Metagenomics Workshop Overview/Discussion

Lab Meeting

Aug 10, 2016

What was EDAMAME 2016?

Explorations in **D**ata **A**nalyses for
Metagenomic **A**dvances in **M**icrobial **E**cology

Michigan State University,
Kellogg Biological Station
July 10-20, 2016

Slide 2

Learning Goals

Ashley Shade edited this page 20 days ago · 7 revisions

Tutorials Organized by Learning Goals

- Computing literacy
  - Shell
  - tmux for remote sessions
  - Getting started with GitHub
  - Computing workflows for biologists - paper
- Cloud computing
  - Amazon EC2 start-up
  - File transfer to the EC2
  - Getting started with the EC2 - from Angus
- Microbial amplicon analysis
  - Assessing sequencing quality with FastQC (bonus: intro to automation, installing software on the EC2)
  - Subsampling a large amplicon dataset for developing an analysis workflow
  - Firing up the QIME AMI
  - QIME workflow overview
  - QIME tutorial
  - mothur workflow
- Microbial shotgun metagenome analysis
  - Examples of installing mg tools on the EC2
  - Demo: using seqtk for subsampling a large metagenome dataset for developing an analysis workflow
  - metaG sequencing preprocessing, quality control, and trimming
  - Digital normalization
  - Assembly with MEGAHit
  - Evaluating assembly
  - Estimating abundance from metagenomes
  - Binning assemblies
  - Annotation of assembled reads
  - Xander for targeted gene assembly
- Ecological Statistics with R
  - R basics
  - Visualizations Demo
  - Quick intro to R for comparative (beta) diversity
- Using Databases
  - Local BLAST
  - Using APIs to access NCBI and MG-RAST data
  - Getting data from NCBI

Slide 3

we covered a lot of things in 10 days and i can't go over all of it so i picked out the key points and will leave you with a link to the individual tutorials so you can try them on your own if you're interested

Major Workshop Themes

Microbial ecology concepts

Metagenomic concepts & tools

Data management & sharing tools

Slide 4

(in my opinion) there were three major workshop themes:

Major Workshop Themes

Microbial ecology concepts

Metagenomics concepts & tools

Data management & sharing tools

Slide 5

Microbial Ecology Concepts

OTUs = **O**perational **T**axonomic **U**nits

Common measures:
richness – number of OTUs present
evenness – abundance of OTUs
composition – taxonomic assignment of OTUs

Slide 6

There is no formal biological classification system for microorganisms and OTUs offer the closest representation of that. Practically you can think of OTUs as microbial "species" but are really groups of similar sequences that have been clustered together based on some parameters we define (usually 97% similar sequence, etc.).

Microbial Ecology Concepts (cont.)

Diversity measures:

within sample/location - includes richness & evenness, aka alpha diversity

between samples/location – aka beta diversity

regional - includes α & β, aka gamma diversity

See A. Shade's preprint: *Diversity is the question, not the answer*
https://peerj.com/preprints/2287/

Slide 7

ecologists have several ways to look at diversity

Vocabulary

OTU
reads (paired vs single, raw vs assembled)
sequencing coverage
reference database
PCR/amplification

Slide 8

before moving any further, i wanted to go through a few definitions so we're all on the same page

OTU – operational taxonomic unit (said this a few min ago)
reads –"chunk" of DNA sequence you get from the sequencing machine (usually 150bp but you can choose)
paired reads – two copies of a sequence
single read – one copy of a sequence
raw – right from sequencing center
assembled – put reads together into longer pieces (=contigs)
sequence coverage – number of times you see a certain sequence
reference database – sequences you know what/who they represent and can use to compare your own sequences to
PCR – polymerase chain reaction is a way to amplify a portion of dna so it is easier to detect b/c you have more copies

Major Workshop Themes

Microbial ecology concepts

Metagenomic concepts & tools

Data management & sharing tools

Slide 9

Activity Intro:

Before sequencing, you can do some back of the envelope calculations.

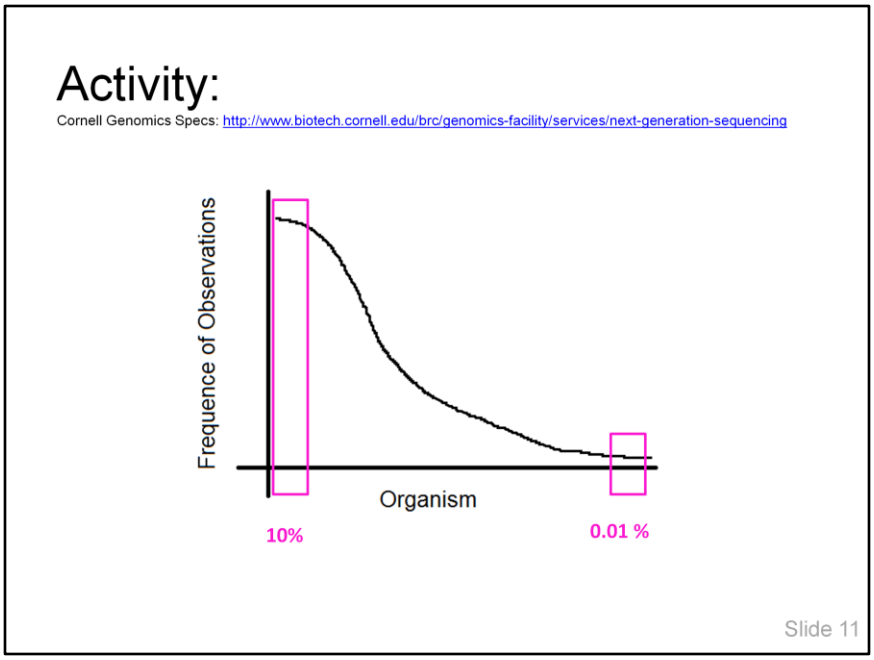
These can help you:

- Get the sequencing info you need
- Save money
- Tailor your experimental design

Slide 10

we had an intro discussion at the workshop that was helpful. we did some back of the envelope calculations to determine the extent of sequencing needed for a defined project. i thought it might be helpful to work through an example together.

doing this can help you…



when you sequence a sample the frequency of observations for each organism will typically look like this, you will catch a lot of the more common organisms and less of the rare ones, for example the more common microbes might make up on average 10% of the population while more rare ones make up 0.01%

Activity:

Cornell Genomics Specs: http://www.biotech.cornell.edu/brc/genomics-facility/services/next-generation-sequencing

You decide to do some shotgun sequencing at the Cornell Genomics Facility using their HiSeq 2500 instrument (on "Rapid Run" mode), which has an output of 35 Gbp per sequencing lane. Assuming the average genome size of an organism you're looking for is 5 Mbp and you want to make sure you see it at least 50 times, what percentage cutoff of the community would you be able to survey?

Try these steps:

1. Multiply 5 Mbp x 50 – this is the number of base pairs you'll be looking for
2. Convert #1 and the HiSeq output to bp to make things easier
3. Divide #1 (in bp) by the output per lane (in bp) - this multiplied by 100 is the percentage cutoff

Now try these with your partner:

1. If you wanted to see the organism 100 times, how would this change the percentage cutoff with one lane?
2. What would you do to survey percentage cutoff equal to that of the practice problem? that you did in the practice question?
3. How would you adjust it to see a rare organism (0.01% of the community) 50 times?

Slide 12

practice:
5Mbp x 50 = 250Mbp
250Mbp/35Gbp = 0.007 x 100 = 0.7%

if 100x you would see organisms that make up 1.4% or more.

add another lane then 5Mbp x 100 = 500 Mbp/70Gbp x 100 = 0.7%

to see a rare organism (0.1%) 50 times, 0.001x=250x106bp where x = 250 Gbp/35 Gbp per lane = ~7 lanes (max is 8 lanes for "High Output" mode)

We might ask…

How does microbial community **structure** and/or **function** affect some sort of **phenotype** or change along a gradient (space, time, perturbation, environment)?

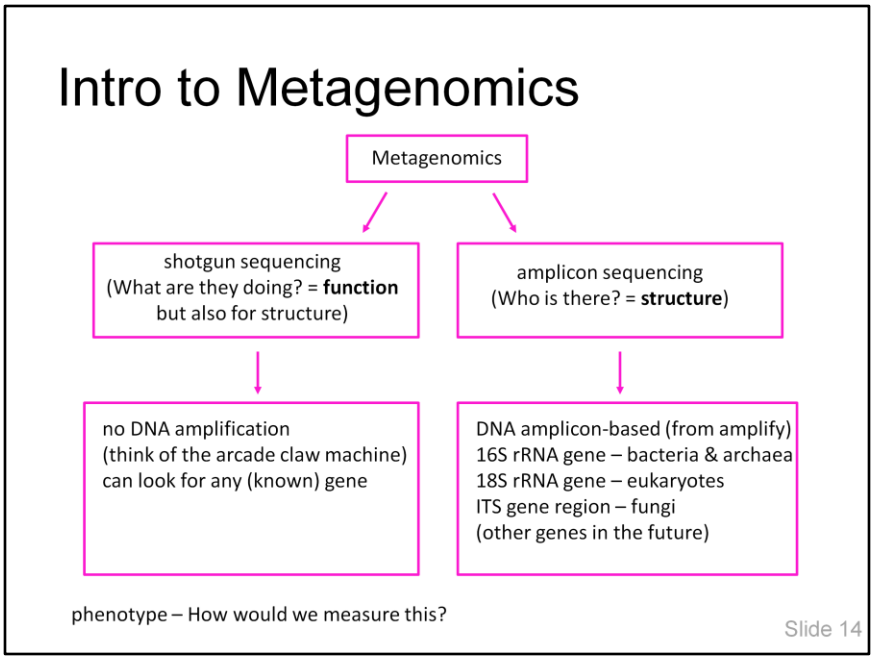
Example:

Is P availability correlated with patterns in ppk gene abundance and diversity across a soil moisture gradient?

Slide 13

most key questions in microbial ecology look something like this…

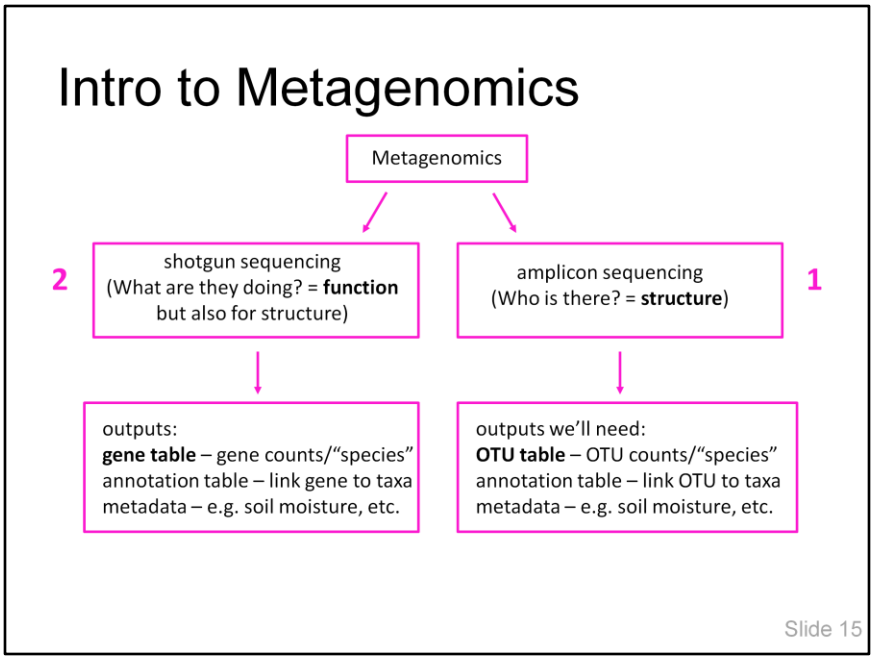
for example, in my project, i'm asking…



two general routes in metagenomics: you can focus on which microbes are there (amplicon seq) or what microbes are doing (shotgun seq)

as the name implies, amplicon sequencing involves the amplification of a target gene which varies with the organism you're trying to id, maybe eventually we can do this with other genes too nirk for example b/c the machines you use to do this are cheaper to use)

shotgun sequencing you don't amplify, can think of it as a claw arcade game where you're picking out sequences from the pile you give to the sequencing center, usually use for functional gene work when you know the gene you're interested in studying



in amplicon sequencing you're looking to generate three things…

for shotgun sequencing you're looking to generate something similar but is gene-based

Amplicon Sequencing

Who's there?/Structure

https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP_v4.md
Qiime tutorial: https://github.com/edamame-course/Amplicon_Analysis/blob/master/final/2016-07-13-QIIME1.md
mothur tutorial: http://www.mothur.org/wiki/MiSeq_SOP

Two popular programs:
Qiime & mothur

Main points:

- Replication (>5 biological reps, 9 is best)
- Include mock communities (to report errors)
- Use paired-end reads (to reduce errors)
- MiSeq runs are inexpensive (compared to HiSeq)

Slide 16

Qiime – more black box approach, easier to visualize data and contribute to development, OTU tables are made by comparing the data to a known db (db dependent)
mother – step by step approach that you can tweak, OTU tables are made by referring to the data itself (db independent)

chimeras – artificially enriched fragment sequences due to the amplification process

Shotgun Sequencing

What are they doing?/Function

https://github.com/edamame-course/Metagenome/blob/master/edamame_metagenomics_overview.pdf?raw=true

Many programs

Same main points (from slide 19) apply

Computational developments make it possible

Slide 17

Shotgun Sequencing (cont.)

What are they doing?/Function
Xander tutorial: https://github.com/edamame-course/Xander/blob/master/Xander.md

Xander

- assembly based on a functional gene of interest
- download files from FunGene (http://fungene.cme.msu.edu/) or work with RDP to make your own gene repository

Slide 18

Shotgun Sequencing (cont.)

What are they
Xander tutoria

Biogeochemical cycles

gene—contributor
amoA_AOA—Feifei Liu
amoA_AOB—RDP
buk—RDP
but—RDP
cbh1—Cheryl Kuske
chb—Fan Yang
coo5—Fan Yang
cydA—Rachel Morris
dsrA—Alexander Loy/Michael Wagner
dsrB—Alexander Loy/Michael Wagner
exc1—Fan Yang
fixN—Rachel Morris
glx—Qichao Tu
hydA—Fan Yang
lcc_ascomycetes—Chris Wright
lcc_basidiomycetes—Chris Wright
ligE—Ryan Penton
lip—Qichao Tu
mcrA—Blaz Stres
mmoX—Qichao Tu
mnp—Qichao Tu
nag3—Fan Yang
napA—Laurent Philippot
narG—Laurent Philippot
nifD—RDP
nifH—RDP
nirA—RDP
nirB—RDP
nirK—Tracy Teal
nirS—Veronica Gruntzig
norB—Gesche Braker
nosZ—Blaz Stres
nosZ_atypical_1—Robert Sanford
nosZ_atypical_2—Robert Sanford

ileS—Scott Santos/Howard Ochman
lepA—Scott Santos/Howard Ochman
leuS—Scott Santos/Howard Ochman
pyrG—Scott Santos/Howard Ochman
recA—Scott Santos/Howard Ochman
recG—Scott Santos/Howard Ochman
rplB—Scott Santos/Howard Ochman
rpoB—Scott Santos/Howard Ochman

Biodegradation

gene—contributor
alkb—Gerben Zylstra/Elyse Rodgers-Vieira
benA—Stephan Gantner
bph—Gerben Zylstra
bphA1—Stephan Gantner
bphA2—Stephan Gantner
BSH—Robert Stedtfeld
carA—Shoko Ivai
cntA—Robert Stedtfeld
cutC—Robert Stedtfeld
dbfA1—Shoko Ivai
dxnA—Shoko Ivai
dxnA-dbfA1—Tim Johnson
HSDH—Robert Stedtfeld
npah—Gerben Zylstra
p450—Gerben Zylstra/Elyse Rodgers-Vieira
ppah—Gerben Zylstra
PSA—Robert Stedtfeld

Metal Cycling

gene—contributor
arsA—PFAM
arsB—PFAM
arsC—PFAM
arsD—PFAM

Slide 19

Xander gene repositories that might be interesting to our lab

Sequencing Data

Comes to you as a .fasta or .fastq file

(4 lines per read, millions+ of reads per file)

@SMS01_R1
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;7;;;;;;;88

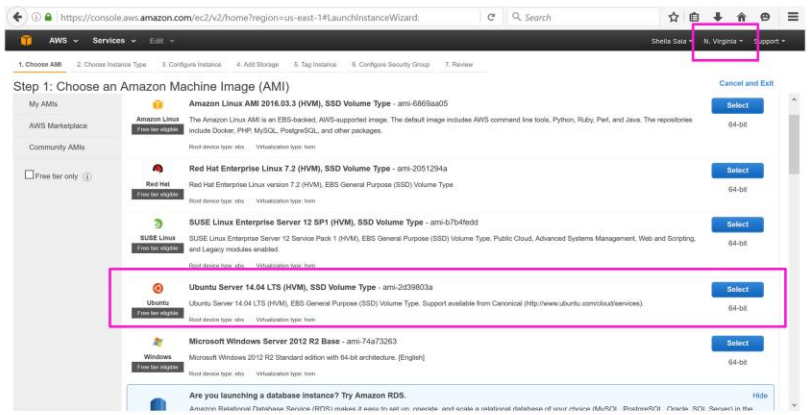
Slide 20

sequence id
sequence
spacer
quality score

most of the programs for analyzing sequence data have to be run in linux, we got some experience navigating around the linux command line

Amazon Cloud Computing

More info: https://aws.amazon.com



Slide 22

for people like us who might not have a dedicated pc for sequencing analysis, we can use the amazon cloud to run our analysis on a computer that is more powerful than our own laptop, we can select the computer and hardware we want – in most cases we want to use a linux computer with the ubuntu operating system

Major Workshop Themes

Microbial ecology concepts

Metagenomic concepts & tools

Data management & sharing tools

Slide 23

Paper Discussion

Link: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002303

1. Take-away messages for hydrologists?
2. Are you using an reproducible workflows? What works and what doesn't?
3. Anything else you agree/disagree with?

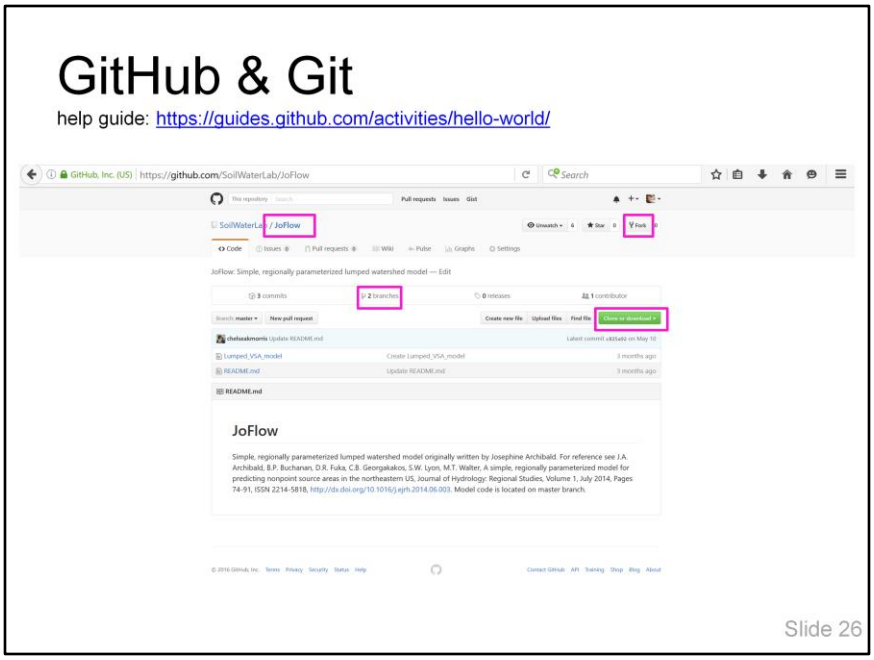
Slide 24

Specific Data Mgmt. Tools

Reproducible Research
- GitHub
- R/RStudio (scripting)
- (markdown, Jupyter notebooks)

Version Control
- Git

Info Sharing/Open Source
- GitHub Repositories
- Mendeley paper groups
- Etherpad

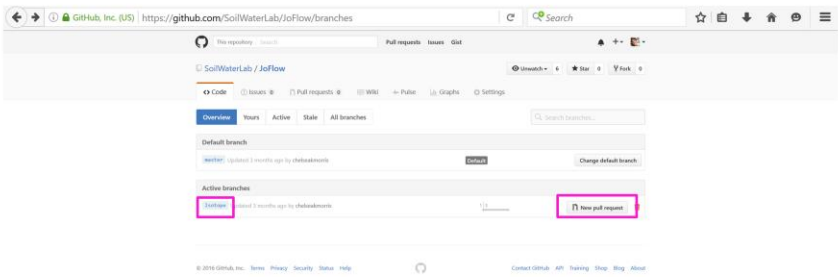


here's an example of repository on the soil and water lab github account, we can copy/clone it to our own computer or we can start a new branch if we want to modify it in another way to modify we would want to use the fork button to create a new branch

git is a version control software that is hosted by github, github is a collection of repositories for storing and sharing code/data/etc.

GitHub & Git

help guide: https://guides.github.com/activities/hello-world/



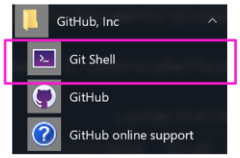
Slide 27

here we can see the branches for the joflow repository, there is the main branch and there is a forked branch where james has been working on a version of the model that includes stable isotope fractionation, this can eventually be joined back to the main branch or if it becomes very different it can be moved to a new repository, to request access to this second branch you can submit a pull request by clicking on the 'new pull request' button

Command Line GitHub & Git

workshop guide: https://github.com/edamame-course/Github/blob/master/Tutorial.md

```
posh-git ~ 2016-tutorials [master]
C:\Users\sheila\Documents\GitHub> ls

    Directory: C:\Users\sheila\Documents\GitHub

Mode                LastWriteTime         Length Name
----                -------------         ------ ----
d-----        7/12/2016  11:40 AM                2016-tutorials
d-----        7/14/2016   4:21 PM                Jones_R_Files
d-----        7/18/2016   1:28 PM                PAPER_LeeSorensen_inprep

C:\Users\sheila\Documents\GitHub> cd .\2016-tutorials
C:\Users\sheila\Documents\GitHub\2016-tutorials [master ≡]> ls

    Directory: C:\Users\sheila\Documents\GitHub\2016-tutorials

Mode                LastWriteTime         Length Name
----                -------------         ------ ----
d-----        7/12/2016  11:05 AM                HandoutsResources
d-----        7/12/2016  11:05 AM                images
d-----        7/12/2016  11:05 AM                lectures
d-----        7/12/2016  11:48 AM                test
d-----        7/12/2016  11:05 AM                Xander
-a----        7/12/2016  11:05 AM           5193 databasesExercise.md
-a----        7/12/2016  11:19 AM            199 README.md

C:\Users\sheila\Documents\GitHub\2016-tutorials [master ≡]>
```



> git status
> git pull

> git add
> git commit
> git push

Slide 28

you can use your mouse to download and update respositories using the github desktop app but you can also use all the git commands to update a github repository in the git shell (command line), you download the git shell when you download the github desktop app

Tutorials Hosted on GitHub

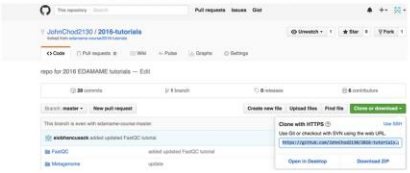
Using markdown (.md)

3. Fork and Clone a repository

The main benefit of forking a repository is that it allows you to copy an existing repository and then have the ability to make your own edits to the repository without changing the original repository.

- Navigate to the EDAMAME 2016-tutorials repository.
- On the upper right-hand side there is a box labeled "Fork". Click on that.

- You'll be re-directed to your github account. The repository is now on your github account but we still need to clone the repository so we have local access to the files.
- In the new window, on the right hand side there is a box labeled "Clone or download".



- Click on that and then copy the link.
- Choose a local directory that you want this repository to be added to. Change into that directory and use git clone with the URL just copied:
- **Note that the below command will not work for you because you need to appropriately edit the URL**

```
git clone https://github.com/**YourGitHubName**/2016-tutorials.git
```

- Directions for 'git clone' can also be found at GitHub.
- This protocol can be used to clone any public repository. For EDAMAME repos, you can `pull` to get the most up-to-date materials from GitHub, but you cannot `push` to edit those resources and have your edits tracked to the main repository, because you are not part of the EDAMAME team. (More details on these commands below).

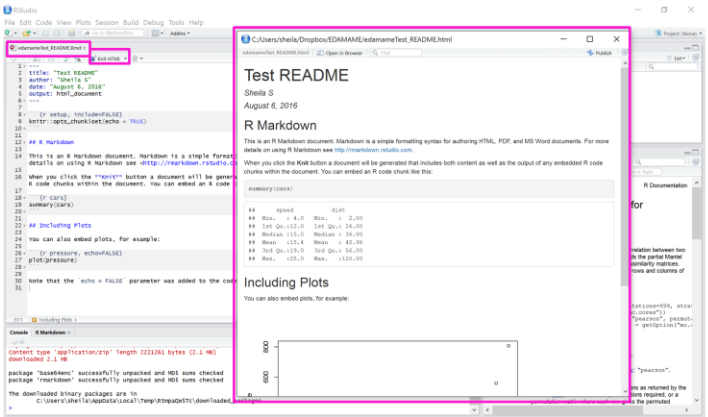
Slide 29

you can post tutorials on github!

R Markdown

help guide: http://rmarkdown.rstudio.com/lesson-1.html

RStudio > File > New File > R Markdown…

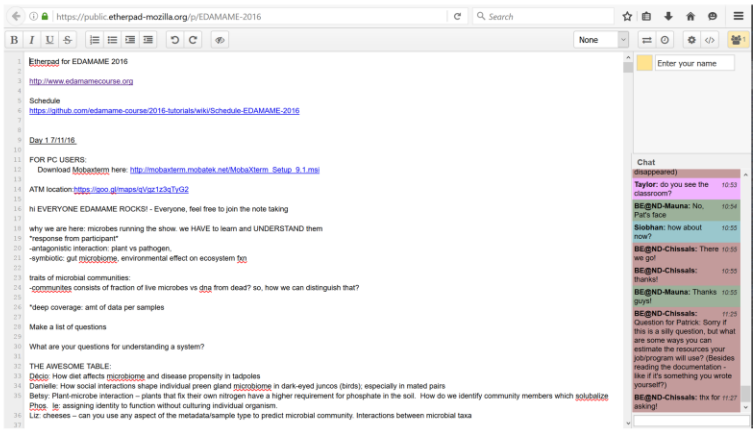


Slide 30

you can make nice looking readme files in RStuio using the R Markdown templates (.Rmd files), follow the path above to start a new template, you might have to install the rmarkdown package before you can proceed

Etherpad

Create a new Etherpad: https://public.etherpad-mozilla.org/
Etherpad info: https://github.com/ether/etherpad-lite#installation

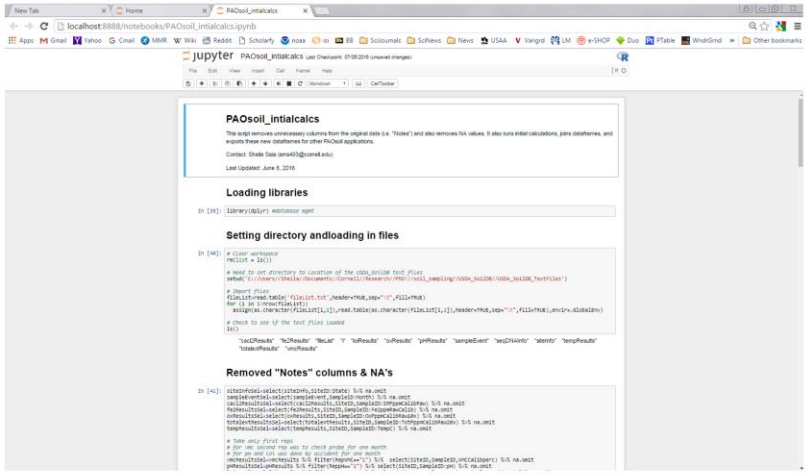


Slide 31

etherpad is a lot like google docs, it's a way to collaboratively take notes or write in real time, there's also a chat box on the side to you can ask questions if you are all working together with the rest of your group.

Jupyter Notebooks

Installation & help: http://jupyter.readthedocs.io/en/latest/install.html



Slide 32

these are another way to keep track of what you do on the computer, you can think of it as a lab notebook but for computational related work – coding/sequencing analysis/etc.

Other Key Themes

Research goals/questions shape your path

Try to focus on hypothesis driven studies

Planning ahead

No need to reinvent the wheel
(wrt. ecological tools & computer programs)

Use statistical tests to relate who/what with process

Slide 33

Other Resources

EDAMAME Tutorials by Subject
https://github.com/edamame-course/2016-tutorials/wiki/Learning-Goals

EDAMAME Schedule (with tutorial links)
https://github.com/edamame-course/2016-tutorials/wiki/Schedule-EDAMAME-2016

MiSeq (Amplicon Sequencing) SOP
https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP_v4.md

Illumina Sequencing Video (paired reads)
https://www.youtube.com/watch?v=womKfikWlxM

de Bruijn graph explanation (used in Xander):
http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Project Templates
http://projecttemplate.net/getting_started.html

R Studio Cheatshes
https://www.rstudio.com/resources/cheatsheets/

Not So Standard Deviation podcast
https://soundcloud.com/nssd-podcast

Extras

Slide 35

www.illumina.com/systems/hiseq_2500_1500/performance_specifications.html

Systems / HiSeq 2500 / Specifications

HiSeq 2500

- Overview
- System
- Applications
- Publications
- Featured Researchers
- Kits
- Specifications
- Workflow
- Literature
- Options & Accessories
- Scientific Data
- Technology
- Software
- Support

Interested in receiving newsletters, case studies, and information on new applications? Enter your email address below.

* First Name:
* Last Name:
* Email:
* Area of Interest: Select...
* Job Function: Select...
* Country: Select...

Sign Up

REQUEST PRICING QUESTIONS

HiSeq 2500 Specifications

HiSeq System Performance Parameters

High Output Run Mode*

	HISEQ SBS V4 SPECIFICATIONS			TRUSEQ SBS V3		
Read length	Dual Flow Cell	Single Flow Cell	Dual Flow Cell Run Time	Dual Flow Cell	Single Flow Cell	Dual Flow Cell Run Time
1×36	128-144 Gb	64-72 Gb	29 hrs	95-105 Gb	47-52 Gb	2 days
2×50	360-400 Gb	180-200 Gb	2.5 days	270-300 Gb	135-150 Gb	5.5 days
2×100	720-800 Gb	360-400 Gb	5 days	540-600 Gb	270-300 Gb	11 days
2×125	900-1 Tb	450-500 Gb	6 days	N/A	N/A	N/A
Reads Passing Filter (8 lanes per flow cell)	Up to 4 billion single read or 8 billion paired-end reads	Up to 2 billion single read or 4 billion paired-end reads		Up to 3 billion single read or 6 billion paired-end reads	Up to 1.5 billion single read or 3 billion paired-end reads	
Quality	Greater than 85% of bases above Q30 at 2×50 bp Greater than 80% of bases above Q30 at 2×100 bp Greater than 80% of bases above Q30 at 2×125 bp			Greater than 85% of bases above Q30 at 2×50 bp Greater than 80% of bases above Q30 at 2×100 bp		

*Install specifications based on Illumina PhiX control library at supported cluster densities (between 610-678 K clusters/mm^2 passing filter using TruSeq v3 Kits or 870-930 K clusters/mm^2 passing filter using HiSeq v4). Run times for high output mode correspond to sequencing only. Performance may vary based on sample quality, cluster density, and other experimental factors.

Rapid Run Mode*

	HISEQ RAPID SBS KIT V2 SPECIFICATIONS		
Read length	Dual Flow Cell	Single Flow Cell	Dual Flow Cell Run Time
1×36	18-22 Gb	9-11 Gb	7 hr
2×50	50-60 Gb	25-30 Gb	16 hr
2×100	100-120 Gb	50-60 Gb	27 hr
2×150	150-180 Gb	75-90 Gb	40 hr
2×250	250-300 Gb	125-150 Gb	60 hr
Reads Passing Filter (2 lanes per flow cell)	Up to 600 million single read or 1.2 billion paired-end reads	Up to 300 million single read or 600 million paired-end reads	
Quality	Greater than 85% of bases above Q30 at 2×50 bp Greater than 80% of bases above Q30 at 2×100 bp Greater than 75% of bases above Q30 at 2×250 bp		

*Install specifications based on Illumina PhiX control library at supported cluster densities (between 700-820 K clusters/mm^2 passing filter using

Slide 36

hiseq 2500 specs for rapid and high output modes

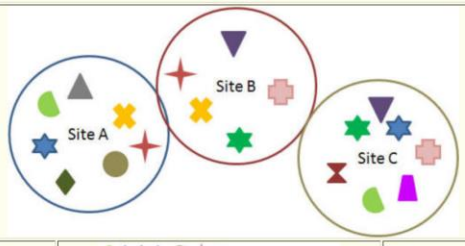
Biodiversity Can be Expressed at Several Scales

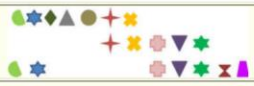
Biodiversity can be measured and monitored at several spatial scales.

Alpha Diversity = richness and evenness of individuals within a habitat unit. For example in the figure below, **Alpha Diversity** of Site A = 7 species, Site B = 5 species, Site C = 7 species.

Beta Diversity = expression of diversity between habitats. In the example below, the greatest **Beta Diversity** is observed between Site A and C with 10 species that differ between them and only 2 species in common.

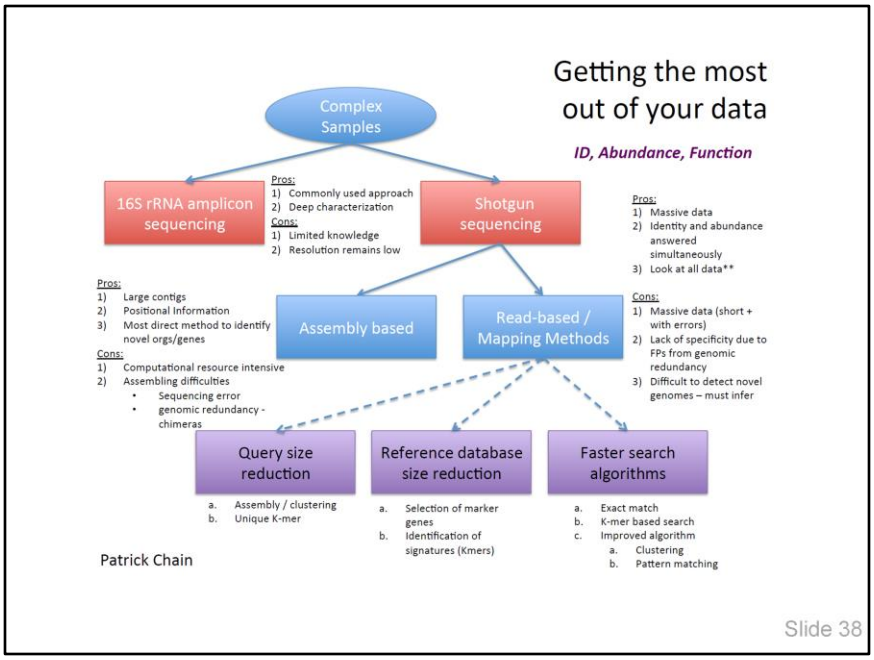
Gamma Diversity = landscape diversity or diversity of habitats within a landscape or region. In this example, the gamma diversity is 3 habitats with 12 species total diversity.



Site A = 7 Species Site B = 5 Species Site C = 7 Species		A vs B = 8 species B vs C = 4 species A vs C = 10 species

http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

Slide 37



pros and cons of different sequencing approaches