An Open and Modular Phylogenetic Tree Construction Pipeline in R

GGBC Workshop | 17th Sept. 2019

Timetable

0900 - 0930	Welcome and Introduction to supersmartR
1000 - 1200	*phylotaR* *outsider* - Base pipeline Coffee/tea, ~1030
1200 - 1300	Lunch
1300 - 1800	- Multigene pipeline *restez* - Large pipeline - Supertree pipeline - Improvements/Free-coding Coffee/tea, ~1500

SUPERSMART*

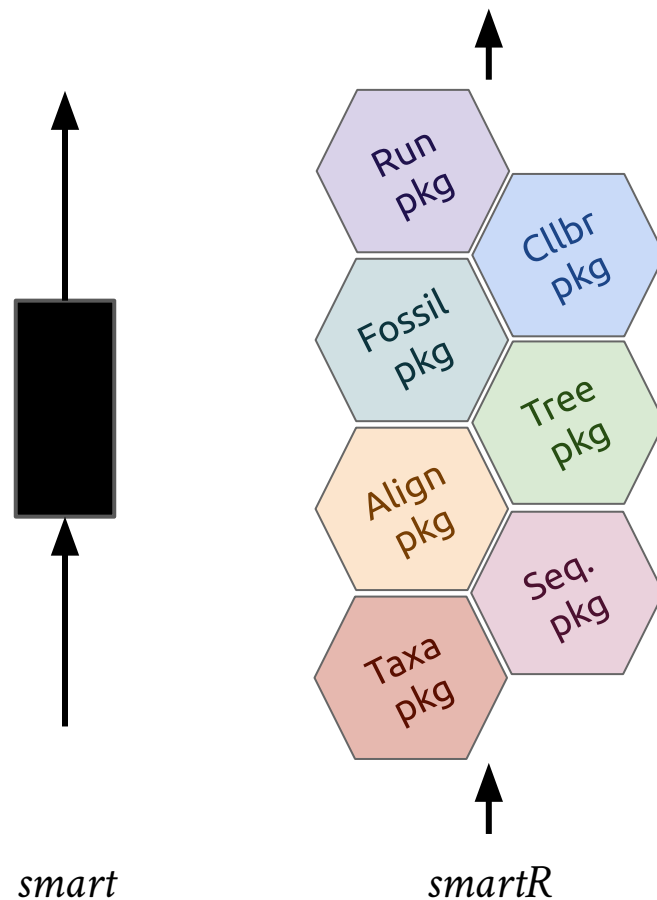Pipeline for constructing phylogenetic trees.

Species-level and backbone approach.

Generic, scalable. Maximise data.
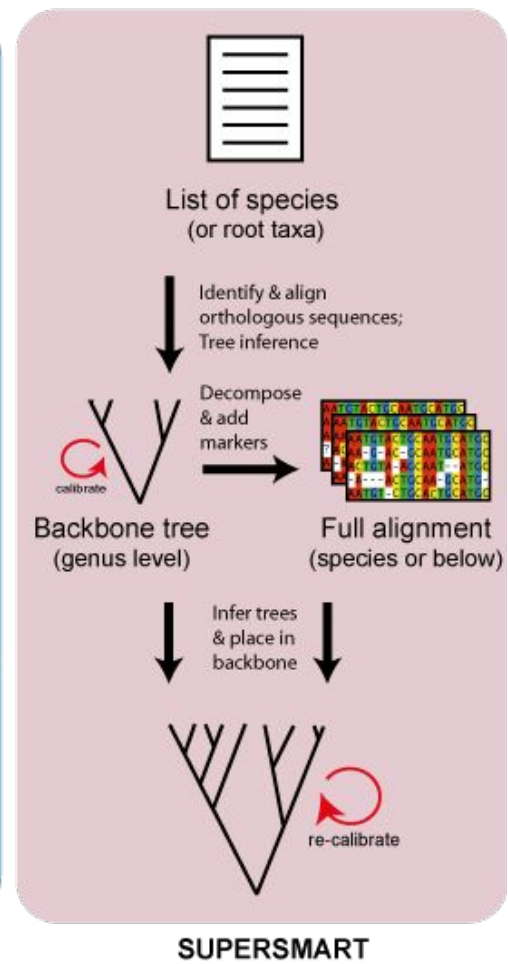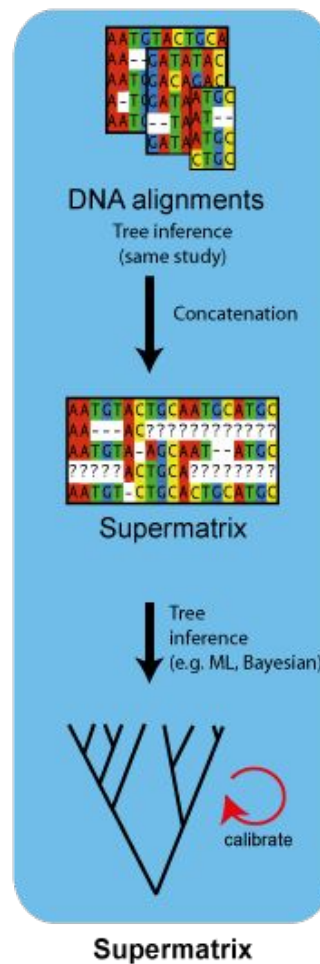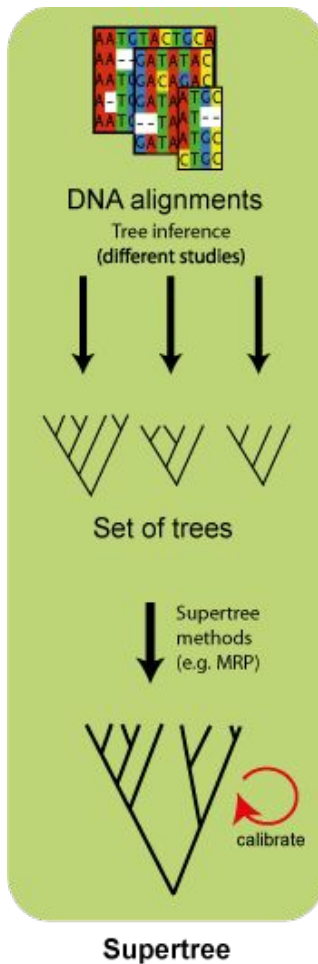
A supersmartR approach

- Break SUPERSMART into separate R packages
- To serve the SUPERSMART pipeline and others
- Banking gains
  - Modular
  - Open-science/software

smart *smartR*

Setting up ....

GGBC supersmartR-workshop | 17th Sept. 2019

Caviomorpha | Image: Viscacha (*Lagidium* [?]), Planet Earth II, BBC ©

Setup

- Please have installed latest versions of:
  - `phylotaR`
  - `outsider`
  - `restez`
  - `gaius`
- Please ensure "docker" is installed
- You should be able to run:

```
library(outsider)
repo <- "dombennett/om..hello.world"
module_install(repo = repo, force = TRUE)
hello_world <- module_import(fname = "hello_world", repo = repo)
hello_world()
```

Setup

Download "supersmartR-workshop" via GitHub:
https://github.com/AntonelliLab/supersmartR-workshop

Setup

Open [RStudio](#), create a new project from the downloaded folder.
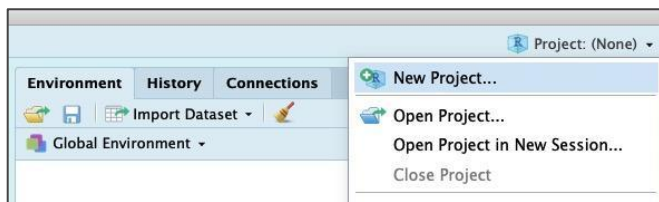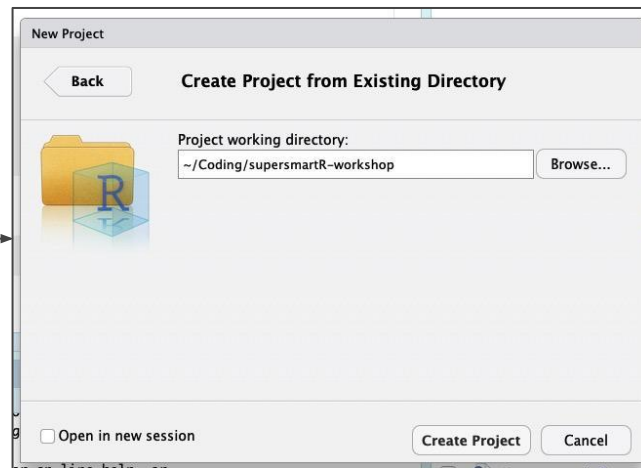
Folder structure

- Package introductions
  - `phylotaR`
  - `restez`
  - ….

  Introduction + exercise per package

- Pipelines
  - `0_base`
  - `1_multigene`
  - …

  Ever more complex pipelines for creating trees for taxa of choice

- Extra topics
  - `ssh`

  Extra topics for improving pipelines

phylotaR

Package introduction | GGBC supersmartR-workshop

Outline

txid

| Taxise | **Stage 1**
Identify and download
taxonomic information | NCBI
Taxonomy |

| Download | **Stage 2**
Get sequences down the
taxonomic hierarchy. | NCBI
GenBank |

| Cluster | **Stage 3**
Generate clusters down
the taxonomic hierarchy. | NCBI
BLAST+ |

| Cluster^2 | **Stage 4**
Identify and join sister
clades together. | NCBI
BLAST+ |

User selects
clusters
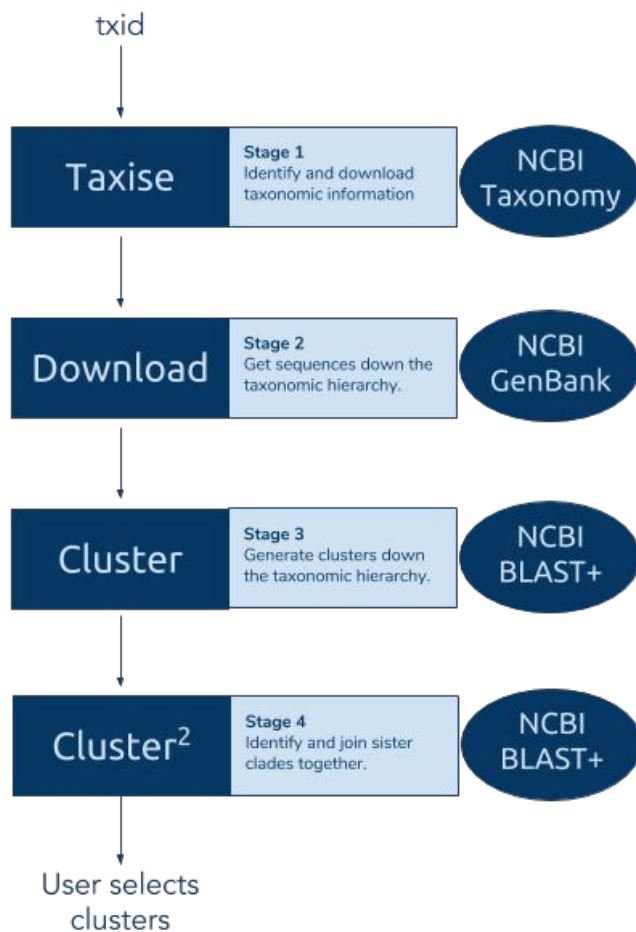
Hands-on

package_introductions/1_phylotaR

outsider

Package introduction | GGBC supersmartR-workshop

Caviomorpha | Image: Viscacha (*Lagidium* [?]), Planet Earth II, BBC ©

Outsider Modules

Local computer

module_install()

```
user/om..program
dockerfiles/
R/
man/
NAMESPACE
DESCRIPTION
inst/om.yml
```

Hostable on GitHub BitBucket GitLab

module_details()

```
library(outsider)
repo <- "user/om..program"
module_install(repo)
module_help(repo)
foo <- module_import('foo',repo)
foo()
```
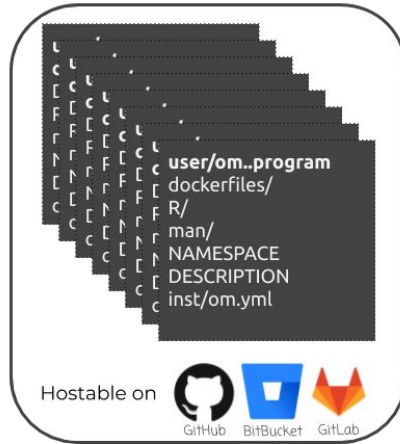
Run via docker

Outline

Hands-on

package_introductions/2_outsider

Pipelines in R

Basic R code for pipelines | GGBC supersmartR-workshop

Pipelines in R

- Structure
  - Stages: independent scripts
  - Defined input/output directories
  - Universal output
- Functions
  - `source()`
  - `file.path()`
- Avoid
  - `setwd()`

```r
# Libs ----
source(file.path("path", "to", "tools.R"))
library("packages")

# Vars ----
wd <- file.path ("path", "to", "parent_dir")
input_dir <- file.path(wd, "previous_stage_name")
output_dir <- file.path(wd, "current_stage_name")

# Run ----
# Run the main code of the stage: download, align ….

# Write out ----
save(results, output_dir)
```

0_base

Phylogenetic pipeline | GGBC supersmartR-workshop

0_base

Generate a phylogenetic tree for a single gene for a single taxon

Stages

1_phylotaR	Fetch sequence clusters
2_clusters	Select gene cluster
3_align	Align sequences
4_phylogeny	Generate tree
5_view	Plot the resulting tree

0_base

Task
- Choose a small taxon*
  - < 10 spp
  - < 1000 nucleotide records
- Edit scripts for your taxon of choice.

Hands-on
- pipelines/0_base

* To choose, visit : https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

1_multigene

Phylogenetic pipeline | GGBC supersmartR-workshop

1_multigene

Generate a phylogenetic tree from multiple gene/clusters for a single taxon.

Stages

1_phylotaR	Fetch sequence clusters
2_clusters	Select multiple clusters
3_align	Align sequences
4_supermatrix	Assemble supermatrix
5_phylogeny	Generate tree
6_view	Plot the resulting tree

1_multigene

Task

- Edit scripts for your taxon of choice from 0_base
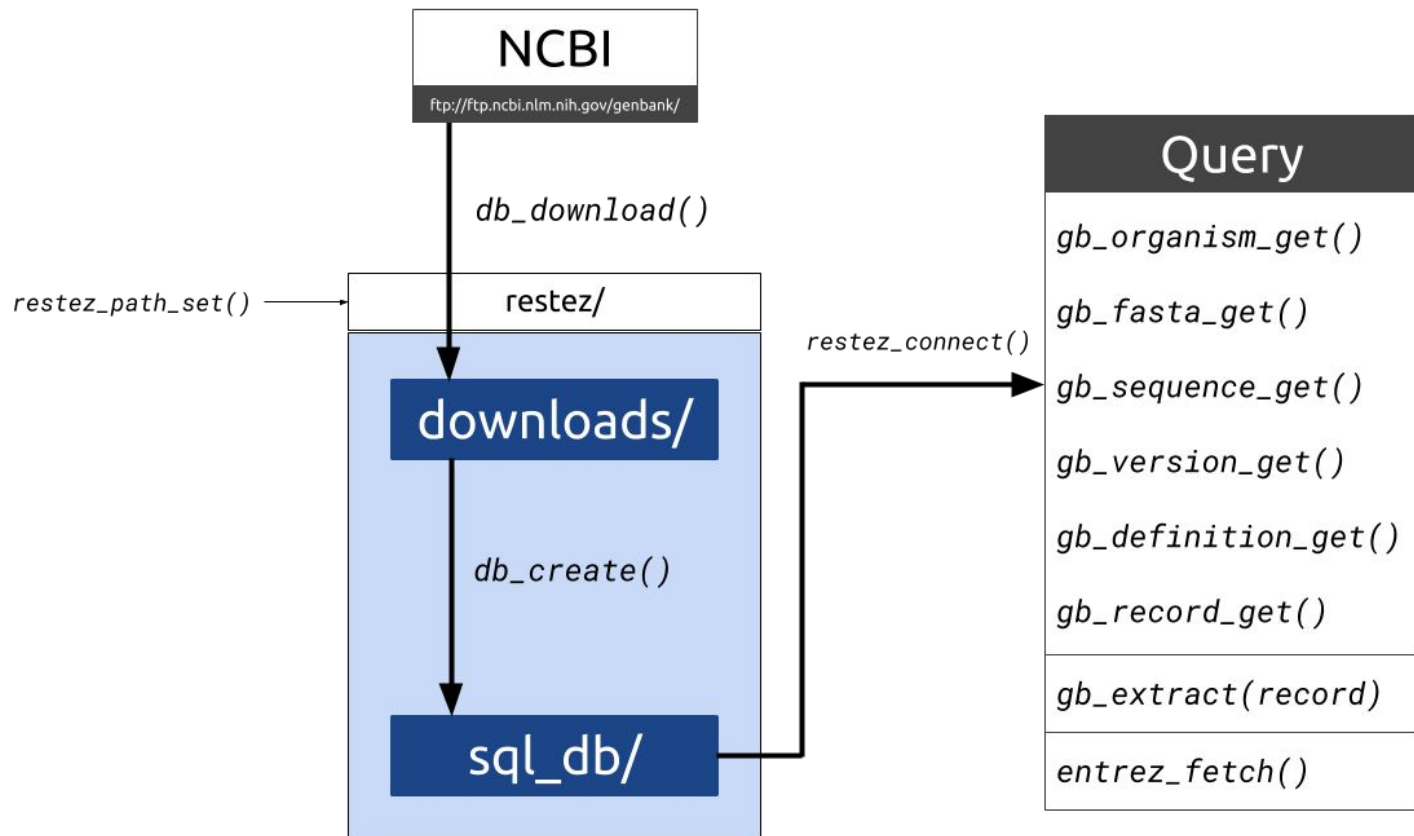
Hands-on

- pipelines/1_multigene

restez

Package introduction | GGBC supersmartR-workshop

Outline

Large

Phylogenetic pipeline | GGBC supersmartR-workshop

2_large

Generate a phylogenetic tree from multiple gene/clusters for a large taxon (> 100 spp.).

Stages

0_restez	Download sections of GenBank
1_phylotaR	Fetch sequence clusters
2_clusters	Select multiple clusters
3_align	Align sequences
4_supermatrix	Assemble supermatrix
5_phylogeny	Generate tree
6_view	Plot the resulting tree

2_large

Task
- Choose a large taxon *
    - < 100 spp
    - < 10000 nucleotide records
- Edit scripts for your taxon of choice

Hands-on
- pipelines/2_large

* To choose, visit : https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

Supertree

Phylogenetic pipeline | GGBC supersmartR-workshop

3_supertree

Generate a phylogenetic tree from multiple gene/clusters for a large taxon (< 500 spp.).

Stages

0_restez	Download sections of GenBank
1_phylotaR	Fetch sequence clusters
2_clusters	Select multiple clusters
3_align	Align sequences
4_supermatrix	Assemble supermatrices (mono. + backbone)
5_phylogeny	Generate trees
6_supertree	Assemble the supertree
7_view	Plot the resulting tree

3_supertree

Task

- Choose a larger taxon still [*]
  - < 500 spp
  - < 100000 nucleotide records
- Edit scripts for your taxon of choice
- (Or: use package_introductions/1_phylotaR data)

Hands-on

- pipelines/3_supertree

[*] To choose, visit : https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
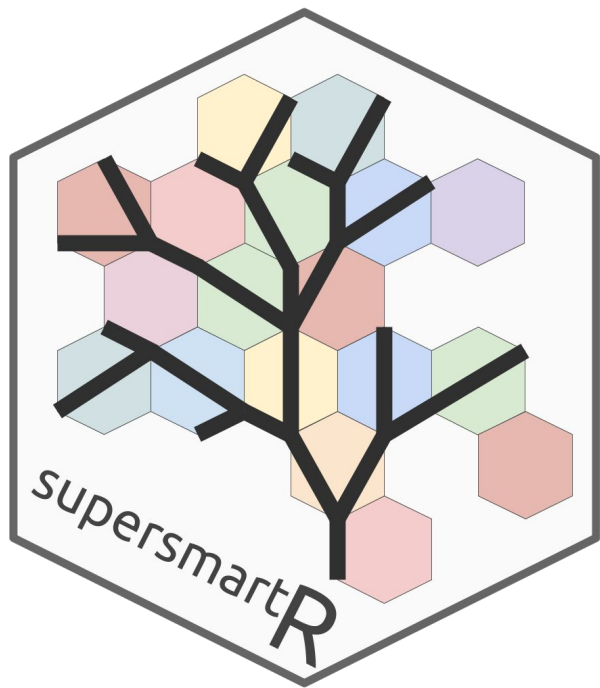
To the next level ...

GGBC supersmartR-workshop | 17th Sept. 2019

Improvements

- Programs (outsider modules)
    - [trimal](#), [partitionFinder](#), [Astral](#) ...
- Taxonomy
    - Names parsing with `taxize`
- Computing power
    - [ssh to servers](#)

supersmartR

An Open and Modular Phylogenetic Tree Construction Pipeline in R

GGBC Workshop | 17th Sept. 2019