GVA 2022 Review

Attempts to add perspectives and additional resources.

Reminder of goals and consider how well they were met

- Teaching goals:
  - Teach the fundamentals of NGS variant analysis.
    - The wiki page
  - Provide context and exposure multiple types of data.
    - SE, PE, MP sequencing
    - Virus, bacteria, plasmid, human in different tutorials
  - Use example commands to familiarize you with variety of programs.
    - The wiki page
  - Provide resources to enable you to do analysis you haven't thought of yet.
    - The wiki page

Stages of NGS analysis

1	2	3	4
Biological Question	Design & Conduct Experiment	Prepare NGS Library & Sequence	Sequencing Analysis

Typical Stages of Variant Analysis

1	2	3	4
Read Quality Control	Map Reads	Identify Variants	Visualize Variants

#1 most common question I get asked

- How much sequencing do I need to do?
  - Most applications 30-50 fold coverage, higher for bacteria/small organisms because they smaller and cheaper.
- How do I change reads or lanes into coverage?

$$\text{Coverage} = \frac{(\text{Read Length}) \times (\text{Sequencing Type}) \times (\text{Number of Reads})}{\text{Size of Genome}}$$

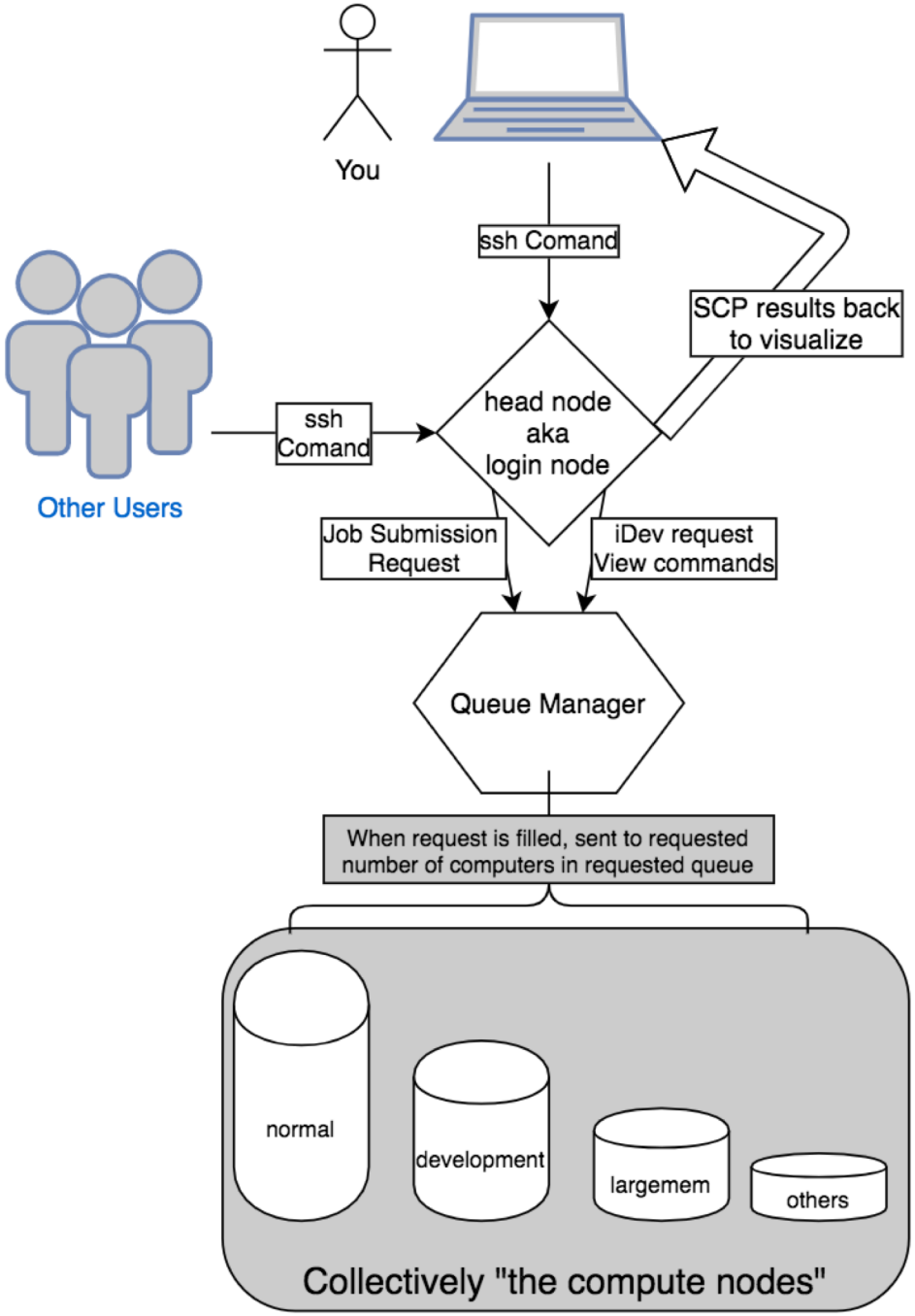
$$30 = \frac{150 \quad \times 2 \text{ (if Pair end or 1 if single)} \quad \times (\text{Number of Reads})}{\text{Size of Genome}}$$

- Min number of reads = ~10% of the genome length
  - If PE 150bp run.
- Max number of reads = ~30% of the genome length

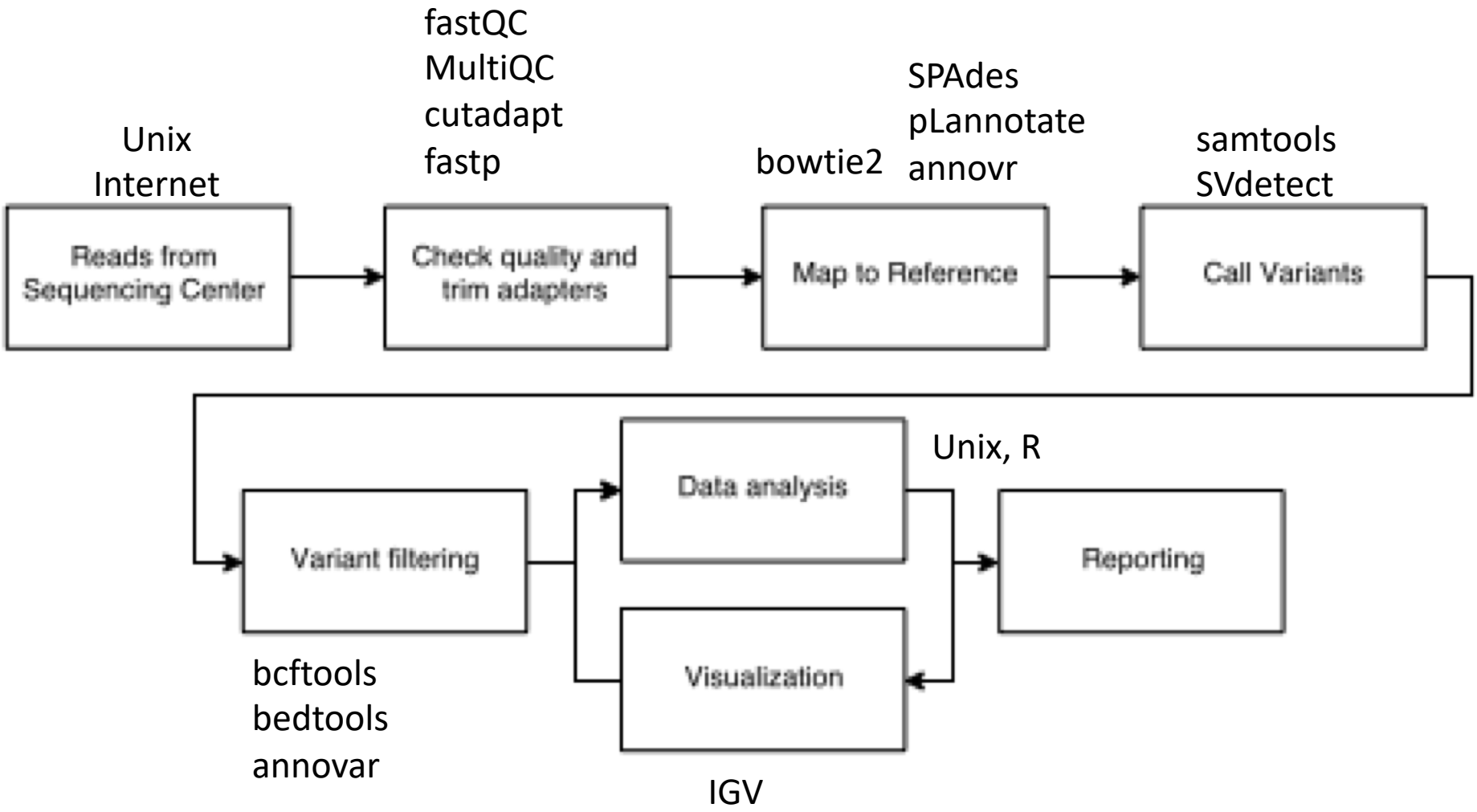
Most common questions about class

1. Do you keep access to the wiki page?
    1. Yes.
2. Can you keep working on the tutorials?
    1. Yes. You will stay on the allocation through at least the end of next month.
3. How do I get my own access to TACC?
    1. Tutorial has info
4. What do I do if I don't or can't get TACC access?
    1. Galaxy https://usegalaxy.org/
    2. AWS https://aws.amazon.com/health/genomics/

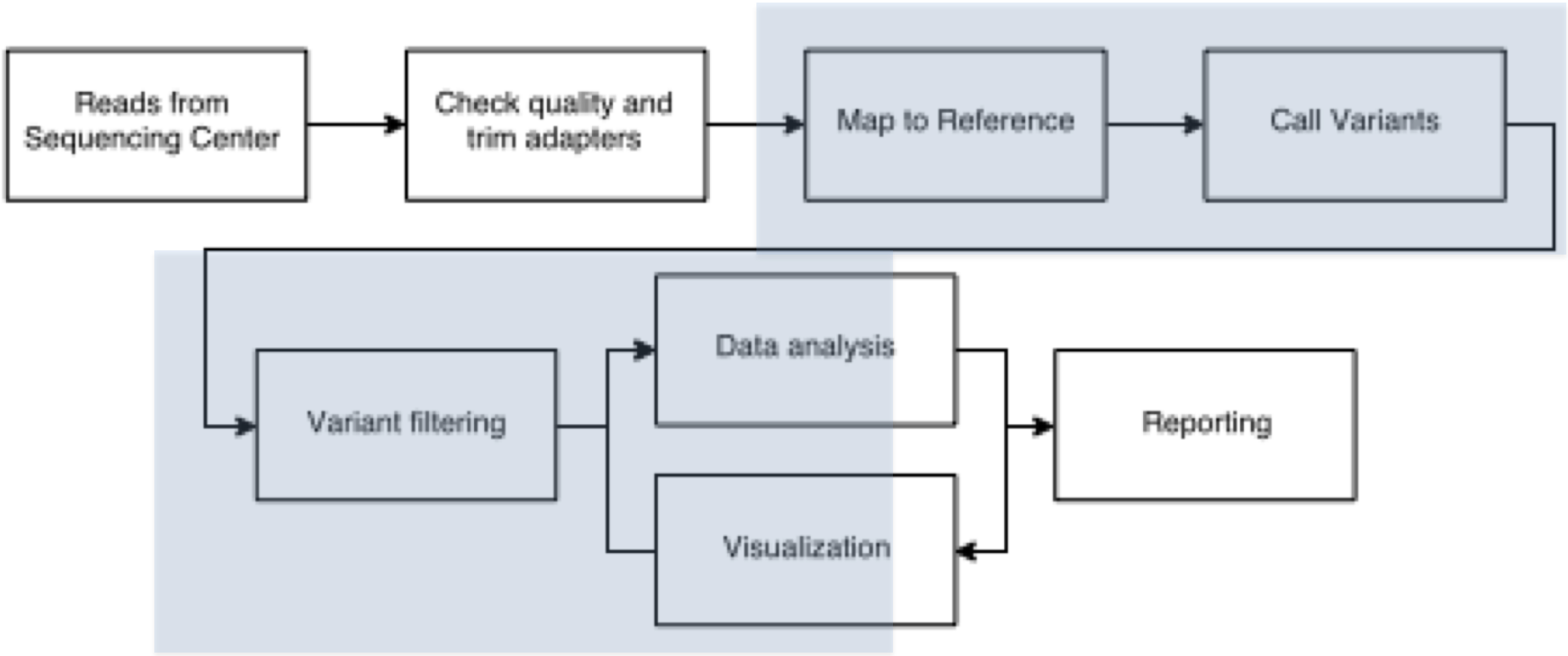
TACC Organization



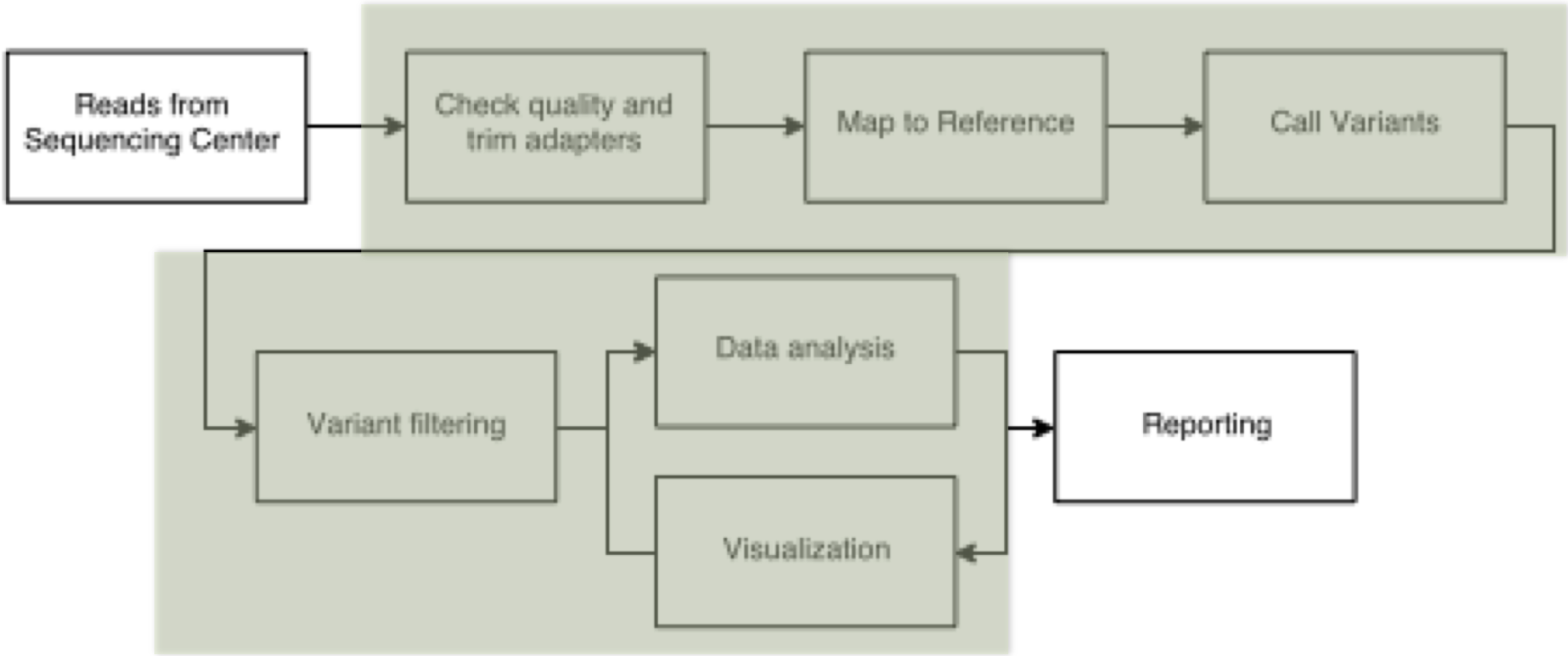
Steps for GVA



microbial all-in-one: breseq



eukaryotic all-in-one: GATK



What to do with all these conda environments?

- Minimal environments:
  - ~Always have correct env active, complicated installs, difficult consistent version control.
- Maximum environments:
  - Easy installs, must activate diff. env at each step
- Project based:
  - Potential to have same program (with diff versions) in multiple env. Maybe good and bad.
- Analysis step:
  - 4 stages variant analysis, “genome assembly”, “population comparison”, “individuals”, “erorr reduction”, etc

Further Resources (online)

- Course wiki: https://wikis.utexas.edu/display/bioiteam/Genome+Variant+Analysis+Course+2022
- Coursera: Genomic data science : https://www.coursera.org/specializations/genomic-data-science
- edX: Python,R: https://www.edx.org/course/subject/computer-science
- Course instructor. You have my email.

What's next

- Today, keep working on tutorials
- Talk to me about what you don't understand about what we have done or why something was important or how it fits together.
- Keep eye out for email from me and from Nicole to review your experience, I really appreciate feedback, it's the only way to make this course better for other people.
- Soon, start analyzing your own data.