
*"Maybe I should have stuck with Excel."*

# Bioinformatics workshop

## Eine kurze Einführung zum *Phyloserver2*

nhm
natural history museum vienna

Martin Kapun - 19/10/2021

# Outline

- GitHub
- UNIX
- Data management
- ATOM
- Bioinformatischer Service
- Der neue Server
- **Hands-on exercise**

# (1) Was ist Github?

- Github ist eine netzbasierter **Versionsverwaltungsdienst** für Softwareprojekte

- Github erlaubt es **kollaborativ** an Softwareprojekten zu arbeiten

- Github ermöglicht es umfangreiche **Dokumentation** für Softwareprojekte anzulegen.

# Wozu brauchen wir Github?

**Repräsentation**
digital science @ NHM

**Publikation**
von neuer Software

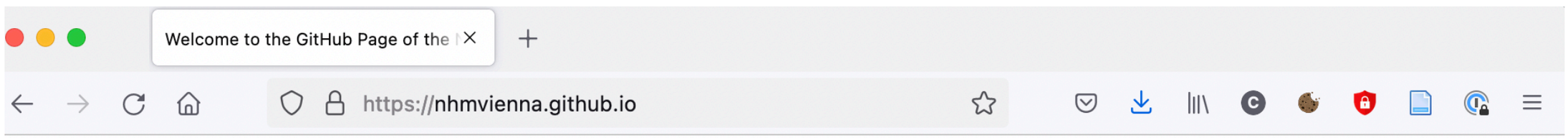
**Tutorials**
für Basiswissen/workflows

**Dokumentation**
von software-basierten Projekten

**Ticketing**
für (bio-)informatischen Service

# Github Page

https://nhmvienna.github.io/




**nhmvienna.github.io**

natural history museum vienna

Official Github page

View My GitHub Profile

Download **ZIP File** | Download **TAR Ball** | View On **GitHub**

Hosted on GitHub Pages — Theme by orderedlist

## Welcome to the GitHub Page of the Natural History Museum Vienna

This page will be filled with life in the near future. Stay tuned!

# Dokumentation

Projektordner als **private oder öffentliche** repositories

capoony / EchinoUCE Private

Code | Issues | Pull requests | Actions | Projects | Security | Insights | Settings

master | 1 branch | 0 tags | Go to file | Add file | Code

| | | |
|---|---|---|
| capoony updates | ff88396 1 hour ago | 44 commits |
| final | updates | 1 hour ago |
| log | updates | yesterday |
| scripts | updates | yesterday |
| shell | updates | 1 hour ago |
| .gitignore | init | 14 days ago |
| README.md | updates | 1 hour ago |

# Dokumentation

Projektordner als **private oder öffentliche** repositories

README.md

## Designing baits of ultraconserved genetic elements in Echinodermata

### Methods

#### Material

Following tutorial IV in Faircloth (2016), I used draft genomes of six echinoderms (Strongylocentrotidae: Strongylocentrotus purpuratus [*Spur*; GS: 921.856Mb; contigs: 1,546; N50: 2,052,140], Hemicentrotus pulcherrimus [*Hpul*; GS: 568.912Mb; contigs: 86,128; N50: 9,641]; Toxopneustidae: Lytechinus pictus [*Lpic*; GS: 998.846Mb; contigs: 11,535; N50: 219,597], Lytechinus variegatus [*Lvar*; GS: 973.87Mb; chromosome scale draft]; Cidaridae: Eucidaris tribuloides [*Etri*; GS: 2187.26Mb; contigs: 1,006,568; N50: 6,630]; and as an outgroup: Ophiotrichidae: Ophiothrix spiculata [*Ospi*; GS: 2764.32Mb; contigs: 644,798; N50: 6,474]), which are already available on GenBank.

I downloaded the raw genomes in FASTA format, simplified their filenames, purged unnecessary information (i.e. space-separated meta-information) in the FASTA headers using a custom script and converted the FASTA files to the 2Bit format.

# Tutorials

nhmvienna / FirstSteps Private

Unwatch 2 Star 0

Code Issues Pull requests Actions Projects Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

capoony Merge pull request #2 from nhmvienna/add-license-1 ... 3cbcf98 6 hours ago 24 commits

| | | |
|---|---|---|
| CodeOfConduct.md | Update CodeOfConduct.md | 6 hours ago |
| GitHubBasics.md | Update GitHubBasics.md | 6 hours ago |
| LICENSE | Create LICENSE | 6 hours ago |
| RBasics.md | Create RBasics.md | 6 hours ago |
| README.md | Update README.md | 6 hours ago |
| UNIXBasics.md | Update UNIXBasics.md | 6 hours ago |

README.md

## FirstSteps

This repository provides information for (1) new users of the NHM Github account and/or (2) new users of the PhyloServers.

Check out the following pages

- Code of Conduct
- Github basics
- UNIX basics
- R basics

### About

This repository provides information for (1) new users of the NHM Github account and/or (2) new users of the PhyloServers.

Readme

GPL-3.0 License

### Releases

No releases published
Create a new release

### Packages

No packages published
Publish your first package

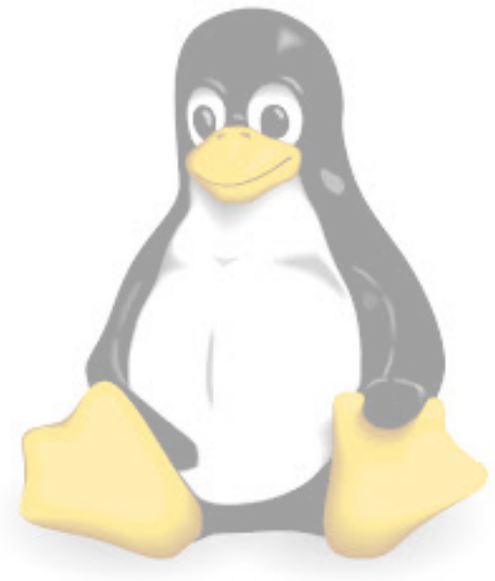
# Let’s have a look:

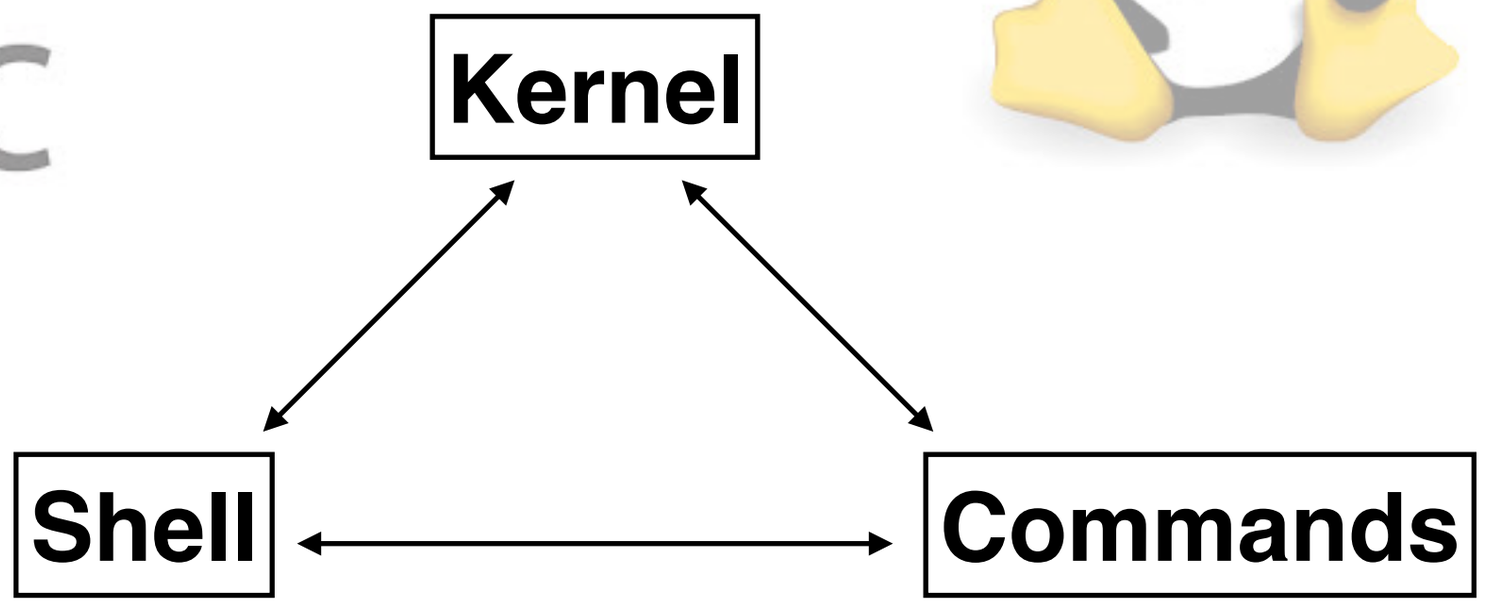
https://github.com/nhmvienna

# All tutorials for bioinformatics at the NHM can be found here:

https://github.com/nhmvienna/FirstSteps

# (2) UNIX



# Was ist UNIX?

**Kernel:**

- master control program of the operating system

**Shell:**

- interactive program that provides an interface between the user and the kernel.

- interprets commands entered by the user or supplied by a shell script, and passes them to the kernel for execution.

# Was ist UNIX?

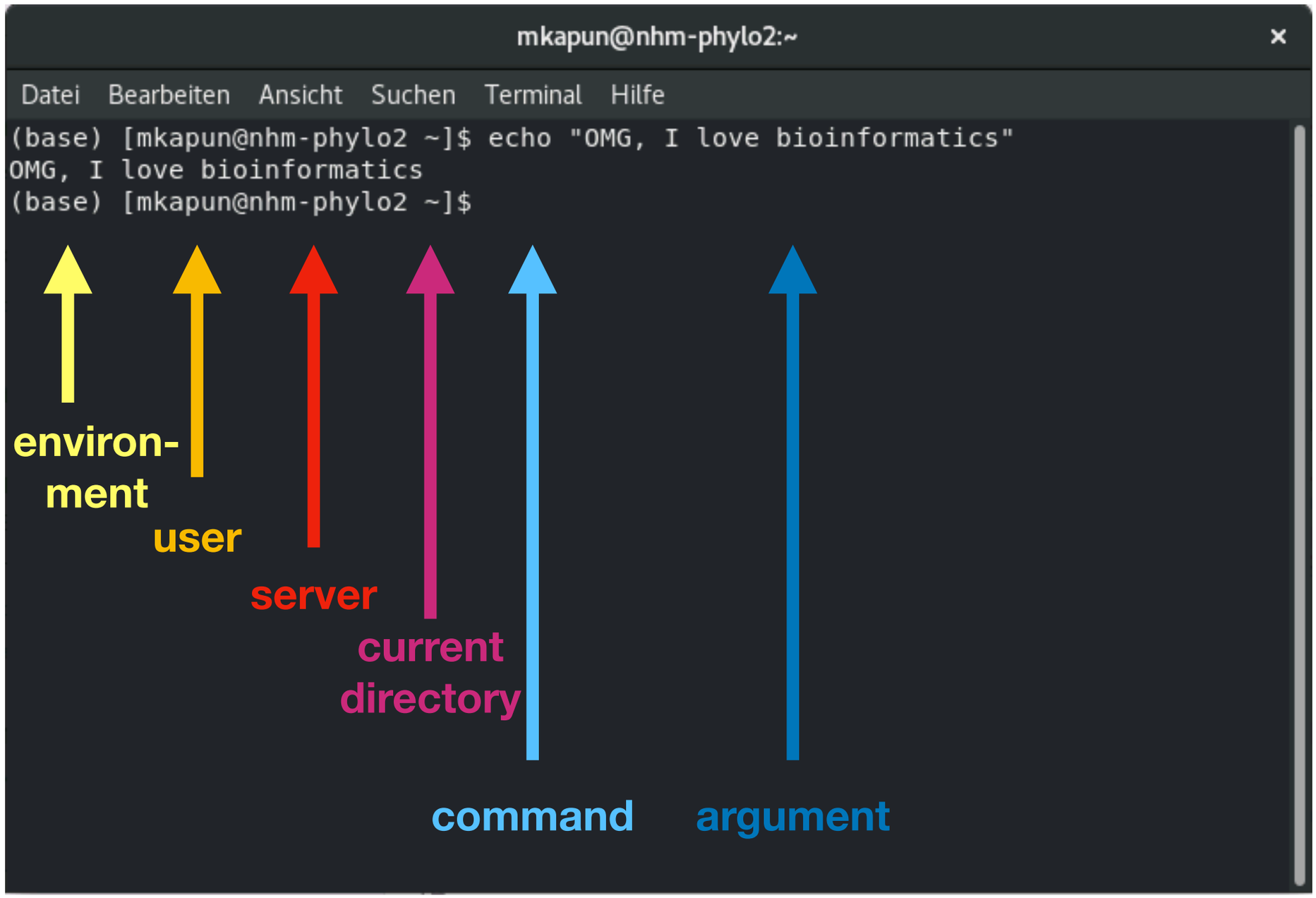
**Commands:**

- Unix and Unix-like systems include a large core of standard utilities for editing text, writing, compiling, and controlling programs, etc.

- Many commands allow arguments (known as options or flags) to modify their default behavior.

- Users enter commands and arguments on the shell command line, and then the shell interprets them and passes them to the kernel for execution.

# Das Terminal Fenster



# The bash shell

**Tricks And Tips**

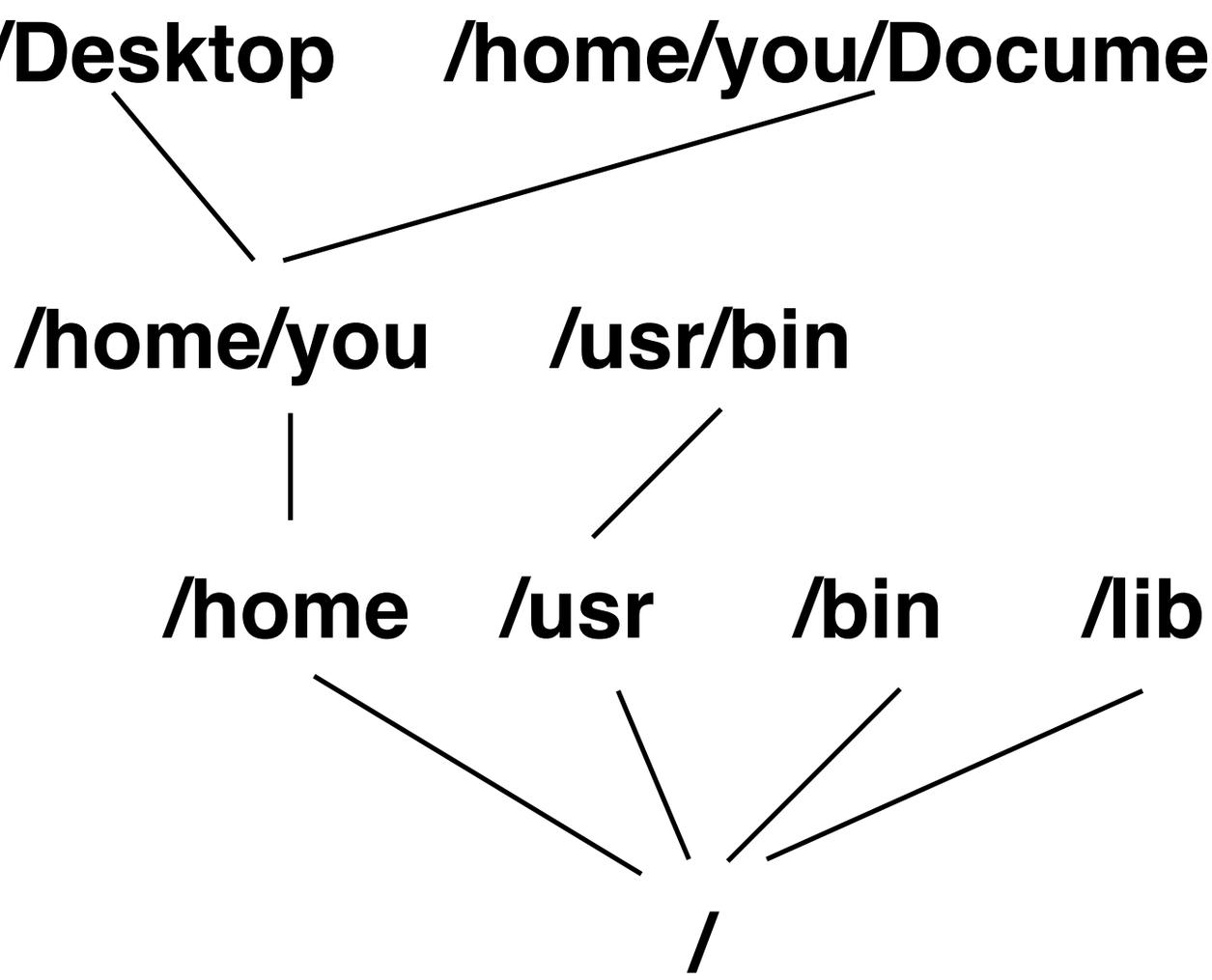
- UNIX is case-sensitive: LS is not the same as ls

- bash “remembers” you previous commands; use up/down arrows to navigate through the history

- Use Tab key to autocomplete command names, path names or file names

# Was ist UNIX?

The tree file system



# directories...

| command | description |
|---|---|
| ls | list files and directories |
| ls -a | list all (hidden) files and directories |
| mkdir | make a directory |
| cd directory | change to named directory |
| cd | change to home-directory |
| cd ~ | change to home-directory |
| cd .. | change to parent directory |
| pwd | display the path of the current directory |

# … and files

| command | description |
|---|---|
| cp file1 file2 | copy file2 and save it as file2 |
| mv file1 file2 | move file1 and rename it to file2 |
| rm file1 | remove file1 |
| wc file1 | count characters, words and lines in file1 |
| head file1 | show first 10 lines of file1 |
| head -100 file1 | show first 100 lines of file1 |
| tail file1 | show last 10 lines of file1 |
| less file1 | powerful preview of file1 (more later) |

# Input/Output

| command | description |
| --- | --- |
| command > file | redirect standard output to a file |
| command >> file | append standard output to a file |
| command < file | redirect standard input from a file |
| command1 \| command2 | pipe the output of command1 to the input of command2 |
| cat file1 file2 > file0 | concatenate file1 and file2 to file0 |

# Tutorials

- Check out: https://github.com/nhmvienna/FirstSteps/blob/main/UNIXBasics/UNIXBasics.md

# (3) Data management

- The key to a successful bioinformatics project is a clean and transparent data structure and documentation

# Ordnerstruktur

**project folder**

**documentation**

daily protocol

Master shell script

individual shell script for every analysis

**data**

raw data

reference genome(s)

cleaned data

mapped data

SNP data

annotations

**programs**

programs

Python scripts

*R* scripts

**analyses**

intermediate results

Tables

Figures

# Documentation

**Shell scripts**

- use Text Editors, e.g. ATOM to generate shell scripts.
- Each shell script should contain full paths and all commands necessary to reconstruct the analysis

```
#!/bin/sh

#  GIS_process.sh
#
#
#  Created by Martin Kapun on 23/06/14.
#
## pass folder name as variable $2 in shell command


#### use R to extract annual and monthly average data from worldclim: THANKS Anna Kosticova

mkdir /Volumes/DATA/nescent/analyses/climate/$2

echo """
# load installed package to the R project
require(raster)
# first load WC bio variables at the resolution of 2.5 deg
biod <- getData('worldclim', var='bio', res=2.5)
tmind <- getData('worldclim', var='tmin', res=2.5)
tmaxd <- getData('worldclim', var='tmax', res=2.5)
precd <- getData('worldclim', var='prec', res=2.5)
```

script name

time stamp

comment starting with a # symbol

# Documentation

## Master shell script

```
###### Master.sh

###### 29/1/2013

############################################################

## in all datasets the indivduals are arranged like this:

#--in2lt 2,3,6,10,12 --in3rmo 2,3,7,11,15 --in3rc 0,1,4,6,10,12 --in3lp 4,12,13,14, --cold-hot 0,1,2,3,13,14,15,16 --all
    0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 --names 52,53,80,100,117,132,136,150,168,85,106,129,143,89,91,96,21

# note that individuals 5 (132) and 9 (85) were excluded because they had too low coverage

# indivdiuals 3 (100) and 6 (136) are males

############################################################

## at first I was testing whether candidate SNPs from the cage experiment are over- or underrepresented within inversion regions in the poolseq data.

#sh /Volumes/Temp/martin/Projects/inversion/shell/overepresented_snps.sh

## what is the residual heterozygosity in the reference strain?

#sh /Volumes/Temp/martin/Projects/inversion/shell/residual_het.sh

################################### at next I tested several different parameters to estimate the sire alleles in the indivduals:

## at first I created an mpileup, a sync file and calculated the coverage

sh /Volumes/Temp/martin/Projects/inversion/shell/sync_and_cov.sh

## then, I extracted the haplotype alleles

sh /Volumes/Temp/martin/Projects/inversion/shell/extract_sire_alleles.sh ## updated

## then I calcualted pi and FST for different combination of datasets and binned the results in 100kb windows

sh /Volumes/Temp/martin/Projects/inversion/shell/pi_fst.sh ## updated

## then I visualized the results in R for all chromosomes:

sh /Volumes/Temp/martin/Projects/inversion/shell/visualize_pi-fst.sh ## updated
```

# Software

- always store the version of all programs used for your project.

- always store ALL custom (Python, Perl, Java, R) scripts used for the project locally in the project programs folder.

# Analyses

- Use comprehensive but readable file/folder names

`Penn_Maine_NormRecomb_concordance_Fst05_FDR001_candidates_test_final1_REAL_final.fet`

- Make use of subfolders rather than storing everything in one big big folder. This also facilitates naming and navigating through the data

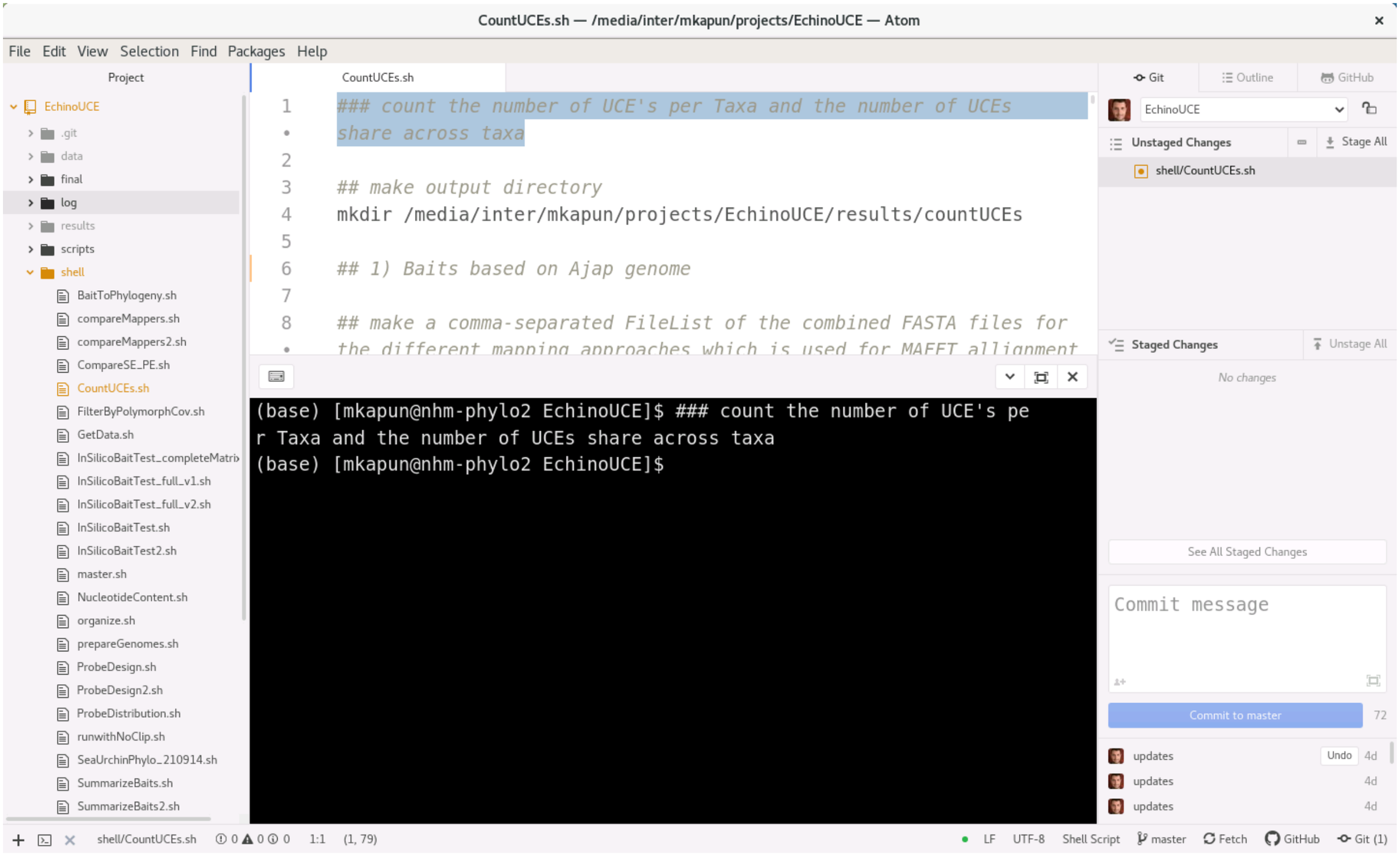
# Data storage

- Storage space is precious!!!

- Retain ALL raw data and final results but (if possible) delete intermediate files (DOCUMENTATION!!)

- Compress raw data and large output files

# (4) Atom editor

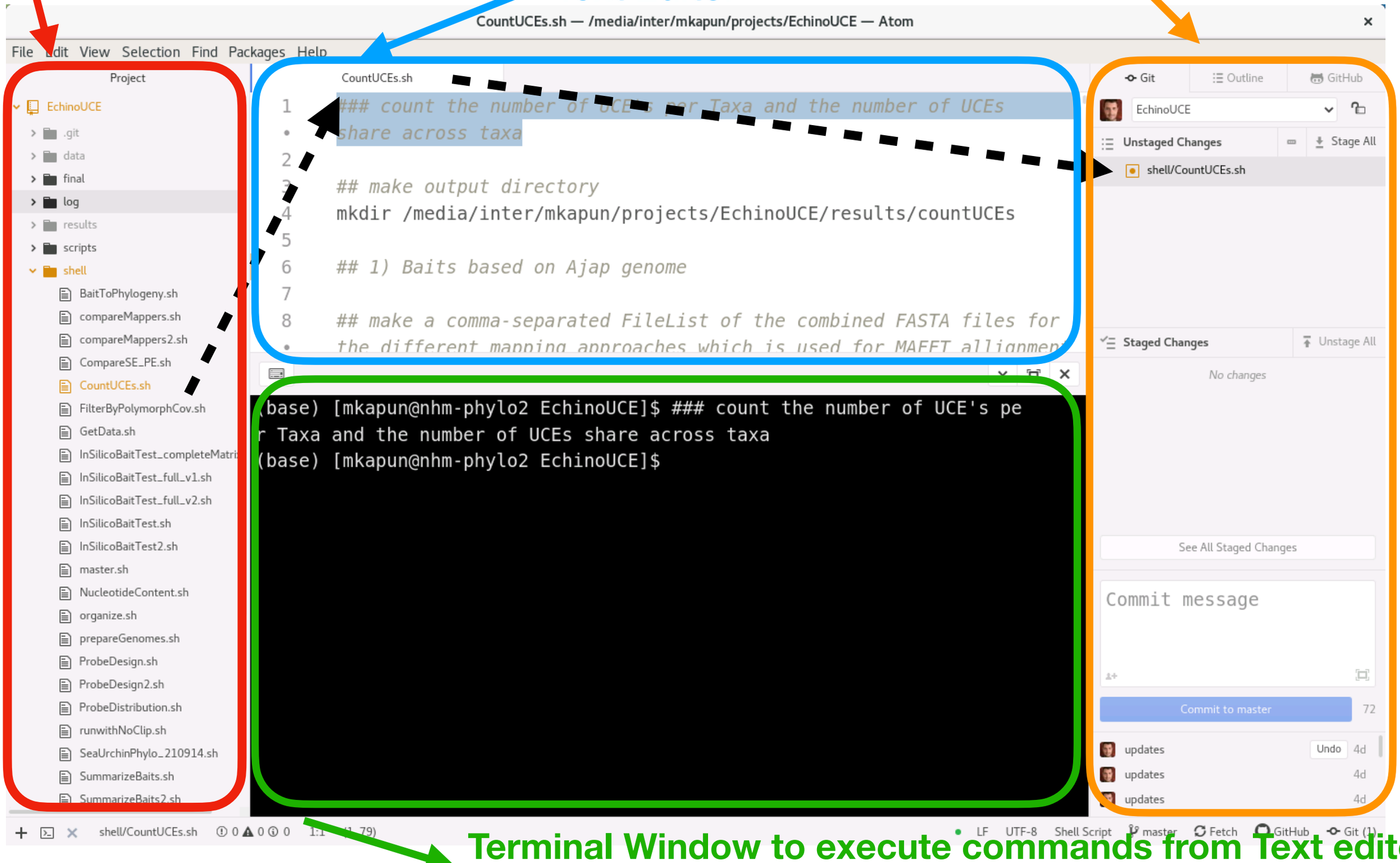


# Atom editor

Project Folder Viewer

Text Editor

GitHub Interface



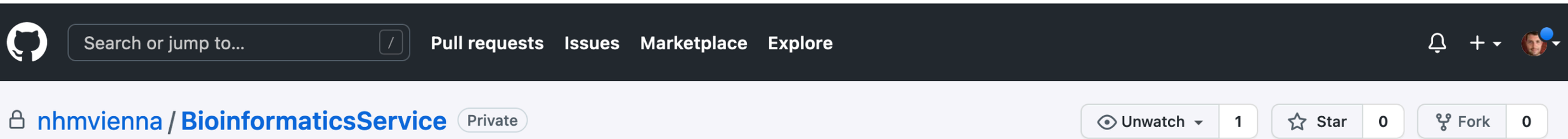

Terminal Window to execute commands from Text editor

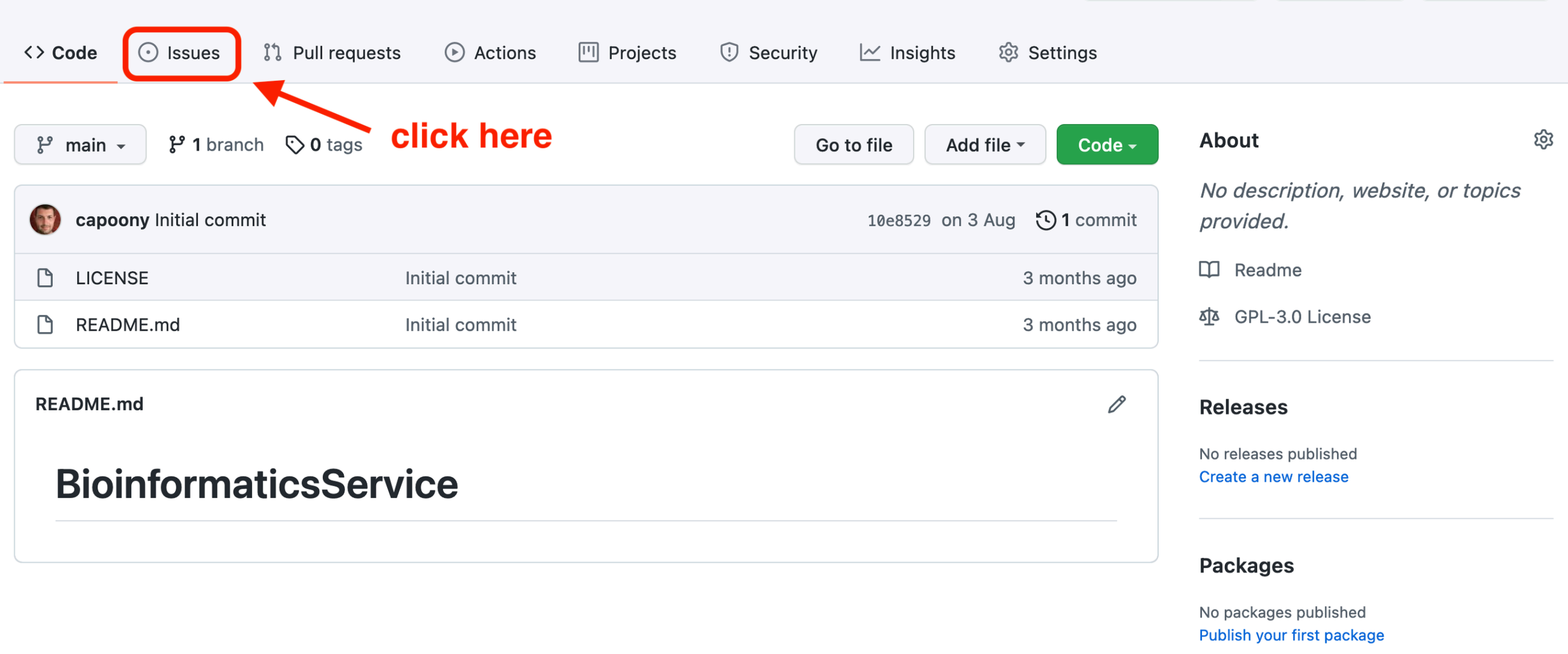
# Tutorials

- Check out: https://github.com/nhmvienna/FirstSteps/blob/main/ATOMbasics.md

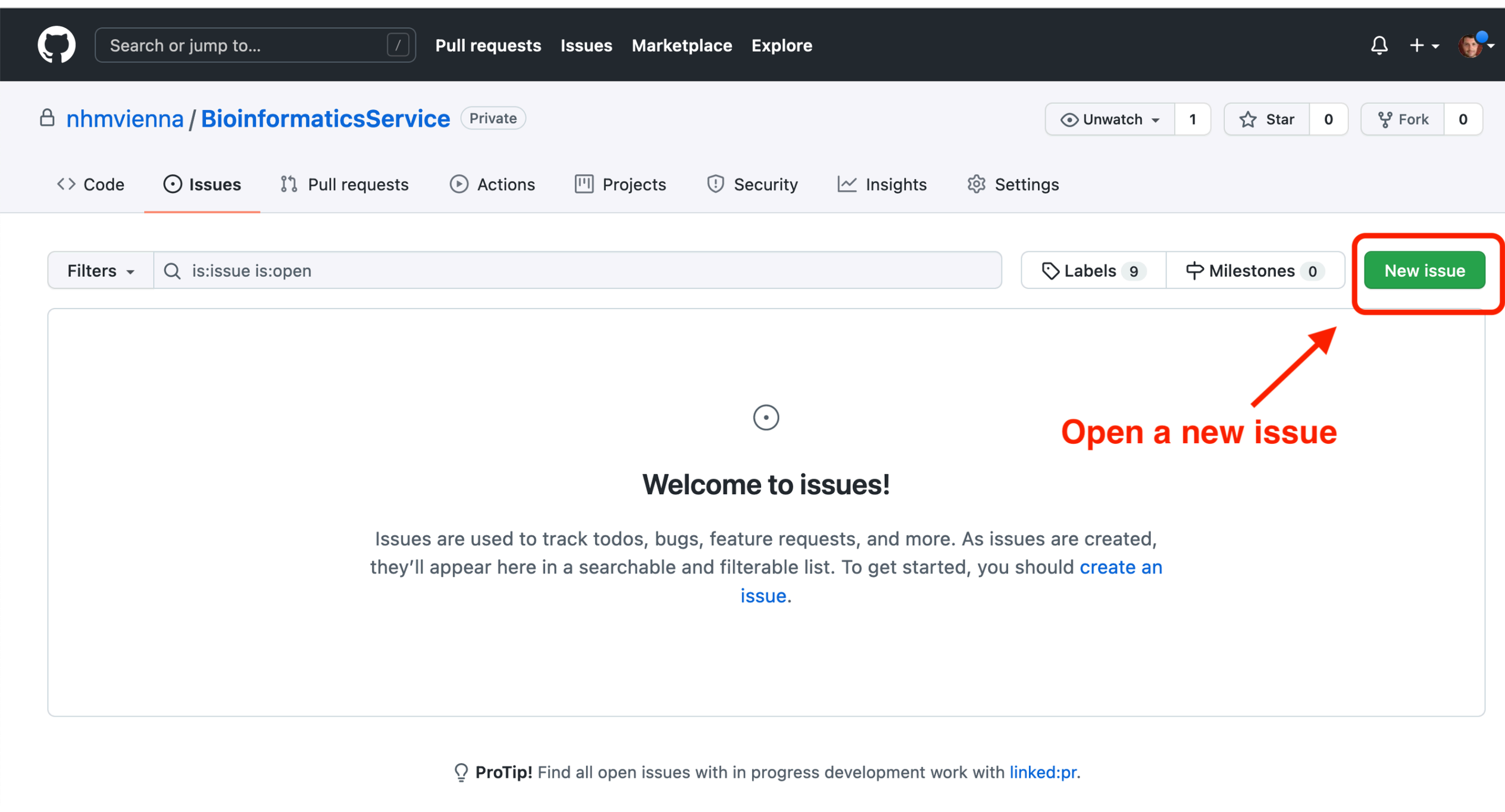
# (5) Hilfe mit Bioinformatik

https://github.com/nhmvienna/BioinformaticsService



# Ticketing


Search or jump to...
Pull requests
Issues
Marketplace
Explore
nhmvienna / BioinformaticsService
Private
Unwatch 1
Star 0
Fork 0
Code
Issues
Pull requests
Actions
Projects
Security
Insights
Settings
Filters
is:issue is:open
Labels 9
Milestones 0
New issue
Open a new issue
Welcome to issues!
Issues are used to track todos, bugs, feature requests, and more. As issues are created, they'll appear here in a searchable and filterable list. To get started, you should create an issue.
ProTip! Find all open issues with in progress development work with linked:pr.

# Ticketing

Search or jump to... / Pull requests Issues Marketplace Explore

nhmvienna / **BioinformaticsService** Private

Unwatch 1 Star 0 Fork 0

Code Issues Pull requests Actions Projects Security Insights Settings

Help with phylogenetic analysis using the R package ape

**(1) Choose a comprehensive title**

Write Preview

Hi Martin,

I try to plot a tree in newick format using the R package ape, but cannot get it to work, could you have a look at my script?

It is attached to this issue.

Thanks a lot
[Rscript.txt](https://github.com/nhmvienna/BioinformaticsService/files/7364668/Rscript.txt)

**(2) Accurately describe your issue**

Attach files by dragging & dropping, selecting or pasting them.

Styling with Markdown is supported

Submit new issue

Remember, contributions to this repository should follow our GitHub Community Guidelines.

**(4) Submit new issue**

Assignees

No one—assign yourself

Labels

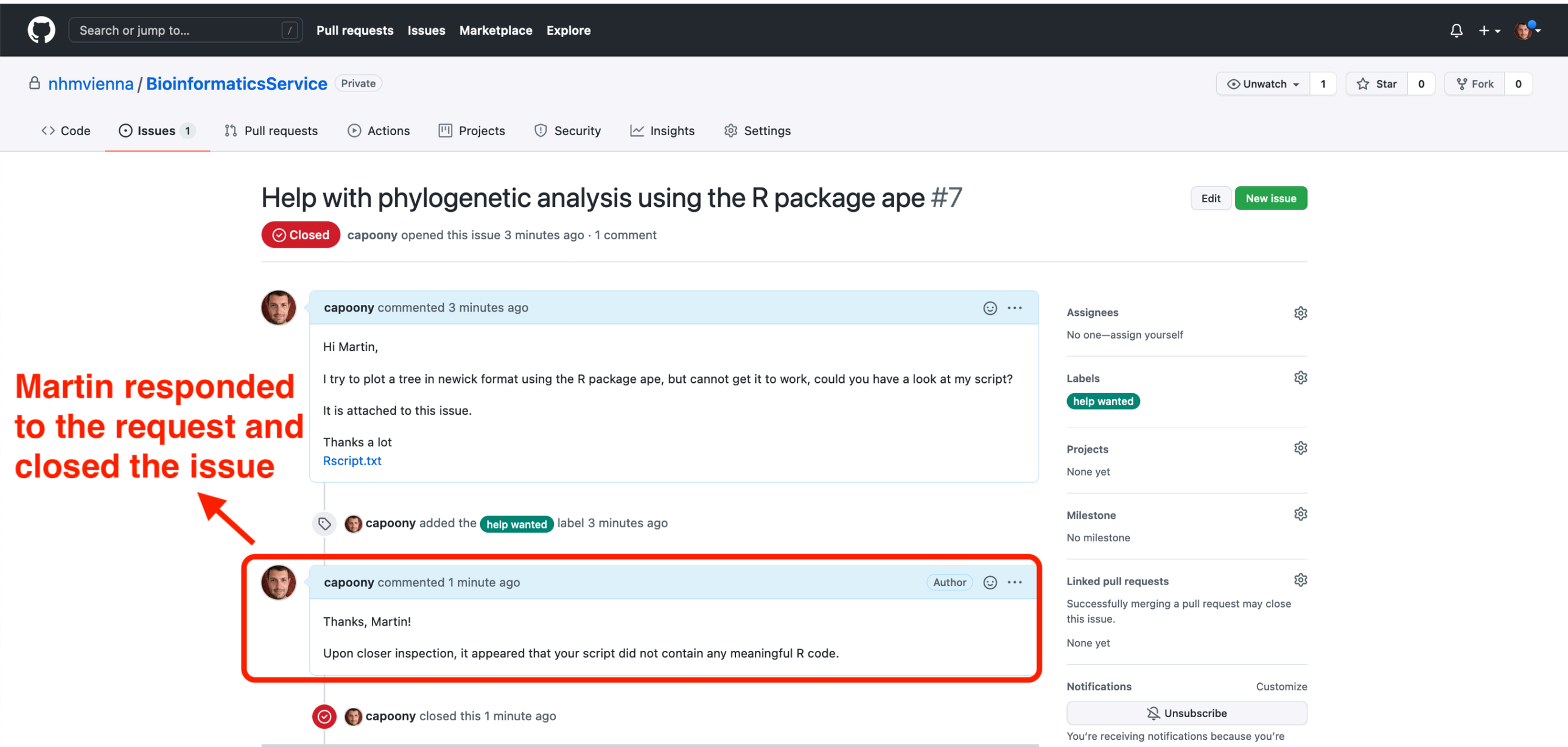
Apply labels to this issue

Filter labels

- help wanted — Extra attention is needed
- bug — Something isn't working
- documentation — Improvements or additions to documentation
- duplicate — This issue or pull request already exists
- enhancement — New feature or request

**(3) Assign label**

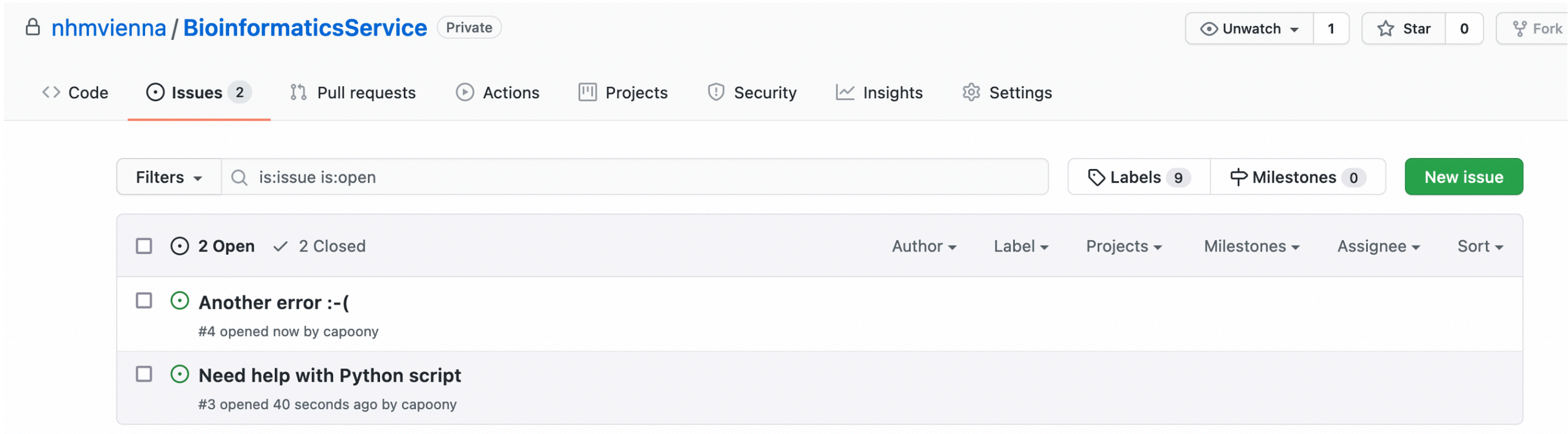
# Ticketing




**Martin responded to the request and closed the issue**

# Ticketing

- Transparente Kommunikation
- Erleichtert Planung
- Erleichtert Statusabfrage



# Tutorials

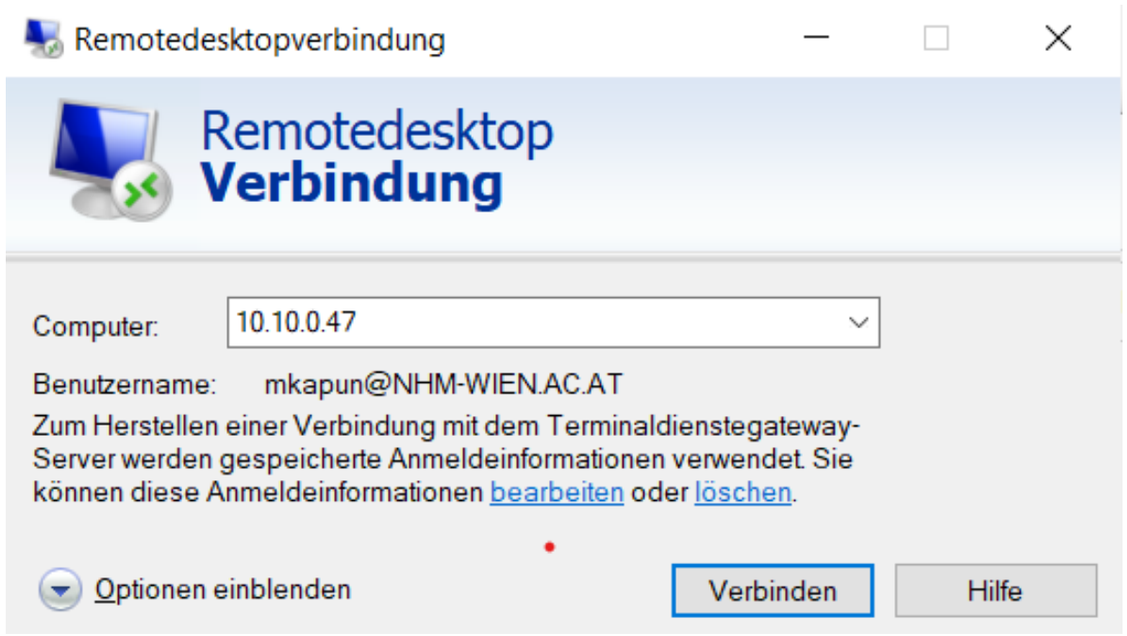
- Check out: https://github.com/nhmvienna/BioinformaticsService

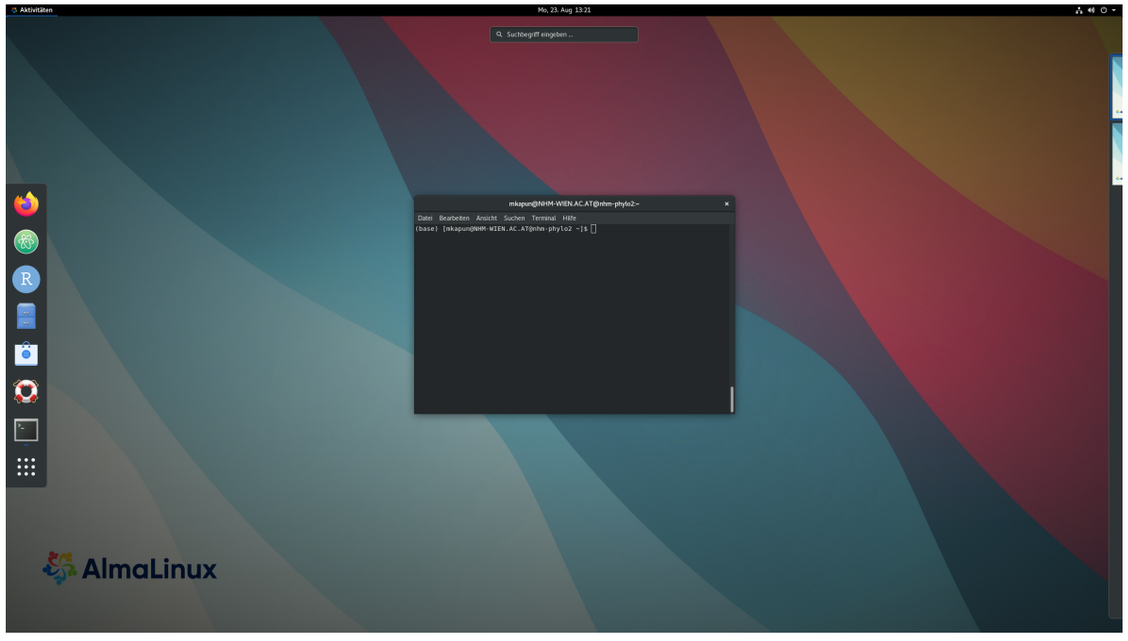
# (6) Phyloserver2

- Type: Dell PowerEdge R7525
- CPU: 2 x AMD 7742 Epyc (2.25GHz; 64cores/128threads; 256MB cache; 225W)
- RAM: 1.5 TB (2 x 750 GB)
- Storage: 3.84 TB (RAID0; 2x 3.84 TB SSD vSAS); 48 TB (RAID5; 4 x 16 TB SAS HD)
- OS: AlmaLinux OS 8

# Easy access

- IP: **10.10.0.47**
- Heute:
  - Username: **test_mmusterfrau**
  - PW: **test123!!**
  - **7 Tage gültig**
- Username: **mmusterfrau**
- PW: selbst gewählt
- Kann nur von innerhalb des NHM Netzwerks benutzt werden



# Installierte Software

- Liste mit aller installierter Software: https://github.com/nhmvienna/FirstSteps/blob/main/Bioinformatics/SoftwareList.md
- Programme sind nicht sofort verfügbar, der Installations-pfad muss erst angegeben werden, z.B.:

```
module load Alignment/ncbi-BLAST-2.12.0

conda activate mafft-7.487
```

# Transparente Dokumentation

- Sämtliche Installationen auf dem Phyloserver2 sind auf GitHub dokumentiert.

- https://github.com/nhmvienna/PhyloserverInstallationDocs

# Working directory

- /home/ directory nur mit 100gb platz -> nicht genug für große Projekte

- /media/inter mit 48TB Speicherplatz

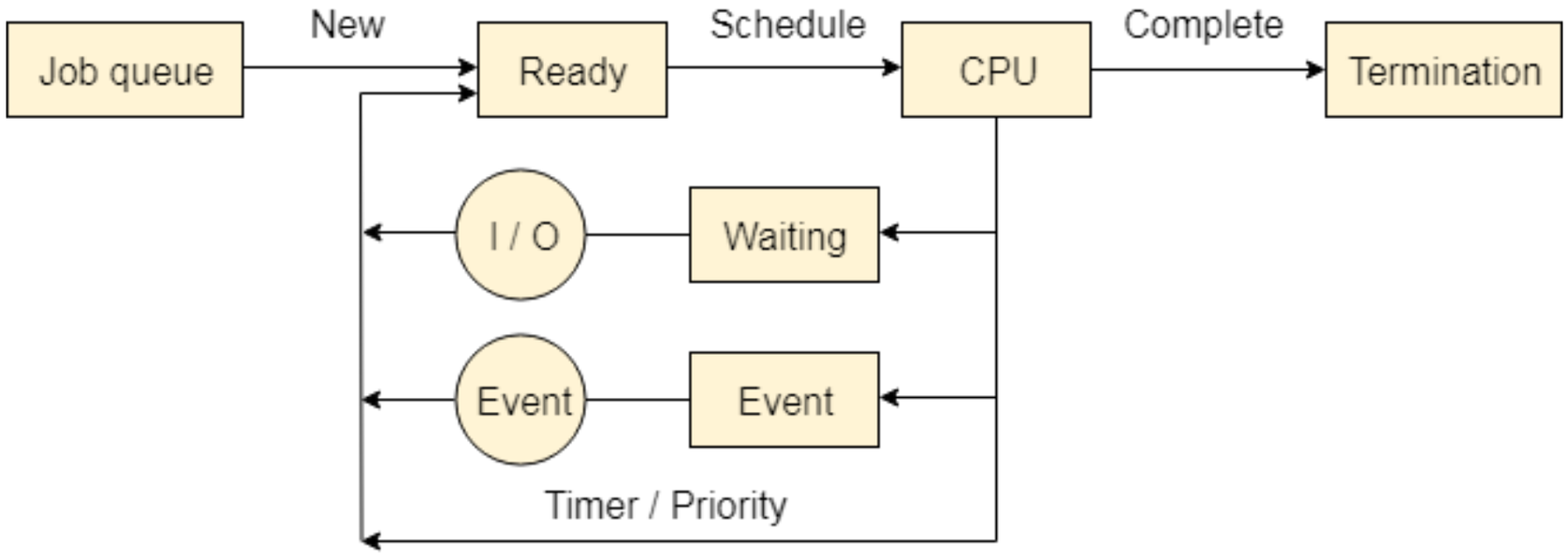
- Benutze daher /media/inter/<username> für Datenanalysen.

# Job queuing

- Siehe tutorial: https://github.com/nhmvienna/FirstSteps/blob/main/Bioinformatics/OpenPBS.md

- Wenn vielen Nutzer den server gleichzeitig nutzen -> Überlastung des Systems.

- Lösung: OpenPBS



# Daten verschieben

- Mounted  Windows H:\ in /home/winusers
- scp in Windows Powershell

```
Windows PowerShell
Copyright (C) Microsoft Corporation. Alle Rechte vorbehalten.

Lernen Sie das neue plattformübergreifende PowerShell kennen – https://aka.ms/pscore6

PS H:\> scp mkapun@10.10.0.47:/home/mkapun/github/FirstSteps/UNIXBasics/TestFileToMoveAround.txt H:\Work\test
mkapun@10.10.0.47's password:
TestFileToMoveAround.txt                                      100%   20     3.8KB/s   00:00
PS H:\> cd .\Work\test\
PS H:\Work\test> ls


    Verzeichnis: H:\Work\test


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        13.10.2021     09:27             20 TestFileToMoveAround.txt


PS H:\Work\test> scp .\TestFileToMoveAround.txt mkapun@10.10.0.47:~/Desktop
mkapun@10.10.0.47's password:
TestFileToMoveAround.txt                                      100%   20    12.2KB/s   00:00
PS H:\Work\test>
```

# Tutorials

- Check out: https://github.com/nhmvienna/FirstSteps/blob/main/Bioinformatics/Phyloserver.md

- READ CAREFULLY BEFORE STARTING TO WORK ON PHYLOSERVER2

# (6) hands-on (finally)

Gehe zu: https://github.com/nhmvienna/MysteriousAbominableYeti

- …und leg los!
- **Viel Spaß!!**

# Vielen Dank!

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 31 (2004) 1–3
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

## Molecular phylogenetic analyses indicate extensive morphological convergence between the "yeti" and primates☆

Michel C. Milinkovitch,[a,b,*] Aldagisa Caccone,[b] and George Amato[c]

[a] *Evolutionary Genetics, Institute of Molecular Biology and Medicine, Free University of Brussels (ULB), CP 300, B-6041 Gosselies, Belgium*
[b] *Molecular Systematics and Conservation Genetics, Yale University, New Haven, CT 06520-8106, USA*
[c] *Wildlife Conservation Society, New York 10460, USA*

Received 5 January 2004

## Conclusions

All our analyses clearly indicate that the yeti is nested several nodes within a specific ungulate group (*i.e.*, the perissodactyls, cf. Fig. 1) and, more specifically, forms a subclade with sequences U02581 and X79547 (cf. figure legend). These results demonstrate that extensive morphological convergences have occurred between the yeti and primates.

**It is quite remarkable that Haddock already identified many years ago the correct phylogenetic position of the yeti (despite he had seen only footprints in the snow) when he yelled at it** *"You odd-toed ungulate !" (Hergé, 1960, p. 26)*.