

Dr. Richard Allen White III
Washington State University
FABI Workshop - Oct 21, 2018

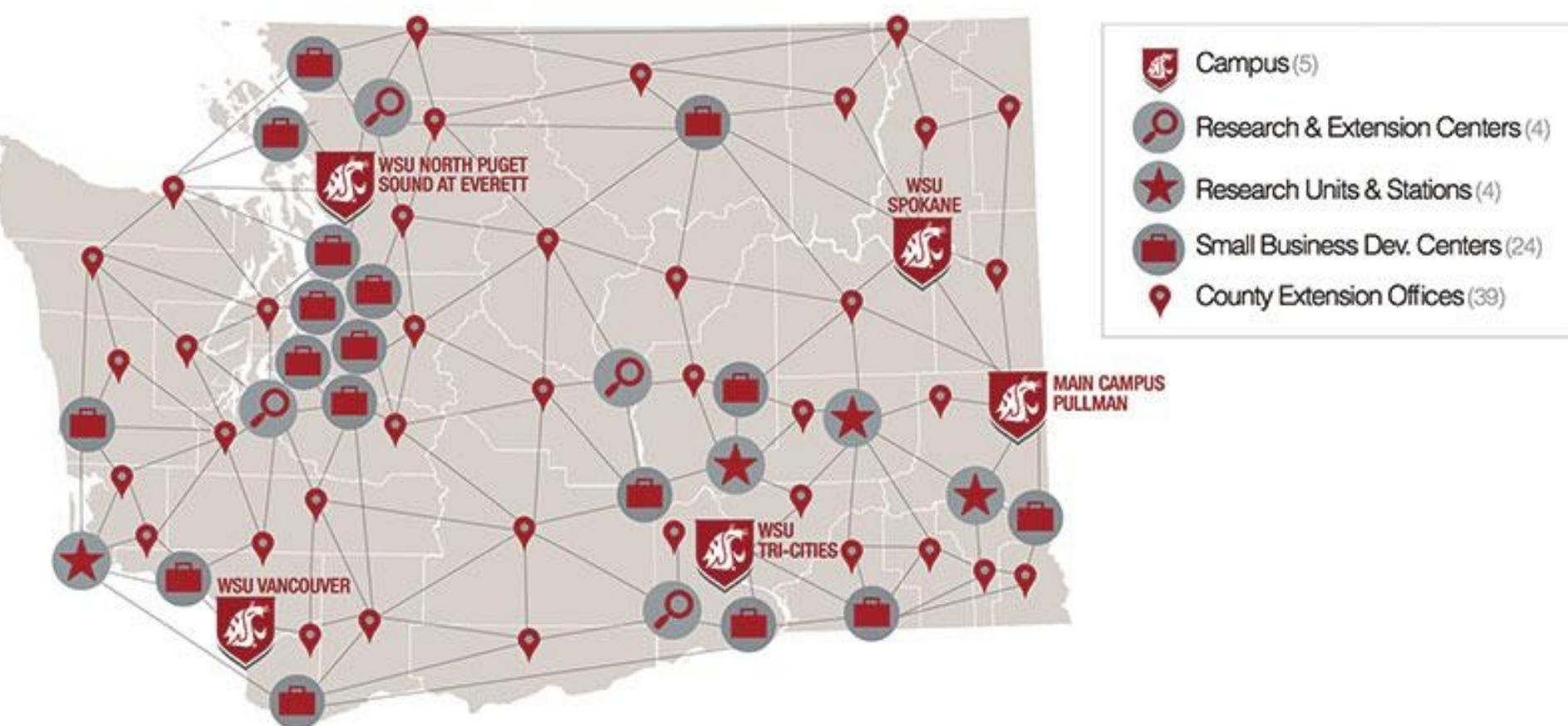
Learning Objectives

1. Brief introduction of me and my research
2. Learn about 'Omics,' technology terms
3. Introduction to multiomics
4. History of genome sequencing
5. Introduction to sequencing technologies
6. Unix/Linux command line
7. Grep -> text wrangling
8. Github
9. R intro
- 10.R ggplot2 intro

Introduction - Dr. RAWIII

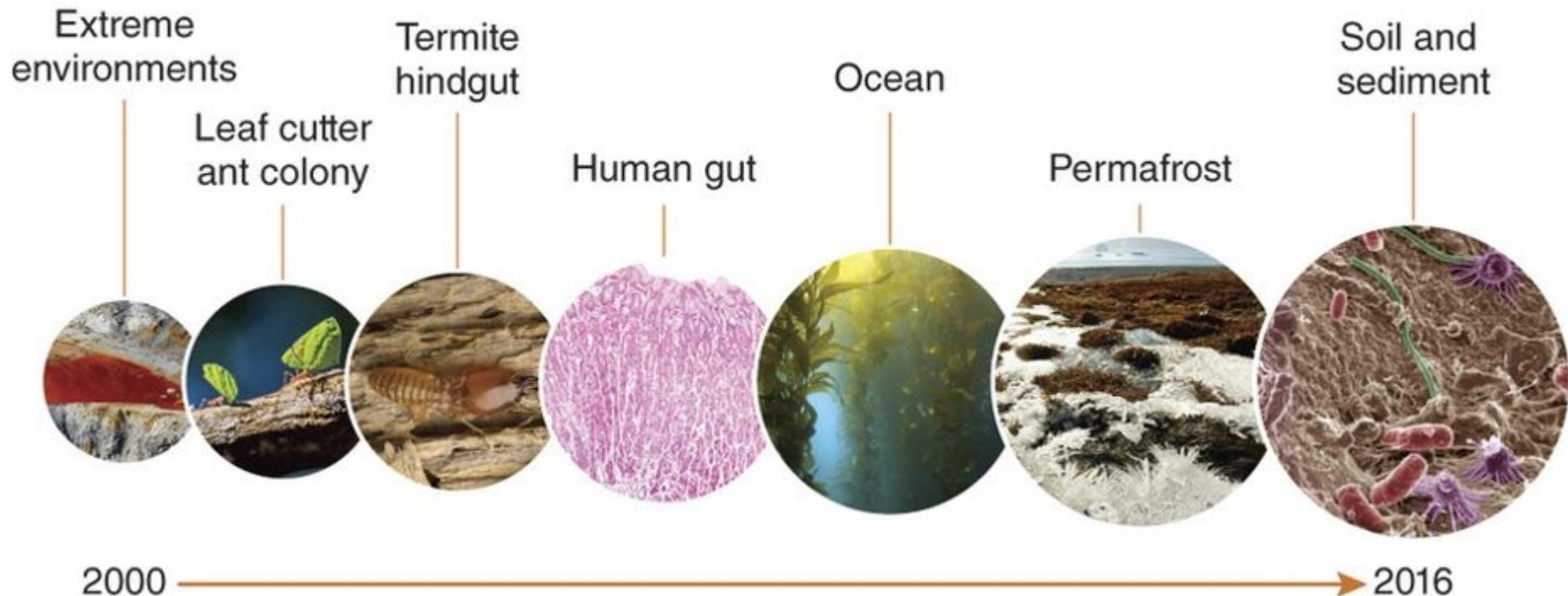


Introduction - Dr. RAWIII



Research - Dr. RAWIII

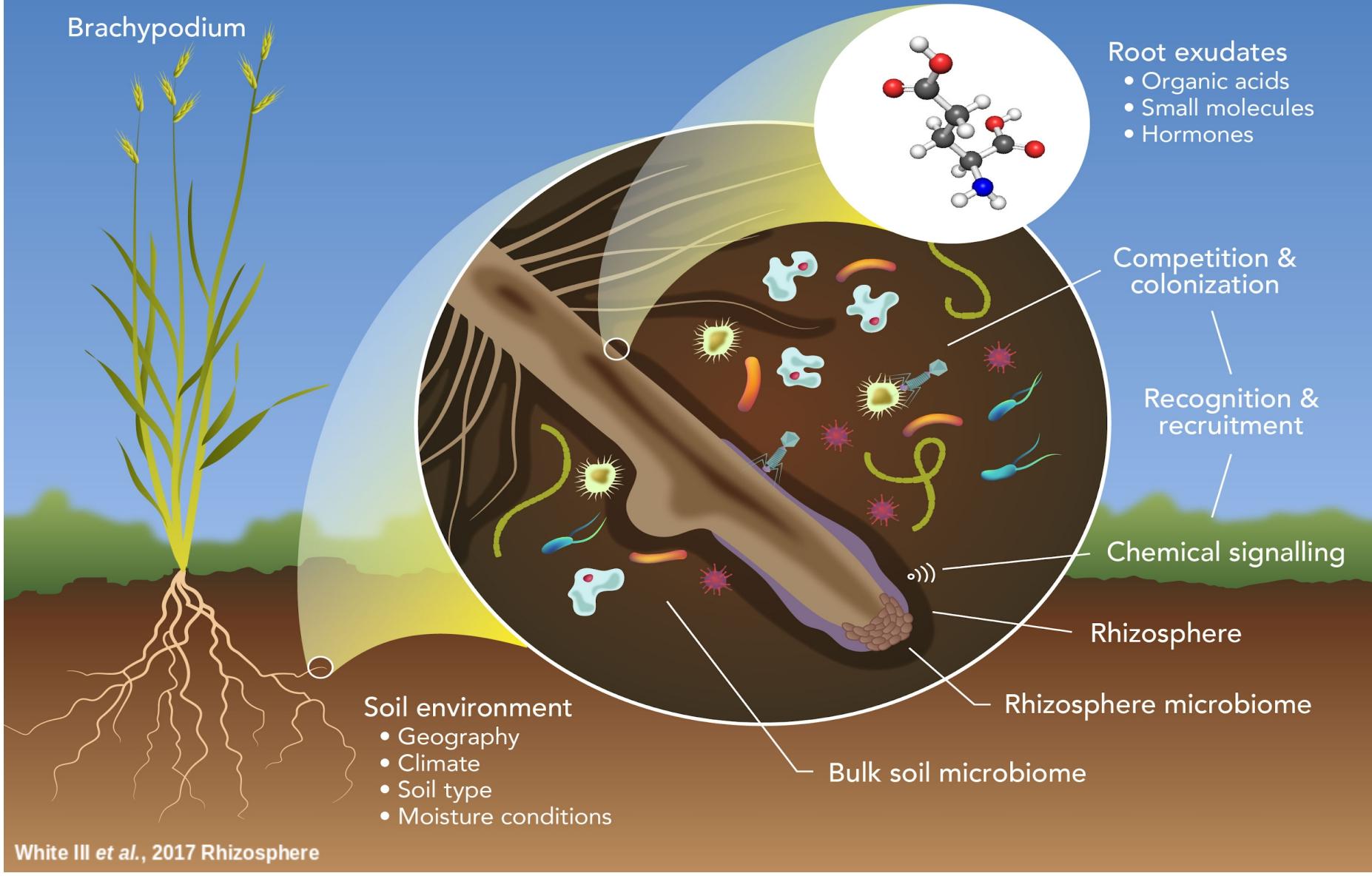
Microbiome complexity and multi-omics analysis timeline



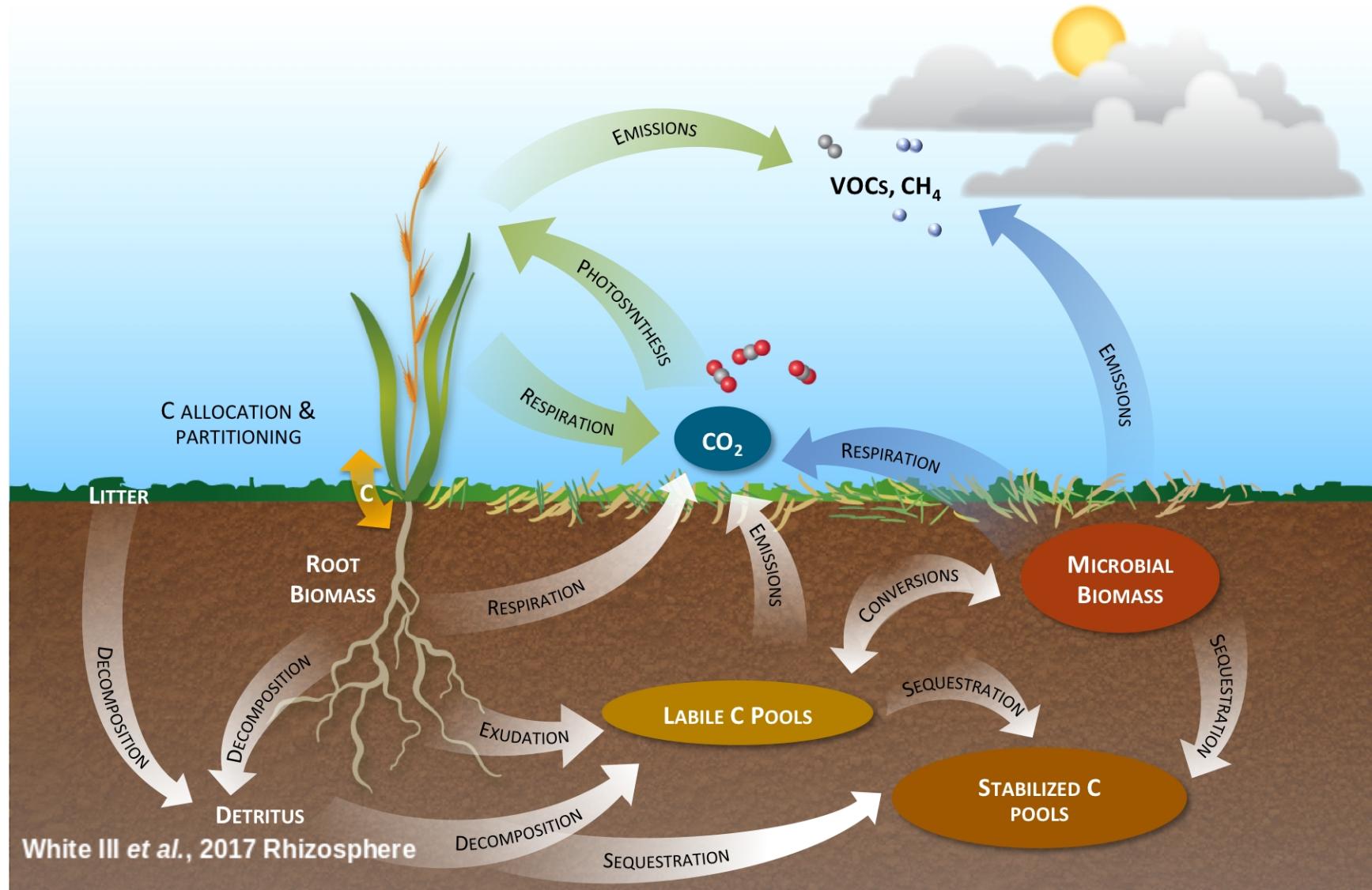
White III et al., 2016. Nature Protocols

Link to article: <https://www.nature.com/articles/nprot.2016.148>

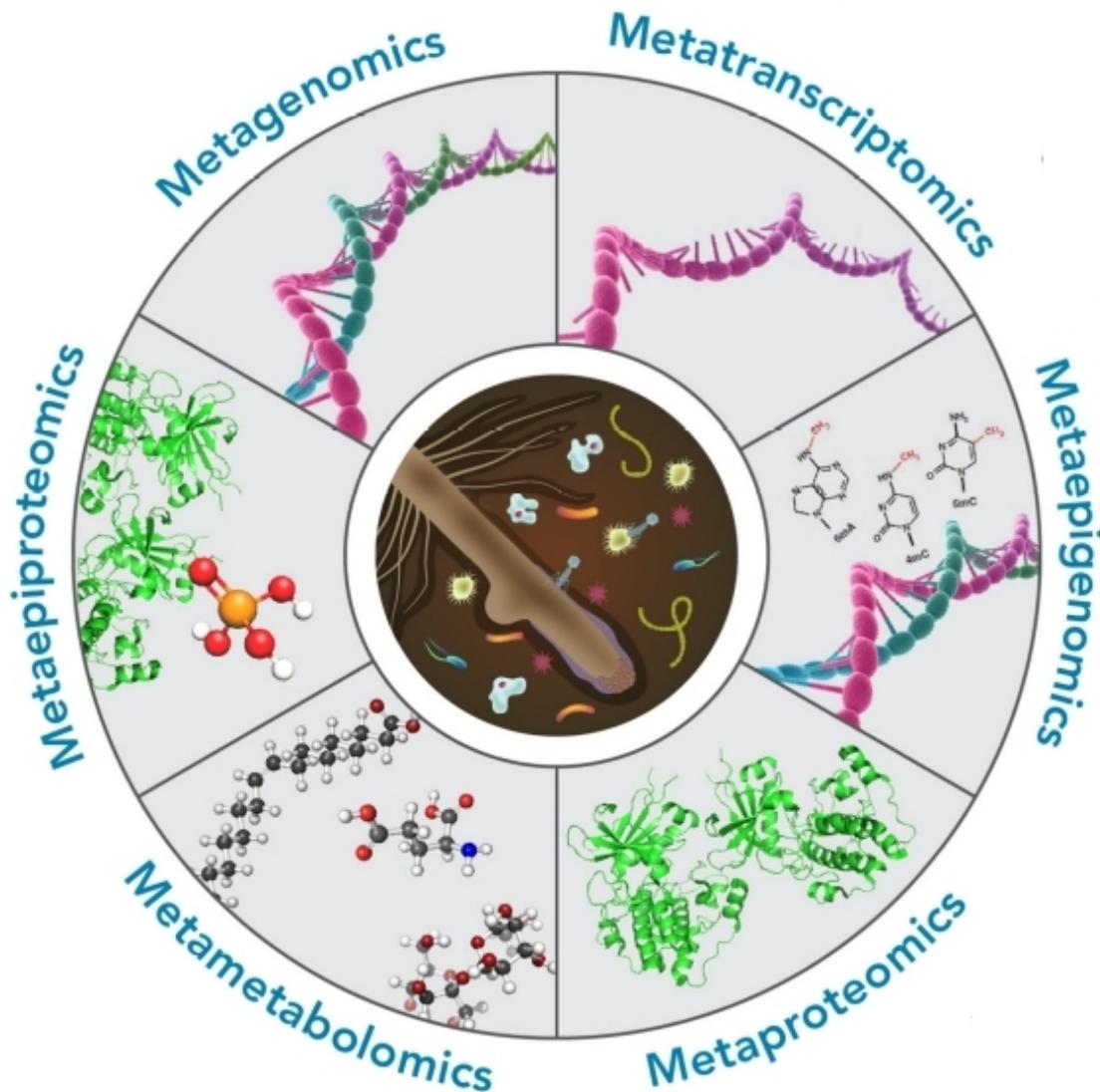
Research - Dr. RAWIII



Research - Dr. RAWIII



Omics terms



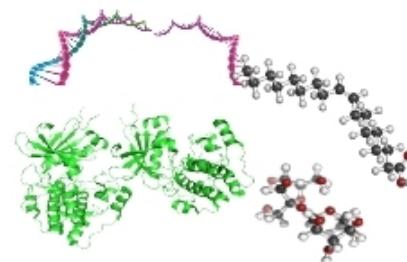
A workflow of a multiomic study



Environmental sample



Single step -
multiomic extraction



Library prep (NGS)
Fractionation (MS)



Sequencing



Mass Spectrometry



Data

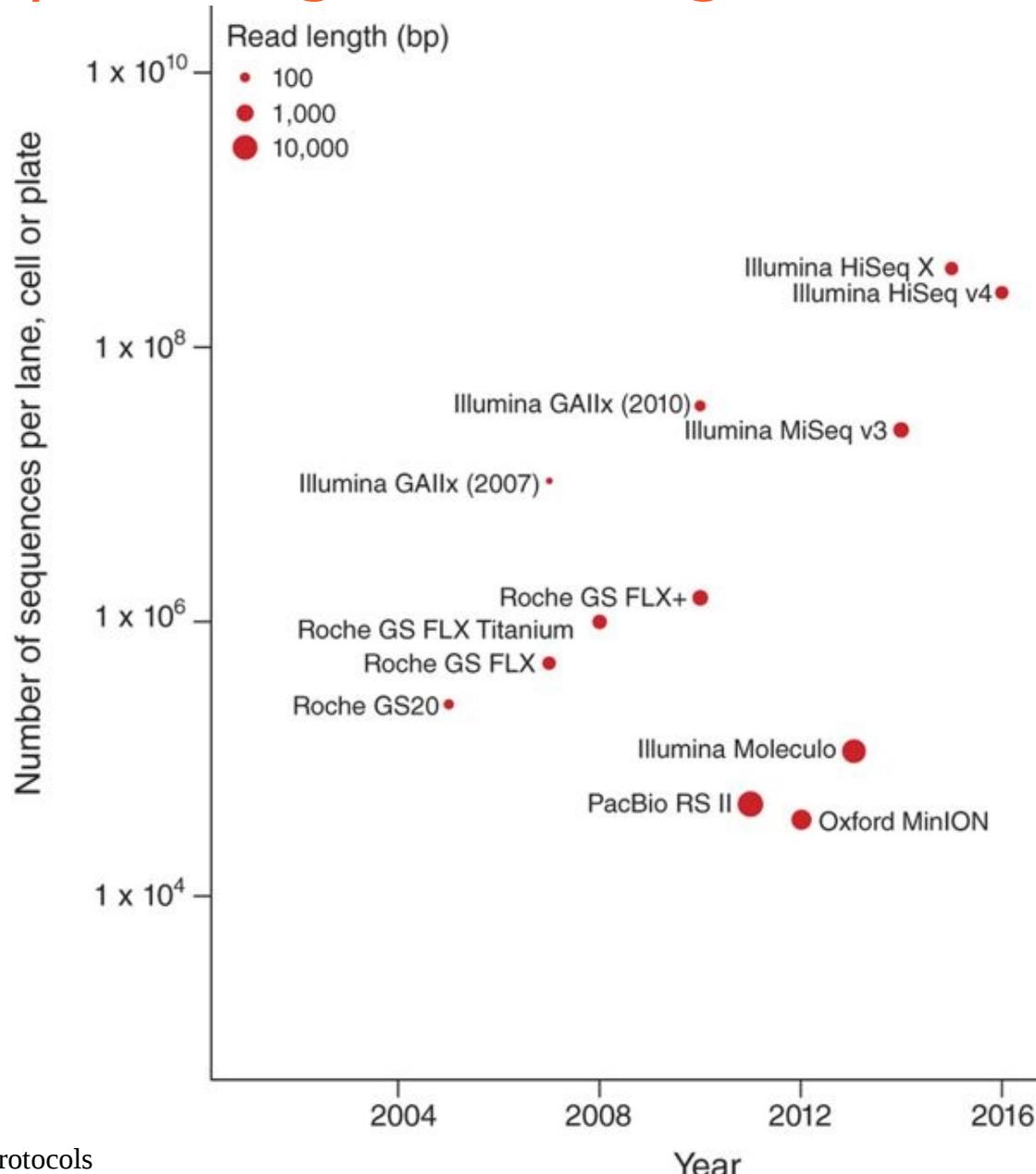


Analysis

Direct Read or
Assembly (NGS)

Spectral features
analysis

Sequencing technologies



History of genome sequenced



- 1972 - First gene to be sequenced was MS2 coat protein (510 bp)
- 1976 - First RNA genome sequenced was Bacteriophage MS2 (3.56 kb)
- 1977 - First DNA genome sequenced was Bacteriophage phiX174 (5.38 kb)
- 1979 - First plasmid to be sequenced was pBR322 (4.3 kb)
- 1984 - First mammalian DNA virus sequenced was Epstein-Barr virus (172 kb)
- 1992 - First chromosome sequenced was Yeast chromosome III (315 kb)
- 1995 - First cellular genome sequenced was *Hemophilus influenzae* (1.8 Mb)
- 1995 - First archaeal genome sequenced was *Methanococcus jannaschii* (1.7 Mb)
- 1996 - First eukaryotic genome sequenced was *Saccharomyces cerevisiae* (12 MB)
- 1998 - First multicellular organism sequenced was *Ceanorhabditis elegans* (97 Mb)
- 2000 - First plant genome sequenced was *Arabidopsis thaliana* (125 Mb)
- 2001 - First mammalian draft genome *Homo sapiens* (3.3 Gb)

Super bonus -> Who sequenced the first cellular genome?



Data format

*.fna, .fasta, .fa ?

*.faa ?

*.gff ?

*.gbk ?

*.fq or fastq ?

*.gtf ?

Read vs. sequence?

Contig?

kmer?



What is Unix?

A multi-task and multi-user Operating System

Developed in 1969 at AT&T's Bell Labs by:

Ken Thompson (Unix)

Dennis Ritchie (C)

Douglas McIlroy (Pipes)

Some other variants: System V, Solaris, SCO Unix, SunOS, 4.4BSD, FreeBSD, NetBSD, OpenBSD, BSDI



What is Linux?

A clone of Unix

Developed in 1991 by Linus Torvalds, a
Finnish graduate student

Inspired by and replacement of Minix

Linus' Minix became Linux

Consist of :

Linux Kernel

GNU (GNU is Not Unix) Software

Software Package management

Others

What is Linux?

Originally developed for 32-bit x86-based PC Ported to other architectures, eg.

-> Alpha, VAX, PowerPC, IBM S/390, MIPS, IA-64 PS2, TiVo, cellphones, watches, Nokia N810, NDS, routers, NAS, GPS, ...





What is Linux?

> 300 Linux Distributions

Ubuntu (**Best for newbies, my fav, South African!**)

Slackware

Redhat RHEL (commercially support)

Fedora (free)

CentOS (free RHEL, based in England)

Debian (one of the few called GNU/Linux)

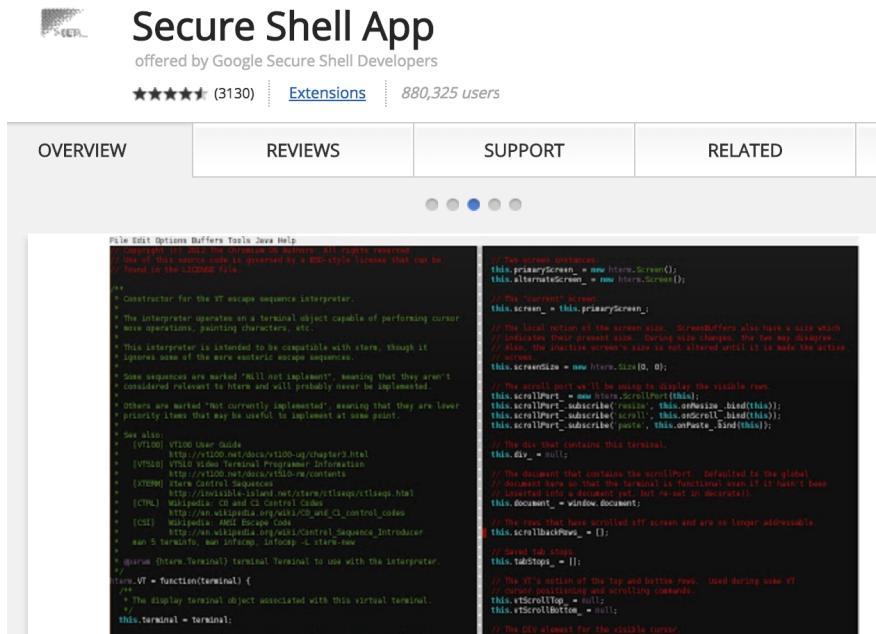
Knoppix (first LiveCD distro.) ...

Terminal for Unix/Bash

Mac: Terminal

Windows: PUTTY

Chrome: Secure Shell (extension; runs command line in your browser! How cool is that?)





Unix/Bash tutorial

>>Master page

https://github.com/raw937/FABI_workshop

https://github.com/raw937/FABI_workshop/blob/master/bin/Bash_tutorial.md

Command line review

cd	?	tar	?
pwd	?	scp	?
mkdir	?	rm	?
head	?	rm -r	?
tail	?	echo	?
more	?		?
less	?	>	?
mv	?	<	?
touch	?	>>	?
cp	?	cat	?

Unix/Bash - check point

- Make a folder with your name
- Make two files in as .txt (labeled 1 and 2)
- Write your name in file 1
- Write your name in file 2 with _a at the end
- Combine those two files to a new file
- Copy the whole folder of your name with a new name



Grep “the hand of the gods”

https://github.com/raw937/FABI_workshop/blob/master/bin/grep_tutorial.md

Grep “the hand of the gods”

<https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=ATCL01#contigs>

NCBI Resources How To Sign in to NCBI

Sequence Set Browser Show help

Project: ATCL01 List of all Projects

ATCL00000000.1 Exiguobacterium chiriquhucha RW-2

Master Contigs Proteins Download

GenBank: ATCL01.1.gbff.gz 2 Mb
FASTA: ATCL01.1.fsa_nt.gz 910.3 kb
ASN.1: ATCL01.1.bbs.gz 1.7 Mb

GETTING STARTED

NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES

Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Variation

POPULAR

PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED

Genetic Testing Registry
GenBank
Reference Sequences
Gene Expression Omnibus
Genome Data Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION

About NCBI
Research at NCBI
NCBI News & Blog
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube
Privacy Policy

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
[Policies and Guidelines](#) | [Contact](#)





Grep “the hand of the gods”

NCBI Resources How To Sign in to NCBI

Sequence Set Browser Show help

Project: ATCL01 List of all Projects

ATCL00000000.1 Exiguobacterium chiriquchacha RW-2

GenBank: [ATCL01.1.gbff.gz](#) 2 Mb
FASTA: [ATCL01.1.fsa_nt.gz](#) 910.3 kb
ASN.1: [ATCL01.1.bbs.gz](#) 1.7 Mb

1. Download fasta file
2. Unzip the file
3. Rename file to “.fasta”
4. Give me the number of lines
5. Tell me the size of the file in bytes
6. Count the number of sequences using a command
7. Export all the headers into a new “.txt” file



Git tutorial

Make an account

<https://github.com/>

Follow tutorial on

https://github.com/raw937/FABI_workshop/edit/master/bin/git%20tutorial.md

R, Rstudio and Rnotebooks

Install Rstudio

<https://www.rstudio.com/products/rstudio/download/>

The screenshot shows the RStudio interface with the following components:

- Top Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Left Sidebar:** Untitled2* (active), Untitled3, Go to file/function, Addins.
- Code Editor (R Notebook):** Displays an R Markdown notebook with code and explanatory text. Lines 1-19 show the initial setup of the notebook, including the title, output type, and a note about previewing results.
- Environment Pane:** Shows the Global Environment, which is currently empty.
- Console Pane:** Displays R command history and output. It shows the installation of the dplyr package and some initial R session logs.
- Bottom Navigation:** Files, Plots, Packages, Help, Viewer.

R, Rstudio and Rnotebooks

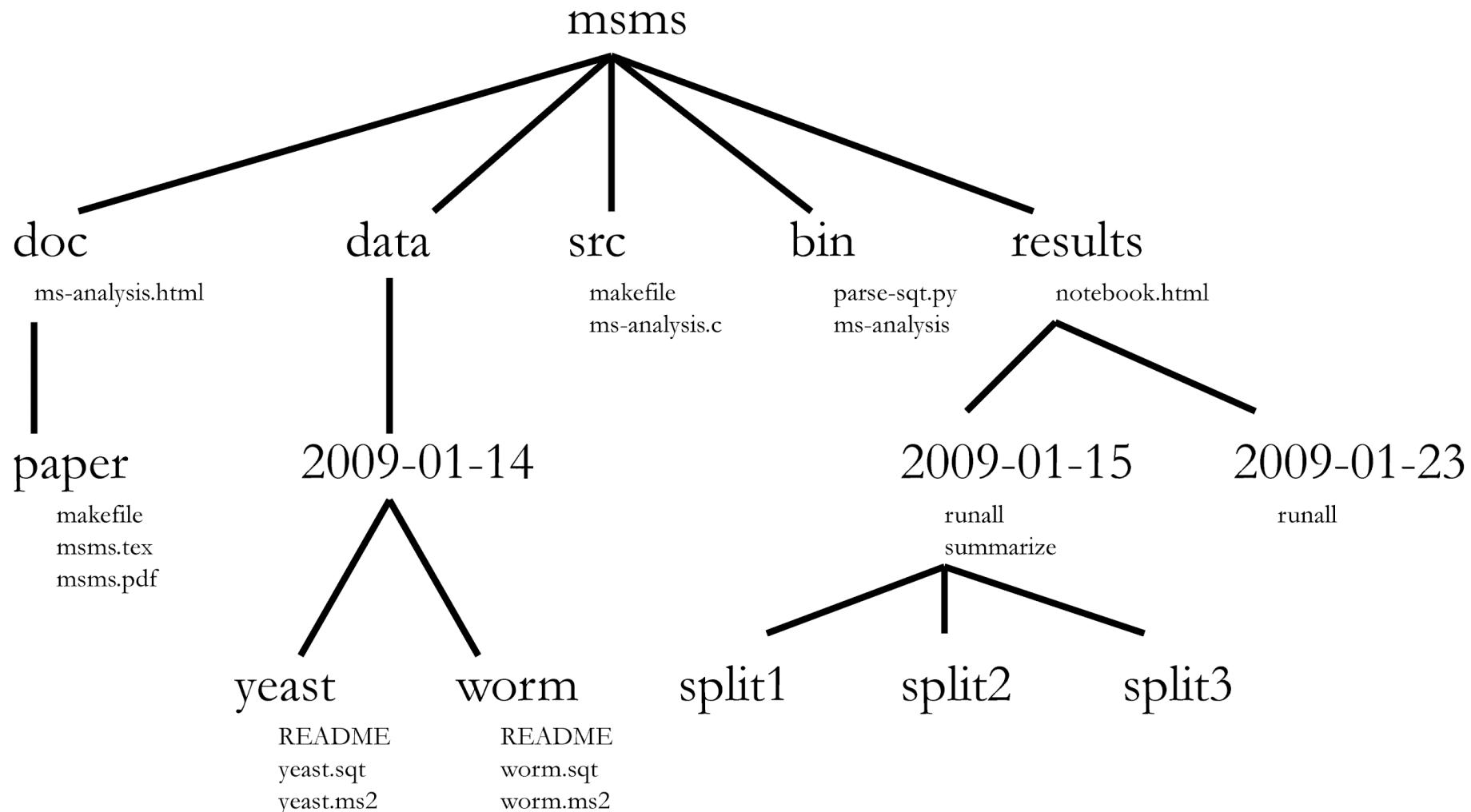
1. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
2. Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.
3. When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).
4. The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

R, Rstudio and Rnotebooks

https://github.com/raw937/FABI_workshop/blob/master/bin/ggplot_tutorial.Rmd

VIEW ON SCREEN (HTML)

Reproducible research



Noble, 2009

Link to article: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424>

Reproducible research

<http://joey711.github.io/phyloseq-demo/Restroom-Biogeography.html>