CompGenomics Wiki and Github

Shashwat Deepali Nagar

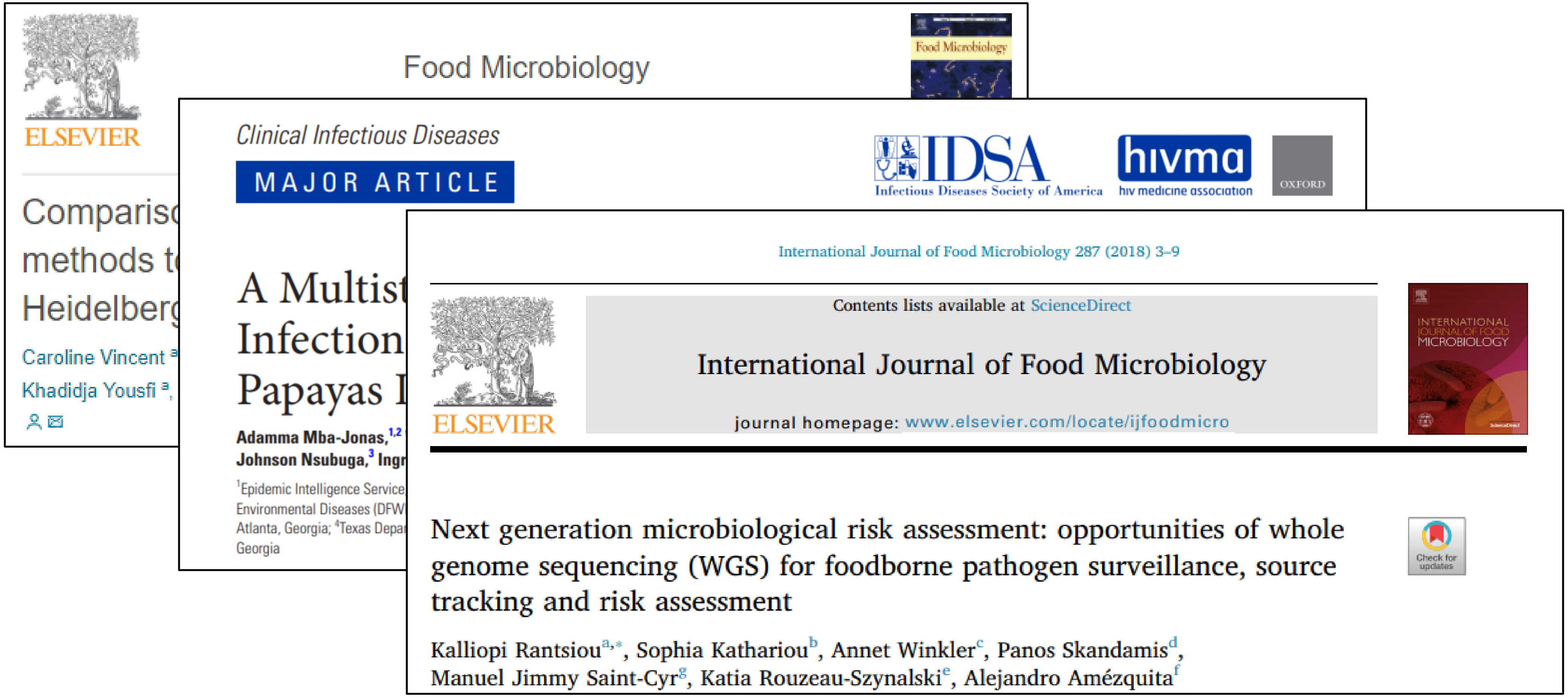
Agenda for today

- Quick intro to Computational Genomics
- Class resources
    - Server
    - Wiki
    - Github
- Team contracts
- Finalizing groups

Computational Genomics @ GATech

- This is the ***thirteenth* year** that this course is being offered
- Students worked with 1 whole genome the first year
- Last year, students works with ~160 genomes
- Bioinformatics is a “big data” field, learning how to analyze large datasets is important

This is a real-world problem, in need of computational solutions



ELSEVIER

Food Microbiology

Compariso
methods t
Heidelberg

Caroline Vincent [a]
Khadidja Yousfi [a],

Clinical Infectious Diseases

MAJOR ARTICLE

A Multist
Infection
Papayas I

Adamma Mba-Jonas,[1,2]
Johnson Nsubuga,[3] Ingr

[1]Epidemic Intelligence Service
Environmental Diseases (DFW
Atlanta, Georgia; [4]Texas Depa
Georgia

International Journal of Food Microbiology 287 (2018) 3–9

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

ELSEVIER

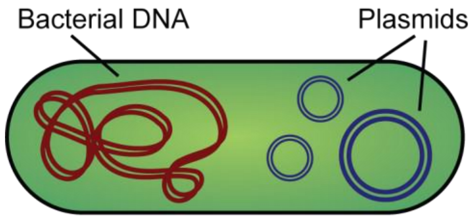
Next generation microbiological risk assessment: opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment

Kalliopi Rantsiou[a,*], Sophia Kathariou[b], Annet Winkler[c], Panos Skandamis[d], Manuel Jimmy Saint-Cyr[g], Katia Rouzeau-Szynalski[e], Alejandro Amézquita[f]

Food Microbiology
IDSA
Infectious Diseases Society of America
hivma
hiv medicine association
OXFORD
INTERNATIONAL JOURNAL OF FOOD MICROBIOLOGY
Check for updates

Very quick overview of the class

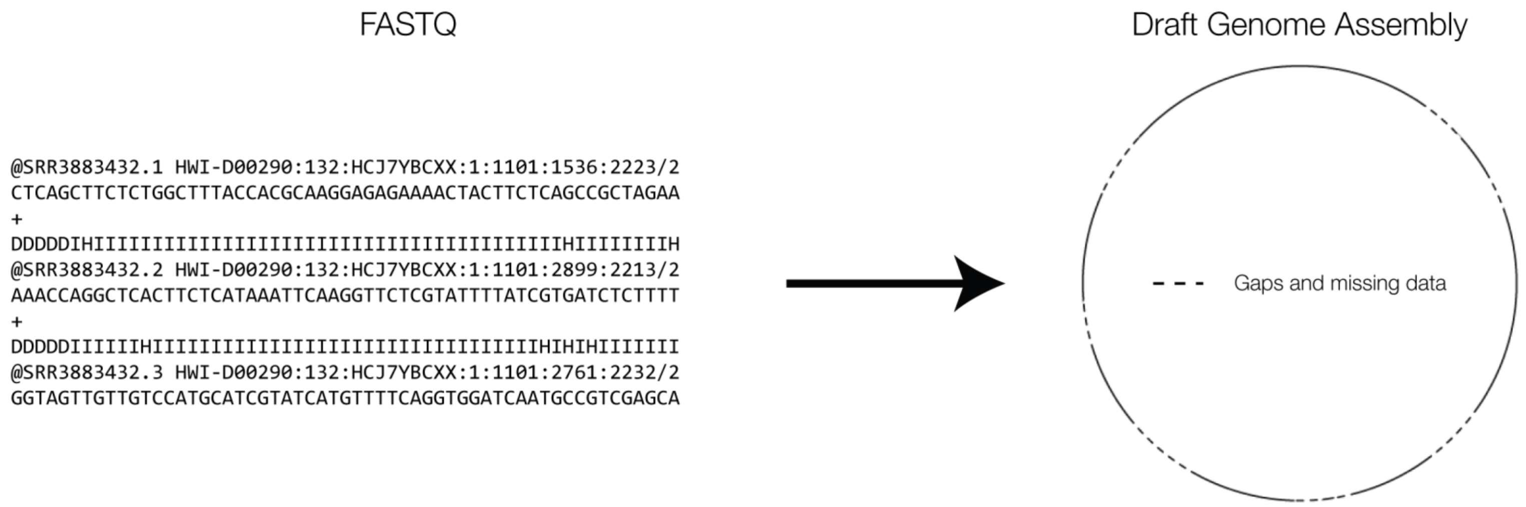
Long time ago in a wet lab far, far away…



https://commons.wikimedia.org/wiki/File:Plasmid_(english).svg

Very quick overview of the class

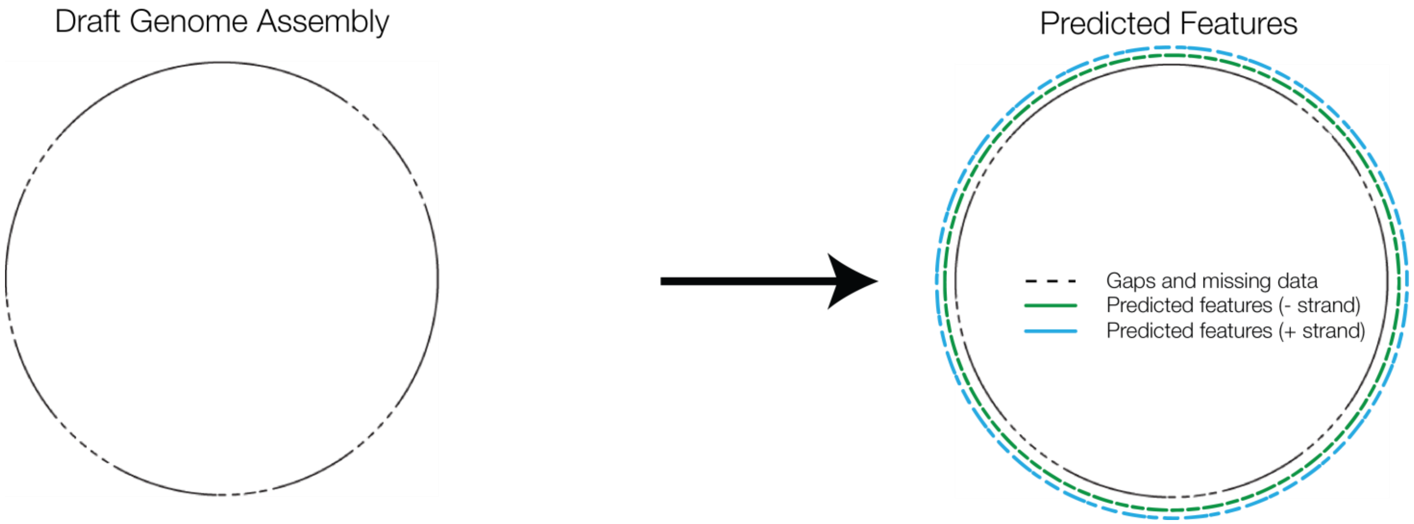
Genome Assembly



```
@SRR3883432.1 HWI-D00290:132:HCJ7YBCXX:1:1101:1536:2223/2
CTCAGCTTCTCTGGCTTTACCACGCAAGGAGAGAAAACTACTTCTCAGCCGCTAGAA
+
DDDDDIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIH
@SRR3883432.2 HWI-D00290:132:HCJ7YBCXX:1:1101:2899:2213/2
AAACCAGGCTCACTTCTCATAAATTCAAGGTTCTCGTATTTTATCGTGATCTCTTTT
+
DDDDDIIIIIIHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIHIHIIIIIII
@SRR3883432.3 HWI-D00290:132:HCJ7YBCXX:1:1101:2761:2232/2
GGTAGTTGTTGTCCATGCATCGTATCATGTTTTCAGGTGGATCAATGCCGTCGAGCA
```

Very quick overview of the class

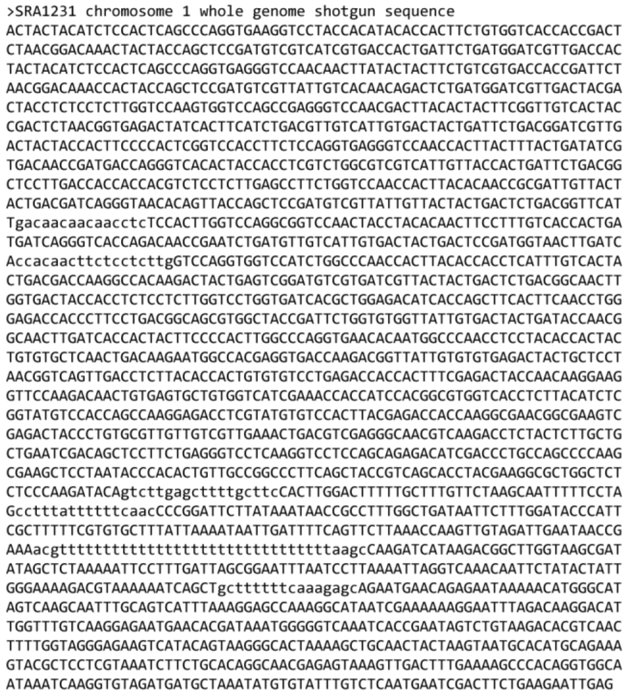
Gene Prediction



Very quick overview of the class

Functional Annotation

Imagine a genome is like a book describing your organism

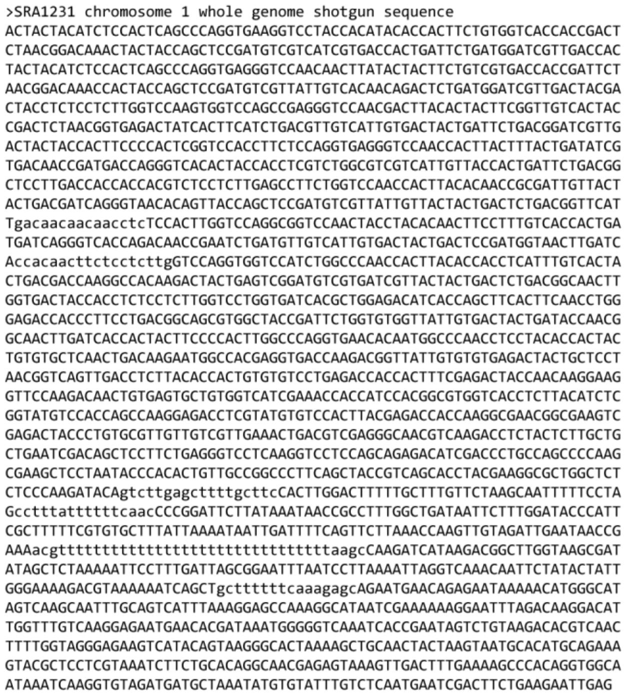


```
>SRA1231 chromosome 1 whole genome shotgun sequence
ACTACTACATCTCCACTCAGCCCAGGTGAAGGTCCTACCACATACACCACTTCTGTGGTCACCACCGACT
CTAACGGACAAACTACTACCAGCTCCGATGTCGTCATCGTGACCACTGATTCTGATGGATCGTTGACCAC
TACTACATCTCCACTCAGCCCAGGTGAGGGTCCAACAACTTATACTACTTCTGTCGTGACCACCGATTCT
AACGGACAAACCACTACCAGCTCCGATGTCGTTATTGTCACAACAGACTCTGATGGATCGTTGACTACGA
CTACCTCTCCTCTTGGTCCAAGTGGTCCAGCCGAGGGTCCAACGACTTACACTACTTCGGTTGTCACTAC
CGACTCTAACGGTGAGACTATCACTTCATCTGACGTTGTCATTGTGACTACTGATTCTGACGGATCGTTG
ACTACTACCACTTCCCCACTCGGTCCACCTTCTCCAGGTGAGGGTCCAACCACTTACTTTACTGATATCG
TGACAACCGATGACCAGGGTCACACTACCACCTCGTCTGGCGTCGTCATTGTTACCACTGATTCTGACGG
CTCCTTGACCACCACCACGTCTCCTCTTGAGCCTTCTGGTCCAACCACTTACACAACCGCGATTGTTACT
ACTGACGATCAGGGTAACACAGTTACCAGCTCCGATGTCGTTATTGTTACTACTGACTCTGACGGTTCAT
TgacaacaacaacctcTCCACTTGGTCCAGGCGGTCCAACTACCTACACAACTTCCTTTGTCACCACTGA
TGATCAGGGTCACCAGACAACCGAATCTGATGTTGTCATTGTGACTACTGACTCCGATGGTAACTTGATC
AccacaacttctcctcttgGTCCAGGTGGTCCATCTGGCCCAACCACTTACACCACCTCATTTGTCACTA
CTGACGACCAAGGCCACAAGACTACTGAGTCGGATGTCGTGATCGTTACTACTGACTCTGACGGCAACTT
GGTGACTACCACCTCTCCTCTTGGTCCTGGTGATCACGCTGGAGACATCACCAGCTTCACTTCAACCTGG
GAGACCACCCTTCCTGACGGCAGCGTGGCTACCGATTCTGGTGTGGTTATTGTGACTACTGATACCAACG
GCAACTTGATCACCACTACTTCCCCACTTGGCCCAGGTGAACACAATGGCCCAACCTCCTACACCACTAC
TGTGTGCTCAACTGACAAGAATGGCCACGAGGTGACCAAGACGGTTATTGTGTGTGAGACTACTGCTCCT
AACGGTCAGTTGACCTCTTACACCACTGTGTGTCCTGAGACCACCACTTTCGAGACTACCAACAAGGAAG
GTTCCAAGACAACTGTGAGTGCTGTGGTCATCGAAACCACCATCCACGGCGTGGTCACCTCTTACATCTC
GGTATGTCCACCAGCCAAGGAGACCTCGTATGTGTCCACTTACGAGACCACCAAGGCGAACGGCGAAGTC
GAGACTACCCTGTGCGTTGTTGTCGTTGAAACTGACGTCGAGGGCAACGTCAAGACCTCTACTCTTGCTG
CTGAATCGACAGCTCCTTCTGAGGGTCCTCAAGGTCCTCCAGCAGAGACATCGACCCTGCCAGCCCCAAG
CGAAGCTCCTAATACCCACACTGTTGCCGGCCCTTCAGCTACCGTCAGCACCTACGAAGGCGCTGGCTCT
CTCCCAAGATACAgtcttgagcttttgcttcCACTTGGACTTTTTGCTTTGTTCTAAGCAATTTTTCCTA
GcctttatttttttcaacCCCGGATTCTTATAAATAACCGCCTTTGGCTGATAATTCTTTGGATACCCATT
CGCTTTTTCGTGTGCTTTATTAAAATAATTGATTTTCAGTTCTTAAACCAAGTTGTAGATTGAATAACCG
AAAacgtttttttttttttttttttttttttttttaagcCAAGATCATAAGACGGCTTGGTAAGCGAT
ATAGCTCTAAAAATTCCTTTGATTAGCGGAATTTAATCCTTAAAATTAGGTCAAACAATTCTATACTATT
GGGAAAAGACGTAAAAAATCAGCTgctttttttcaaagagcAGAATGAACAGAGAATAAAAACATGGGCAT
AGTCAAGCAATTTGCAGTCATTTAAAGGAGCCAAAGGCATAATCGAAAAAAGGAATTTAGACAAGGACAT
TGGTTTGTCAAGGAGAATGAACACGATAAATGGGGGTCAAATCACCGAATAGTCTGTAAGACACGTCAAC
TTTTGGTAGGGAGAAGTCATACAGTAAGGGCACTAAAAGCTGCAACTACTAAGTAATGCACATGCAGAAA
GTACGCTCCTCGTAAATCTTCTGCACAGGCAACGAGAGTAAAGTTGACTTTGAAAAGCCCACAGGTGGCA
ATAAATCAAGGTGTAGATGATGCTAAATATGTGTATTTGTCTCAATGAATCGACTTCTGAAGAATTGAG
```

This is the genome of Klebsiella pneunomiae strain XYZ, isolated from a patient in the Emory Heathcare. This genome contains many interesting antimicrobial resistance genes, as well other genes for invasion, host evasion, and adherance. Antimicrobial compounds such as beta-lactams, penicillins, and even carbepenems have no effect on me.

Very quick overview of the class

Some words, or **genomic features**, in the book have special meaning



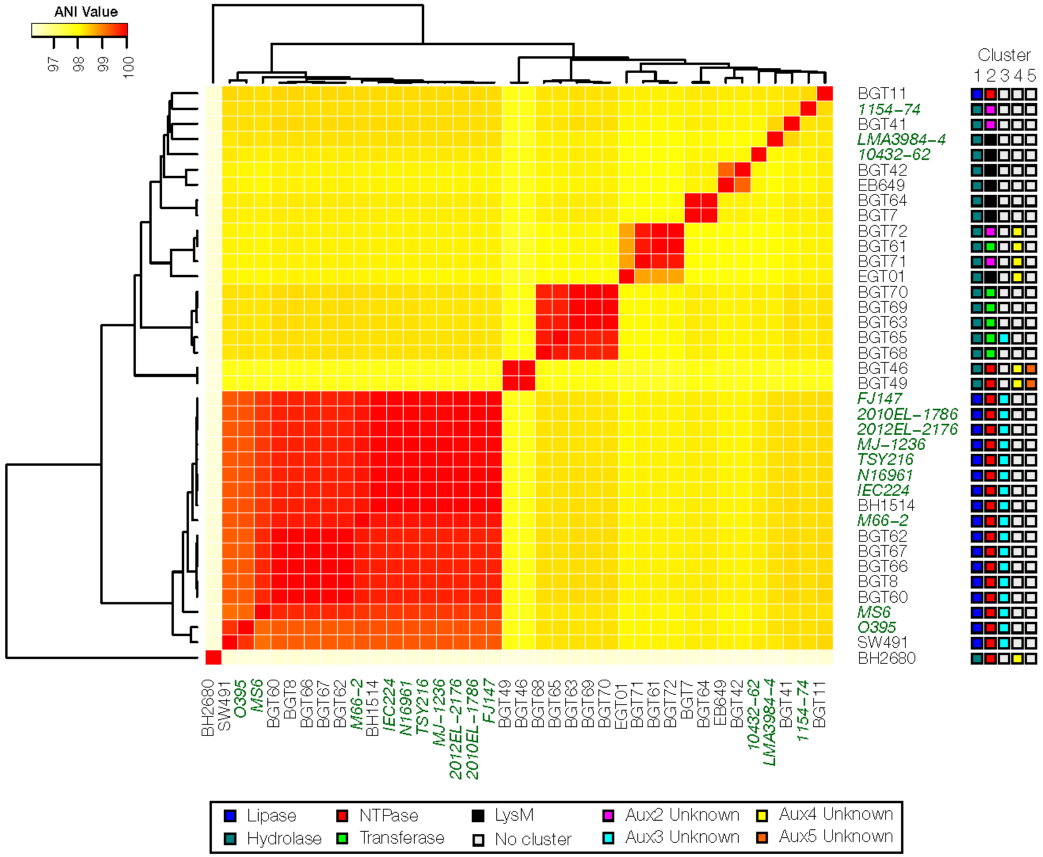
```
>SRA1231 chromosome 1 whole genome shotgun sequence
ACTACTACATCTCCACTCAGCCCAGGTGAAGGTCCTACCACATACACCACTTCTGTGGTCACCACCGACT
CTAACGGACAAACTACTACCAGCTCCGATGTCGTCATCGTGACCACTGATTCTGATGGATCGTTGACCAC
TACTACATCTCCACTCAGCCCAGGTGAGGGTCCAACAACTTATACTACTTCTGTCGTGACCACCGATTCT
AACGGACAAACCACTACCAGCTCCGATGTCGTTATTGTCACAACAGACTCTGATGGATCGTTGACTACGA
CTACCTCTCCTCTTGGTCCAAGTGGTCCAGCCGAGGGTCCAACGACTTACACTACTTCGGTTGTCACTAC
CGACTCTAACGGTGAGACTATCACTTCATCTGACGTTGTCATTGTGACTACTGATTCTGACGGATCGTTG
ACTACTACCACTTCCCCACTCGGTCCACCTTCTCCAGGTGAGGGTCCAACCACTTACTTTACTGATATCG
TGACAACCGATGACCAGGGTCACACTACCACCTCGTCTGGCGTCGTCATTGTTACCACTGATTCTGACGG
CTCCTTGACCACCACCACGTCTCCTCTTGAGCCTTCTGGTCCAACCACTTACACAACCGCGATTGTTACT
ACTGACGATCAGGGTAACACAGTTACCAGCTCCGATGTCGTTATTGTTACTACTGACTCTGACGGTTCAT
TgacaacaacaacctcTCCACTTGGTCCAGGCGGTCCAACTACCTACACAACTTCCTTTGTCACCACTGA
TGATCAGGGTCACCAGACAACCGAATCTGATGTTGTCATTGTGACTACTGACTCCGATGGTAACTTGATC
AccacaacttctcctcttgGTCCAGGTGGTCCATCTGGCCCAACCACTTACACCACCTCATTTGTCACTA
CTGACGACCAAGGCCACAAGACTACTGAGTCGGATGTCGTGATCGTTACTACTGACTCTGACGGCAACTT
GGTGACTACCACCTCTCCTCTTGGTCCTGGTGATCACGCTGGAGACATCACCAGCTTCACTTCAACCTGG
GAGACCACCCTTCCTGACGGCAGCGTGGCTACCGATTCTGGTGTGGTTATTGTGACTACTGATACCAACG
GCAACTTGATCACCACTACTTCCCCACTTGGCCCAGGTGAACACAATGGCCCAACCTCCTACACCACTAC
TGTGTGCTCAACTGACAAGAATGGCCACGAGGTGACCAAGACGGTTATTGTGTGTGAGACTACTGCTCCT
AACGGTCAGTTGACCTCTTACACCACTGTGTGTCCTGAGACCACCACTTTCGAGACTACCAACAAGGAAG
GTTCCAAGACAACTGTGAGTGCTGTGGTCATCGAAACCACCATCCACGGCGTGGTCACCTCTTACATCTC
GGTATGTCCACCAGCCAAGGAGACCTCGTATGTGTCCACTTACGAGACCACCAAGGCGAACGGCGAAGTC
GAGACTACCCTGTGCGTTGTTGTCGTTGAAACTGACGTCGAGGGCAACGTCAAGACCTCTACTCTTGCTG
CTGAATCGACAGCTCCTTCTGAGGGTCCTCAAGGTCCTCCAGCAGAGACATCGACCCTGCCAGCCCCAAG
CGAAGCTCCTAATACCCACACTGTTGCCGGCCCTTCAGCTACCGTCAGCACCTACGAAGGCGCTGGCTCT
CTCCCAAGATACAgtcttgagcttttgcttcCACTTGGACTTTTTGCTTTGTTCTAAGCAATTTTTCCTA
GcctttattttttcaacCCCGGATTCTTATAAATAACCGCCTTTGGCTGATAATTCTTTGGATACCCATT
CGCTTTTTCGTGTGCTTTATTAAAATAATTGATTTTCAGTTCTTAAACCAAGTTGTAGATTGAATAACCG
AAAacgtttttttttttttttttttttttttttttaagcCAAGATCATAAGACGGCTTGGTAAGCGAT
ATAGCTCTAAAAATTCCTTTGATTAGCGGAATTTAATCCTTAAAATTAGGTCAAACAATTCTATACTATT
GGGAAAAGACGTAAAAAATCAGCTgcttttttcaaagagcAGAATGAACAGAGAATAAAAACATGGGCAT
AGTCAAGCAATTTGCAGTCATTTAAAGGAGCCAAAGGCATAATCGAAAAAAGGAATTTAGACAAGGACAT
TGGTTTGTCAAGGAGAATGAACACGATAAATGGGGGTCAAATCACCGAATAGTCTGTAAGACACGTCAAC
TTTTGGTAGGGAGAAGTCATACAGTAAGGGCACTAAAAGCTGCAACTACTAAGTAATGCACATGCAGAAA
GTACGCTCCTCGTAAATCTTCTGCACAGGCAACGAGAGTAAAGTTGACTTTGAAAAGCCCACAGGTGGCA
ATAAATCAAGGTGTAGATGATGCTAAATATGTGTATTTGTCTCAATGAATCGACTTCTGAAGAATTGAG
```

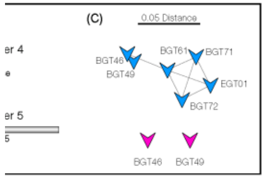
→

This is the genome of Klebsiella pneunomiae strain XYZ, isolated from a patient in the Emory Heathcare. This genome contains many interesting antimicrobial resistance genes, as well other genes for invasion, host evasion, and adherance. Antimicrobial compounds such as beta-lactams, penicillins, and even carbepenems have no effect on me.

Very quick overview of the class

Comparative genomics



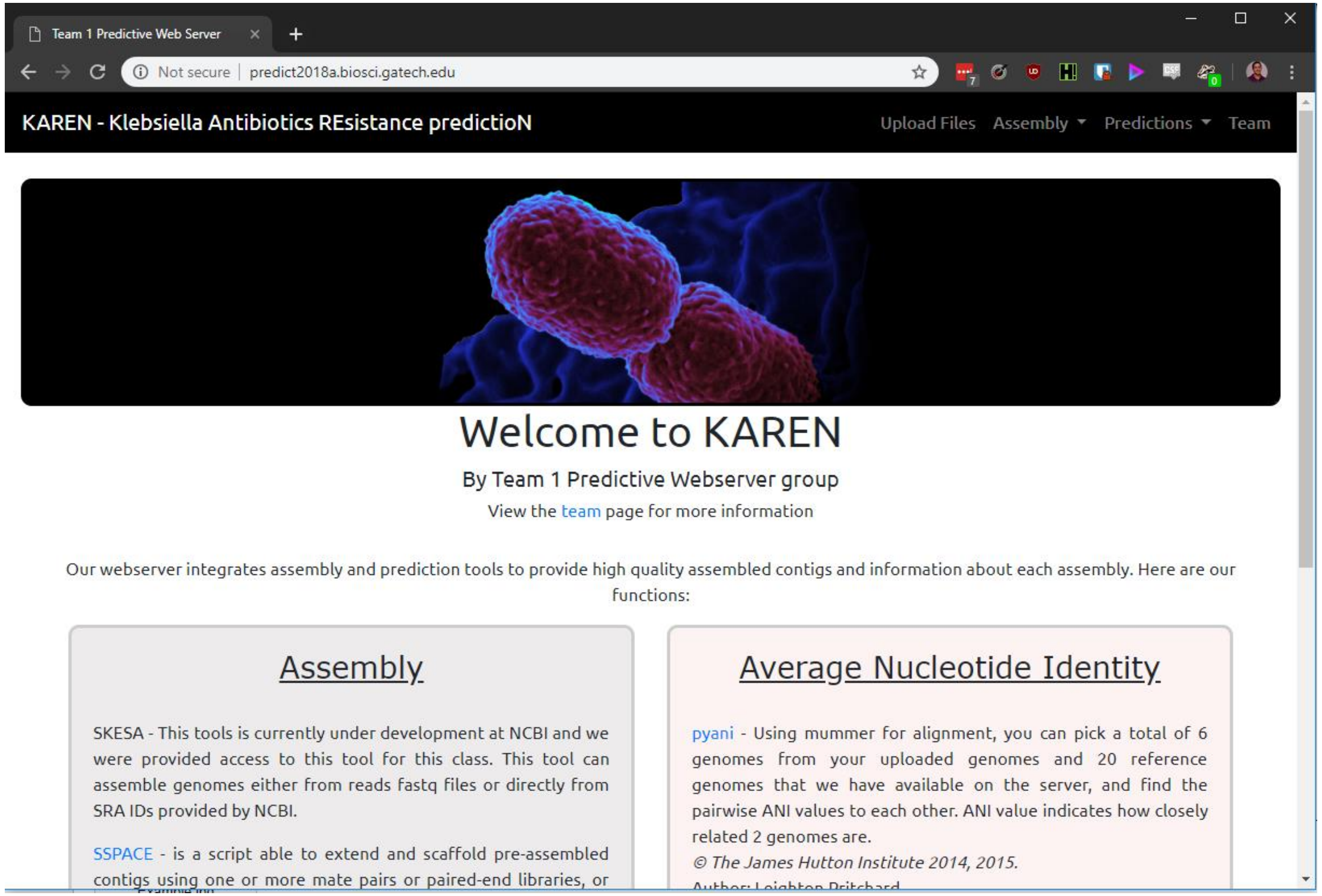


Crisan et al. (2019). Analysis of *Vibrio cholerae* genomes using a novel bioinformatic tool identifies new, active Type VI Secretion System gene clusters. *Submitted*

Very quick overview of the class

In the end you will produce something like this →

http://predict2018a.biosci.gatech.edu/


Team 1 Predictive Web Server
Not secure | predict2018a.biosci.gatech.edu
KAREN - Klebsiella Antibiotics REsistance predictioN
Upload Files
Assembly
Predictions
Team

Welcome to KAREN

By Team 1 Predictive Webserver group

View the team page for more information

Our webserver integrates assembly and prediction tools to provide high quality assembled contigs and information about each assembly. Here are our functions:

Assembly

SKESA - This tools is currently under development at NCBI and we were provided access to this tool for this class. This tool can assemble genomes either from reads fastq files or directly from SRA IDs provided by NCBI.

SSPACE - is a script able to extend and scaffold pre-assembled contigs using one or more mate pairs or paired-end libraries, or

Average Nucleotide Identity

pyani - Using mummer for alignment, you can pick a total of 6 genomes from your uploaded genomes and 20 reference genomes that we have available on the server, and find the pairwise ANI values to each other. ANI value indicates how closely related 2 genomes are.

© The James Hutton Institute 2014, 2015.

Author: Leighton Pritchard

Some Tips

- Start Early – Most of these processes will take time and decent computational resources. All-nighters aren't all that fun.
  - Server people – you should start even earlier. Setting up the server can be extremely difficult depending on how much you've read and understood about the tool (and your skill level).

Some Tips

- Document continuously. It is tempting to leave documentation of your work and results to the end – don't! Document your code, analysis, and results as you go. This results in better documentation and won't leave you awake at 3am the day before your group final presentation
    - Wiki-writing is just as import as analysis documentation.
    - I'll talk about what kind of documentation we are looking for next week

Some Tips

- Stick to an accession convention across the groups when processing genes/proteins – this will avoid confusion, unnecessary mapping, and will make sure you do not wind up with duplicates

- Have proper channels for exchange of data – again, this avoids confusion

Some Tips

- **Everyone likes the new stuff** – Try to search for what is recent in the literature and attempt to use it along with the classical tools

- **Understand what you are doing** – don’t blindly follow what other classes have done, have a proper understanding and basis of what you are doing

Final considerations

- Biology is messy, there are lots of exceptions to the rules – follow the data, even when things look a little weird… just don't go down the rabbit hole

- You will need to attempt many analyses, *many* of them won't be informative

- Don't get discouraged if something doesn't work, iterate and move on

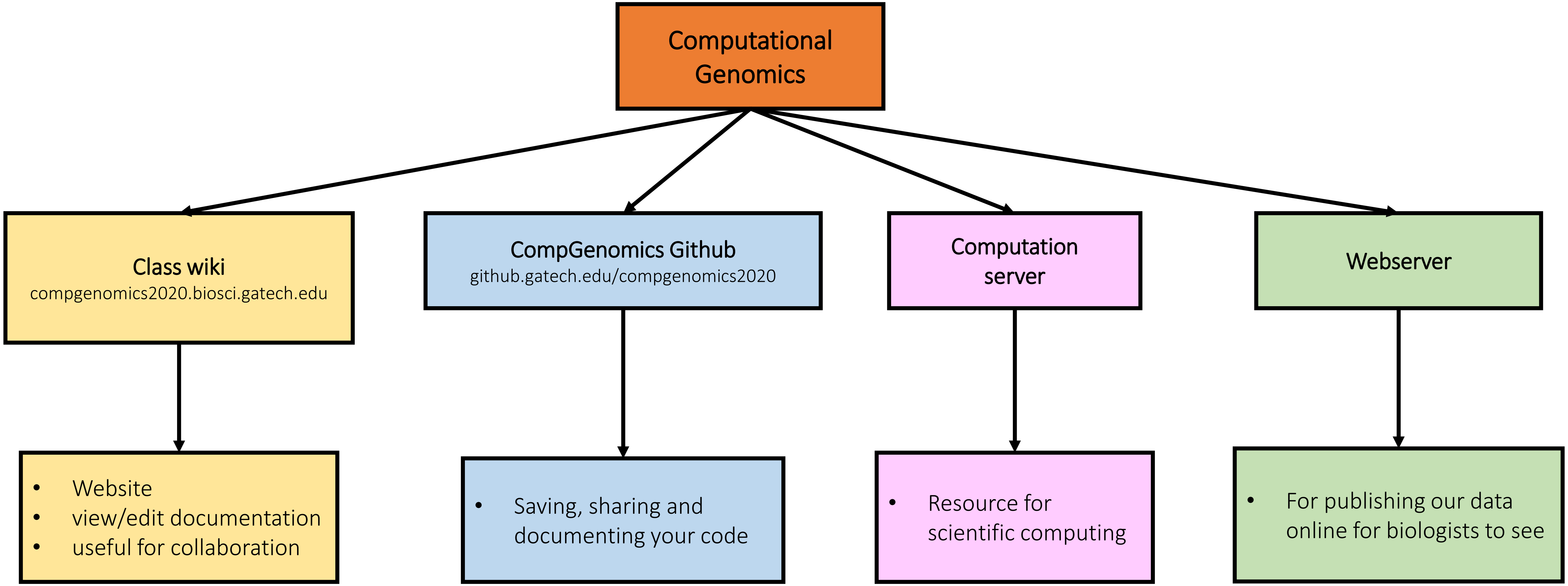
Final considerations

- If you get stuck, you can reach out to me and
  - You want to talk through an analysis idea
  - You're not sure about how to interpret a result

- Don't wait for final results
  - You don't need the final results of the group before you to begin working
  - You can refine your knowledge over time with the help of groups before you

Class resources

CompGenomics Resources



Computational Genomics

Class wiki
compgenomics2020.biosci.gatech.edu

- Website
- view/edit documentation
- useful for collaboration

CompGenomics Github
github.gatech.edu/compgenomics2020

- Saving, sharing and documenting your code

Computation server

- Resource for scientific computing

Webserver

- For publishing our data online for biologists to see

Shared computational server

- The server will be made available for you at the same time your data is released
- You will not have root access on the server. Unlike your local VMs, you cannot `sudo` your way out of trouble
- There are enough resources for you to accomplish your tasks – if you are judicious

  `16 cores | 64GB RAM | 1.2 TB disk space`

Shared computational server

Ground rules for shared computational resources:

1. Be a good neighbor – Do not use all the cores, all the RAM, or all the disk
2. Emphasis on *shared* – Some tools you will use required tens of GB of data files (e.g. protein databases), share these between teams/groups
3. Practice good hygiene – don’t create clutter or unnecessary copies of data; hard- and *symlinks* are your friends
4. Do not (ever) run `‘rm -rf *’`, even if you think it’s safe

If you accidentally delete all your group’s data, you will be responsible for recreating it before the due dates listed on the syllabus. We do not keep backups.

Course Wiki page

https://compgenomics2020.biosci.gatech.edu/

- The course wiki page is a public display of your work

- Lectures, readings, and all other course material can be found here

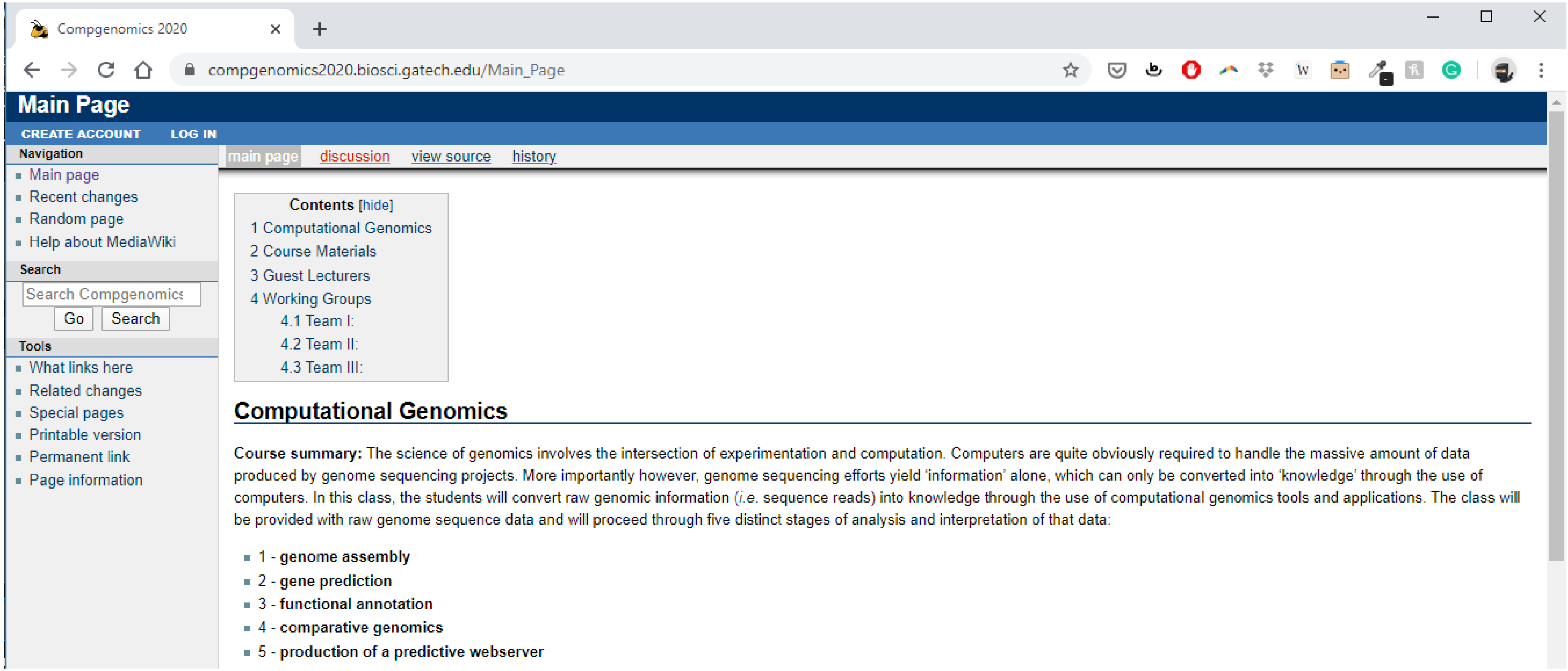
- We ***will not*** be using Canvas, other than to send email announcements to the class and submit exercises

Don’t reinvent the wheel

Or, perhaps, don’t reinvent it every time

- Scientific research is building and expanding on existing knowledge
- For some of the tasks you’ll be given, there may only be one “great” way to do them
  - We won’t penalize doing the same thing as previous years ***if*** you provide your own, data-driven reasoning

The Compgenomics Wiki



Main Page

Contents [hide]

Computational Genomics

Course summary: The science of genomics involves the intersection of experimentation and computation. Computers are quite obviously required to handle the massive amount of data produced by genome sequencing projects. More importantly however, genome sequencing efforts yield 'information' alone, which can only be converted into 'knowledge' through the use of computers. In this class, the students will convert raw genomic information (*i.e.* sequence reads) into knowledge through the use of computational genomics tools and applications. The class will be provided with raw genome sequence data and will proceed through five distinct stages of analysis and interpretation of that data:

- 1 - **genome assembly**
- 2 - **gene prediction**
- 3 - **functional annotation**
- 4 - **comparative genomics**
- 5 - **production of a predictive webserver**

Wiki accounts

- All students are required to create a wiki account and contribute to the wiki

- In order to register you must use your “user@gatech.edu” email address (e.g. snagar9@gatech.edu)

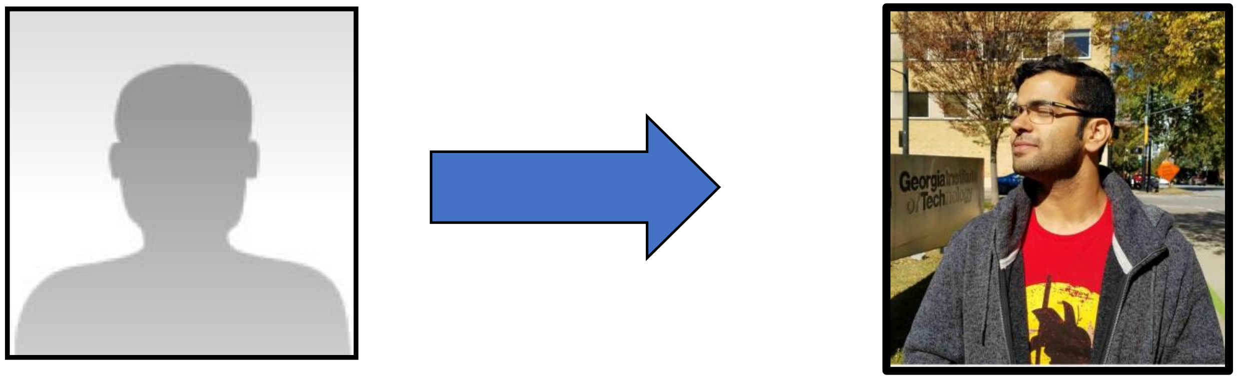
- You have until Tuesday, January 14th to register an account
  - Timely registration of your wiki account is worth 10 points

Wiki Accounts

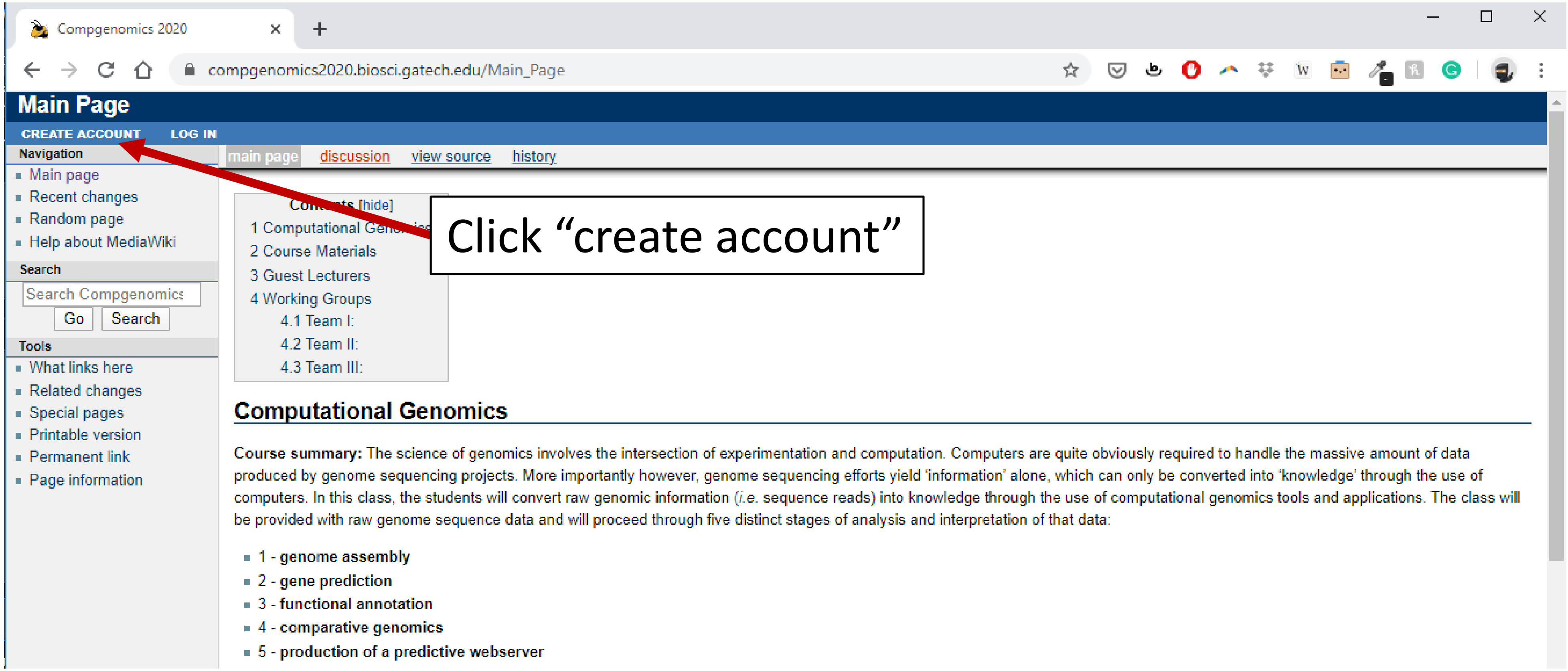
- The wiki is public and where your collaborators, other scientists, and maybe future employers will see your work

- Add yourself on the Profiles page. Be sure to include a picture.

https://compgenomics2020.biosci.gatech.edu/Profiles



Account registration



Click “create account”

Main Page

Contents [hide]

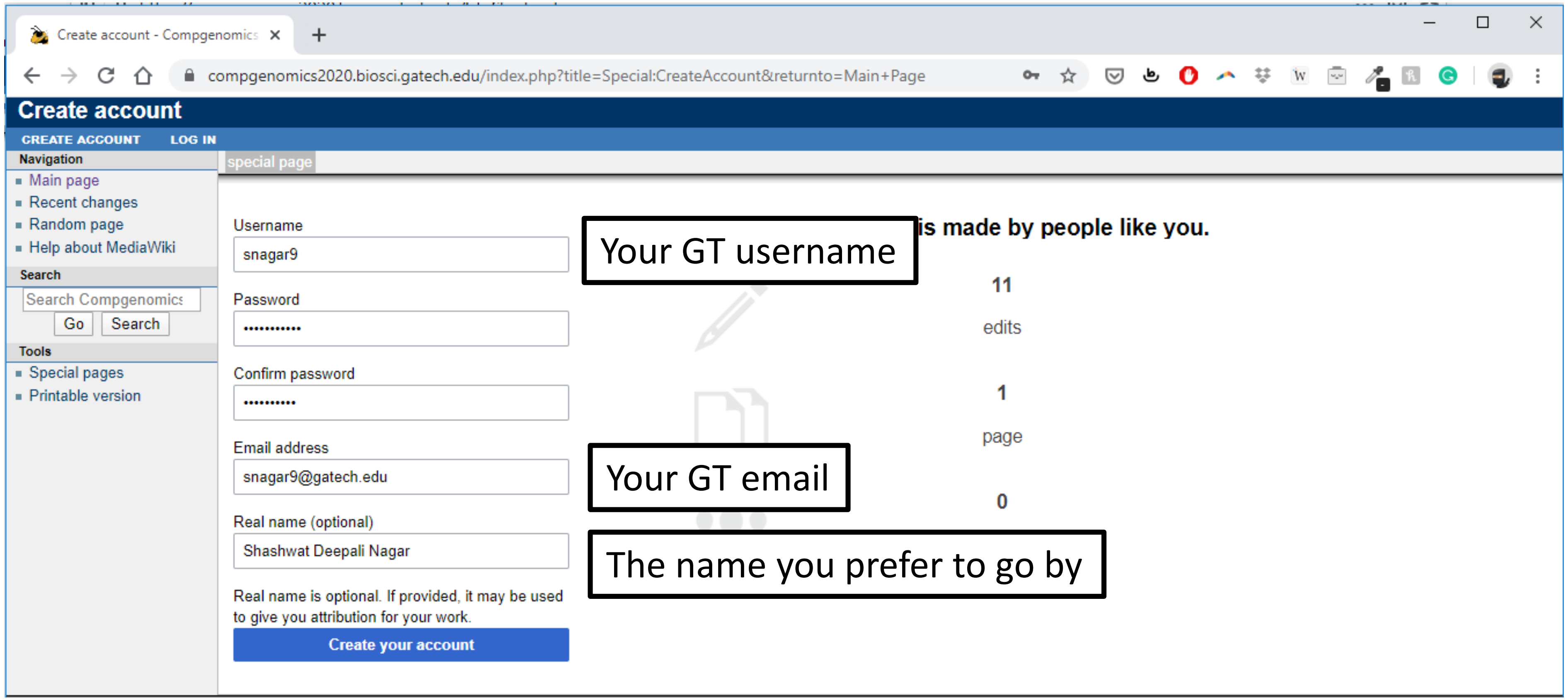
1 Computational Genomics
2 Course Materials
3 Guest Lecturers
4 Working Groups
4.1 Team I:
4.2 Team II:
4.3 Team III:

Computational Genomics

Course summary: The science of genomics involves the intersection of experimentation and computation. Computers are quite obviously required to handle the massive amount of data produced by genome sequencing projects. More importantly however, genome sequencing efforts yield ‘information’ alone, which can only be converted into ‘knowledge’ through the use of computers. In this class, the students will convert raw genomic information (*i.e.* sequence reads) into knowledge through the use of computational genomics tools and applications. The class will be provided with raw genome sequence data and will proceed through five distinct stages of analysis and interpretation of that data:

- 1 - **genome assembly**
- 2 - **gene prediction**
- 3 - **functional annotation**
- 4 - **comparative genomics**
- 5 - **production of a predictive webserver**

Account registration



Create account - Compgenomics

compgenomics2020.biosci.gatech.edu/index.php?title=Special:CreateAccount&returnto=Main+Page

Create account

Username: snagar9

Password: ••••••••••

Confirm password: •••••••••

Email address: snagar9@gatech.edu

Real name (optional): Shashwat Deepali Nagar

Real name is optional. If provided, it may be used to give you attribution for your work.

Create your account

is made by people like you.

11 edits

1 page

0

Your GT username

Your GT email

The name you prefer to go by

Class Github

https://github.gatech.edu/compgenomics2020/

- I'll talk more about what `git` and source code management are next week on Tuesday
- This site will contain *all* your code, raw analysis, and other products of your work
- During the semester, these data will be private, afterwards they will be open-sourced (on Github.com)

Why use git?

- Git and other source code management (SCM) tools are very useful for storing and tracking the files and scripts you create when working

- It provides an audit trail of ***who*** made ***what*** changes, ***when***

- Use and proficiency with SCM is quickly becoming a required skill in both industry and the public sector (i.e. CDC)

Why use git?

- Git keeps everyone accountable
- Keep collaborative files in one place
- Provides a platform to organize, document, and share your work

Github student pack

https://education.github.com/pack

- Get free and discounted access to AWS (Cloud computing resources), DigitalOcean (Virtual servers), and other useful software tools

- Great for learning and refining some of the skills you'll learn this semester

- Particularly useful for Web Server folks to prototype their work

Team Contract

Team contracts

- Team contracts will help you set expectations about frequency and quality of work, along with ground rules for transgressions.

- Things you might want to consider
    - How many times do you want to meet each week?

    - Channels of communication

    - Roles outside of bioinformatics analysis: Liaising, Documenting, Presenting etc.
        - You can either decide that specific people will be responsible for tasks like this or that everyone should document things as they perform specific analyses.

    - Grade distribution
        - Are you all okay with everyone in the group getting the same grade?
        - Would you want to fill out peer review forms at the end?
            - What are the penalties for failure to participate? Will the non-participating student (as determined by peers) accept a lower grade compared to their peers?

Group assignments