# Exercises in Marine Ecological Genetics

## 01. Introduction

- General info and course outline
- Connect to the HPC cluster
- Work on the command line
- Test for HWE in R

Martin Helmkampf

# General course info

- Suggestion: we start at 14:00 and finish at 15:30
- Course language will be English, but questions can always be asked in German
- There will be no tests or grades
- Slides will be provided, but please do not post them online
- Use of personal laptops is recommended, but workstations are available
- Contact: martin.helmkampf (at) uni-oldenburg.de

# Objectives

- Apply theory and concepts of population genetics, genomics and DNA barcoding in practice
- Learn to analyze, visualize and interpret real world data
- Learn how to work on a high performance computing cluster (using bash) / a scripting environment (using R)
- Become familiar with the most common data types and file formats

# Course outline

Preliminary, may be subject to change

| Class | Date | Topics | Script |
|---|---|---|---|
| 01 | Apr 11 | Introduction, setup | 01_intro.R |
| – | Apr 18 | Good Friday | |
| 02 | Apr 25 | Hardy-Weinberg equilibrium / $N_e$ | |
| 03 | May 02 | Population structure and gene flow | |
| 04 | May 09 | Genome assembly and metrics | |
| 05 | May 16 | SNPs and population genomics | |
| 06 | May 23 | Recombination and linkage disequilibrium | |
| – | May 30 | Himmelfahrt break | |
| 07 | Jun 06 | Measures of genetic diversity | |
| – | Jun 13 | Selection and Mutation | |
| 08 | Jun 20 | Student presentations – no exercises | |
| 09 | Jun 27 | DNA barcoding | |
| 10 | Jul 04 | Metabarcoding / eDNA | |
| 11 | Jul 11 | Metabarcoding / eDNA II | |

# Required software

- Browser
- Text editor (e.g. Notepad, TextEdit, VSCodium)
- Terminal / ssh client (e.g. git bash, Terminal)

To connect to the high performance computing cluster ROSA, a shh client is required

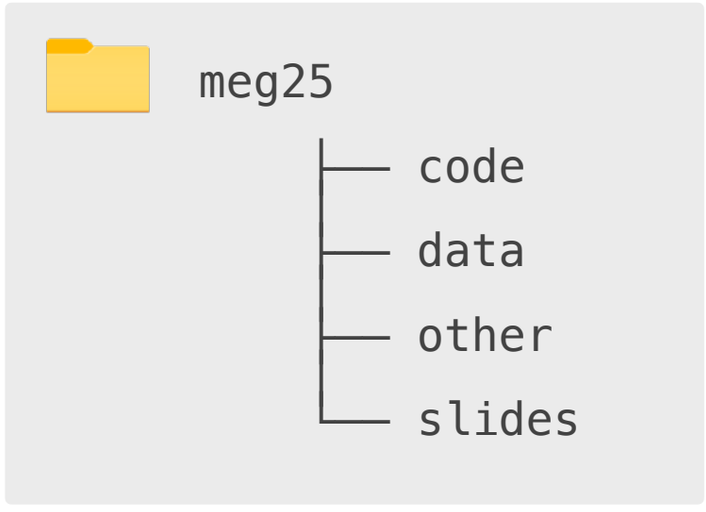
On Windows, a good option is git bash, which is part of Git for Windows. Install from https://gitforwindows.org with default settings (alternative: WSL)

On macOS or Linux, a terminal (and git) come preinstalled

# Accessing the course materials

All code and data for the course will be provided through a Git repository:

https://github.com/mhelmkampf/meg25



# High performance computing on ROSA at UOL


TOP 500 The List.

CERTIFICATE

CARL - Lenovo NeXtScale nx360M5, Xeon E5-2650v4 12C 2.2GHz, Infiniband FDR

Carl von Ossietzky University of Oldenburg, Germany

is ranked

No. 363

among the World's TOP500 Supercomputers

with 457.23 Tflop/s Linpack Performance

in the 48th TOP500 List published at SC16, Salt Lake City, UT on Novem

Congratulations from the TOP500 Editors

Erich Strohmaier
NERSC/Berkeley Lab

University of Tennessee

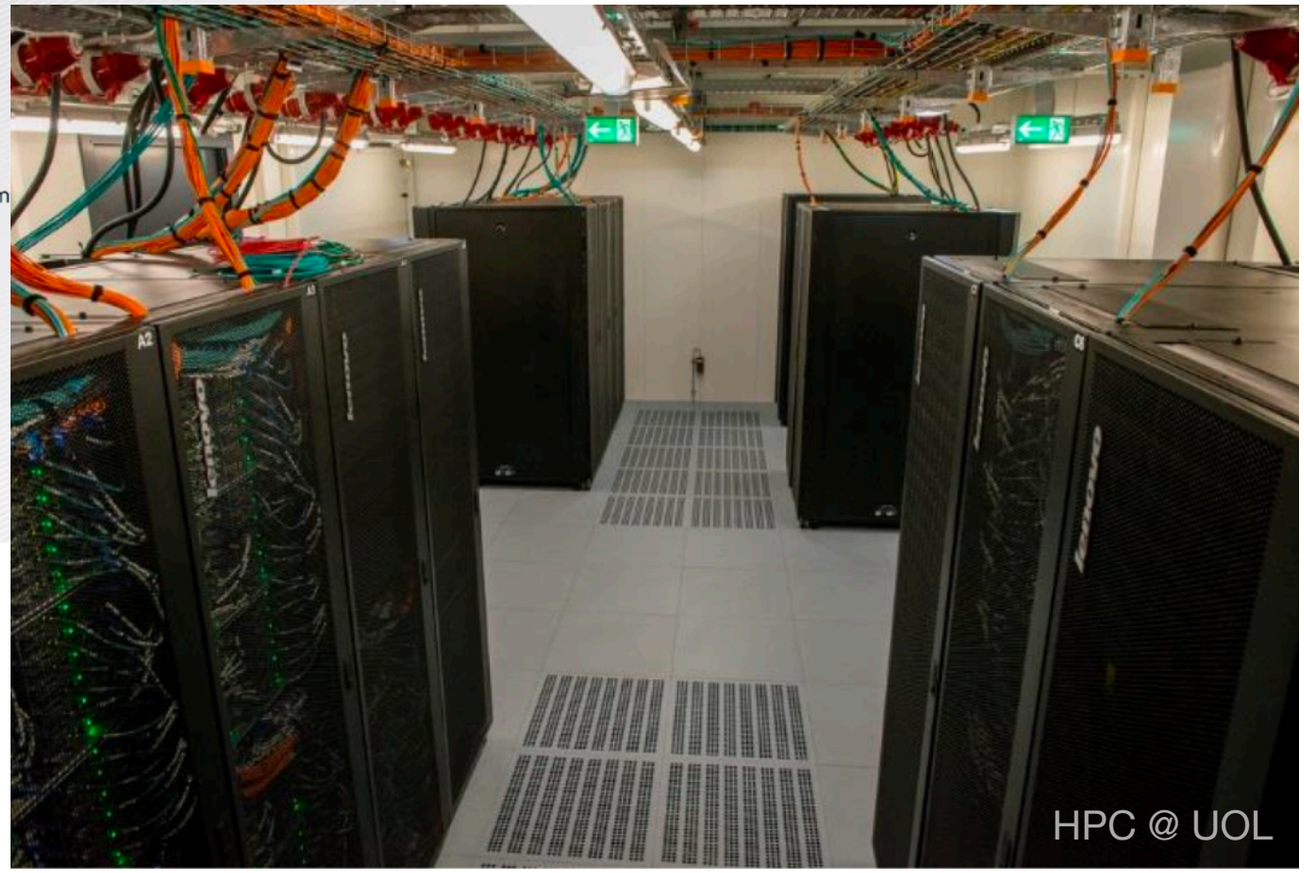
Horst Simon
NERSC/Berkeley Lab

91 compute nodes

11,648 cores (CPUs)

91 TB RAM total

625 TFlop/s ($R_{max}$)


HPC @ UOL

https://hpcwiki.uni-oldenburg.de/

# Advantages of command line / scripting tools

- Highly flexible
- Can be automated and combined into complex workflows
- Reproducible, easy to document
- Can run on high performance computers

# Accessing the cluster from the command line (ssh)

Windows

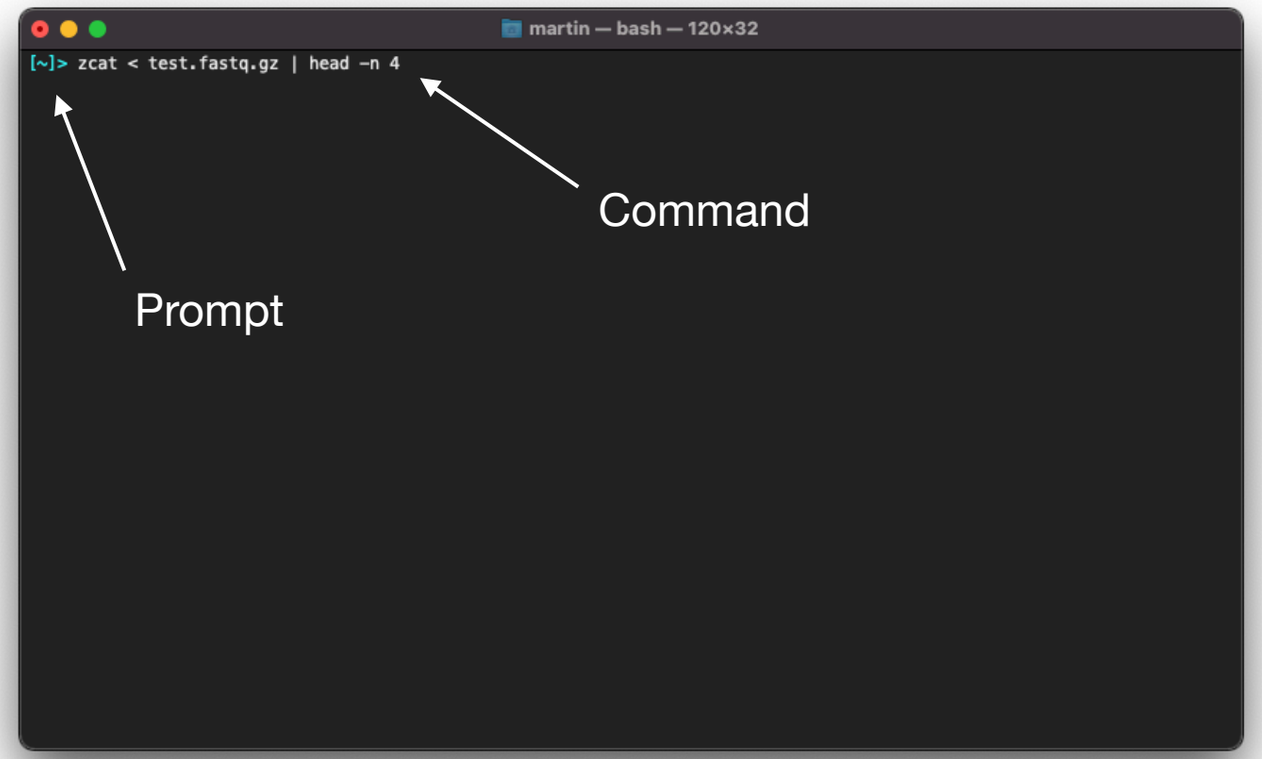
- launch git bash from Start menu
- alternatively, launch terminal in Windows Subsystem for Linux (WSL) on Windows 10 or above (see https://learn.microsoft.com/en-us/windows/wsl/install)

macOS

- open Terminal app
  in /Applications/Utilities,
  type and execute “bash”

Typical usage

```
command [-options] [file]
```



# Get set up on the HPC cluster

Connect to login node

```
# Pick and write down a course account user name and password (passed around)
ssh user1234@rosa.hpc.uni-oldenburg.de
```

Download course materials to cluster account using git

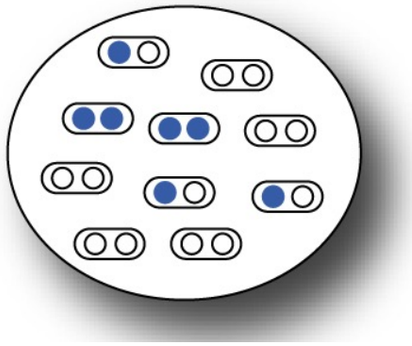
```
git clone https://github.com/mhelmkampf/meg25.git
```

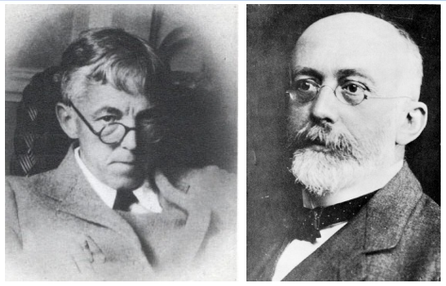
# Is this population in Hardy-Weinberg equilibrium?

Exercise 1

Diploid, sexual reproduction, 1 locus, 2 co-dominant alleles (yellow, blue)

# *HARDY-WEINBERG (1908)*





Godfrey H. Hardy (1877-1947)

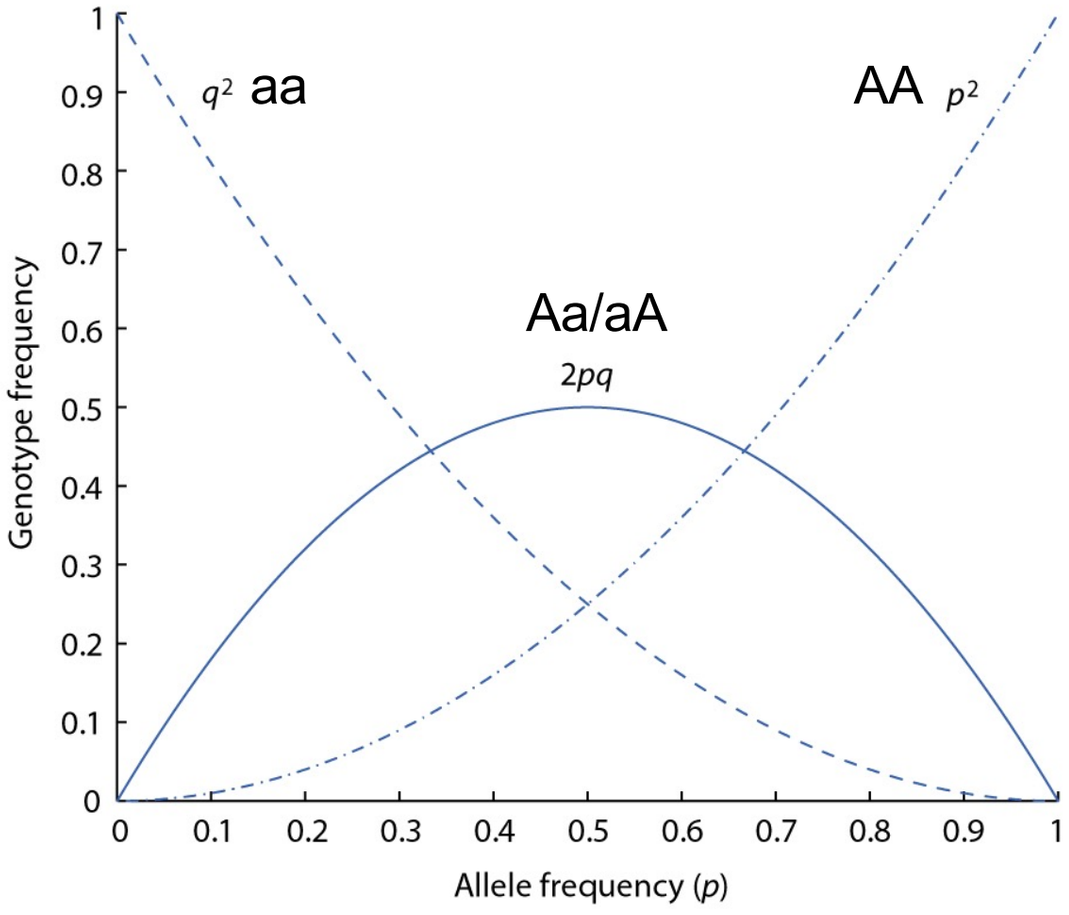
Wilhelm Weinberg (1862-1937)

*Establish the relationship between allele frequencies and genotype frequencies in a population*

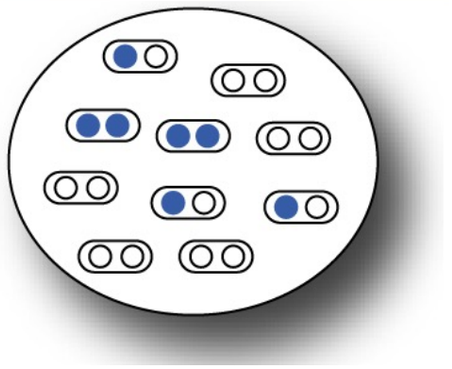
$$p^2 + 2pq + q^2 = 1$$

AA Aa/aA aa

*p* and *q*: allele frequencies for a locus with two alleles (A and a) (*p* + *q* = 1)



# *HETEROZYGOSITY*



In one population

$H_o$ = proportion of heterozygote individuals, observed heterozygosity

$H_e = 2pq = 1 - p^2 - q^2$, expected heterozygosity (assuming HW equilibrium)

$$F = \frac{H_e - H_o}{H_e}$$

Fixation index: *proportion by which heterozygosity is reduced or increased relative to the heterozygosity of a population at HW equilibrium with the same allele frequencies*.

Divided by $H_e$ −> *proportion* (of expected heterozygosity)

Varies between -1 and 1

$F < 0$: heterozygote excess

$F > 0$ heterozygote deficit (homozygote excess)

May be averaged over several loci −> reduces bias

May be extended to *k* alleles

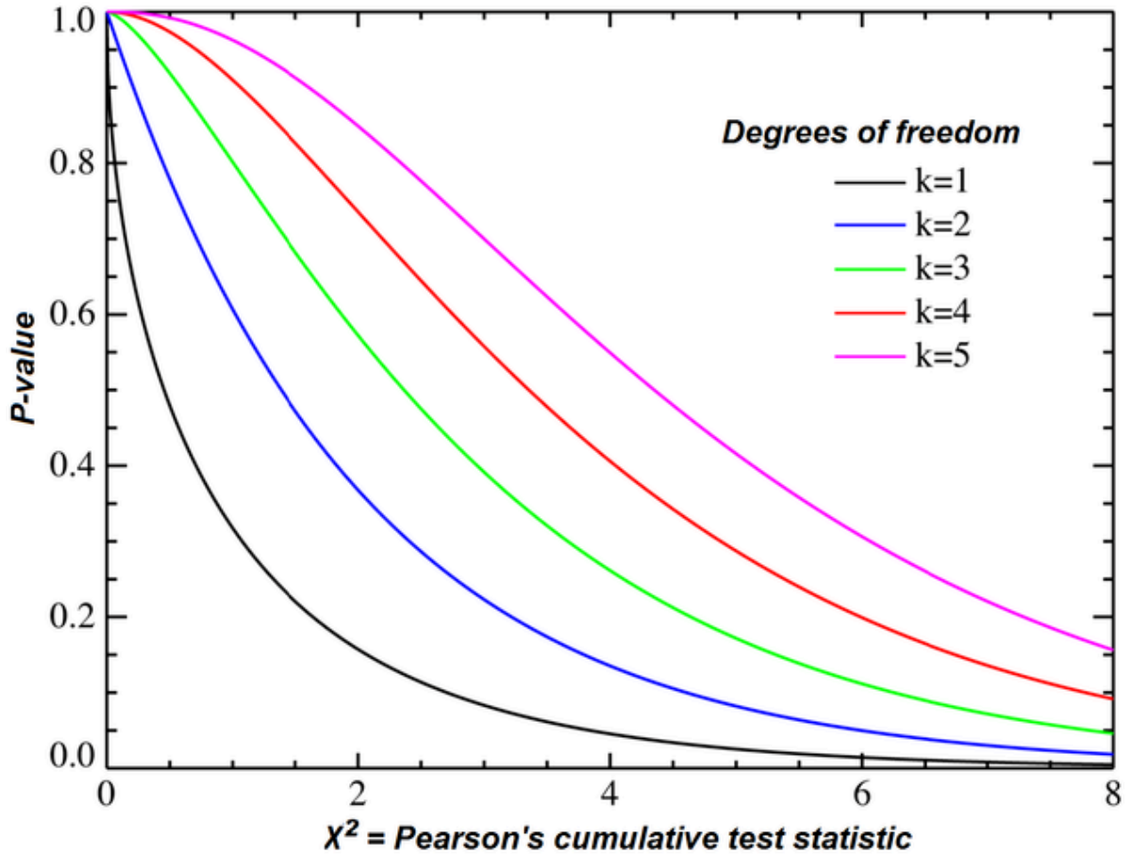
# Pearson’s chi-squared test

Exercise 1

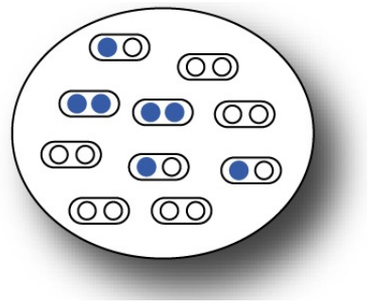
Chi-square statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Chi-square distribution:



A *single generation of reproduction will result in a population that meets the expected Hardy-Weinberg frequencies*, i.e. is at *Hardy-Weinberg* (HW) *equilibrium*



Assuming an “ideal” population, i.e. :

- Diploid organisms
- Sexual reproduction (as opposed to clonal)
- Random mating (as opposed to e.g. assortative) with respect to genotype
- Random union of gametes
- Discrete, non-overlapping generations
- Very large (infinite) population
- No migration
- No population structure
- No natural selection
- Two alleles
- Identical allele frequencies in both sexes

-> Departures from HW equilibrium may indicate:

- Inbreeding
- Assortative mating
- Self-fertilization
- Natural selection
- Population structure
- ...