# How to do serious computing for biological data
# - UNR HPC resources for life scientists –

**Richard Tillett**

June 20th, 2018

HPC Research Computing
University of Nevada, Reno

University of Nevada, Reno
Research & Innovation
*Where big ideas ignite.*℠

NEVADA INBRE

NATIONAL INSTITUTES OF HEALTH
NIGMS

University of Nevada, Reno

# Outline

- Our migration to Pronghorn (Nevada Center for Bioinformatics)
- Getting software working on HPC
- A benchmark case study (RNA-seq)
- Training opportunities for you

# The Nevada Center for Bioinformatics move to Pronghorn

# CFB setup before pronghorn

- 3 computational servers in a mini-cluster
- Shared storage arrays between them
- Software job scheduler to manage jobs among the machines

# CFB setup after pronghorn

- Same mini-cluster, but now it is for development and testing
- Invested in 3 node equivalents and storage
- Step by step, moved tools and pipelines onto Pronghorn

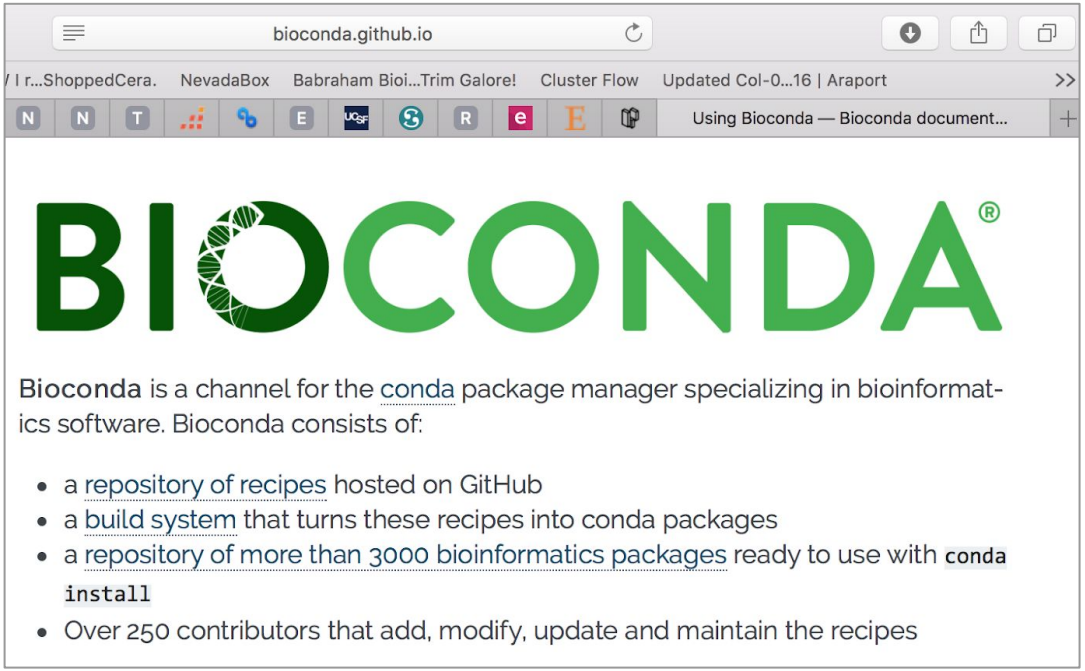
# Getting software running on HPC

# Installing software on HPC

- Compiling from source
  - Advantages:
    - Very Unix-y
    - Simple when it works
  - Disadvantages
    - Dependency hell when it doesn’t
    - Missing libraries, static vs. dynamic libs, version requirement mismatches, etc.
    - On someone else’s machine, you can’t just `sudo apt install libxml2-dev` troubles away

# Installing software on HPC

- Conda & bioconda
- Act similar to package managers (yum, apt), but does not need root permissions



# Installing software on HPC

- Singularity containers on Pronghorn
  - http://singularity.lbl.gov
- Fairly advanced topic
- But essentially guarantees code portability

# Benchmarking RNA-seq on Pronghorn

# A benchmarking case study

- RNA-seq (12 samples)
  - 4 conditions x 3 biological replicates
  - Paired-end sequences, human
  - Illumina NextSeq 500 output
- To run on Pronghorn using our automated pipeline

| id | read pairs | |
|---:|---:|---|
| sample_01 | 30,050,856 | |
| sample_02 | 23,684,972 | |
| sample_03 | 27,036,920 | |
| sample_04 | 25,451,184 | |
| sample_05 | 32,251,584 | |
| sample_06 | 33,040,002 | |
| sample_07 | 26,295,200 | |
| sample_08 | 27,975,013 | |
| sample_09 | 20,152,064 | |
| sample_10 | 33,819,297 | |
| sample_11 | 31,798,509 | |
| sample_12 | 33,071,288 | |
| **total** | **344,626,889** | **35 GB (zipped)** |

*Data courtesy Heather Burkin*

# Our pipeline: Lahontan

- Reproducible
- Modular
- Portable

Features Business Explore Marketplace Pricing This repository Search

UNR-CFB / lahontan Watch 5

Code Issues 15 Pull requests 0 Projects 0 Wiki Insights

Lahontan: a flexible, multiscale RNA-seq pipeline

295 commits 2 branches 5 releases 3 contributors

Branch: master New pull request

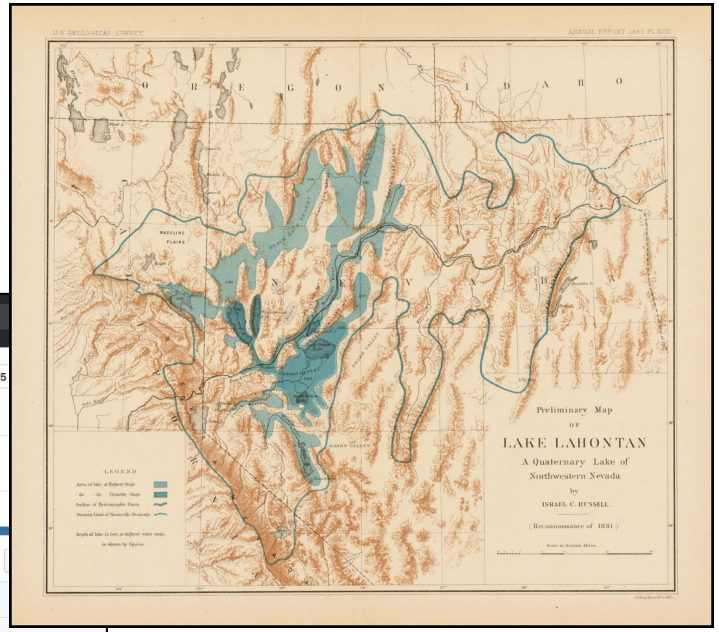
Alberto024 Fixed bug in slurm batch script

| | |
|---|---|
| lib | Fixed bug in slurm batch script |
| report | Moved scripts to lib; updated Singularity; reorganization |
| src | Added gclib as a submodule in order to satisfy gffread/gffcompare dep... |
| .gitignore | Added __pycache__ to gitignore |
| .gitmodules | Added gclib as a submodule in order to satisfy gffread/gffcompare dep... |
| LICENSE | Added LICENSE |
| README.md | Reorganized README |
| Singularity | Added bzip2 library to singularity recipe |

README.md

## Lahontan

Lahontan is a set of tools built around an object-oriented pipeline used in qualifying and processing sets of paired-end next-gen RNA sequencing data. It takes fastq files from a high-throughput sequencer and performs

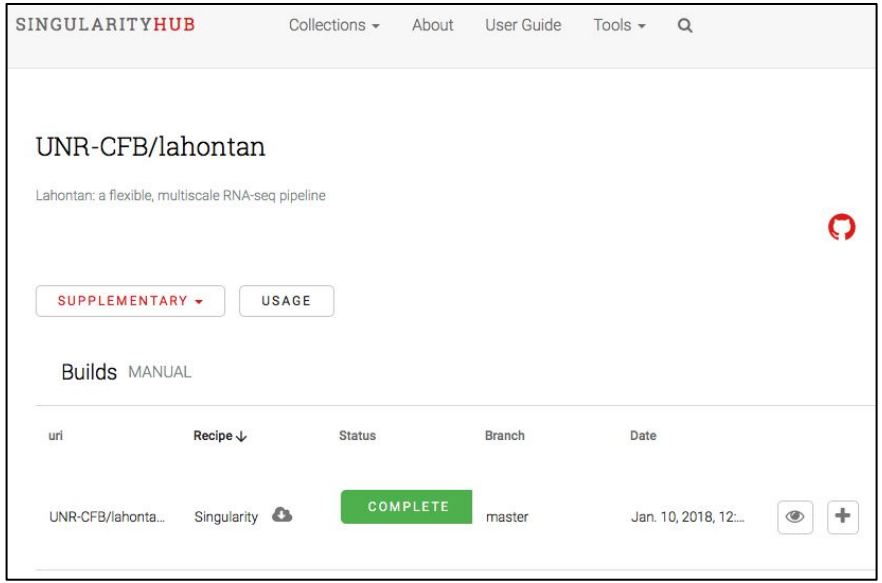



*Lake Lahontan, US Geological Survey, 1881*

*https://github.com/UNR-CFB/lahontan*

# Lahontan is scalable and deployable

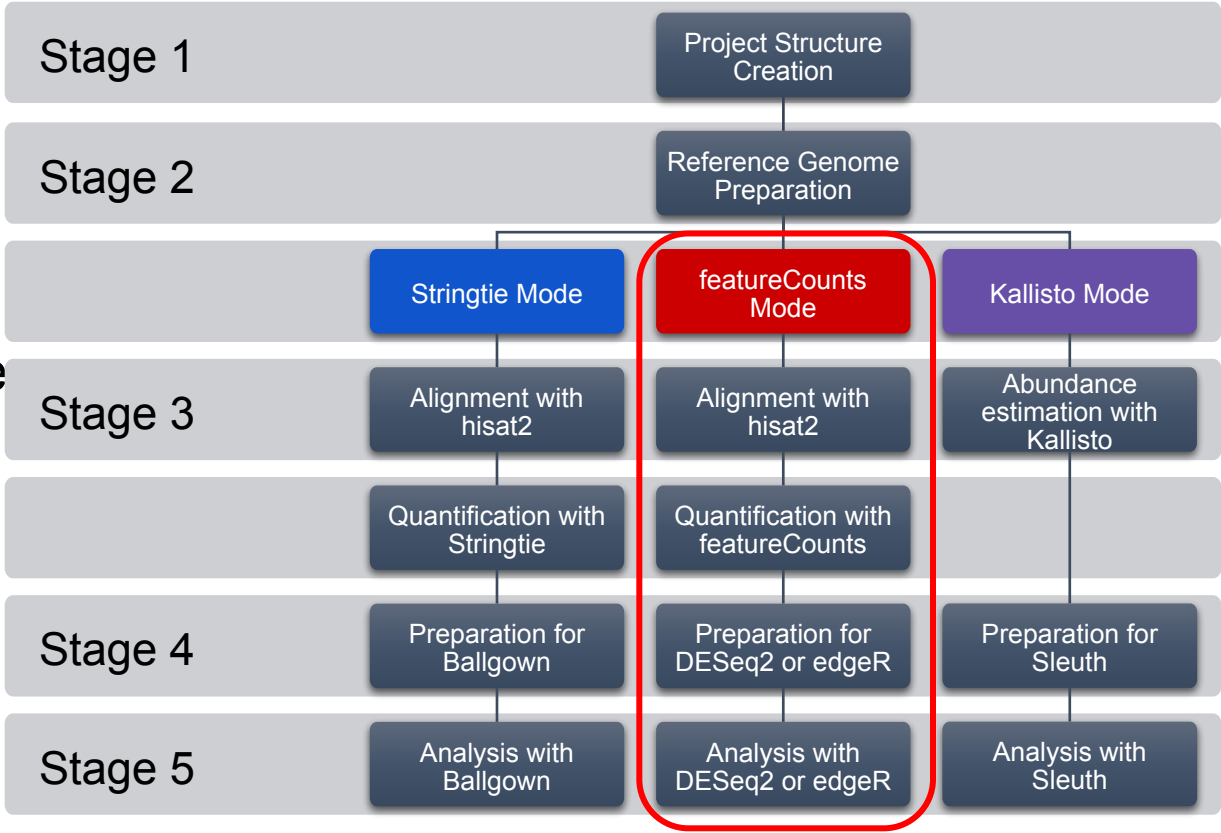
- It runs on a single linux workstation
- It runs on mid-sized parallel computing clusters
  - (w/ slurm queue scheduling)
- It runs on Pronghorn without installation
  - Download the Singularity image and it just works
  - HPC/cloud hybrid model




*https://www.singularity-hub.org/collections/388*

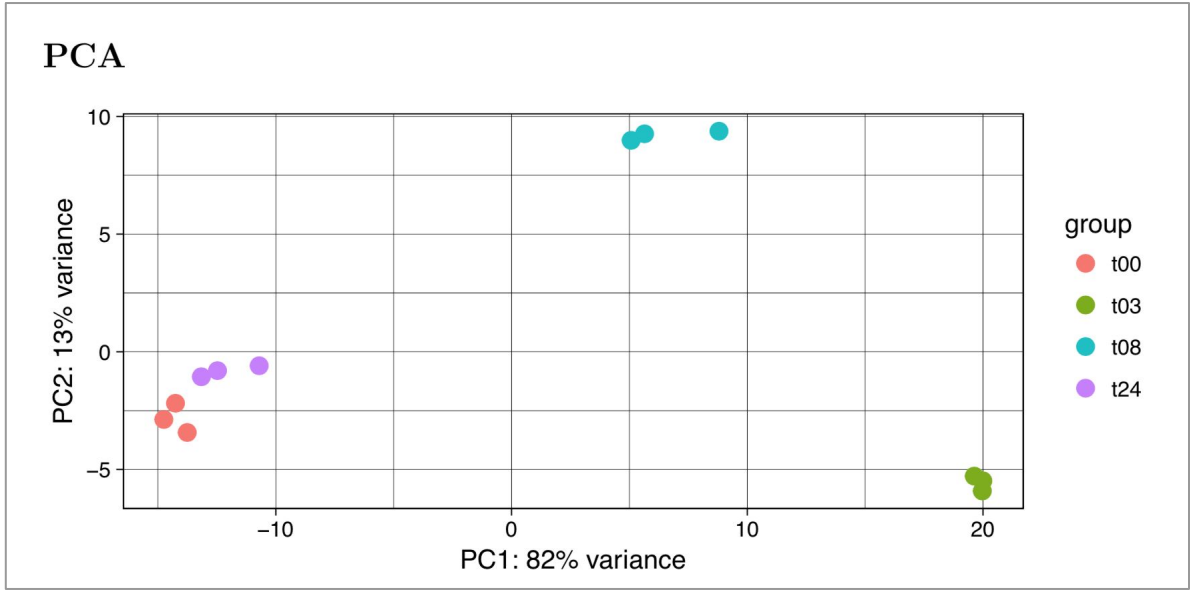
# Lahontan is automated & modular

- Edit a few text files to tell lahontan where the input data and reference genome is
- Then invoke lahontan with the desired pipeline mode
- And it trims, does QC, aligns, quantifies, tests DEGs, generates reports



# Lahonton generates:

- QC reports
- Trimmed reads
- Aligned reads (BAM files)
- Gene-level count tables
- DEG tests w/ DeSeq2 and EdgeR tools & the source R scripts to alter the tests, as needed
- PDF reports for stats results (w/ useful descriptive statistical figures too)




*EdgeR report PCA plot for the benchmark set*

# Resources + Costs

**Resources used**

- CPUs used: 96 physical CPU cores (3 full nodes)
- Run time: 58 mins
- Total disk use: 123 GB

```
35G     bench_input
76G     bench_run.m50.e
3.6G    lahontan
8.8G    Reference
123G    total
[rltillett@login-0 benchmark]$
```

**Costs (Pay-as-you-go rate)**

- At compute rate of $0.01 per CPU-hour
  - compute cost: $0.96
- At storage rate of $8 per Tb per month
  - Storage cost: $8.00
- **total cost: $8.96**

Interested, maybe?

# Training opportunities

- HPC Hackathons, Fridays at 2pm, MIKC 405 lab
- Help choose the next seminar/workshop topic. It could cover:
  - Unix basics
  - Migrating to Pronghorn
  - RNA-seq w/ Lahontan
  - R stats topics (deseq2 and edgeR)
- Apply for a Service Award (deadline June 29th)

# INBRE Service Awards

- May request up to $5000 in total value award between Bioinformatics, Proteomics, Genomics.
- Proposals could request a mix of analysis service and training for the same
- (Proposals are not required to do so. You may request analysis services without any training component)
- Grab a flyer after the talk!

# Many thanks to:

*This work was funded by the Nevada INBRE, with a grant from the National Institute of General Medical Sciences (8 P20 GM103440) from the National Institutes of Health.*

Sebastian Smith (Research Computing)

Heather Burkin (Pharmacology)

Juli Petereit (CTRIN, Nevada Center for Bioinformatics)

Alberto Nava (UC Berkeley)

HPC Research Computing University of Nevada, Reno

University of Nevada, Reno Research & Innovation *Where big ideas ignite.*℠

NEVADA INBRE

NATIONAL INSTITUTES OF HEALTH NIGMS

University of Nevada, Reno