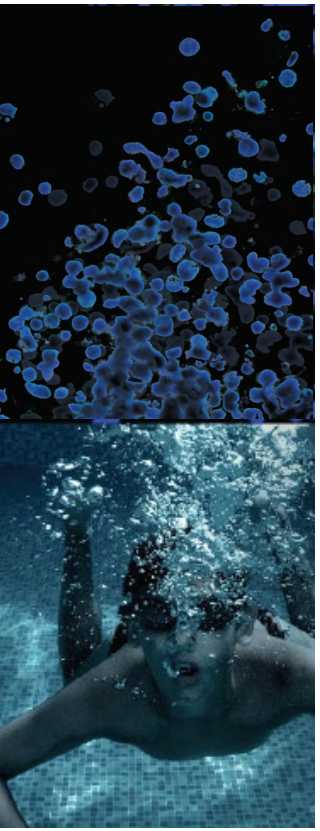
Advancing Authentication through Credible Standards and Robust Next-generation Sequencing Workflows



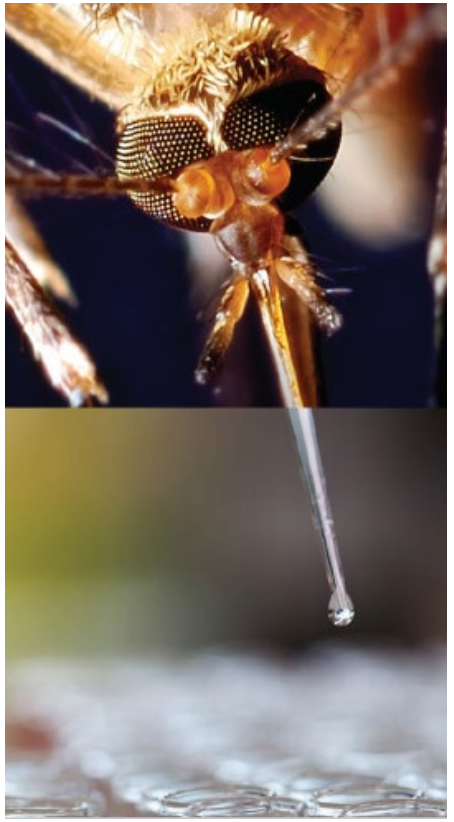
Briana Benton, BS
Technical Manager, ATCC

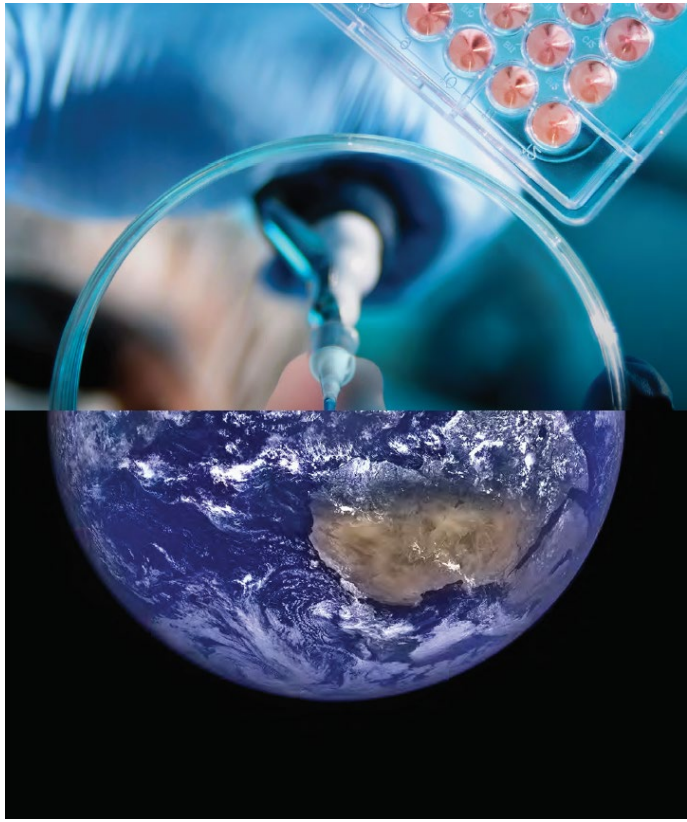
Sam Minot, PhD
Scientist, Fred Hutchinson Cancer Research Center

Nick Greenfield, MA
Founder and CEO, One Codex

Marco Riojas, PhD
Scientist, ATCC

Credible Leads to InCredible™

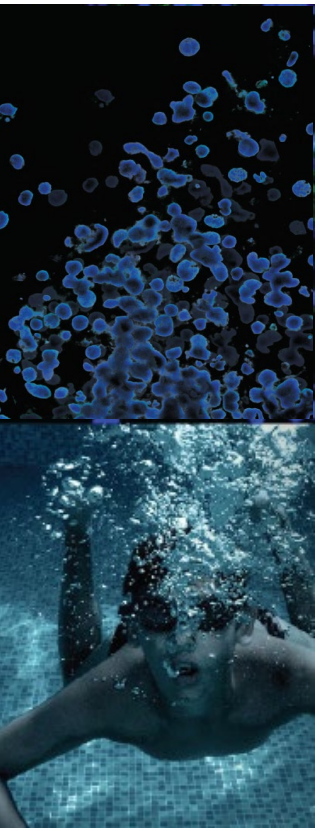


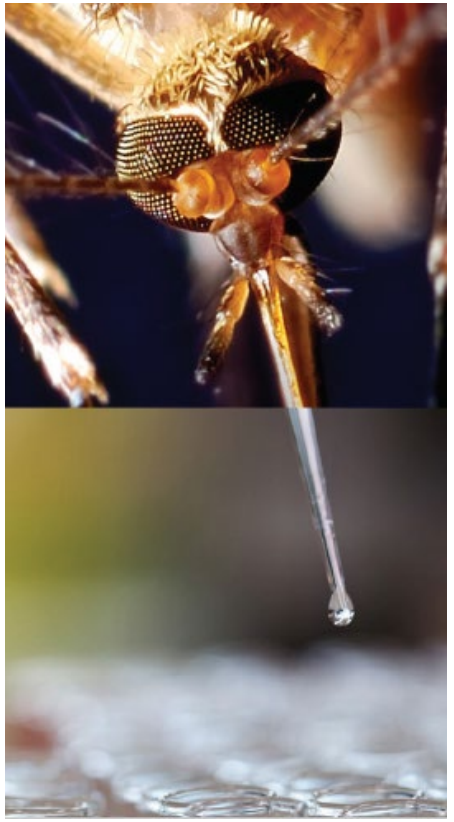


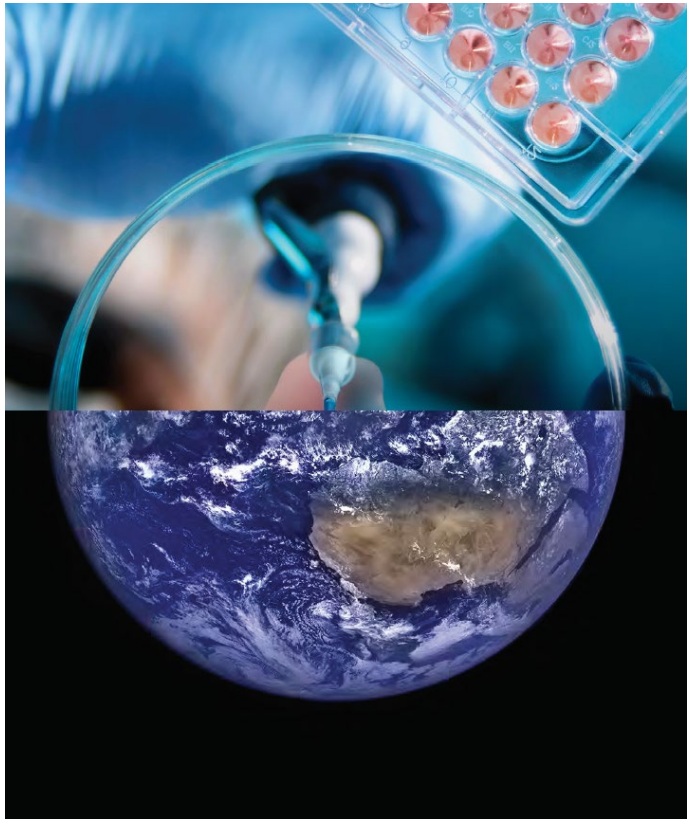
Making Sense Out of Microbiome Data – The Importance of Standards

Briana Benton, BS
Technical Manager, ATCC

Credible Leads to Incredible™



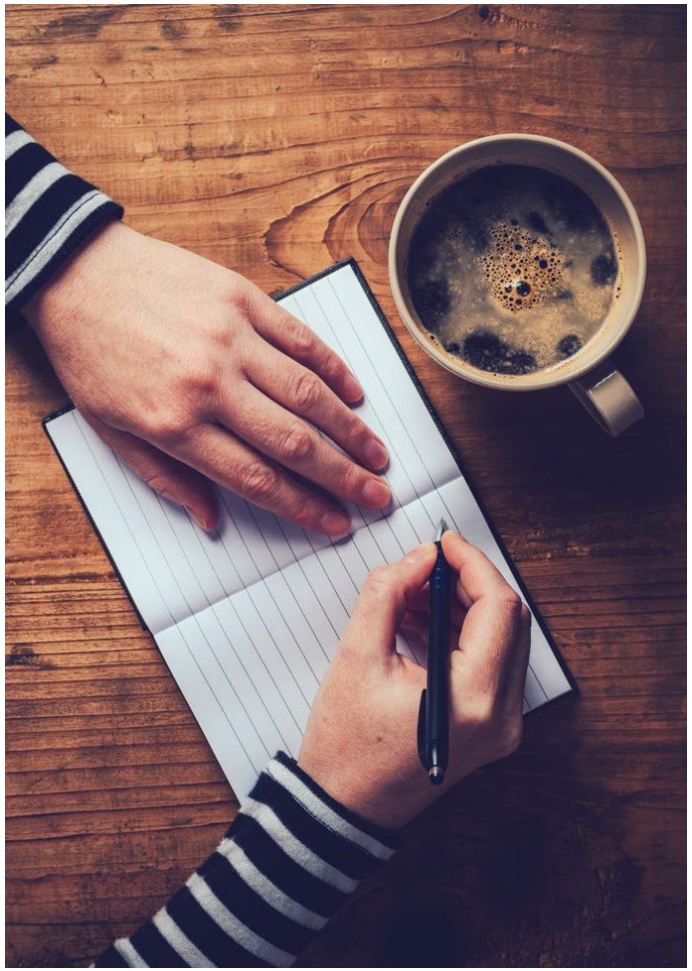




Agenda

- ✓ Challenges in microbiome analysis and development of standards
- ✓ The ATCC® Microbiome Standards portfolio and upcoming new products
- ✓ Applications of standards in microbiome research

- ✗ Microbiome assay development
- ✗ Show the best data
- ✗ Recommend any specific assay, kit, protocol, or instrument



Microbiome Research

The microbiome field is rapidly moving toward translational research pertinent to human health and disease, therapeutics, and personalized medicine

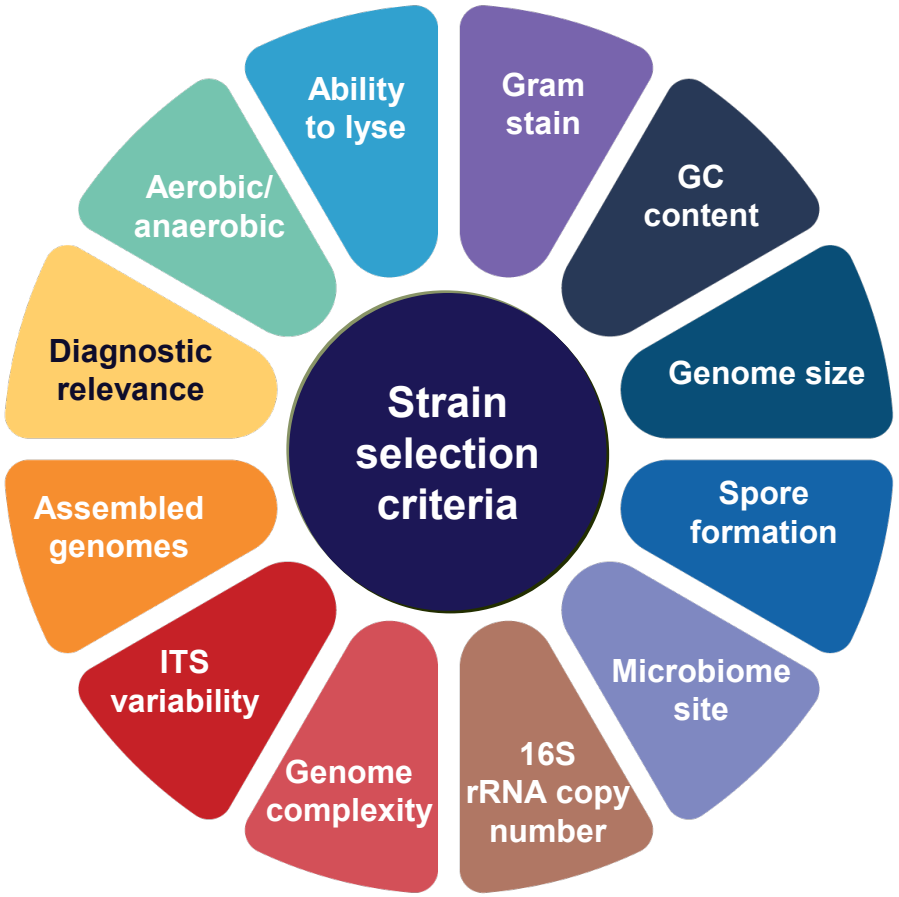
ATCC®

Challenges in Microbiome Research

MICROBIOME WORKFLOW

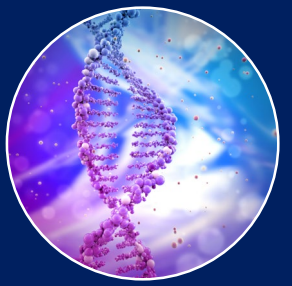
- Sample Collection
  - Storage
  - Handling
  - Processing
- DNA Extraction
  - Cell lysis
  - DNA/RNA recovery
  - Quality & quantity
- 16S / Shotgun
  - Amplicon vs non-amplicon
  - Choice of primers
  - Library preparation
- Library Preparation
- Sequencing
  - Platform
  - Chemistry
  - Depth
- Data Analysis
  - Read quality
  - Algorithm
  - Database

Development of Mock Microbial Communities



Whole Cell Standards

- Authenticated ATCC cultures
- Growth and image cytometry cell counting
- Mixed in even proportion based cell numbers cells
- Storage at 4°C

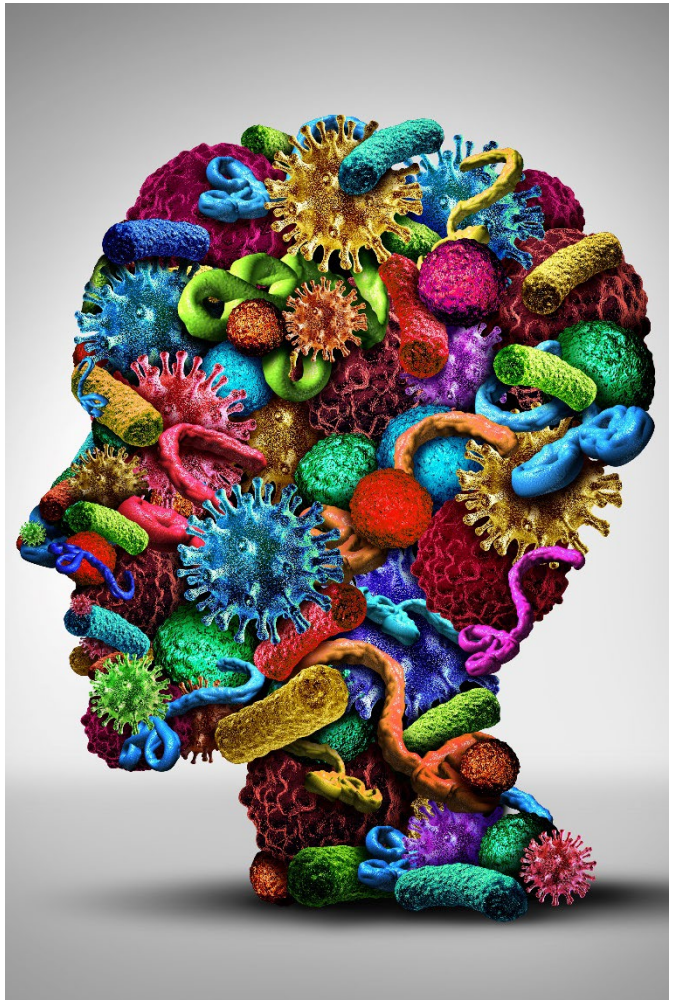


Genomic DNA Standards

- Authenticated ATCC nucleic acids
- Fluorescent dye-based quantification
- Mixed in even proportions based genome copy number
- Storage at -20°C

Assay development, optimization, verification, and quality control

ATCC® Microbiome Standards Portfolio

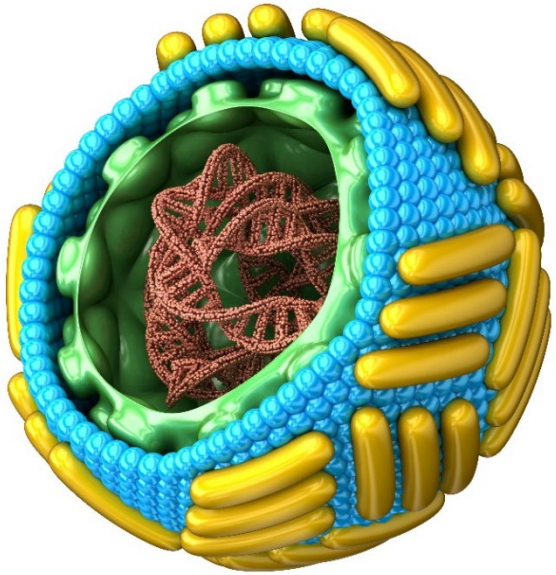


Preparation	ATCC® Catalog No.	Number of Organisms	Composition	Complexity	Importance
Genomic DNA	MSA-1000™	10	Even	Medium	Standards for assay development and optimization
	MSA-1001™	10	Staggered	Medium	
	MSA-1002™	20	Even	High	
	MSA-1003™	20	Staggered	High	
Whole cell	MSA-2003™	10	Even	Medium	
	MSA-2002™	20	Even	High	
Genomic DNA	MSA-4000™	11	Staggered	Medium	NGS-based pathogen detection
Genomic DNA	MSA-3000™	6	Even	Low	Environmental studies
	MSA-3001™	10	Even	Medium	
	MSA-3002™	10	Staggered	Medium	

Site-specific Microbiome Standards

Standard	Preparation	ATCC® Catalog No.	Number of Organisms	Importance
Oral	Whole cell	MSA-2004™	6	• Mock microbial communities representing the oral, skin, gut, and vaginal microbiomes • Comprises normal and atypical flora • Anaerobic and aerobic microbial strains • A combination of Gram-positive and -negative bacterial cultures • Even composition
	Genomic DNA	MSA-1004™		
Skin	Whole cell	MSA-2005™	6	
	Genomic DNA	MSA-1005™		
Gut	Whole cell	MSA-2006™	12	
	Genomic DNA	MSA-1006™		
Vaginal	Whole cell	MSA-2007™	6	
	Genomic DNA	MSA-1007™		

ATCC Virome Standards



Composition of Virome Standards
Human herpesvirus 5 strain AD169 (ATCC® VR-538™)
Human mastadenovirus strain F (ATCC® VR-931™)
Influenza B virus strain B/Florida/4/2006 (ATCC® VR-1804™)
Zika virus strain MR 766 (ATCC® VR-1838™)
Reovirus 3 strain Dearing (ATCC® VR-824™)
Human respiratory syncytial virus strain A2 (ATCC® VR-1540™)

Standard	Preparation	ATCC® Catalog No.	Number of Organisms	Specification (ddPCR™)	Applications
Virome	Virus Mix	MSA-2008™	6	2 x 10^3 genome copies/µL per virus	Standards for virome assay development, optimization, verification, and validation; evaluating reproducibility; and use as a daily run quality control
	Nucleic Acid Mix	MSA-1008™	6	2 x 10^4 genome copies/µL per virus	

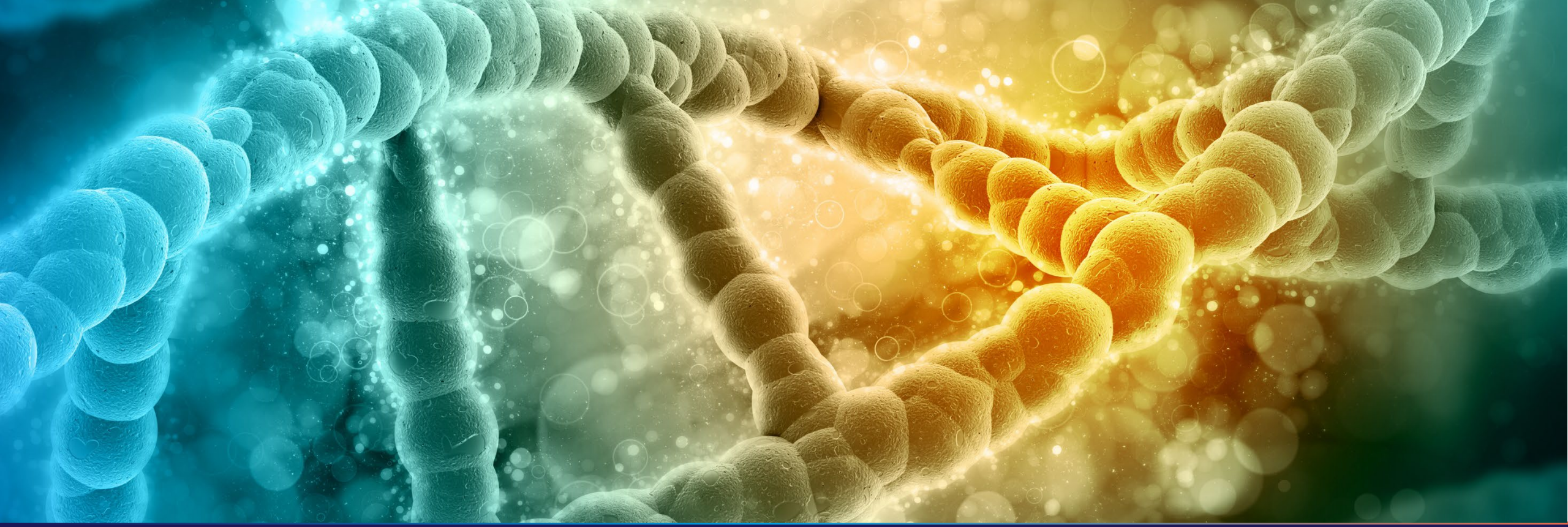
Spike-in and Mycobiome Standards

Standard	Preparation	ATCC® Catalog No.	Number of Organisms	Importance
Spike-in	Whole cell	MSA-2014™	3	• Microbiome measurements and data normalization • 16S rRNA and shotgun assay verification, validation, and quality control
	Genomic	MSA-1014™		

Standard	Preparation	ATCC® Catalog No.	Number of Organisms	Importance
Mycobiome	Whole cell	MSA-2010™	10	• Fungal mock community standards for assay development, optimization, verification, and validation; evaluating reproducibility; and use as a daily run quality control
	Genomic	MSA-1010™		

Utility and Application of Microbiome Standards

ATCC®



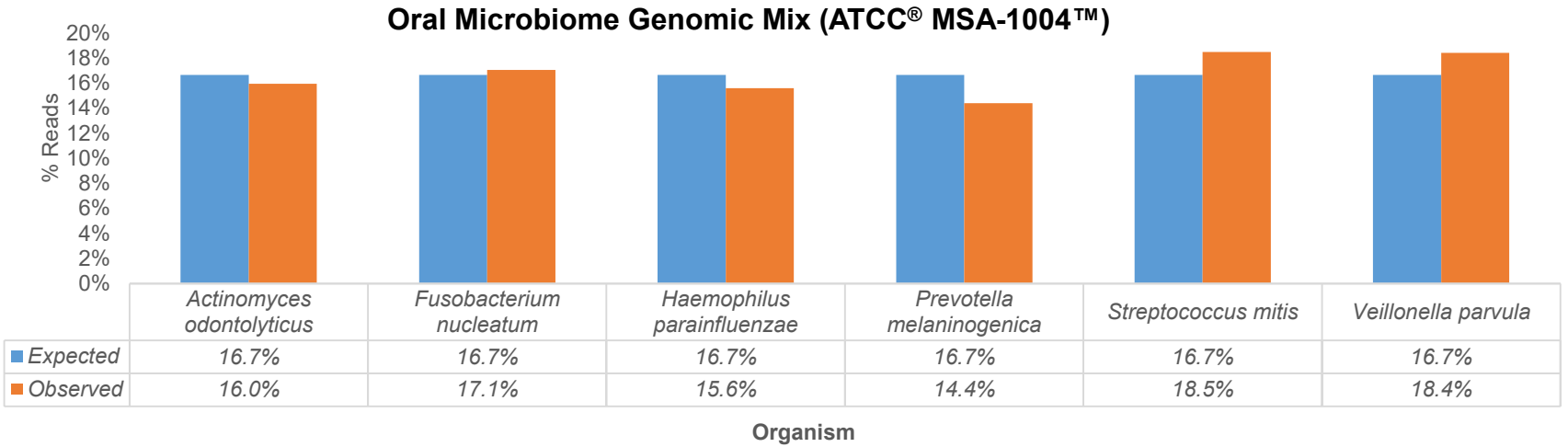
Evaluating DNA Extraction Methods and Kits

ATCC®

Genomic Versus Whole Cell Standards

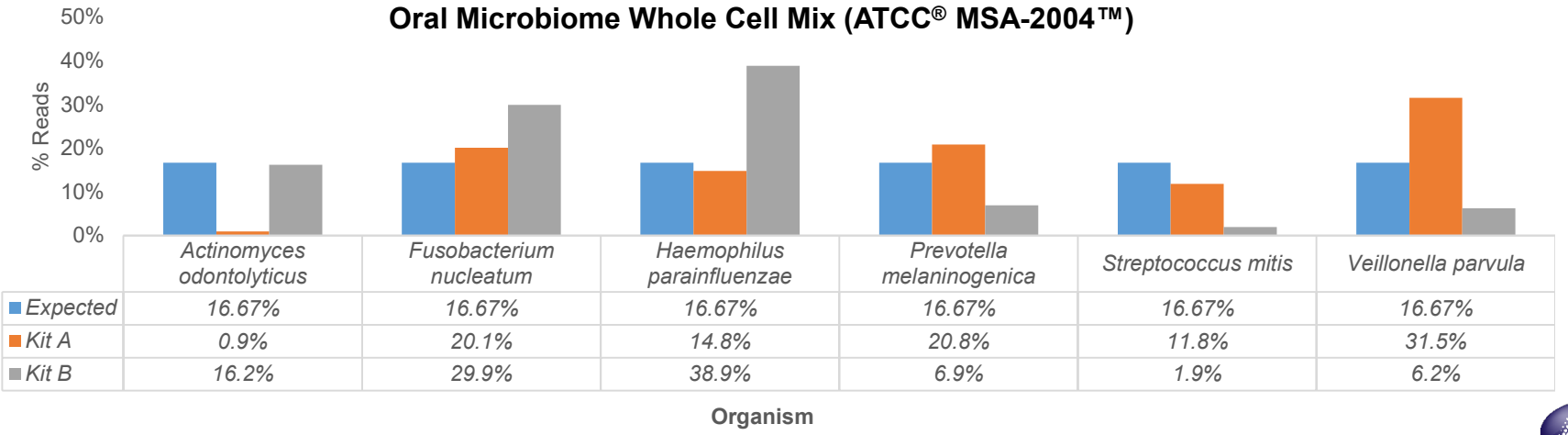
DNA extraction methods are not perfect

Shotgun metagenomic analysis of the Oral Microbiome Genomic Mix



	Actinomyces odontolyticus	Fusobacterium nucleatum	Haemophilus parainfluenzae	Prevotella melaninogenica	Streptococcus mitis	Veillonella parvula
Expected	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%
Observed	16.0%	17.1%	15.6%	14.4%	18.5%	18.4%

DNA extraction from the Oral Microbiome Whole Cell Mix with two different kits followed by shotgun metagenomic analysis

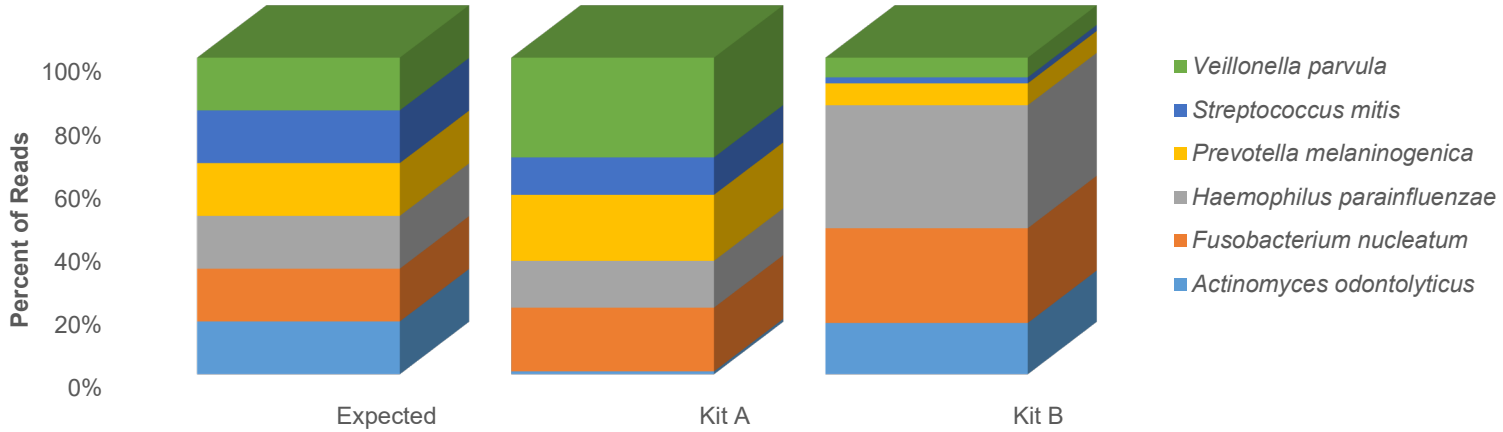


	Actinomyces odontolyticus	Fusobacterium nucleatum	Haemophilus parainfluenzae	Prevotella melaninogenica	Streptococcus mitis	Veillonella parvula
Expected	16.67%	16.67%	16.67%	16.67%	16.67%	16.67%
Kit A	0.9%	20.1%	14.8%	20.8%	11.8%	31.5%
Kit B	16.2%	29.9%	38.9%	6.9%	1.9%	6.2%

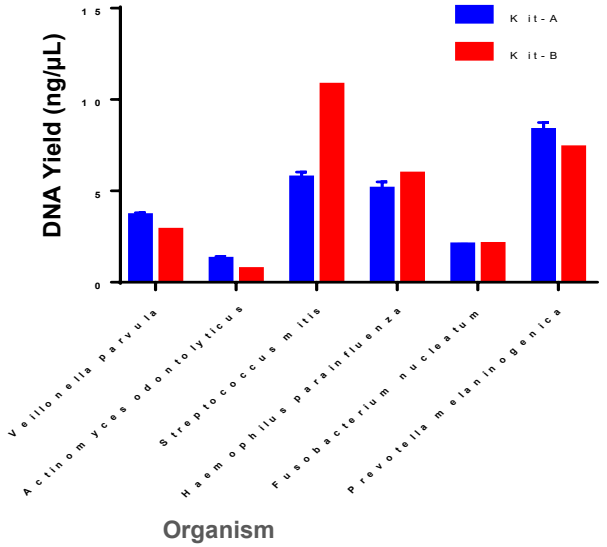
Assess Biases in DNA Extraction

Compare different pre-treatments and extraction methods, optimize protocols, and validate different kits

DNA extraction from the Oral Whole Cell Mix with two different kits followed by shotgun metagenomic analysis



DNA extraction from individual strains that are components of the Oral Whole Cell Mix



Organism	Number of Cells per Component	Gram Stain	Genome size	%GC
Actinomyces odontolyticus	~$2x10^7$	+	2.39	65.5
Fusobacterium nucleatum		-	2.17	27.2
Haemophilus parainfluenzae		-	2.12	39.3
Prevotella melaninogenica		-	3.17	35.1
Streptococcus mitis		+	1.83	40.5
Veillonella parvula		-	2.16	38.6

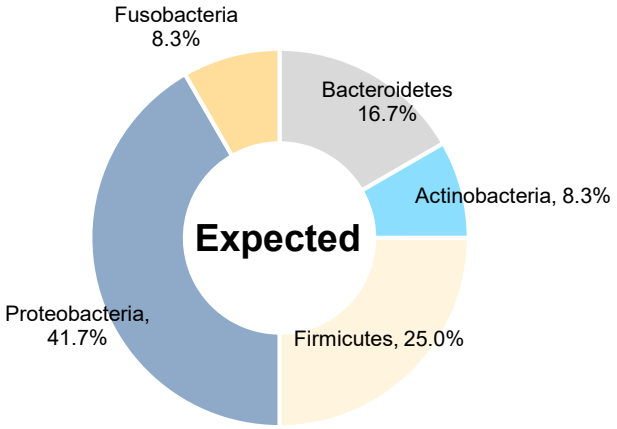
Extraction Kit

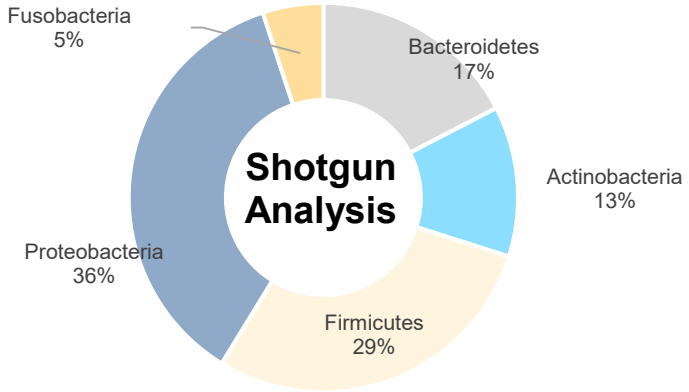
Gut Whole Cell Standard

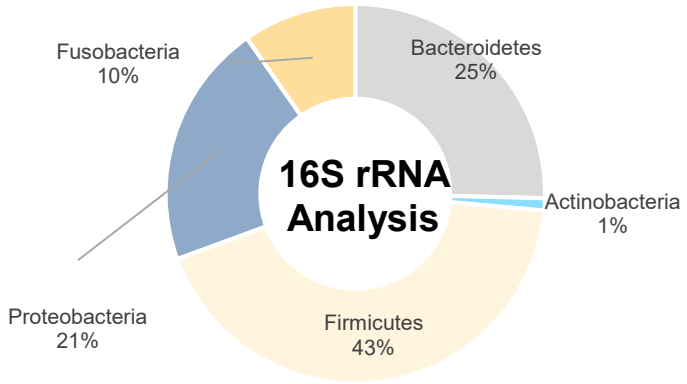
Profiling of gut microbiome standard at the phylum, genus, and species level

The Gut Whole Cell Microbiome Standard (ATCC® MSA-2006™) can be used as a full process control for shotgun and 16S rRNA assays

Phylum	Expected	Strain	Expected	Observed-Shotgun	Observed-16S (V1V2)
Bacteroidetes	16.6%	*Bacteroides fragilis*	8.3%	12.3%	17.0%
		Bacteroides vulgatus	8.3%	8.6%	8.4%
Actinobacteria	8.33%	*Bifidobacterium adolescentis*	8.3%	12.0%	1.0%
Firmicutes	25.0%	*Clostridium difficile*	8.3%	16.5%	29.0%
		Enterococcus faecalis	8.3%	3.1%	1.6%
		Lactobacilus plantarum	8.3%	8.1%	12.3%
Proteobacteria	41.7%	*Enterobacter cloacae*	8.3%	10.6%	4.0%
		Escherichia coli	8.3%	6.6%	3.3%
		Helicobacter pylori	8.3%	3.8%	7.7%
		Salmonella enterica	8.3%	4.9%	2.2%
		Yersinia enterocolitica	8.3%	8.8%	3.6%
Fusobacteria	8.3%	*Fusobacterium nucleatum*	8.3%	4.8%	9.7%







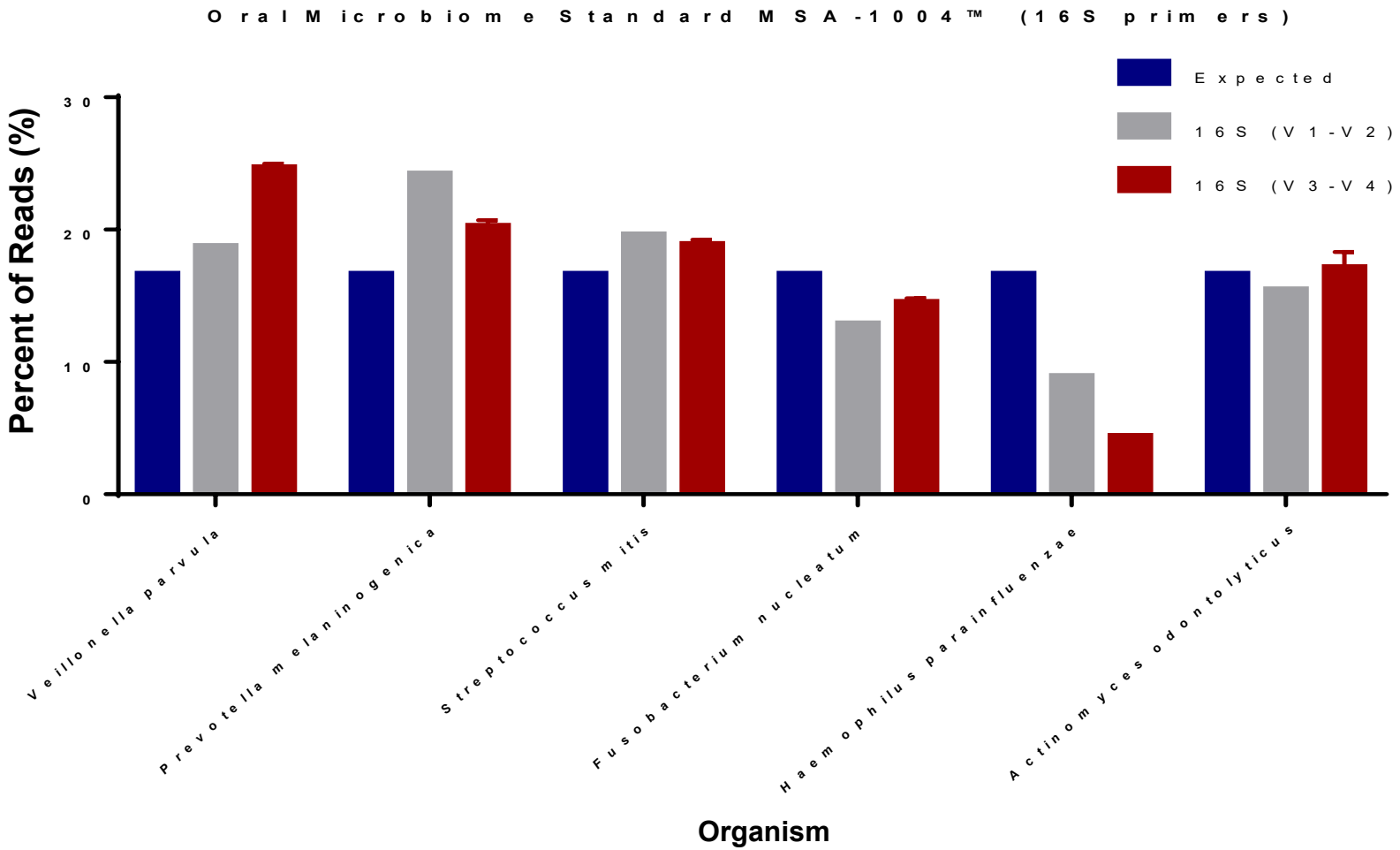
Evaluating 16S rRNA and WGS Library Kits

ATCC®

16S Amplicon-based Analysis: Primer Selection

Compare different primer sets, optimize amplification steps, and validate 16S analysis methods

16S rRNA analysis of the Oral Genomic DNA Standard via two primer sets



Comparing Library Preparation Kits

Nextera Flex enables uniform coverage of genomes of low GC content

Oral Microbiome Genomic DNA (ATCC® MSA-1004™)

Nextera™ XT Workflow

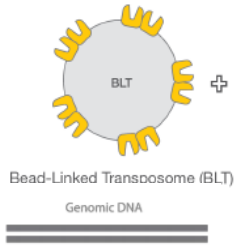
ONE CODEX

Free enzyme Tagmentation → **PCR Amplification** → **Normalization** → **Sequencing** → **Analysis**

Sample Composition

Name	Estimated Abundance
Actinomyces odontolyticus	46.07%
Prevotella melaninogenica	16.09%
Streptococcus mitis	16.04%
Veillonella parvula	12.65%
Haemophilus parainfluenzae	7.80%
Fusobacterium nucleatum	1.34%

Nextera™ Flex Workflow



ONE CODEX

Bead-Linked Tagmentation → **PCR Amplification** → **Sequencing** → **Analysis**

Sample Composition

Name	Estimated Abundance
Streptococcus mitis	18.83%
Veillonella parvula	18.42%
Fusobacterium nucleatum	17.13%
Haemophilus parainfluenzae	15.68%
Actinomyces odontolyticus	15.54%
Prevotella melaninogenica	14.40%

Data courtesy of Illumina

ATCC®

Comparing Library Preparation Kits

The LoopSeq™ 16S rRNA long-read method allows highest sequence accuracy and species-level taxonomy

Loop Genomics

Overall Score – 99%

True Positives: 100% | Relative Abundance: 98% | False Positives: 100%

True Positives
Detection of organisms in the control
Expand
100%
20 true positives detected (of 20 total)

Relative Abundance
Quantification of organisms in the control
Expand
98%
20 organisms in control

False Positives
Detection of organisms not in the control
Expand
100%
0 false positives

Sample Metadata
Sample preparation details & metadata
Expand
16S sequencing

Short Reads

Overall Score – 94%

True Positives: 100% | Relative Abundance: 93% | False Positives: 90%

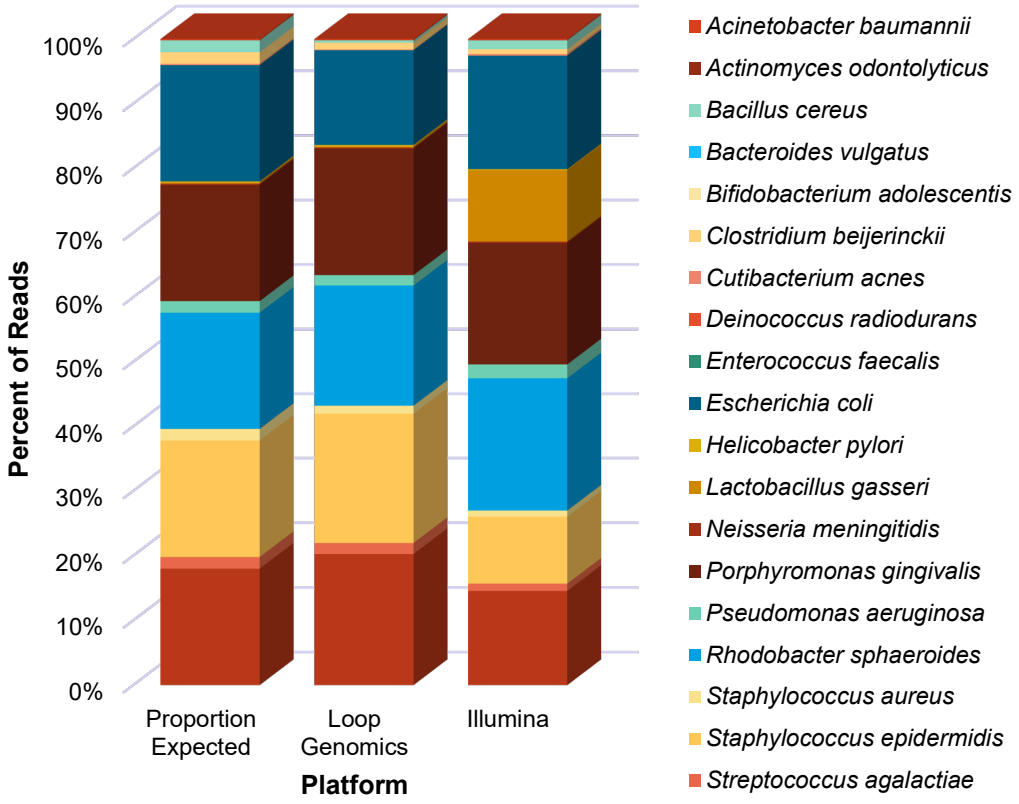
True Positives
Detection of organisms in the control
Expand
100%
20 true positives detected (of 20 total)

Relative Abundance
Quantification of organisms in the control
Expand
93%
20 organisms in control

False Positives
Detection of organisms not in the control
Expand
90%
10 false positives

Sample Metadata
Sample preparation details & metadata
Expand
16S sequencing of MSA-1003™

Genomic DNA (ATCC® MSA-1003™)



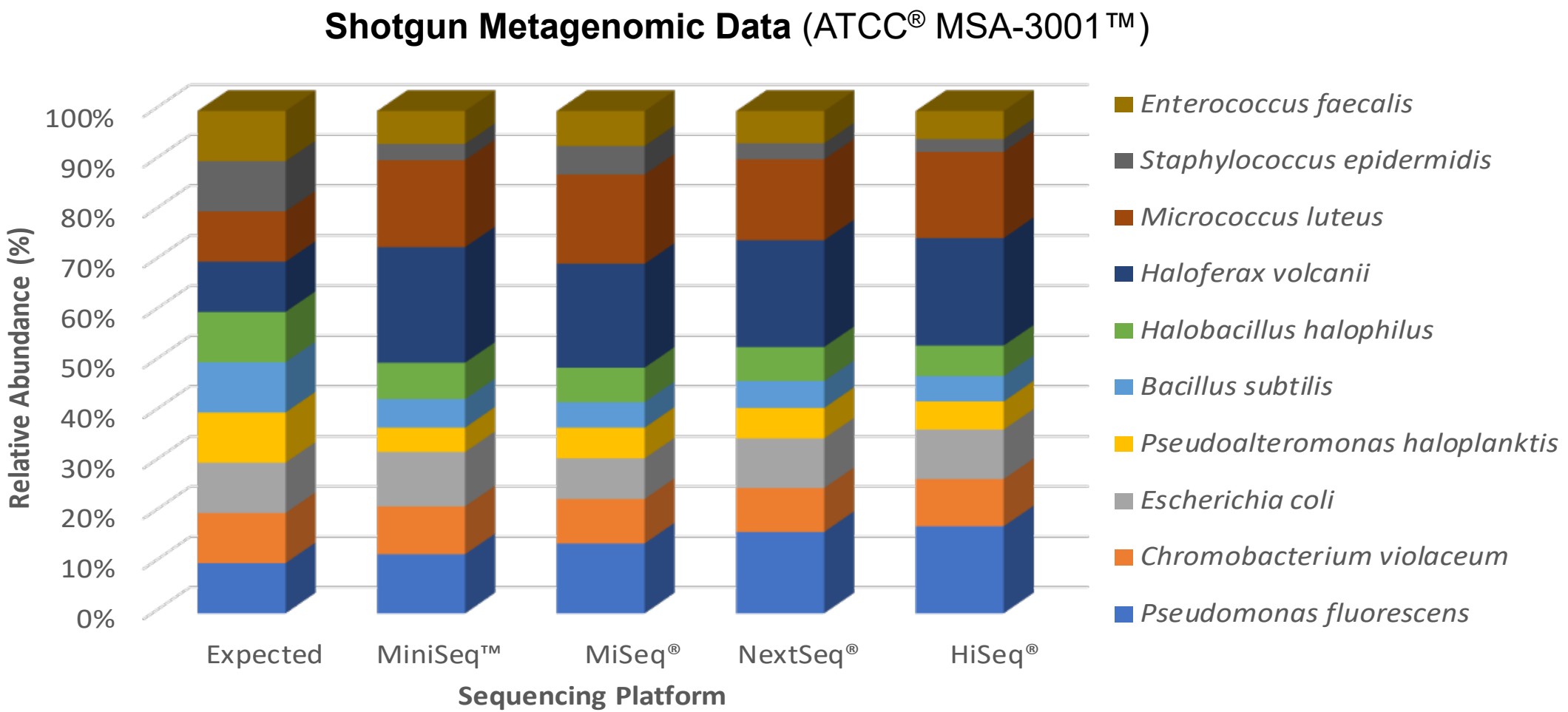
LoopSeq data courtesy of Tony Lialin, Loop Genomics

Evaluating NGS Platforms

ATCC

Short-read Sequencing Platform: Illumina®

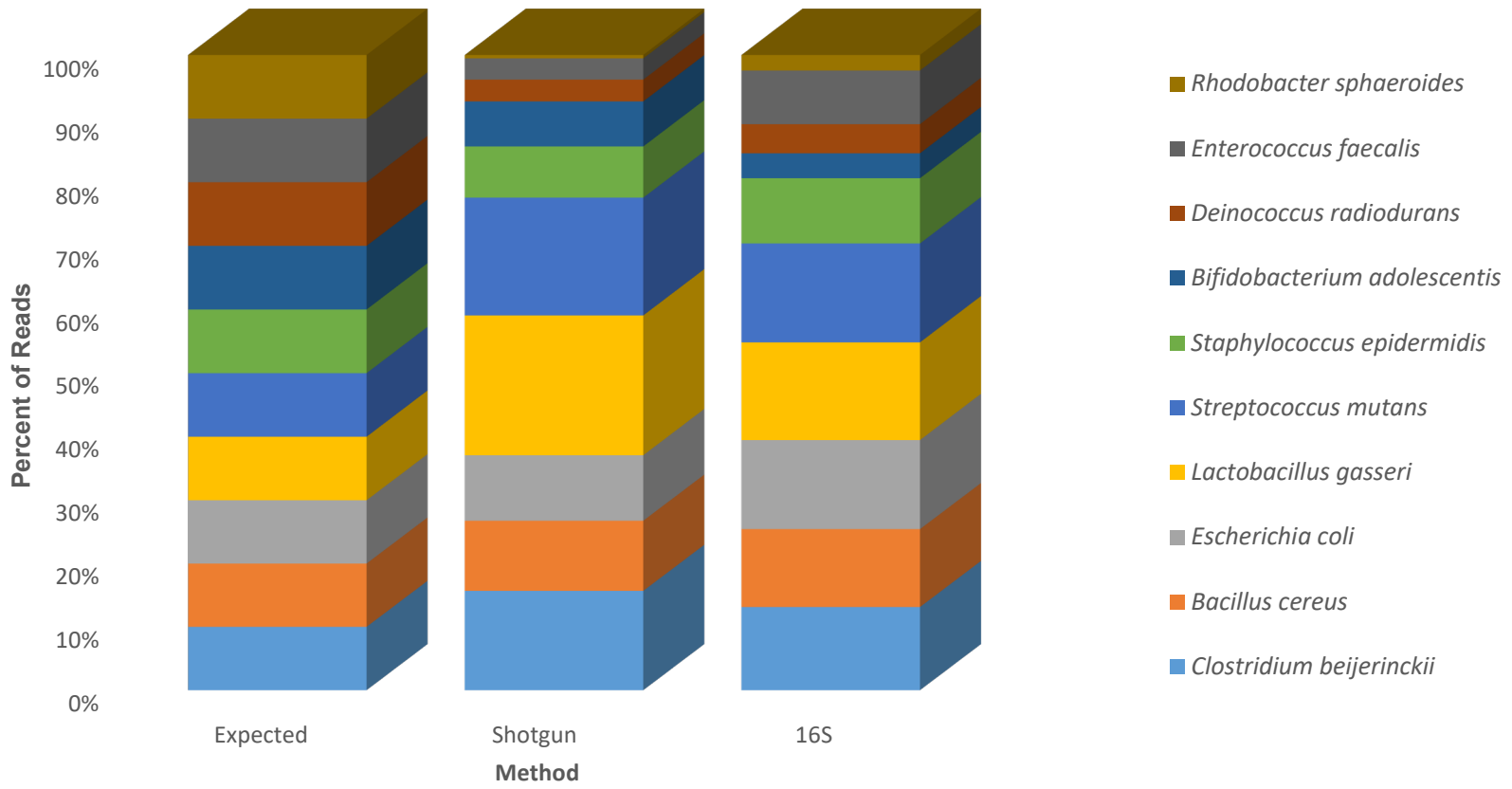
Assay reproducibility through different Illumina sequencing platforms



Data courtesy of Dr. Stefan Green, UIC (ABRF-MGRG)

Short-read Sequencing Platform: Ion Torrent™

16S rRNA and shotgun data on the Ion GPM Platform (ATCC® MSA-1000™)

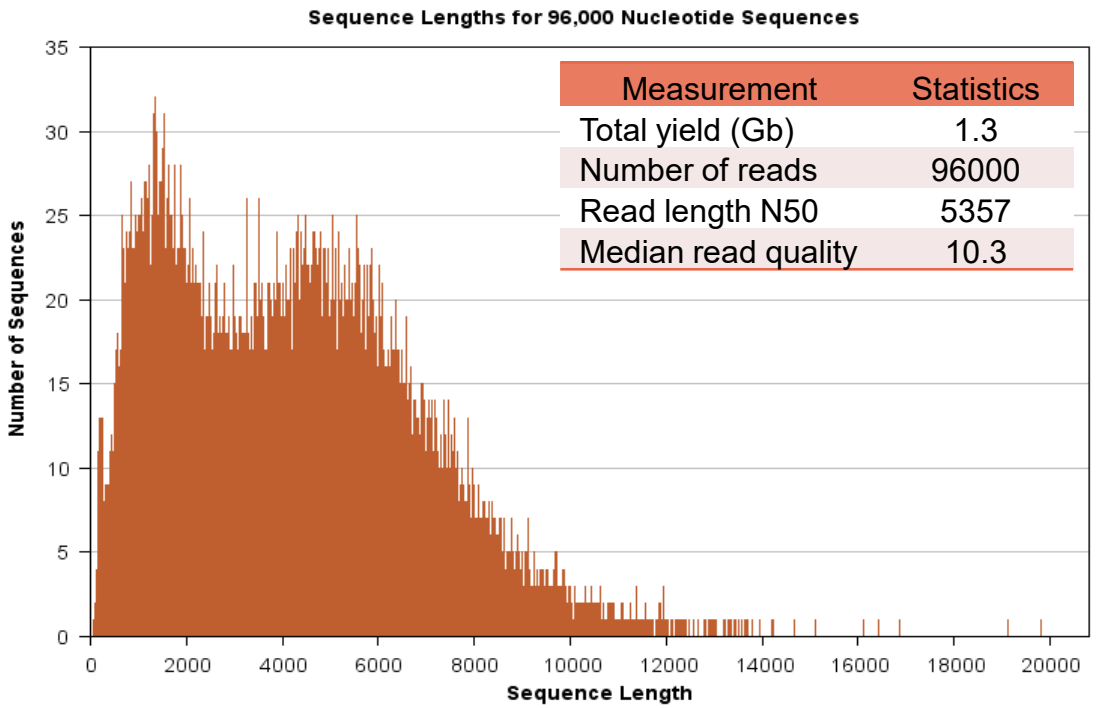


Data courtesy of Dr. Pat Gillevet and Rohan Patil (Microbiome Analysis Center, GMU)

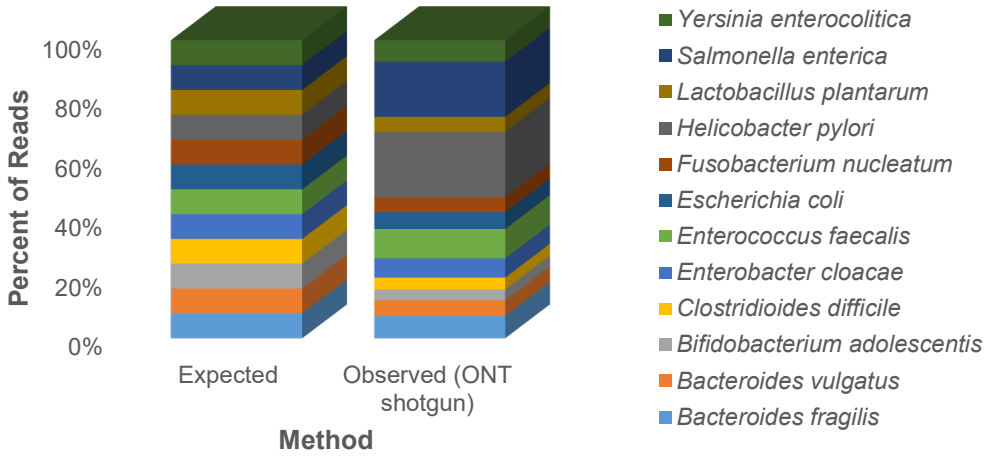
Long-read Sequencing Platform: Nanopore®

One hour sequencing coverage was enough to identify all organisms in the mix with sufficient genome coverage

The Gut Microbiome Whole Cell Standard (ATCC® MSA-1006™) was analyzed via shotgun sequencing on the MinION platform



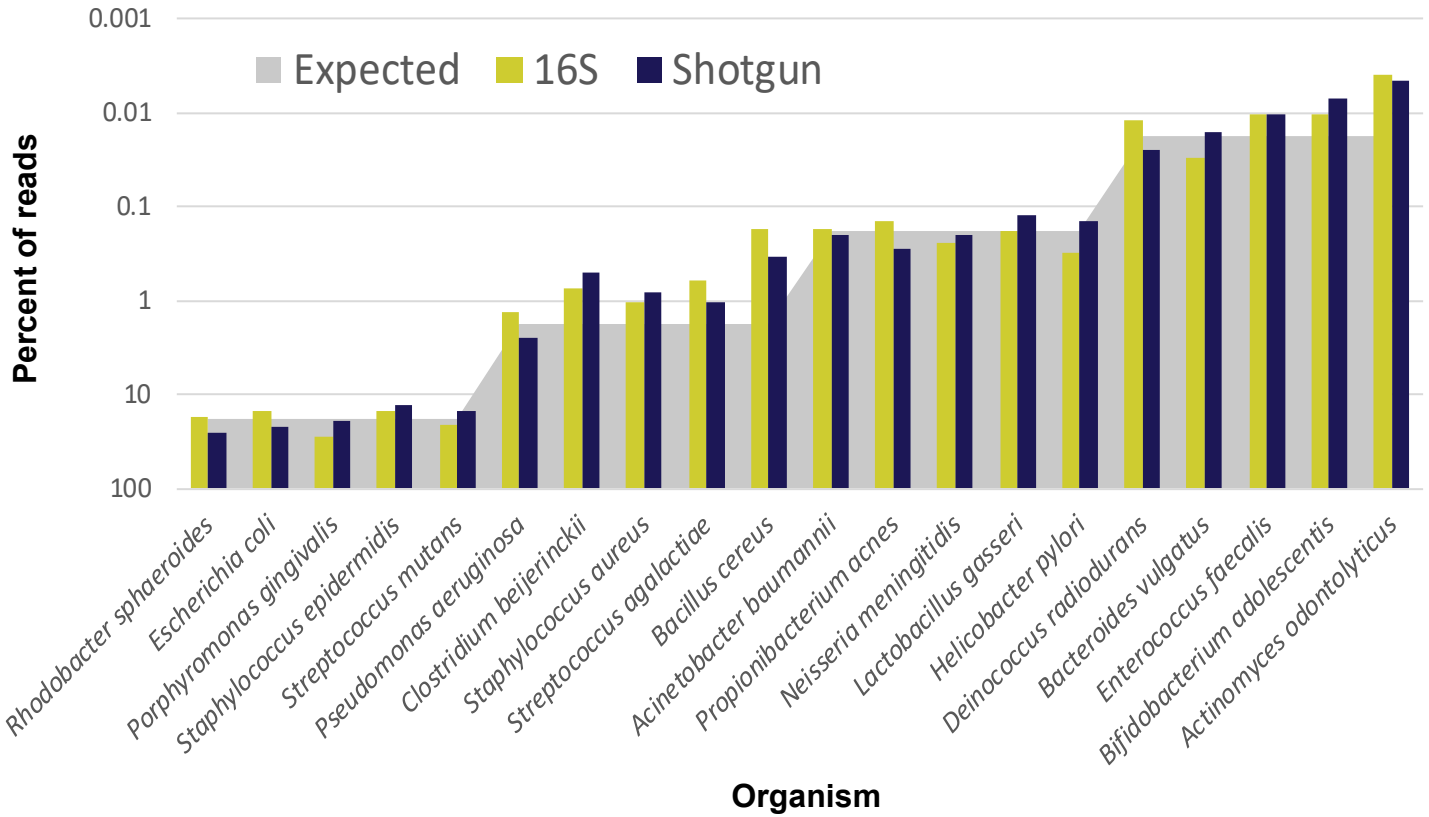
Measurement	Statistics
Total yield (Gb)	1.3
Number of reads	96000
Read length N50	5357
Median read quality	10.3



Organism	Genome Coverage (x)
Enterobacter cloacae	9.1
Enterococcus faecalis	14.1
Bacteroides fragilis	8.1
Bacteroides vulgatus	6.6
Bifidobacterium adolescentis	1.8
Clostridioides difficile	7.4
Escherichia coli	6.5
Fusobacterium-nucleatum	4.6
Helicobacter pylori	16.6
Lactobacillus plantarum	6.0
Salmonella enterica	11.1
Yersinia enterocolitica	11.3

Long-read Sequencing Platform: PACBIO®

16S rRNA (full-length) and shotgun data on the PacBio Sequel Platform (ATCC® MSA-1003™)



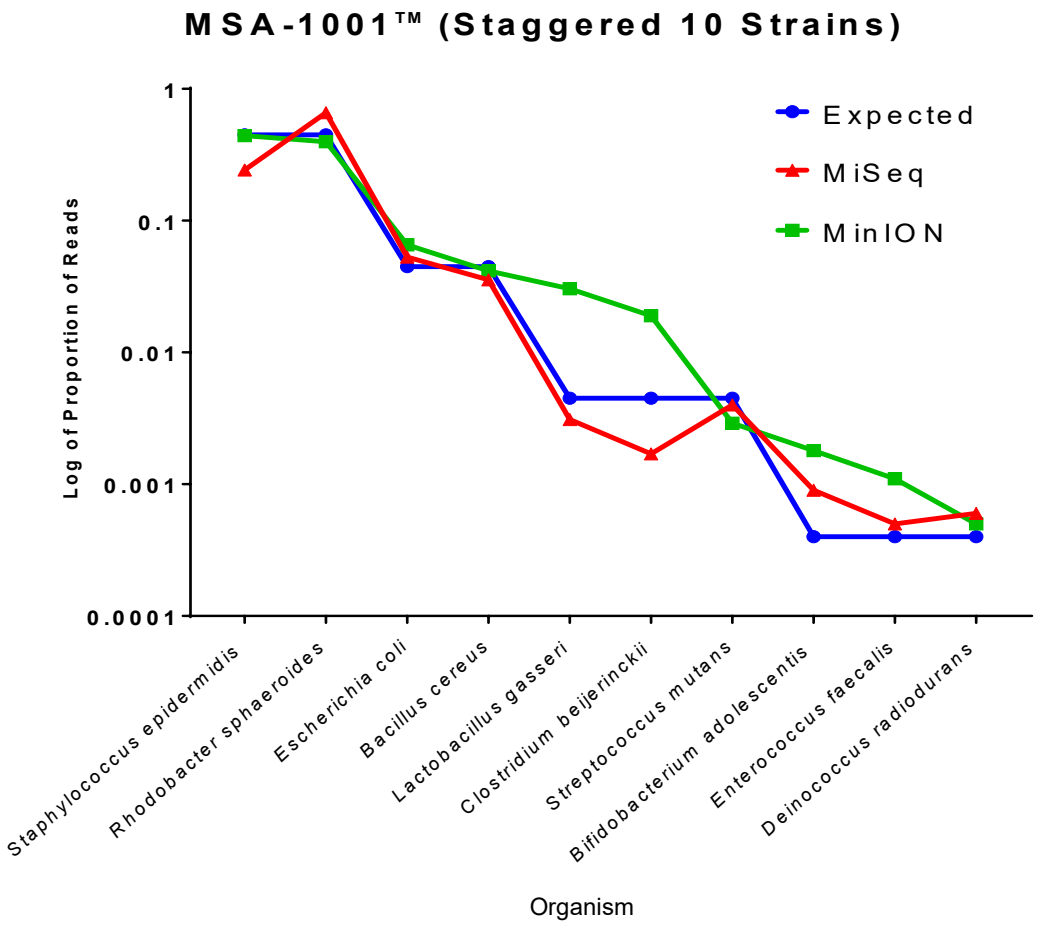
ATCC quality control score (One Codex)

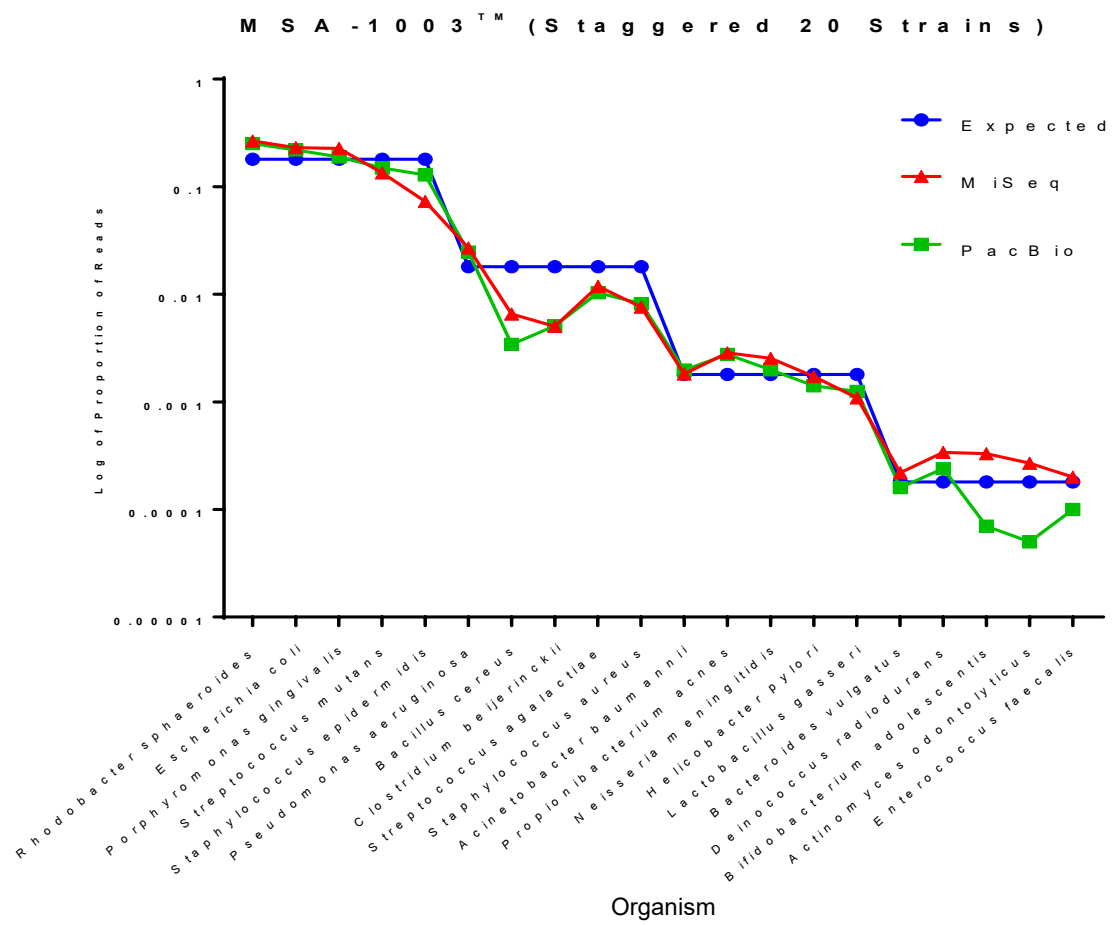
One Codex Analysis	16S rRNA run 1	16S rRNA run 2	Shotgun run 1	Shotgun run 2
True positives	100%	100%	100%	100%
Relative abundance	95%	95%	97%	97%
False positives	100%	100%	88%	84%
Overall score	98%	98%	95%	95%

Data courtesy of Dr. Joan Wong, PACBIO®

Shotgun Metagenomic Analysis: Short vs Long Reads

ATCC Microbiome Standards are technology agnostic



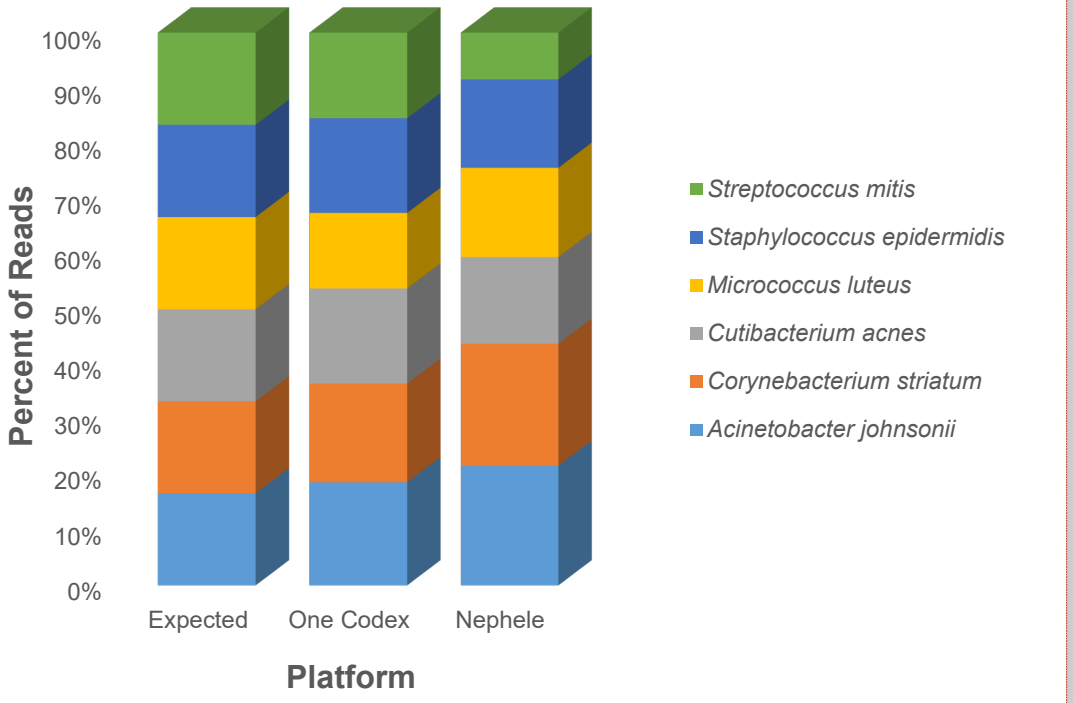


Comparing Bioinformatics and Databases

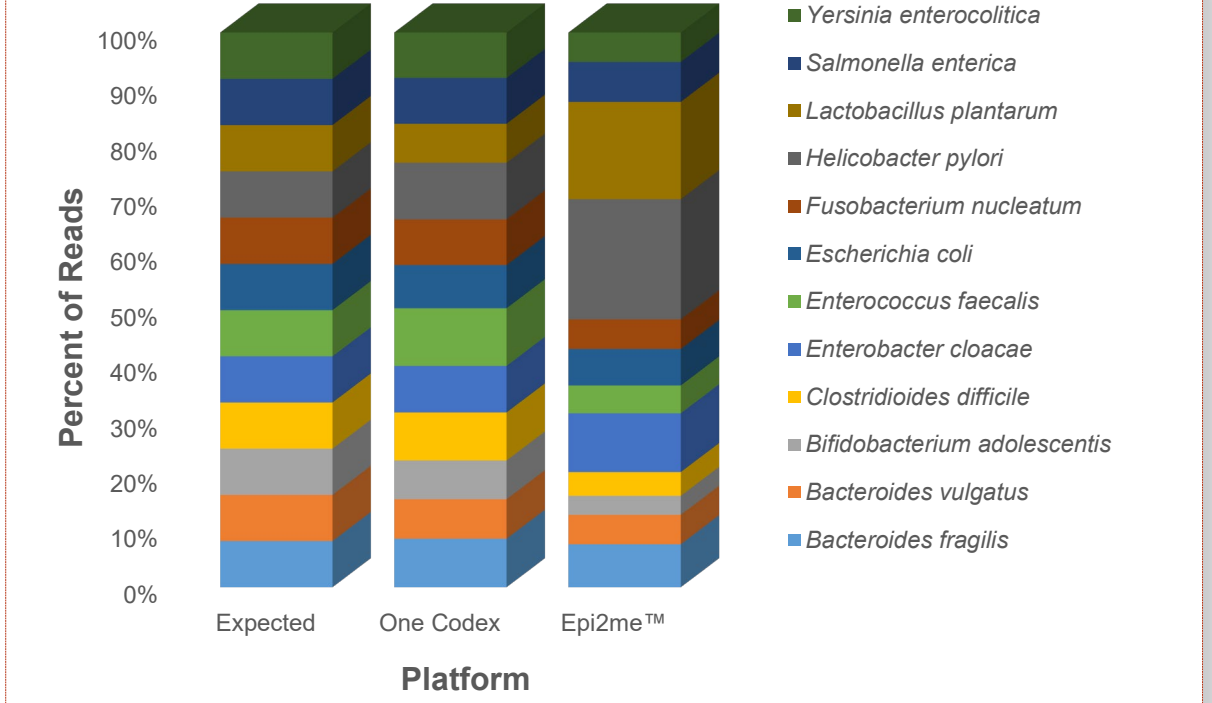
Data Analysis Using Different Databases

Evaluation of NGS data from microbiome standards in multiple analysis platforms and databases

Nephele vs One Codex
Short-read sequencing data from the Skin Genomic DNA Mix (ATCC® MSA-1005™)



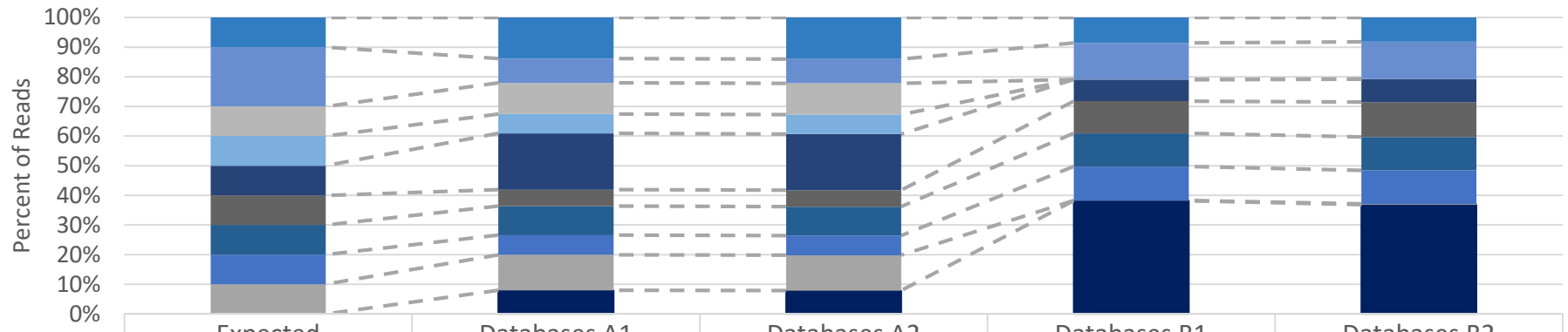
Epi2Me vs One Codex
Long-read sequencing data from the Gut Genomic DNA Mix (ATCC® MSA-1006™)



Mycobiome Standards

Data analysis platform impacts strain identification and taxonomic resolution

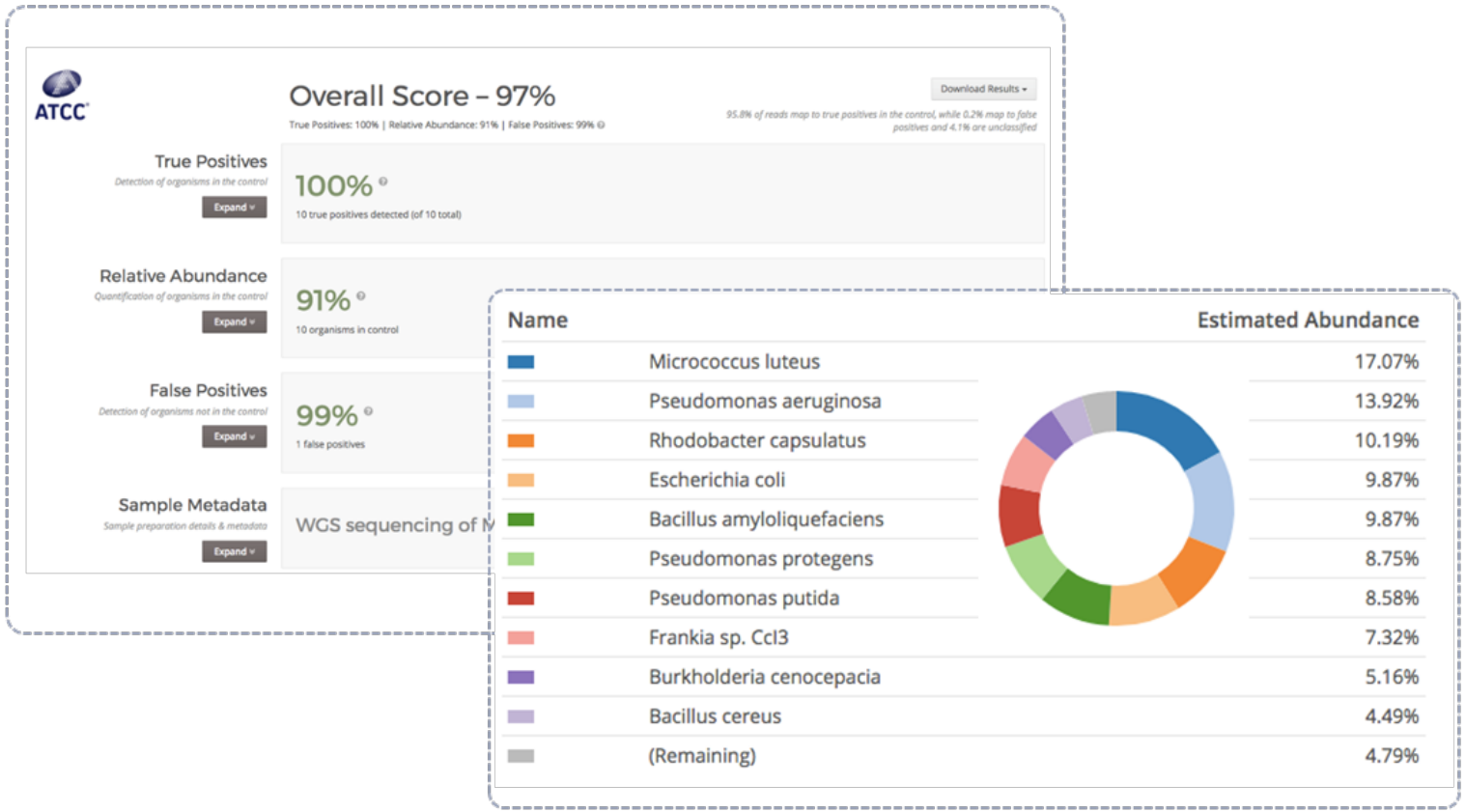
Shotgun Analysis of Genomic DNA standards (ATCC® MSA-1010™)



	Expected	Databases A1	Databases A2	Databases B1	Databases B2
Aspergillus	10%	13.88%	13.99%	8.60	8.16
Candida	20%	8.17%	8.22%	12.31	12.61
Cryptococcus	10%	10.52%	10.51%	0.00	0.00
Cutaneotrichosporon	10%	6.53%	6.61%	0.00	0.00
Fusarium	10%	19.00%	18.94%	7.32	7.78
Malassezia	10%	5.53%	5.55%	10.86	11.74
Penicillium	10%	9.81%	9.76%	11.17	11.25
Saccharomyces	10%	6.63%	6.62%	11.48	11.35
Trichophyton	10%	12.01%	11.98%	0.05	0.29
Unclasified	0%	7.93%	7.82%	38.21	36.82

ATCC Data Analysis Solution

ONE CODEX



WORKFLOW:

1. Drag and drop Fastq files or export via cloud
2. Choose your ATCC product and analysis (16s and shotgun)
3. Download your reports

RESULTS ARE PROVIDED ON A SCORECARD REPORTS:

1. **True positives:** Percentage of organisms detected from the control
2. **False positives:** Detection of organisms not in the control
3. **Relative abundance:** Quantification of organisms in the control

ATCC Data Analysis Solution

Mock Microbial Communities

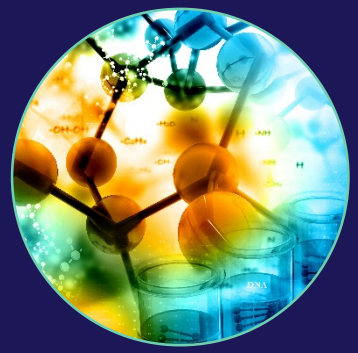
- Genomic DNA and whole cell standards
- Even and staggered mixtures comprising 10 or 20 strains
- Environmental and pathogen mixtures

Site-specific Standards

- Genomic DNA and whole cell standards
- Even mixtures of 6-12 strains
- Bacterial strains prevalent in the oral, skin, gut, and vaginal microbiome

Spike-In Standards

- Recombinant strains with a unique DNA tag stably integrated into the chromosome
- Recombinant standards include the Gram-negative and Gram-positive bacteria



New Products

- Genomic DNA and whole cell mock communities representing:
  - Virome
  - Mycobiome

Bundled with data analysis on the One Codex platform

Posters at ASM Microbe 2019

Development of Fungal Mock Community Standards for Mycobiome Studies
Poster Board Number: FRIDAY – MBP-74
Date: Friday, June 21, 2019
Time: 11:00 AM-12:00PM, 4:00 PM-5:00 PM

Utility of Recombinant Bacteria with Unique Tags as Spike-In Controls for Microbiome Studies
Poster Board Number: FRIDAY – CPHM-940
Date: Friday, June 21, 2019
Time: 11:00 AM-12:00PM, 4:00 PM-5:00 PM

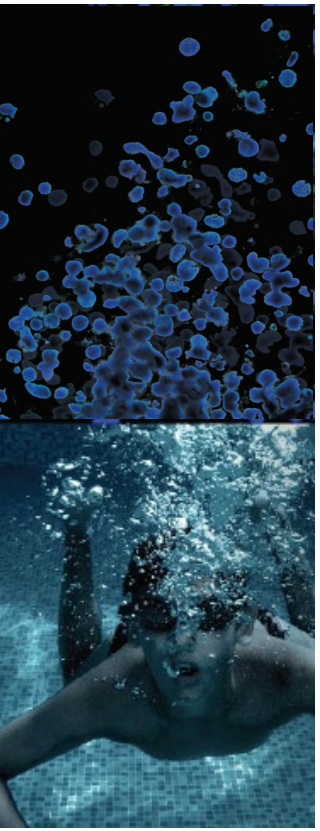
Evaluation of ATCC® Site-Specific Microbiome Standards on Long-Read Sequencing Platforms
Poster Board Number: SATURDAY – MBP-7
Date: Saturday, June 22, 2019
Time: 11:00 AM-12:00PM, 4:00 PM-5:00 PM

Acknowledgements

- Monique Hunter, MS
- Anna McCluskey, BS
- Stephen King, MS
- Juan Lopera, PhD
- Cara Wilder, PhD
- Dev Mittar, PhD
- Scott Tighe, PhD, UVM
- Denise O'Sullivan, PhD, LGC
- Nick Greenfield, MA, One Codex
- Pat Gillevet, PhD, Microbiome Analysis Center, GMU
- Rohan Patil, Microbiome Analysis Center, GMU
- Stefan Green, PhD, UIC (ABRF-MGRG)
- Joan Wong, PhD, PACBIO®
- Tony Lialin, Loop Genomics

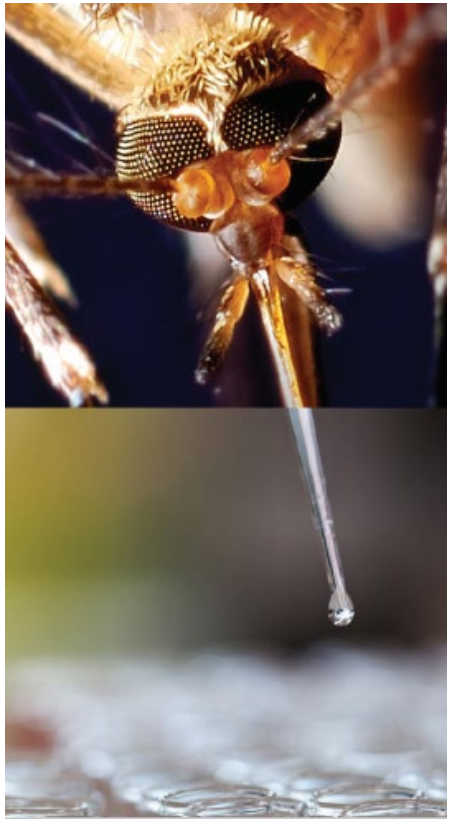
ATCC®

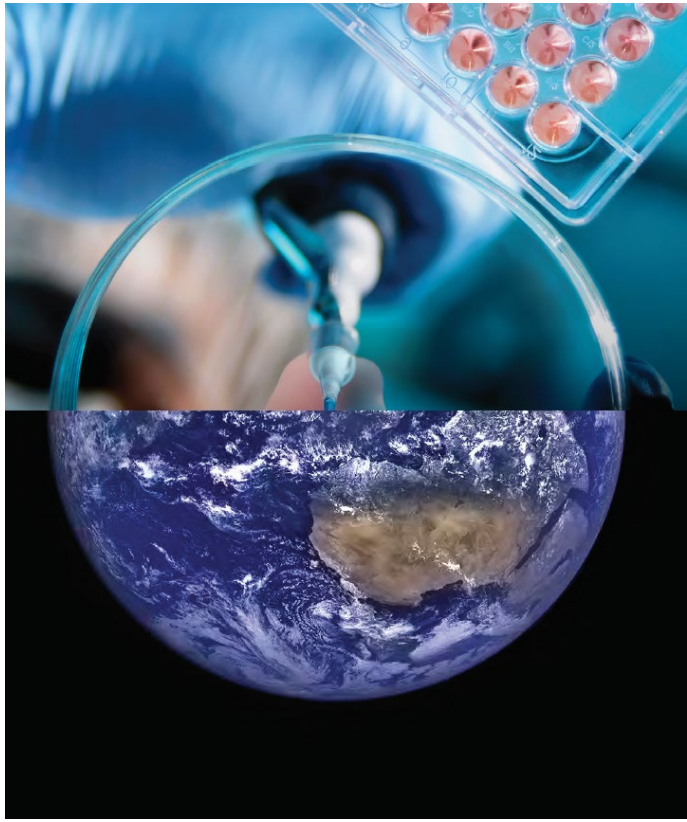
ATCC®



Questions?

Credible Leads to Incredible™





Gene-Level Microbiome Analysis Identifies Culturable Strains Consistently Associated with Human Cancer across Independent Cohorts

Samuel Minot, PhD
Microbiome Research Initiative
Fred Hutch Cancer Research Center
Seattle, Washington, USA

FRED HUTCH

Twitter: @sminot
Blog: www.minot.bio

Disclosures

ATCC – Consulting
One Codex – Financial Interest

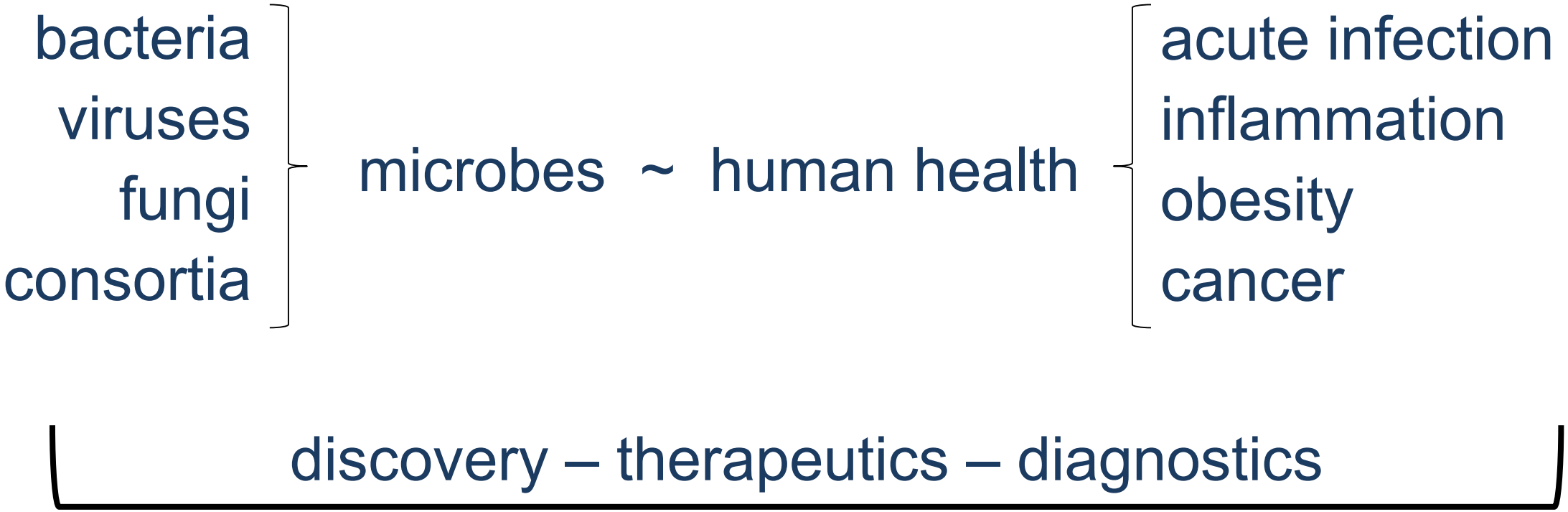
Collaborators

Amy Willis, PhD
Professor of Biostatistics
University of Washington

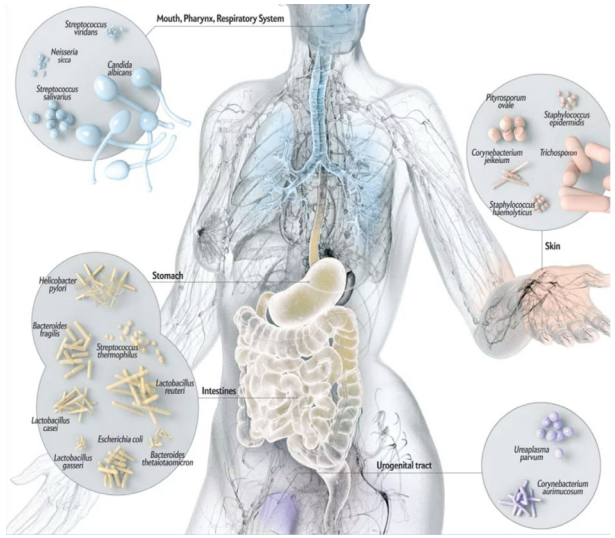
Jonathan Golob, MD, PhD
Infectious Disease & Internal Medicine
University of Michigan

gene-level metagenomics

Applied Microbiology

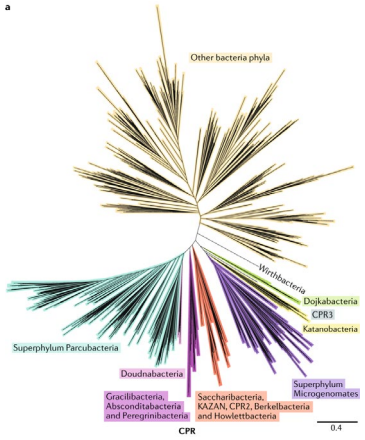


Ontologies of Microbiome Analysis



WGS

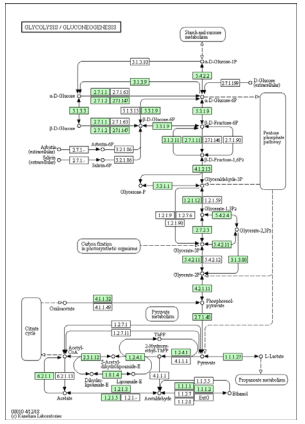
Microbial Taxa



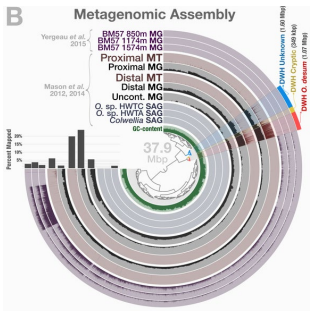
Microbial genes

geneX geneY geneZ

Metabolic Pathways



Assembled Genomes



Scientific American (2012-05-15)
Castelle, *et al*. 2018
Eren, *et al*. 2015
genome.jp/kegg

FRED HUTCH

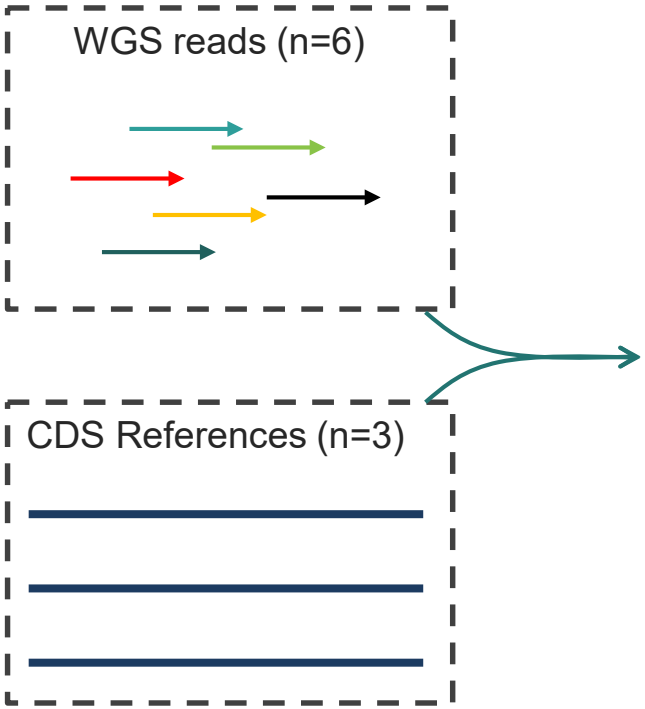
Gene-Level Analysis of the Microbiome

1. Computational methods development
    - Single-sample analysis
    - Cross-sample analysis
2. Application to CRC datasets
3. Validation in mouse model

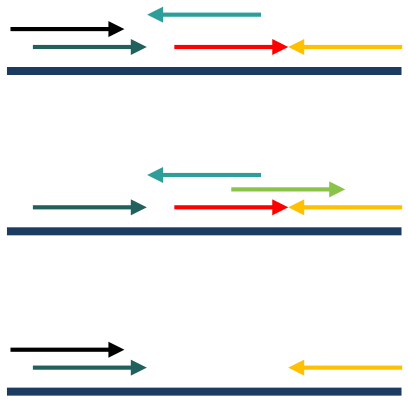
Gene-Level Analysis of the Microbiome

1. Computational methods development
    - **Single-sample analysis**
    - Cross-sample analysis
2. Application to CRC datasets
3. Validation in mouse model

Detecting Genes from WGS Data

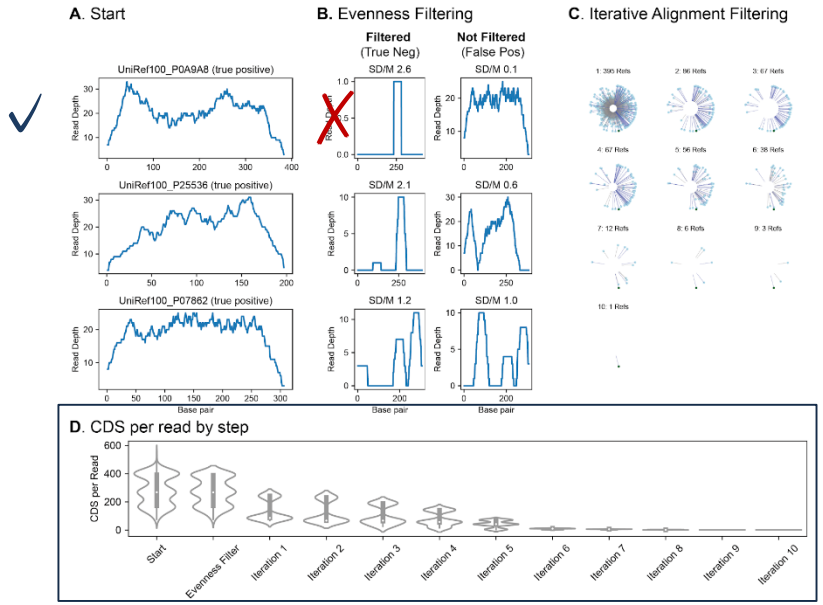


Alignments

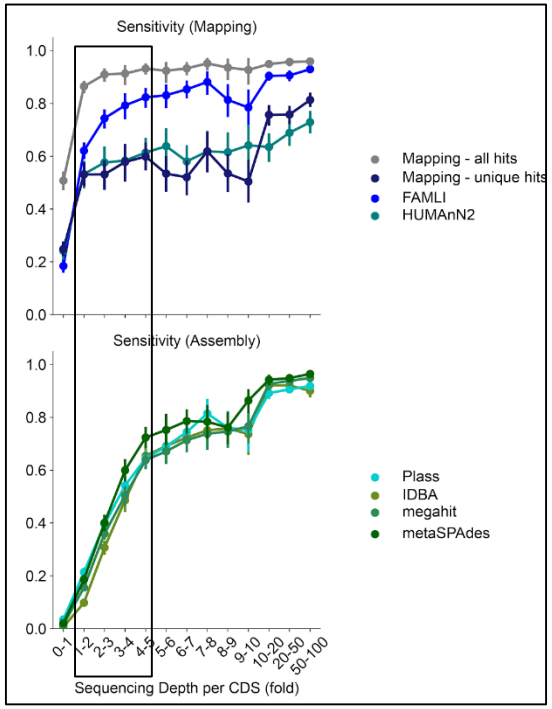


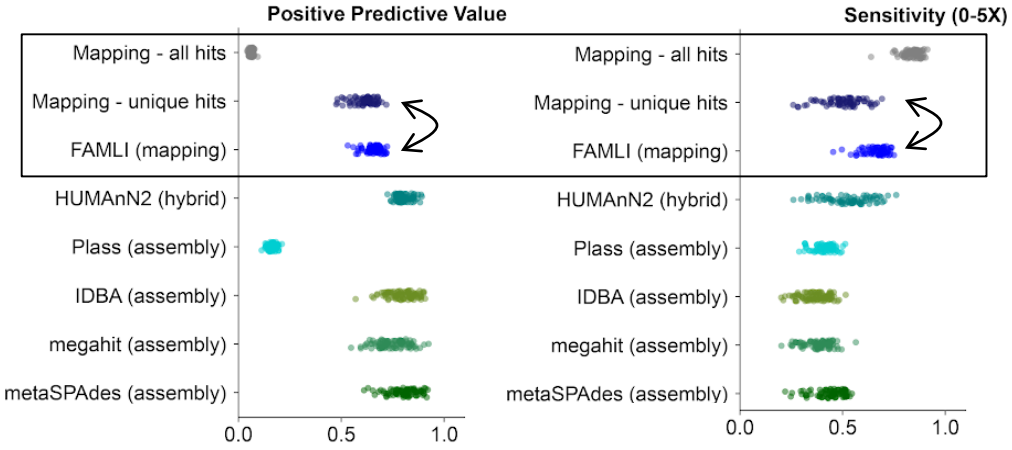
Take all hits?
Take unique hits?

FAMLI Filtering Algorithm

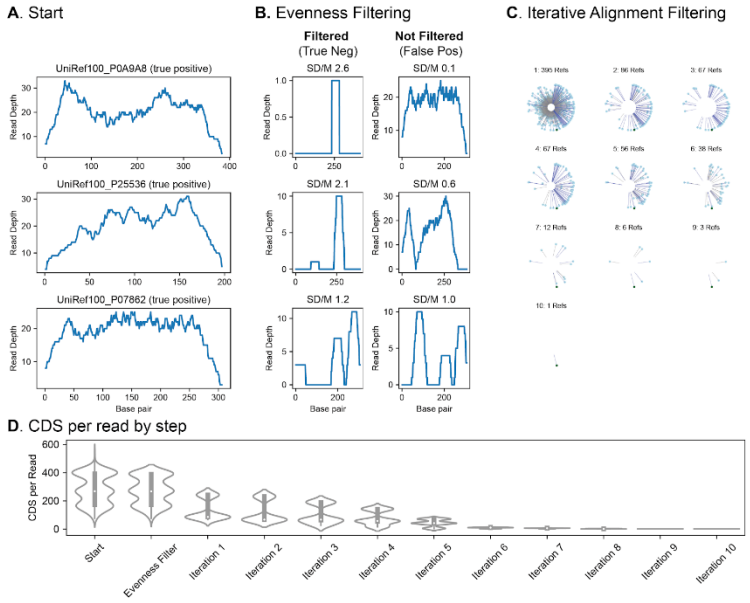


Detecting Genes from WGS Data





FAMLI Filtering Algorithm



Performance Evaluation

Gene-Level Analysis of the Microbiome

1. Computational methods development
    - Single-sample analysis
    - **Cross-sample analysis**
2. Application to CRC datasets
3. Validation in mouse model

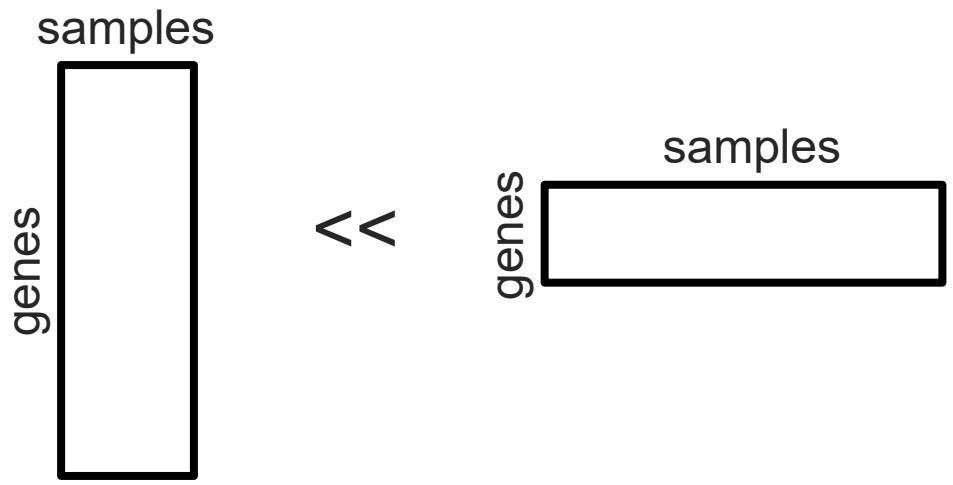
Efficiently Comparing Gene-Level Communities

Single stool metagenome:

10-50M reads → 100k – 1M genes

Aggregate metagenome:

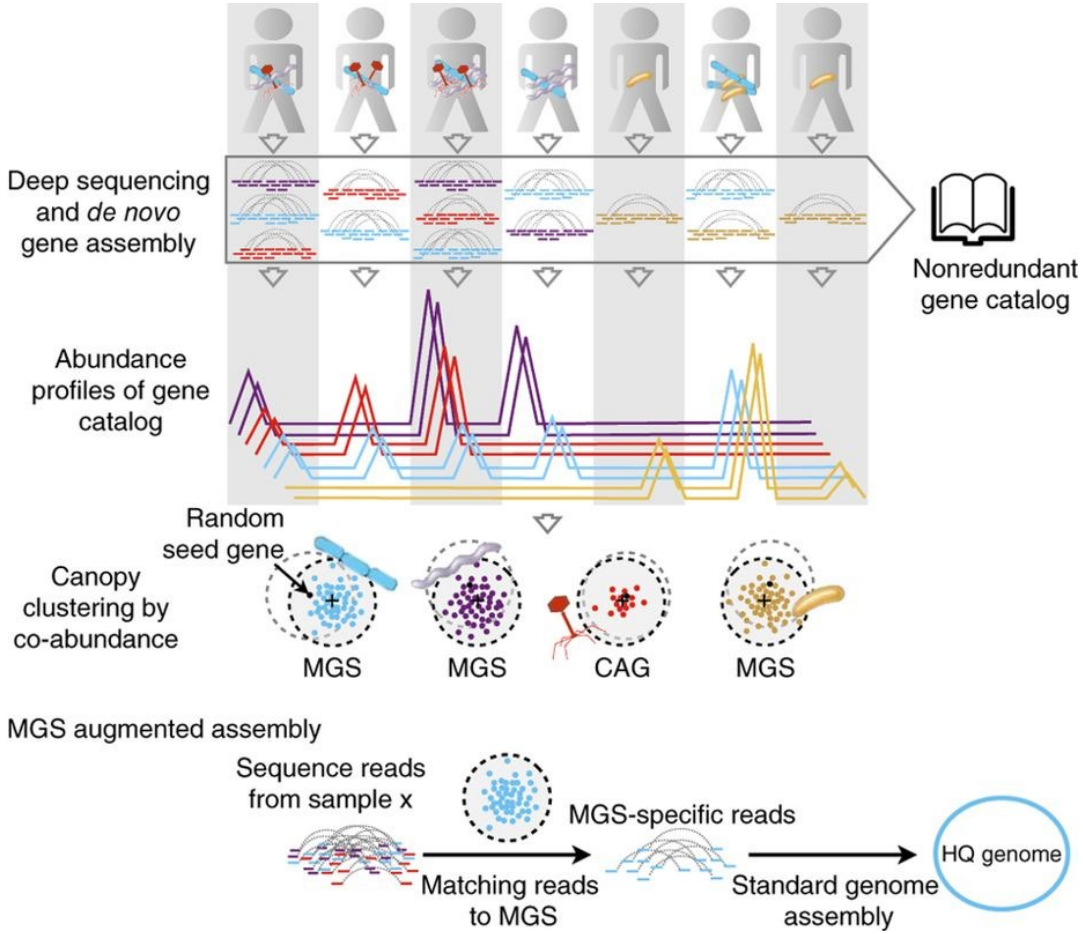
~100 people → 1M – 10M genes



Efficiently Comparing Gene-Level Communities

Approach: Co-Abundance Clustering

- Genes → Co-Abundant Gene Groups (CAGs)



Nielsen HB, et al. Nature Biotech 2014

Previous state-of-the-art:

Clustering heuristic to identify **species**

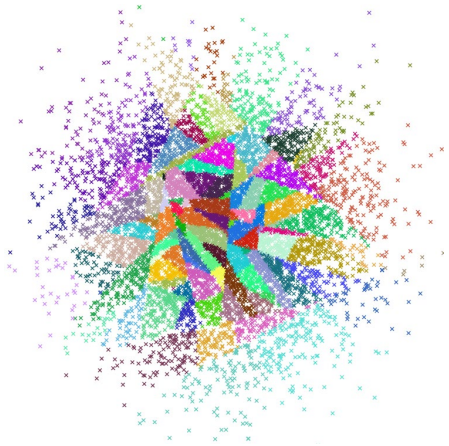
Goal:

Clustering of co-abundant **genes**
(operons, HGT, phages, auxiliary genome)

Enabling Technology:

Approximate Nearest Neighbor Algorithm

Efficiently Comparing Gene-Level Communities



Approximate Nearest Neighbor Algorithm efficiently partitions densely populated high-dimensional space

Previous state-of-the-art:

Clustering heuristic to identify **species**

Goal:

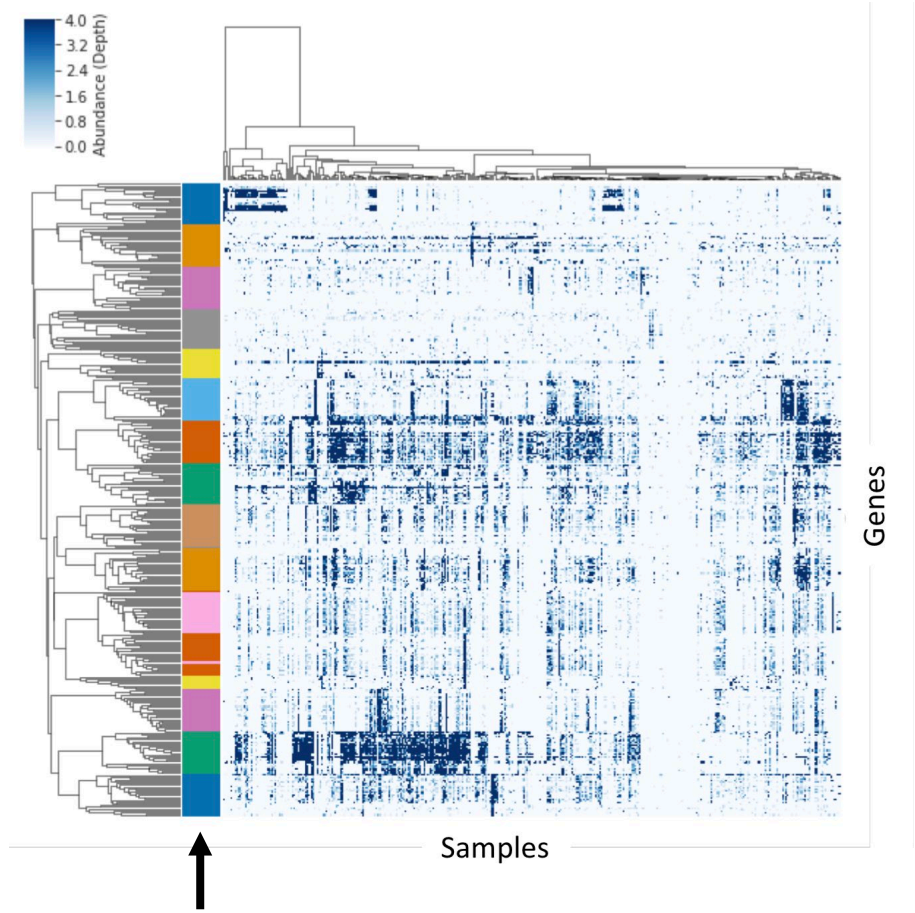
Clustering of co-abundant **genes**
(operons, HGT, phages, auxiliary genome)

Enabling Technology:

Approximate Nearest Neighbor Algorithm

Efficiently Comparing Gene-Level Communities

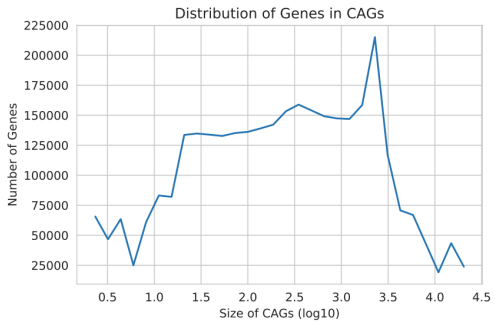
- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1

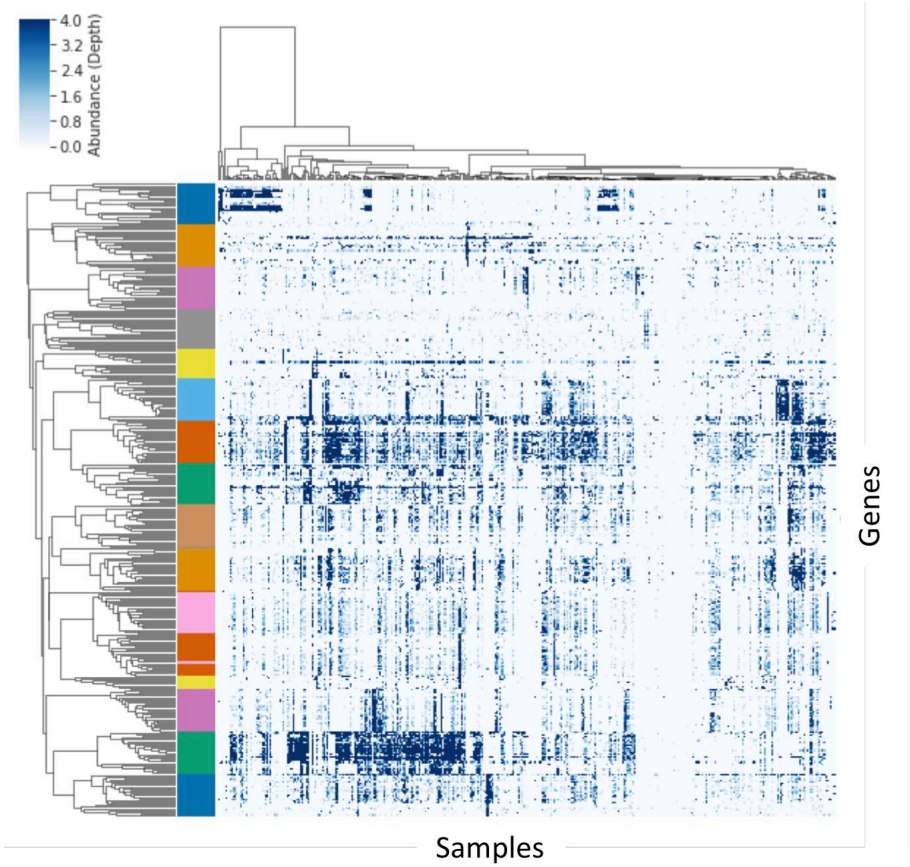


Gene Grouping Captures Co-Abundance

Efficiently Comparing Gene-Level Communities

- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1

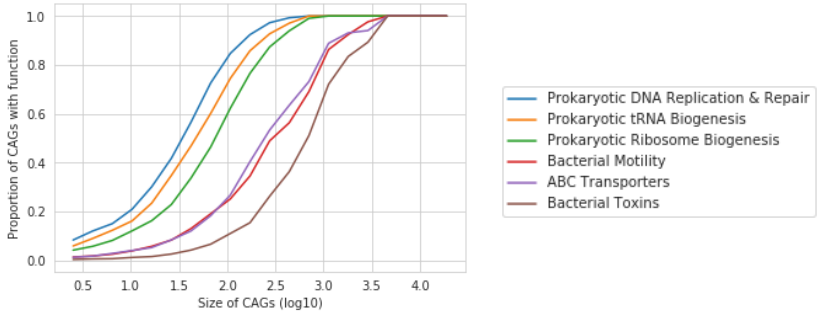


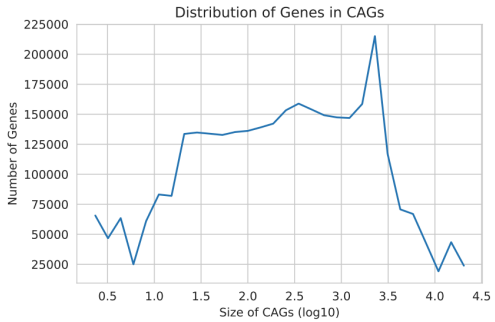


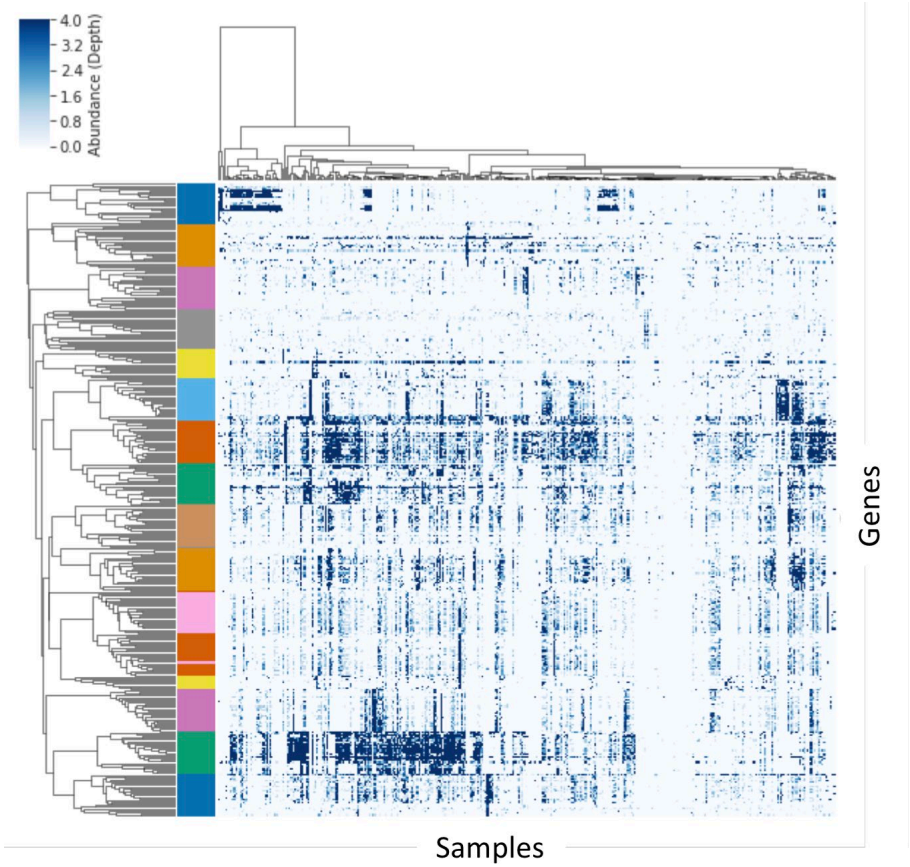
Co-Abundant Gene Groups (CAGs) range from 20-3,000 genes

Efficiently Comparing Gene-Level Communities

- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1





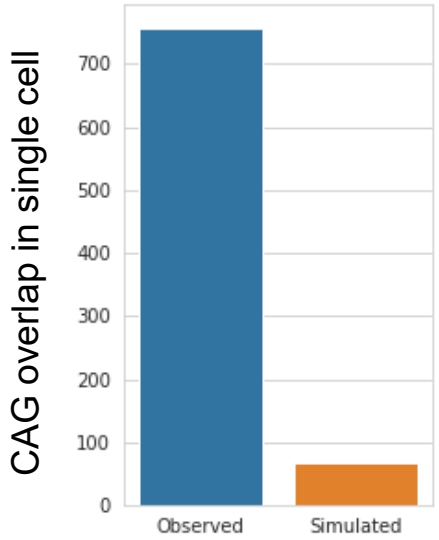


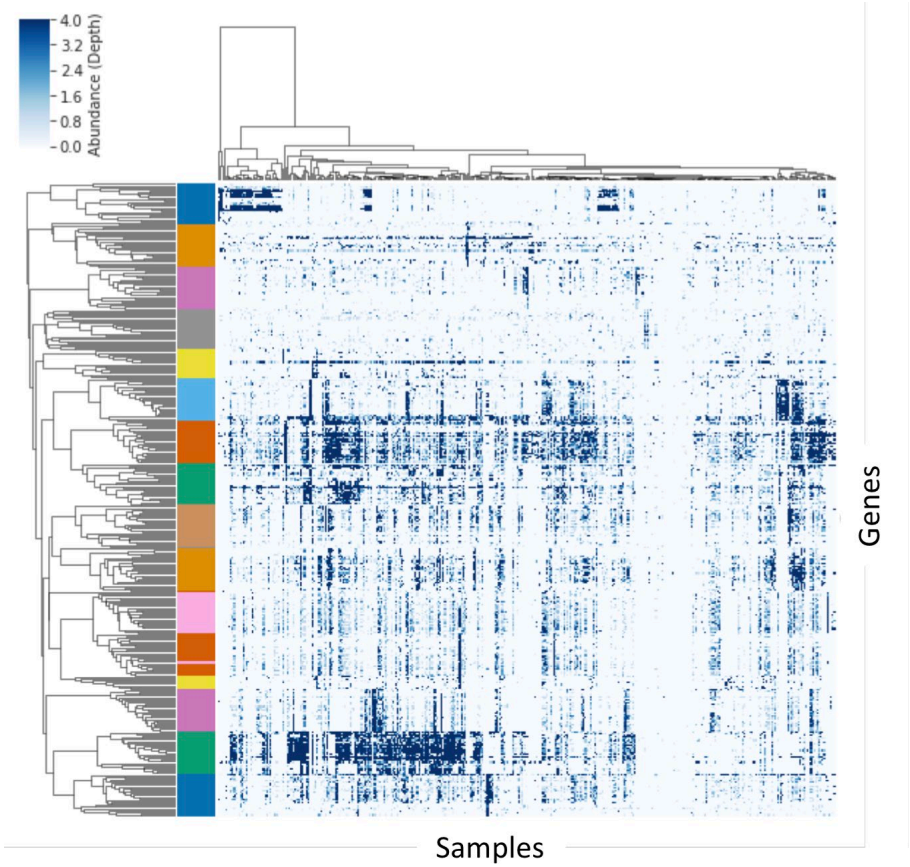
Co-Abundant Gene Groups (CAGs) contain functions required for bacterial life

Efficiently Comparing Gene-Level Communities

- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1

10X Genomics – Microbiome Data

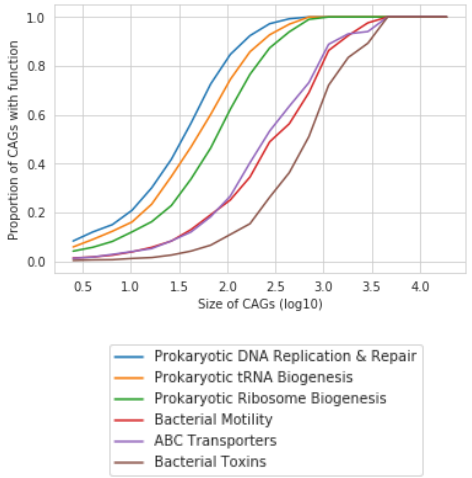


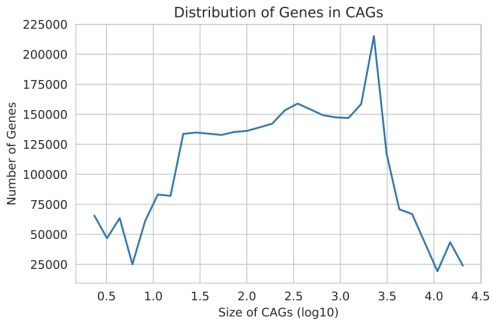


Co-Abundant Gene Groups (CAGs) co-occur in single-cell sequencing datasets

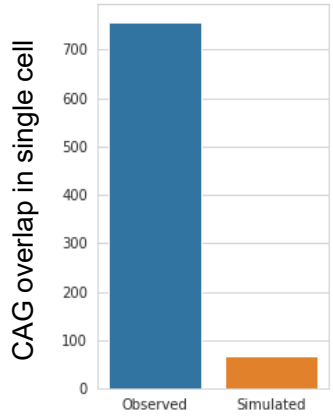
Efficiently Comparing Gene-Level Communities

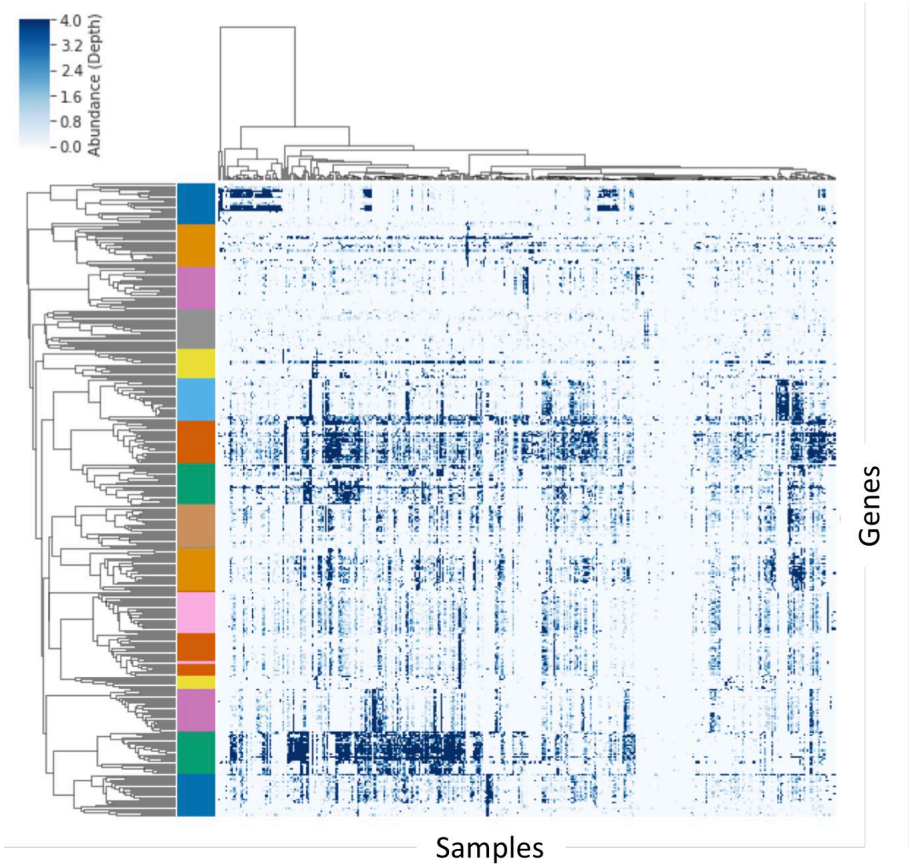
- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1





10X Genomics – Microbiome Data





Performance: 5M genes ~ 5,000 samples (240GB RAM) → ~10 hours

Efficiently Comparing Gene-Level Communities

- Source code: https://github.com/FredHutch/find-cags/
- Python package: `pip install ann_linkage_clustering`
- Docker container: quay.io/fhcrc-microbiome/find-cags
- Manuscript: biorxiv.org/content/10.1101/567818v1

Grouping *millions* of genes → *10,000's* of CAGs

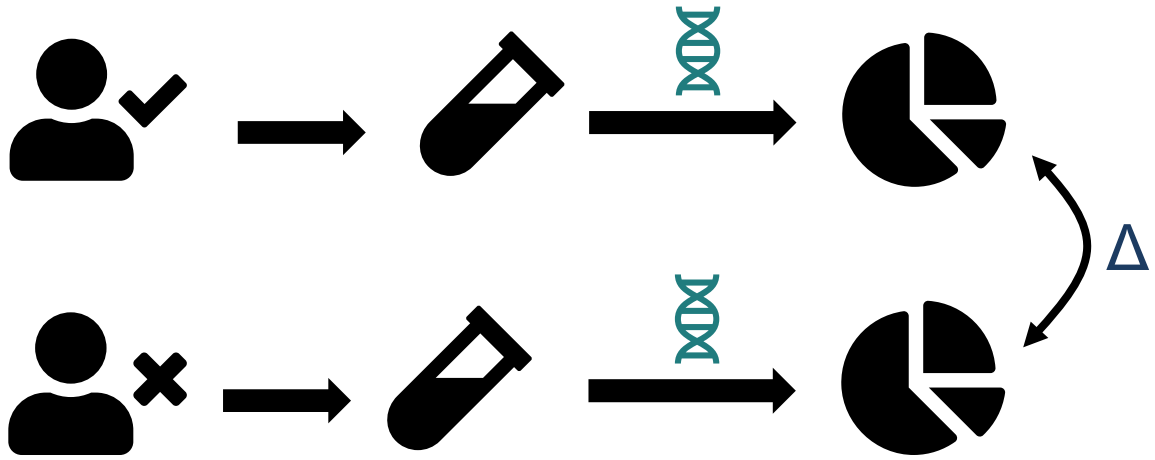
Enables efficient cross-sample comparison

Preserves biological complexity

Gene-Level Analysis of the Microbiome

1. Computational methods development
    - Single-sample analysis
    - Cross-sample analysis
2. **Application to CRC datasets**
3. Validation in mouse model

Identifying Strains Associated with CRC



nature medicine ARTICLES https://doi.org/10.1038/s41591-019-0405-7

Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation

Andrew Maltez Thomas[1,2,3,32], Paolo Manghi[1,32], Francesco Asnicar[1], Edoardo Pasolli[1], Federica Armanini[1], Moreno Zolfo[1], Francesco Beghini[1], Serena Manara[1], Nicolai Karcher[1], Chiara Pozzi[4], Sara Gandini[4], Davide Serrano[4], Sonia Tarallo[5], Antonio Francavilla[5], Gaetano Gallo[6,7], Mario Trompetto[7], Giulio Ferrero[8], Sayaka Mizutani[9,10], Hirotsugu Shiroma[9], Satoshi Shiba[11], Tatsuhiro Shibata[11,12], Shinichi Yachida[11,13], Takuji Yamada[9,14], Jakob Wirbel[15], Petra Schrotz-King[16], Cornelia M. Ulrich[17], Hermann Brenner[16,18,19], Manimozhiyan Arumugam[20,21], Peer Bork[15,22,23,24], Georg Zeller[15], Francesca Cordero[8], Emmanuel Dias-Neto[3,25], João Carlos Setubal[2,26], Adrian Tett[1], Barbara Pardini[5,27], Maria Rescigno[28], Levi Waldron[29,30,33], Alessio Naccarati[5,31,33] and Nicola Segata[1,33*]

nature medicine ARTICLES https://doi.org/10.1038/s41591-019-0406-6

Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer

Jakob Wirbel[1,31], Paul Theodor Pyl[2,3,31], Ece Kartal[1,4], Konrad Zych[1], Alireza Kashani[2], Alessio Milanese[1], Jonas S. Fleck[1], Anita Y. Voigt[1,5], Albert Palleja[2], Ruby Ponnudurai[2], Shinichi Sunagawa[1,6], Luis Pedro Coelho[1,30], Petra Schrotz-King[7], Emily Vogtmann[8], Nina Habermann[9], Emma Niméus[3,10], Andrew M. Thomas[11,12], Paolo Manghi[11], Sara Gandini[13], Davide Serrano[13], Sayaka Mizutani[14,15], Hirotsugu Shiroma[14], Satoshi Shiba[16], Tatsuhiro Shibata[16,17], Shinichi Yachida[16,18], Takuji Yamada[14,19], Levi Waldron[20,21], Alessio Naccarati[22,23], Nicola Segata[11], Rashmi Sinha[8], Cornelia M. Ulrich[24], Hermann Brenner[7,25,26], Manimozhiyan Arumugam[2,27,32*], Peer Bork[1,4,28,29,32*] and Georg Zeller[1,32*]

Discovery cohort:

Zeller (2014) Molecular Systems Biology

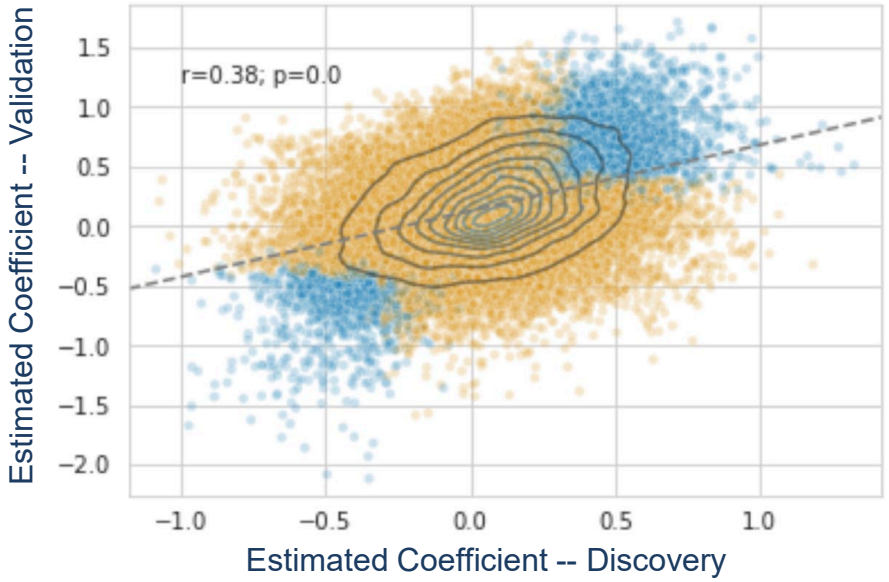
156 participants – France

Validation cohort:

Yu (2015) Gut

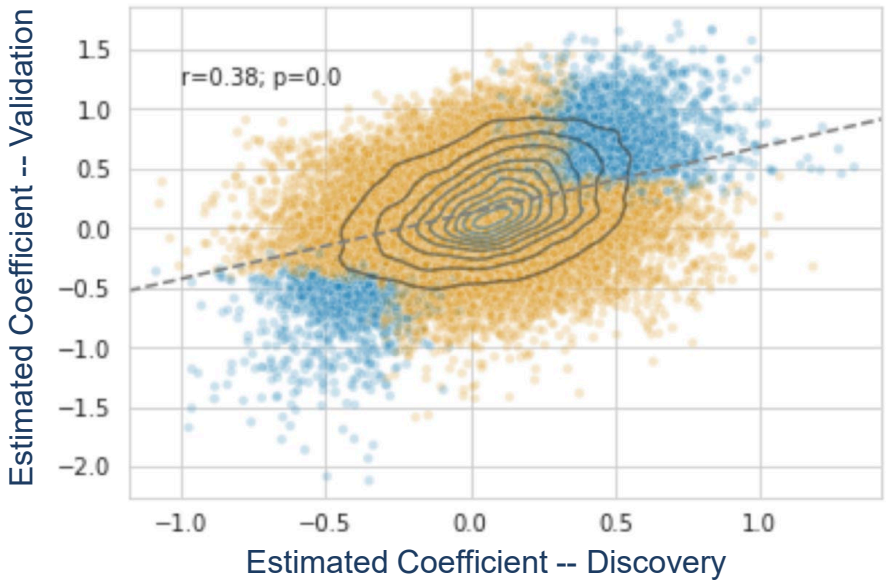
128 participants – China

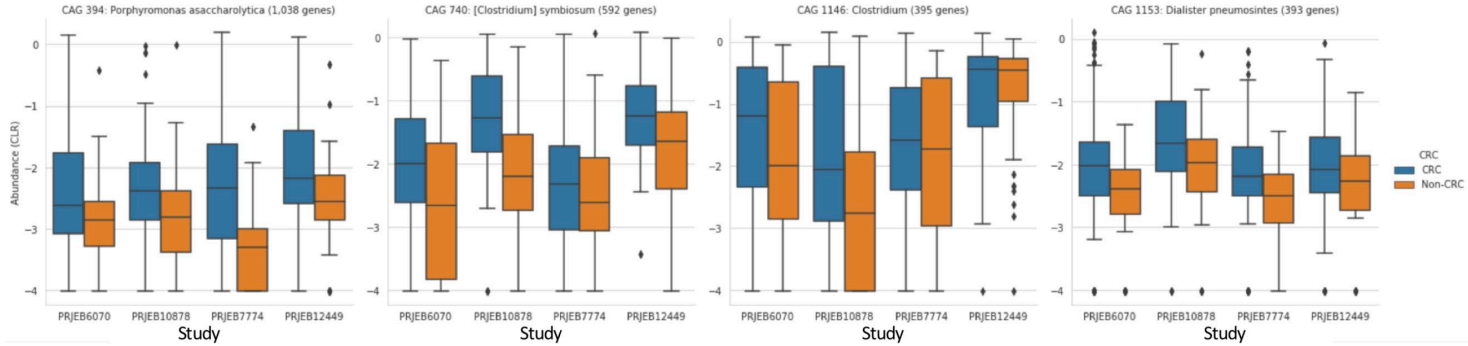
Identifying Strains Associated with CRC



Association of CAG abundance with CRC is **highly reproducible**

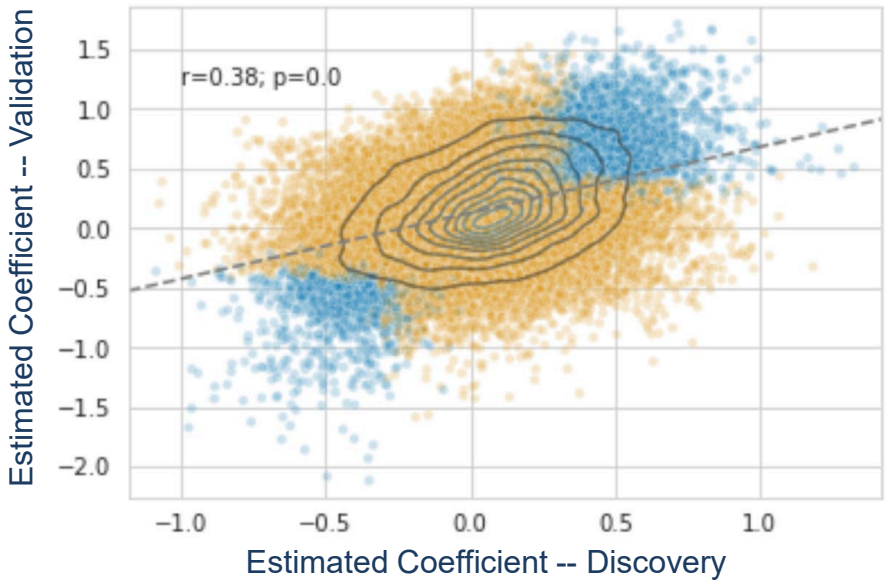
Identifying Strains Associated with CRC

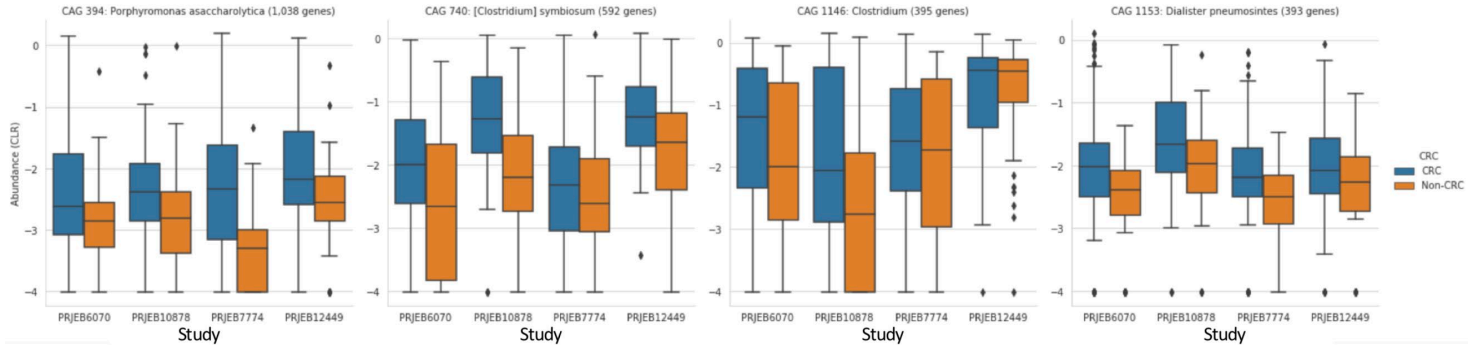




Consistent association with CRC across **four independent cohorts**

Identifying Strains Associated with CRC





Disease → CAGs

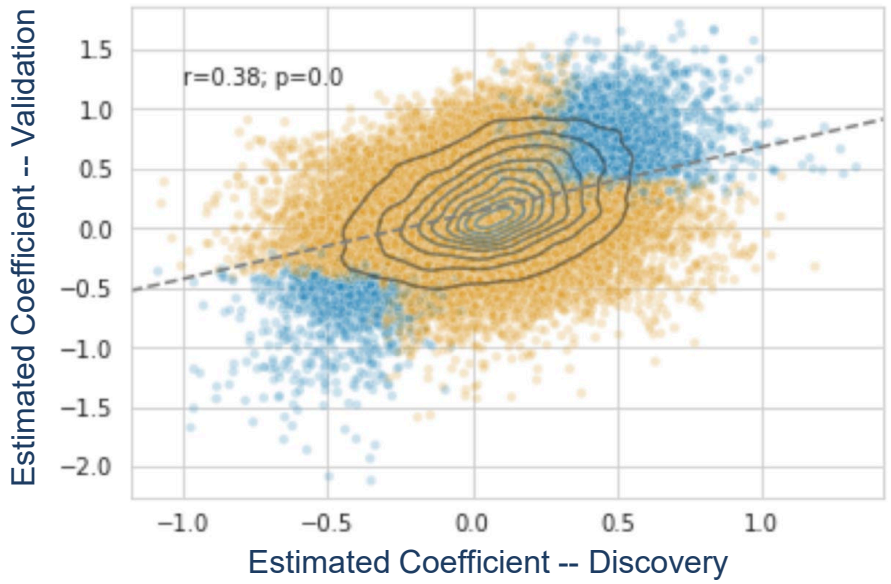
CAGs → Genes

Genes → Function (KEGG)

Genes → Taxonomy (NCBI)

Genes → Genomes (ATCC)

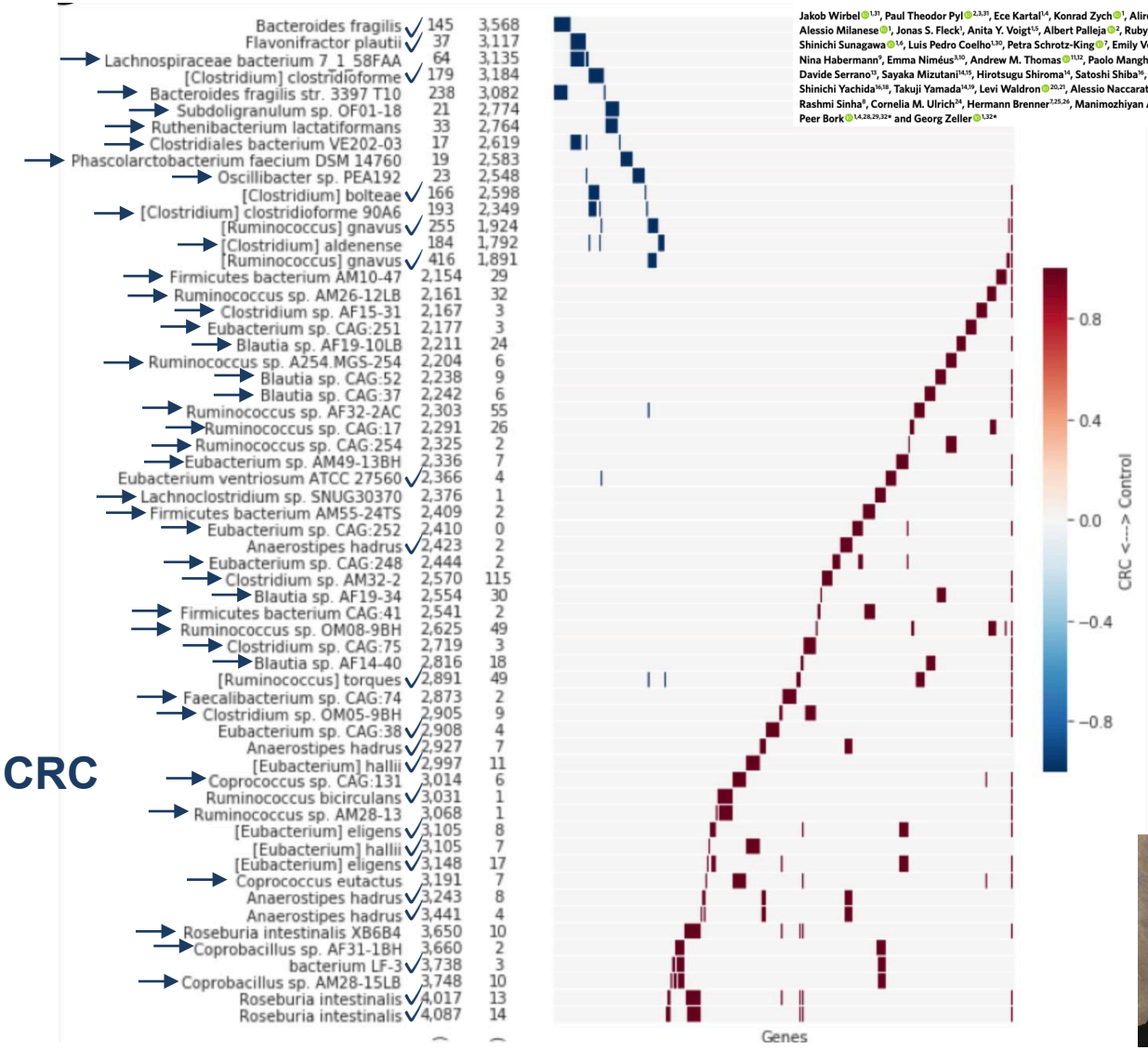
Identifying Strains Associated with CRC



nature medicine ARTICLES https://doi.org/10.1038/s41591-019-0406-6

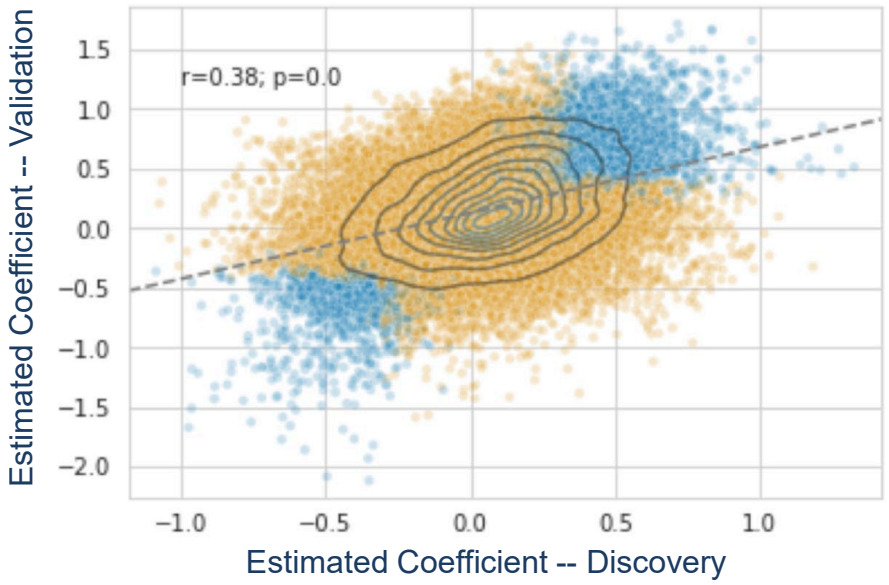
Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer

Jakob Wirbel[1,3†], Paul Theodor Pyl[2,3,3†], Ece Kartal[1,4], Konrad Zych[1], Alireza Kashani[2], Alessio Milanese[1], Jonas S. Fleck[1], Anita Y. Voigt[1,5], Albert Palleja[2], Ruby Ponnudurai[1], Shinichi Sunagawa[1,6], Luis Pedro Coelho[1,30], Petra Schrotz-King[7], Emily Vogtmann[8], Nina Habermann[9], Emma Niméus[3,10], Andrew M. Thomas[11,12], Paolo Manghi[11], Sara Gandini[13], Davide Serrano[13], Sayaka Mizutani[14,15], Hirotsugu Shiroma[14], Satoshi Shiba[16], Tatsuhiro Shibata[16,17], Shinichi Yachida[16,18], Takuji Yamada[14,19], Levi Waldron[20,21], Alessio Naccarati[22,23], Nicola Segata[11], Rashmi Sinha[8], Cornelia M. Ulrich[24], Hermann Brenner[7,25,26], Manimozhiyan Arumugam[2,27,32*], Peer Bork[1,4,28,29,32*] and Georg Zeller[1,32*]

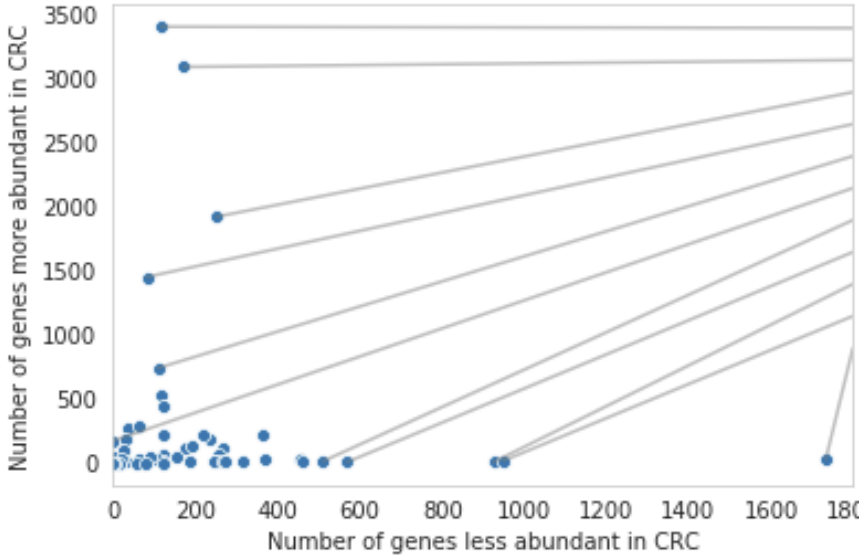


Genes identify microbial strains associated with CRC

Identifying Strains Associated with CRC



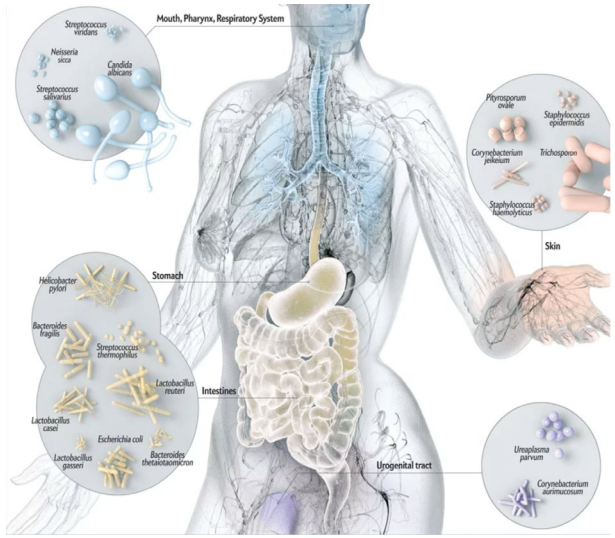
Experimental Validation – Neel Dey MD (Fred Hutch)



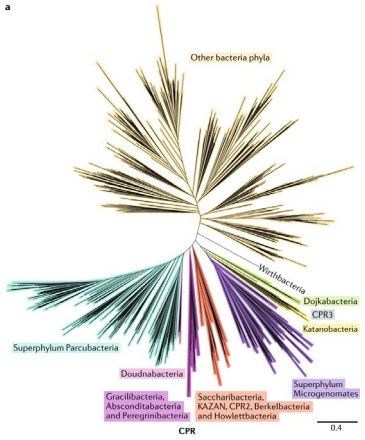
Bacteroides fragilis strain A
Clostridium bolteae strain B
Ruminococcus gnavus strain C
Clostridium symbiosum strain D
Clostridium asparagiforme strain E
Fusobacterium nucleatum strain F
Dorea longicatena strain G
Bifidobacterium psudocatenulatum strain H
Eubacterium rectale strain I
Ruminococcus obeum strain J
Coprococcus comes strain K

Bacterial Isolate → Mouse Model (*apc*) → Cancer Phenotype

Gene-Level Metagenomics for Microbiome Research

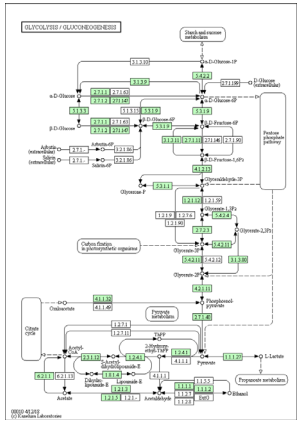


Microbial Taxa

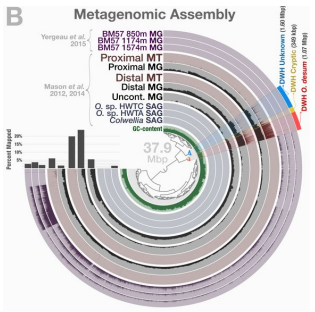


Microbial genes

Metabolic Pathways



Assembled Genomes



Scientific American (2012-05-15)
Castelle, et al. 2018
Eren, et al. 2015
genome.jp/kegg

FRED HUTCH

THANK YOU

FRED HUTCH
CURES START HERE®

fredhutch.org


CREDIBLE
STANDARDS
INCREDIBLE
RESULTS

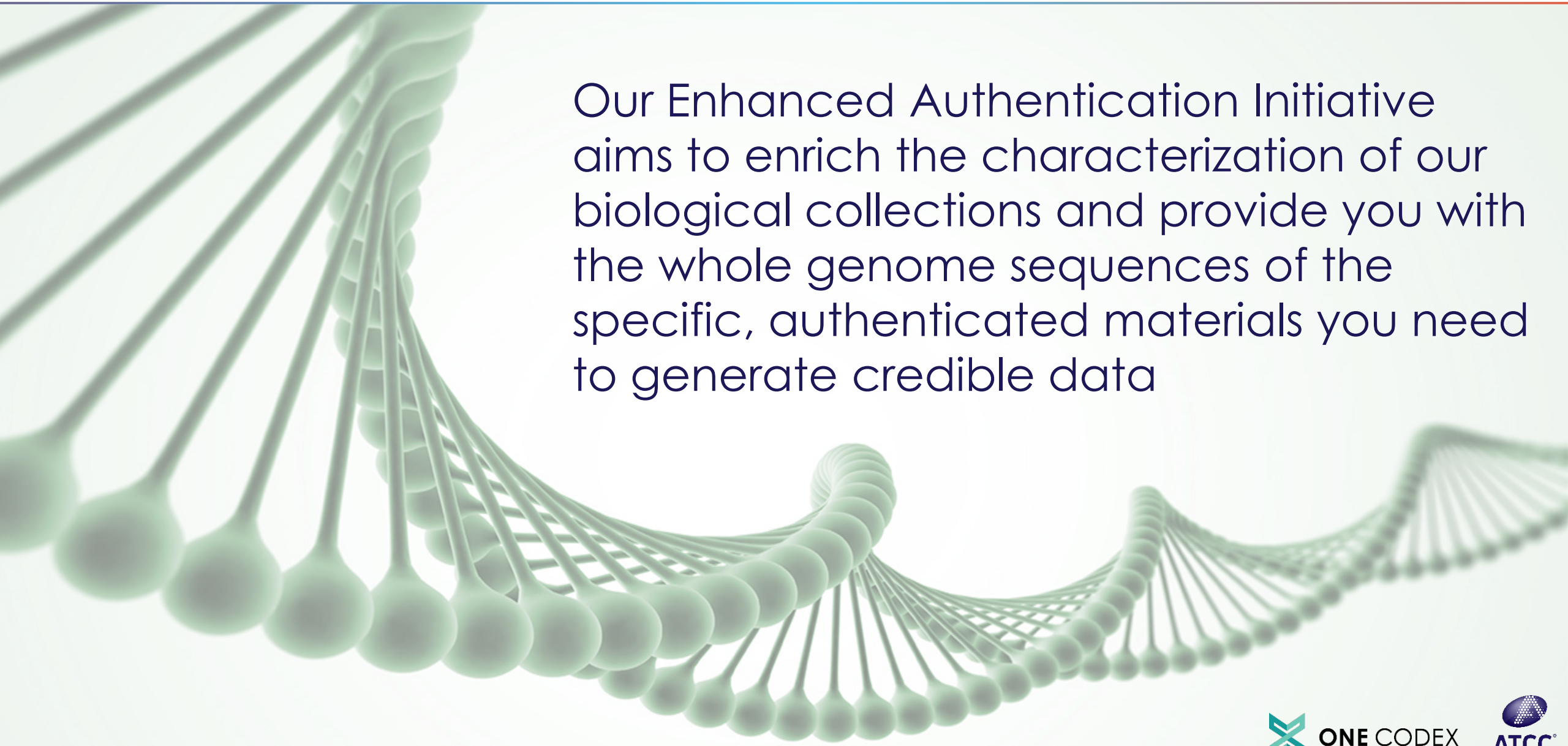
Advancing Authentication

Whole genome sequencing has generated data at an unprecedented scale in the biological sciences. However, existing public genomic databases can lack quality, completeness, authenticity, and traceability.

ONE CODEX

ATCC

Advancing Authentication



Our Enhanced Authentication Initiative aims to enrich the characterization of our biological collections and provide you with the whole genome sequences of the specific, authenticated materials you need to generate credible data

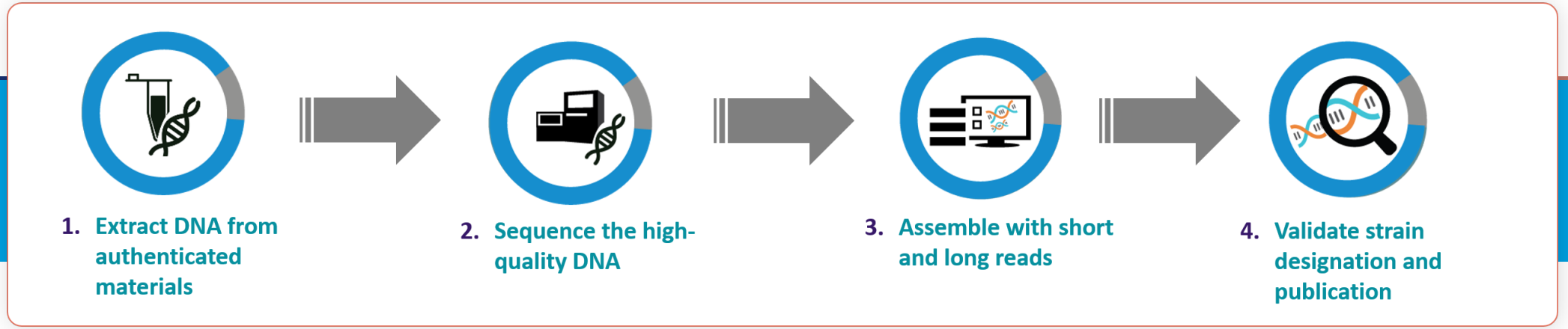
ONE CODEX

ATCC

We are giving you the first look at our

ATCC Genome Portal

Your resource for reference-quality genomes from authenticated ATCC products



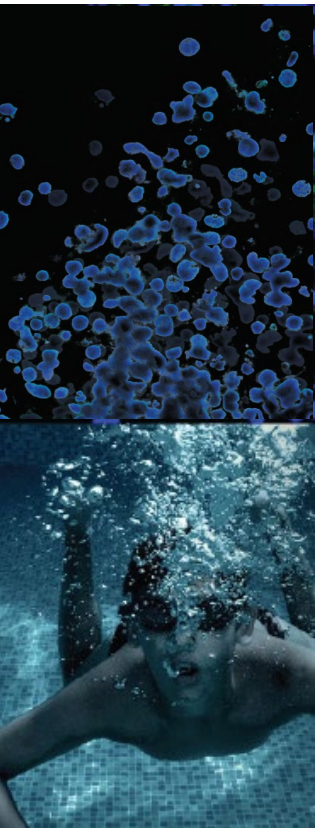
ONE CODEX

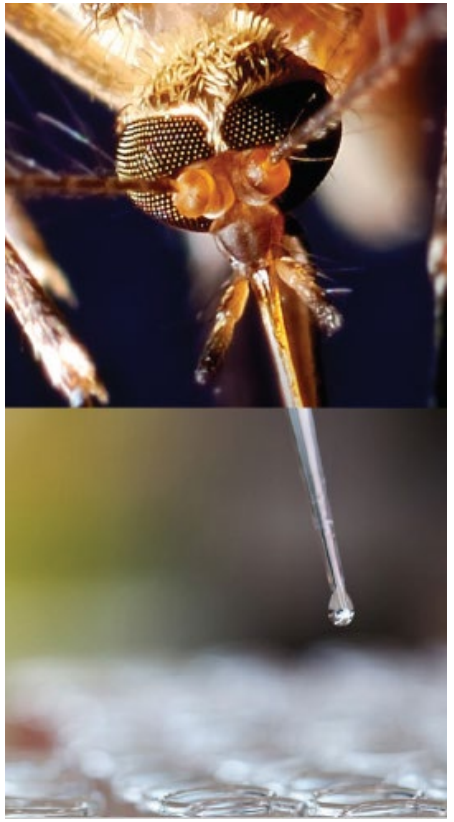
ATCC

Enhanced Authentication Initiative

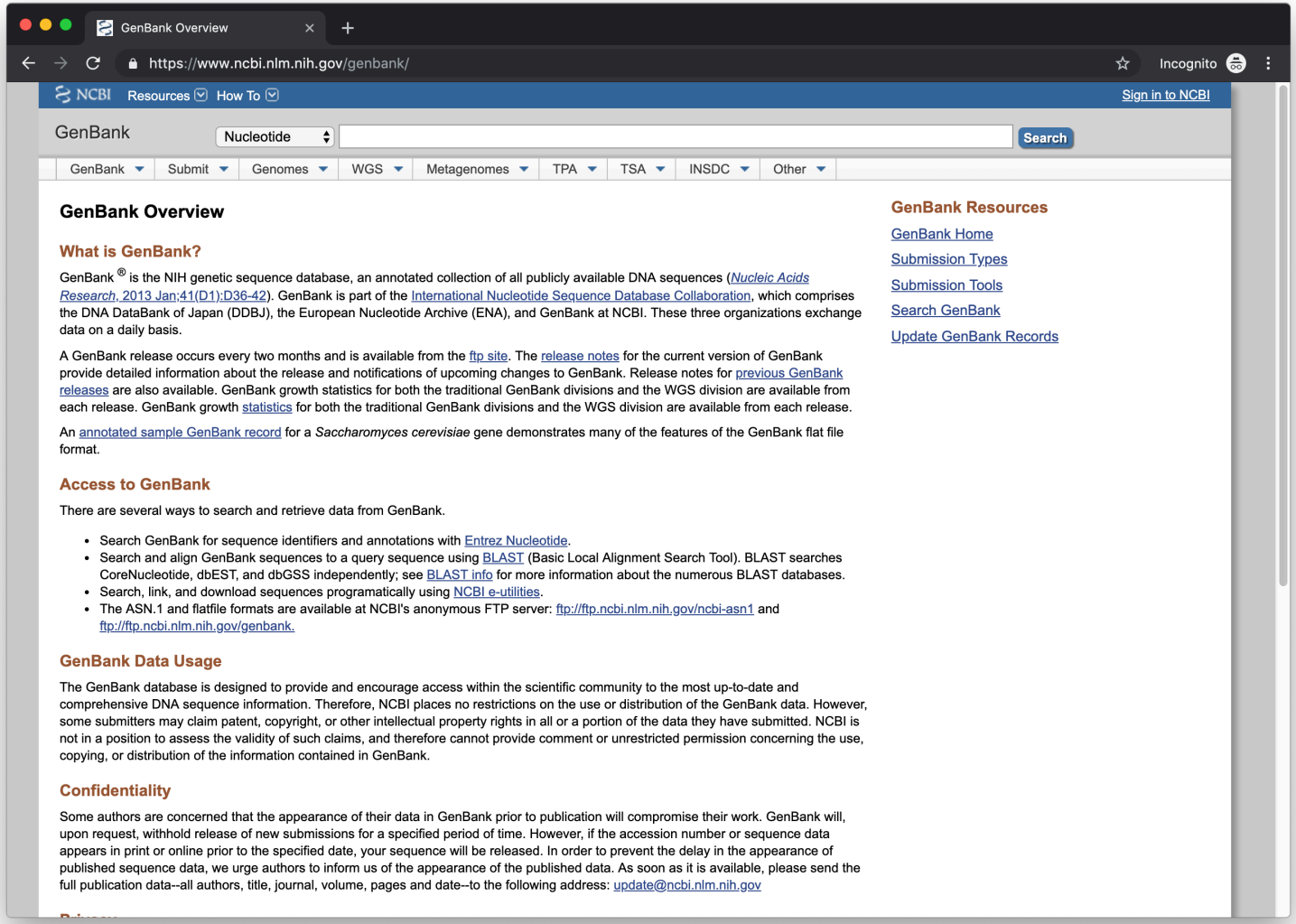
Nick Greenfield, MA
Founder and CEO, One Codex

Credible Leads to Incredible™





Current microbial genomics references



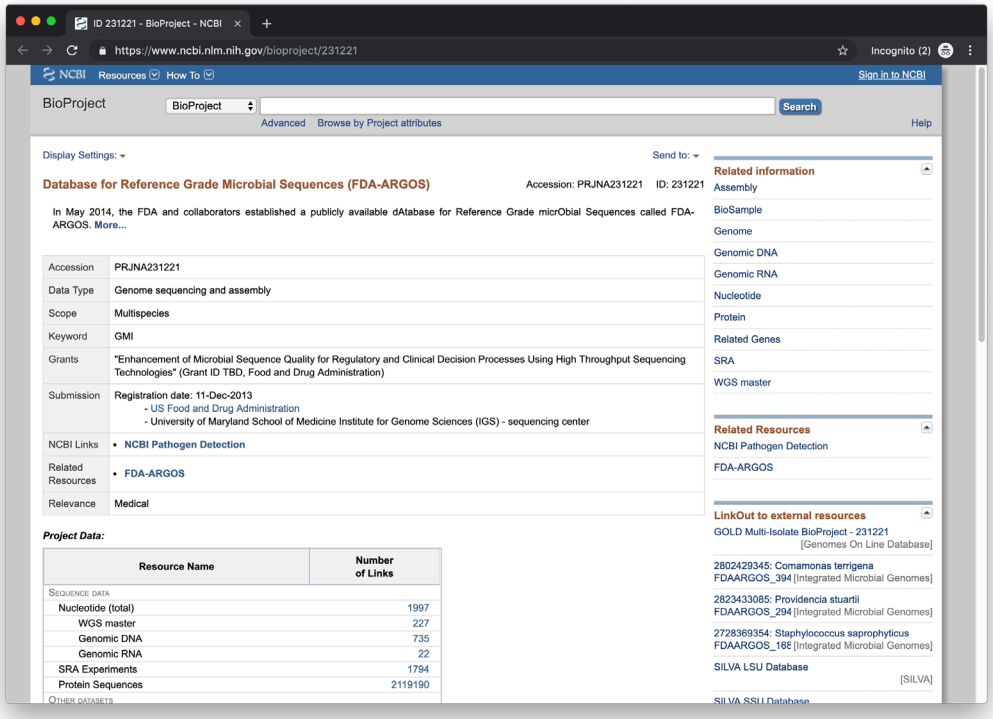
GenBank®

- De facto standard
- *Lots* of genomes
- But relatively little curation
- And highly variable quality

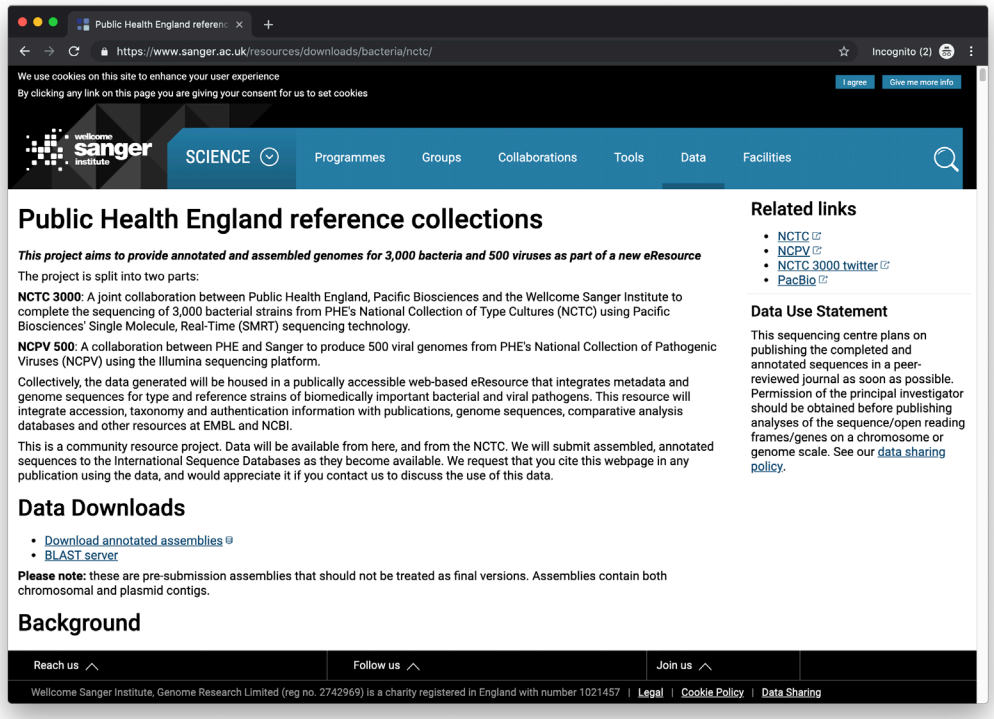
ONE CODEX ATCC®

Current microbial genomics references

FDA ARGOS



NCTC

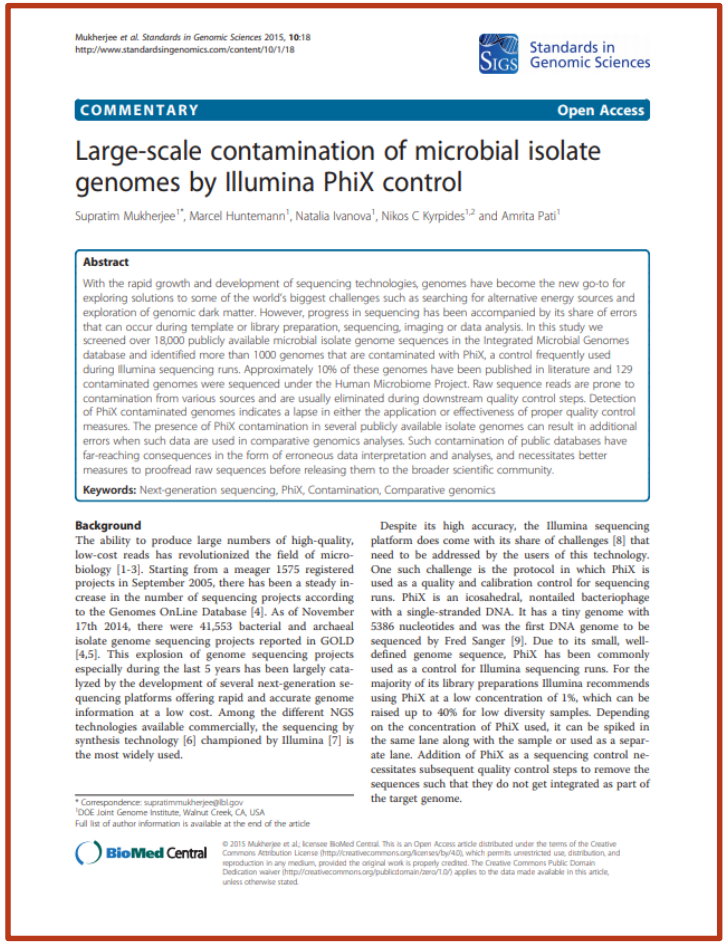


… and numerous other specialty collections

ONE CODEX ATCC

Current database challenges

#1 Quality



Mukherjee *et al. Standards in Genomic Sciences* 2015, **10**:18
http://www.standardsingenomics.com/content/10/1/18

Standards in Genomic Sciences

COMMENTARY Open Access

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee[1*], Marcel Huntemann[1], Natalia Ivanova[1], Nikos C Kyrpides[1,2] and Amrita Pati[1]

Abstract

With the rapid growth and development of sequencing technologies, genomes have become the new go-to for exploring solutions to some of the world's biggest challenges such as searching for alternative energy sources and exploration of genomic dark matter. However, progress in sequencing has been accompanied by its share of errors that can occur during template or library preparation, sequencing, imaging or data analysis. In this study we screened over 18,000 publicly available microbial isolate genome sequences in the Integrated Microbial Genomes database and identified more than 1000 genomes that are contaminated with PhiX, a control frequently used during Illumina sequencing runs. Approximately 10% of these genomes have been published in literature and 129 contaminated genomes were sequenced under the Human Microbiome Project. Raw sequence reads are prone to contamination from various sources and are usually eliminated during downstream quality control steps. Detection of PhiX contaminated genomes indicates a lapse in either the application or effectiveness of proper quality control measures. The presence of PhiX contamination in several publicly available isolate genomes can result in additional errors when such data are used in comparative genomics analyses. Such contamination of public databases have far-reaching consequences in the form of erroneous data interpretation and analyses, and necessitates better measures to proofread raw sequences before releasing them to the broader scientific community.

Keywords: Next-generation sequencing, PhiX, Contamination, Comparative genomics

Background

The ability to produce large numbers of high-quality, low-cost reads has revolutionized the field of microbiology [1-3]. Starting from a meager 1575 registered projects in September 2005, there has been a steady increase in the number of sequencing projects according to the Genomes OnLine Database [4]. As of November 17th 2014, there were 41,553 bacterial and archaeal isolate genome sequencing projects reported in GOLD [4,5]. This explosion of genome sequencing projects especially during the last 5 years has been largely catalyzed by the development of several next-generation sequencing platforms offering rapid and accurate genome information at a low cost. Among the different NGS technologies available commercially, the sequencing by synthesis technology [6] championed by Illumina [7] is the most widely used.

Despite its high accuracy, the Illumina sequencing platform does come with its share of challenges [8] that need to be addressed by the users of this technology. One such challenge is the protocol in which PhiX is used as a quality and calibration control for sequencing runs. PhiX is an icosahedral, nontailed bacteriophage with a single-stranded DNA. It has a tiny genome with 5386 nucleotides and was the first DNA genome to be sequenced by Fred Sanger [9]. Due to its small, well-defined genome sequence, PhiX has been commonly used as a control for Illumina sequencing runs. For the majority of its library preparations Illumina recommends using PhiX at a low concentration of 1%, which can be raised up to 40% for low diversity samples. Depending on the concentration of PhiX used, it can be spiked in the same lane along with the sample or used as a separate lane. Addition of PhiX as a sequencing control necessitates subsequent quality control steps to remove the sequences such that they do not get integrated as part of the target genome.

* Correspondence: supratimmukherjee@lbl.gov
[1]DOE Joint Genome Institute, Walnut Creek, CA, USA
Full list of author information is available at the end of the article

"…[we] identified more than 1000 genomes that are contaminated with PhiX…"

ONE CODEX ATCC

Current database challenges

#1 Quality

Research

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,[1] Mihaela Pertea,[1,2] Aleksey V. Zimin,[1,3] and Steven L. Salzberg[1,2,3,4]

[1]*Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [3]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [4]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA*

Contaminant sequences that appear in published genomes can cause numerous problems for downstream analyses, particularly for evolutionary studies and metagenomics projects. Our large-scale scan of complete and draft bacterial and archaeal genomes in the NCBI RefSeq database reveals that 2250 genomes are contaminated by human sequence. The contaminant sequences derive primarily from high-copy human repeat regions, which themselves are not adequately represented in the current human reference genome, GRCh38. The absence of the sequences from the human assembly offers a likely explanation for their presence in bacterial assemblies. In some cases, the contaminating contigs have been erroneously annotated as containing protein-coding sequences, which over time have propagated to create spurious protein "families" across multiple prokaryotic and eukaryotic genomes. As a result, 3437 spurious protein entries are currently present in the widely used nr and TrEMBL protein databases. We report here an extensive list of contaminant sequences in bacterial genome assemblies and the proteins associated with them. We found that nearly all contaminants occurred in small contigs in draft genomes, which suggests that filtering out small contigs from draft genome assemblies may mitigate the issue of contamination while still keeping nearly all of the genuine genomic sequences.

[Supplemental material is available for this article.]

Over the past two decades, the number of publicly available genomes has grown from just a handful of species to well over 100,000 genomes today. These genomes are pivotal resources for countless biomedical research questions, including microbiome studies that use them to identify species in complex samples (Breitwieser et al. 2017). Ideally, all genomes in reference databases would be complete and accurate (Fraser et al. 2002), but for practical reasons, the vast majority of genomes available today are still "drafts." A draft genome consists of multiple contigs or scaffolds that are typically unordered and not assigned into chromosomes (Ghurye et al. 2016). A genome is not truly complete or "finished" until every base pair has been determined for every chromosome and organelle, end-to-end, with no gaps. Even the human genome, although far more complete than most other animal genomes, is still unfinished: The current human assembly, GRCh38.p13 (released Feb. 28, 2019), has 473 scaffolds that contain 875 internal gaps. While most of the human sequence has been placed on chromosomes, some highly repetitive regions are underrepresented (Altemose et al. 2014), leading to problems that we discuss below. Draft genomes of other species vary widely in quality as well as contiguity, with some having thousands of contigs and others having a much smaller number.

Contamination of genome assemblies with sequences from other species is not uncommon, especially in draft genomes (Longo et al. 2011; Merchant et al. 2014; Delmont and Eren 2016; Kryukov and Imanishi 2016; Lu and Salzberg 2018). In 2011, researchers reported that over 10% of selected nonprimate assemblies in the NCBI and UCSC Genome Browser databases were contaminated with the primate-specific *AluY* repeats (Longo et al. 2011). Although validation pipelines have improved substantially since then (Tatusova et al. 2016; Haft et al. 2018), some contaminants still remain, as we describe below. Furthermore, when open reading frames (ORFs) in the contaminated contigs get annotated as protein-coding genes, their protein sequence may be added to other databases. Once in those databases, these spurious proteins may in turn be used in future annotation, leading to the so-called "transitive catastrophe" problem where errors are propagated widely (Karp 1998; Salzberg 2007; Danchin et al. 2018). Indeed, one study found that the percentage of misannotated entries in the NCBI nonredundant (nr) protein collection, which is used for thousands of BLAST searches every day, has been increasing over time (Schnoes et al. 2009).

Contamination of genomic sequences can be particularly problematic for metagenomic studies. For example, if a genome labeled as species X contains fragments of the human genome, then any sample containing human DNA might erroneously be identified as also containing species X. Since human DNA is virtually always present in the environment of sequencing laboratories, human contamination is very common in sequencing experiments of all types. Contamination of laboratory reagents with DNA from other organisms can also lead to serious misinterpretations, such as the supposed detection of the novel virus NIH-CQV

Corresponding authors: florian.bw@gmail.com, salzberg@jhu.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.245373.118.

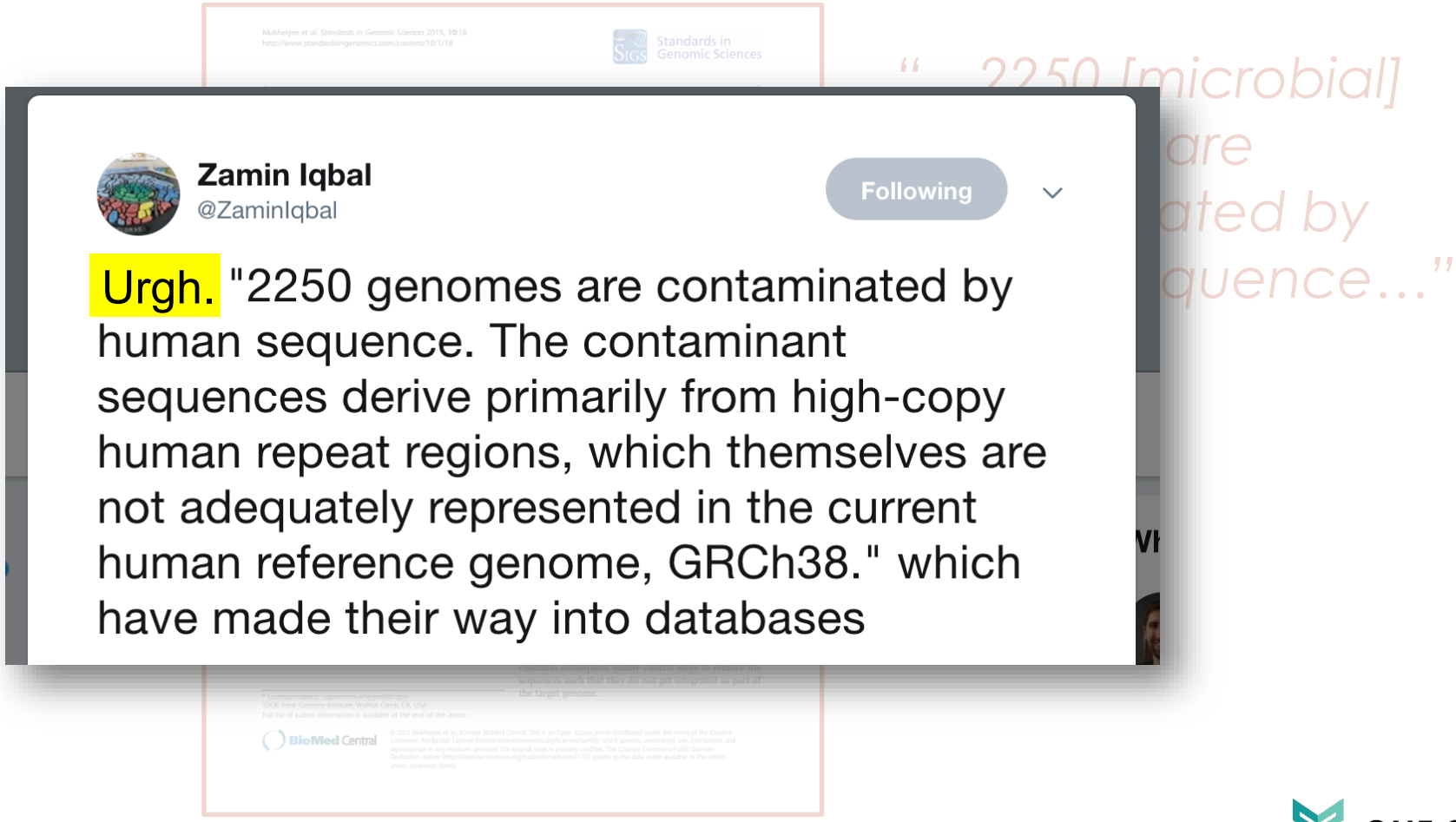
29:954–960 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

"…2250 [microbial] genomes are contaminated by human sequence…"

ONE CODEX ATCC

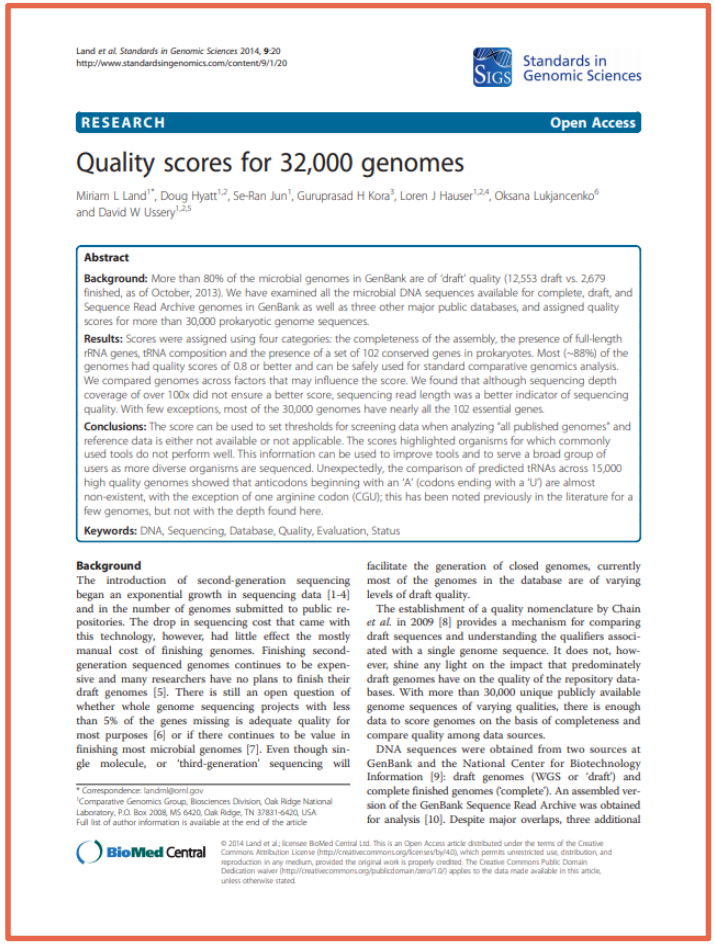
Current database challenges

#1 Quality



Current database challenges

#2 Completeness



Land *et al. Standards in Genomic Sciences* 2014, **9**:20
http://www.standardsingenomics.com/content/9/1/20

Standards in Genomic Sciences

RESEARCH Open Access

Quality scores for 32,000 genomes

Miriam L Land[1*], Doug Hyatt[1,2], Se-Ran Jun[1], Guruprasad H Kora[3], Loren J Hauser[1,2,4], Oksana Lukjancenko[6] and David W Ussery[1,2,5]

Abstract

Background: More than 80% of the microbial genomes in GenBank are of 'draft' quality (12,553 draft vs. 2,679 finished, as of October, 2013). We have examined all the microbial DNA sequences available for complete, draft, and Sequence Read Archive genomes in GenBank as well as three other major public databases, and assigned quality scores for more than 30,000 prokaryotic genome sequences.

Results: Scores were assigned using four categories: the completeness of the assembly, the presence of full-length rRNA genes, tRNA composition and the presence of a set of 102 conserved genes in prokaryotes. Most (~88%) of the genomes had quality scores of 0.8 or better and can be safely used for standard comparative genomics analysis. We compared genomes across factors that may influence the score. We found that although sequencing depth coverage of over 100x did not ensure a better score, sequencing read length was a better indicator of sequencing quality. With few exceptions, most of the 30,000 genomes have nearly all the 102 essential genes.

Conclusions: The score can be used to set thresholds for screening data when analyzing "all published genomes" and reference data is either not available or not applicable. The scores highlighted organisms for which commonly used tools do not perform well. This information can be used to improve tools and to serve a broad group of users as more diverse organisms are sequenced. Unexpectedly, the comparison of predicted tRNAs across 15,000 high quality genomes showed that anticodons beginning with an 'A' (codons ending with a 'U') are almost non-existent, with the exception of one arginine codon (CGU); this has been noted previously in the literature for a few genomes, but not with the depth found here.

Keywords: DNA, Sequencing, Database, Quality, Evaluation, Status

Background

The introduction of second-generation sequencing began an exponential growth in sequencing data [1-4] and in the number of genomes submitted to public repositories. The drop in sequencing cost that came with this technology, however, had little effect the mostly manual cost of finishing genomes. Finishing second-generation sequenced genomes continues to be expensive and many researchers have no plans to finish their draft genomes [5]. There is still an open question of whether whole genome sequencing projects with less than 5% of the genes missing is adequate quality for most purposes [6] or if there continues to be value in finishing most microbial genomes [7]. Even though single molecule, or 'third-generation' sequencing will facilitate the generation of closed genomes, currently most of the genomes in the database are of varying levels of draft quality.

The establishment of a quality nomenclature by Chain *et al.* in 2009 [8] provides a mechanism for comparing draft sequences and understanding the qualifiers associated with a single genome sequence. It does not, however, shine any light on the impact that predominately draft genomes have on the quality of the repository databases. With more than 30,000 unique publicly available genome sequences of varying qualities, there is enough data to score genomes on the basis of completeness and compare quality among data sources.

DNA sequences were obtained from two sources at GenBank and the National Center for Biotechnology Information [9]: draft genomes (WGS or 'draft') and complete finished genomes ('complete'). An assembled version of the GenBank Sequence Read Archive was obtained for analysis [10]. Despite major overlaps, three additional

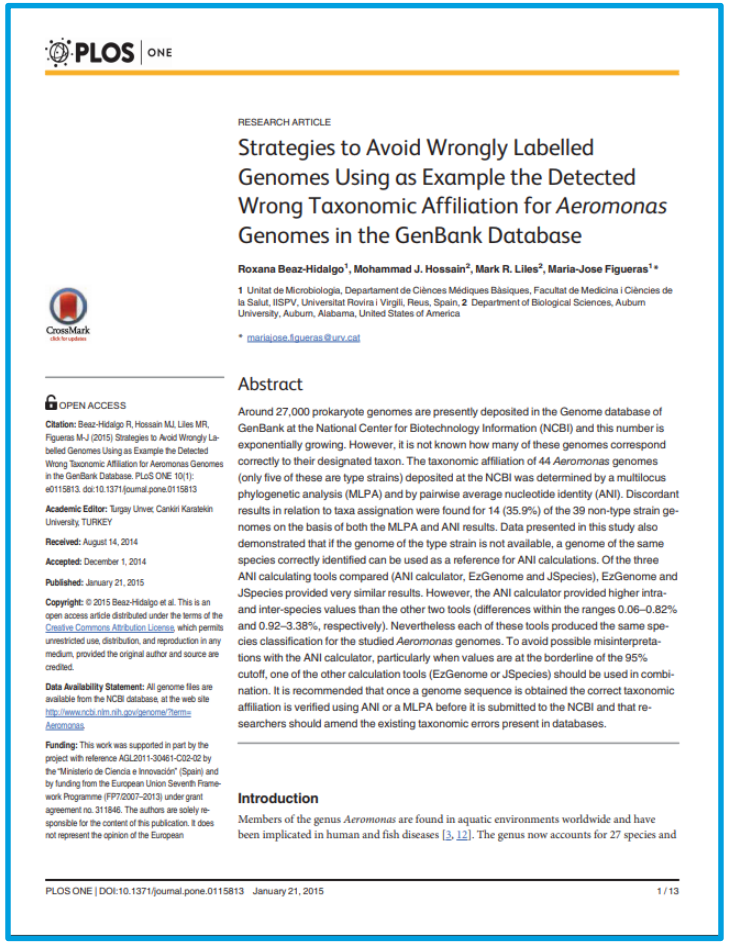
* Correspondence: landml@ornl.gov
[1]Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6420, Oak Ridge, TN 37831-6420, USA
Full list of author information is available at the end of the article

"More than 80% of the microbial genomes in GenBank are of 'draft' quality…"

ONE CODEX ATCC

Current database challenges

#3 Authenticity



PLOS | ONE

RESEARCH ARTICLE

Strategies to Avoid Wrongly Labelled Genomes Using as Example the Detected Wrong Taxonomic Affiliation for *Aeromonas* Genomes in the GenBank Database

Roxana Beaz-Hidalgo[1], Mohammad J. Hossain[2], Mark R. Liles[2], Maria-Jose Figueras[1]*

1 Unitat de Microbiologia, Departament de Ciències Mèdiques Bàsiques, Facultat de Medicina i Ciències de la Salut, IISPV, Universitat Rovira i Virgili, Reus, Spain, **2** Department of Biological Sciences, Auburn University, Auburn, Alabama, United States of America

* mariajose.figueras@urv.cat

OPEN ACCESS

Citation: Beaz-Hidalgo R, Hossain MJ, Liles MR, Figueras M-J (2015) Strategies to Avoid Wrongly Labelled Genomes Using as Example the Detected Wrong Taxonomic Affiliation for Aeromonas Genomes in the GenBank Database. PLoS ONE 10(1): e0115813. doi:10.1371/journal.pone.0115813

Academic Editor: Turgay Unver, Cankiri Karatekin University, TURKEY

Received: August 14, 2014

Accepted: December 1, 2014

Published: January 21, 2015

Data Availability Statement: All genome files are available from the NCBI database, at the web site http://www.ncbi.nlm.nih.gov/genome/?term=Aeromonas

Funding: This work was supported in part by the project with reference AGL2011-30461-C02-02 by the "Ministerio de Ciencia e Innovación" (Spain) and by funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 311846. The authors are solely responsible for the content of this publication. It does not represent the opinion of the European

Abstract

Around 27,000 prokaryote genomes are presently deposited in the Genome database of GenBank at the National Center for Biotechnology Information (NCBI) and this number is exponentially growing. However, it is not known how many of these genomes correspond correctly to their designated taxon. The taxonomic affiliation of 44 *Aeromonas* genomes (only five of these are type strains) deposited at the NCBI was determined by a multilocus phylogenetic analysis (MLPA) and by pairwise average nucleotide identity (ANI). Discordant results in relation to taxa assignation were found for 14 (35.9%) of the 39 non-type strain genomes on the basis of both the MLPA and ANI results. Data presented in this study also demonstrated that if the genome of the type strain is not available, a genome of the same species correctly identified can be used as a reference for ANI calculations. Of the three ANI calculating tools compared (ANI calculator, EzGenome and JSpecies), EzGenome and JSpecies provided very similar results. However, the ANI calculator provided higher intra- and inter-species values than the other two tools (differences within the ranges 0.06–0.82% and 0.92–3.38%, respectively). Nevertheless each of these tools produced the same species classification for the studied *Aeromonas* genomes. To avoid possible misinterpretations with the ANI calculator, particularly when values are at the borderline of the 95% cutoff, one of the other calculation tools (EzGenome or JSpecies) should be used in combination. It is recommended that once a genome sequence is obtained the correct taxonomic affiliation is verified using ANI or a MLPA before it is submitted to the NCBI and that researchers should amend the existing taxonomic errors present in databases.

Introduction

Members of the genus *Aeromonas* are found in aquatic environments worldwide and have been implicated in human and fish diseases [3, 12]. The genus now accounts for 27 species and

"Discordant results in relation to taxa assignation were found for 14 (35.9%) of the 39 non-type strain genomes…"

Current database challenges

#4 Traceability

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

LETTER TO THE EDITOR

Beware of False "Type Strain" Genome Sequences

Francisco Salvà-Serra,[a,b,c,d,e] Daniel Jaén-Luchoro,[a,b,c,d] Roger Karlsson,[a,b,d,f] Antoni Bennasar-Figueras,[a,g] Hedvig E. Jakobsson,[b,d] Edward R. B. Moore[a,b,c,d]

[a]Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[b]Clinical Microbiology, Sahlgrenska University Hospital, Gothenburg, Sweden
[c]Culture Collection University of Gothenburg, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[d]Centre for Antibiotic Resistance Research (CARe), University of Gothenburg, Gothenburg, Sweden
[e]Microbiology, Department of Biology, University of the Balearic Islands, Palma de Mallorca, Spain
[f]Nanoxis Consulting AB, Gothenburg, Sweden
[g]Area of Infectious Diseases, Research Institute of Health Sciences (IUNICS-UIB), University of the Balearic Islands, Palma de Mallorca, Spain

With this letter, we warn users of bacterial DNA sequence data about recent cases misusing the term "type strain" in bacterial genome sequence reports and highlight the importance that the term is used in the correct context.

In recent articles published in the journal *Genome Announcements* (GA), (now *Microbiology Resource Announcements* [MRA]) (1–5), the complete genome sequences of five strains of five species of bacteria were reported. The titles of the articles stated that the strains represent the respective "type strains" of the five species, although the articles do not present details about the processes of defining the type strains. At this point, it is important to point out that the concept and description of "type strain" are not arbitrary; the type strains of bacterial species are defined by rule 18a of the International Code of Nomenclature of Prokaryotes as follows: "The type strain is made up of living cultures of an organism, which are descended from a strain designated as the nomenclatural type" (6). Strains serving as nomenclatural type material are designated precisely, and this may be done in various ways (7); typically, a "holotype strain" is designated at the time of valid publication of a new species name (rule 18b), in which the type strain must conform to defined conditions (rule 30). The five species described in the GA publications had been validly published previously, and the type strains for those species were already defined, preserved, and publicly available in numerous culture collections. Meanwhile, the strains reported in the five GA articles were recently isolated independently. Therefore, since they are not descended from the already defined type strains, they cannot represent the authentic type strains of the species.

To confirm these observations, we performed average nucleotide identity (ANI) analyses using JSpeciesWS (8, 9) between the five genome sequences reported in the GA articles and the genome sequences of the documented type strains of the five species (Table 1). The ANI values range from 83.10 to 98.96% (percentages aligned, 75.56 to 91.18%), confirming that the described strains are different and, furthermore, confirming that one of them, *Lelliottia nimipressuralis* SGAir0187, is misclassified and is not a strain of this species.

Representing genome sequences from false type strains of given species has the potential to lead to erroneous conclusions in future studies that may rely on the published misinformation, as they may be used as the wrong reference points. We encourage the authors of the five GA publications and the journal to publish *corrigenda* and remove the words "type strain" from the titles of the publications. We also warn users of publicly available genome sequence data to be cautious in accepting the metadata associated with genome sequences; recent studies have clearly demonstrated the presence of high numbers of misclassified genome se-

Citation Salvà-Serra F, Jaén-Luchoro D, Karlsson R, Bennasar-Figueras A, Jakobsson HE, Moore ERB. 2019. Beware of false "type strain" genome sequences. Microbiol Resour Announc 8:e00369-19. https://doi.org/10.1128/MRA.00369-19.

Editor Irene L. G. Newton, Indiana University, Bloomington

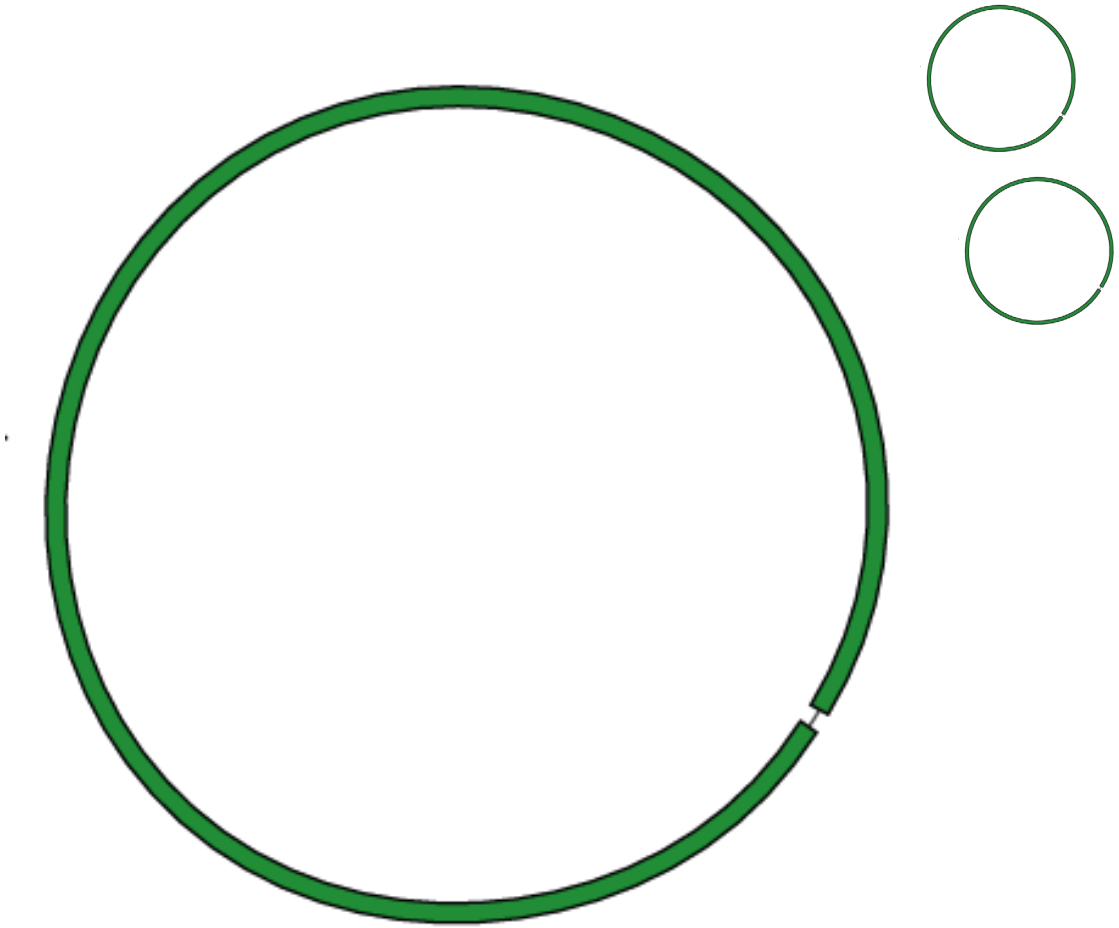
Address correspondence to Francisco Salvà-Serra, francisco.salva.serra@gu.se, or Edward R. B. Moore, erbmoore@ccug.se.

Ed. Note: The authors of the published articles did not respond.

Published 30 May 2019

"We also warn users of publicly available genome sequence data to be cautious in accepting the metadata associated with genome sequences…"

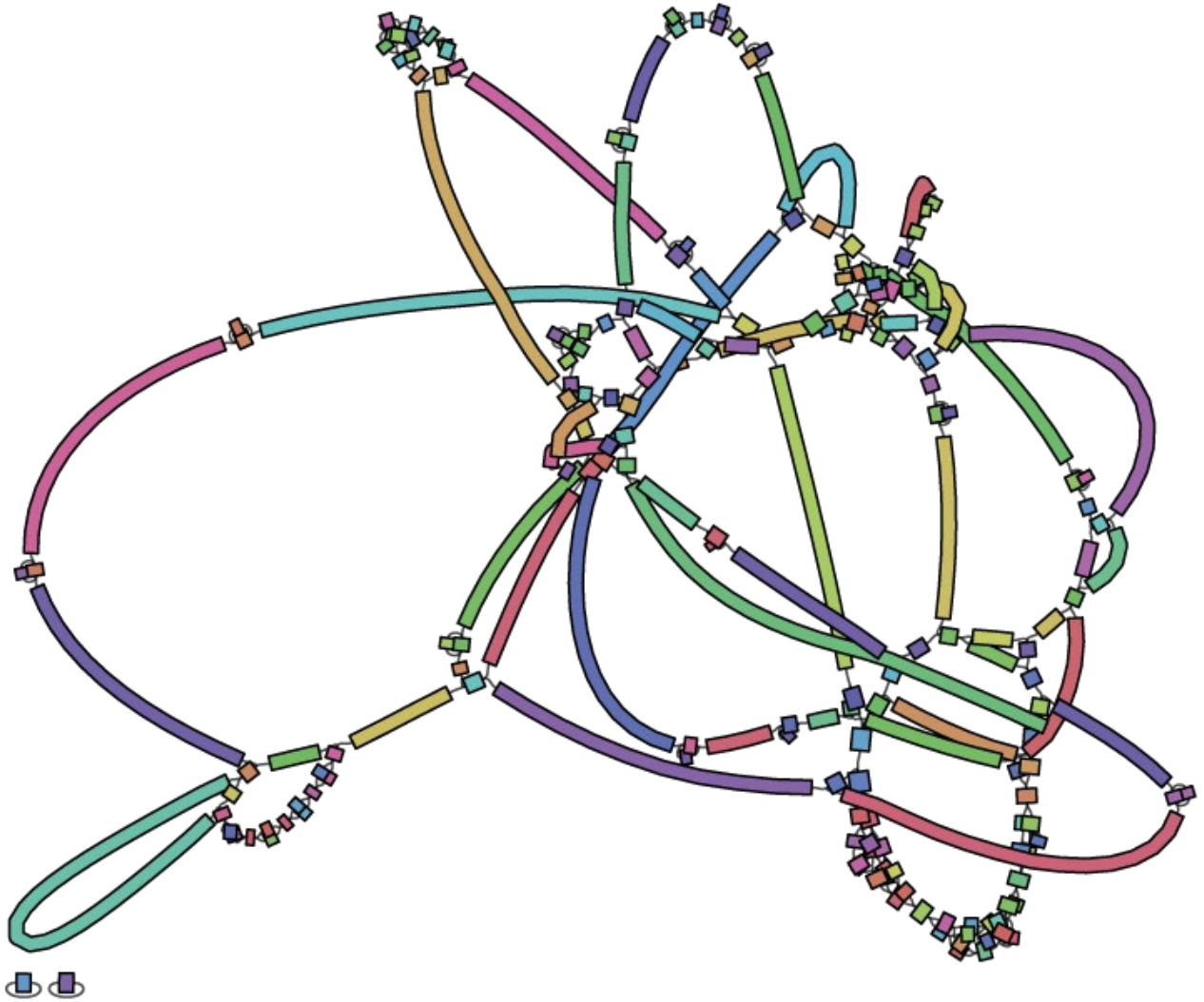
What does a good genome assembly look like?



Images from Ryan Wick https://github.com/rrwick/Unicycler

ONE CODEX ATCC

What does a bad (less good) assembly look like?

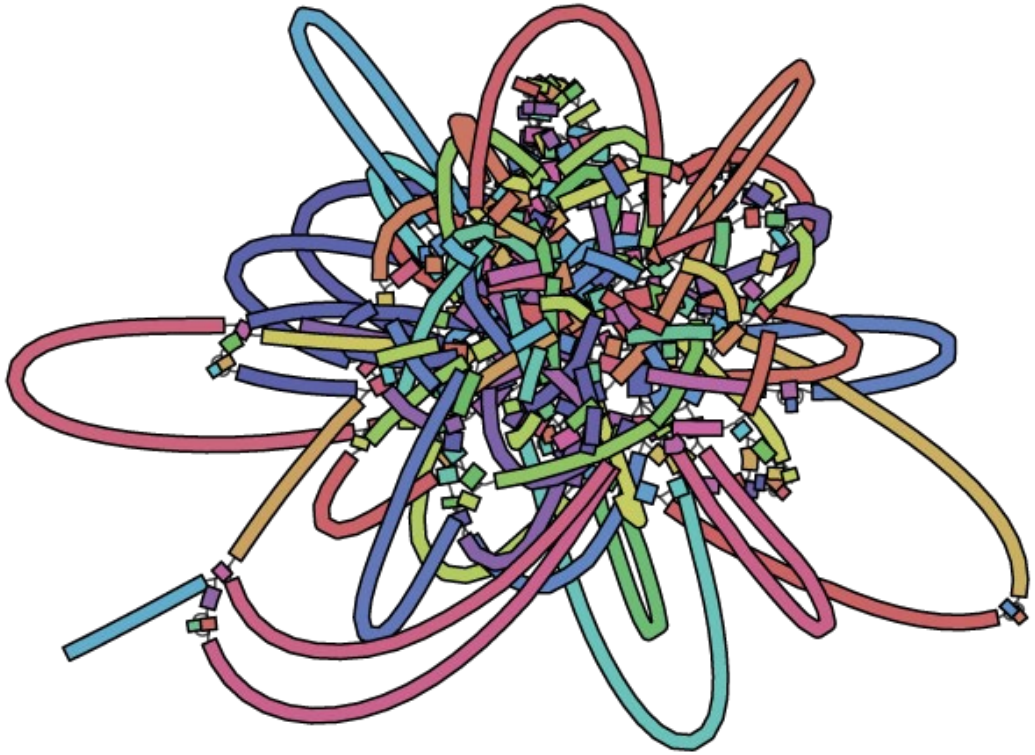


Good Illumina-only assembly

Images from Ryan Wick https://github.com/rrwick/Unicycler

ONE CODEX ATCC

What does a bad (less good) assembly look like?

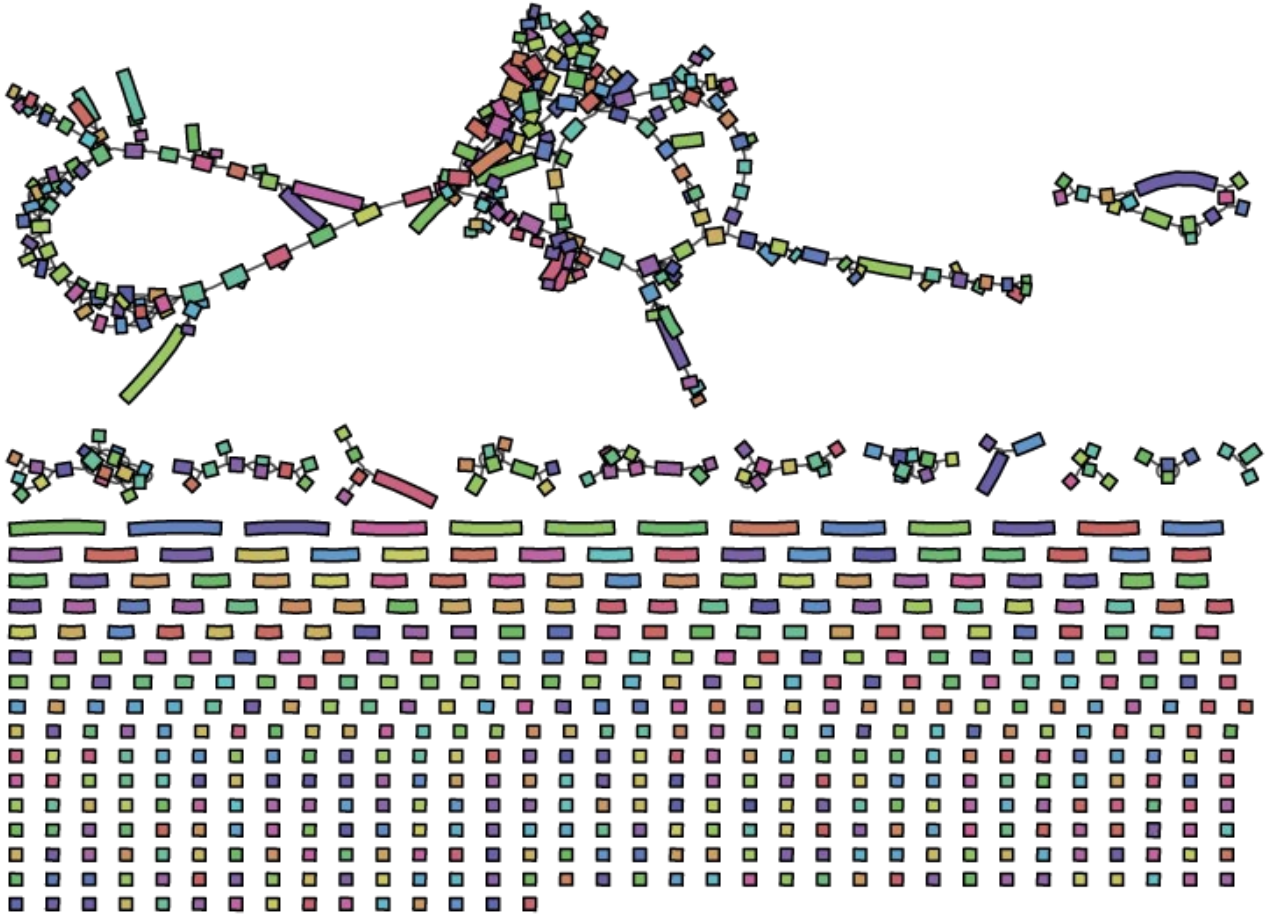


OK Illumina-only assembly

Images from Ryan Wick https://github.com/rrwick/Unicycler

ONE CODEX ATCC

What does a bad (less good) assembly look like?

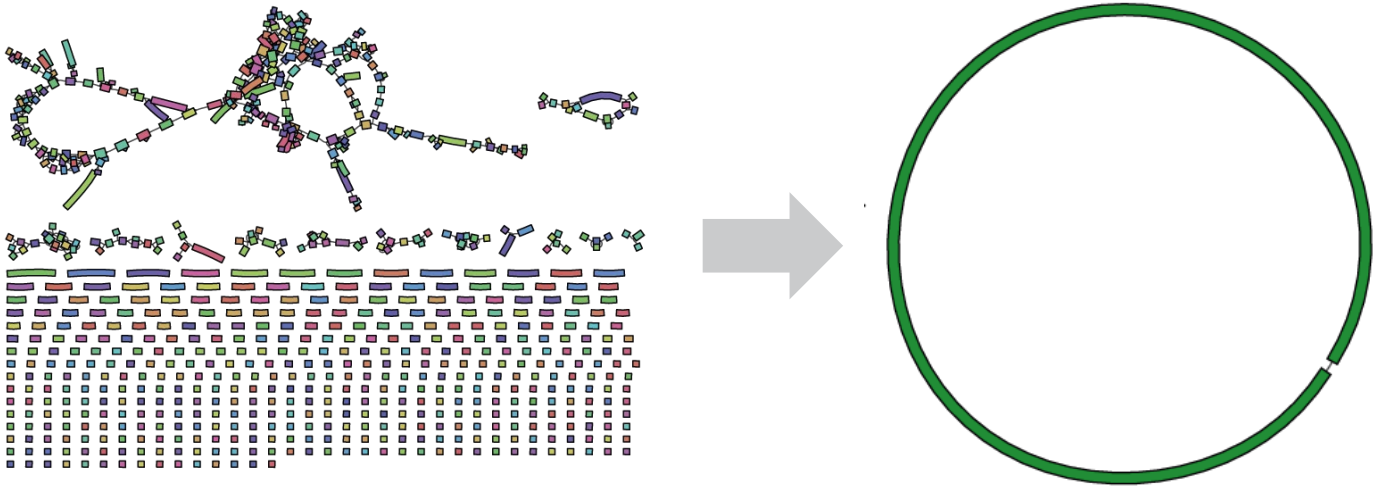


Bad Illumina-only assembly

Images from Ryan Wick https://github.com/rrwick/Unicycler

ONE CODEX ATCC

Enhanced Authentication Initiative – Goals

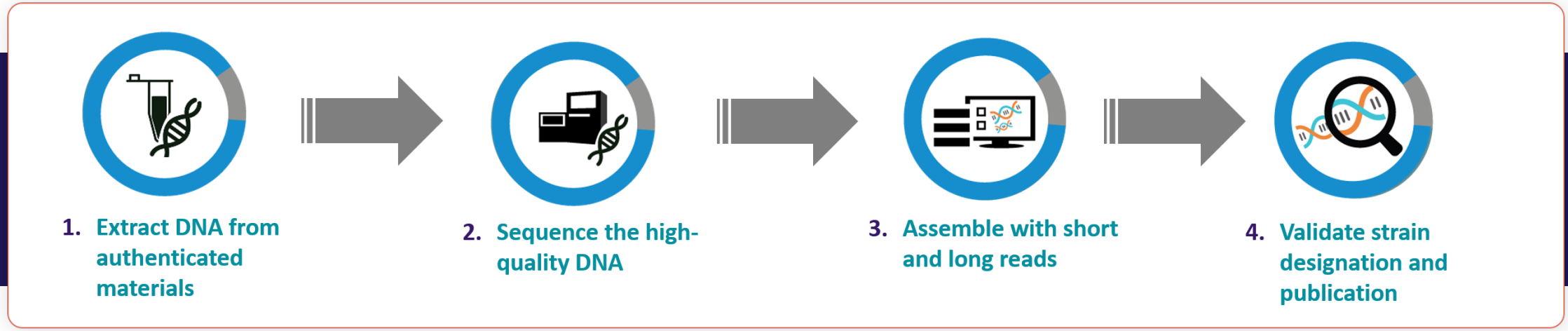


Improve

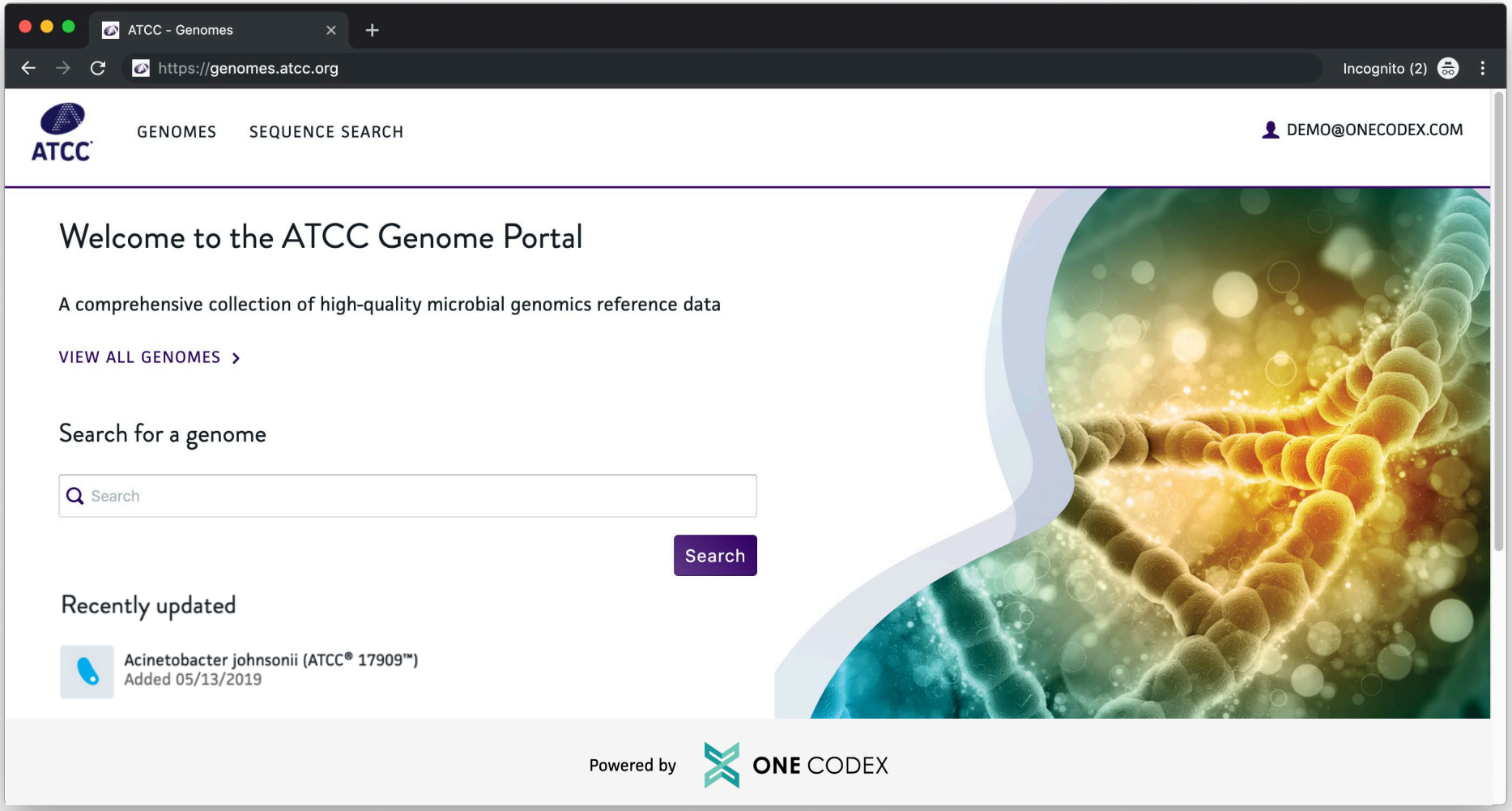
- **Quality**
- **Completeness**
- **Authenticity**
- **Traceability**

ONE CODEX ATCC

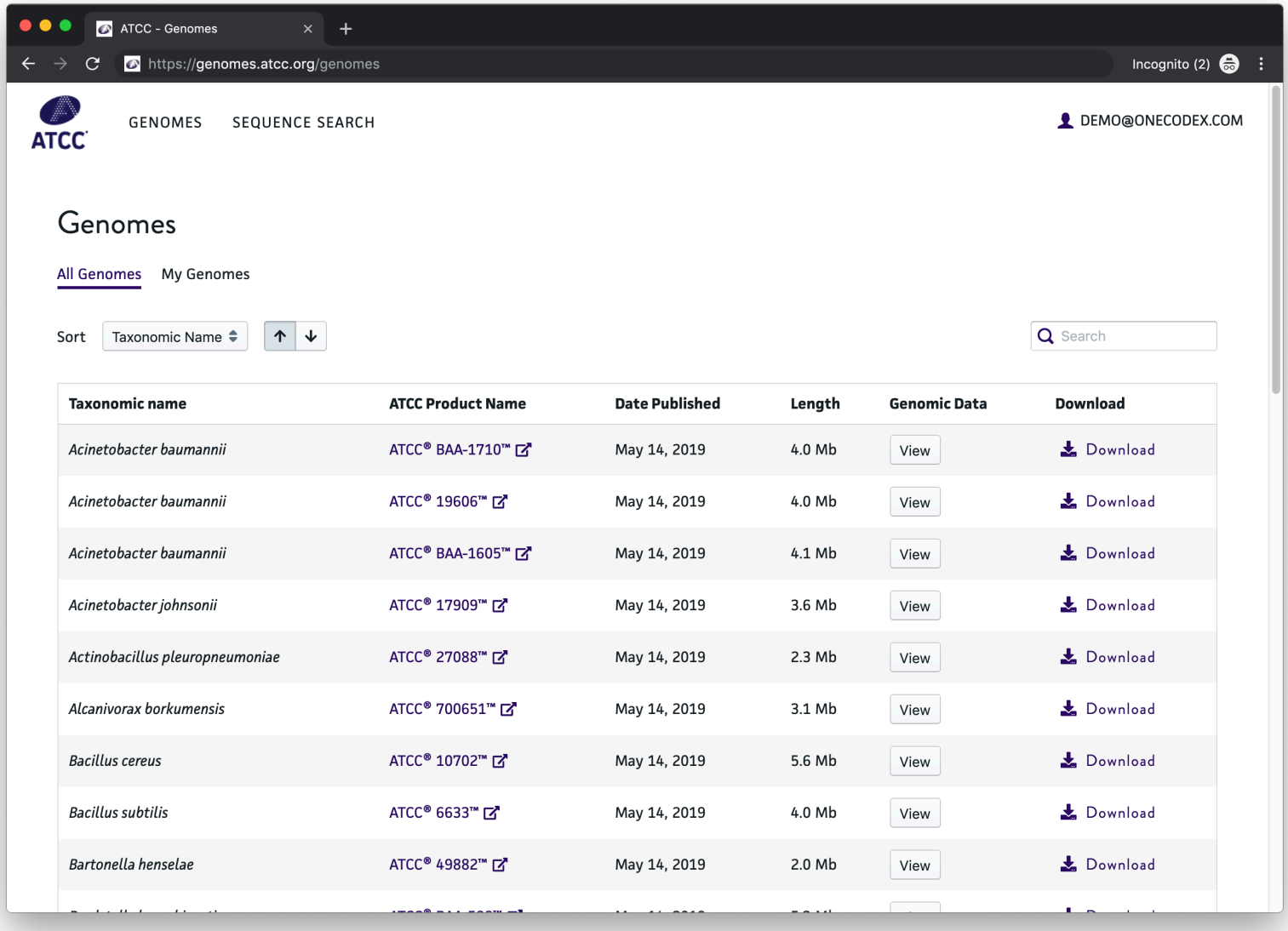
Enhanced Authentication Initiative – Overview



Enhanced Authentication Initiative – Overview

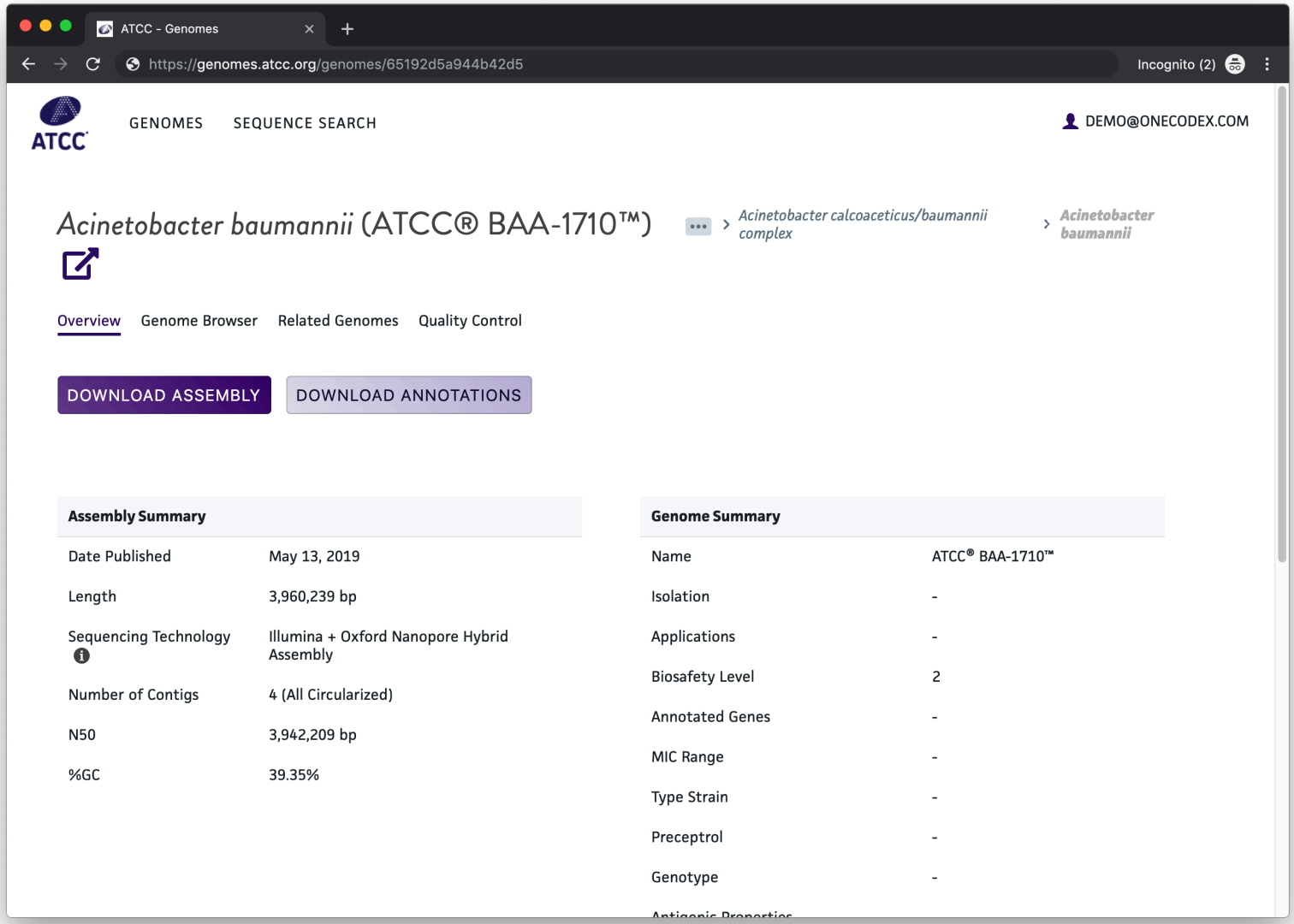


Enhanced Authentication Initiative – Quick Tour

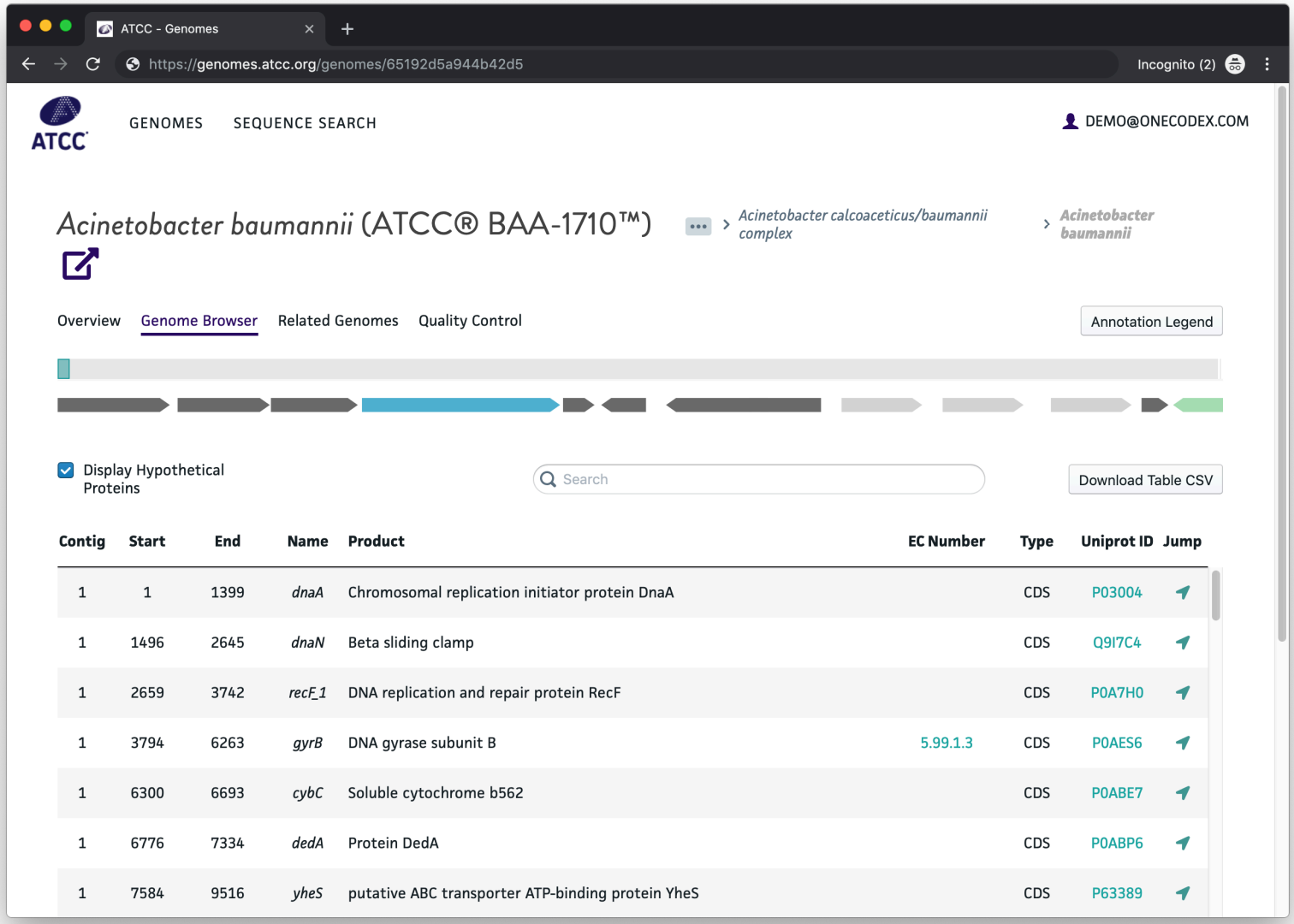


ONE CODEX ATCC

Enhanced Authentication Initiative – Quick Tour

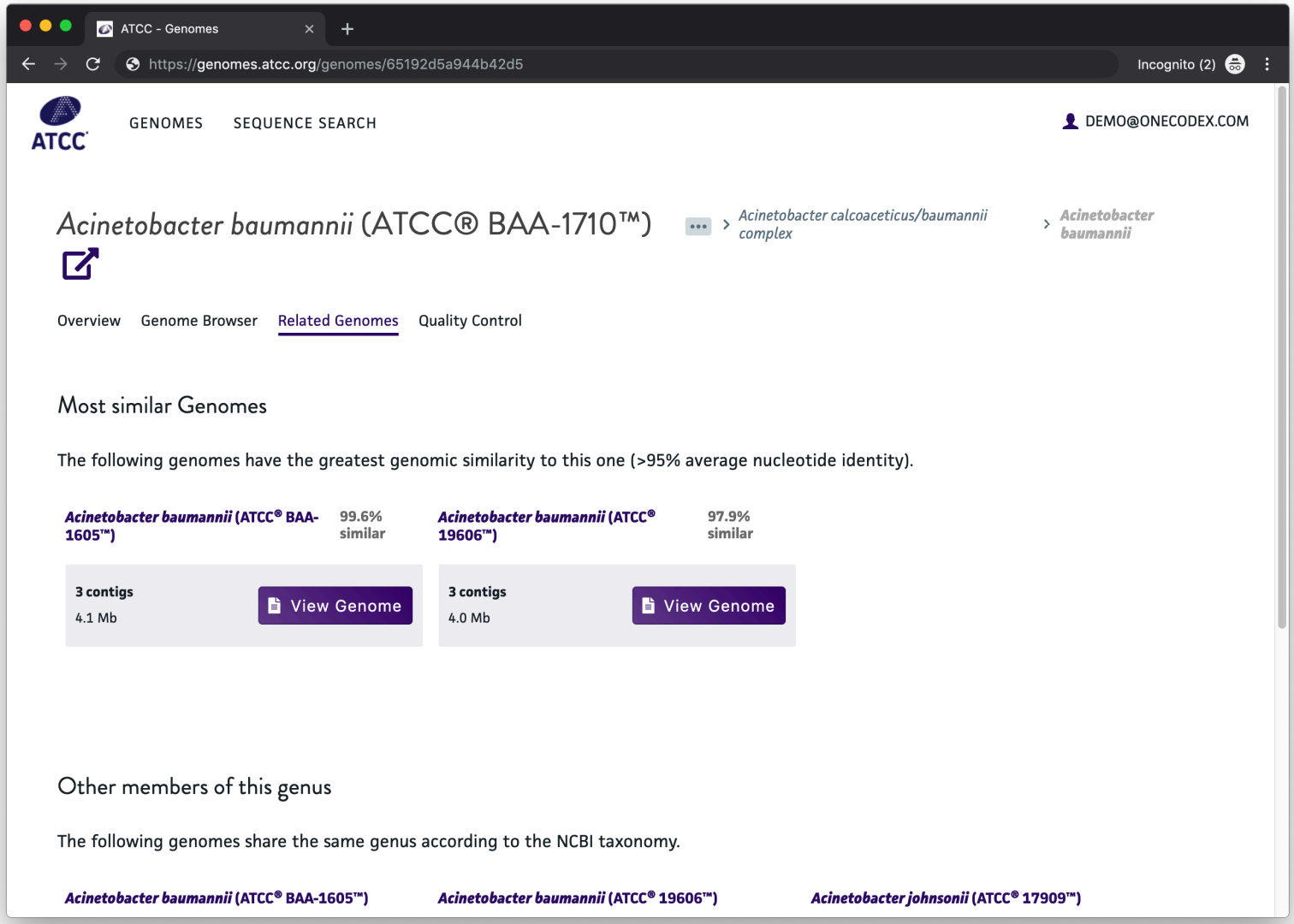


Enhanced Authentication Initiative – Quick Tour

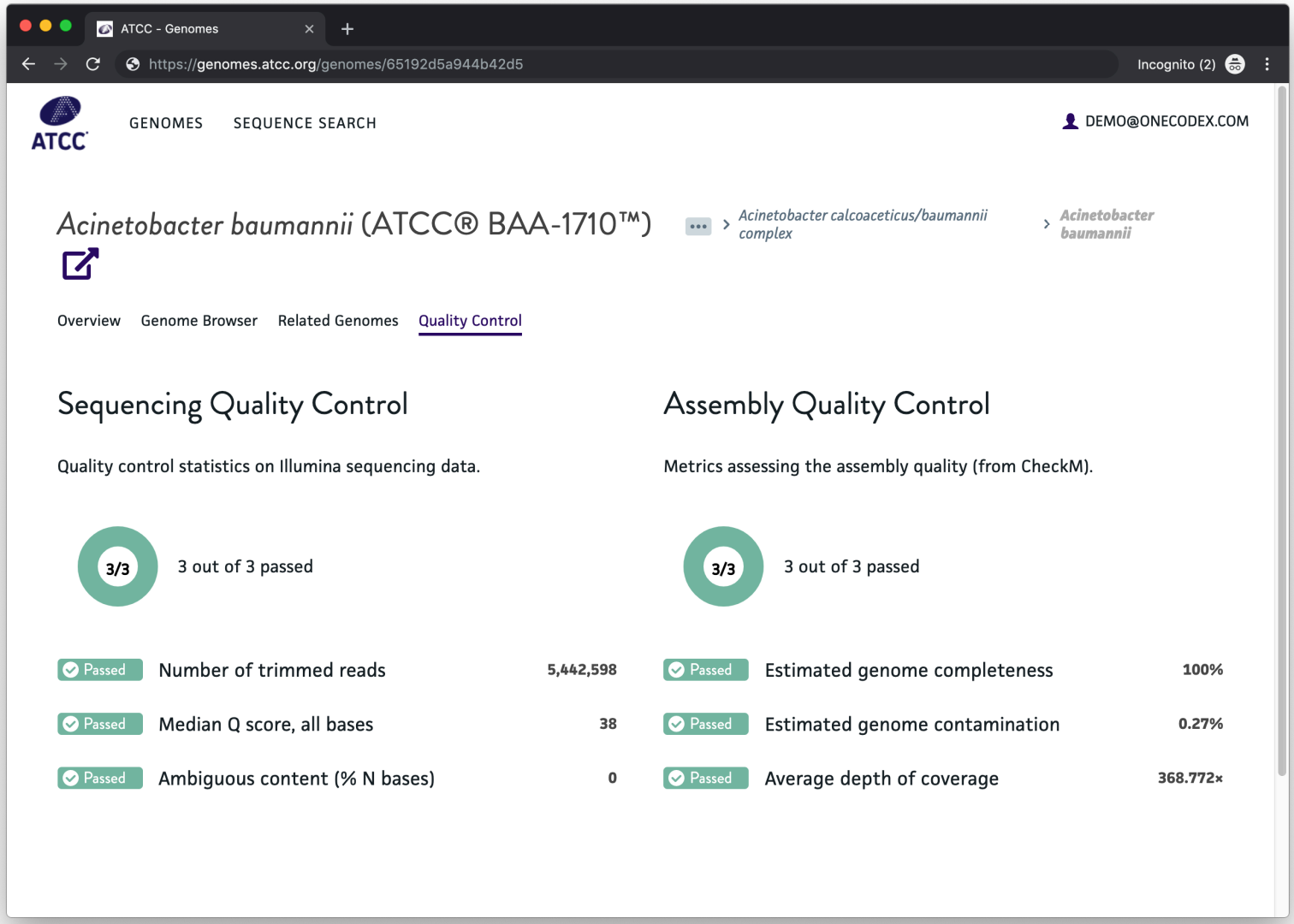


ONE CODEX ATCC

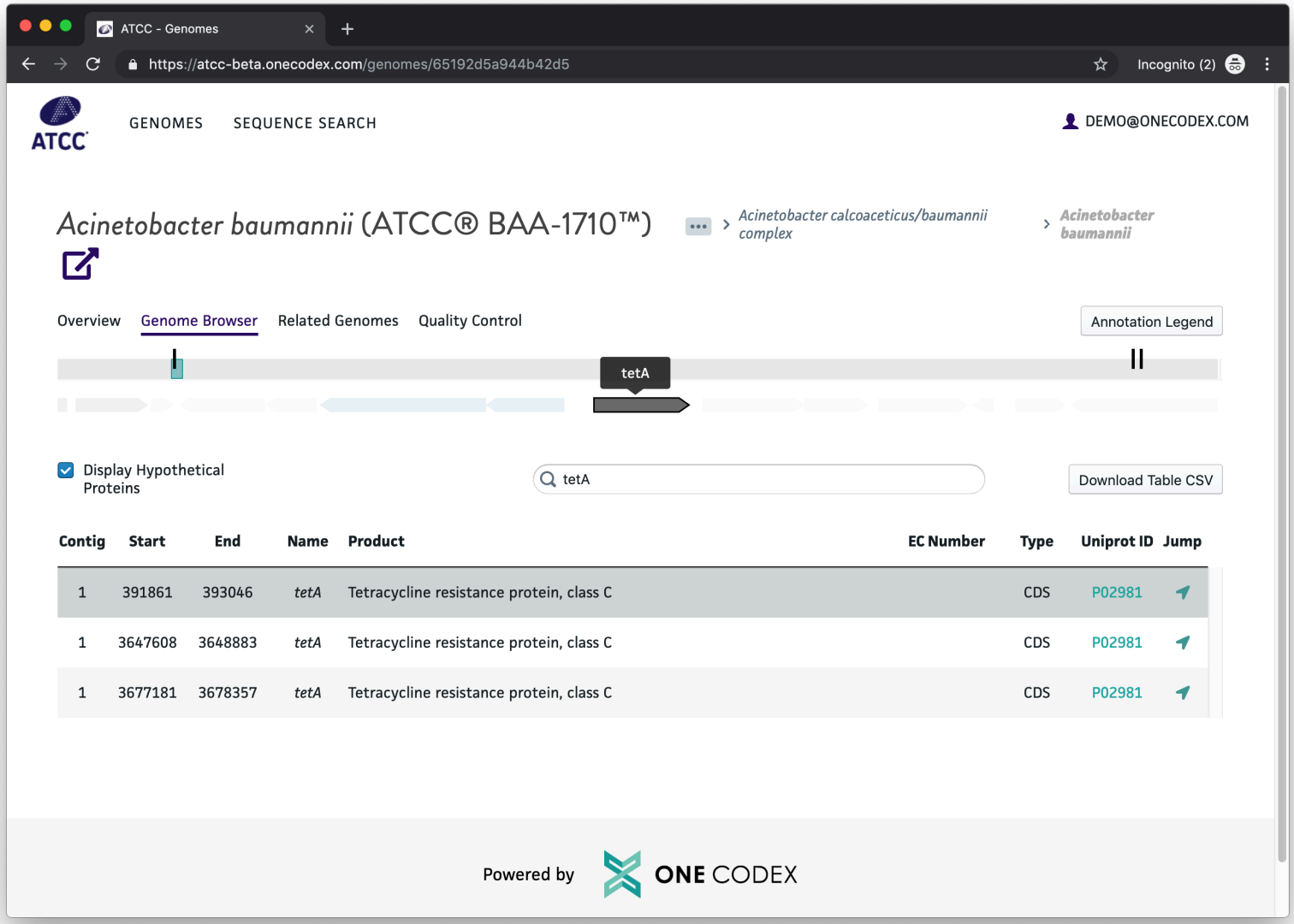
Enhanced Authentication Initiative – Quick Tour



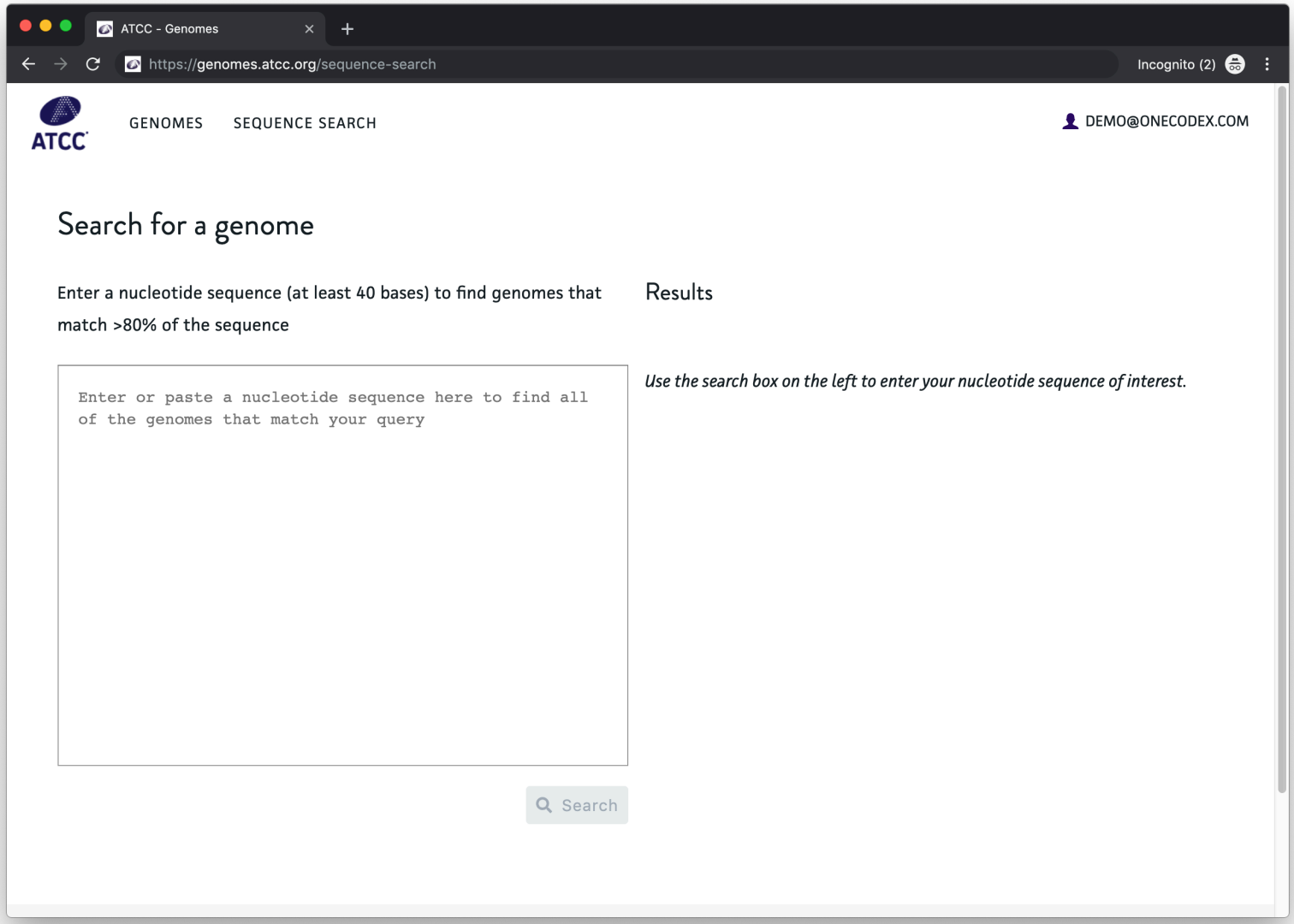
Enhanced Authentication Initiative – Quick Tour



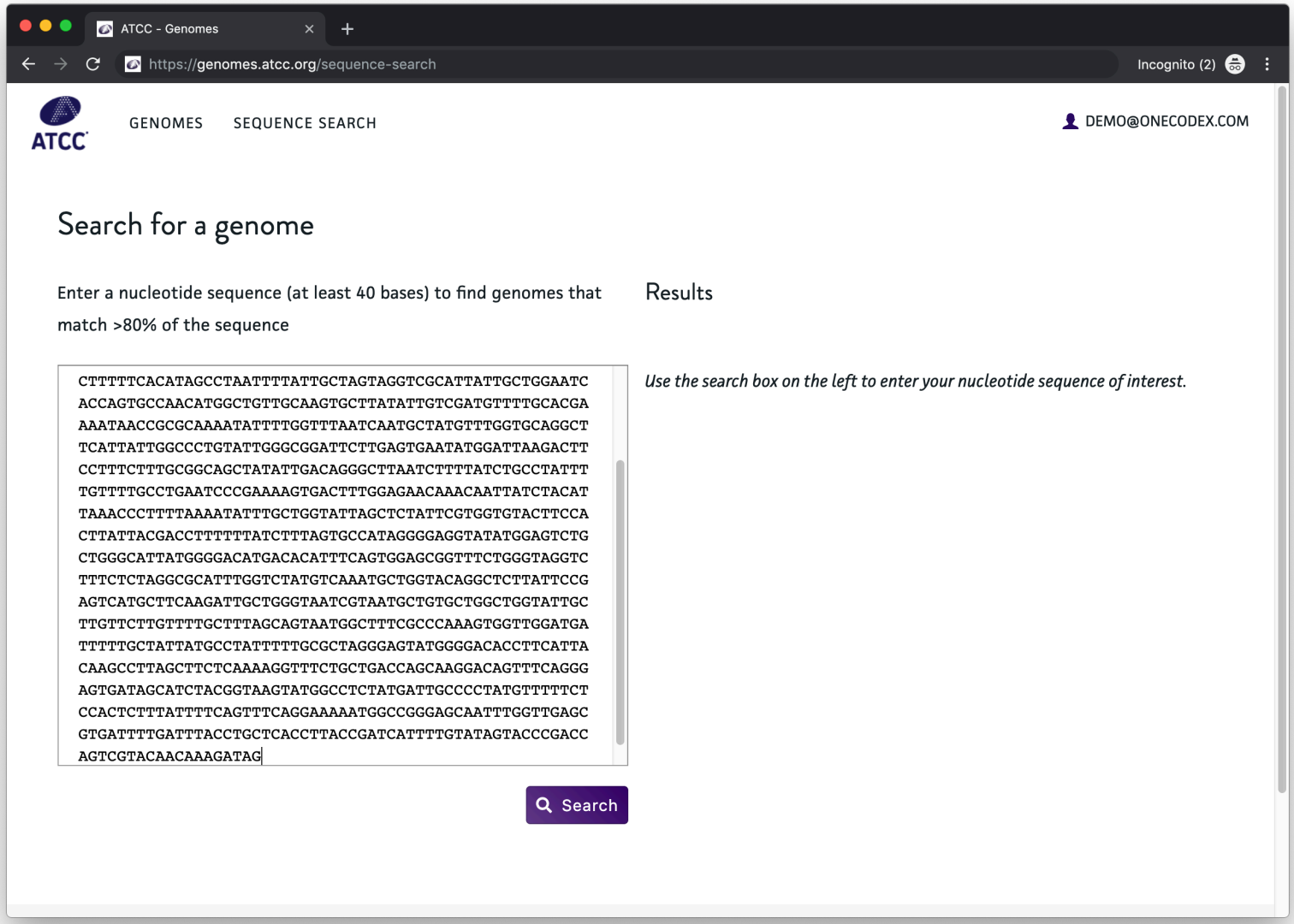
Enhanced Authentication Initiative – Quick Tour



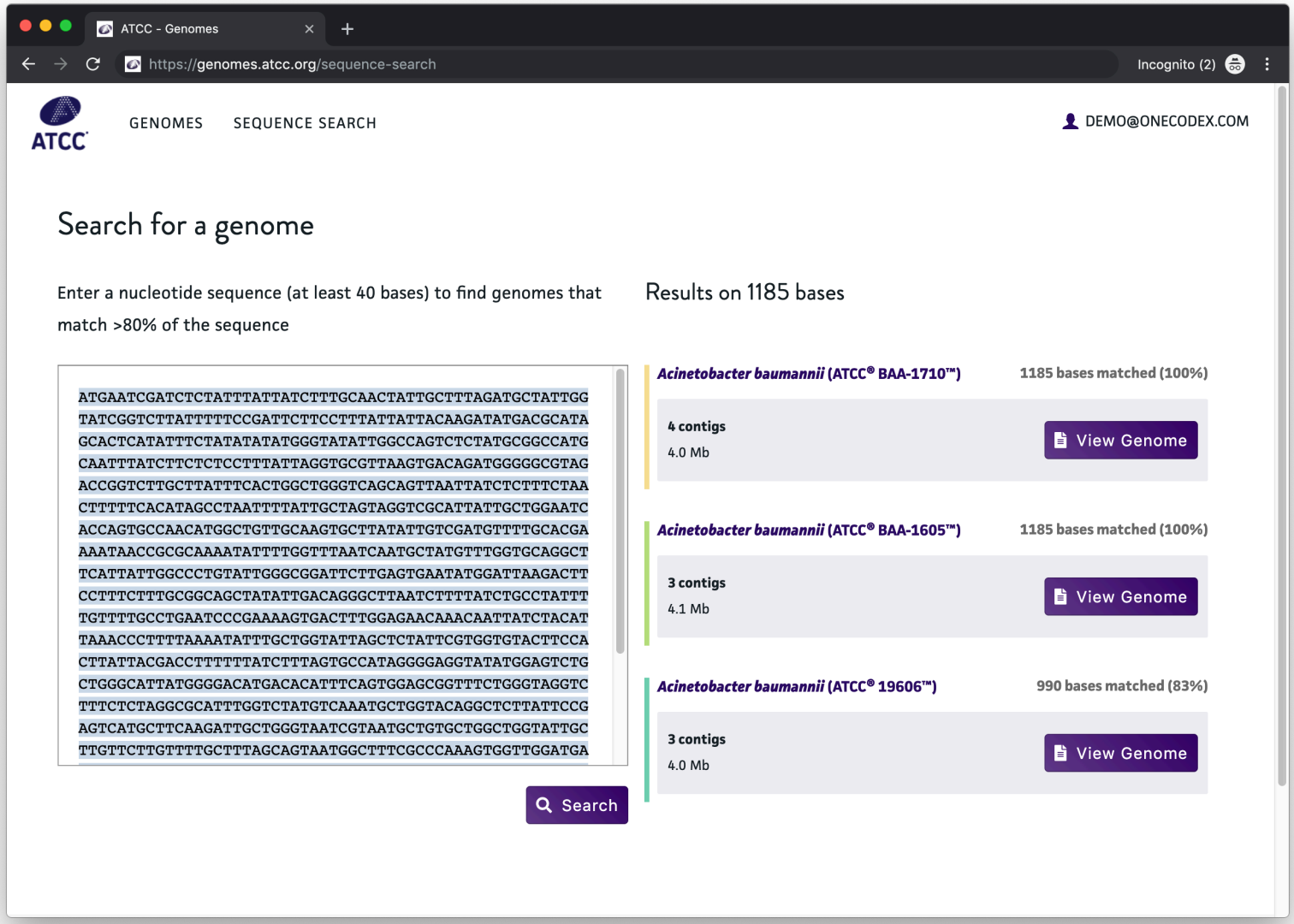
Enhanced Authentication Initiative – Quick Tour



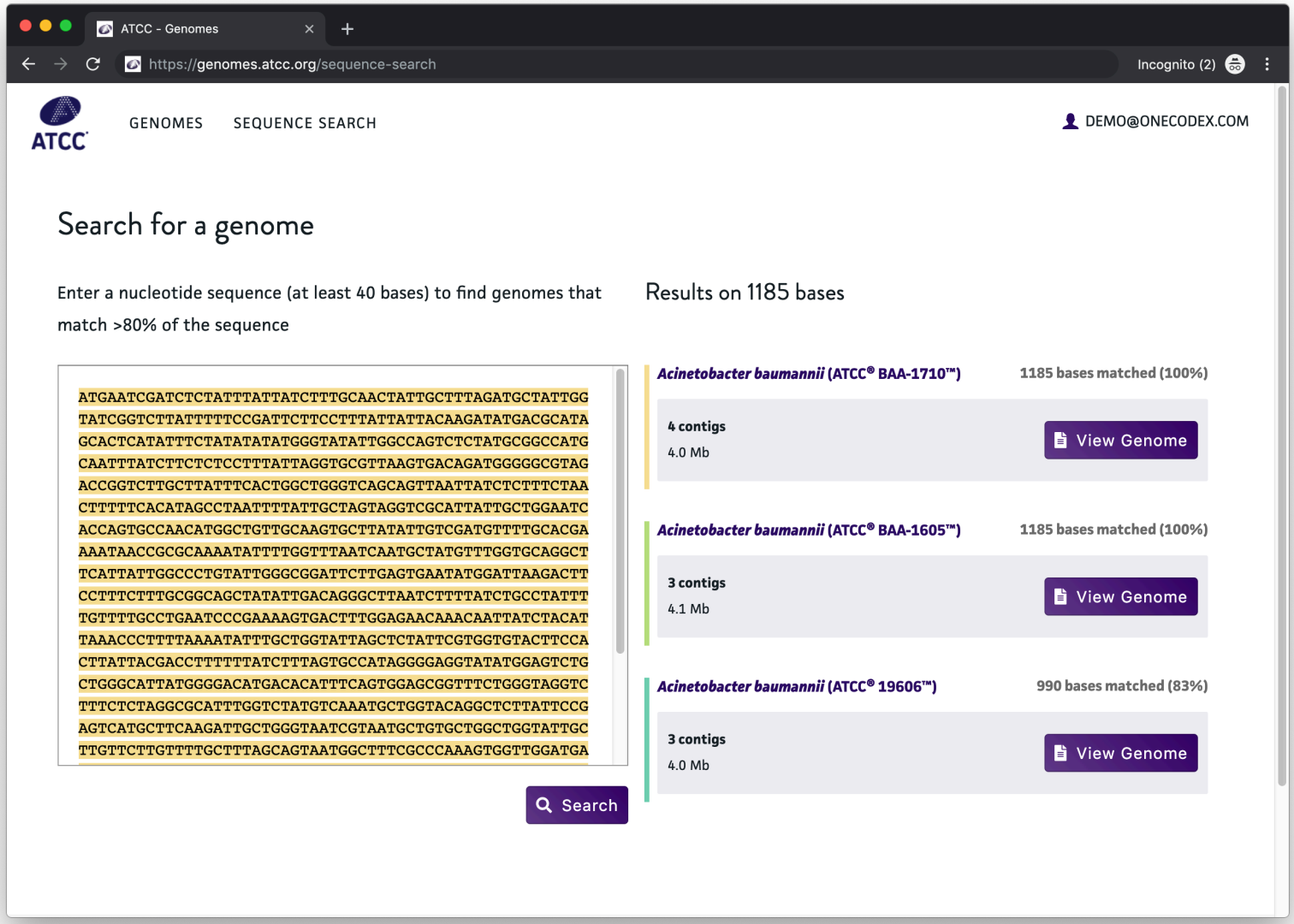
Enhanced Authentication Initiative – Quick Tour



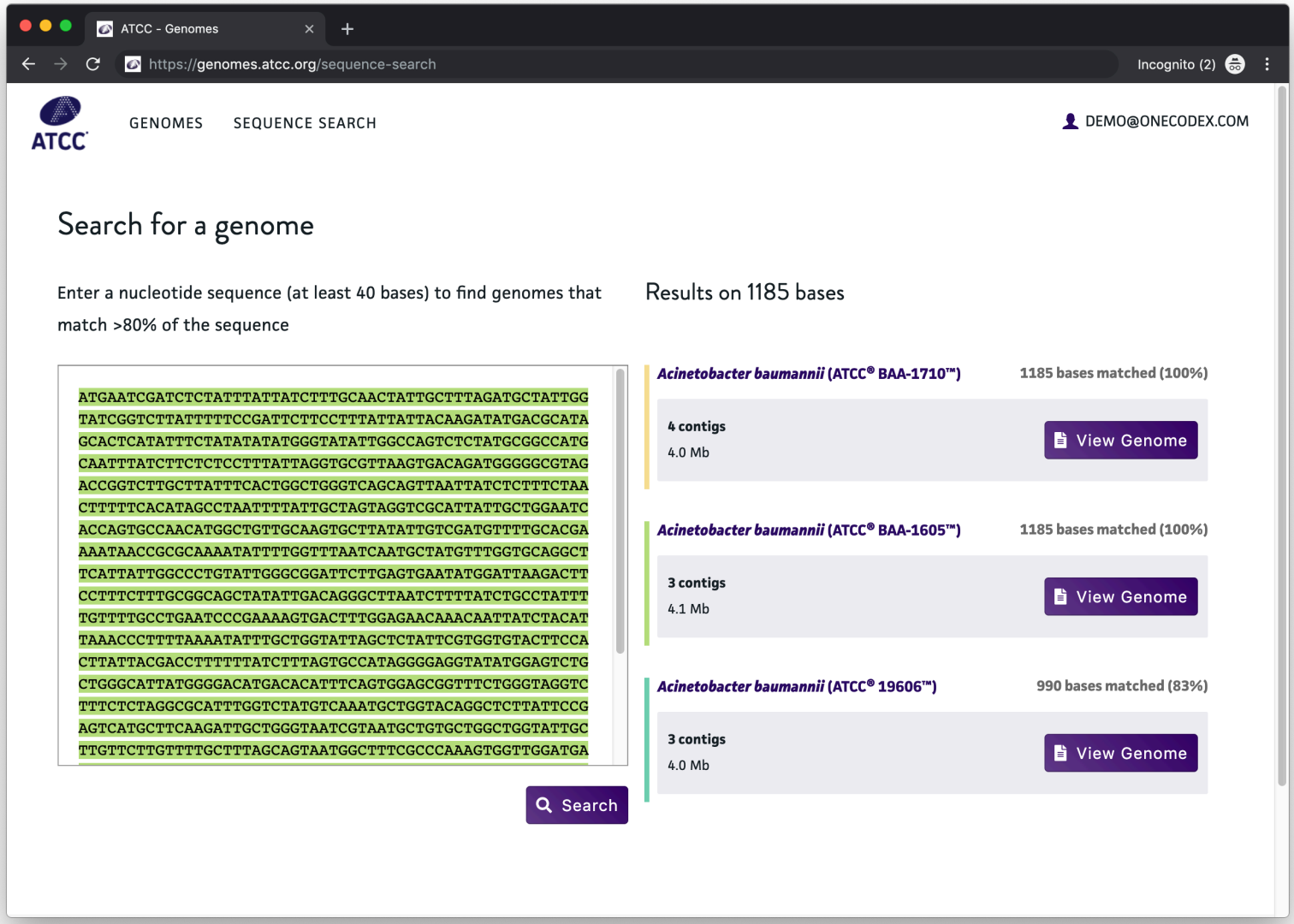
Enhanced Authentication Initiative – Quick Tour



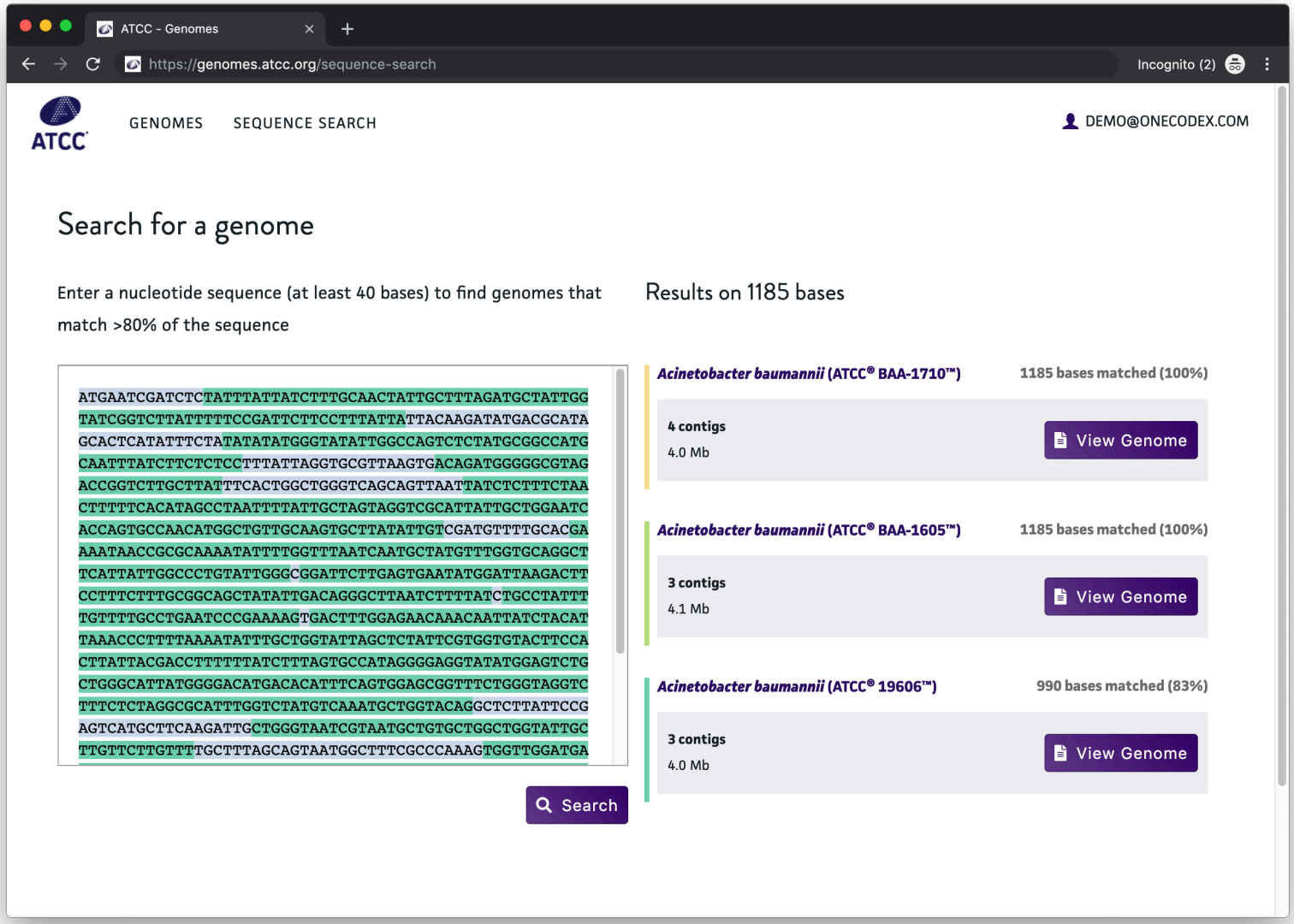
Enhanced Authentication Initiative – Quick Tour



Enhanced Authentication Initiative – Quick Tour



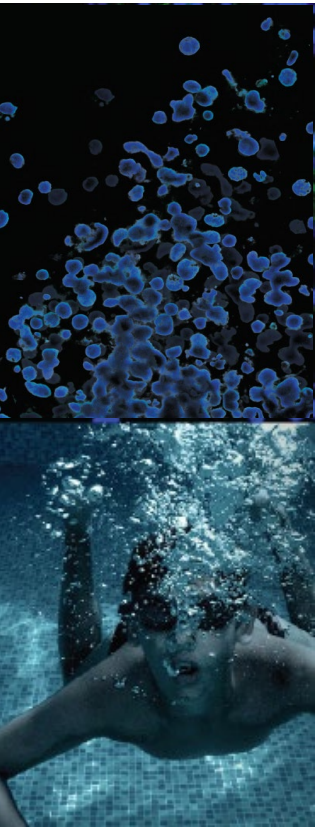
Enhanced Authentication Initiative – Quick Tour



Enhanced Authentication Initiative – Outcomes

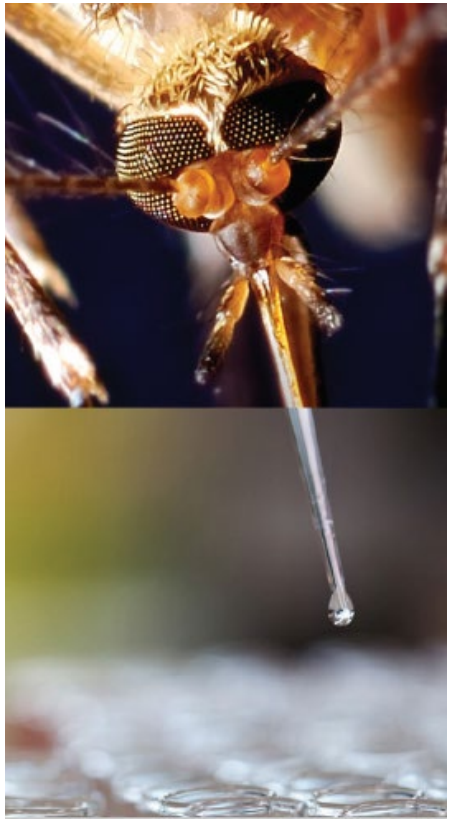
- Modern, easy-to-use genome browser and data platform
- Improved quality, completeness, authenticity, traceability
- Publicly available September 2019

ATCC®



Questions?

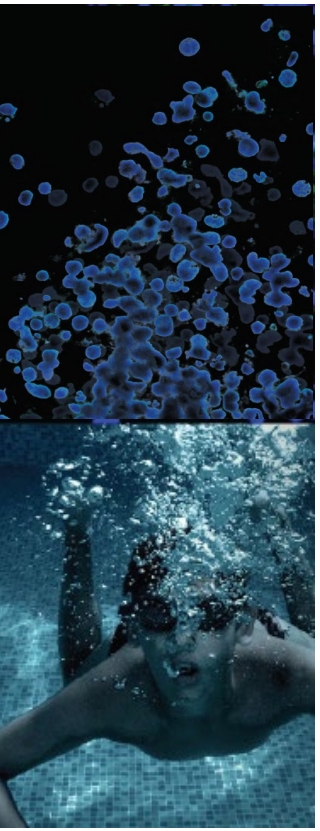
Credible Leads to Incredible™

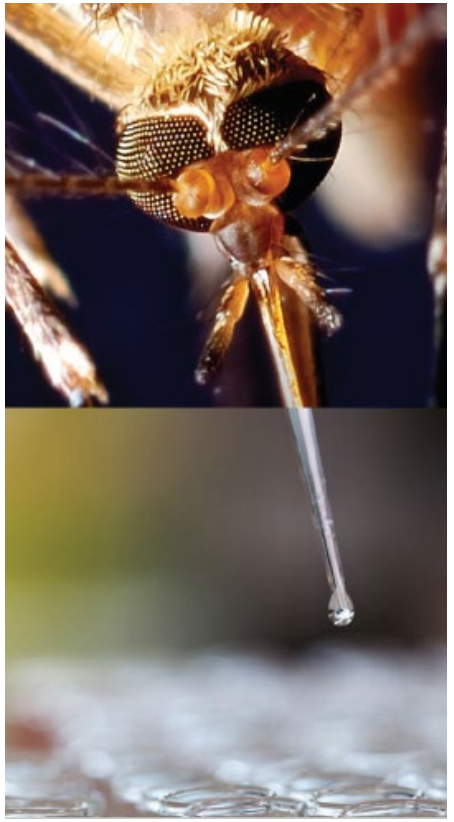


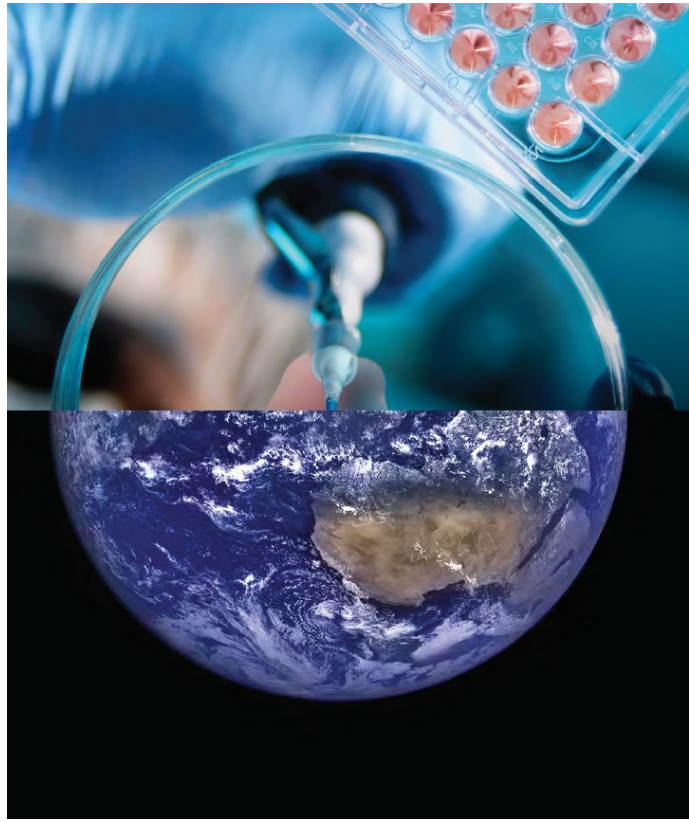
Using Whole-Genome Sequencing for the Enhanced Authentication of ATCC's *Bacillus cereus* Group Strains

Marco A. Riojas, PhD
Scientist, ATCC

Credible Leads to Incredible™

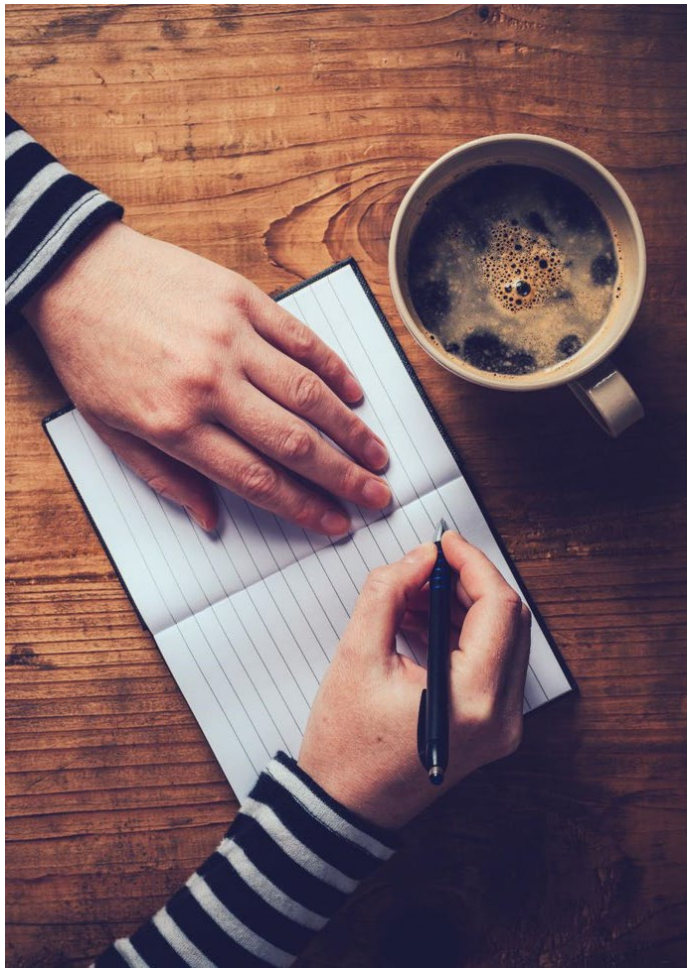






Agenda

- *Bacillus* and *Bacillus cereus* Group (BcG)
- Type strains and species definitions
- BcG strain analysis



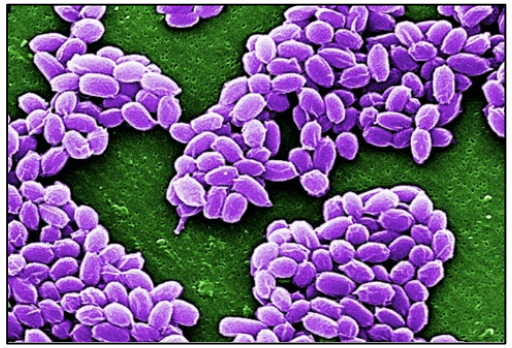
Bacillus and *Bacillus cereus* Group (BcG)

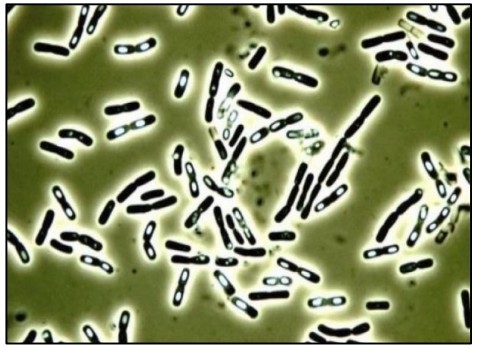
The Genus *Bacillus*

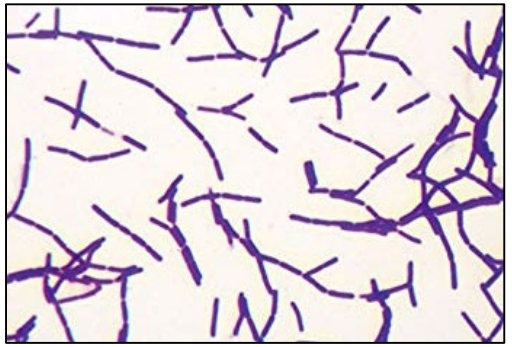
- Composed of Gram-positive, aerobic, endospore-forming bacteria
- Formed in 1872 with the description of *Bacillus subtilis* (the type species) and *B. anthracis*
- Currently a total of 281 species and subspecies

The *Bacillus cereus* Group (BcG)

Currently 18 species, some of which are pathogenic to various plant and animal species







B. anthracis

B. cereus

B. thuringiensis

The *Bacillus cereus* Group (BcG)

A group of closely related species, including some important to health and biotechnology

Species	Year Identified
B. anthracis	1872
B. mycoides	1886
B. cereus	1887
B. thuringiensis	1915
B. pseudomycoides	1998
*B. weihenstephanensis**	1998
B. cytotoxicus	2013
B. toyonensis	2014
B. weidmannii	2016

Species	Year Identified
B. paranthracis	2017
B. pacificus	2017
B. tropicus	2017
B. albus	2017
B. mobilis	2017
B. luti	2017
B. proteolyticus	2017
B. nitratireductans	2017
B. paramycoides	2017

ATCC

Type Strains and Species Definitions

How Are Species Defined?

Background

- Each species is represented by a type strain and a description of that strain
  - Usually the first strain identified
  - Not necessarily the most typical or representative of the species
  - **The type strain is essentially the "definition" of a species**

- If the strain upon which the description was based upon cannot be found or is no longer appropriate, a neotype strain may be proposed

- *Mycobacterium tuberculosis*
  - Identified in 1882 by Robert Koch
  - Strain H37Rv was proposed as the neotype strain in 1972 (Kubica, *et al*.)
  - Now *Mycobacterium tuberculosis* H37Rv^{T} is the type strain

Kubica, *et al*. *Int J Syst Bacteriol*. 1972, 22(2):99-106.

ATCC

How is a New Strain Assigned to a Species?

Background

- The characteristics of the new strain are compared to the characteristics of species type strains
  - Historically: phenotypic characteristics
- A strain that shares enough of the essential characteristics of a type strain is said to be within the circumscription of that species/type strain
  - Therefore, it belongs to that species
- Recognizing that phenotypes can be quite unreliable, today we rely more heavily on genotypic comparisons
  - 16S rRNA genes, *hsp65*, *rpoB*
- Small numbers of genes can still provide misleading results
  - Most accurate comparison would be between whole genomes

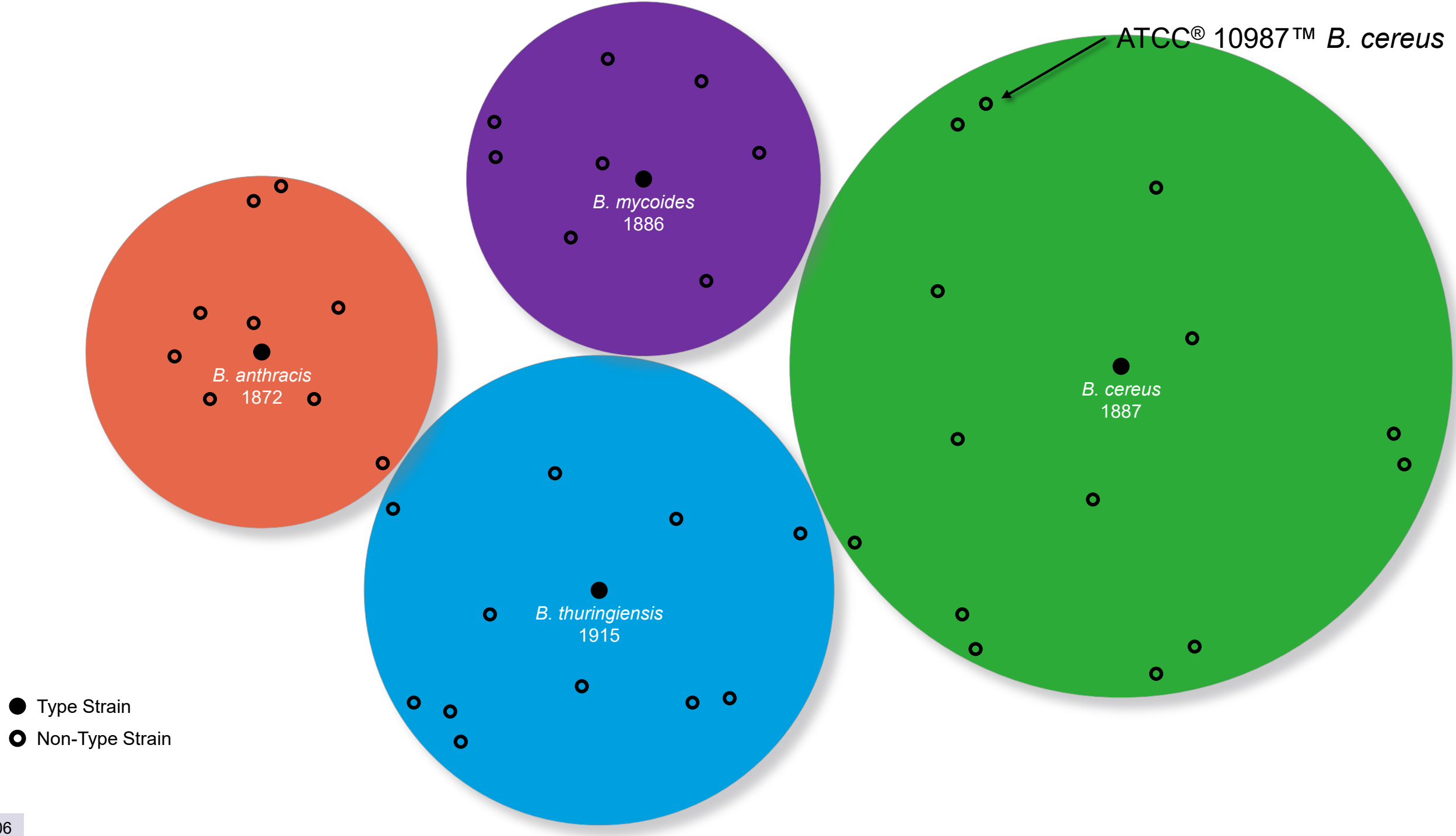
ATCC

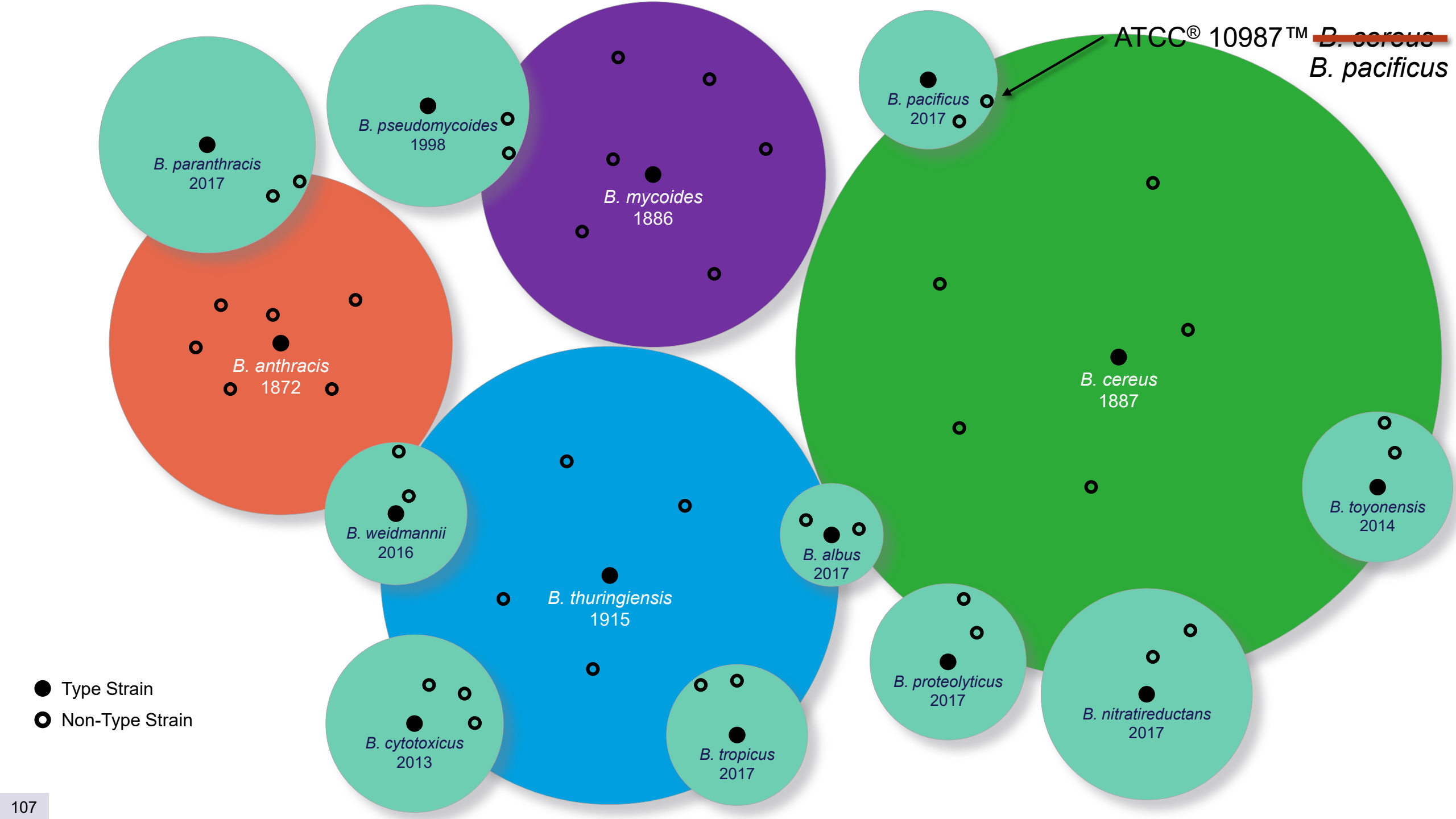
Bacillus cereus Group (BcG) Strains at ATCC

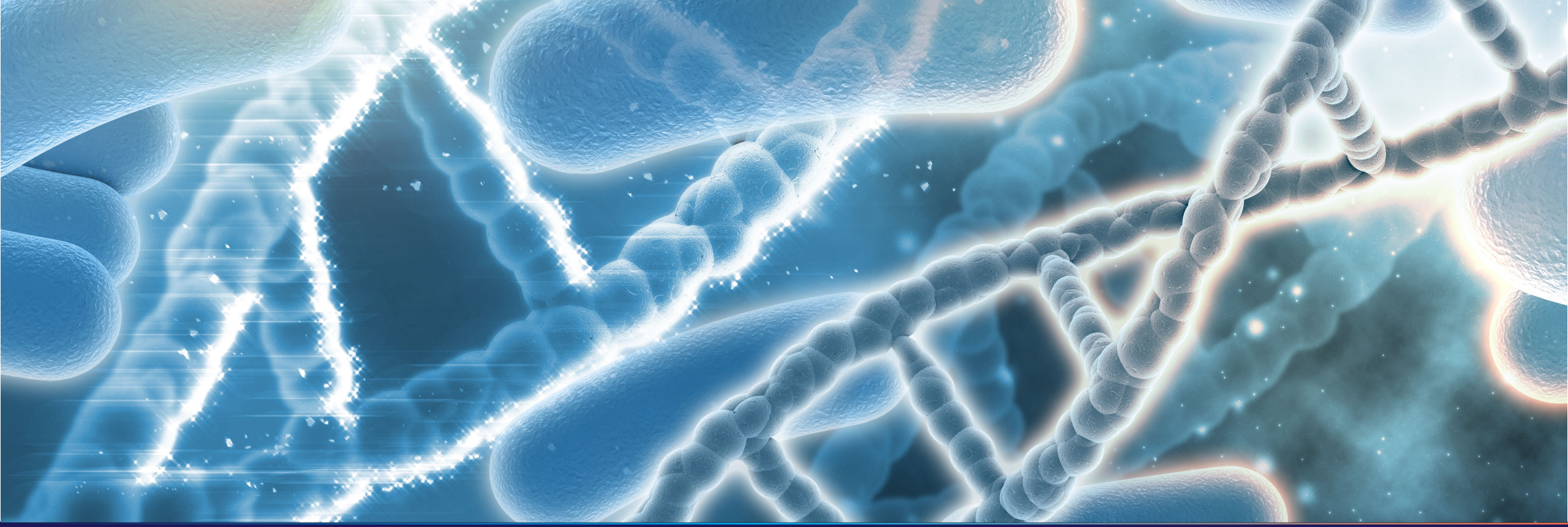
Background

- Many of our strains were deposited many decades ago
  - *E.g.*, ATCC® 246™ *Bacillus cereus* was deposited in 1925
  - New species have been discovered
  - Previous classifications of our strains were based on less accurate/comprehensive methods, such as phenotypic observations or biochemical testing
- Do our *Bacillus* strains match up with current taxonomy?
  - GOAL: Identify our strains using the most current techniques and definitions
- Whole-genome sequencing (WGS) of expertly authenticated material
  - Primary focus on BcG strains from ATCC and BEI Resources*
  - Secondary focus on the BcG species at large

* BEI Resources material is the property of the National Institute of Allergy and Infectious Diseases (NIAID), NIH


ATCC® 10987™ B. cereus
B. mycoides
1886
B. anthracis
1872
B. cereus
1887
B. thuringiensis
1915
Type Strain
Non-Type Strain


ATCC® 10987™ B. cereus
B. pacificus
B. paranthracis
2017
B. pseudomycoides
1998
B. mycoides
1886
B. pacificus
2017
B. anthracis
1872
B. cereus
1887
B. weidmannii
2016
B. thuringiensis
1915
B. albus
2017
B. toyonensis
2014
B. proteolyticus
2017
B. nitratireductans
2017
B. cytotoxicus
2013
B. tropicus
2017
Type Strain
Non-Type Strain



BcG Strain Analysis

Bacillus cereus Group (BcG) Strains

Methodology

- Selected a subset of BcG and *Bacillus* strains
  - Priority given to lower accession numbers (*e.g.*, ATCC® 246™, 4342™, 6463™, etc.)
    - Older deposits more likely to be misclassified
  - Also obtained strains from BEI Resources, a NIAID funded repository
- Whole-Genome Sequencing (WGS)
  - Illumina® MiSeq® v3 (2×300)
- Genome Assembly
  - SPAdes
- Genomic Comparison of Taxonomy
  - Digital DNA-DNA Hybridization (dDDH)
    - Genome-to-Genome Distance Calculator (GGDC)
  - Average Nucleotide Identity (ANI)
    - OrthoANIb

dDDH Range	Interpretation	
≥ 80%	Same species	Same subspecies
70 – 80%	Same species	Different subspecies
< 70%	Different species	

Meier-Kolthoff JP, *et al*. (2014); Meier-Kolthoff JP, *et al*. (2013); Auch, *et al*. (2013a); Auch, *et al*. (2013b).

B. anthracis and *B. mycoides*

Species/Subspecies	Strain	Source	A0488^T	NR-1041	NR-1202	NR-1389	NR-1408	NR-36073	NR-36091	NR-3838	NR-411	NR-41	NR-46	NR-51483	NR-51484	NR-51485	NR-51486	ATCC 6462^T	ATCC 6462^T	NR-613	NR-22171	NBRC 101238^T	KBAB4	DSM 20231^T
B. anthracis	A0488^T	ABJC01	100	99.8	99.7	99.7	99.6	99.7	99.6	99.7	99.8	99.5	99.6	99.6	99.8	99.8	99.8	38.2	38	39.7	37.8	37.9	38.2	28
B. anthracis	NR-1041	This Work	99.92	100	99.8	99.8	99.7	99.7	99.7	99.8	99.8	99.5	99.7	99.7	99.8	99.9	99.8	38.3	38.1	39.8	38	38	38.2	23
B. anthracis	NR-1202	This Work	99.93	99.92	100	99.8	99.8	99.8	99.7	99.9	99.9	99.6	99.8	99.9	99.8	99.8	99.8	38.2	38.1	39.8	37.9	37.9	38.2	23
B. anthracis	NR-1389	This Work	99.89	99.91	99.92	100	99.6	99.9	99.8	99.8	99.8	99.7	99.7	99.7	99.8	99.8	99.8	38.3	38.1	39.8	37.9	38	38.2	23
B. anthracis	NR-1408	This Work	99.87	99.89	99.92	99.92	100	99.7	99.6	99.8	99.8	99.5	99.7	99.7	99.6	99.7	99.6	38.8	38.6	39.9	38.4	38.5	38.8	23
B. anthracis	NR-36073	This Work	99.90	99.93	99.90	99.93	99.87	100	99.8	99.8	99.7	99.7	99.7	99.7	99.7	99.7	99.7	38.2	38.1	39.8	37.9	38	38.2	22
B. anthracis	NR-36091	This Work	99.89	99.88	99.88	99.90	99.88	99.91	100	99.7	99.7	99.6	99.7	99.6	99.7	99.7	99.6	38.3	38.2	39.8	38	38.1	38.3	23
B. anthracis	NR-3838	This Work	99.93	99.91	99.97	99.91	99.91	99.92	99.89	100	100	99.6	99.8	99.8	99.8	99.8	99.8	38.2	38.1	39.8	37.9	38	38.2	22
B. anthracis	NR-411	This Work	99.91	99.91	99.93	99.93	99.94	99.92	99.90	99.94	100	99.6	99.8	99.8	99.8	99.8	99.8	38.2	38.1	39.8	37.9	38	38.2	23
B. anthracis	NR-41	This Work	99.92	99.91	99.93	99.92	99.91	99.93	99.88	99.93	99.91	100	99.5	99.6	99.5	99.5	99.5	38.3	38.1	39.8	38	38	38.2	23
B. anthracis	NR-46	This Work	99.89	99.85	99.89	99.90	99.90	99.89	99.87	99.91	99.88	99.87	100	99.7	99.7	99.7	99.7	38.6	38.5	39.9	38.2	38.3	38.6	23
B. anthracis	NR-51483	This Work	99.91	99.91	99.95	99.91	99.93	99.92	99.90	99.93	99.95	99.94	99.88	100	99.7	99.7	99.7	38.2	38.1	39.8	37.9	38	38.2	23
B. anthracis	NR-51484	This Work	99.92	99.95	99.94	99.91	99.89	99.91	99.89	99.92	99.93	99.92	99.89	99.90	100	99.9	100	38.2	38.1	39.8	37.9	38	38.2	23
B. anthracis	NR-51485	This Work	99.91	99.95	99.93	99.91	99.87	99.94	99.91	99.92	99.91	99.91	99.86	99.92	99.96	100	99.9	38.2	38.1	39.8	37.9	38	38.2	23
B. anthracis	NR-51486	This Work	99.92	99.94	99.93	99.92	99.86	99.91	99.89	99.94	99.89	99.91	99.85	99.93	99.98	99.94	100	38.2	38.1	39.8	37.9	38	38.2	23
B. mycoides	ATCC 6462^T	CP009692.1	89.35	89.47	89.37	89.37	89.49	89.34	89.27	89.39	89.40	89.45	89.45	89.38	89.31	89.34	89.41	100	96.8	93.2	89.9	78.7	78.8	27
B. mycoides	ATCC 6462^T	This Work	89.24	89.47	89.38	89.37	89.53	89.40	89.31	89.33	89.35	89.40	89.47	89.38	89.38	89.41	89.36	99.94	100	92.3	88.8	78	78.4	25
B. mycoides	NR-613	This Work	89.73	89.86	89.69	89.88	89.89	89.81	89.88	89.84	89.83	89.88	89.85	89.87	89.86	89.86	89.71	99.08	99.13	100	87.7	79.7	80.7	23
B. cereus	NR-22171	This Work	89.28	89.46	89.45	89.38	89.42	89.33	89.32	89.34	89.38	89.37	89.47	89.29	89.32	89.31	89.32	98.81	98.65	98.33	100	79.8	79	23
B. weihenstephanensis	NBRC 101238^T	BAUY01	89.27	89.38	89.31	89.32	89.53	89.30	89.36	89.38	89.31	89.37	89.45	89.28	89.35	89.31	89.29	97.71	97.62	97.54	97.78	100	84.4	22
B. weihenstephanensis	KBAB4	CP000903.1	89.31	89.37	89.40	89.45	89.51	89.48	89.39	89.41	89.37	89.51	89.53	89.36	89.35	89.34	89.34	97.53	97.52	97.58	97.75	98.33	100	27
Staphylococcus aureus subsp. *aureus*	DSM 20231^T	CP011526.1	67.29	67.62	67.65	67.06	67.27	67.22	67.26	67.61	67.47	67.33	67.62	67.38	67.49	67.58	67.50	68.13	67.27	67.74	67.52	67.80	68.00	100

ANI	Interpretation	dDDH
98.0 - 100	Same species and subspecies	80.0 - 100
96.5 - 97.999	Same species, different subspecies	70.0 - 79.9
92.0 - 96.499	Different species	50.0 - 69.9
85.0 - 91.999		30.0 - 49.9
0.0 - 84.999		0.0 - 29.9

Type Strain

Non-Type Strain

Subspecies Circumscription

Species Circumscription

B. anthracis and *B. mycoides*

Species/Subspecies	Strain	Source
B. anthracis	A0488^{T}	ABJC01
B. anthracis	NR-1041	This Work
B. anthracis	NR-1202	This Work
B. anthracis	NR-1389	This Work
B. anthracis	NR-1408	This Work
B. anthracis	NR-36073	This Work
B. anthracis	NR-36091	This Work
B. anthracis	NR-3838	This Work
B. anthracis	NR-411	This Work
B. anthracis	NR-41	This Work
B. anthracis	NR-46	This Work
B. anthracis	NR-51483	This Work
B. anthracis	NR-51484	This Work
B. anthracis	NR-51485	This Work
B. anthracis	NR-51486	This Work
B. mycoides	ATCC 6462^{T}	CP009692.1
B. mycoides	ATCC 6462^{T}	This Work
B. mycoides	NR-613	This Work
B. cereus	NR-22171	This Work
B. weihenstephanensis	NBRC 101238^{T}	BAUY01
B. weihenstephanensis	KBAB4	CP000903.1
Staphylococcus aureus subsp. *aureus*	DSM 20231^{T}	CP011526.1

A0488^{T}	NR-1041	NR-1202	NR-1389	NR-1408	NR-36073	NR-36091	NR-3838	NR-411	NR-41	NR-46	NR-51483	NR-51484	NR-51485	NR-51486	ATCC 6462^{T}	ATCC 6462^{T}	NR-613	NR-22171	NBRC 101238^{T}	KBAB4	DSM 20231^{T}
100	99.8	99.7	99.7	99.6	99.7	99.6	99.7	99.8	99.5	99.6	99.6	99.8	99.8	99.8	38.2	38	39.7	37.8	37.9	38.2	28
	100	99.8	99.8	99.7	99.7	99.7	99.8	99.8	99.5	99.7	99.7	99.8	99.9	99.8	38.3	38.1	39.8	38	38	38.2	23
		100	99.8	99.8	99.8	99.7	99.9	99.9	99.6	99.8	99.9	99.8	99.8	99.8	38.2	38.1	39.8	37.9	37.9	38.2	23
			100	99.6	99.9	99.8	99.8	99.8	99.7	99.7	99.7	99.8	99.8	99.8	38.3	38.1	39.8	37.9	38	38.2	23
				100	99.7	99.6	99.8	99.8	99.5	99.7	99.7	99.							38.5	38.8	23
					100	99.8	99.8	99.7	99.7	99.7	99.7	99.							38	38.2	22
						100	99.7	99.7	99.6	99.7	99.6	99.							38.1	38.3	23
							100	100	99.6	99.8	99.8	99.							38	38.2	22
								100	99.6	99.8	99.8	99.							38	38.2	23
									100	99.5	99.6	99.							38	38.2	23
										100	99.7	99.7	99.7	99.7	38.6	38.5	39.9	38.2	38.3	38.6	23
											100	99.7	99.7	99.7	38.2	38.1	39.8	37.9	38	38.2	23
												100	99.9	100	38.2	38.1	39.8	37.9	38	38.2	23
													100	99.9	38.2	38.1	39.8	37.9	38	38.2	23
														100	38.2	38.1	39.8	37.9	38	38.2	23
															100	96.8	93.2	89.9	78.7	78.8	27
																100	92.3	88.8	78	78.4	25
																	100	87.7	79.7	80.7	23
																		100	79.8	79	23
																			100	84.4	22
																				100	27
																					100

dDDH

ANI	Interpretation	dDDH
98.0 - 100	Same species and subspecies	80.0 - 100
96.5 - 97.999	Same species, different subspecies	70.0 - 79.9
92.0 - 96.499	Different species	50.0 - 69.9
85.0 - 91.999		30.0 - 49.9
0.0 - 84.999		0.0 - 29.9

Type Strain

Non-Type Strain

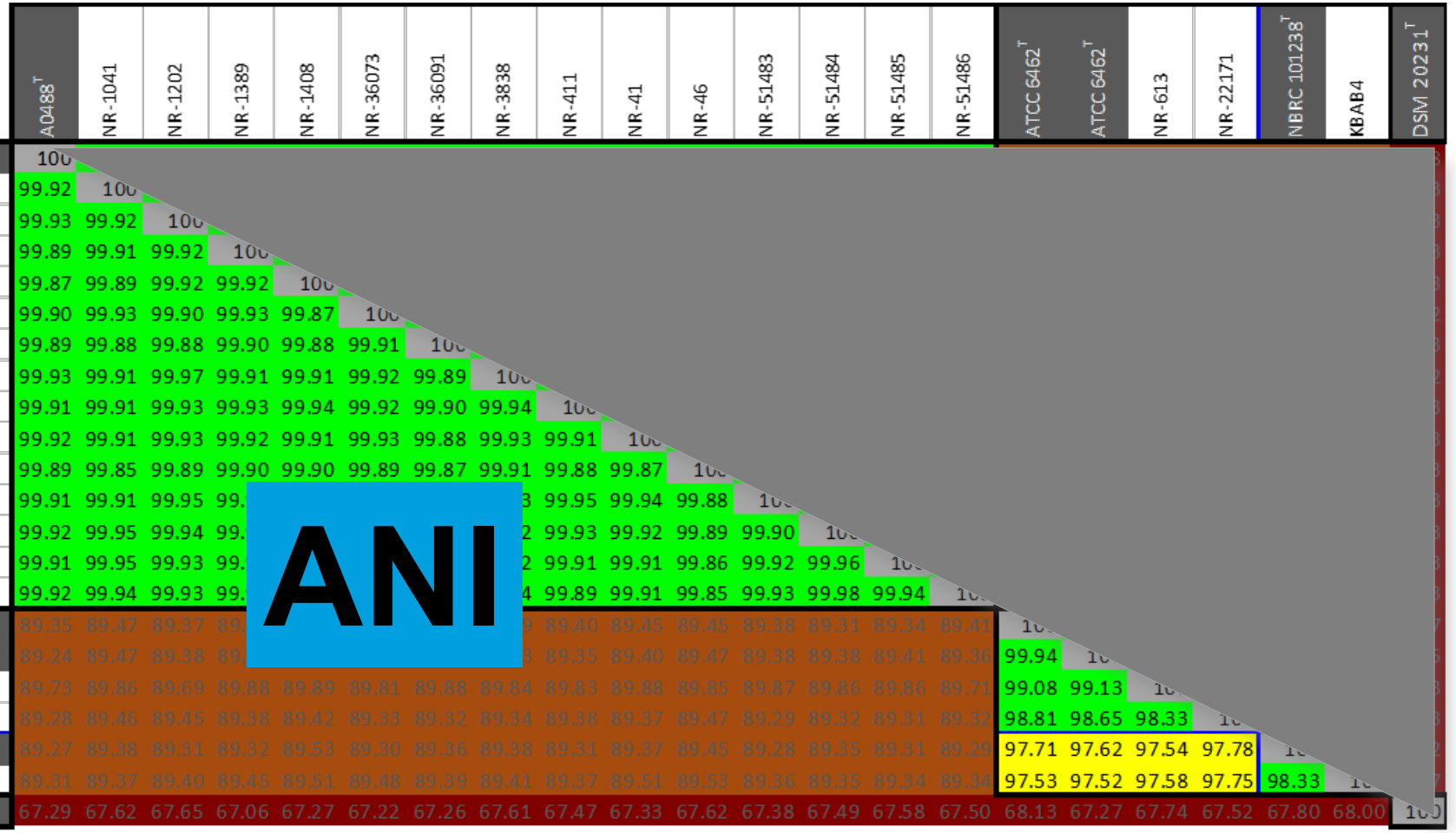
Subspecies Circumscription

Species Circumscription

ATCC

B. anthracis and *B. mycoides*

Species/Subspecies	Strain	Source
B. anthracis	A0488^{T}	ABJC01
B. anthracis	NR-1041	This Work
B. anthracis	NR-1202	This Work
B. anthracis	NR-1389	This Work
B. anthracis	NR-1408	This Work
B. anthracis	NR-36073	This Work
B. anthracis	NR-36091	This Work
B. anthracis	NR-3838	This Work
B. anthracis	NR-411	This Work
B. anthracis	NR-41	This Work
B. anthracis	NR-46	This Work
B. anthracis	NR-51483	This Work
B. anthracis	NR-51484	This Work
B. anthracis	NR-51485	This Work
B. anthracis	NR-51486	This Work
B. mycoides	ATCC 6462^{T}	CP009692.1
B. mycoides	ATCC 6462^{T}	This Work
B. mycoides	NR-613	This Work
B. cereus	NR-22171	This Work
B. weihenstephanensis	NBRC 101238^{T}	BAUY01
B. weihenstephanensis	KBAB4	CP000903.1
Staphylococcus aureus subsp. *aureus*	DSM 20231^{T}	CP011526.1



ANI	Interpretation	dDDH
98.0 - 100	Same species and subspecies	80.0 - 100
96.5 - 97.999	Same species, different subspecies	70.0 - 79.9
92.0 - 96.499	Different species	50.0 - 69.9
85.0 - 91.999		30.0 - 49.9
0.0 - 84.999		0.0 - 29.9

Type Strain

Non-Type Strain

Subspecies Circumscription

Species Circumscription

Bacillus cereus and *Bacillus thuringiensis*

Species/Subspecies	Strain	Source	ATCC 14579^T	ATCC 13367	ATCC 11778	ATCC 33019	ATCC 10792^T	ATCC 10792^T	NR-28583	ATCC 246	NR-610	ATCC 33679	ATCC 10876	ATCC 7039	ATCC 9592	ATCC 700872	ATCC 35646	ATCC 19266
B. cereus	ATCC 14579^T	NC_004722.1	100	82.8	82.5	81.8	71.1	71.3	71	74.1	73.2	73.2	72.5	65.2	65.2	65.8	65.7	65.2
B. thuringiensis	ATCC 13367	This Work	98.15	100	89.4	87.3	69.8	69.5	69.6	75.2	73.2	73.3	73.6	64.	Subspecies 1			64.5
B. cereus	ATCC 11778	This Work	98.06	98.84	100	92.1	71.1	71.2	71	77.7	76.8	76.7	75.7	66.				66.9
B. cereus	ATCC 33019	This Work	97.96	98.67	99.10	100	70.7	70.8	70.5	77.5	75.7	75.5	74.8	65.9	66	66.3	66	66.1
B. thuringiensis	ATCC 10792^T	CP020754.1	96.65	96.46	96.58	96.69	100	99.1	82.6	68.1	67	67.1	67.2	68.9	69	69.3	68.9	68.8
B. thuringiensis	ATCC 10792^T	This Work	96.60	96.44	96.73	96.75	99.83	100	82.2	67.9	67.2	67.3	66.8	68.	Subspecies 2			68
B. thuringiensis	NR-28583	This Work	96.62	96.38	96.66	96.65	98.02	98.01	100	68.7	68	68	67.9	70	70.2	70.3	69.8	69.9
B. cereus	ATCC 246	This Work	97.13	97.28	97.51	97.54	96.21	96.14	96.35	100	87.9	87.7	81.6	65.1	65.2	65.7	65.4	64.7
B. thuringiensis	NR-610	This Work	97.04	97.16	97.42	97.32	96.15	96.21	96.20	98.80	100	96.7	79.5	63.	Subspecies 3			63.6
B. thuringiensis	ATCC 33679	This Work	97.04	97.13	97.41	97.34	96.17	96.23	96.33	98.78	99.92	100	79.6	63.				63.5
B. cereus	ATCC 10876	This Work	96.91	97.11	97.32	97.26	96.10	96.04	96.31	98.04	97.85	97.86	100	64.2	64.3	65.2	64.8	63.5
B. cereus	ATCC 7039	This Work	95.88	95.78	96.04	96.11	96.42	96.25	96.49	95.82	95.66	95.63	95.69	100	95.6	93.6	92.6	92.5
B. cereus	ATCC 9592	This Work	95.83	95.82	96.06	96.04	96.42	96.27	96.59	95.80	95.65	95.62	95.70	99.92	100	94	93	92.8
B. thuringiensis	ATCC 700872	This Work	95.97	95.91	96.10	96.04	96.42	96.32	96.50	95.93	95.74	95.67	95.89	99.27	99.32	100	99.7	91.8
B. thuringiensis	ATCC 35646	This Work	95.91	95.80	96.11	96.02	96.38	96.21	96.53	95.83	95.65	95.60	95.75	99.20	99.16	99.93	100	91.1
B. thuringiensis	ATCC 19266	This Work	95.83	95.73	96.15	96.00	96.33	96.28	96.53	95.83	95.61	95.61	95.57	99.12	99.11	99.06	99.02	100

ANI	Interpretation	dDDH
98.0 - 100	Same species and subspecies	80.0 - 100
96.5 - 97.999	Same species, different subspecies	70.0 - 79.9
92.0 - 96.499	Different species	50.0 - 69.9
85.0 - 91.999		30.0 - 49.9
0.0 - 84.999		0.0 - 29.9

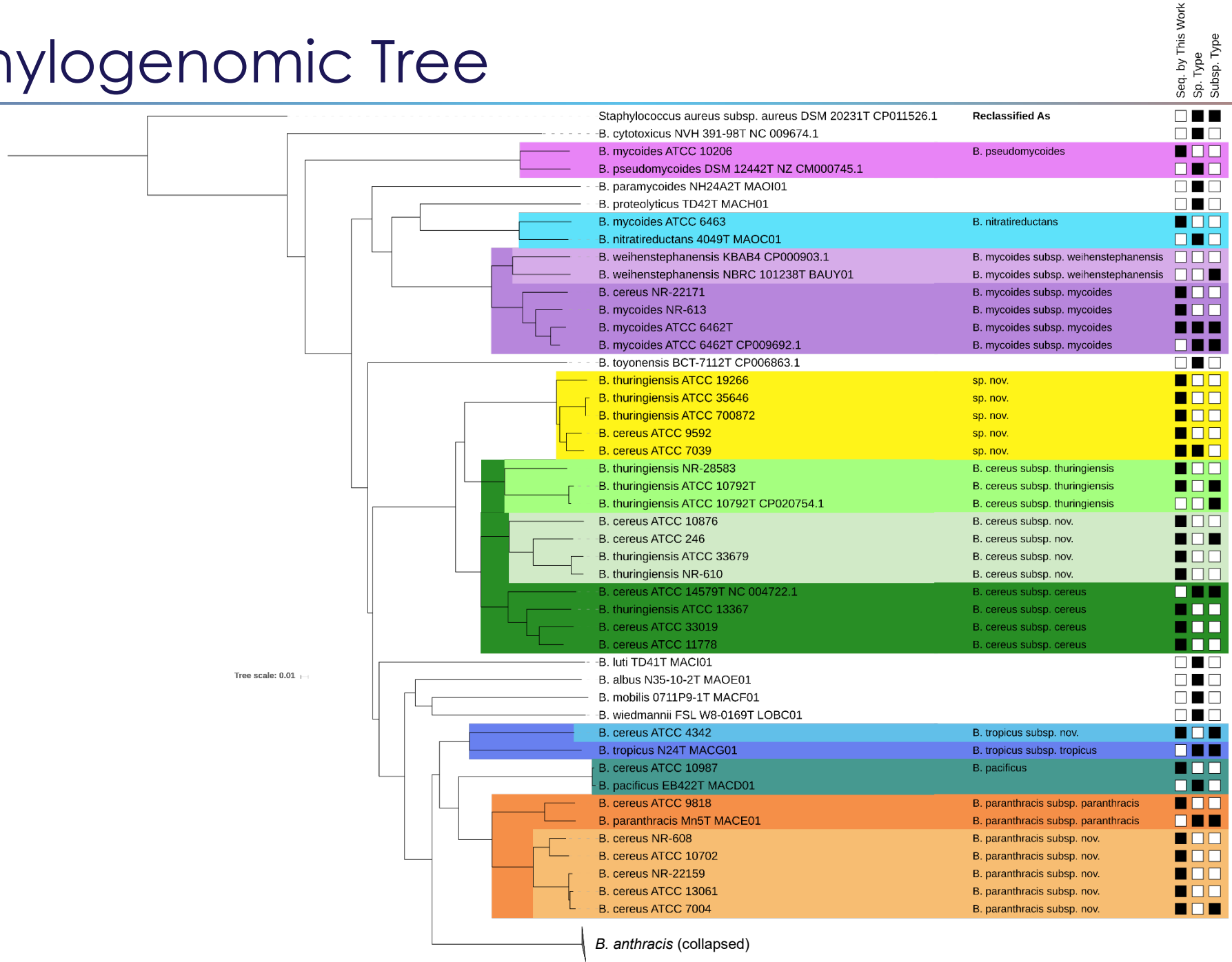
Novel Species

Other BcG Strains

Species/Subspecies	Strain	Source	BCT-7112^T	FSL W8-0169^T	Mn5^T	ATCC 9818	ATCC 10702	NR-608	NR-22159	ATCC 13061	ATCC 7004	EB422^T	ATCC 10987	N24^T	ATCC 4342	N35-10-2^T	0711P9-1^T	TD41^T	TD42^T	4049^T	ATCC 6463	NH24A2^T	DSM 12442^T	ATCC 10206	NVH 391-98^T
B. toyonensis	BCT-7112^T	CP006863.1	100	44	42.8	42.8	42.7	43.2	42.7	42.8	42.8	42.5	42.5	43.2	43.1	42.6	42.7	43.6	42.4	42	42	39.3	27.3	27.3	26.2
B. wiedmannii	FSL W8-0169^T	LOBC01	91.30	100	50.9	50.8	51.3	51.8	51.1	51.1	51.1	51.6	51.7	52	51.6	54.1	58.1	45.2	40.9	42.2	42.2	37.7	27	27	25.5
B. paranthracis	Mn5^T	MACE01	90.98	93.18	100	91.7	78.6	79	77.4	77.2	77.4	65.4	65.3	61.9	64.3	52.7	49.9	44.6	39.1	39.7	39.6	36.6	26.8	26.9	25.4
B. cereus	ATCC 9818	This Work	90.92	93.14	99.08	100	78.5	79	77.8	77.5	77.2	65.7	65.5	61.6	64.4	52.8	49.9	44.7	39.2	39.6	39.6	36.7	26.8	26.9	25.6
B. cereus	ATCC 10702	This Work	90.81	93.15	97.58	97.60	100	95.2	90	89.9	89.4	66	67	62.6	65.4	52.1	51	44.3	39.3	39.6	39.6	36.8	27	27.4	25.5
B. cereus	NR-608	This Work	90.95	93.33	97.69	97.67	99.93	100	90.3	90.2	89.8	66.6	67.6	63.1	65.8	52.6	51.6	44.9	39.9	40.1	40.1	37.3	27.2	27.4	25.8
B. cereus	NR-22159	This Work	90.88	93.24	97.47	97.55	98.90	98.86	100	99	98.6	66.7	66.8	62.6	65.3	51.9	50.7	44.3	39.2	39.6	39.5	36.6	26.8	27	25.5
B. cereus	ATCC 13061	This Work	90.84	93.23	97.46	97.54	98.87	98.82	99.78	100	98.9	66.4	66.7	62.6	65.1	52	50.2	44.3	39.2	39.6	39.7	36.7	26.8	27	25.5
B. cereus	ATCC 7004	This Work	90.89	93.16	97.48	97.36	98.79	98.79	99.80	99.82	100	66.4	66.5	62.4	65.1	52	50	44.3	39.2	39.6	39.6	36.7	26.9	27.1	25.5
B. pacificus	EB422^T	MACD01	90.77	93.23	95.94	95.82	95.88	96.01	96.01	96.00	95.93	100	99.3	60	59.7	52.4	51.7	44.3	39.2	39.9	39.8	36.7	26.8	26.8	25.6
B. cereus	ATCC 10987	This Work	90.77	93.24	95.82	95.87	96.08	96.15	96.06	96.07	96.00	99.86	100	60.3	59.9	52.6	52	44.4	39.2	39.9	39.9	36.7	26.8	26.9	25.8
B. tropicus	N24^T	MACG01	91.09	93.36	95.33	95.27	95.37	95.51	95.48	95.44	95.40	94.97	95.04	100	70.7	52.2	50.7	45.3	39.8	39.9	40	37	26.9	26.9	25.5
B. cereus	ATCC 4342	This Work	90.96	93.30	95.66	95.68	95.76	95.90	95.84	95.85	95.80	94.89	94.94	96.70	100	53.1	50.5	44.8	39.6	40	39.9	36.9	27	27.3	25.6
B. albus	N35-10-2^T	MAOE01	90.95	93.77	93.45	93.46	93.32	93.54	93.35	93.42	93.28	93.45	93.44	93.37	93.58	100	54.4	44.1	39.1	40.1	40.1	36.6	26.9	26.8	25.4
B. mobilis	0711P9-1^T	MACF01	90.90	94.65	92.96	92.93	93.18	93.37	93.16	92.95	92.97	93.28	93.39	93.10	93.07	93.94	100	44.1	39.9	41.3	41.3	37.3	26.7	26.7	25.5
B. luti	TD41^T	MACI01	91.13	91.62	91.46	91.50	91.51	91.66	91.50	91.43	91.45	91.44	91.43	91.84	91.64	91.35	91.41	100	40.7	40	40.1	38	26.9	26.9	25.5
B. proteolyticus	TD42^T	MACH01	90.71	90.26	89.72	89.66	89.73	89.89	89.75	89.67	89.67	89.66	89.80	89.81	89.74	89.55	89.86	90.15	100	58.4	57.6	43.8	28.1	28	25.6
B. nitratireductans	4049^T	MAOC01	90.66	90.74	90.03	89.95	89.85	90.10	89.99	89.91	89.91	90.03	90.06	90.07	90.06	90.07	90.47	90.10	94.75	100	86.3	43.7	27.7	27.6	25.6
B. mycoides	ATCC 6463	This Work	90.64	90.70	89.80	89.81	89.85	90.04	89.80	89.85	89.93	89.93	89.98	90.05	89.93	90.03	90.42	90.11	94.59	98.55	100	43.6	28	27.7	25.7
B. paramycoides	NH24A2^T	MAOI01	89.66	89.17	88.73	88.84	88.79	88.91	88.94	88.89	88.76	88.74	88.86	89.02	88.69	88.73	89.05	89.15	91.20	91.25	91.29	100	28.1	28	25.9
B. pseudomycoides	DSM 12442^T	NZ_CM000745.1	82.65	82.50	82.32	82.36	82.23	82.38	82.37	82.39	82.40	82.32	82.30	82.37	82.47	82.33	82.18	82.53	83.17	82.95	83.11	83.41	100	86.5	28
B. mycoides	ATCC 10206	This Work	82.61	82.21	82.23	82.34	82.16	82.49	82.29	82.29	82.22	82.33	82.32	82.24	82.29	82.24	82.22	82.45	82.92	82.80	82.88	83.14	98.59	100	27.9
B. cytotoxicus	NVH 391-98^T	NC_009674.1	81.44	81.29	81.15	81.15	81.24	81.30	81.38	81.22	81.23	81.32	81.39	81.27	81.16	81.31	81.27	81.37	81.31	81.58	81.48	81.86	83.20	83.07	100

ANI	Interpretation	dDDH
98.0 - 100	Same species and subspecies	80.0 - 100
96.5 - 97.999	Same species, different subspecies	70.0 - 79.9
92.0 - 96.499	Different species	50.0 - 69.9
85.0 - 91.999		30.0 - 49.9
0.0 - 84.999		0.0 - 29.9

BcG Phylogenomic Tree



Summary of Strains from ATCC and BEI Resources

41 in-house strains sequenced

- 27 items should receive some form of name change
- 23 transferred to a different species (*e.g.*, *B. cereus* to *B. pacificus*)
- 19 added to a subspecies for additional specificity

Publication of Results

- **Poster: ASM Microbe 2019, San Francisco, CA**
  - EEB02 - Microbial Evolution and Comparative Genomics 1
  - Poster Board Number: **FRIDAY - EEB-488**
  - Friday, June 21, 2019
    - 11 AM - 12 PM PT
    - 4 PM - 5 PM PT

- **Scientific Manuscript Submitted: *Int J Syst Evol Microbiol***
  - Phylogenomic Reclassification of ATCC *Bacillus* Strains and Various Taxa within the Genus *Bacillus*
    - Marco A. Riojas, Andrew M. Frank, Samantha L. Fenn, Stephen King, Sonia Brower, Manzour Hernando Hazbón

Acknowledgements

- Joe Leonelli, PhD

- Briana Benton

- Samantha Fenn

- Andrew Frank

- Manzour Hazbón, PhD

- Stephen King

- Juan Lopera, PhD

- Anna McCluskey

NIH National Institute of Allergy and Infectious Diseases

bei RESOURCES

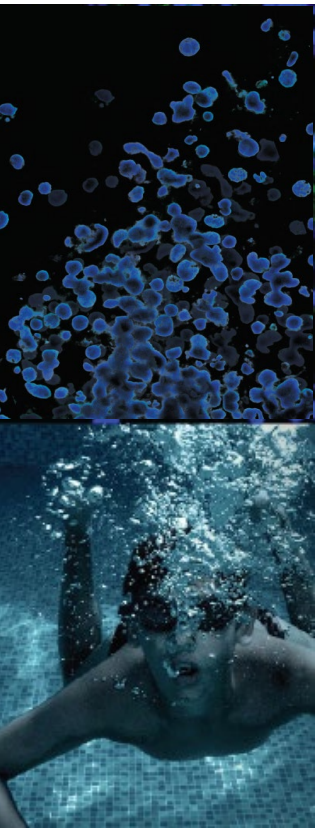
SUPPORTING INFECTIOUS DISEASE RESEARCH

National Institute of Allergy and Infectious Diseases (NIAID)

The following reagents were obtained through BEI Resources, NIAID, NIH: NR-41, NR-46, NR-411, NR-608, NR-610, NR-613, NR-1041, NR-1202, NR-1389, NR-1408, NR-3838, NR-22159, NR-22171, NR-28583, NR-36073, NR-36091, NR-51483, NR-51484, NR-51485, and NR-51486.

ATCC

ATCC®



Questions?

Credible Leads to Incredible™