The phyloseq package and analysis of high throughput amplicon sequencing data

Paul McMurdie
Statistics Department
Stanford University

BioC 2012

July 24-25, 2012 (Developer Day: July 23)
Fred Hutchinson Cancer Research Center - Seattle, WA

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Why we wrote phyloseq

"phyloseq" <-> phylogenetic sequencing

Context

There are already several ecology and phylogenetic packages available in R, including the vegan, ade4, ape, phangorn, picante, etc. To varying degrees, these packages already leverage the many powerful statistical and graphics tools available in R.

However, prior to phyloseq there was no standard within Bioconductor (or R generally) for storing or sharing the suite of related data objects that describe a phylogenetic sequencing project, leading to a common (and usually poorly documented) hurdle to using R for phylogenetic sequencing analysis.

Why we wrote phyloseq

"phyloseq" <-> phylogenetic sequencing

Design Philosophy

Data Infrastructure:
The first goal of the phyloseq package is to provide an infrastructure for importing and representing the data from phylogenetic sequencing experiments in a manner that is convenient, concise, and complete; such that it is ultimately very easy to share -- and reproduce -- the complex multivariate statistical analyses often required of these experiments.

Data Interpretation:
We further aim to provide enough tools, extensions of existing tools, and examples such that the phyloseq package can be considered an important first step in the interpretation of phylogenetic sequencing data.

Graphics:
We provide many custom graphics functions built upon ggplot2 that are very useful in both exploration and interpretation of the data, but also highly customizable and in many cases of a quality suitable for publication.

phyloseq Design and Features

- A single, explicitly-defined S4 class that can store the different data types of a phylogenetic sequencing experiment in a single object.
- Importers all create this special “phyloseq” class
- Most phyloseq functions will act on this experiment-level object. It doesn't need to be diced-up to work. (Keep data together)
- Internal tools check validity and agreement among components of an experiment. Helps prevent mistakes.
- Plotting tools for creating quality graphics, built using ggplot2.
- Example datasets from real published data, with references, documentation, and examples.
- Examples using other R tools after importing with phyloseq

Workshop Outline:

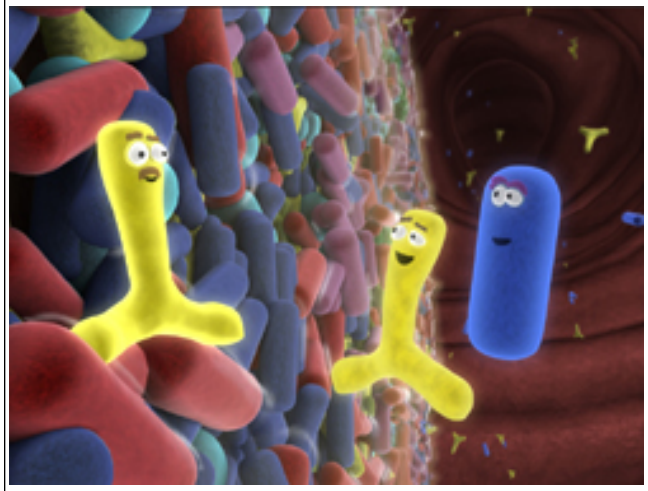
- Background of modern phylogenetic sequencing (if needed).
- Motivation, Design, and Philosophy of phyloseq

live code

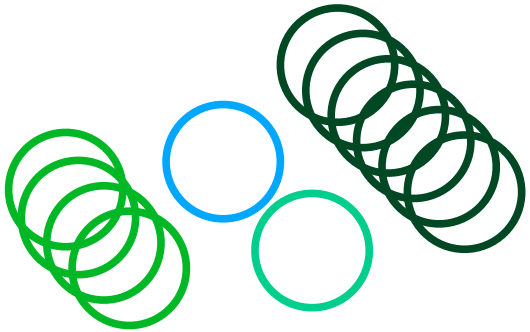
- How to import data with phyloseq
- Basic interaction with data and simple summary graphics
- Data preprocessing using phyloseq tools
- More complex exploratory/summary graphics, including ordination
- Validation tools supported in phyloseq
- Additional validation/testing using other R tools (getting data components to other R functions)

Amplicon Sequencing

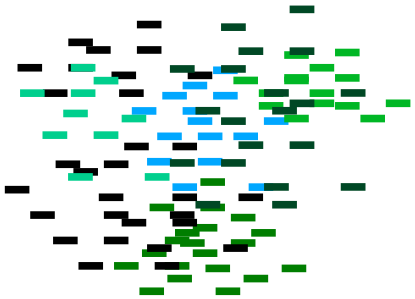
Goal: Infer original abundance of different types of target gene



biological sample
e.g. bacterial community



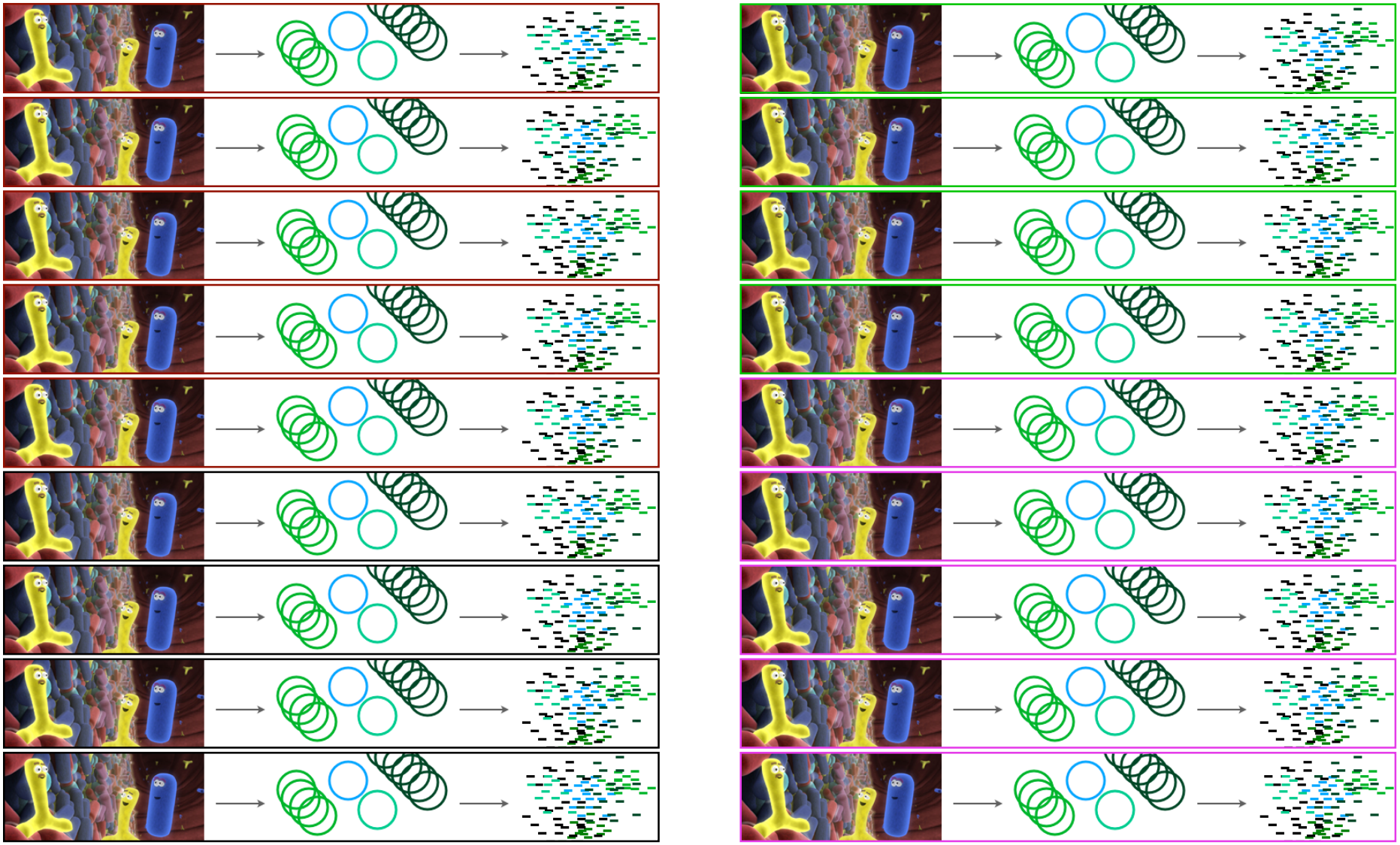
Extract DNA
(mixture)



Amplify single gene
of interest.
Sequence products

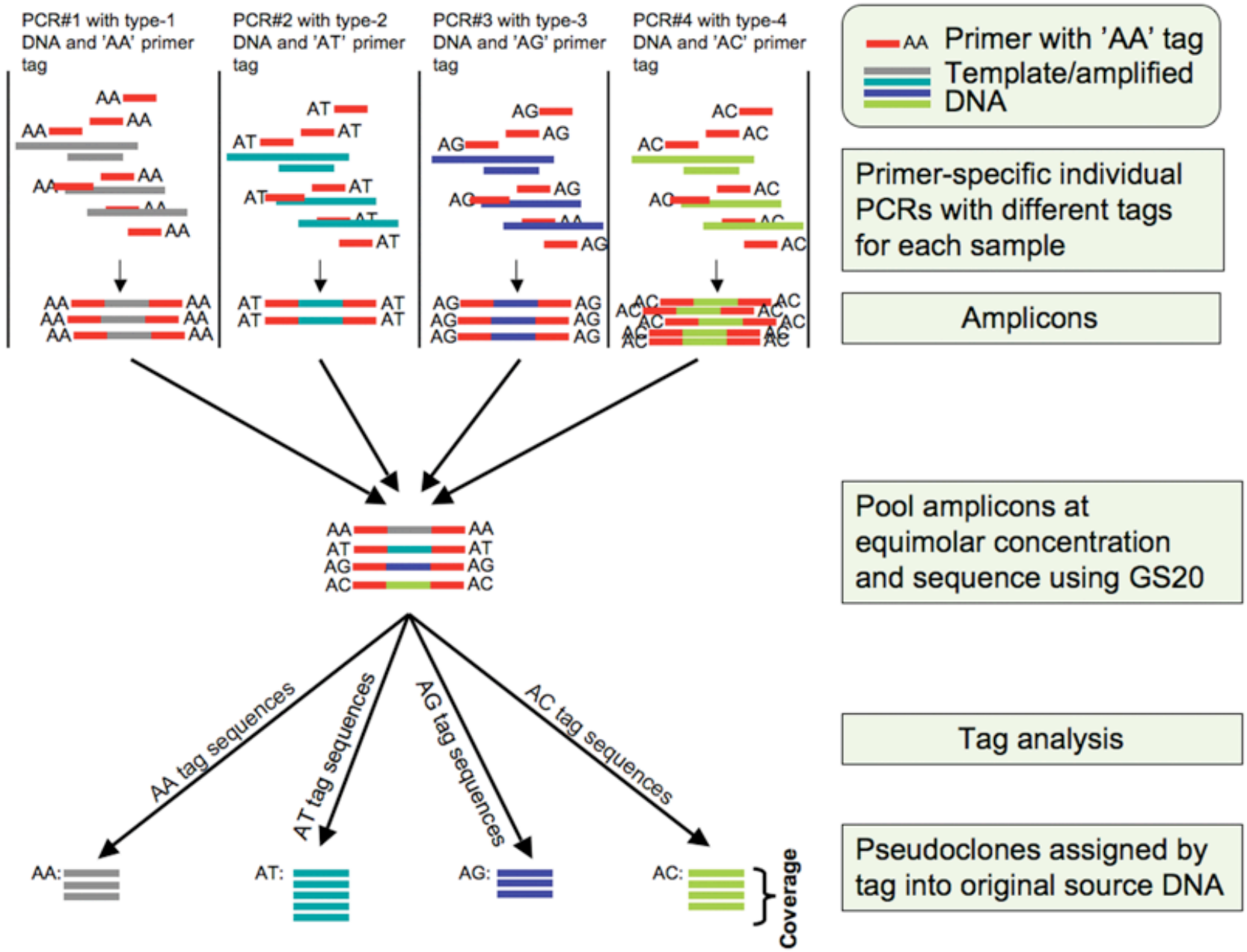
Amplicon Sequencing

Repeat many times with different samples/replicates



Amplicon Sequencing

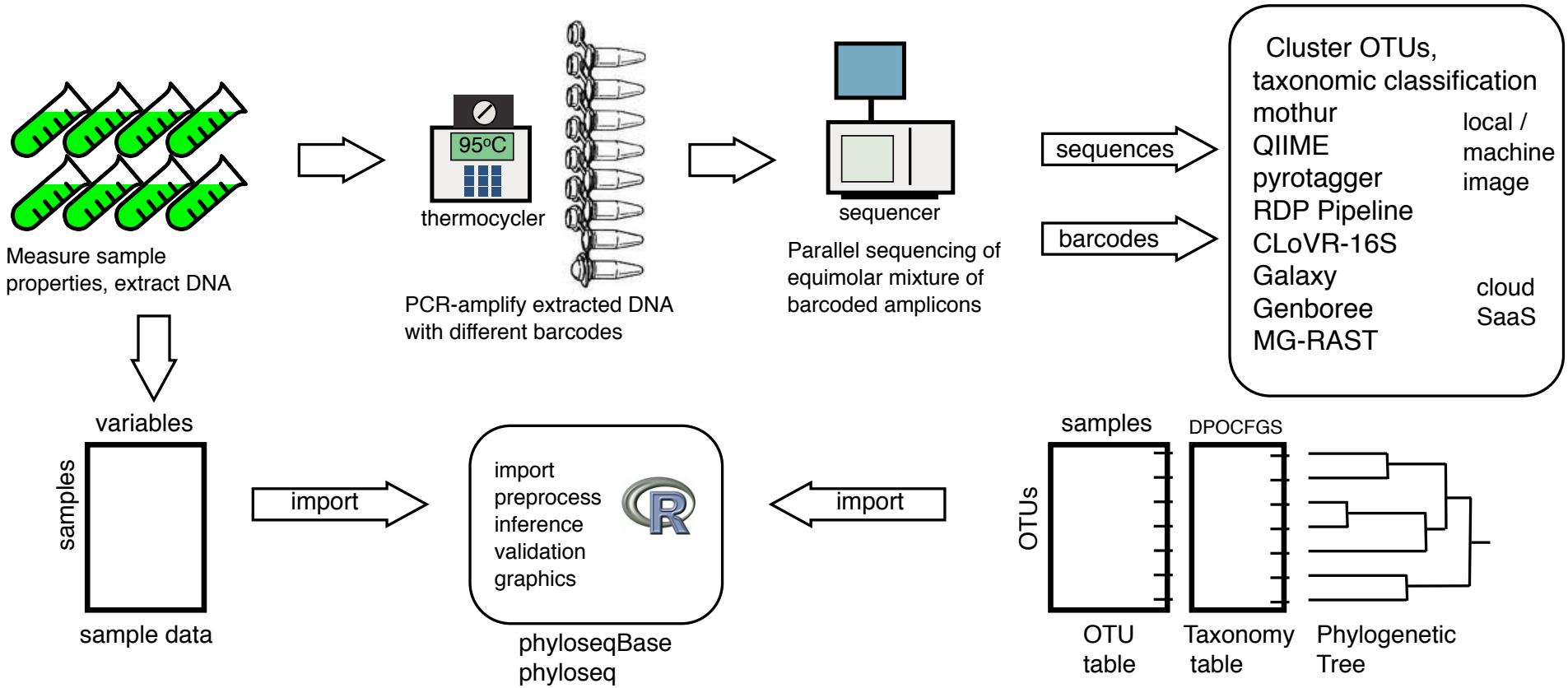
parallel tagged sequencing
"bar-coded" sequences



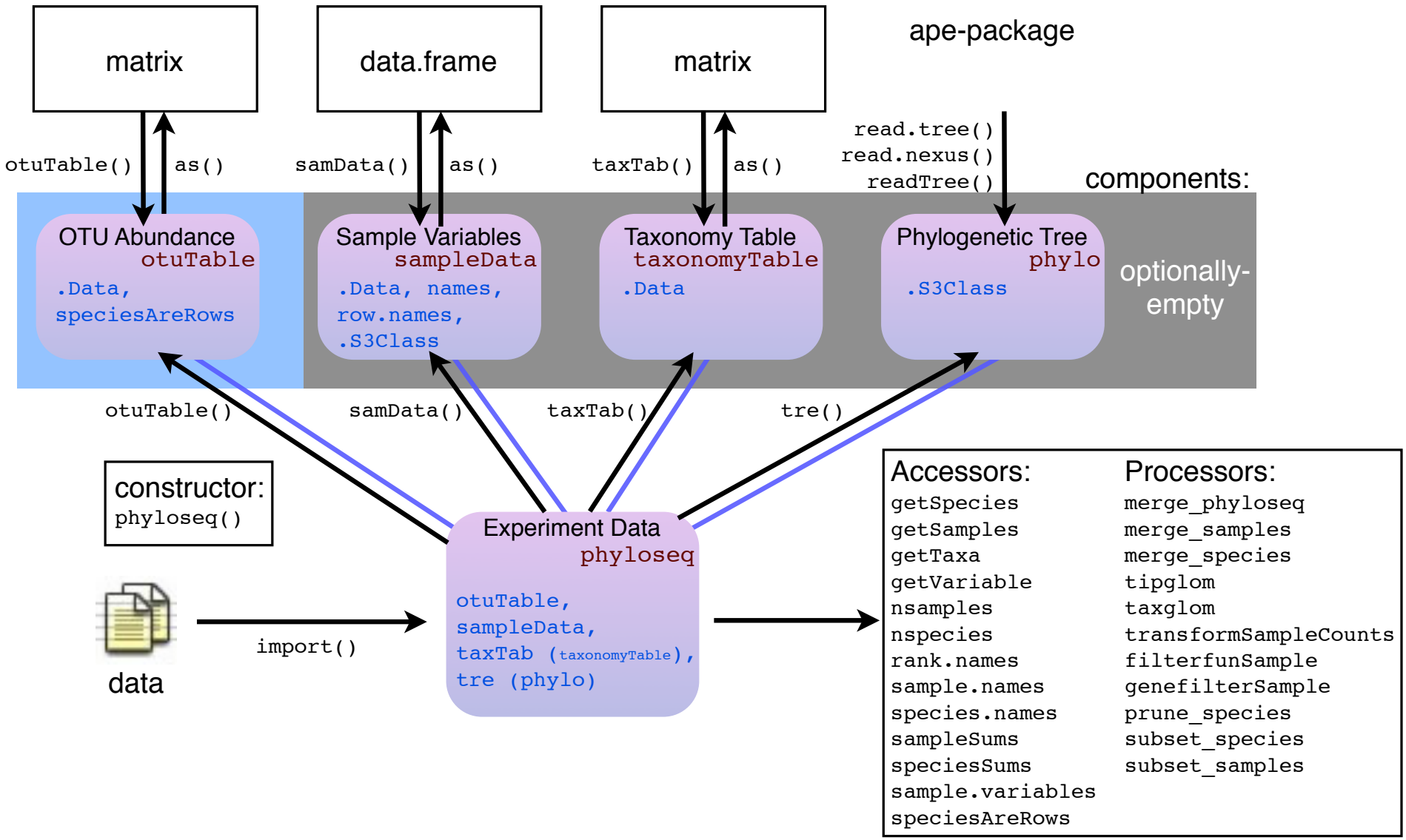
Binladen, *et al* (2007). The use of coded PCR primers enables. *PLoS ONE*, *2*(2), e197. doi:10.1371/journal.pone.0000197

Amplicon Sequencing

Overview of amplicon sequencing and analysis



phyloseq classes and "data infrastructure"



phyloseq accessors

Function	Description
`[`	Standard extraction operator. works on otuTable, sampleData, and taxonomyTable
`access`	General slot accessor function for phyloseq-package
`getslots.phyloseq`	Return the slot names of phyloseq objects
`getSpecies`	Returns the abundance values of sample 'i' for all species in 'x'
`getSamples`	Returns the abundance values of species 'i' for all samples in 'x'
`getTaxa`	Get a unique vector of the observed taxa at a particular taxonomic rank
`getVariable`	Returns an individual sample variable vector/factor
`nsamples`	Get the number of samples described by an object
`nspecies`	Get the number of species (taxa) described by an object
`otuTable`	Build or access otuTable objects
`rank.names`	Get the names of the available taxonomic ranks
`sampleData`	Build or access sampleData objects
`sample.names`	Return the names of the samples described by an object
`species.names`	Return the names of the species described by an object
`sampleSums`	Returns the total number of individuals observed from each sample
`sample.variables`	Returns the names of sample variables in an object
`speciesSums`	Returns the total number of individuals observed from each species
`speciesAreRows`	Returns the orientation of the abundance table
`taxTab`	Build or access taxTab objects
`tre`	Access the tree contained in a phyloseq object

phyloseq constructors

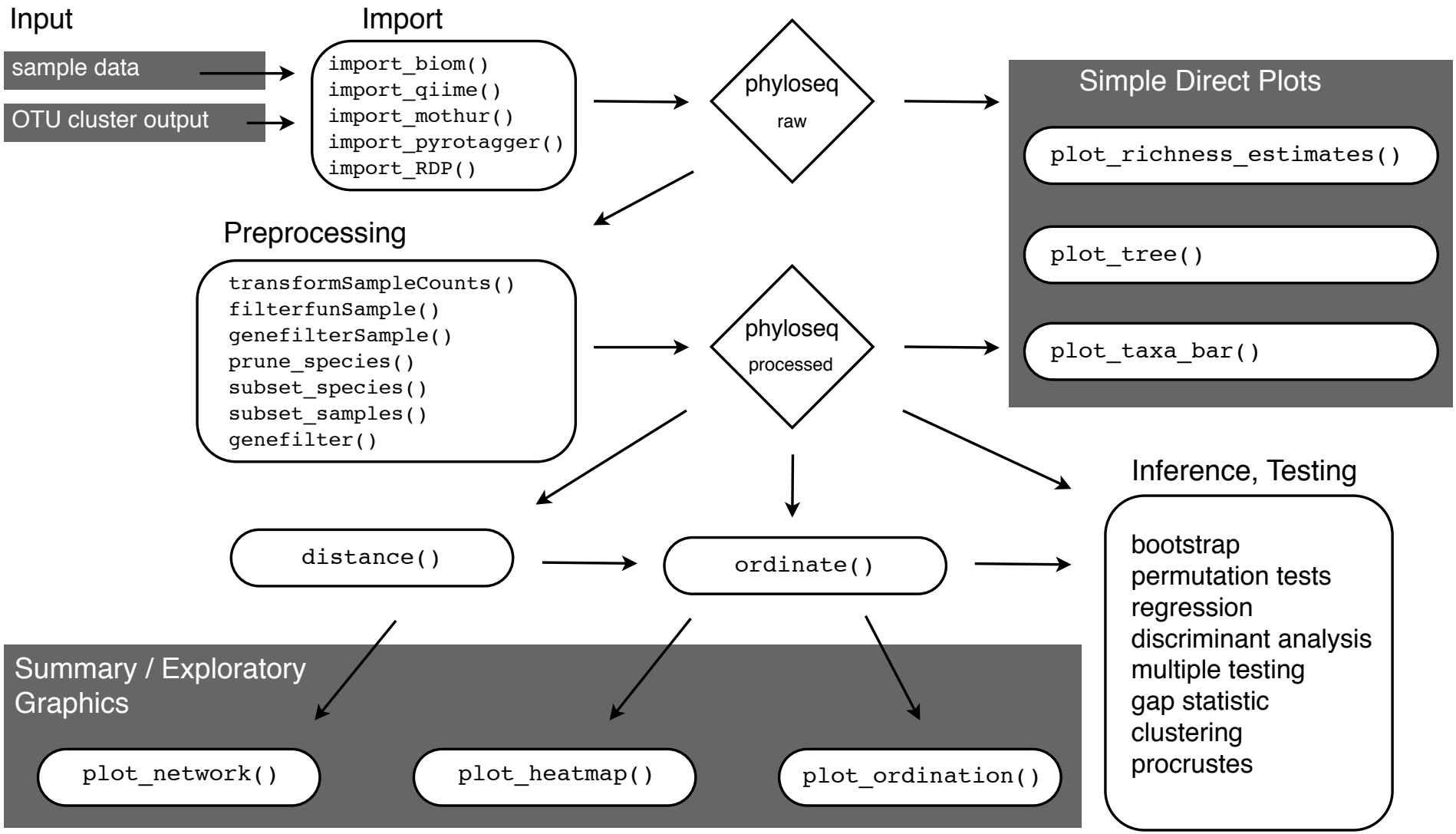
Functions for building component data objects

Function	Input Class	Output Description
`otuTable`	numeric matrix	`otuTable` object storing taxa abundance
`otuTable`	data.frame	`otuTable` object storing taxa abundance
`sampleData`	data.frame	`sampleData` object storing sample variables
`taxTab`	character string	`taxonomyTable` object storing taxonomic identities
`tre`	file path char	phylo4-class tree, read from file
`tre`	phylo-class tree	phylo4-class tree, converted from argument
`read.table`	table file path	A matrix or data.frame (Std Rcore function)
`read.tree`	Newick file path	phylo-class tree object (ape)
`read.nexus`	Nexus file path	phylo-class tree object (ape)
`readNexus`	Nexus file path	phylo4-class tree object (phylobase)

Functions for building complex data objects

Function	Input Class	Output Description
`phyloseq`	2 or more component objects	phyloseq-class, "experiment-level" object
`merge_phyloseq`	2 or more component or phyloseq-class objects	Combined instance of phyloseq-class

Overview of analysis using phyloseq



phyloseq wiki

https://github.com/joey711/phyloseq/wiki

github | Search... | Explore Gist Blog Help | joey711

joey711 / phyloseq | Admin | Pull Request | Unwatch 9 | Fork 4

Code | Network | Pull Requests 0 | Issues 25 | Wiki | Graphs

Home | Pages | Wiki History | Git Access

Home

New Page | Edit Page | Page History

Welcome to the phyloseq wiki!

This wiki is meant to provide additional support/documentation to what is already provided as in-package documentation in `phyloseq` itself. Once installed and loaded (`library("phyloseq")`), a lot of documentation is immediately available to you through the standard R documentation interface. Type `?"phyloseq-package"` in an R session to see the "front page" of this documentation. The documentation index should also provided links to phyloseq vignettes.

Quick links:

Installation

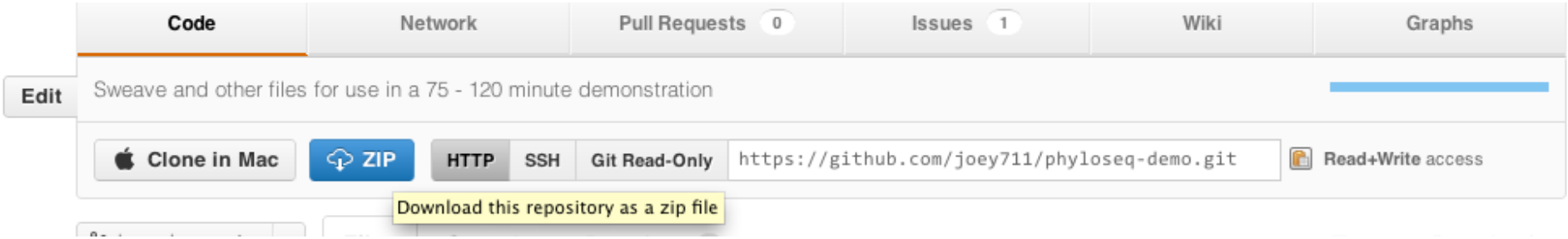
Vignettes

Graphics Examples

UniFrac

Example Data

https://github.com/joey711/phyloseq-demo



(Begin live demo)