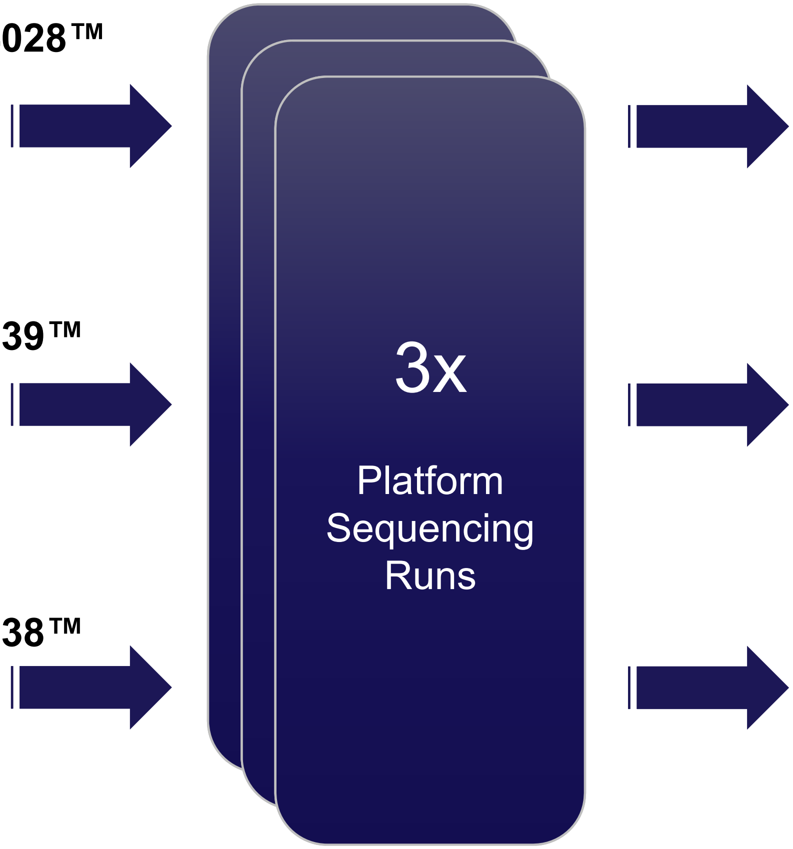


ATCC® CRM-14028™
Salmonella enterica

ATCC® CRM-8739™
Escherichia coli



ATCC® CRM-6538™
Staphylococcus aureus



3x

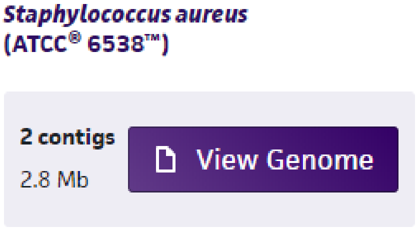
Platform Sequencing Runs

Salmonella enterica **(ATCC® 14028™)**

2 contigs
5.0 Mb

Escherichia coli **(ATCC® 8739™)**

1 contig
4.7 Mb



Staphylococcus aureus **(ATCC® 6538™)**

2 contigs
2.8 Mb

ATCC Bioinformatics Best Practices

Extract DNA with optimized, proprietary protocols to get the best quality input material
Avoid “garbage in, garbage out” data and results

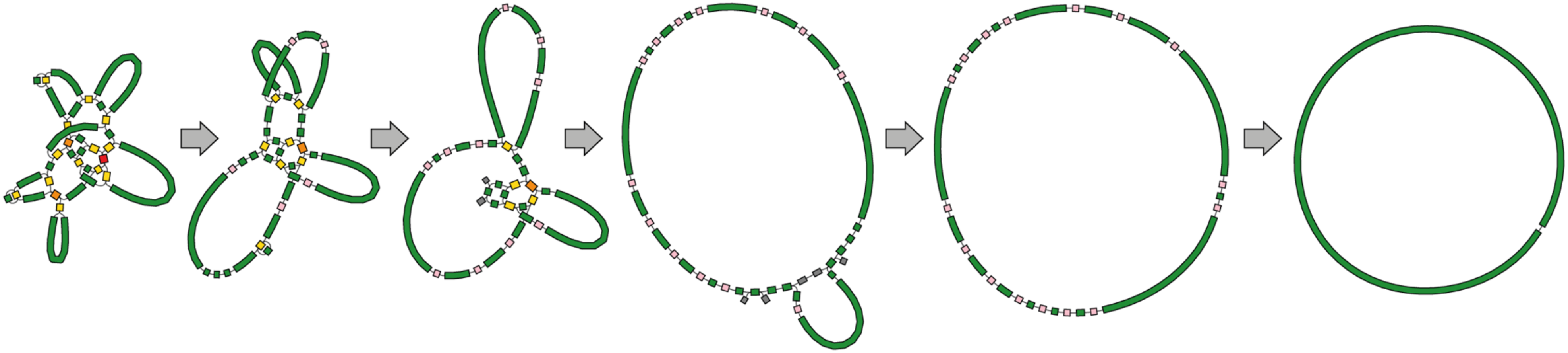
Sequence on both Illumina® and Oxford Nanopore Technologies®
Impose strict quality control thresholds to save only highest quality reads per instrument

Combine data from both technologies to achieve high-quality, complete genomes
We perform hybrid genome assembly, leveraging the strengths of both platforms while avoiding biases and weaknesses of them individually

Annotate genomes to enable gene-level analyses
Provide users with a reliable annotation, so they can identify genes of interest rapidly to enable gene-specific research

Confirm strain identity against highly curated databases
We’ve partnered with One Codex, an industry leader in bioinformatic bacterial identification, to verify our strain designations by using *k*-mer and genomic distance approaches

Hybrid Genome Assembly



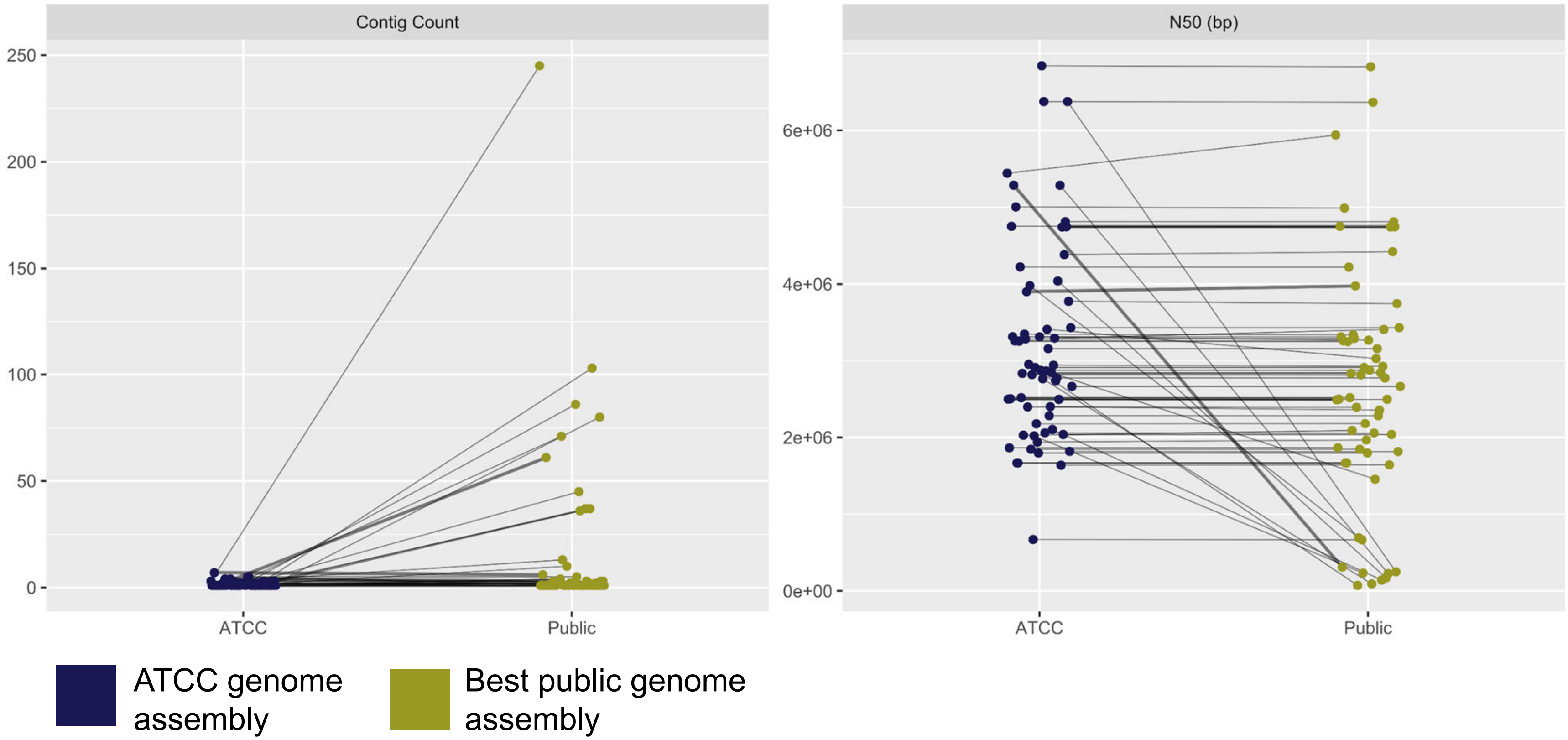
Illumina only genome assembly

Long reads mapped to a tangled region creates a resolved bridge
Successively applying bridges resolve the structure of the genome

Completed hybrid assembly

Image reproduced from https://github.com/rrwick/Unicycler

ATCC's Assemblies Substantially Improve Upon Public Assemblies

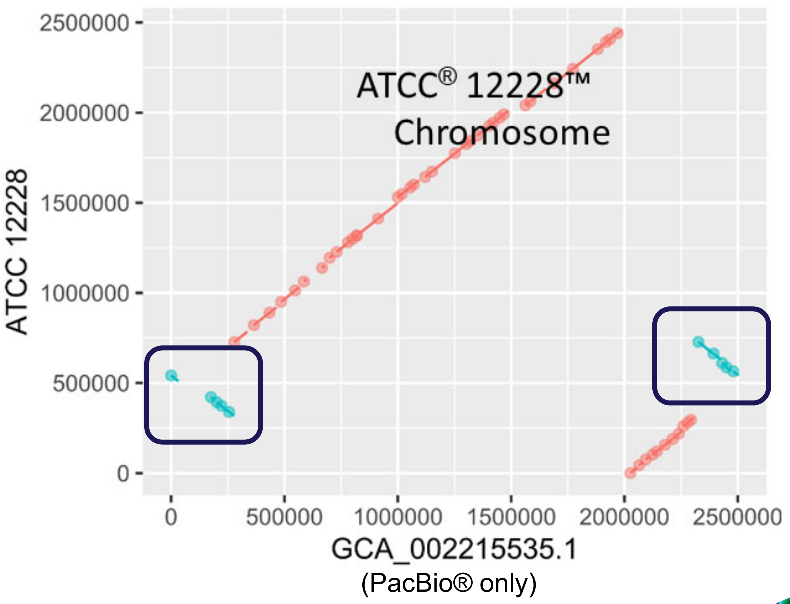


Downward trend in contig count and **upward** trend in N50 indicate ATCC is producing more complete genomes than the best publicly available alternative genome

ATCC’s Assemblies Substantially Improve Upon Public Assemblies

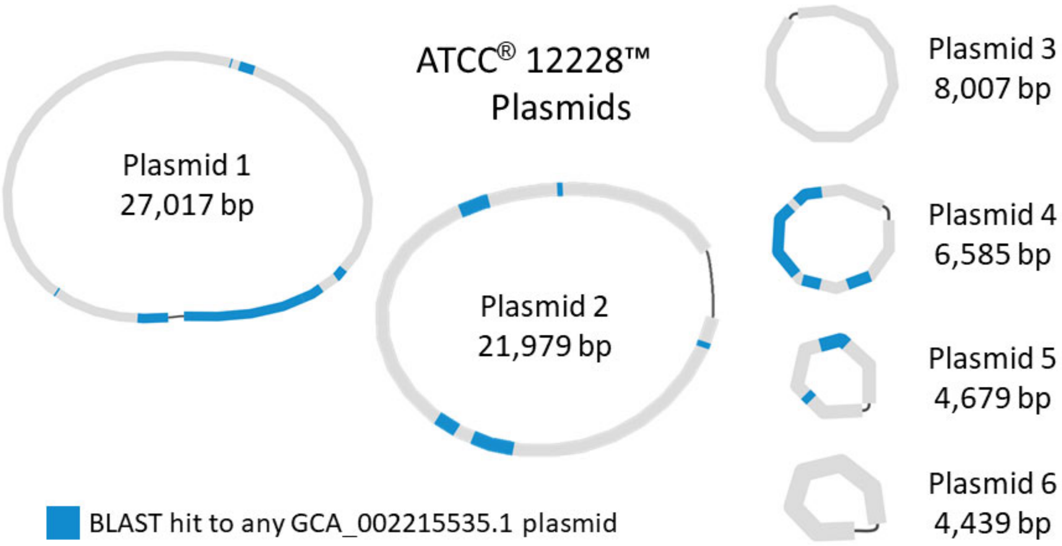
Structural Differences

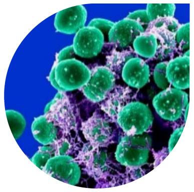
Major structural inversion



Plasmid Content Differences

No shared plasmid content





ATCC® 12228™
Staphylococcus epidermidis

ATCC Bioinformatics Best Practices

Extract DNA with optimized, proprietary protocols to get the best quality input material
Avoid “garbage in, garbage out” data and results

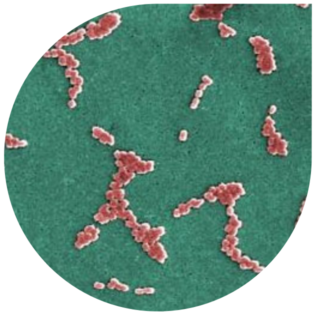
Sequence on both Illumina® and Oxford Nanopore Technologies®
Impose strict quality control thresholds to save only highest quality reads per instrument

Combine data from both technologies to achieve high-quality, complete genomes
We perform hybrid genome assembly, leveraging the strengths of both platforms while avoiding biases and weaknesses of them individually

Annotate genomes to enable gene-level analyses
Provide users with a reliable annotation, so they can identify genes of interest rapidly to enable gene-specific research

Confirm strain identity against highly curated databases
We’ve partnered with One Codex, an industry leader in bioinformatic bacterial identification, to verify our strain designations by using *k*-mer and genomic distance approaches

Application – Variation in AMR Genes



ATCC® BAA-2803™
Acinetobacter baumannii
Global Priority Superbug

- ATCC hosts 57 strains of WHO priority pathogens – the **Global Priority Superbug** collection
  - Genomes being made available on the ATCC Genome Portal
  - Rich source data
  - Susceptibility data including MIC values and susceptibility profiles for targeted drugs
- More info at **atcc.org/globalprioritysuperbugs**

- In *A. baumannii,* mutations of *gyrA* can confer high levels of resistance to fluoroquinolone antibiotics (i.e. Ciprofloxacin)
- The ATCC Genome Portal can rapidly search for sequences of interest within our genome collection, up to 4000X faster than BLAST

ATCC Bioinformatics Best Practices

Extract DNA with optimized, proprietary protocols to get the best quality input material
Avoid “garbage in, garbage out” data and results

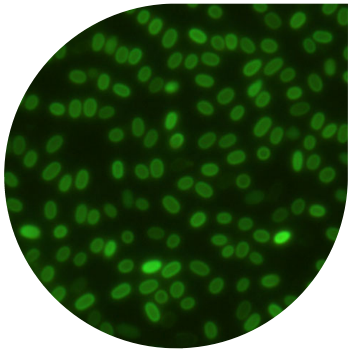
Sequence on both Illumina® and Oxford Nanopore Technologies®
Impose strict quality control thresholds to save only highest quality reads per instrument

Combine data from both technologies to achieve high-quality, complete genomes
We perform hybrid genome assembly, leveraging the strengths of both platforms while avoiding biases and weaknesses of them individually

Annotate genomes to enable gene-level analyses
Provide users with a reliable annotation, so they can identify genes of interest rapidly to enable gene-specific research

Confirm strain identity against highly curated databases
We’ve partnered with One Codex, an industry leader in bioinformatic bacterial identification, to verify our strain designations by using *k*-mer and genomic distance approaches

Enhanced Authentication Reveals Hidden Biodiversity



- Taxonomy of *Bacillus* continues to be rapidly updated this decade
- Sequenced and assembled genomes for 61 of ATCC's *Bacillus cereus* and *Bacillus subtilis* group strains, some of our oldest strains
- Measured genomic distance to type strains of all taxonomically valid *Bacillus*
  - **12** items are being moved into *Bacillus* species described only in the last decade
  - **1** item is being moved into an older existing *Bacillus* species
  - **9** items are being moved to taxonomically undescribed species

Application – Ascertain Strain Identity

Type Strain

ATCC® 33397™
Streptococcus anginosus

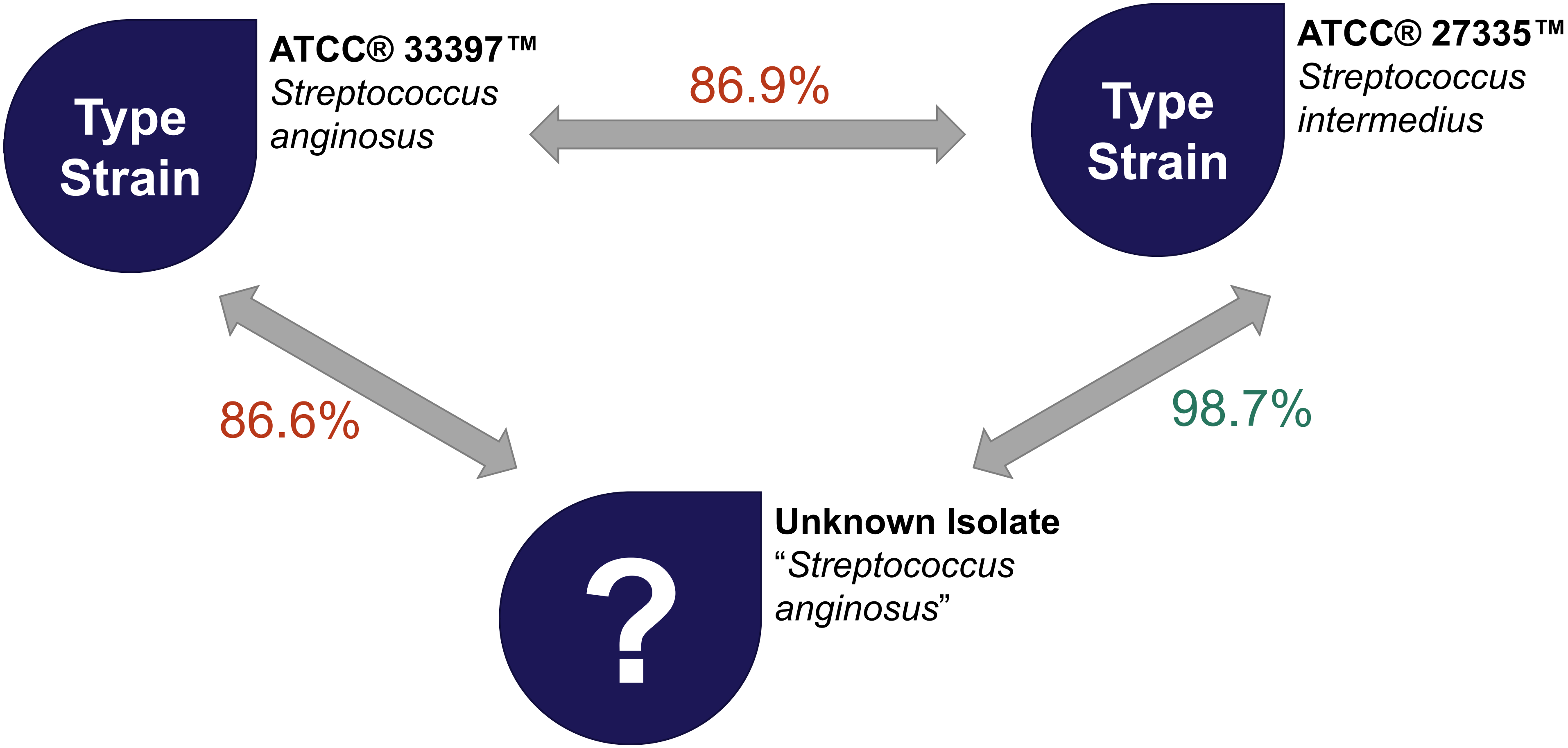
Type Strain

ATCC® 27335™
Streptococcus intermedius

?

Unknown Isolate
“*Streptococcus anginosus*”

Application – Ascertain Strain Identity



Type Strain

ATCC® 33397™
Streptococcus anginosus

86.9%

Type Strain

ATCC® 27335™
Streptococcus intermedius

86.6%

98.7%

?

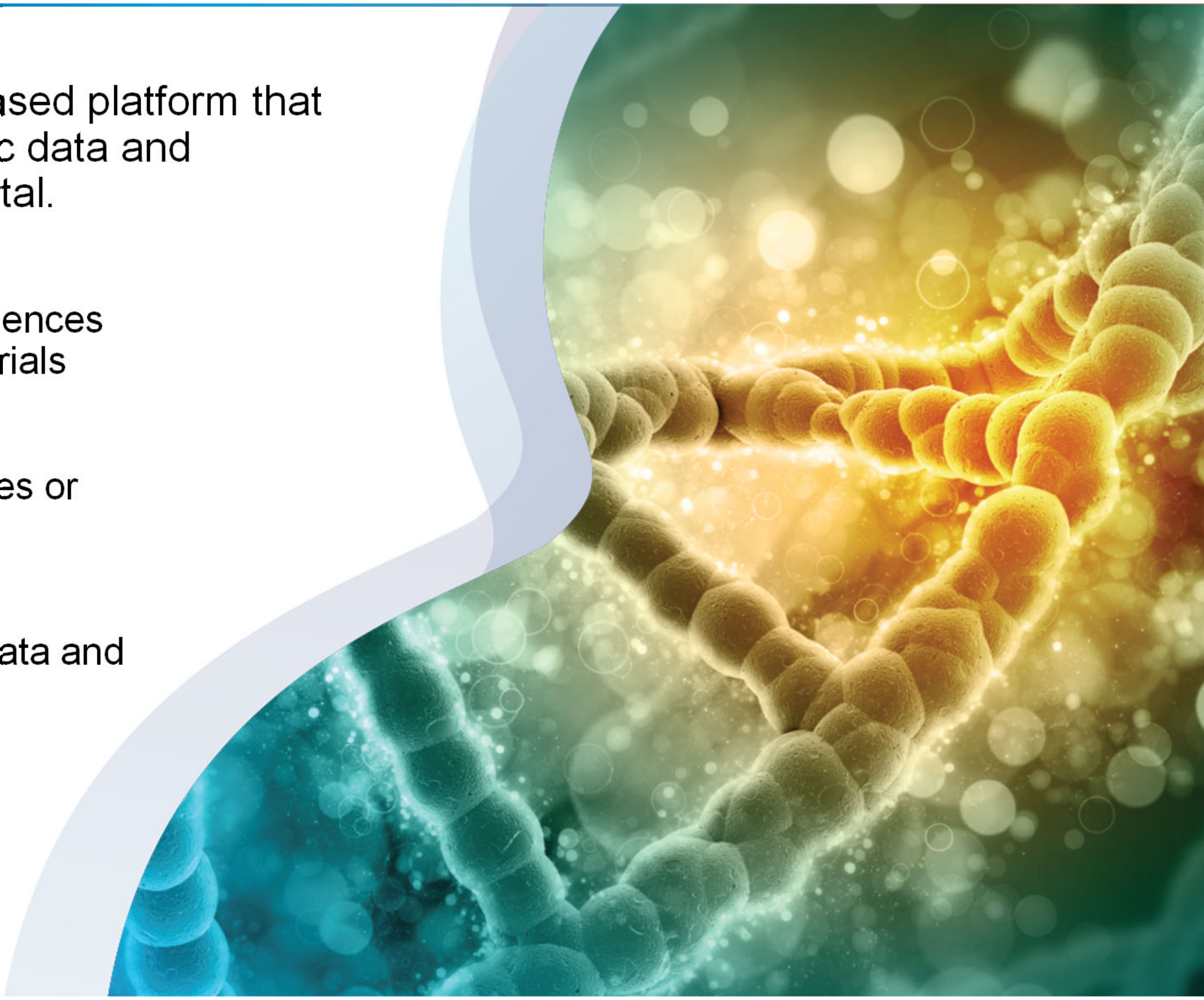
Unknown Isolate
“*Streptococcus anginosus*”

ATCC Genome Portal

The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal.

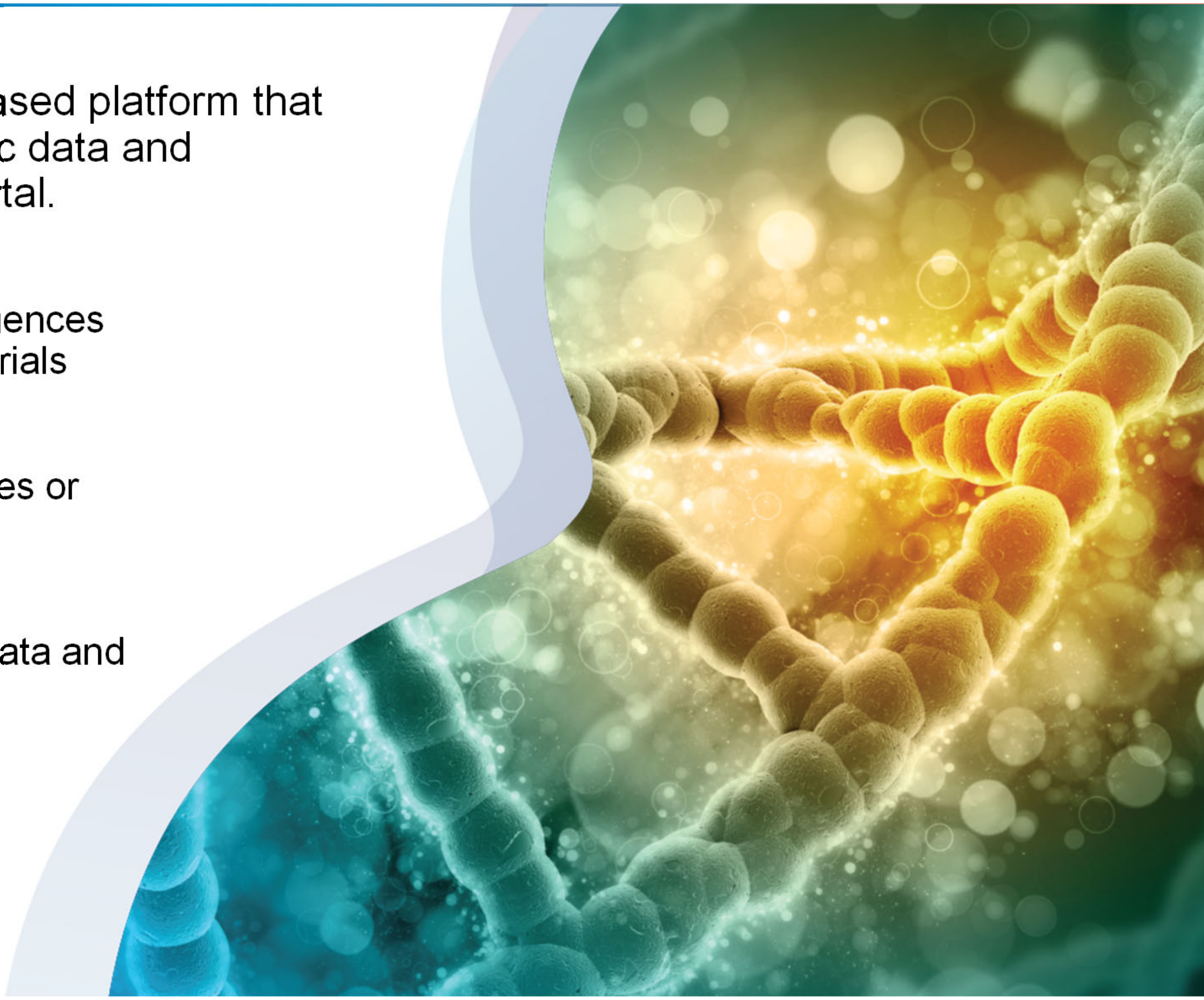
- Download whole-genome sequences and annotations of ATCC materials
- Search for nucleotide sequences or genes within genomes
- View genome assembly metadata and quality metrics

genomes.atcc.org



Thank you

ATCC Genome Portal



The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal.

Download whole-genome sequences and annotations of ATCC materials

Search for nucleotide sequences or genes within genomes

View genome assembly metadata and quality metrics

genomes.atcc.org

© 2019 American Type Culture Collection. The ATCC trademark and trade name, and any other trademarks listed in this publication are trademarks owned by the American Type Culture Collection unless indicated otherwise.