



"Maybe I should have stuck with Excel."

# Bioinformatics workshop

Eine kurze Einführung zum *Phyloserver2*

# 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](https://nhmvienna.github.io)



natural history museum vienna

Official Github page

[View My GitHub Profile](#)

Download  
ZIP File

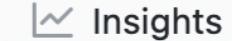
Download  
TAR Ball

View On  
GitHub

# 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](#)

File/Folder	Last Update
 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](#) [*Hpu*; 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 ... 3bcf98 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

**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](#)

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](#)

Let's have a look:

<https://github.com/nhmvienna>

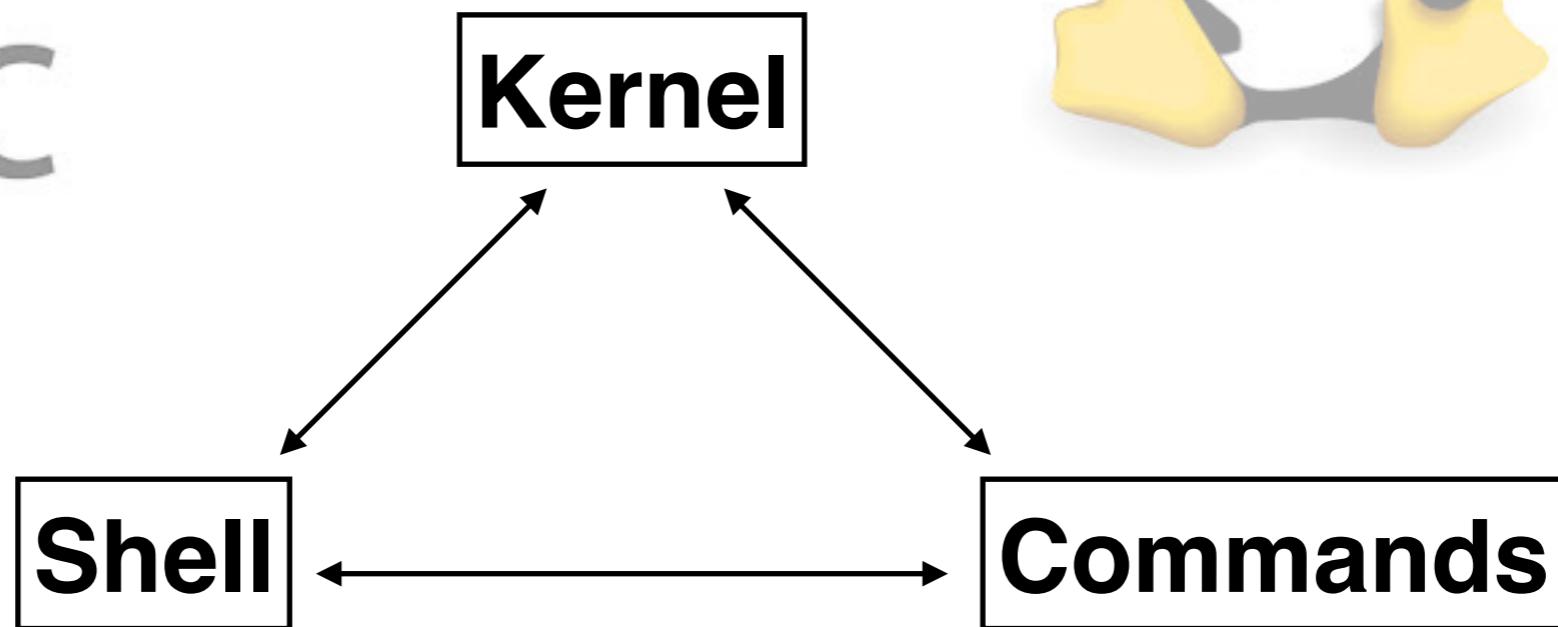
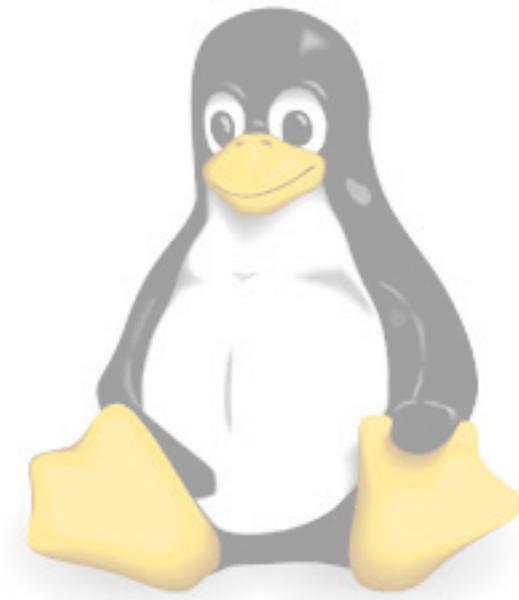
All tutorials for bioinformatics  
at the NHM can be found here:

<https://github.com/nhmvienna/FirstSteps>

# (2) UNIX



Mac



# Was ist UNIX?

## **Kernel:**

- master control program of the operating system

## **Shell:**

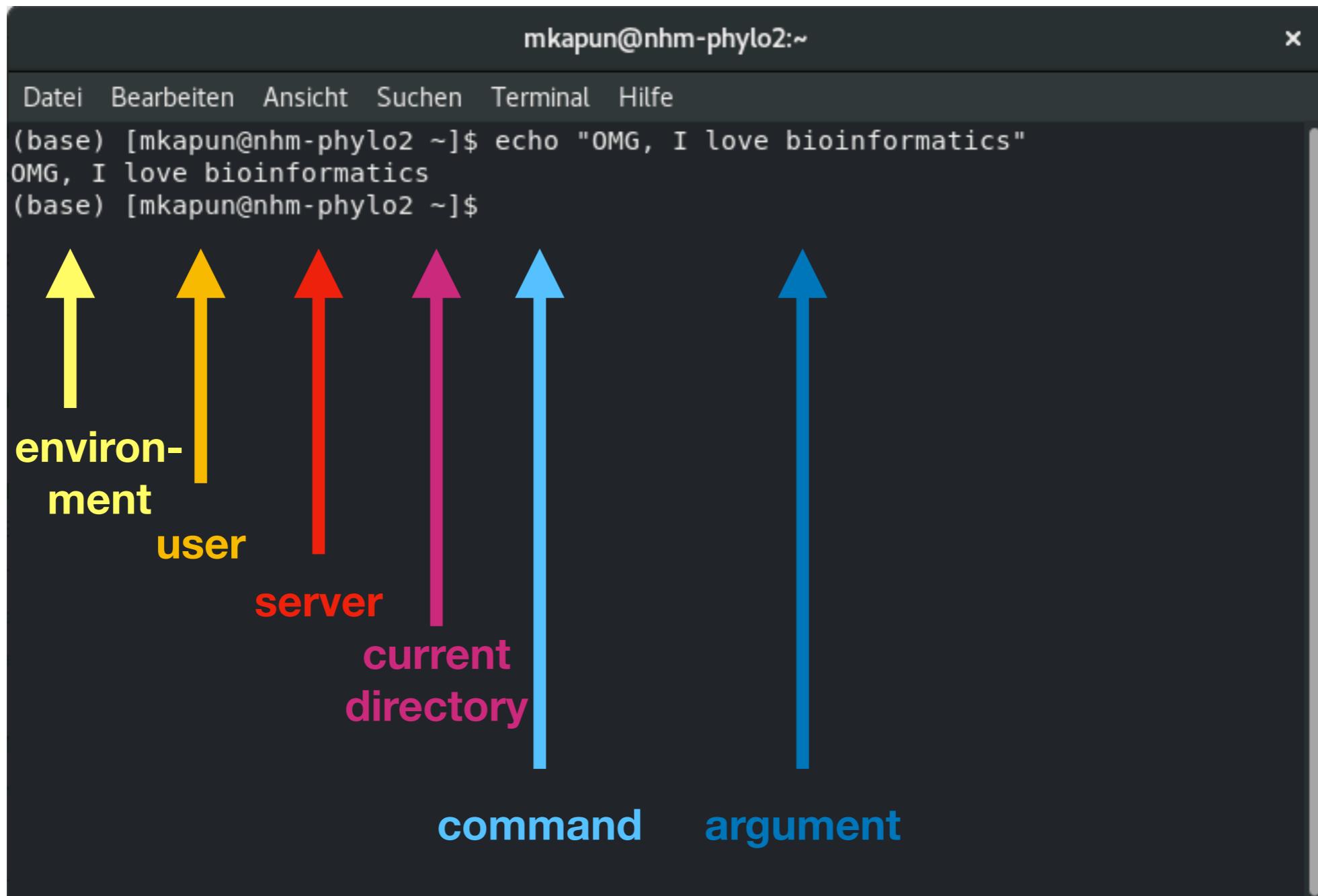
- interactive program that provides an interface between the user and the kernel.
- interprets commands entered by the user or supplied by a shell script, and passes them to the kernel for execution.

# 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

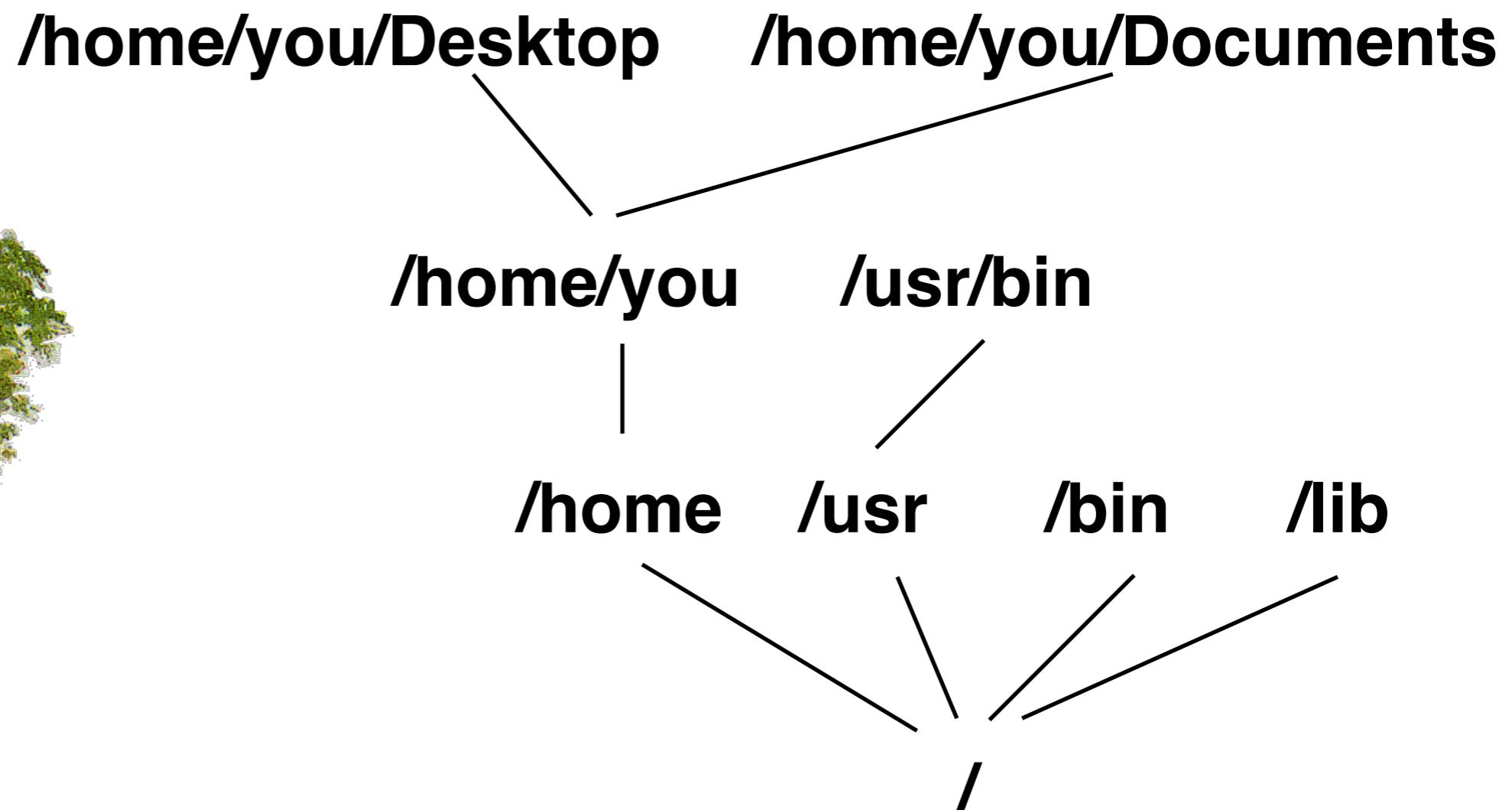


# The bash shell

## Tricks And Tips

- UNIX is case-sensitive: LS is not the same as ls
- bash “remembers” your 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
<code>ls</code>	<b>list files and directories</b>
<code>ls -a</code>	<b>list all (hidden) files and directories</b>
<code>mkdir</code>	<b>make a directory</b>
<code>cd directory</code>	<b>change to named directory</b>
<code>cd</code>	<b>change to home-directory</b>
<code>cd ~</code>	<b>change to home-directory</b>
<code>cd ..</code>	<b>change to parent directory</b>
<code>pwd</code>	<b>display the path of the current directory</b>

# ... and files

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

# Input/Output

command	description
<code>command &gt; file</code>	<b>redirect standard output to a file</b>
<code>command &gt;&gt; file</code>	<b>append standard output to a file</b>
<code>command &lt; file</code>	<b>redirect standard input from a file</b>
<code>command1   command2</code>	<b>pipe the output of command1 to the input of command2</b>
<code>cat file1 file2 &gt; file0</code>	<b>concatenate file1 and file2 to file0</b>

# Tutorials

- Check out: [https://github.com/nhmvienna/FirstSteps/  
blob/main/UNIXBasics/UNIXBasics.md](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 ← script name
#
#
# Created by Martin Kapun on 23/06/14. ← time stamp
#
## 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)
```

comment starting with a # symbol

# Documentation

## Master shell script

```
##### Master.sh
#####
## in all datasets the individuals 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
# individuals 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/overrepresented_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 individuals:
## 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 calculated 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\_FDR00  
1\_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

CountUCEs.sh — /media/inter/mkapun/projects/EchinoUCE — Atom

File Edit View Selection Find Packages Help

Project

EchinoUCE

- .git
- data
- final
- log
- results
- scripts
- shell
- BaitToPhylogeny.sh
- compareMappers.sh
- compareMappers2.sh
- CompareSE\_PE.sh
- CountUCEs.sh
- FilterByPolymorphCov.sh
- GetData.sh
- InSilicoBaitTest\_completeMatrix.sh
- InSilicoBaitTest\_full\_v1.sh
- InSilicoBaitTest\_full\_v2.sh
- InSilicoBaitTest.sh
- InSilicoBaitTest2.sh
- master.sh
- NucleotideContent.sh
- organize.sh
- prepareGenomes.sh
- ProbeDesign.sh
- ProbeDesign2.sh
- ProbeDistribution.sh
- runwithNoClip.sh
- SeaUrchinPhylo\_210914.sh
- SummarizeBaits.sh
- SummarizeBaits2.sh

CountUCEs.sh

```
1  ##### count the number of UCE's per Taxa and the number of UCEs
2  share across taxa
3
4  ## make output directory
5  mkdir /media/inter/mkapun/projects/EchinoUCE/results/countUCEs
6
7  ## 1) Baits based on Ajap genome
8
9  ## make a comma-separated FileList of the combined FASTA files for
10 * the different mapping approaches which is used for MAFFT alignment
```

(base) [mkapun@nhm-phylo2 EchinoUCE]\$ ##### count the number of UCE's per Taxa and the number of UCEs share across taxa  
(base) [mkapun@nhm-phylo2 EchinoUCE]\$

Git Outline GitHub

EchinoUCE

Unstaged Changes shell/CountUCEs.sh

Staged Changes Unstage All

No changes

See All Staged Changes

Commit message

Commit to master 72

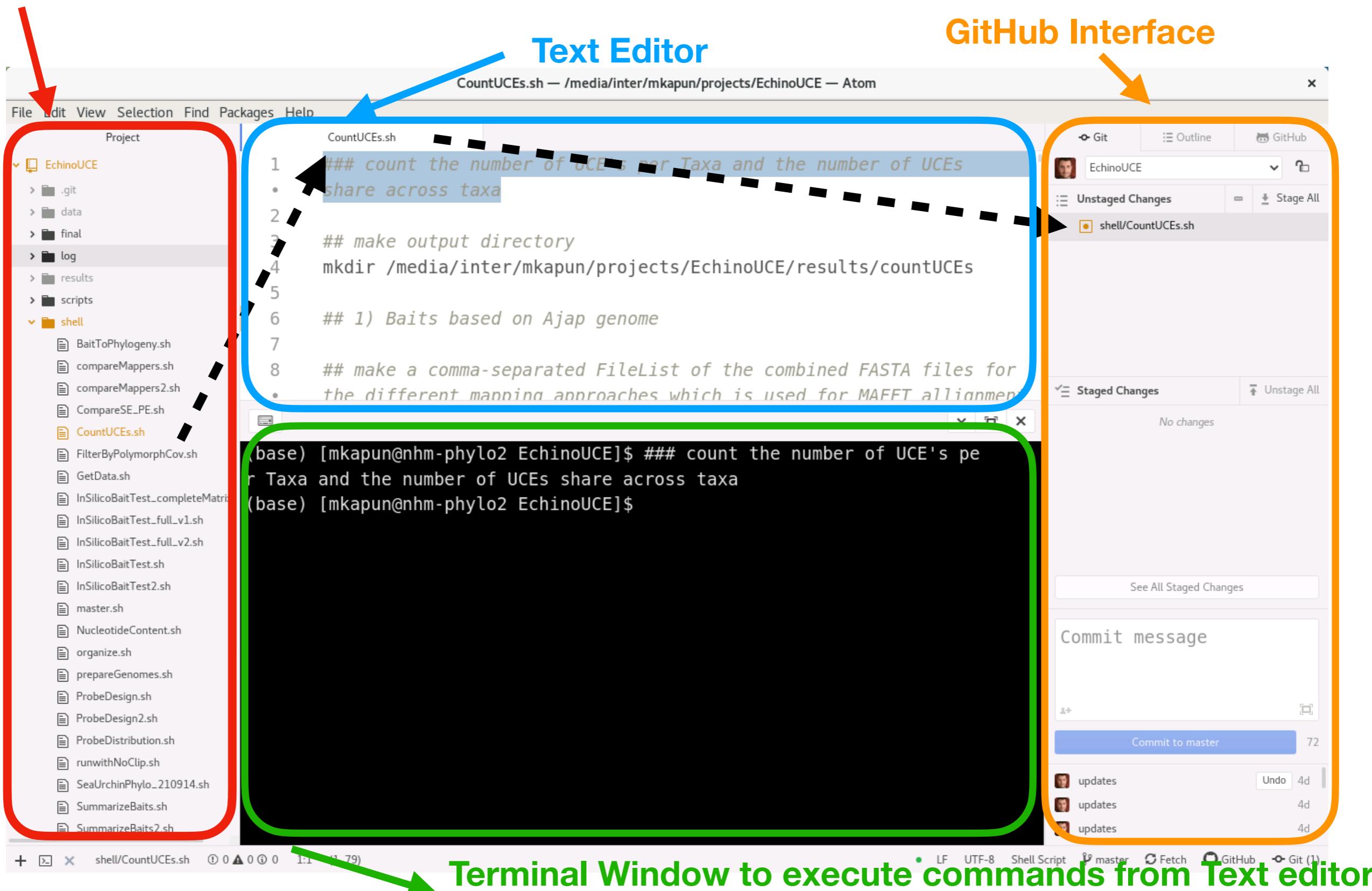
updates 4d

updates 4d

updates 4d

+ X shell/CountUCEs.sh ① 0 ▲ 0 ① 0 1:1 (1, 79) • LF UTF-8 Shell Script master Fetch GitHub Git (1)

# Project Folder Viewer Atom editor



# Tutorials

- Check out: [https://github.com/nhmvienna/FirstSteps/  
blob/main/ATOMbasics.md](https://github.com/nhmvienna/FirstSteps/blob/main/ATOMbasics.md)

# (5) Hilfe mit Bioinformatik

<https://github.com/nhmvienna/BioinformaticsService>

The screenshot shows a GitHub repository page for 'BioinformaticsService'. The repository is private, as indicated by the 'Private' badge. The navigation bar includes links for Pull requests, Issues, Marketplace, and Explore. The main content area shows a commit from 'capoony' with the message 'Initial commit'. Below the commit, there are links for 'LICENSE' and 'README.md', both of which show 'Initial commit' and were made '3 months ago'. On the right side, there are sections for 'About', 'Readme', 'GPL-3.0 License', 'Releases', and 'Packages'. The 'Issues' tab is highlighted with a red box and an arrow points to it with the text 'click here'.

nhmvienna / **BioinformaticsService** Private

Code Issues Pull requests Actions Projects Security Insights Settings

Unwatch 1 Star 0 Fork 0

Code Issues Pull requests Actions Projects Security Insights Settings

main 1 branch 0 tags

click here

Go to file Add file Code

**About**

No description, website, or topics provided.

Readme

GPL-3.0 License

**Releases**

No releases published

Create a new release

**Packages**

No packages published

Publish your first package

# Ticketing

A screenshot of a GitHub repository page for 'nhmvienna / BioinformaticsService'. The page shows the repository details: 'Private', 1 unwatched, 0 stars, 0 forks. The 'Issues' tab is selected. A search bar at the top contains the query 'is:issue is:open'. Below the search bar are filters for 'Labels' (9) and 'Milestones' (0). A prominent green button labeled 'New issue' is highlighted with a red box and an arrow pointing to it from the bottom right. The text 'Open a new issue' is overlaid in red. The main content area features a large 'Welcome to issues!' heading and a descriptive paragraph about using issues for tracking tasks.

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

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

The screenshot shows the GitHub interface for creating a new issue in the repository `nhmvienna/BioinformaticsService`.

**(1) Choose a comprehensive title**: The title field contains the text "Help with phylogenetic analysis using the R package ape".

**(2) Accurately describe your issue**: The issue body contains a message to Martin asking for help with plotting a tree in Newick format using the R package `ape`. It also mentions that the script is attached and thanks Martin for his help.

**(3) Assign label**: The labels sidebar shows a list of available labels: `help wanted` (selected), `bug`, `documentation`, `duplicate`, and `enhancement`.

**(4) Submit new issue**: The green "Submit new issue" button is highlighted.

Annotations:

- Red box around the title input field.
- Red box around the issue body text area.
- Red box around the "Submit new issue" button.
- Red box around the labels sidebar.

# Ticketing

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

Unwatch 1 Star 0 Fork 0

Code Issues 1 Pull requests Actions Projects Security Insights Settings

## Help with phylogenetic analysis using the R package ape #7

Closed capoony opened this issue 3 minutes ago · 1 comment

capoony commented 3 minutes ago

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](#)

capoony added the help wanted label 3 minutes ago

capoony commented 1 minute ago

Thanks, Martin!  
Upon closer inspection, it appeared that your script did not contain any meaningful R code.

capoony closed this 1 minute ago

Assignees: No one—assign yourself

Labels: help wanted

Projects: None yet

Milestone: No milestone

Linked pull requests: Successfully merging a pull request may close this issue.  
None yet

Notifications: Customize

Unsubscribe

You're receiving notifications because you're

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

# Ticketing

- Transparente Kommunikation
- Erleichtert Planung
- Erleichtert Statusabfrage

The screenshot shows a GitHub repository page for 'nhmvienna / BioinformaticsService'. The repository is private, has 1 unwatched issue, 0 stars, and 0 forks. The 'Issues' tab is selected, showing 2 open issues. The search bar filters for 'is:issue is:open'. The issues listed are:

- Another error :-)**  
#4 opened now by capoony
- Need help with Python script**  
#3 opened 40 seconds ago by capoony

Navigation and filtering options include: Filters (dropdown), Labels (9), Milestones (0), New issue, Author dropdown, Label dropdown, Projects dropdown, Milestones dropdown, Assignee dropdown, and Sort dropdown.

# Tutorials

- Check out: [https://github.com/nhmvienna/  
BioinformaticsService](https://github.com/nhmvienna/BioinformaticsService)

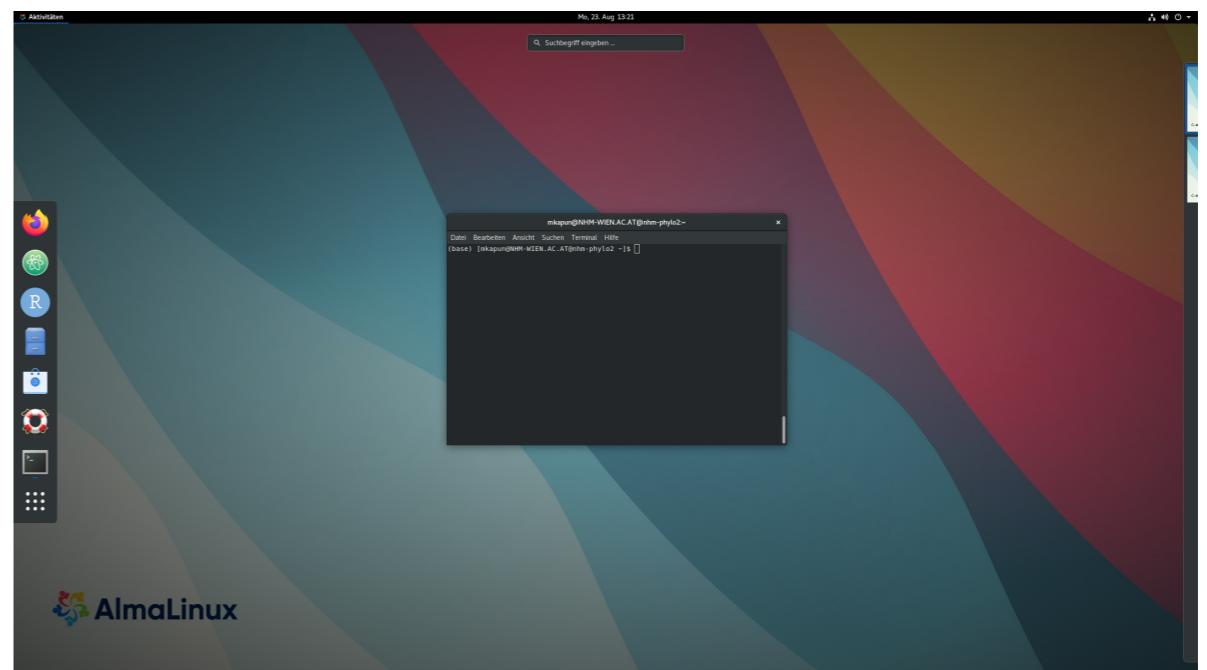
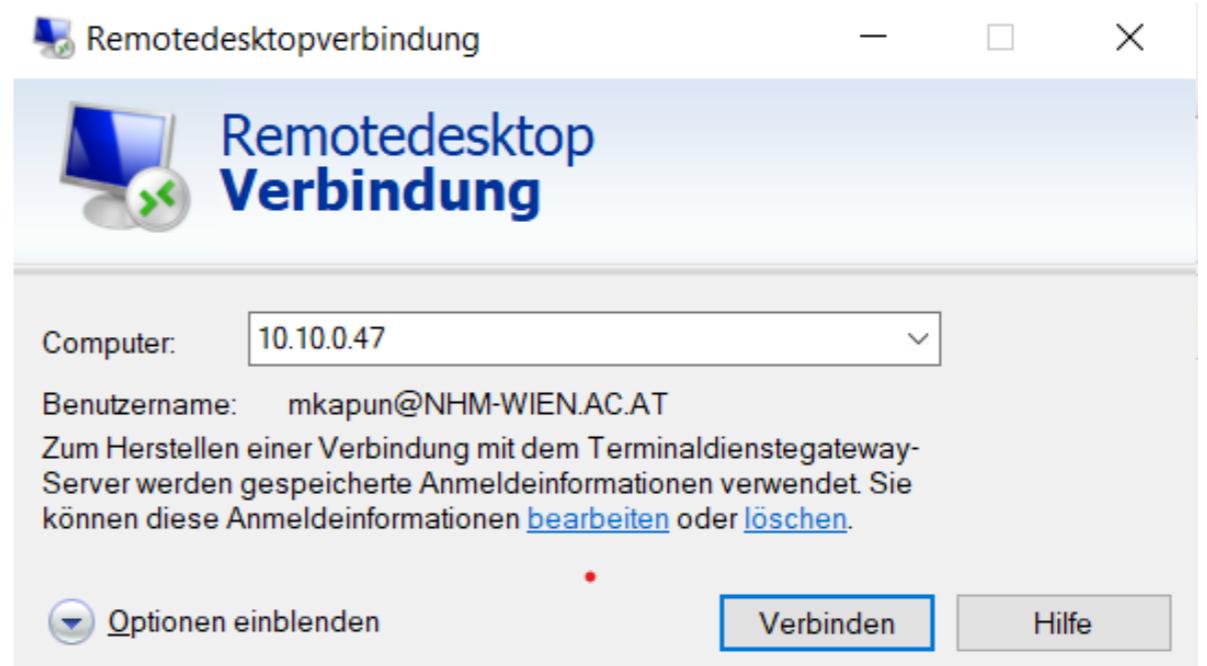
# (6) Phyloserver2

- 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 Installationspfad muss erst angegeben werden, z.B.:

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

```
conda activate mafft-7.487
```

# Transparente Dokumentation

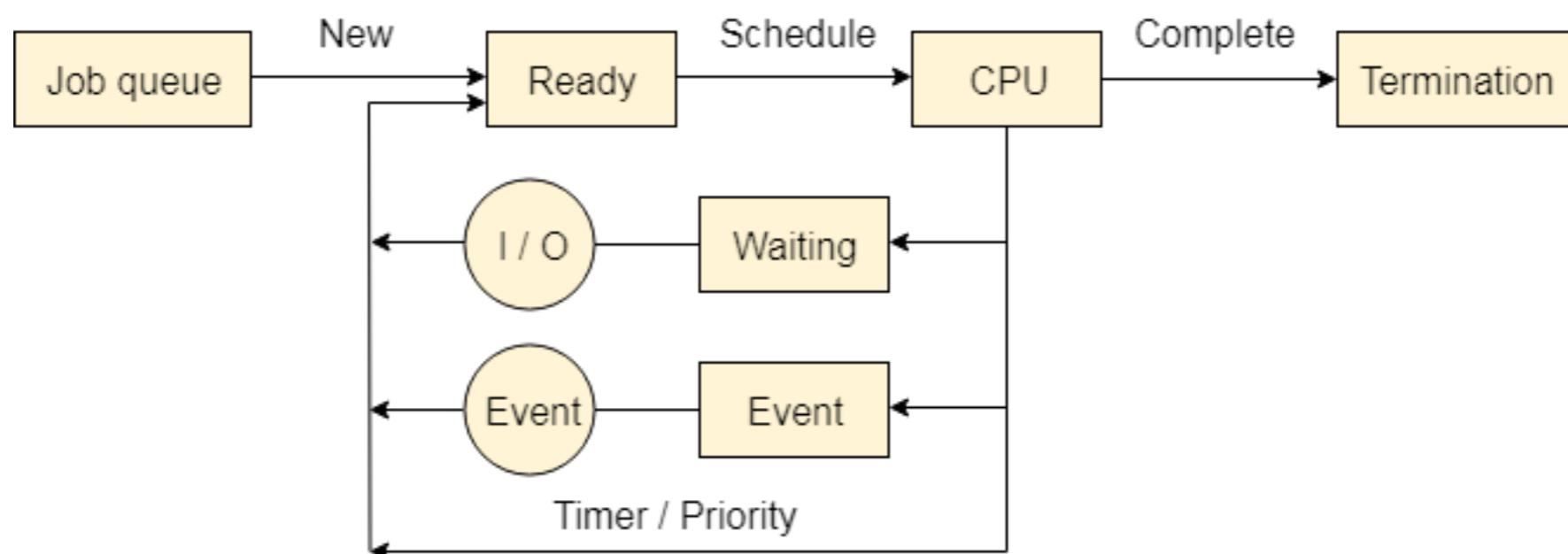
- Sämtliche Installationen auf dem Phyloserver2 sind auf GitHub dokumentiert.
- [https://github.com/nhmvienna/  
PhyloserverInstallationDocs](https://github.com/nhmvienna/PhyloserverInstallationDocs)

# 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.

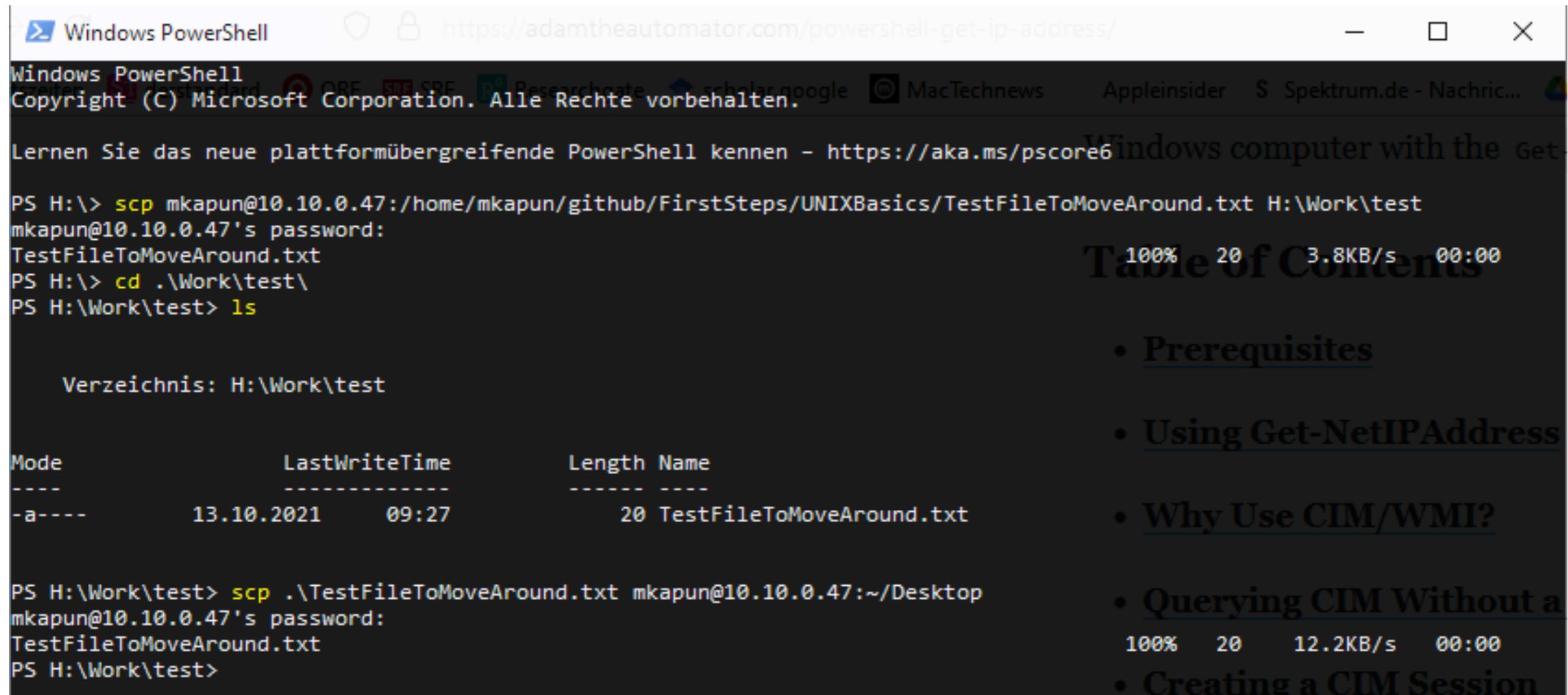
# Job queuing

- 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



A screenshot of a Windows PowerShell window. The title bar says "Windows PowerShell". The URL in the address bar is "https://adamtheautomator.com/powershell-get-ip-address/". The PowerShell window shows the following command and its execution:

```
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  
PS H:\> cd .\Work\test\  
PS H:\Work\test> ls
```

The output shows a single file "TestFileToMoveAround.txt" with the following details:

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

Below this, another SCP command is shown:

```
PS H:\Work\test> scp .\TestFileToMoveAround.txt mkapun@10.10.0.47:~/Desktop  
mkapun@10.10.0.47's password:  
TestFileToMoveAround.txt  
PS H:\Work\test>
```

On the right side of the slide, there is a vertical list of bullet points:

- Prerequisites
- Using Get-NetIPAddress
- Why Use CIM/WMI?
- Querying CIM Without a
- Creating a CIM Session

# Tutorials

- Check out: [https://github.com/nhmvienna/FirstSteps/  
blob/main/Bioinformatics/Phylosolver.md](https://github.com/nhmvienna/FirstSteps/blob/main/Bioinformatics/Phylosolver.md)
- READ CAREFULLY BEFORE STARTING TO WORK ON  
**PHYLOSERVER2**

# (6) hands-on (finally)

Gehe zu: [https://github.com/nhmvienna/  
MysteriousAbominableYeti](https://github.com/nhmvienna/MysteriousAbominableYeti)

- ...und leg los!
- **Viel Spaß!!**



# Vielen Dank!



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



ELSEVIER

Molecular Phylogenetics and Evolution 31 (2004) 1–3

---

MOLECULAR  
PHYLOGENETICS  
AND  
EVOLUTION

---

[www.elsevier.com/locate/ympev](http://www.elsevier.com/locate/ympev)

Molecular phylogenetic analyses indicate extensive morphological convergence between the “yeti” and primates<sup>☆</sup>

Michel C. Milinkovitch,<sup>a,b,\*</sup> Aldagisa Caccone,<sup>b</sup> and George Amato<sup>c</sup>

<sup>a</sup> Evolutionary Genetics, Institute of Molecular Biology and Medicine, Free University of Brussels (ULB), CP 300, B-6041 Gosselies, Belgium

<sup>b</sup> Molecular Systematics and Conservation Genetics, Yale University, New Haven, CT 06520-8106, USA

<sup>c</sup> 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).

