Metagenomic assembly (mostly), Binning (briefly), and more on Quality Control.

Todd J Treangen
Associate Professor, Computer Science
Rice University, Houston TX

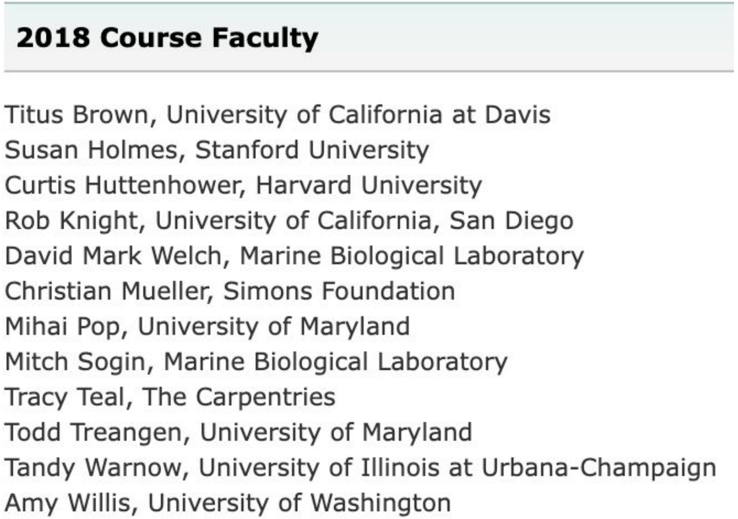
Michael Nute
Research Scientist, Computer Science
Rice University, Houston TX

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)

-Very happy to be back for 2024, and joined by research scientist and metagenomics guru Mike Nute



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, **60 minutes of hands on/tutorials**, **25 minutes of Q&A/breaks**

- **9:00am to 9:20am**: Introduction to Monday Metagenomics + kickoff
- **9:20am to 9:35am**: QC and front matter
- **9:35am to 9:50am**: Metagenomic assembly part 1
- **9:50am to 10:00am**: Q&A/Break
- **10:00am to 10:30am**: Metagenomic assembly part 2
- **10:30am** to **10:45am**: Assembly validation and Assembly questions
- **10:45am to 11:00am:** Q&A/Break
- **11:00am to 12:00pm**: hands on tutorials
- **11:00am to 11:30am**: Assembly game (Todd)
- **11:30am to 11:40am**: QC/cleaning
- **11:40am to 12:00pm**: De novo assembly tutorial

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 assembly and bin a real metagenome
- Settled on a game that 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

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

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

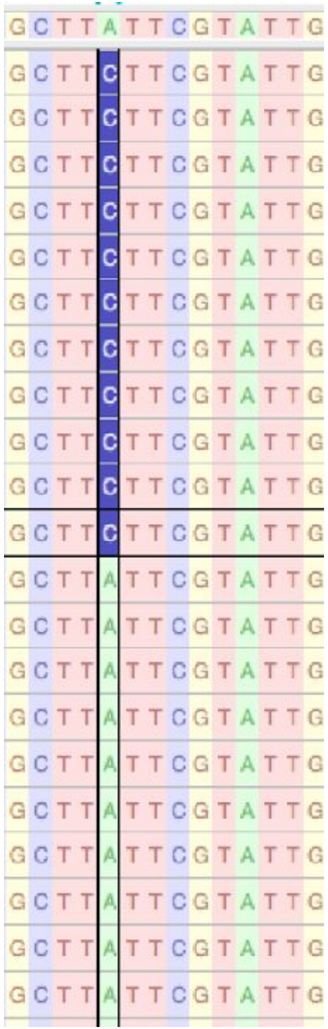
Email: **treangen@rice.edu**

Web: **www.treangenlab.com**

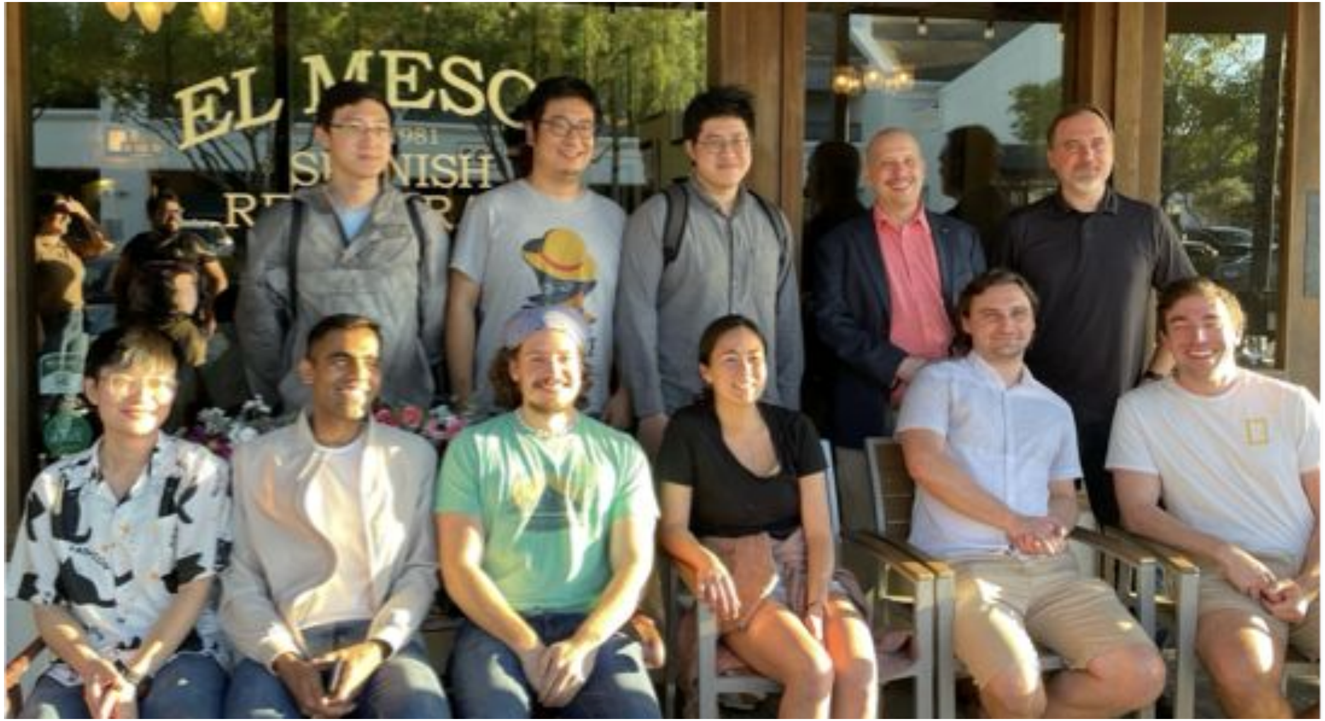
Research Interests: Metagenomics, Engineering detection, DNA screening, Infectious disease transmission, biodefense, microbial forensics

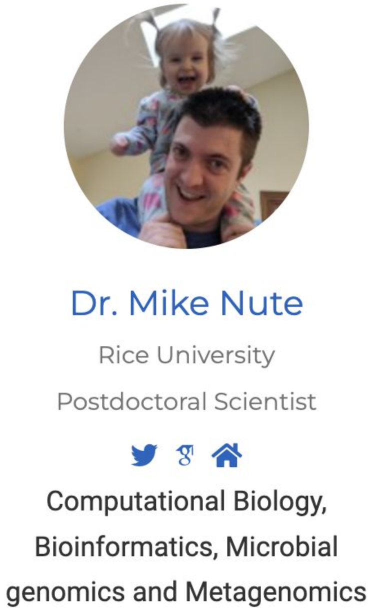
Prior to Rice

- 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++)



```
GCTTATTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTCTTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
GCTTATTCGTATTG
```




Dr. Mike Nute

Rice University

Postdoctoral Scientist

Computational Biology, Bioinformatics, Microbial genomics and Metagenomics

Treangen lab circa May 2024

Attending STAMPS

Active Research areas

Data structures and algorithms

Software engineering

Pathogen diagnostics and detection

Types of research questions my group focuses on

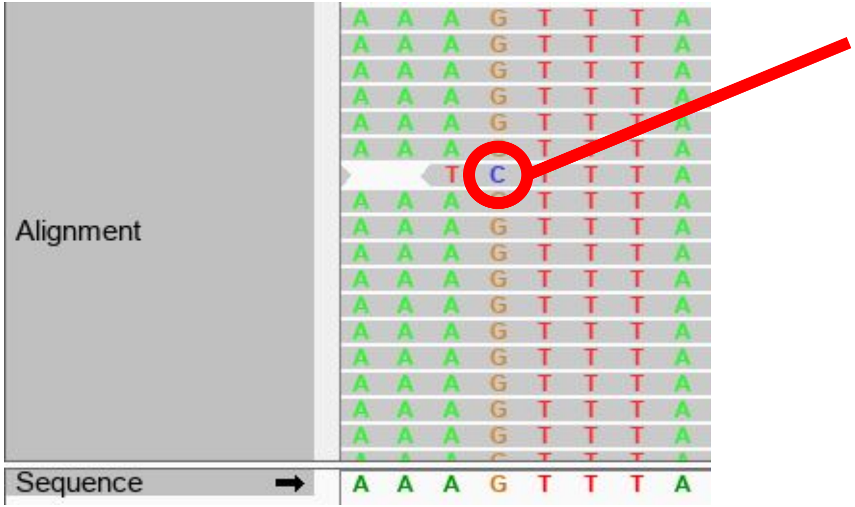
Computational microbial genomics:

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 environmental 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 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?

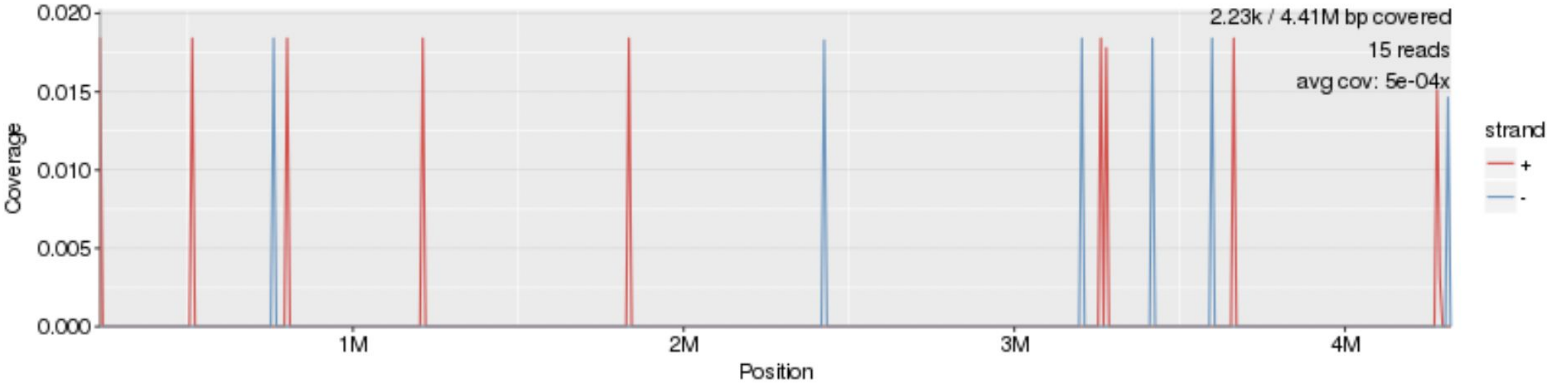
Sequencing error rates varies significantly between technologies, runs, lanes, multiplexes, genomic location as well as substitution types

Types of research questions my group focuses on

Computational microbial forensics:

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

C: Mycobacterium tuberculosis in PT8 (NC_000962.3)

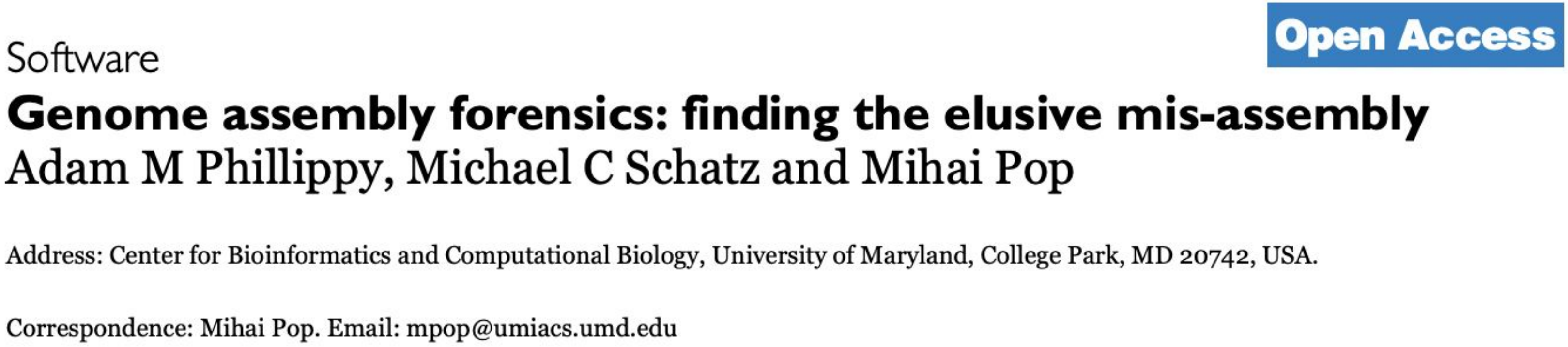


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1568-0

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

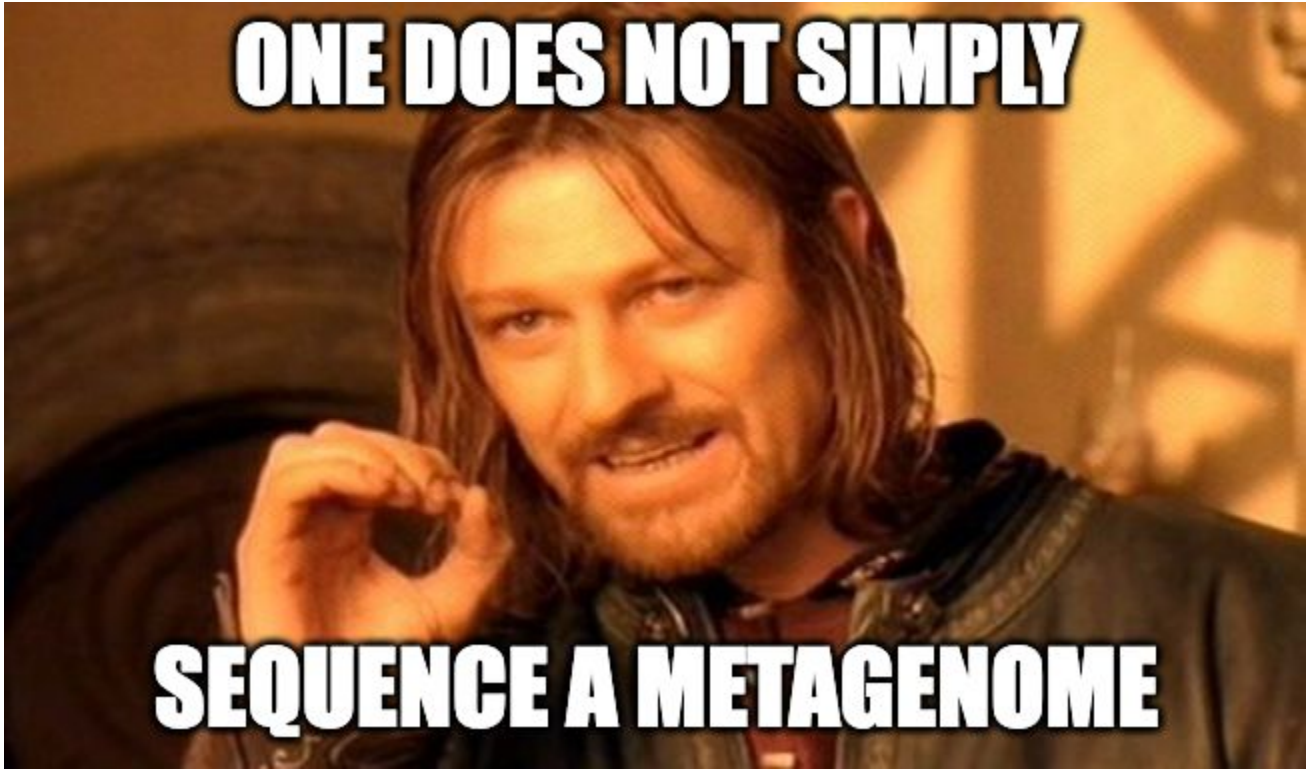
Adam M Phillippy, Michael C Schatz and Mihai Pop

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!

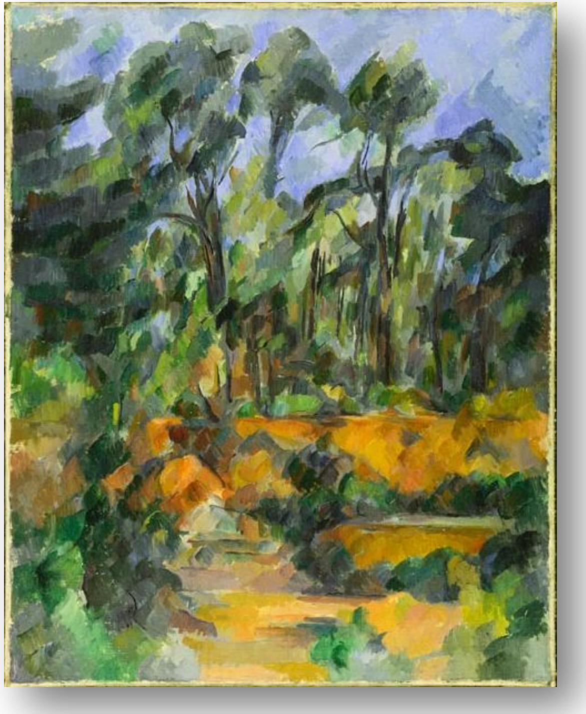

ONE DOES NOT SIMPLY
SEQUENCE A METAGENOME

Sequencing machines turn forests into twigs

Paul Cézanne, circa 1902-1904

Current tools turn twigs into wood chips..

Goal is to turn wood chips back into forest



Paul Cézanne, circa 1902-1904

...while avoiding misassemblies!

*African baobab tree