# Long read 16S (briefly), plus reference guided metagenomic assembly, genome alignment, & visualization

Natalie Kokroko
PhD student, Treangen lab/quest
Rice University, Houston TX

Michael Nute
Research Scientist, Treangen lab/quest
Rice University, Houston TX

Todd J Treangen
Associate Professor, Rice University
Bioengineering & Computer Science

Name: **Todd J. Treangen / Associate Professor**

Institution: **Rice University (Computer Science) – since July 2018**

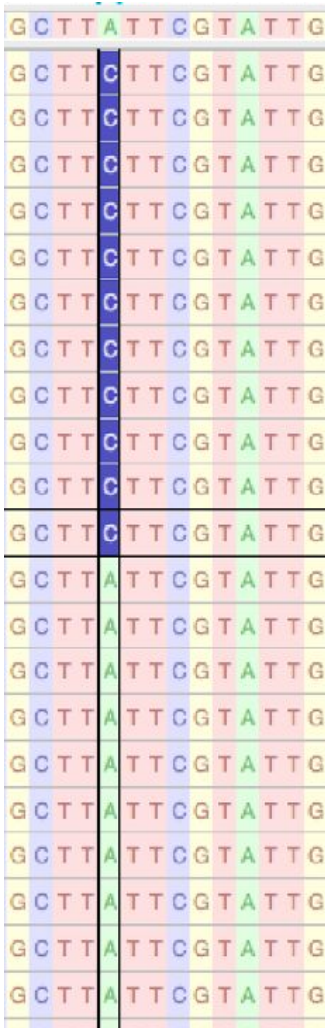
Email: **treangen@rice.edu**

Web: **www.treangenlab.com**

**Research Interests:** Metagenomics, Engineering detection, DNA screening, Infectious disease transmission, biodefense, microbial forensics

**Prior to Rice**

- 2016-2018: Research Assistant Professor (University of Maryland) with Mihai
- 2012-2016: PI, Genomics, NBFAC
- 2010-2012: Postdoctoral Scientist, Johns Hopkins & UMD
- 2003-2008: PhD in Computer Science, Polytechnic University of Catalonia
- 1999-2003: Software engineer (python, C++)




- **Office**: DH 3103
- **Web**: www.treangenlab.com
- **Email**: treangen@rice.edu

# Treangen lab May 2025

# Rice University

# My STAMPS experience (and related)

-First heard lots of great things about STAMPS back in 2016, while a member Mihai Pop's research group at University of Maryland College Park.

-Participated in my first STAMPS as instructor back in 2018, then again in 2019, 2022, (not offered in 2020, 2021, and I missed 2023 sadly), back in 2024!

-Very happy to be back for 2025 and looking forward to hanging out for a few days

**2018 Course Faculty**

Titus Brown, University of California at Davis
Susan Holmes, Stanford University
Curtis Huttenhower, Harvard University
Rob Knight, University of California, San Diego
David Mark Welch, Marine Biological Laboratory
Christian Mueller, Simons Foundation
Mihai Pop, University of Maryland
Mitch Sogin, Marine Biological Laboratory
Tracy Teal, The Carpentries
Todd Treangen, University of Maryland
Tandy Warnow, University of Illinois at Urbana-Champaign
Amy Willis, University of Washington

# Agenda/overview for the next ~3 hours

**75 minutes of lecture**, **75 minutes of tutorials/games**, **30 minutes of Q&A/breaks**

- **9:00am to 9:20am**: Introduction + kickoff
- **9:20am to 9:35am**: Egg break game
- **9:35am to 9:50am**: Brief long-read 16S lecture
- **9:50am to 10:05am:** Q&A/Break 1
- **10:05am to 10:25am**: Emu hands-on tutorial
- **10:25am to 10:40am**: De novo vs reference guided metagenomic assembly
- **10:40am** to **11:00am**: Assembly game
- **11:00am to 11:10am:** Q&A/Break 2
- **11:10am to 11:35am**: Strain analysis lecture
- **11:35am to 11:55am**: Parsnp/Gingr hands-on tutorial
- **11:55am to noon**: Recap/Overflow
- Bonus material (for evening): De novo Assembly + Binning hands-on tutorial
  - Roughly should take 30-40 minutes to get through

# Thoughts when brainstorming for today

- Setup very nicely thanks to previous lectures and tutorials (Thank you Titus!)
- I briefly considered making this an escape room game, where you'd have to accurately assemble and bin a real metagenome
- Settled on two games that I have play tested previously at STAMPS, and I hope you all enjoy (more on that later)
- Much of what will be presented today is inspired by previous STAMPS interactions and discussions with Mihai, Titus, Amy, Mike Lee, and many others!

# Active Research areas

**Data structures and algorithms**

**Software engineering**

**Pathogen diagnostics and detection**

# Types of research questions my group focuses on

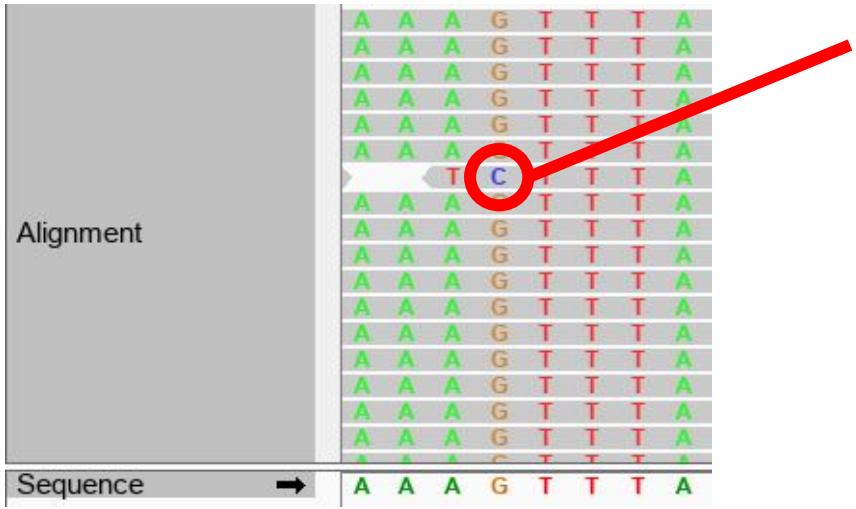
Computational microbial genomics:

1. Is this a mutation or is it a sequencing error?
2. Is this microbe *really* in the sample or is it a contaminant?
3. Is this horizontal gene transfer or chimeric assembly artifact/error?
4. Is this microbe detected in an metagenomic sample harmful to human health?
5. Is it possible to develop methods that can scale up to terabyte to petabyte scale datasets without huge accuracy/sensitivity tradeoffs?

# Types of research questions my group focuses on

Computational microbial forensics:

1. Is this a legit mutation or is it a sequencing error?



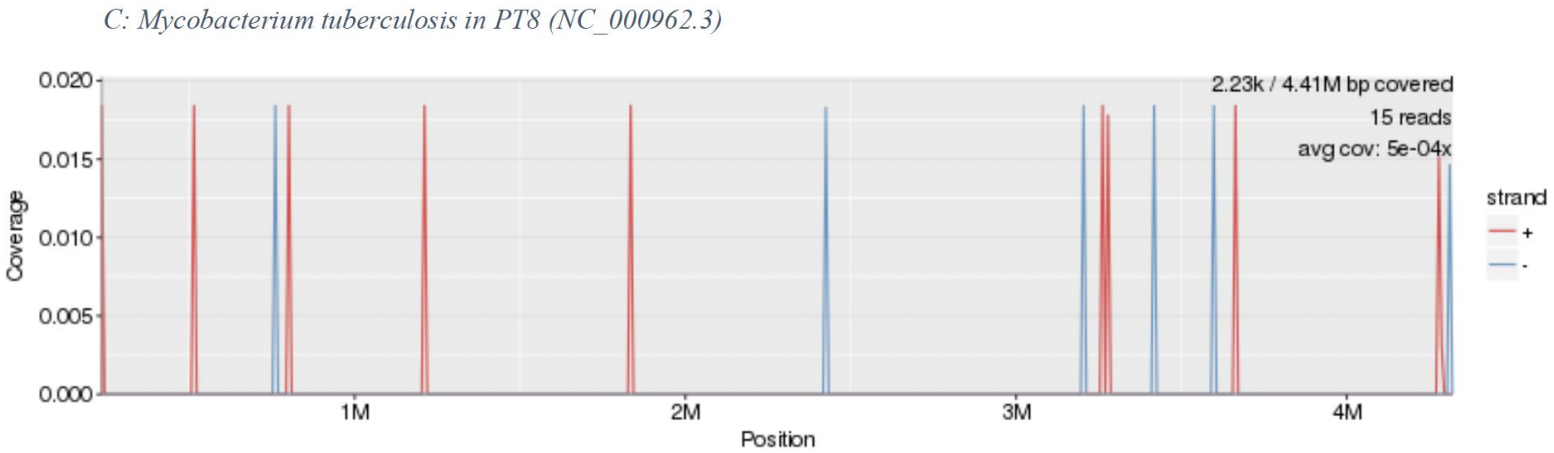

Is this base a true variant or a sequencing error?

Sequencing error rates varies significantly between technologies, runs, lanes, multiplexes, genomic location as well as substitution types

# Types of research questions my group focuses on

Computational microbial forensics:

2. Is this microbe *really* in the sample or is it a contaminant?




https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0

# Types of research questions my group focuses on

Computational microbial forensics:

3. Is this horizontal gene transfer or misassembly/chimeric contig?

Software

Open Access

**Genome assembly forensics: finding the elusive mis-assembly**

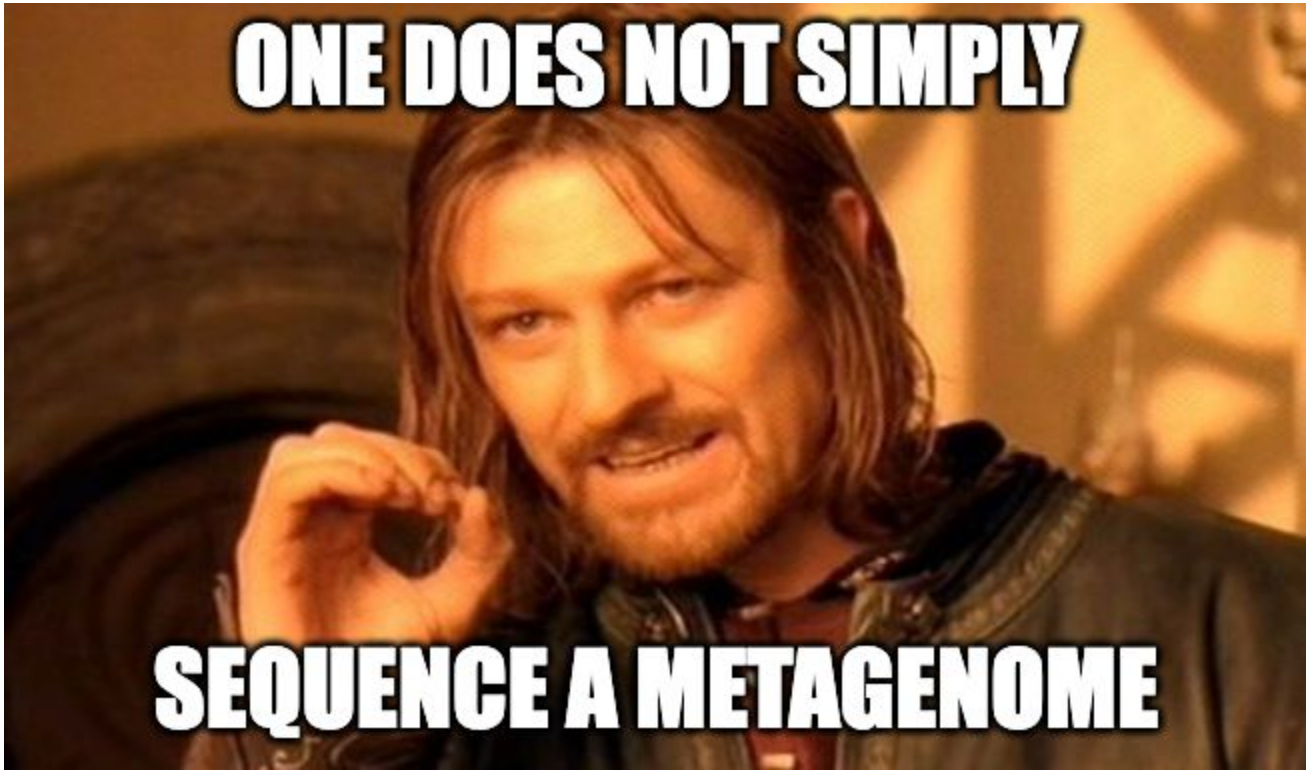
Adam M Phillippy, Michael C Schatz and Mihai Pop

Address: Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA.

Correspondence: Mihai Pop. Email: mpop@umiacs.umd.edu

**Ten issues to be aware of when sequencing and analyzing metagenomes:**

1. Sample storage and prep can influence results!
2. Hard to lyse vs easy to lyse microbes can create biased community profiles!
3. Underrepresentation of extreme GC content microbes
4. Kit contamination/Cross-contamination/Environmental contamination
5. Uneven coverage/coverage gaps for diverse microbial communities
6. Running out of $$$ (unbiased is expensive)
7. Running out of time/patience/storage to analyze 100s/1000s of samples
8. Not enough input DNA/RNA for sequencing platform, and none left
9. Intra vs inter genomic repeats can bias counts/observations, snarl assemblies
10. Lots of different ways to analyze the data!


ONE DOES NOT SIMPLY
SEQUENCE A METAGENOME

# A microbiome is a "forest" of microbes

Paul Cézanne, circa 1902-1904

# Sequencing machines turn these forests into twigs

Paul Cézanne, circa 1902-1904

# Current computational tools turn twigs into wooden puzzle pieces

# Goal is to turn puzzle pieces back into forest

Paul Cézanne, circa 1902-1904

# ...while avoiding misassemblies!

*African baobab tree

# A brief detour on the size of the puzzle piece (kmer)

JOURNAL ARTICLE

# Informed and automated *k*-mer size selection for genome assembly



Rayan Chikhi , Paul Medvedev ✉ Author Notes

*Bioinformatics*, Volume 30, Issue 1, January 2014, Pages 31–37, https://doi.org/10.1093/bioinformatics/btt310

**Published:** 03 June 2013 **Article history** ▾

# Computational approaches for analyzing genome/metagenomes

**A**
**0**
**1**

# Computational approaches for analyzing genome/metagenomes

**AG**
**00**
**12**

# Computational approaches for analyzing genome/metagenomes

```
AGC
000
123
```

# Computational approaches for analyzing genome/metagenomes

```
AGCT
0000
1234
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTC
00000
12345
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCG
000000
123456
```

# Computational approaches for analyzing genome/metagenomes

Note: Common kmer size (7) for viral genome analysis

AGCTCGA
000
123

RESEARCH ARTICLE

## Identification and Genomic Analysis of a Novel Group C Orthobunyavirus Isolated from a Mosquito Captured near Iquitos, Peru

Todd J. Treangen[1]*, George Schoeler[2¤a], Adam M. Phillippy[1¤b], Nicholas H. Bergman[1], Michael J. Turell[3]

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGAT
00000000
12345678
```

# Computational approaches for analyzing genome/metagenomes

Note: Common kmer size (9) for functional profiling

```
AGCTCGATT
000000000
123456789
```

Balaji *et al. Genome Biology* (2022) 23:133
https://doi.org/10.1186/s13059-022-02695-x

Genome Biology

SOFTWARE Open Access

## SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning

Advait Balaji[1†], Bryce Kille[1†], Anthony D. Kappell[2], Gene D. Godbold[3], Madeline Diep[4], R. A. Leo Elworth[1], Zhiqin Qian[1], Dreycey Albin[1], Daniel J. Nasko[5], Nidhi Shah[5], Mihai Pop[5], Santiago Segarra[6], Krista L. Ternus[2*] and Todd J. Treangen[1*]

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTA
0000000001
1234567890
```

# Computational approaches for analyzing genome/metagenomes

Note: Default kmer size (11) for blastn

```
AGCTCGATTAC
00000000011
12345678901
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACA
000000000111
123456789012
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAG
0000000001111
1234567890123
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGG
00000000011111
12345678901234
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGT
000000000111111
123456789012345
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTA
0000000001111111
1234567890123456
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAA
00000000011111111
12345678901234567
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAA
0000000001111111
1234567890123456
```

# Computational approaches for analyzing genome/metagenomes

Note: Common minimum size (19) for maximal unique match length

```
AGCTCGATTACAGGTAAAT
00000000000000000000
12345678901234567890
```

*Bioinformatics*, 2024, **40(5)**, btae311
https://doi.org/10.1093/bioinformatics/btae311
Advance Access Publication Date: 9 May 2024
**Applications Note**

OXFORD

Genome analysis

## Parsnp 2.0: scalable core-genome alignment for massive microbial datasets

**Bryce Kille [1,*], Michael G. Nute[1], Victor Huang[1], Eddie Kim[1], Adam M. Phillippy [2], Todd J. Treangen [1,3,*]**

[1]Department of Computer Science, Rice University, Houston, TX 77005, United States
[2]Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, United States
[3]Department of Bioengineering, Rice University, Houston, TX 77030, United States

*Corresponding authors. Department of Computer Science, Rice University, 6100 Main St., Houston, TX 77005, United States. E-mails: blk6@rice.edu (B.K.) and treangen@rice.edu (T.J.T.)

Associate Editor: Russell Schwartz

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATC
00000000011111111112
12345678901234567890
```

# Computational approaches for analyzing genome/metagenomes

Note: Common kmer size (21) for minhash analysis of bacteria

```
AGCTCGATTACAGGTAAATCT
0000000
1234567
```

Ondov *et al. Genome Biology* (2016) 17:132
DOI 10.1186/s13059-016-0997-x

Genome Biology

SOFTWARE Open Access

## Mash: fast genome and metagenome distance estimation using MinHash

Brian D. Ondov[1], Todd J. Treangen[1], Páll Melsted[2], Adam B. Mallonee[1], Nicholas H. Bergman[1], Sergey Koren[3] and Adam M. Phillippy[3*]

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTG
00000000011111111111222
12345678901234567890123
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGG
000000000111111111122222
123456789012345678901 23
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGGC
000000000111111111122222
123456789012345678901234
```

# Computational approaches for analyzing genome/metagenomes

Note: Common match size (25) for primer design

AGCTCGATTACAGGTAAATCTGGCT

nature communications

Article

https://doi.org/10.1038/s41467-024-49957-9

## Olivar: towards automated variant aware primer design for multiplex tiled amplicon sequencing of pathogens

Received: 2 August 2023

Accepted: 25 June 2024

Published online: 26 July 2024

Michael X. Wang[1], Esther G. Lou[2], Nicolae Sapoval[3], Eddie Kim[3], Prashant Kalvapalle[2], Bryce Kille[3], R. A. Leo Elworth[3], Yunxi Liu[3], Yilei Fu[3], Lauren B. Stadler[2] ✉ & Todd J. Treangen[1,3] ✉

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGGCTA
00000000011111111112222222
12345678901234567890123456
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGGCTAT
00000000011111111122222222
12345678901234567890123456
```

# Computational approaches for analyzing genome/metagenomes

Note: Default kmer size (28) for megablast

```
AGCTCGATTACAGGTAAATCTGGCTATC
000000000111111111222222222
123456789012345678901234567 8
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGGCTATCA
000000000111111111222222222
123456789012345678901234567 89
```

# Computational approaches for analyzing genome/metagenomes

```
AGCTCGATTACAGGTAAATCTGGCTATCAT
00000000011111111122222222223
12345678901234567890123456789O
```

# Computational approaches for analyzing genome/metagenomes

Note: Common kmer size (31) for taxonomic profiling

**AGCTCGATTACAGGTAAATCTGGCTATCATG**

Nasko *et al. Genome Biology* (2018) 19:165
https://doi.org/10.1186/s13059-018-1554-6

Genome Biology

OPEN LETTER Open Access

## RefSeq database growth influences the accuracy of *k*-mer-based lowest common ancestor species identification

Daniel J. Nasko[1], Sergey Koren[2], Adam M. Phillippy[2] and Todd J. Treangen[3*]

As RefSeq grows, is pathogen identification with k-mer based methods getting better, or worse?

Nasko et al, 2018 (Genome Biology)

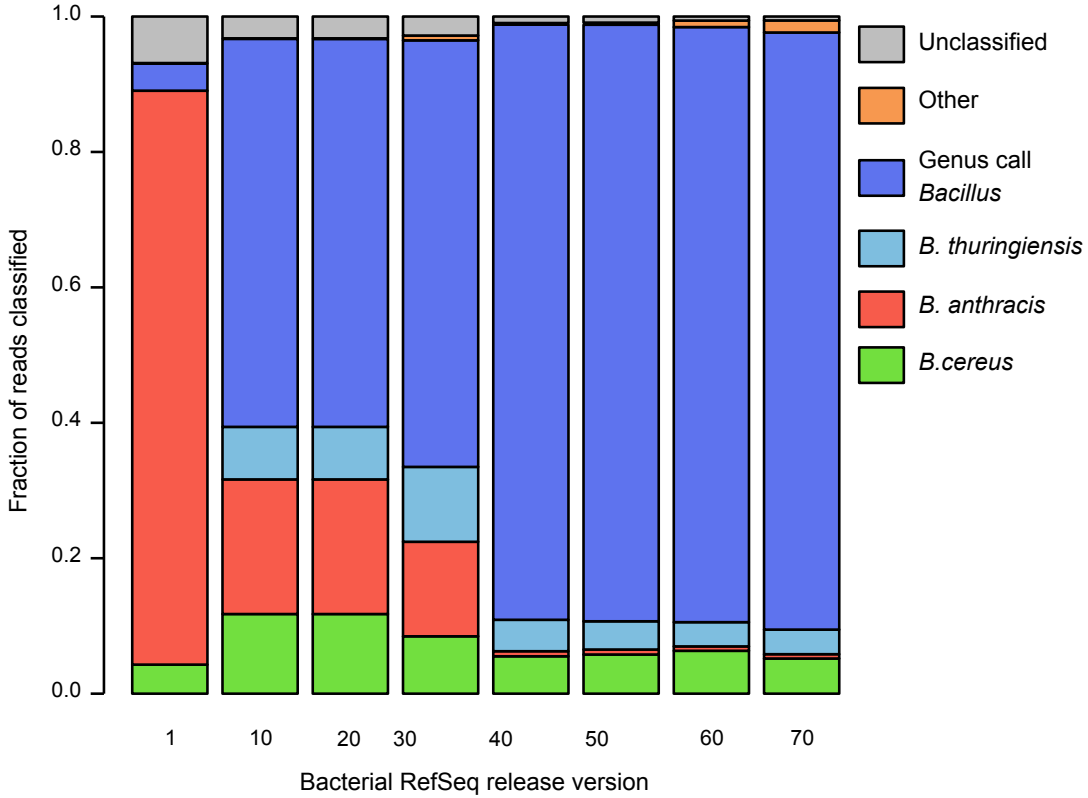
# Testing with one novel genome

Simulate 10,000 Illumina reads using a genome not in RefSeq versions 1-70

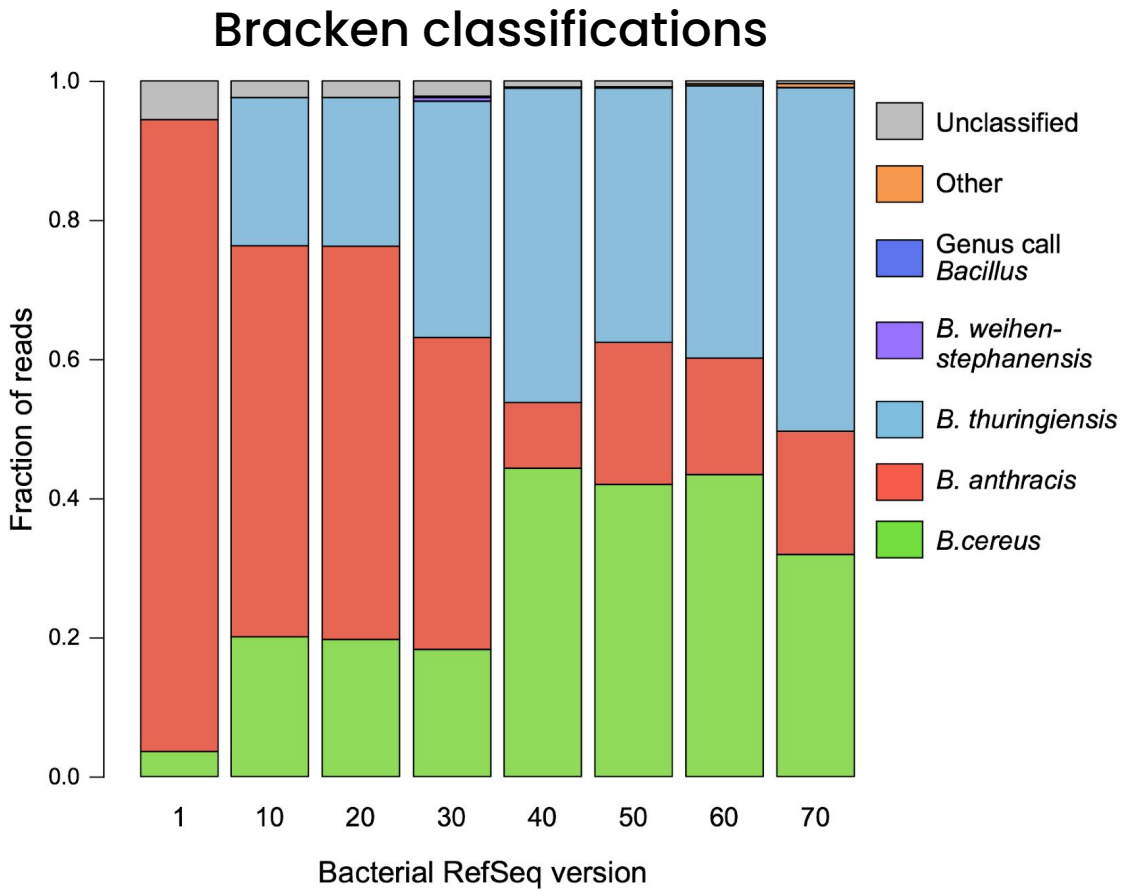
*Bacillus cereus* strain ISSFR-23F

Run this query set against bacterial RefSeq 1,10,20..70 using Kraken and Bracken

# Kraken classifications




Genus-level calls increase for Kraken as the DB grows (again)




Bracken improves the number of *B. cereus* (correct) classifications

# Agenda/overview for the next ~3 hours

**75 minutes of lecture**, **75 minutes of tutorials/games**, **30 minutes of Q&A/breaks**

- **9:00am to 9:20am**: Introduction + kickoff
- **9:20am to 9:35am**: Egg break game
- **9:35am to 9:50am**: Brief long-read 16S lecture
- **9:50am to 10:05am:** Q&A/Break 1
- **10:05am to 10:25am**: Emu hands-on tutorial
- **10:25am to 10:40am**: De novo vs reference guided metagenomic assembly
- **10:40am** to **11:00am**: Assembly game
- **11:00am to 11:10am:** Q&A/Break 2
- **11:10am to 11:35am**: Strain analysis lecture
- **11:35am to 11:55am**: Parsnp/Gingr hands-on tutorial
- **11:55am to noon**: Recap/Overflow
- Bonus material (for evening): De novo Assembly + Binning hands-on tutorial
  - Roughly should take 30-40 minutes to get through