Utrecht University

A Bayesian inference method to estimate transmission trees with multiple introductions

UU, UMCU & RIVM

Bastiaan van der Roest, Martin Bootsma, Egil Fischer, Don Klinkenberg and Mirjam Kretzschmar

Nordita, June 15 2023

BioRxiv: https://doi.org/10.1101/2023.02.07.527429

Introduction

- ‘Who infected whom’ vital to understand infectious disease outbreaks
  - Understanding risk factors
  - Design targeted intervention strategies
- Multiple transmission routes?
- Single/multiple introduction?

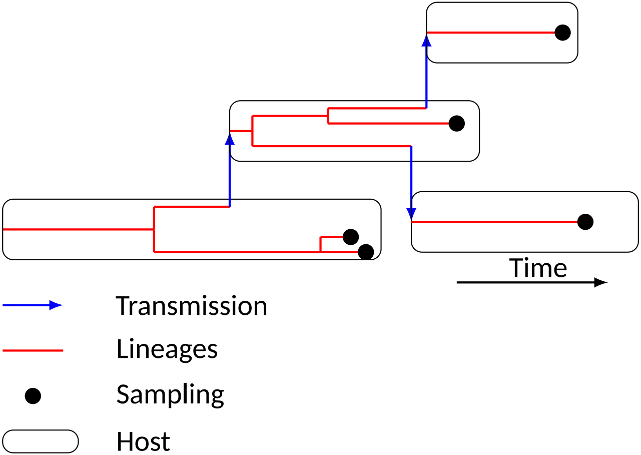
Example: SARS-CoV-2 in mink farms

- SARS-CoV-2 outbreak among humans
- Simultaneously outbreak among mink farms
- Transmission farm→farm or human→farm?

Available data

- Epidemiological data
  - Day of detection
  - Risk factors host
  - (Geographical) distance
  - Generation time interval
- Genetic data of the pathogen (sequence data)
  - Possibly multiple samples from the same host
- Both sources provide information

Phylogenetic tree vs transmission tree



Time

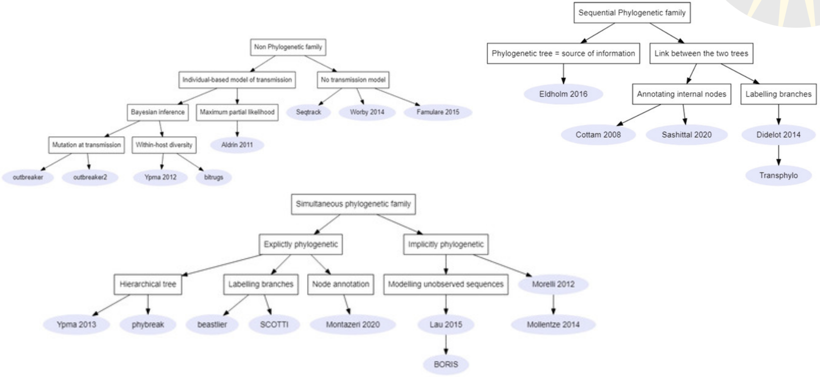
→ Transmission

— Lineages

● Sampling

▭ Host

Existing methods



Duault et al. Pathogens, 2022

Short description likelihood model

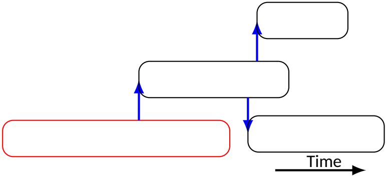
- Likelihood based: 4 components
  1. Epidemiological
     - Chance of exact transmission times,modes, infectors…
  2. Detection
     - Chance that detections occurred at observed moments
  3. Phylogenetic tree
     - How are the observed sequences related to each other
  4. Genetic likelihood
     - Chance of mutation over phylogenetic tree that match observations

Klinkenberg et al. PLoS Comp Biol. 2017

Short description likelihood model

- Likelihood based: 4 components
    1. Epidemiological likelihood

Epidemiological likelihood: Basic version

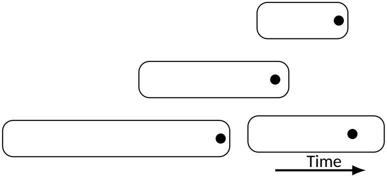


- N infected host
- Index case known ($i = 1$)
- Suppose we know infector M_i of each host $1 \leq i \leq N$ ($M_1 = 0$)
- Infection times I_i known.
- Let $\Lambda(t)$ be the infection pressure at time t.
- Let $\Lambda_i^j(t)$ be infection pressure on host i by host j at time t.
- $L = e^{-\int\limits_0^T \Lambda(\tau)d\tau} \prod\limits_{i:M_i \geq 0} \Lambda_i^{M_i}(I_i)$
- Model needed for infection pressure (transmission)

Short description likelihood model

- Likelihood based: 4 components
    1. Epidemiological likelihood
    2. Detection

Detection likelihood

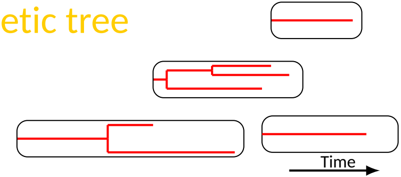


- First sample per host determines moment detection
- Other samples only relevant for genetic likelihood.
- Simplest setting: (time detection - time infection) i.i.d.
- Time dectection - time infection: Gamma-distribution (μ_d, σ_d).

Short description likelihood model

- Likelihood based: 4 components
    1. Epidemiological likelihood
    2. Detection
    3. Phylogenetic tree

Phylogenetic tree



- Within-host dynamics $w(\tau, r)$ describes the product of pathogen generation time and effective population size,
- Inverse of $w(\tau, r)$ determines coalescent rate.
- We choose $w(\tau, r) = r\tau$.
- Thus, the likelihood for the phylogenetic tree in host i becomes

$$P(P_i|S_i, I, M, \theta) = e^{-\int_0^\infty \binom{L_i(\theta)}{2} \frac{1}{w(\tau,r)} d\tau} \prod_{x|x\in P_i \& n<x<2n} \frac{1}{w(\tau_x, r)}$$

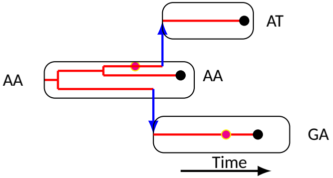
Likelihood full phylogenetic tree:

Product of likelihood phylogenetic tree per host

Short description likelihood model

- Likelihood based: 4 components
    1. Epidemiological likelihood
    2. Detection
    3. Phylogenetic tree
    4. Genetic Likelihood

Genetic likelihood



- Jukes Cantor model
- Constant mutation rate: all mutations equally likely

$$P(G|P,\theta) = \prod_{loci} \sum_{\{A,C,T,G\}^{3n-1}} \prod_{x} \left(\frac{1}{4} - \frac{1}{4}e^{-\mu(t_x - t_{v_x})}\right)^{l_{mut}(1-g)} \cdot \left(\frac{1}{4} + \frac{3}{4}e^{-\mu(t_x - t_{v_x})}\right)^{1-l_{mut})(1-g)}$$

∀ locus, all possible mutations are calculated

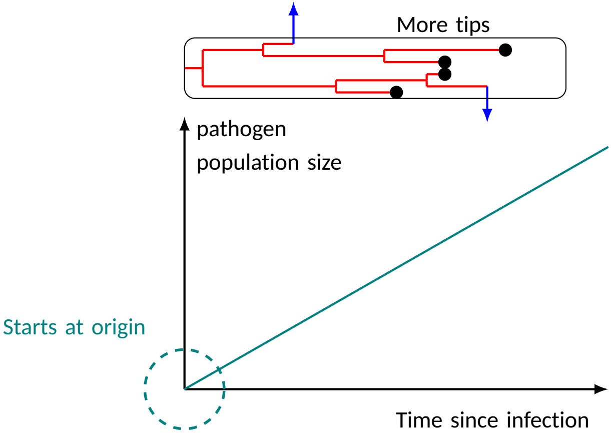
- likelihood calculated using Felsenstein’s pruning algorithm.
- actual rate of nucleotide change is 0.75μ.

Limitations

- one phylogenetic tree
  - by linking ’minitrees’ in each host
- Consequence of choice
  - clonal evolution (point mutations)
  - no recombination/reassortment
  - restriction on data that can be handled by phybreak!

The phybreak model

- Multiple samples/patient, strict bottleneck



The phybreak model

- Wide bottleneck

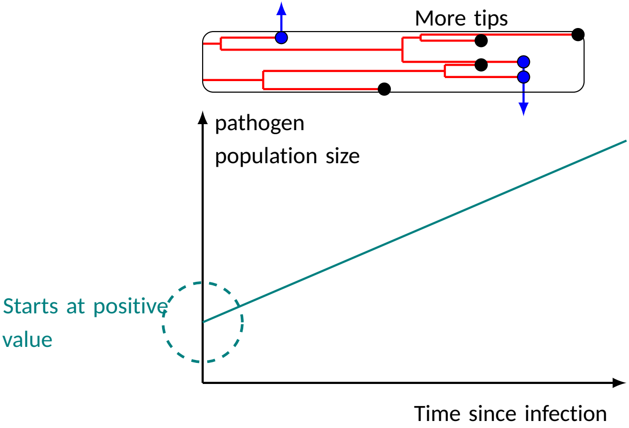
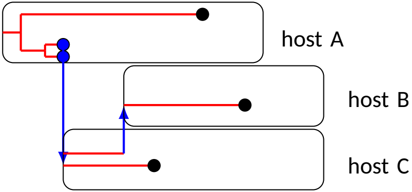


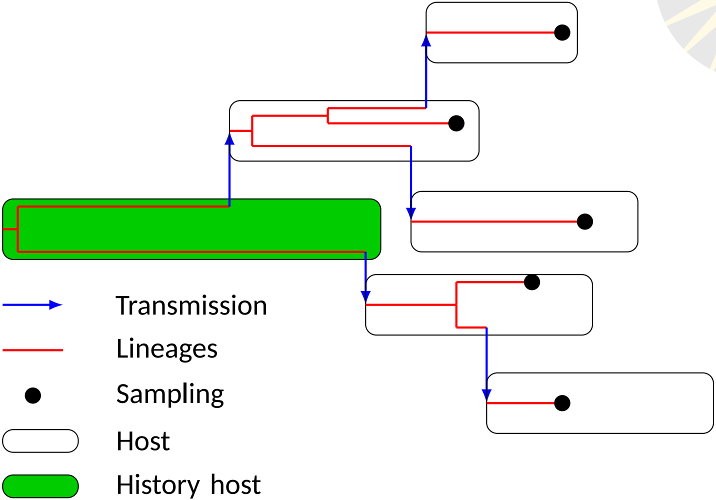
Illustration of uncertainty: ghost infector



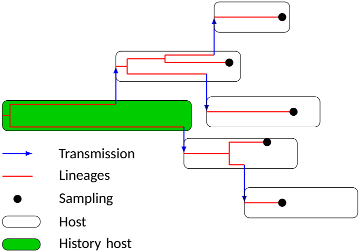
What is the aim of the new model?

- Allow for multiple introductions in the host population
  - SARS-CoV-2 outbreak among mink farms co-occurred with human pandemic
- Determine transmission trees of each introduction
- Determine epidemiological parameters
  - R_0, generation time interval, effectiveness of interventions

History host



History host



- If sequences very divers (compared to mutation rate)
  - A priori split the outbreak into several distinct outbreaks
- Interesting case: # introductions not obvious based on data
- Concept history host
  - Infector of all index cases
  - Different coalescent rate
  - Time last coalescent: Time till MRCA
  - Introductions not related

Rate of infection from history host important

- Single introduction:
  - Relative infectiousness of host important to determine infector
- Multiple introductions:
  - Relative importance history host vs other hosts important

Force of infection of Poisson process upon host *i*.

$$\Lambda_i(t) = \alpha + \sum_{j \neq i} g_j^\theta(t - I_j) K_{ij}^\theta(t)$$

α: Introduction rate from history host: Spatially homogeneous

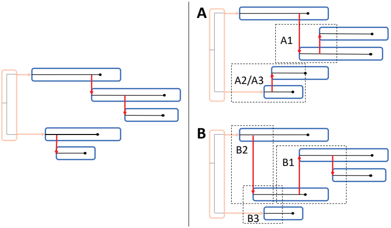
$g_j^\theta(\tau)$: Generation time of host *j* as function of time since infection τ

$K_{ij}^\theta(t)$: 'Connectivity' from host *j* to host *i*.

Numerics

- $(\mathrm{MC})^3$ (Metropolis coupled Markov chain Monte Carlo)
- Starting configuration
- Still numerically expensive

Error types



type A Estimated infector belongs to same cluster as true infector

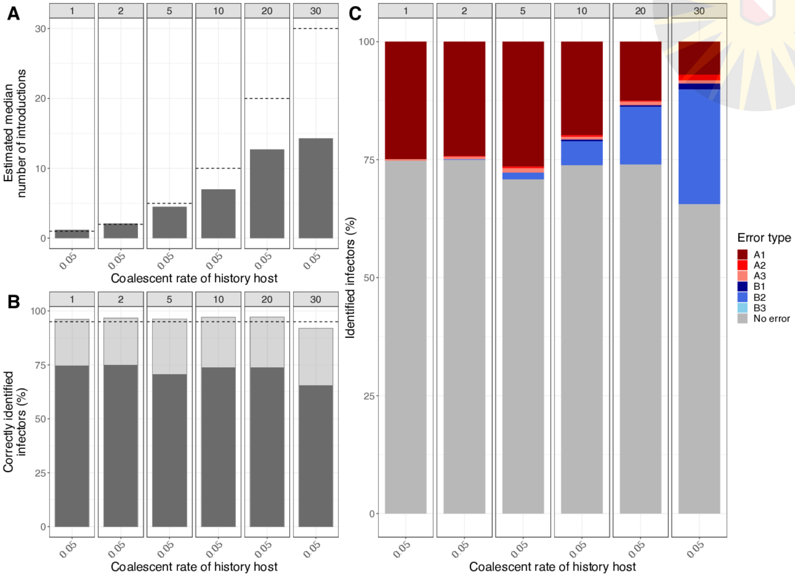
type B Estimated infector belongs to different cluster as true infector

Type 1 Neither true infector nor identified infector is index case.

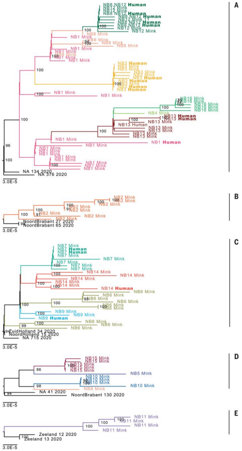
Type 2 Host index in simulated, not in estimated outbreak.

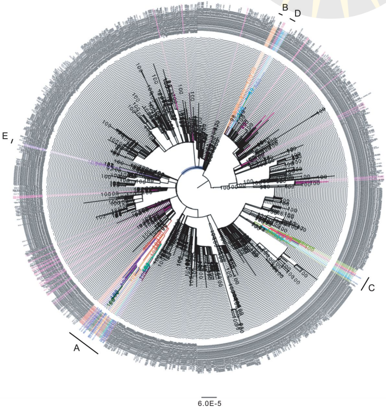
Type 3 Host not index in simulated but index in estimated outbreak

Results Simulation studies: 63 hosts



Data Phylogeny[1]





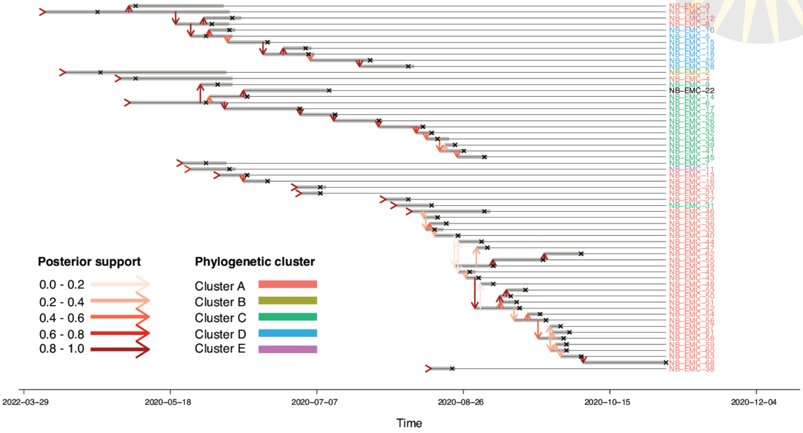
Phylogenetic analysis of 88 mink and 18 mink-related human SARS-CoV-2 sequences in the five mink farm clusters.

Sequences from different farms are depicted in different colors.

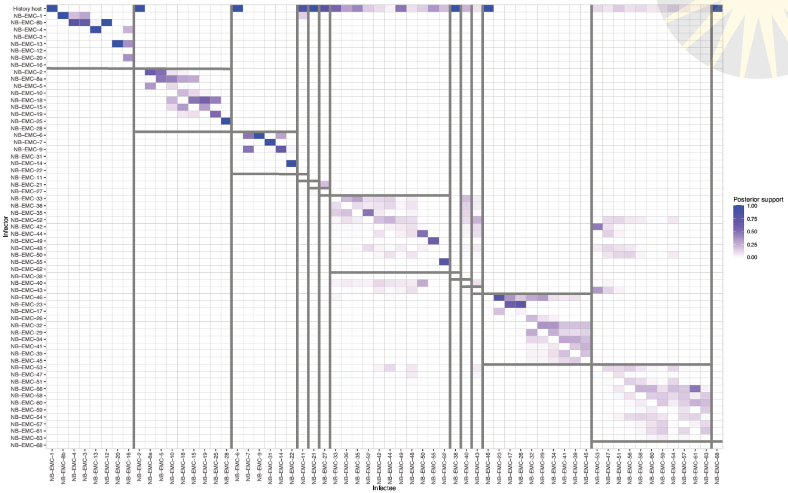
The scale bar represents units of substitutions per site.

[1] Oude Munnik et al Science. 2020

Results SARS-CoV-2 mink farms



Results SARS-CoV-2 mink farms



Discussion

- Results mink farms different from phylogenetic analysis
- Missing cases
- Uninfected population
- Open population (hospital)
- More detailed description of history host

Questions?

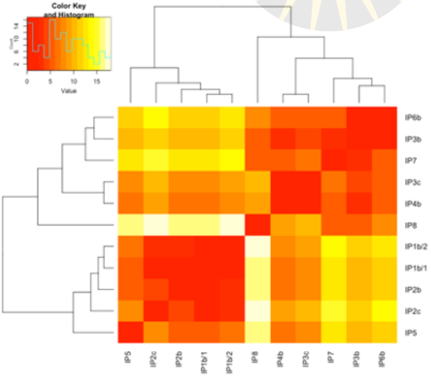
Workshop

Which method to choose?

- Genetic methods:
  - Distance based methods
  - Hierarchical clustering
  - Minimum spanning tree
  - Neighbour-joining tree
  - Evolutionary tree from phylodynamic analysis (BEAST)
- Genetic + Transmission methods
  - Non-phylogenetic
  - Phylogenetic
    - Sequential or simultaneous

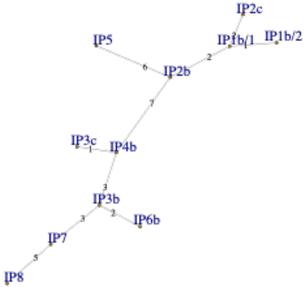
Genetic Methods

- Sequence distance = total number of nucleotides different between two sequences (i.e., number of SNPs between two sequences)
- Hierarchical clustering
  - Building tree by merging samples with closest distance



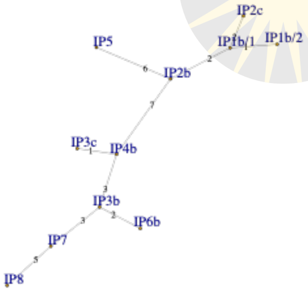
Genetic Methods

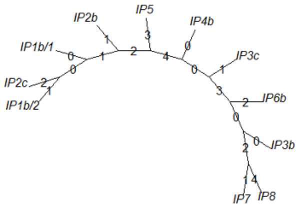
- Minimum spanning tree (MST)
  - Nodes are samples/cases
  - Edges are distances between samples
  - Connecting samples with least overall distance



Genetic Methods

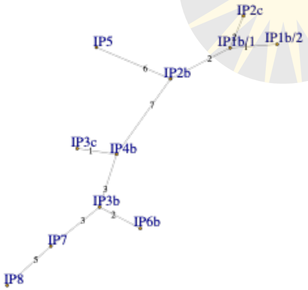
- Minimum spanning tree (MST)
  - Nodes are samples/cases
  - Edges are distances between samples
  - Connecting samples with least overall distance
- Neighbour-joining tree (NJ tree)
  - Tips are samples/cases
  - Nodes are unobserved intermediate sequences
  - Edges are distances between nodes
  - Represent evolutionary relationships between biological sequences





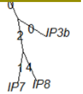
Genetic Methods

- Minimum spanning tree (MST)
  - Nodes are samples/cases
  - Edges are distances between samples
  - Connecting samples with least overall distance



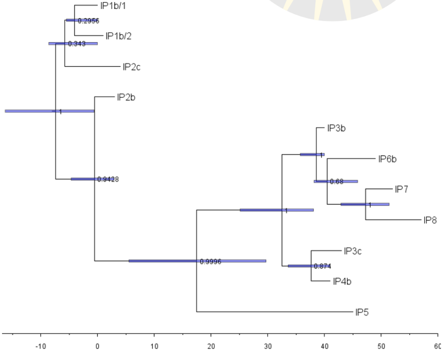
- Neighbour-joining tree (NJ tree)
  - Tips are samples/cases
  - Represent evolutionary relationships between biological sequences

Time not involved



Genetic Methods

- Phylogenetic tree (BEAST)
  - Rooted, time-measured phylogenies
  - Model for mutations
  - Use time of sampling as extra data source



Introduction to phybreak

- Workflow
  - provide data
  - initialise analysis
  - run the analysis
  - inspect and summarize the results

Phybreak: Workflow

- provide data

```
phybreakdata( sequences, sample.times, spatial = NULL,
sample.names = NULL,host.names = sample.names,
culling.times = NULL, external.sequence = FALSE,
sim.infection.times = NULL, sim.infectors = NULL,
sim.tree = NULL )
```

- necessary: sequences and sample times in common R formats
- optional: e.g. host names if there are multiple samples per host

Phybreak: Workflow

- initialise analysis

```
phybreak( dataset, times = NULL, mu = NULL, gen.shape = 3, gen.mean = 1, trans.model = "gamma",
infectivity_file = NULL, sample.shape = 3, sample.mean = 1, multiple.introductions = FALSE,
introductions = 1, intro.rate = 1, wh.model = "linear", wh.bottleneck = "auto", wh.history = 1,
wh.slope = 1, wh.exponent = 1, wh.level = 0.1, dist.model = "power", dist.exponent = 2,
dist.scale = 1, dist.mean = 1, est.mu = TRUE, prior.mu.mean = 0, prior.mu.sd = 100, est.gen.mean
= TRUE, prior.gen.mean.mean = 1, prior.gen.mean.sd = Inf, est.sample.mean = TRUE,
prior.sample.mean.mean = 1, prior.sample.mean.sd = Inf, est.intro.rate = TRUE,
prior.intro.rate.mean = 1, prior.intro.rate.shape = 1, est.trans.growth = TRUE, est.trans.sample
= TRUE, est.wh.slope = TRUE, prior.wh.slope.shape = 3, prior.wh.slope.mean = 1, est.wh.exponent =
TRUE, prior.wh.exponent.shape = 1, prior.wh.exponent.mean = 1, est.wh.level = TRUE,
prior.wh.level.shape = 1, prior.wh.level.mean = 0.1, est.wh.history = TRUE,
prior.wh.history.shape = 1, prior.wh.history.mean = 100, est.dist.exponent = TRUE,
prior.dist.exponent.shape = 1, prior.dist.exponent.mean = 1, est.dist.scale = TRUE,
prior.dist.scale.shape = 1, prior.dist.scale.mean = 1, est.dist.mean = TRUE,
prior.dist.mean.shape = 1, prior.dist.mean.mean = 1, use.tree = FALSE, use.NJtree = TRUE, ... )
```

- oops...

Phybreak: Workflow

- initialise analysis

```
phybreak( dataset, times = NULL, mu = NULL, gen.shape = 3, gen.mean = 1, trans.model = "gamma",
infectivity_file = NULL, sample.shape = 3, sample.mean = 1, multiple.introductions = FALSE,
introductions = 1, intro.rate = 1,wh.model = "linear", wh.bottleneck = "auto", wh.history = 1,
wh.slope = 1, wh.exponent = 1, wh.level = 0.1, dist.model = "power", dist.exponent = 2,
dist.scale = 1, dist.mean = 1, est.mu = TRUE, prior.mu.mean = 0, prior.mu.sd = 100, est.gen.mean
= TRUE, prior.gen.mean.mean = 1, prior.gen.mean.sd = Inf, est.sample.mean = TRUE,
prior.sample.mean.mean = 1, prior.sample.mean.sd = Inf, est.intro.rate = TRUE,
prior.intro.rate.mean = 1, prior.intro.rate.shape = 1, est.trans.growth = TRUE, est.trans.sample
= TRUE, est.wh.slope = TRUE, prior.wh.slope.shape = 3, prior.wh.slope.mean = 1, est.wh.exponent =
TRUE, prior.wh.exponent.shape = 1, prior.wh.exponent.mean = 1, est.wh.level = TRUE,
prior.wh.level.shape = 1,prior.wh.level.mean = 0.1, est.wh.history = TRUE, prior.wh.history.shape
= 1, prior.wh.history.mean = 100, est.dist.exponent = TRUE, prior.dist.exponent.shape = 1,
prior.dist.exponent.mean = 1, est.dist.scale = TRUE, prior.dist.scale.shape = 1,
prior.dist.scale.mean = 1, est.dist.mean = TRUE, prior.dist.mean.shape = 1, prior.dist.mean.mean
= 1, use.tree = FALSE, use.NJtree = TRUE, ... )
```

- ingredients
  - data
  - model options
  - which parameters to estimate
  - prior distributions
  - starting conditions

Taking a step back: the phybreak model

Phybreak: outbreak analysis with sequence data

Seems specific, but still many possible settings…

- Norovirus outbreak during a youth camp
- Legionnaires' disease linked to a water tower
- MRSA on intensive care unit
- H7N7 influenza on poultry farms

…and many possible datasets (even with only sequences)

- environmental samples
- longitudinal sampling
- WGS, MLST, short-reads

Phybreak: model, data, parameters

Narrowing down: what situation are we looking at?

Role of the model:

- the model serves to link the data to what we want to know
- the model is an exact description of how the data were generated: what happened during the outbreak, resulting in the data
  - what we want to know
  - additional stuff
    - could be interesting, or not
    - could be used as a check of model validity

Phybreak: model, data, parameters

Narrowing down: what situation are we looking at?

Role of the model: (specific to phybreak)

- the model serves to link the data to what we want to know
  link sequences + sampling times to who infected whom
- the model is an exact description of how the data were generated: what happened during the outbreak, resulting in the data
  - what we want to know who infected whom
  - additional stuff (e.g. infection times, generation interval distribution, mutation rate)
    - could be interesting, or not
    - could be used as a check of model validity

Phybreak: model, data, parameters

Consequence of "what we want to know" = "who infected whom"

- person-to-person infections (or farm-to-farm)
- all cases should have been observed (or at least, most)
- most cases should have been sampled
- many datasets not suitable for phybreak!

What if you use phybreak nevertheless?

- know how to interpret the results => know the model!!

Phybreak: workflow

- provide data
- initialise analysis
- run the analysis
- inspect and summarize the results

Phybreak: workflow

- run the analysis (= mcmc sampling)

```
 burnin_phybreak( x, ncycles, classic = 0, keepphylo = 0, withinhost_only = 0,
parameter_frequency = 1, status_interval = 10, historydist = 0.5, nchains = 1, heats = NULL, swap
= 1 )

 sample_phybreak( x, nsample, thin = 1, thinswap = 1, classic = 0, keepphylo = 0, withinhost_only
= 0, parameter_frequency = 1, status_interval = 10, verbose = 1, historydist = 0.5, nchains = 1,
heats = NULL, all_chains = FALSE, parallel = FALSE, ... )
```

- burnin_phybreak: running mcmc without keeping samples
  - convergence

Phybreak: workflow

run the analysis (= mcmc sampling)

```
 sample_phybreak(x, nsample, thin = n, thinswap = 1, classic = 0, keepphylo = 0, withinhost_only
= 0, parameter_frequency = 1, status_interval = 10, verbose = 1, historydist = 0.5, nchains = 1,
heats = NULL, all_chains = FALSE, parallel = FALSE, ... )
```

Ingredients:

- initialised model (previous step)
- number of samples to keep (from the mcmc)
- keep every n’th sample (throw away the rest ->mixing)
- technical stuff about how to run the mcmc

Phybreak: workflow

- provide data
- initialise analysis
- run the analysis
- inspect and summarize the results

Phybreak: workflow

- Inspect and summarize the results
  - checking convergence and mixing (use with package ’coda’)
  - get_mcmc(x)
  - ESS(x)

- the ‘mean’ phylogenetic tree and ‘mean’ transmission tree

```
phylotree( x, samplesize = Inf, support = c("proportion", "count"), phylo.class = FALSE )

transtree( x, method = c("count", "edmonds", "mpc", "mtcc"), samplesize = Inf, infector.name = TRUE, support =
c("proportion", "count"), infection.times = c("all", "infector", "infector.sd"), time.quantiles = c(0.025, 0.5, 0.975),
show.besttree = FALSE, phylo.class = FALSE )
```

- the most likely infectors for each case

```
infectorsets( x, which.hosts = "all", percentile = 0.95, minsupport = 0, samplesize = Inf, infector.name = TRUE, support
= c("proportion", "count"), output = c("list", "matrix") )
```

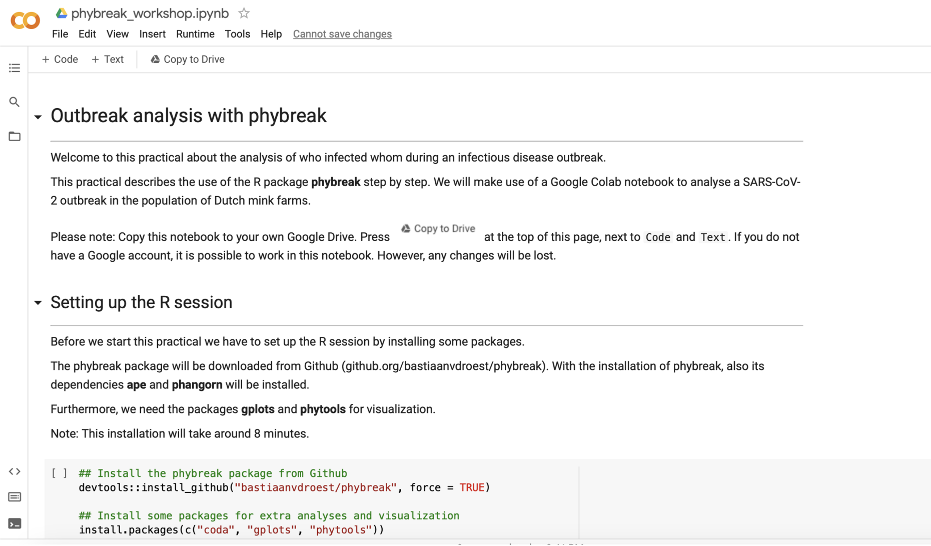
- plotting the ’mean’ phylogenetic and/or transmission tree

```
plotPhylo(x, plot.which = c("sample", "mpc", "mtcc", "mcc"), samplenr = 0, ...)
plotTrans( x, ... )
plotPhyloTrans( x,... )
```

Practical Session using Phybreak

Before we start…

Go to: www.bit.ly/phybreakworkshop_colab



Outbreak analysis with phybreak

Welcome to this practical about the analysis of who infected whom during an infectious disease outbreak.

This practical describes the use of the R package **phybreak** step by step. We will make use of a Google Colab notebook to analyse a SARS-CoV-2 outbreak in the population of Dutch mink farms.

Please note: Copy this notebook to your own Google Drive. Press Copy to Drive at the top of this page, next to `Code` and `Text`. If you do not have a Google account, it is possible to work in this notebook. However, any changes will be lost.

Setting up the R session

Before we start this practical we have to set up the R session by installing some packages.

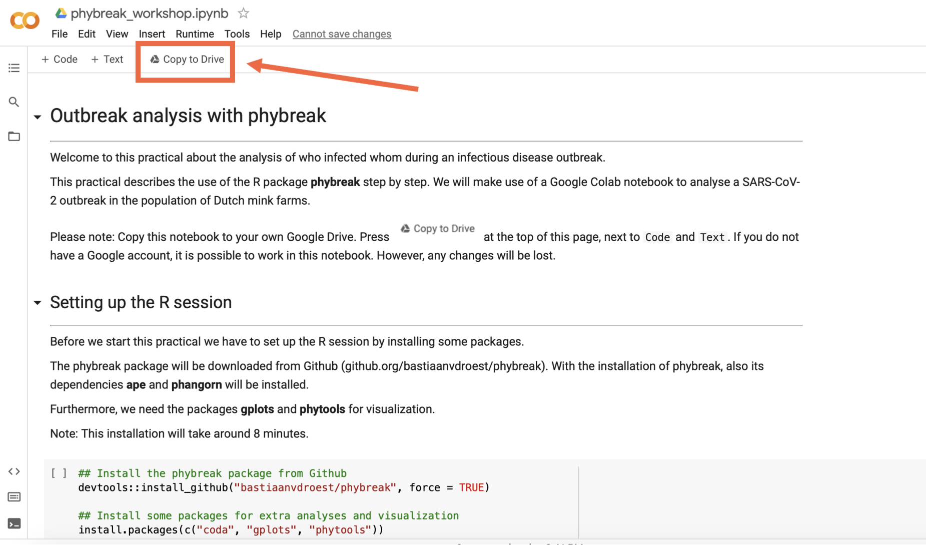
The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

Note: This installation will take around 8 minutes.

```
[ ] ## Install the phybreak package from Github
    devtools::install_github("bastiaanvdroest/phybreak", force = TRUE)

    ## Install some packages for extra analyses and visualization
    install.packages(c("coda", "gplots", "phytools"))
```

www.bit.ly/phybreakworkshop_colab



CO
phybreak_workshop.ipynb
File Edit View Insert Runtime Tools Help Cannot save changes
+ Code + Text
Copy to Drive

Outbreak analysis with phybreak

Welcome to this practical about the analysis of who infected whom during an infectious disease outbreak.

This practical describes the use of the R package **phybreak** step by step. We will make use of a Google Colab notebook to analyse a SARS-CoV-2 outbreak in the population of Dutch mink farms.

Please note: Copy this notebook to your own Google Drive. Press Copy to Drive at the top of this page, next to `Code` and `Text`. If you do not have a Google account, it is possible to work in this notebook. However, any changes will be lost.

Setting up the R session

Before we start this practical we have to set up the R session by installing some packages.

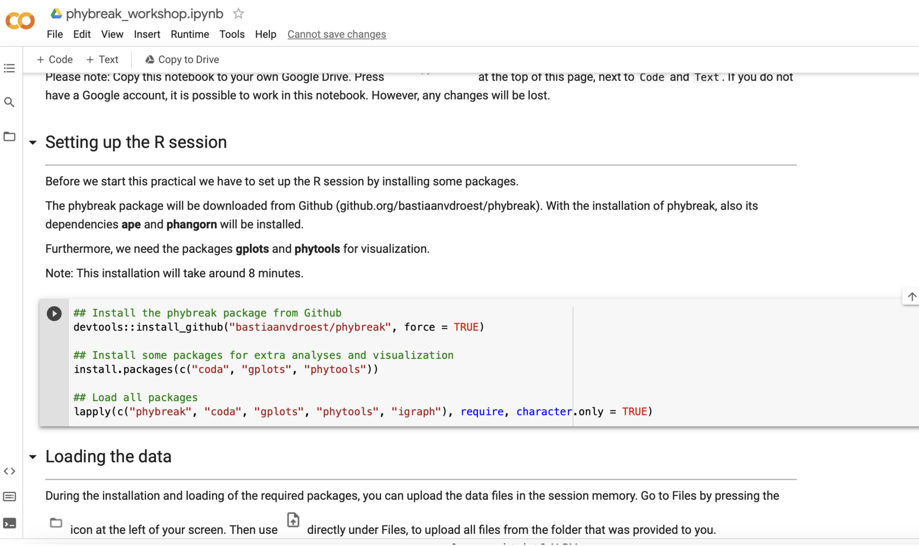
The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

Note: This installation will take around 8 minutes.

```
[ ]  ## Install the phybreak package from Github
     devtools::install_github("bastiaanvdroest/phybreak", force = TRUE)

     ## Install some packages for extra analyses and visualization
     install.packages(c("coda", "gplots", "phytools"))
```

www.bit.ly/phybreakworkshop_colab



Please note: Copy this notebook to your own Google Drive. Press at the top of this page, next to `Code` and `Text`. If you do not have a Google account, it is possible to work in this notebook. However, any changes will be lost.

Setting up the R session

Before we start this practical we have to set up the R session by installing some packages.

The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

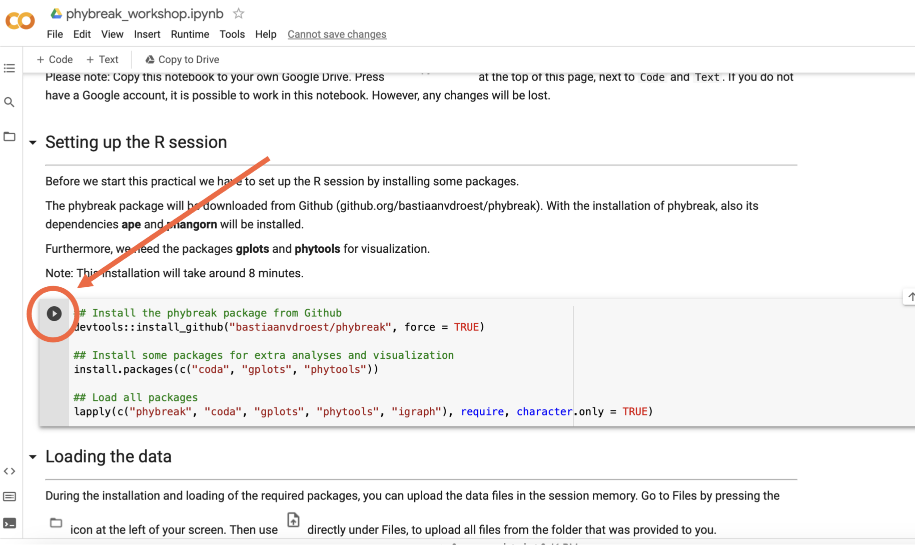
Note: This installation will take around 8 minutes.

```
## Install the phybreak package from Github
devtools::install_github("bastiaanvdroest/phybreak", force = TRUE)

## Install some packages for extra analyses and visualization
install.packages(c("coda", "gplots", "phytools"))

## Load all packages
lapply(c("phybreak", "coda", "gplots", "phytools", "igraph"), require, character.only = TRUE)
```

Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the 🗀 icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

www.bit.ly/phybreakworkshop_colab


phybreak_workshop.ipynb
File Edit View Insert Runtime Tools Help Cannot save changes
+ Code + Text
Copy to Drive

Please note: Copy this notebook to your own Google Drive. Press at the top of this page, next to Code and Text. If you do not have a Google account, it is possible to work in this notebook. However, any changes will be lost.

Setting up the R session

Before we start this practical we have to set up the R session by installing some packages.

The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

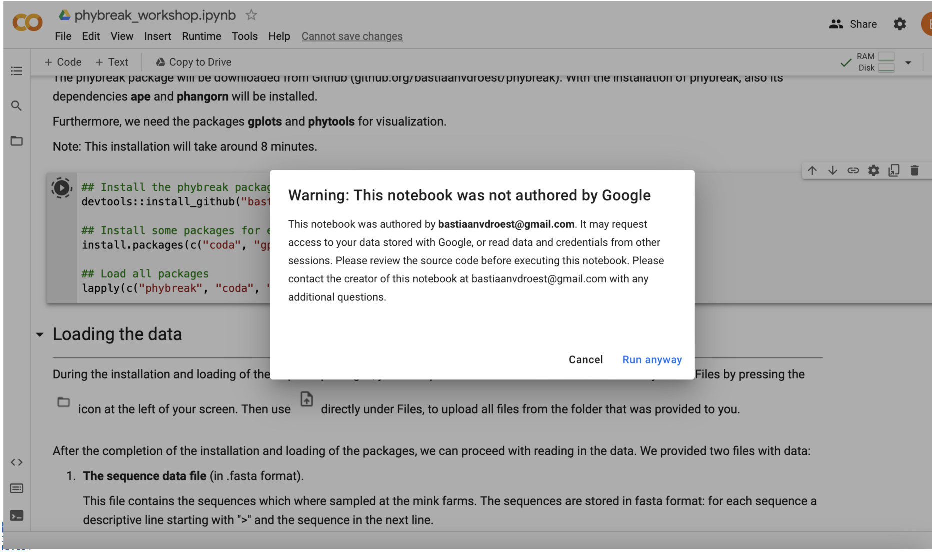
Note: This installation will take around 8 minutes.

```
## Install the phybreak package from Github
devtools::install_github("bastiaanvdroest/phybreak", force = TRUE)

## Install some packages for extra analyses and visualization
install.packages(c("coda", "gplots", "phytools"))

## Load all packages
lapply(c("phybreak", "coda", "gplots", "phytools", "igraph"), require, character.only = TRUE)
```

Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

www.bit.ly/phybreakworkshop_colab



The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

Note: This installation will take around 8 minutes.

```
## Install the phybreak packag
devtools::install_github("bast

## Install some packages for e
install.packages(c("coda", "gp

## Load all packages
lapply(c("phybreak", "coda", "
```

Warning: This notebook was not authored by Google

This notebook was authored by **bastiaanvdroest@gmail.com**. It may request access to your data stored with Google, or read data and credentials from other sessions. Please review the source code before executing this notebook. Please contact the creator of this notebook at bastiaanvdroest@gmail.com with any additional questions.

Loading the data

During the installation and loading of the Files by pressing the

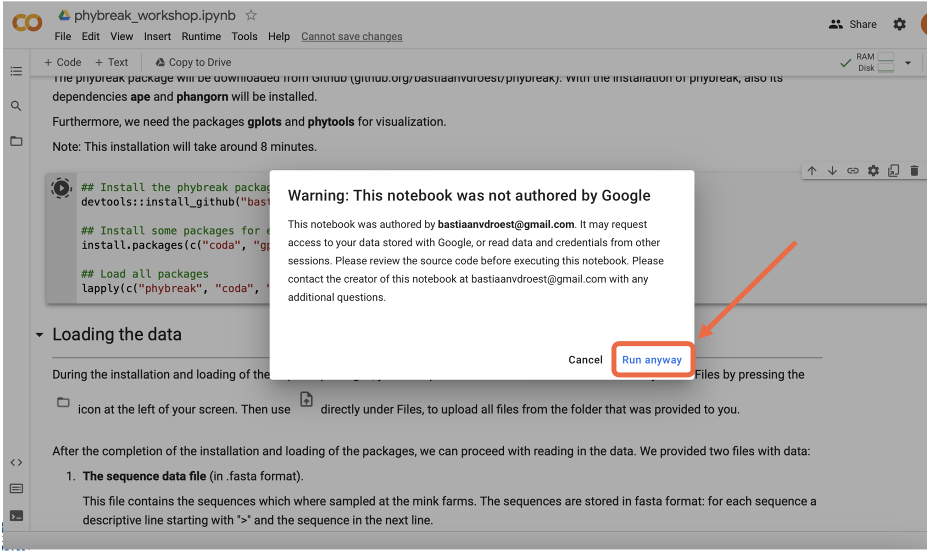
icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

www.bit.ly/phybreakworkshop_colab



The phybreak package will be downloaded from Github (github.org/bastiaanvdroest/phybreak). With the installation of phybreak, also its dependencies **ape** and **phangorn** will be installed.

Furthermore, we need the packages **gplots** and **phytools** for visualization.

Note: This installation will take around 8 minutes.

```
## Install the phybreak packag
devtools::install_github("bast

## Install some packages for e
install.packages(c("coda", "gp

## Load all packages
lapply(c("phybreak", "coda", '
```

Warning: This notebook was not authored by Google

This notebook was authored by **bastiaanvdroest@gmail.com**. It may request access to your data stored with Google, or read data and credentials from other sessions. Please review the source code before executing this notebook. Please contact the creator of this notebook at bastiaanvdroest@gmail.com with any additional questions.

Loading the data

During the installation and loading of the Files by pressing the

icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

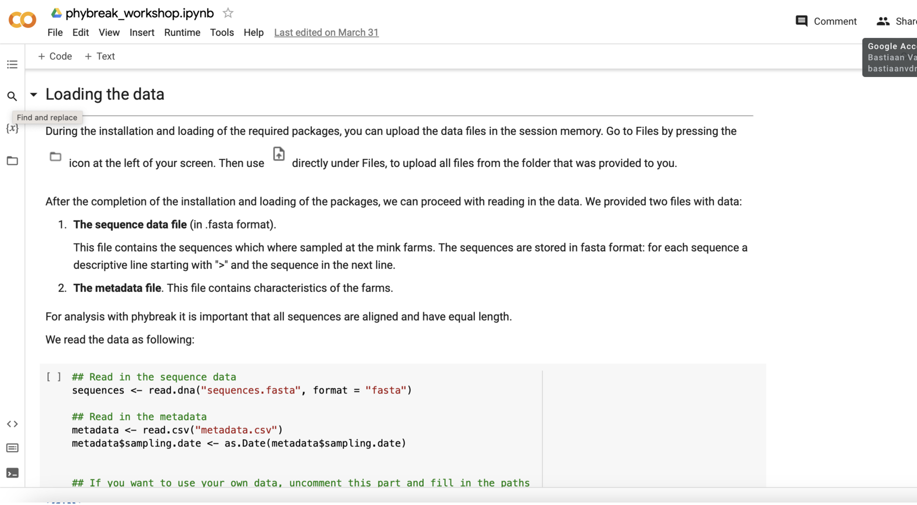
After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

Time for coffee

www.bit.ly/phybreakworkshop_colab



Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the 🗀 icon at the left of your screen. Then use 🗎 directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

2. **The metadata file**. This file contains characteristics of the farms.

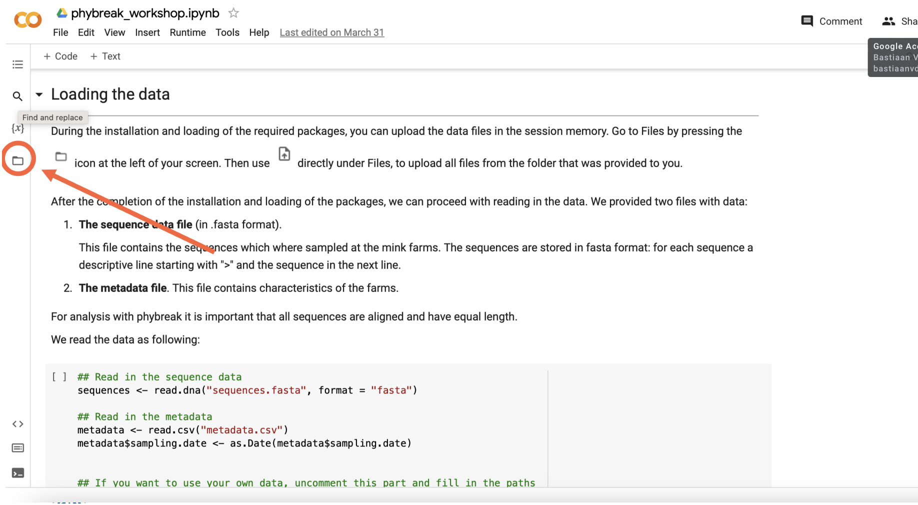
For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
    metadata$sampling.date <- as.Date(metadata$sampling.date)


    ## If you want to use your own data, uncomment this part and fill in the paths
```

Data on Nordita website: Timetable: today’s session

www.bit.ly/phybreakworkshop_colab



Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

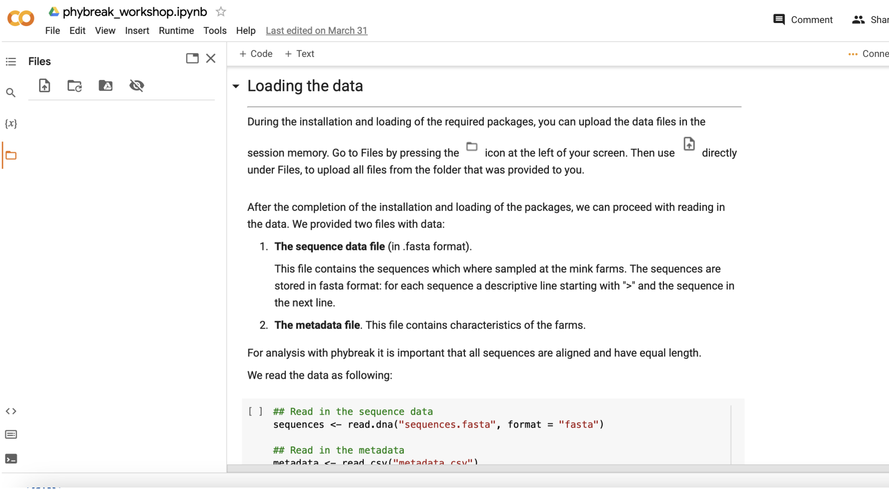
2. **The metadata file**. This file contains characteristics of the farms.

For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
    metadata$sampling.date <- as.Date(metadata$sampling.date)


    ## If you want to use your own data, uncomment this part and fill in the paths
```

www.bit.ly/phybreakworkshop_colab



Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the 🗀 icon at the left of your screen. Then use 🗎 directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

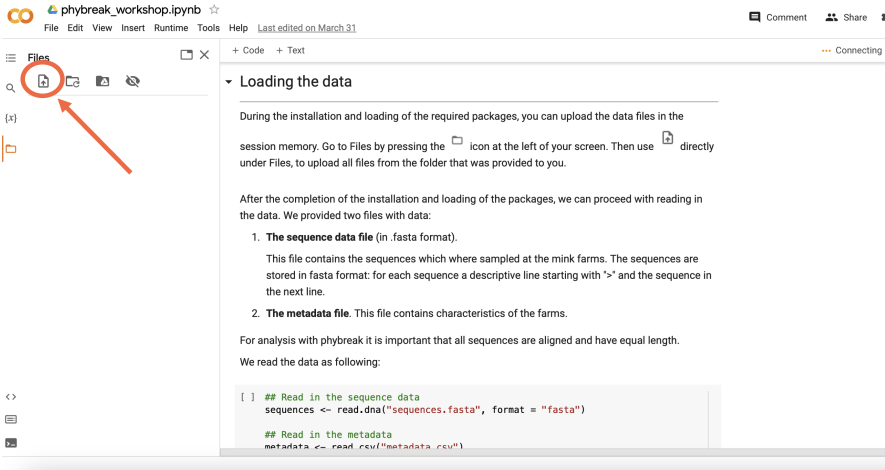
2. **The metadata file**. This file contains characteristics of the farms.

For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
```

www.bit.ly/phybreakworkshop_colab


CO
phybreak_workshop.ipynb ☆
File Edit View Insert Runtime Tools Help Last edited on March 31
Comment
Share
Files
+ Code + Text
··· Connecting
{x}

Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the 🗀 icon at the left of your screen. Then use 🗎 directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading of the packages, we can proceed with reading in the data. We provided two files with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

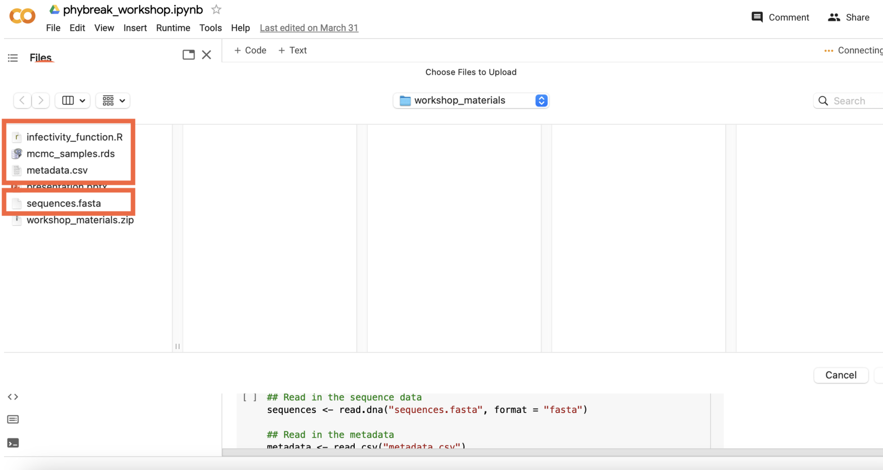
2. **The metadata file**. This file contains characteristics of the farms.

For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

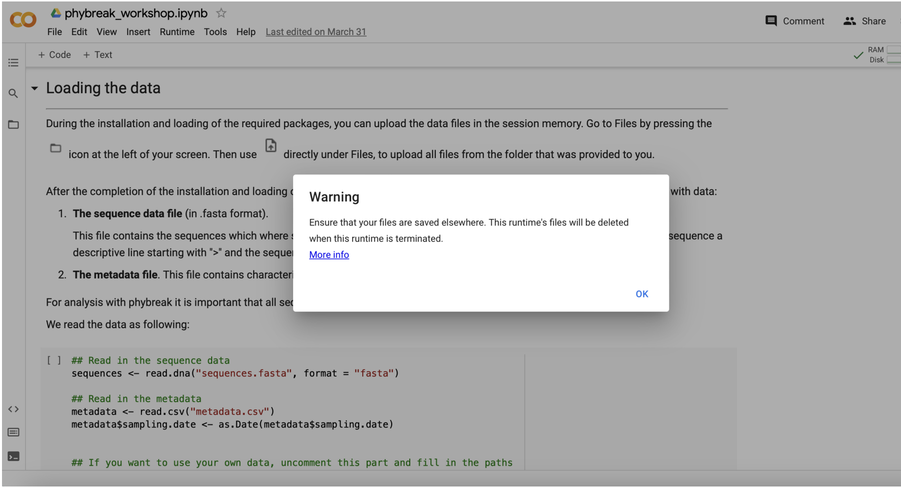
```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
```

www.bit.ly/phybreakworkshop_colab



```
[ ]  ## Read in the sequence data
     sequences <- read.dna("sequences.fasta", format = "fasta")

     ## Read in the metadata
     metadata <- read.csv("metadata.csv")
```

www.bit.ly/phybreakworkshop_colab



Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the

icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading c with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where s sequence a descriptive line starting with ">" and the seque

2. **The metadata file**. This file contains characteri

For analysis with phybreak it is important that all sec

We read the data as following:

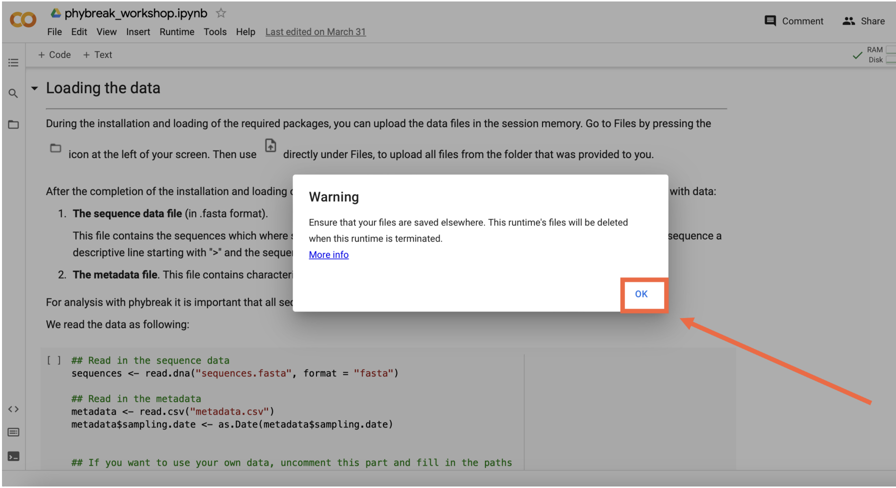
```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
    metadata$sampling.date <- as.Date(metadata$sampling.date)


    ## If you want to use your own data, uncomment this part and fill in the paths
```

Warning

Ensure that your files are saved elsewhere. This runtime's files will be deleted when this runtime is terminated.

More info

www.bit.ly/phybreakworkshop_colab


CO
phybreak_workshop.ipynb ☆
File Edit View Insert Runtime Tools Help Last edited on March 31
Comment
Share
+ Code + Text
RAM
Disk

Loading the data

During the installation and loading of the required packages, you can upload the data files in the session memory. Go to Files by pressing the icon at the left of your screen. Then use directly under Files, to upload all files from the folder that was provided to you.

After the completion of the installation and loading c with data:

1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where s sequence a descriptive line starting with ">" and the sequen

2. **The metadata file**. This file contains characteri

For analysis with phybreak it is important that all seq

We read the data as following:

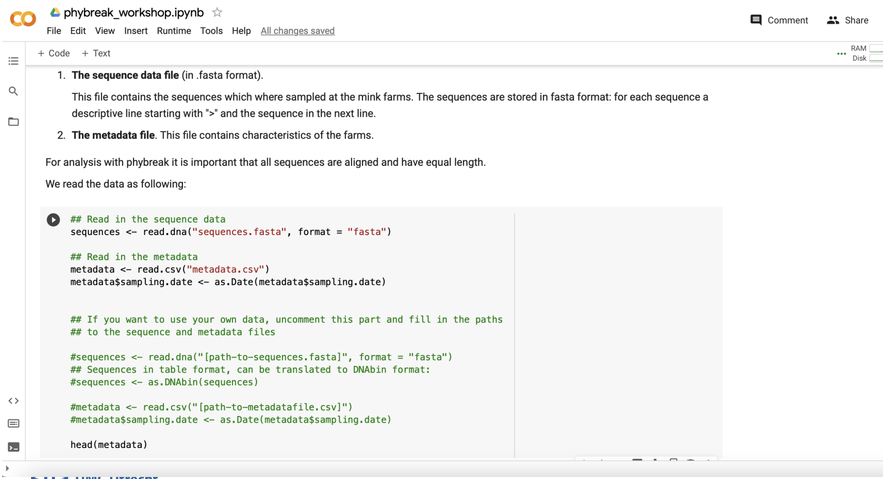
```
[ ] ## Read in the sequence data
    sequences <- read.dna("sequences.fasta", format = "fasta")

    ## Read in the metadata
    metadata <- read.csv("metadata.csv")
    metadata$sampling.date <- as.Date(metadata$sampling.date)


    ## If you want to use your own data, uncomment this part and fill in the paths
```

Warning

Ensure that your files are saved elsewhere. This runtime's files will be deleted when this runtime is terminated.

More info

www.bit.ly/phybreakworkshop_colab



1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

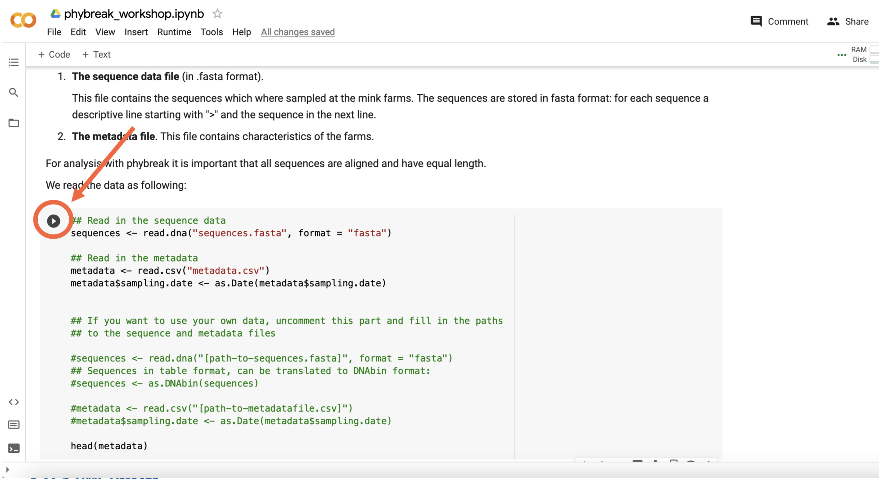
2. **The metadata file**. This file contains characteristics of the farms.

For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

```
## Read in the sequence data
sequences <- read.dna("sequences.fasta", format = "fasta")

## Read in the metadata
metadata <- read.csv("metadata.csv")
metadata$sampling.date <- as.Date(metadata$sampling.date)


## If you want to use your own data, uncomment this part and fill in the paths
## to the sequence and metadata files

#sequences <- read.dna("[path-to-sequences.fasta]", format = "fasta")
## Sequences in table format, can be translated to DNAbin format:
#sequences <- as.DNAbin(sequences)

#metadata <- read.csv("[path-to-metadatafile.csv]")
#metadata$sampling.date <- as.Date(metadata$sampling.date)

head(metadata)
```

www.bit.ly/phybreakworkshop_colab



1. **The sequence data file** (in .fasta format).

   This file contains the sequences which where sampled at the mink farms. The sequences are stored in fasta format: for each sequence a descriptive line starting with ">" and the sequence in the next line.

2. **The metadata file**. This file contains characteristics of the farms.

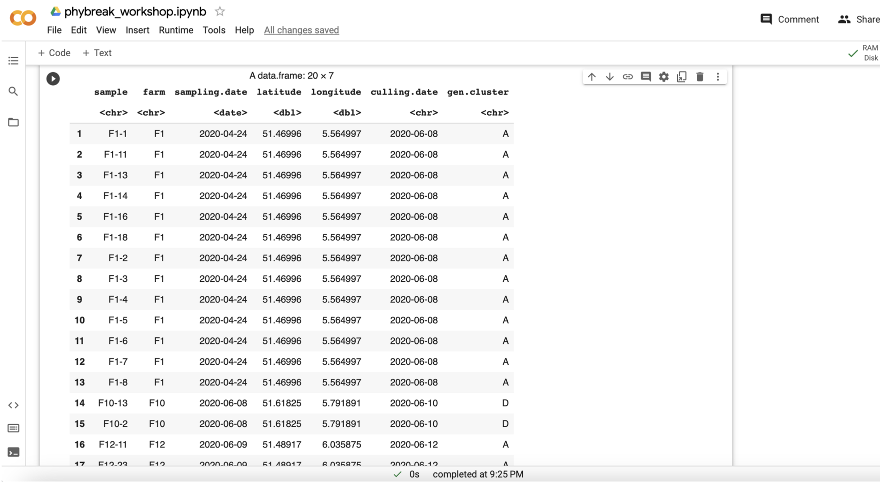
For analysis with phybreak it is important that all sequences are aligned and have equal length.

We read the data as following:

```
## Read in the sequence data
sequences <- read.dna("sequences.fasta", format = "fasta")

## Read in the metadata
metadata <- read.csv("metadata.csv")
metadata$sampling.date <- as.Date(metadata$sampling.date)


## If you want to use your own data, uncomment this part and fill in the paths
## to the sequence and metadata files

#sequences <- read.dna("[path-to-sequences.fasta]", format = "fasta")
## Sequences in table format, can be translated to DNAbin format:
#sequences <- as.DNAbin(sequences)

#metadata <- read.csv("[path-to-metadatafile.csv]")
#metadata$sampling.date <- as.Date(metadata$sampling.date)

head(metadata)
```

UMC Utrecht

www.bit.ly/phybreakworkshop_colab



A data.frame: 20 × 7

	sample	farm	sampling.date	latitude	longitude	culling.date	gen.cluster
	<chr>	<chr>	<date>	<dbl>	<dbl>	<chr>	<chr>
1	F1-1	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
2	F1-11	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
3	F1-13	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
4	F1-14	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
5	F1-16	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
6	F1-18	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
7	F1-2	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
8	F1-3	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
9	F1-4	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
10	F1-5	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
11	F1-6	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
12	F1-7	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
13	F1-8	F1	2020-04-24	51.46996	5.564997	2020-06-08	A
14	F10-13	F10	2020-06-08	51.61825	5.791891	2020-06-10	D
15	F10-2	F10	2020-06-08	51.61825	5.791891	2020-06-10	D
16	F12-11	F12	2020-06-09	51.48917	6.035875	2020-06-12	A
17	F12-23	F12	2020-06-09	51.48917	6.035875	2020-06-12	A

✓ 0s completed at 9:25 PM

www.bit.ly/phybreakworkshop_colab