Computing in an empirical biology lab
Integrating Oscar into your lab’s research
Yes, Oscar is a computational tool
Yes, Oscar is a computational tool
Yes, Oscar is a data storage tool
Yes, Oscar is a computational tool
Yes, Oscar is a data storage tool
Oscar is a powerful collaborative tool
Eliminate the need to transfer large data files between computers and users
A structured repository for raw data
A structured repository for analyses
Organization is part of the analyses, rather than something that comes after
Share analyses and data by telling people where they are rather than creating a copy.
In the short term, it makes sharing really easy.
In the long term, PI’s know write where the data, tools, and analyses generated in the lab are.
My Oscar workflow
- Oscar
My Oscar workflow
- Oscar
- Google docs
My Oscar workflow

- Oscar
- Google docs
- git
Analyses and storage on Oscar
Constrained file tree

All raw data go into a central repository

Common analyses (e.g., assembly) go into a central repository

Users have their own folders for specialized stuff
Shared access

Set permissions so that all users have access to all files in the data directory.
Constrained file tree

/gpfs/data/cdunn/
  # user data directories
  cdunn/
  ssmith/

repo/ # git repositories, with src and executables for lab-generated tools
archive/ # Snap-shots of old analyses
bin/ # third-party binaries
src/ # third-party source

analyses/
  assembly/ # Assemblies of transcriptomes and genomes
  mapping/ # Mapping analyses
  library/ # Common reference files

sequences/
  # Raw data from sequencing centers
  454/
  illumina/
  helicos/
  solid/
  sanger/

  collaborators/ # Data collected by collaborators for shared projects
  public/ # Data downloaded from public archives
  processed/ # Derivative data files, eg filtered sequences
Always be thinking about how to minimize your disk footprint

Storage is about $1 per GB

Never copy data files to analysis directories

Don’t save analyses you don’t need

Delete intermediate analysis files

Far cheaper to keep scripts and regenerate analyses than to keep analyses
Use scratch for analyses!

Doesn’t count towards quota

Higher performance?

Move analyses to data when you know you want to keep them
Documentation on Google Docs

<table>
<thead>
<tr>
<th>TITLE</th>
<th>OWNER</th>
<th>LAST MODIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly Notebook.doc</td>
<td>Shared notebooks me</td>
<td>10:50 pm me</td>
</tr>
<tr>
<td>To order</td>
<td>Shared logistics me</td>
<td>Oct 17 Freya Goetz</td>
</tr>
<tr>
<td>Mollusc Taxon Sampling</td>
<td>Shared specimen me</td>
<td>Oct 17 Freya Goetz</td>
</tr>
<tr>
<td>StellaExpressionOctober2011</td>
<td>Shared ana</td>
<td>Rebecca Helm</td>
</tr>
<tr>
<td>Smith Lab Expression</td>
<td>Shared</td>
<td>Oct 17 Rebecca Helm</td>
</tr>
<tr>
<td>Deep sequencing run stats</td>
<td>Shared analyser me</td>
<td>Oct 17 Freya Goetz</td>
</tr>
<tr>
<td>rna extractions</td>
<td>Shared specimen data</td>
<td>Oct 17 Freya Goetz</td>
</tr>
<tr>
<td>Specimen data</td>
<td>Shared specimen data</td>
<td>Oct 17 me</td>
</tr>
<tr>
<td>specimen data</td>
<td>Shared Dunn Lab</td>
<td>Oct 17 me</td>
</tr>
</tbody>
</table>
Documentation

Data and analyses are a liability rather than an asset if they aren’t well documented.

Documentation should be realtime, not something that is done after analyses.

Good documentation is a powerful teaching and learning tool.
Create a collection in google docs, share it with your lab members.

Create new analysis notebooks for each project.

Add each notebook to the lab collection.
Paste commands, notes, plots, etc, right into the doc
Code management with git
git is a:
- Distributed software revision control system
- Allows you to organize all lab software in a single central repository
- Can write and use software in the repository on any computer
git allows you to:

- Revert to an earlier version
- Store verbose comments on your code development
- Fork and merge projects
Multiple options for git central repositories

Plans & Pricing

All plans include unlimited public & private repositories.

- **5 USERS**
  - **FREE**
  - **CURRENT PLAN**

- **10 USERS**
  - $10 / month
  - CHOOSE

- **25 USERS**
  - $20 / month
  - CHOOSE

- **50 USERS**
  - $40 / month
  - CHOOSE

- **UNLIMITED**
  - $80 / month
  - CHOOSE