Welcome to the New Workshop 11

THIS WORKSHOP HAS BEEN SIGNIFICANTLY REWRITTEN

- This workshop has an increased focus on understanding the process from end-to-end
- The input datasets are new
  - The use of a standardized mock microbial community can give us a known input with an expected output
  - Also contains a stool sequencing run for a more complex community
- I am making this up as I go along (sorta)
  - We are scheduled for 3 hours of class time per day, but nobody really wants that
  - I aim to have everybody out between 2 and 2.5 hours
  - You are my first class, so slides distributed before class are draft versions. Slides that are posted after a session are the final product. I need to work out the timing of the days
  - I am still adding and cutting material for the next two days as I see the timing
- If nothing else, I have removed all content teaching you about 454 sequencing.

UCLA

QCBio

Metagenome Analysis Workshop

Michael Weinstein
UCLA Collaboratory

Topic

INTRODUCTION TO THE INSTRUCTOR

Who Am I?

RELEVANT TRAINING/EXPERIENCE

Bench scientist with an interest in molecular human genetics and host/pathogen interactions and genetic disease pathology who (partially) transitioned these interests to bioinformatics and computational biology

- Virus/virus interactions in the human immune landscape
- Genetics and molecular pathology of atherosclerosis
- Molecular physiology of lipoprotein metabolism in mammals
- Genetics and molecular pathology of skeletal dysplasia (dwarfisms)
- Cancer immunology and immune therapy
- CRISPR target design (and off-targeting behavior)
- Microbiome analysis and reproducibility
- Scientific computer program design and usability
- Cybersecurity

Who Am I?

AFFILIATIONS

I am affiliated with UCLA and Zymo Research

- I have an appointment of some kind in MCDB
  - Exact nature of this appointment is currently in transition, ask if interested
  - Primary role is teaching, the study of how to teach bioinformatics, and supporting the research of others with expertise in molecular and computational methods
- I am a scientist at Zymo Research on the Bioinformatics and Microbiome teams
  - This does create a conflict of interest in this course: Zymo Research sells products for the application that is the topic of this class
  - For this reason, I will try to avoid discussing specific product brands during lecture, with one exception:
    - We will be using sequencing data from the ZymoBIOMICS mock microbial community
  - I will discuss free open-source software provided by Zymo Research
    - No financial interest, we provide it for the good of the field.

Who Are You?

ASSUMPTIONS I MAKE ABOUT YOU (PREREQUISITES)

- You have some experience working on the command line
- You are able to log in to the Hoffman2 computing cluster
  - You know how to do this
  - Your account is currently in good standing and active
  - If you have not logged in recently, please do so now
  - If you are unable to log in, please put in a help request to the cluster admins
- You have some understanding of high-throughput sequencing fundamentals
  - You can understand a FASTQ file somewhat
  - You understand, roughly, how sequence is generated and stored

Who Are You?

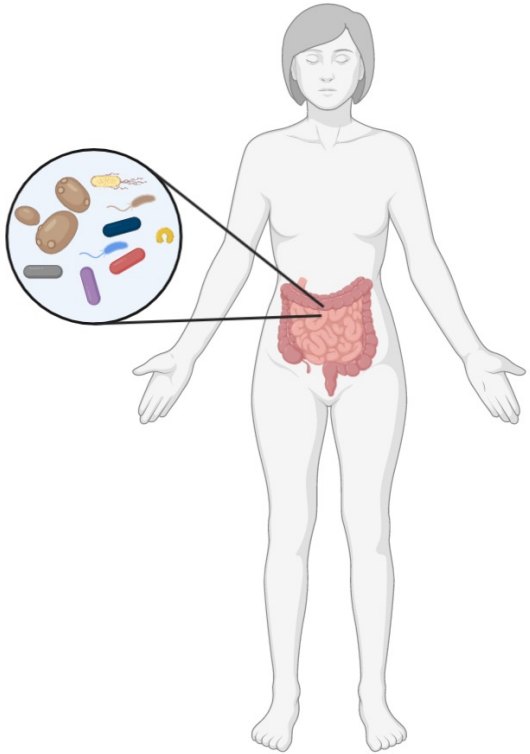
WE ARE ALL HERE TO HAVE A POSITIVE IMPACT IN THE BIOMEDICAL FIELD AND TO CONTRIBUTE TO THE GREATER GOOD OF HUMANITY (STARTING WITH OUR CLASSMATES)

- What is your name?
- Are you a UCLA affiliate?
  - What stage are you in (undergrad, grad, postdoc, etc.)?
  - What department and lab are you in?
  - What are you studying or planning to study?
- Are you from outside UCLA?
  - What is your affiliation?
  - What is your role?
  - What are you studying or planning to study?

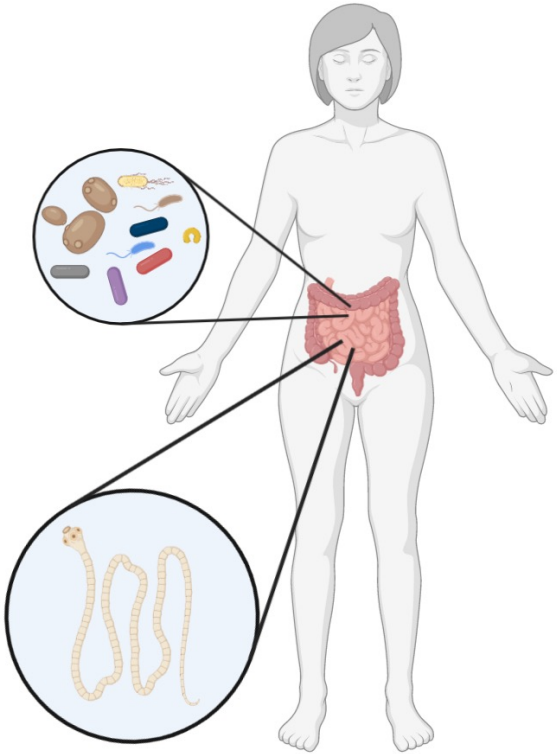
Topic

INTRODUCTION TO THE METAGENOME

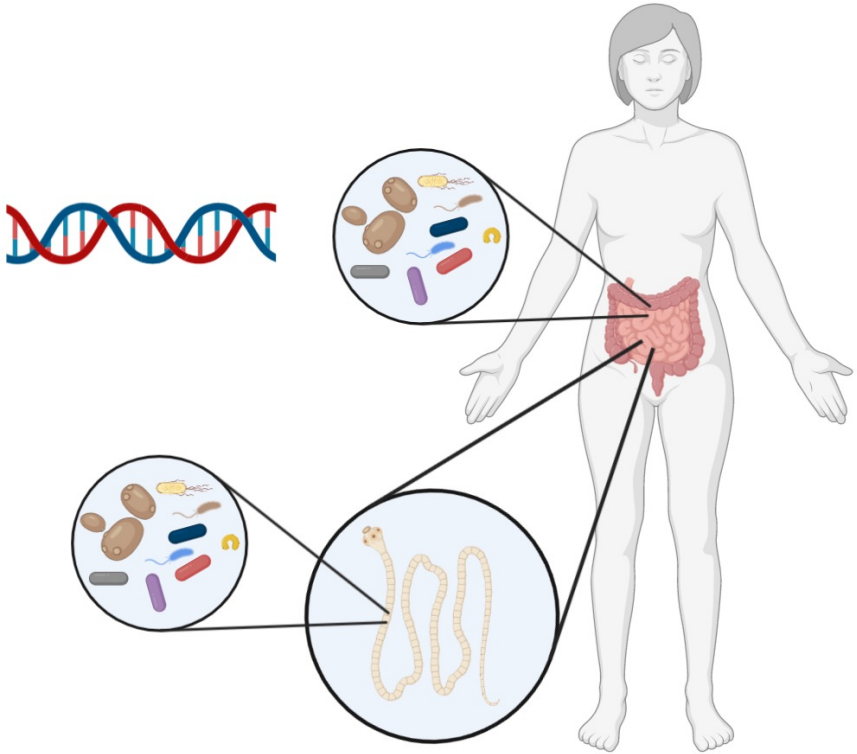
Metagenome *vs.* Microbiome



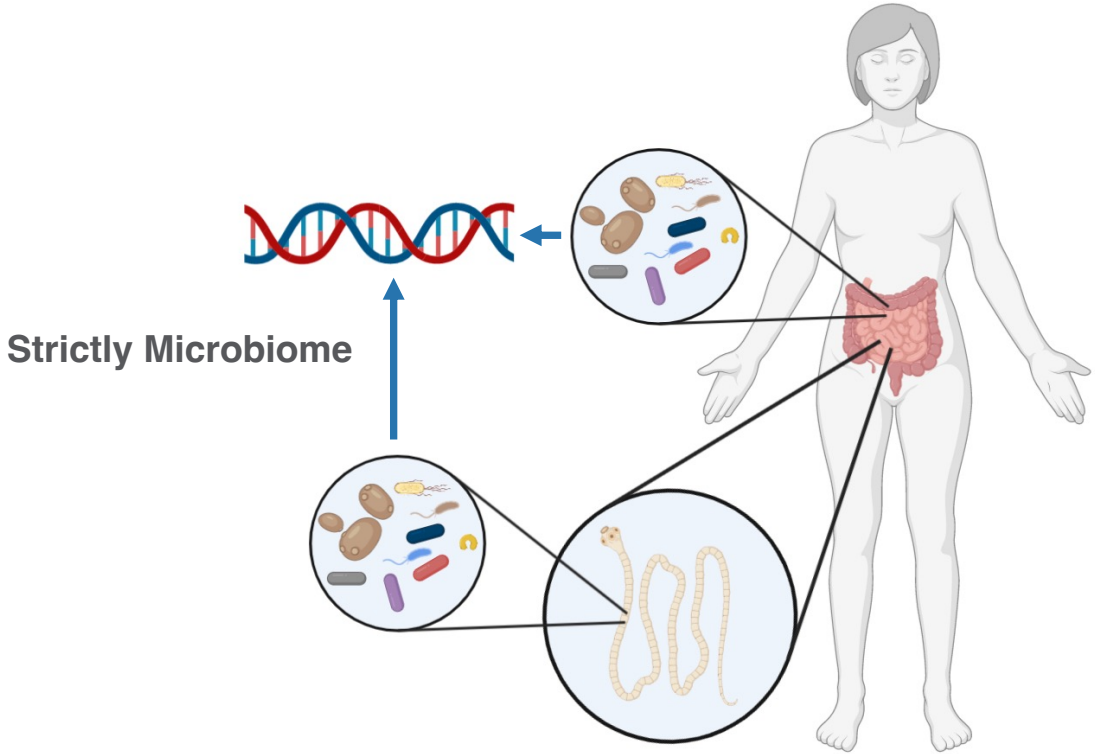
Metagenome *vs.* Microbiome



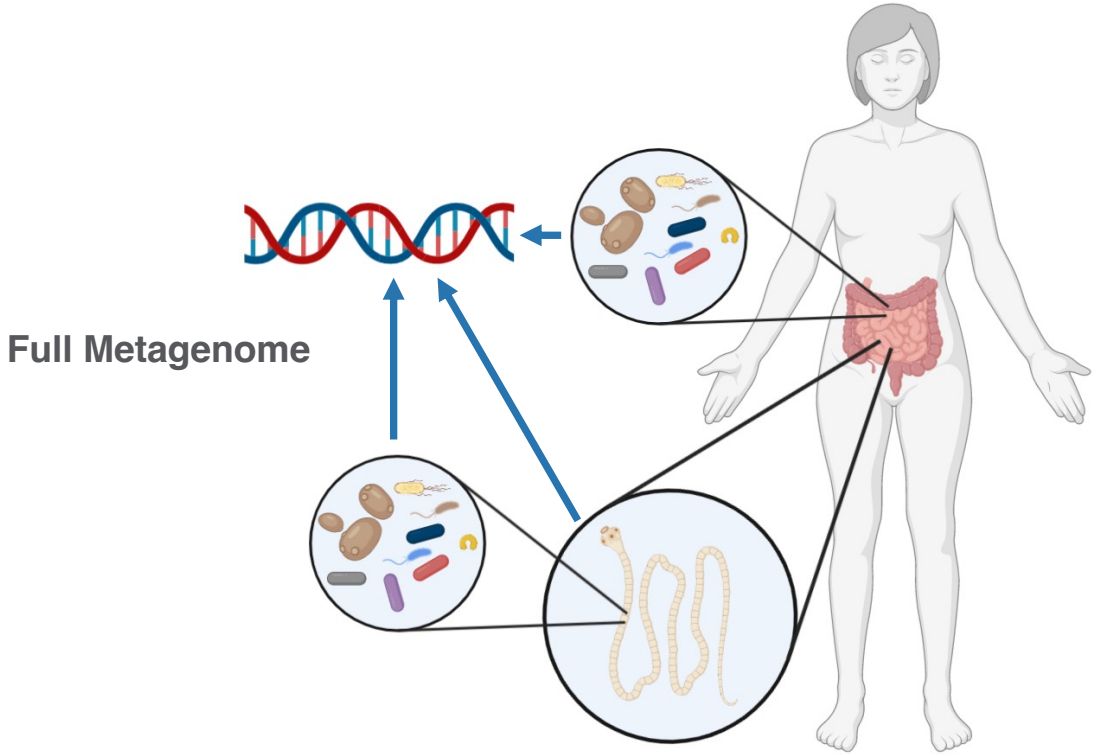
Metagenome *vs.* Microbiome



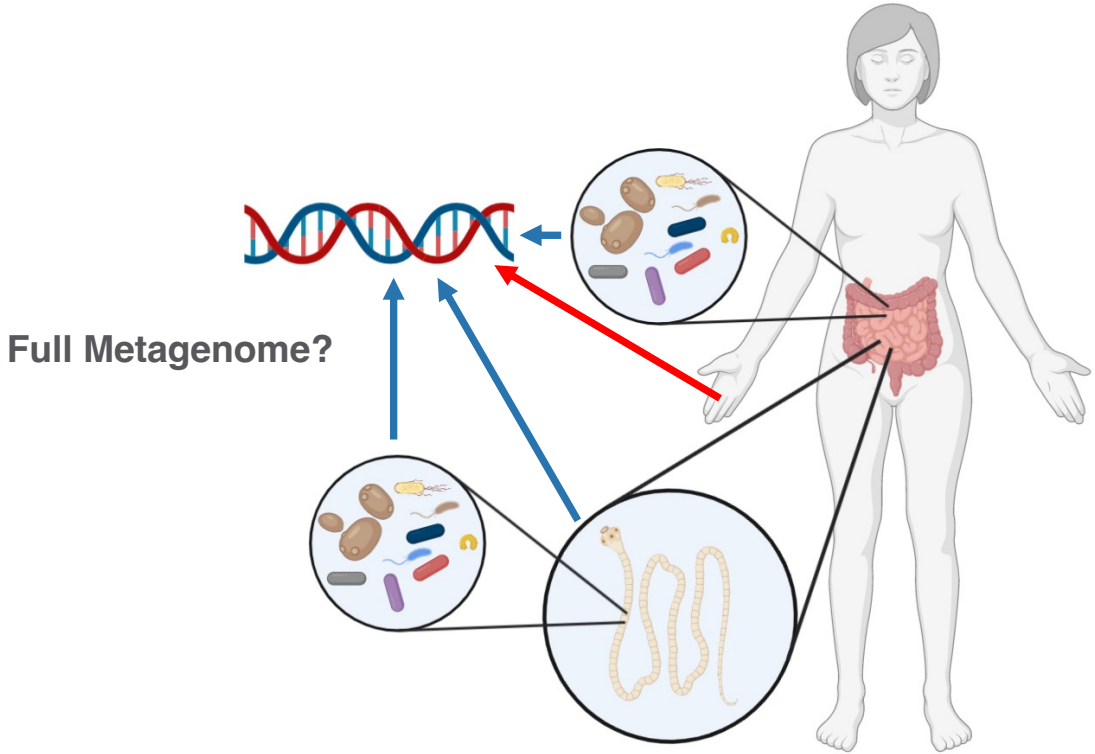
Metagenome *vs.* Microbiome



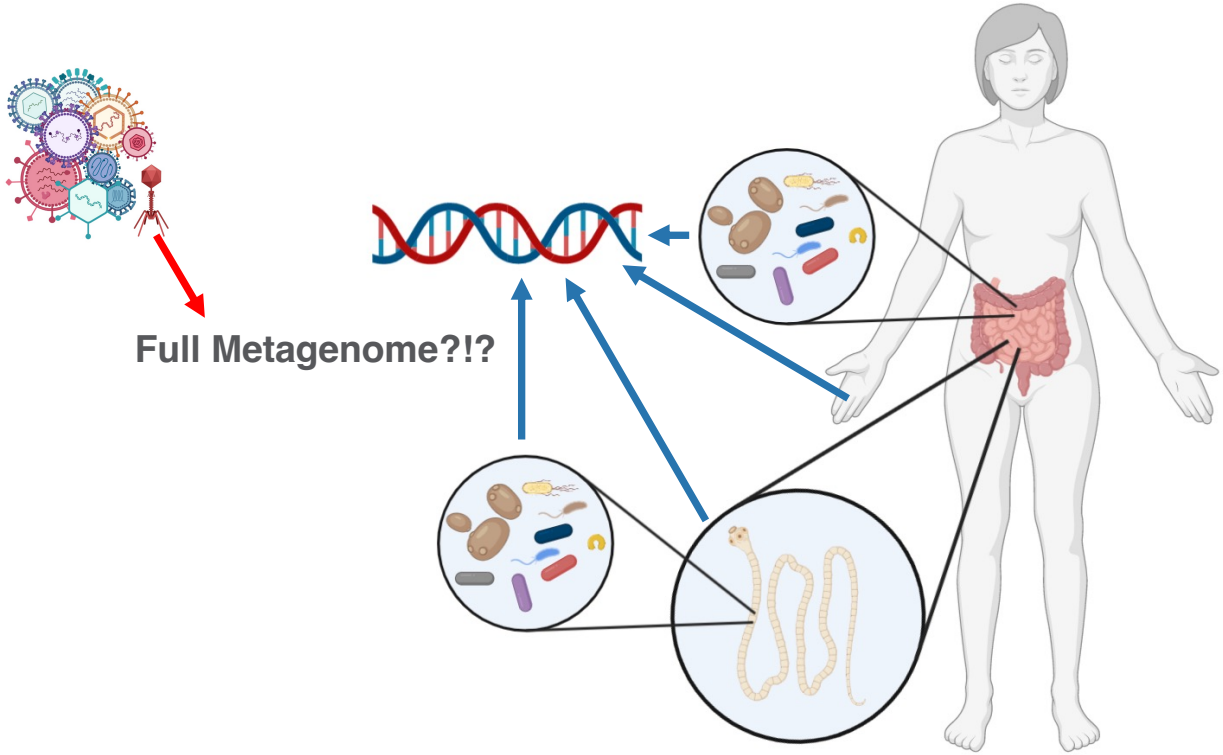
Metagenome *vs.* Microbiome



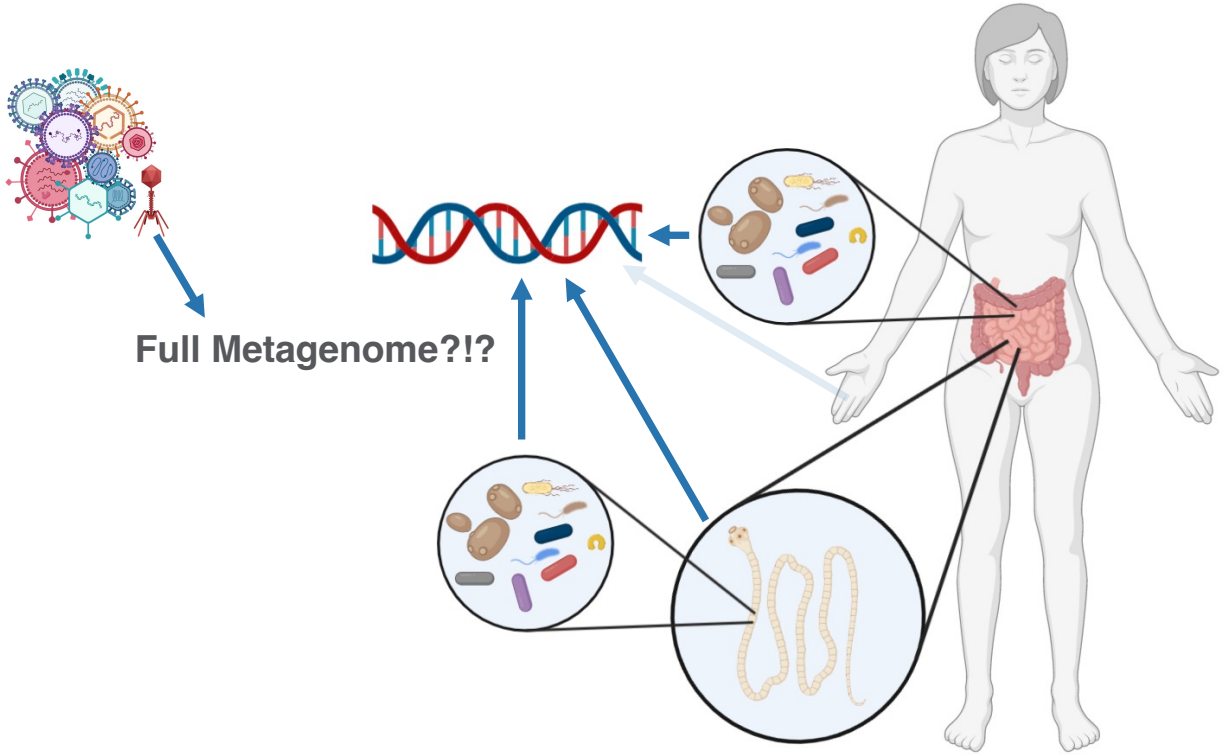
Metagenome *vs.* Microbiome



Metagenome *vs.* Microbiome

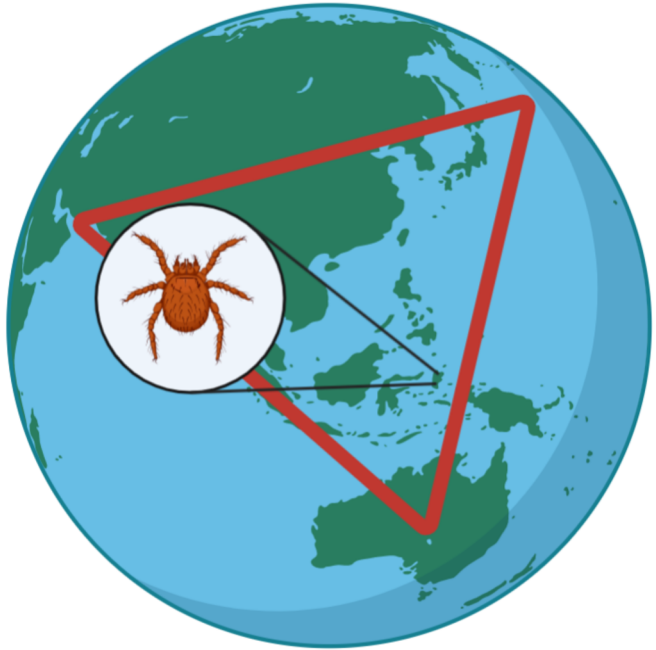


Metagenome *vs.* Microbiome

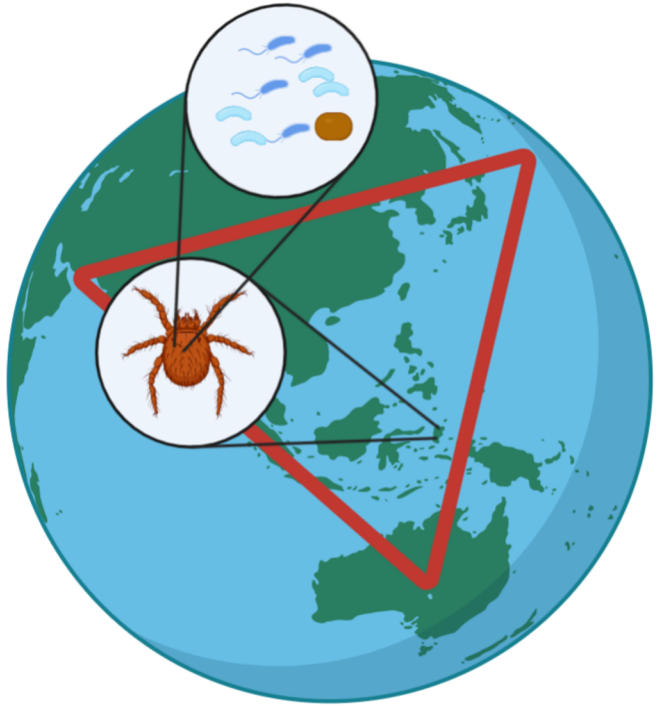


Everything Is Somewhere

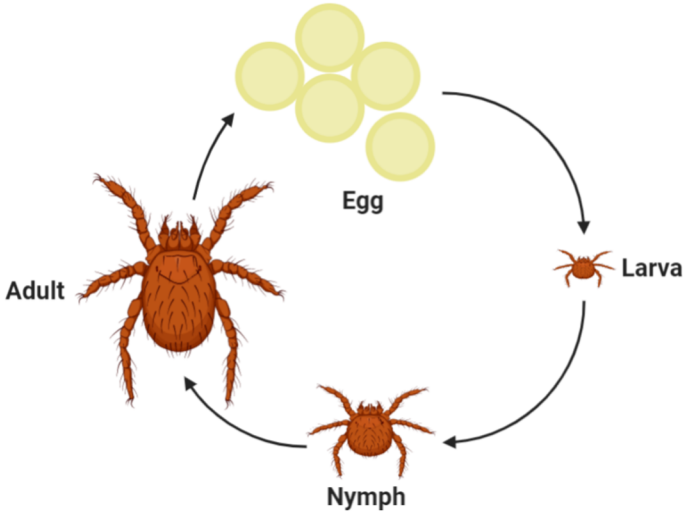
Everything Is Somewhere



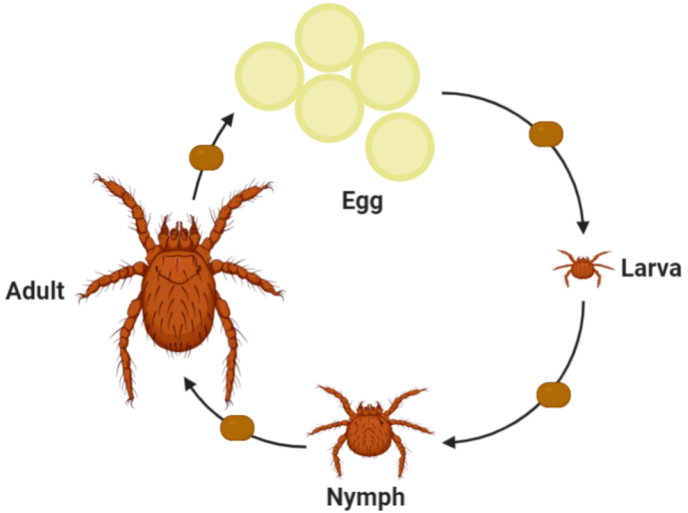
Everything Is Somewhere



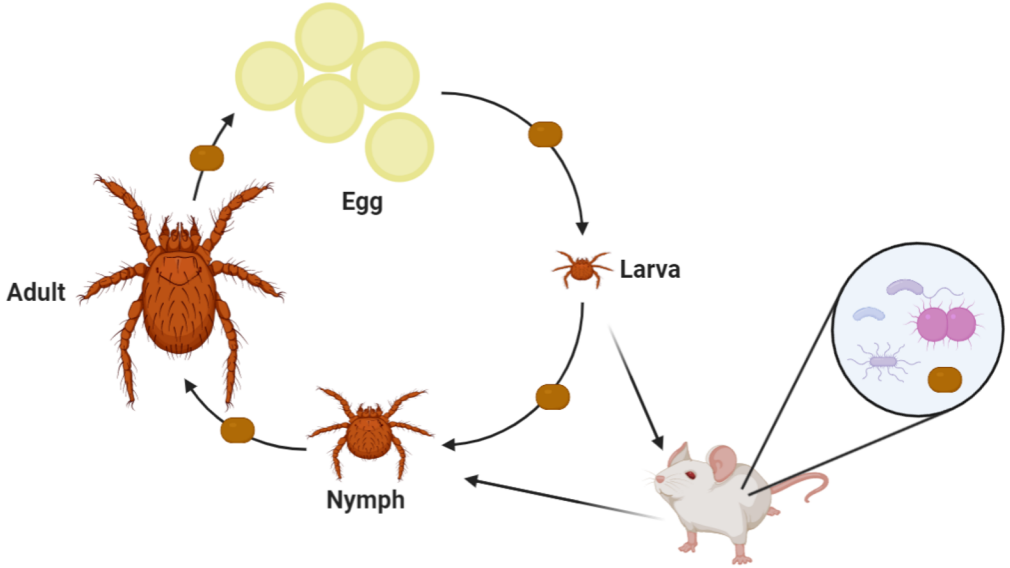
Microbiomes Move



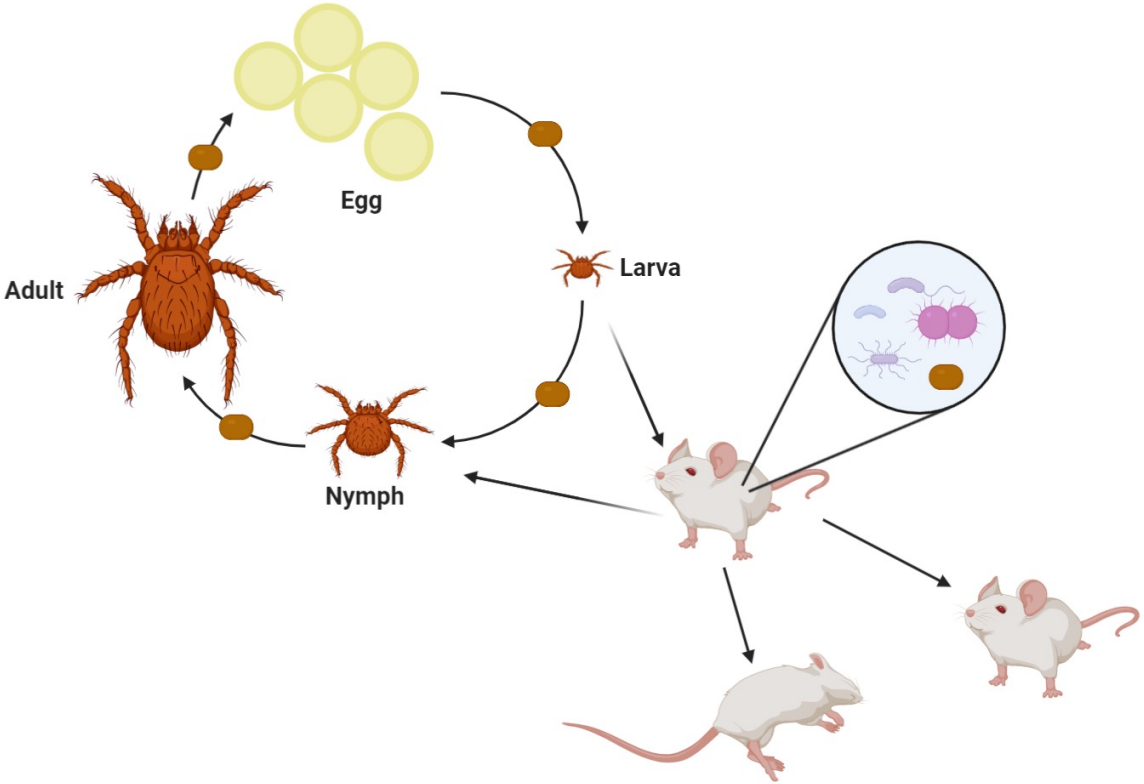
Microbiomes Move



Microbiomes Move



Microbiomes Move



Course Overview

Day 1

- Introduction
- Bioinformatics Review
  - Sequencing Review
  - Computing Review
- Course Goals
- Types of Analysis
  - Target *vs.* Shotgun
  - Blur *vs.* Sharpen
- Starting Materials

Day 2

- Day 1 Recap
- Sample Prep
  - Collection
  - Preservation/transport
  - Lysis
  - Library Prep
- Identifying Microbes
  - DADA2 via QIIME2
- Basic Sample Comparisons

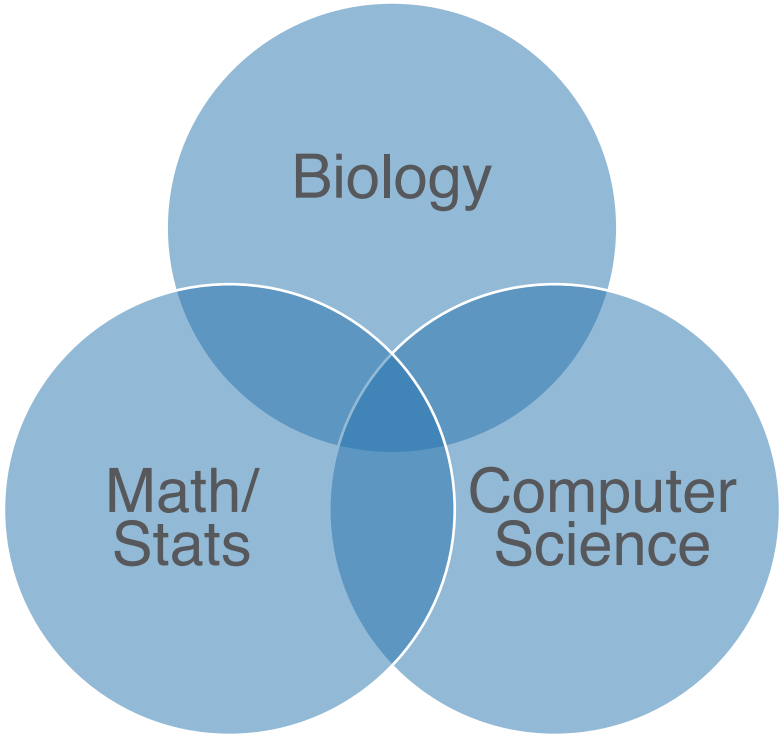
Day 3

- Day 2 Recap

Topic

BIOINFORMATICS REVIEW

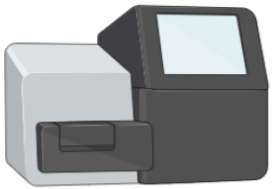
What Is Bioinformatics?



Topic

BIOINFORMATICS REVIEW: SEQUENCING TECHNOLOGY

Sequencers



Tech/Machine	Nanopore	MiSeq	HiSeq	NovaSeq	PacBio
Setup Cost	Painless	Ouch	You're probably a core facility (over half a million)		
Cost/Base	Not cheap	Not cheap	Cheap	Very Cheap	Not cheap
Run Time*	Up to 2 days	About 2.5 days			A workday
Read Length	5-7 figures	Up to 2x300	Up to 2x150	Up to 2x150	4-5 figures
Error Rate	High	Medium to Low	Low	Low	Very low**

* Most sequencers can read less length in exchange for a shorter run time

** PacBio can trade off error rate for read length; very low error assumes a short insert.

What's In a (File) Name?

- QSEQ
  - s_3_1_0065_qseq.txt.gz
  - s_3: Lane Number
  - 1: Read Direction
  - 0065: Tile Number
  - qseq.txt.gz: File Format

- FASTQ
  - sample_S1_L001_R1_001.fastq.gz
    - sample: Sample Name
    - S1: Sample Number
    - L001 : Lane Number
    - R1: Read Direction
      - If starting with "I": Index read direction
    - 001: File Number
      - Always 001 on modern systems

QSEQ Format

```
MM123   002   3   2208   33.30   98.40   0   2   ATGCAGGG   aababc`_   1
MM123   002   3   2208   33.40   98.10   0   2   GAATGTCA   ba_\[JI@   0
MM123   002   3   2208   63.30   98.20   0   2   CGTACCGC   hfeZWTSJ   1
MM123   002   3   2208   30.10   98.60   0   2   TCCCTAAG   KJKHDBA@   0
```

QSEQ Format

```
MM123   002   3   2208   33.30   98.40   0   2   ATGCAGGG   aababc`_   1
MM123   002   3   2208   33.40   98.10   0   2   GAATGTCA   ba_\[JI@   0
MM123   002   3   2208   63.30   98.20   0   2   CGTACCGC   hfeZWTSJ   1
MM123   002   3   2208   30.10   98.60   0   2   TCCCTAAG   KJKHDBA@   0

MM123   002   3   2208    :  Machine ID, Run Number, Lane, Tile
33.30   98.40   0   2     :  X-pos, Y-pos, Index, Direction
ATGCAGGG   aababc`_   1 :  Sequence, Quality String, Pass Filter
(Pass Illumina filtering values: 1 = Pass, 0 = Fail)
```

A Long Time Ago,
In a Sequencing Core Far Away…

It is a period of format war.
QSEQ programmers who like everything
on one line are fighting the FASTQ rebels
who prefer more lines with less data.

FASTQ won the decisive victory
by being easier on the human eyes
But QSEQ still holds a few hidden bases…

FASTQ Format

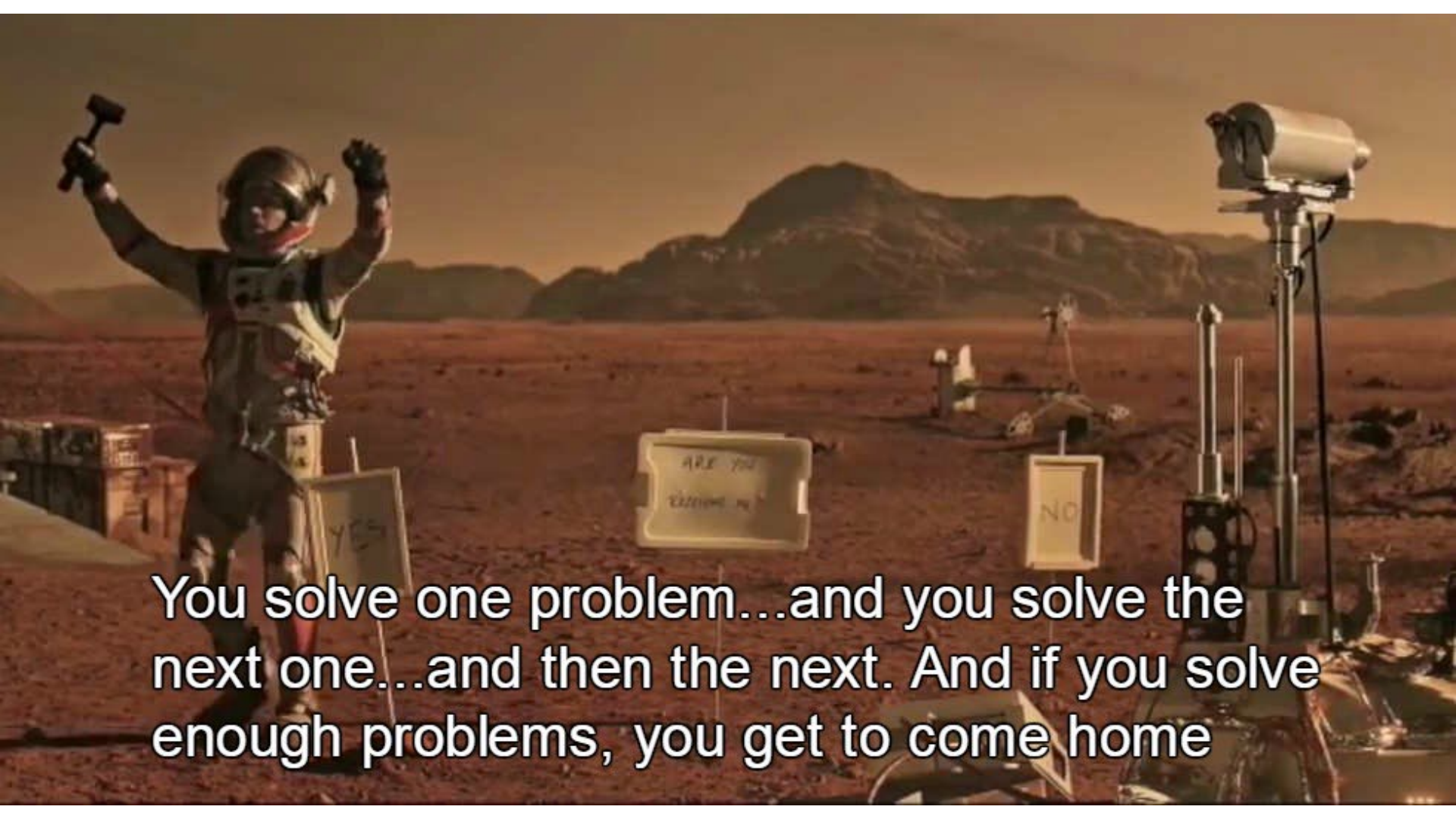
```
@MM123:002:FC123AB:3:2208:3330:9840 2:Y:18:ATCACG
AGGATACTAGCATAGATACCCTAGATAGTCATAGATCATGATAGGGAGATCTA
+
IJJJJJJIIIIIJIIIIIFFFEEEEEDDDDDDCABBBBB@@00))))*(*&%!
```

FASTQ Format

```
@MM123:002:FC123AB:3:2208:3330:9840 2:Y:18:ATCACG
AGGATACTAGCATAGATACCCTAGATAGTCATAGATCATGATAGGGAGATCTA
+
IJJJJJIIIIIJIIIIIFFFEEEEEDDDDDDCABBBBB@@00))))*(*&%!
Reads always start with "@"
MM123:002:FC123AB:3 - Machine, Run, Flowcell, Lane
2208:3330:9840      - Tile, X-pos, Y-pos
2:Y:18:ATCACG       - Direction, Filtered?,
                    - Control bits, Index/Sample
Separator between sequence and quality starts with "+"
```

Understanding Quality Strings

- Quality strings use ASCII values to encode a two-digit integer using a single character
  - Keeps files smaller
  - Keeps scores in alignment with bases
  - Character > Number > Base Value Adjustment > Phred > Confidence


You solve one problem…and you solve the next one…and then the next. And if you solve enough problems, you get to come home

Character > Number (ASCII)

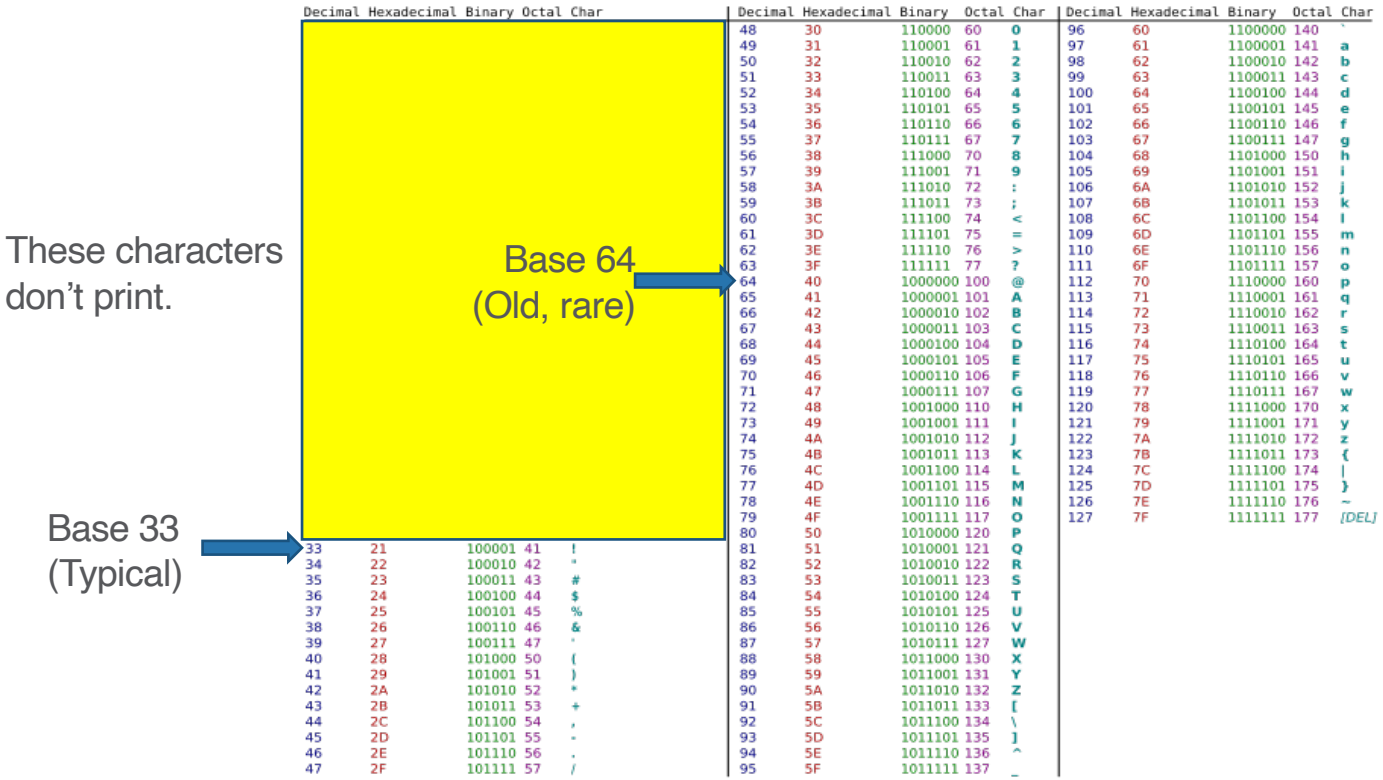
Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	*[NULL]*
1	1	1	1	*[START OF HEADING]*
2	2	10	2	*[START OF TEXT]*
3	3	11	3	*[END OF TEXT]*
4	4	100	4	*[END OF TRANSMISSION]*
5	5	101	5	*[ENQUIRY]*
6	6	110	6	*[ACKNOWLEDGE]*
7	7	111	7	*[BELL]*
8	8	1000	10	*[BACKSPACE]*
9	9	1001	11	*[HORIZONTAL TAB]*
10	A	1010	12	*[LINE FEED]*
11	B	1011	13	*[VERTICAL TAB]*
12	C	1100	14	*[FORM FEED]*
13	D	1101	15	*[CARRIAGE RETURN]*
14	E	1110	16	*[SHIFT OUT]*
15	F	1111	17	*[SHIFT IN]*
16	10	10000	20	*[DATA LINK ESCAPE]*
17	11	10001	21	*[DEVICE CONTROL 1]*
18	12	10010	22	*[DEVICE CONTROL 2]*
19	13	10011	23	*[DEVICE CONTROL 3]*
20	14	10100	24	*[DEVICE CONTROL 4]*
21	15	10101	25	*[NEGATIVE ACKNOWLEDGE]*
22	16	10110	26	*[SYNCHRONOUS IDLE]*
23	17	10111	27	*[ENG OF TRANS. BLOCK]*
24	18	11000	30	*[CANCEL]*
25	19	11001	31	*[END OF MEDIUM]*
26	1A	11010	32	*[SUBSTITUTE]*
27	1B	11011	33	*[ESCAPE]*
28	1C	11100	34	*[FILE SEPARATOR]*
29	1D	11101	35	*[GROUP SEPARATOR]*
30	1E	11110	36	*[RECORD SEPARATOR]*
31	1F	11111	37	*[UNIT SEPARATOR]*
32	20	100000	40	*[SPACE]*
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	**0**
49	31	110001	61	**1**
50	32	110010	62	**2**
51	33	110011	63	**3**
52	34	110100	64	**4**
53	35	110101	65	**5**
54	36	110110	66	**6**
55	37	110111	67	**7**
56	38	111000	70	**8**
57	39	111001	71	**9**
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	**A**
66	42	1000010	102	**B**
67	43	1000011	103	**C**
68	44	1000100	104	**D**
69	45	1000101	105	**E**
70	46	1000110	106	**F**
71	47	1000111	107	**G**
72	48	1001000	110	**H**
73	49	1001001	111	**I**
74	4A	1001010	112	**J**
75	4B	1001011	113	**K**
76	4C	1001100	114	**L**
77	4D	1001101	115	**M**
78	4E	1001110	116	**N**
79	4F	1001111	117	**O**
80	50	1010000	120	**P**
81	51	1010001	121	**Q**
82	52	1010010	122	**R**
83	53	1010011	123	**S**
84	54	1010100	124	**T**
85	55	1010101	125	**U**
86	56	1010110	126	**V**
87	57	1010111	127	**W**
88	58	1011000	130	**X**
89	59	1011001	131	**Y**
90	5A	1011010	132	**Z**
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	**a**
98	62	1100010	142	**b**
99	63	1100011	143	**c**
100	64	1100100	144	**d**
101	65	1100101	145	**e**
102	66	1100110	146	**f**
103	67	1100111	147	**g**
104	68	1101000	150	**h**
105	69	1101001	151	**i**
106	6A	1101010	152	**j**
107	6B	1101011	153	**k**
108	6C	1101100	154	**l**
109	6D	1101101	155	**m**
110	6E	1101110	156	**n**
111	6F	1101111	157	**o**
112	70	1110000	160	**p**
113	71	1110001	161	**q**
114	72	1110010	162	**r**
115	73	1110011	163	**s**
116	74	1110100	164	**t**
117	75	1110101	165	**u**
118	76	1110110	166	**v**
119	77	1110111	167	**w**
120	78	1111000	170	**x**
121	79	1111001	171	**y**
122	7A	1111010	172	**z**
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	*[DEL]*

Number > Base Value Adjustment

These characters don't print.

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]
1	1	1	1	[START OF HEADING]
2	2	10	2	[START OF TEXT]
3	3	11	3	[END OF TEXT]
4	4	100	4	[END OF TRANSMISSION]
5	5	101	5	[ENQUIRY]
6	6	110	6	[ACKNOWLEDGE]
7	7	111	7	[BELL]
8	8	1000	10	[BACKSPACE]
9	9	1001	11	[HORIZONTAL TAB]
10	A	1010	12	[LINE FEED]
11	B	1011	13	[VERTICAL TAB]
12	C	1100	14	[FORM FEED]
13	D	1101	15	[CARRIAGE RETURN]
14	E	1110	16	[SHIFT OUT]
15	F	1111	17	[SHIFT IN]
16	10	10000	20	[DATA LINK ESCAPE]
17	11	10001	21	[DEVICE CONTROL 1]
18	12	10010	22	[DEVICE CONTROL 2]
19	13	10011	23	[DEVICE CONTROL 3]
20	14	10100	24	[DEVICE CONTROL 4]
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]
22	16	10110	26	[SYNCHRONOUS IDLE]
23	17	10111	27	[ENG OF TRANS. BLOCK]
24	18	11000	30	[CANCEL]
25	19	11001	31	[END OF MEDIUM]
26	1A	11010	32	[SUBSTITUTE]
27	1B	11011	33	[ESCAPE]
28	1C	11100	34	[FILE SEPARATOR]
29	1D	11101	35	[GROUP SEPARATOR]
30	1E	11110	36	[RECORD SEPARATOR]
31	1F	11111	37	[UNIT SEPARATOR]
32	20	100000	40	[SPACE]
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Number > Base Value Adjustment



Decimal	Hexadecimal	Binary	Octal	Char
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Number > Base Value Adjustment > Phred

Decimal	Hexadecimal	Binary	Octal	Char
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Number > Base Value Adjustment > Phred

Decimal	Hexadecimal	Binary	Octal	Char
73	49	1001001	111	I

$$\begin{array}{rcrcl} \text{Value} & - & \text{base} & = & \text{Phred} \\ 73 & - & 33 & = & \underline{\text{Phred}} \\ & & & & 40 \end{array}$$

Phred > Confidence

$$\text{Value} - \text{base} = \text{Phred}$$

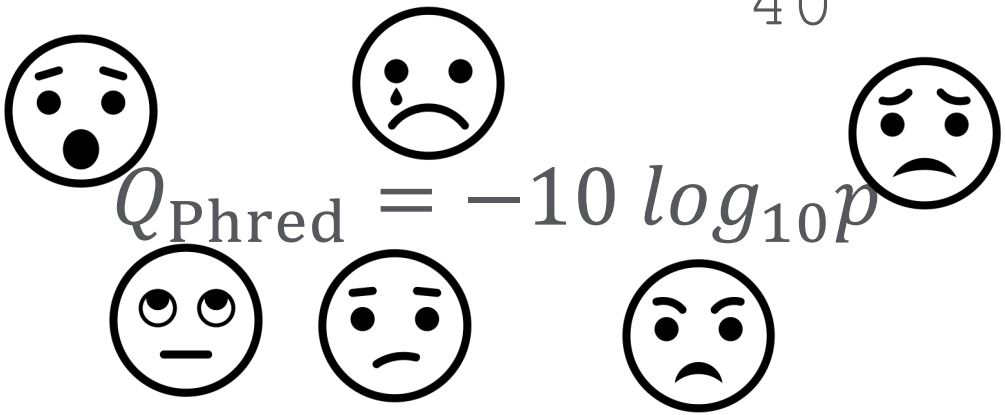
$$73 - 33 = \frac{\text{Phred}}{40}$$

$$Q_{\text{Phred}} = -10\, log_{10} p$$

Phred > Confidence

$$\text{Value} - \text{base} = \text{Phred}$$

$$73 - 33 = \frac{\text{Phred}}{40}$$



$$Q_{\text{Phred}} = -10\, log_{10} p$$

Phred > Confidence

$$\text{Value} - \text{base} = \text{Phred}$$

$$73 - 33 = \frac{\text{Phred}}{40}$$

$$Q_{\text{Phred}} = -10\, log_{10} p$$

	Phred	Error Probability	Confidence
Base Call = N	0	1 / 1	0%
	10	1 / 10	90%
	20	1 / 100	99%
	30	1 / 1000	99.9%
Illumina Base Max	40	1 / 10000	99.99%
	50	1 / 100000	99.999%
Align Max	60	1 / 1000000	99.9999%

Phred > Confidence

$$\frac{\text{Value} - \text{base}}{40} = \text{Phred}$$

$$\frac{73 - 33}{40} = \text{Phred}$$

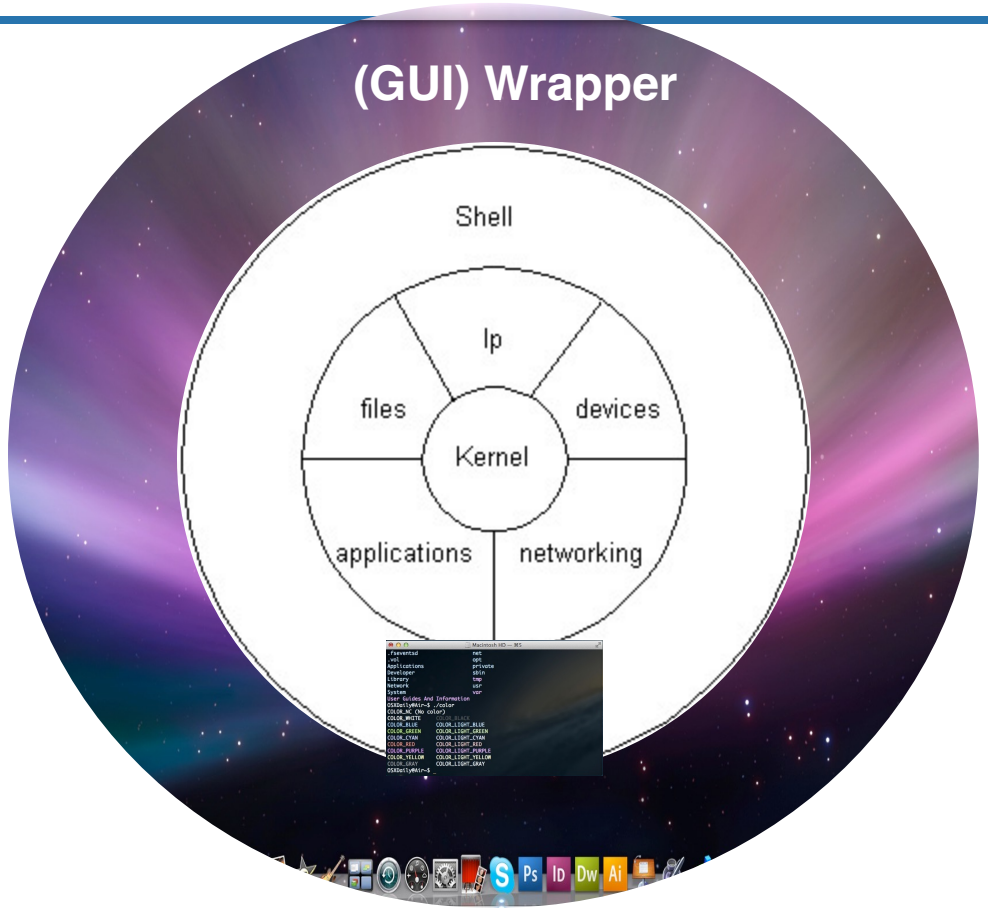
$$Q_{\text{Phred}} = -10\, log_{10} p$$

Phred	Error Probability	Confidence
0	1 / 1	0%
10	1 / 10	90%
20	1 / 100	99%
30	1 / 1000	99.9%
40	1 / 10000	99.99%
50	1 / 100000	99.999%
60	1 / 1000000	99.9999%

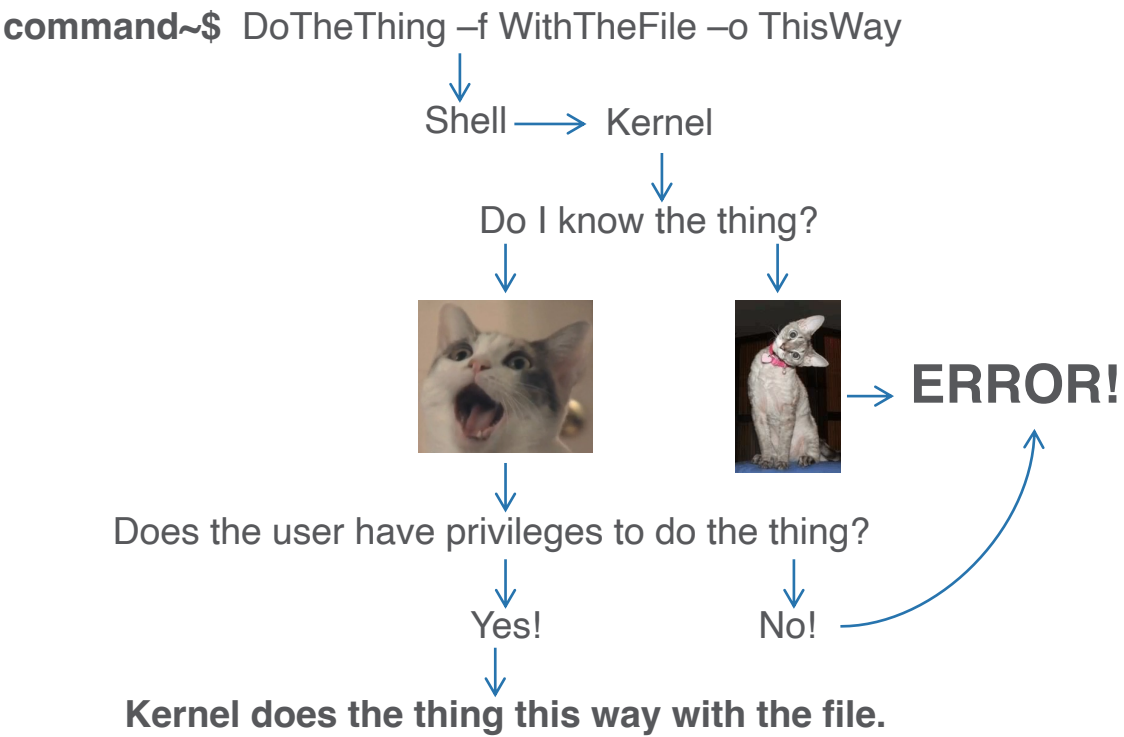
Topic

BIOINFORMATICS REVIEW: COMPUTING

Unix in 60 Seconds



Unix in 60 Seconds



Basic Unix Commands

Where am I?	`pwd`
Change directory	`cd` `cd ~/data`
Move up one level	`cd ..`
Look at a file	`less -S fileName`
Copy a file	`cp /data/file /otherDir/file`
Delete a file	`rm file`
Delete a directory	`rmdir ~/dirName/`
Secure copy	`scp user1@host1:file user@host2:file`
Compress a file	`gzip -c file > file.gz`
Uncompress file	`gzip -d -c file.gz > file`
Uncompress a file	`gunzip file.gz`
Make a new folder	`mkdir folderName`
Current directory	`./`
Home directory	`~/`
List all files in folder	`ls *`
Count lines in a file	`wc fileX`

Intro to Hoffman2 & Unix Command Line

What software is available?

Software is located at: /u/local/apps/

To list all software

```
$ ls /u/local/apps/
```

Software

But I want to use software that isn't installed.

- Put the executable in your home directory.

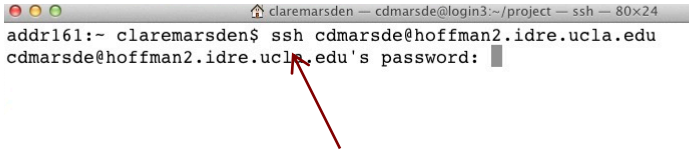
- If you need admin rights to install - email user support, and request they install it for you.

Intro to Hoffman2

Login to Hoffman2

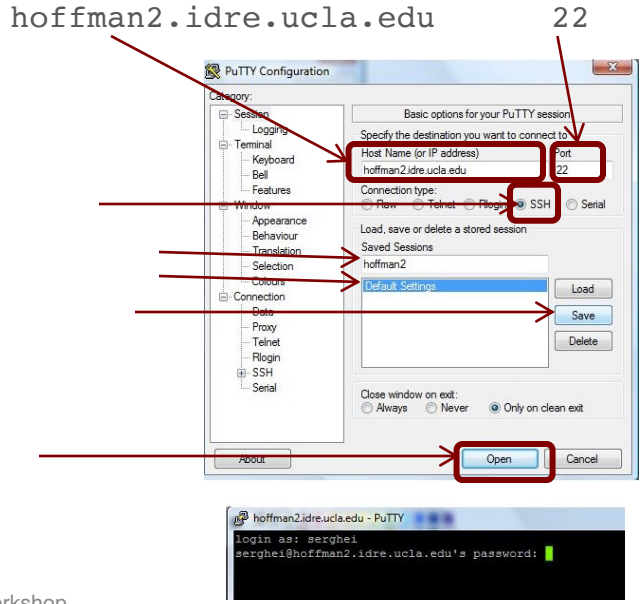
Apple/Mac

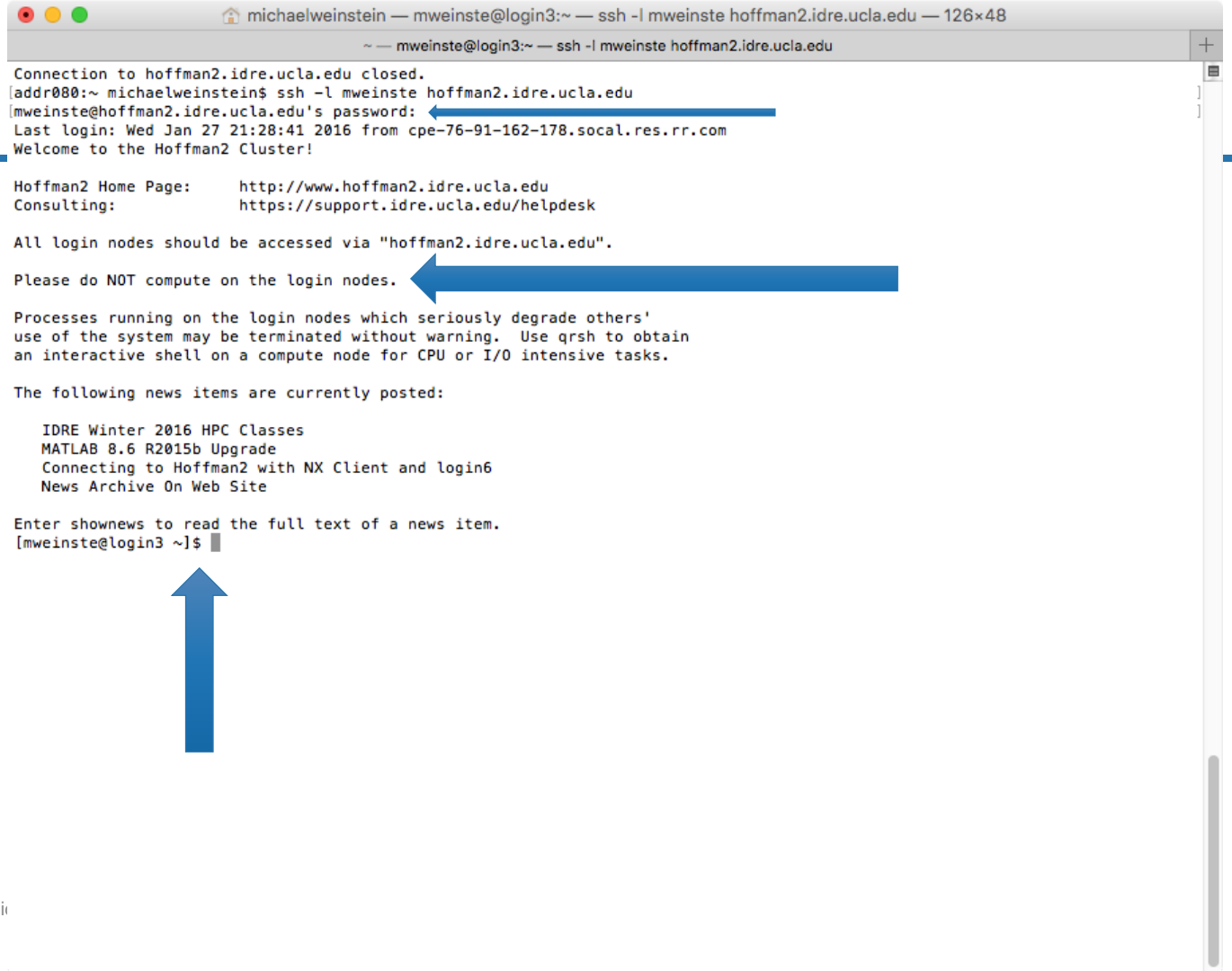
Open a terminal

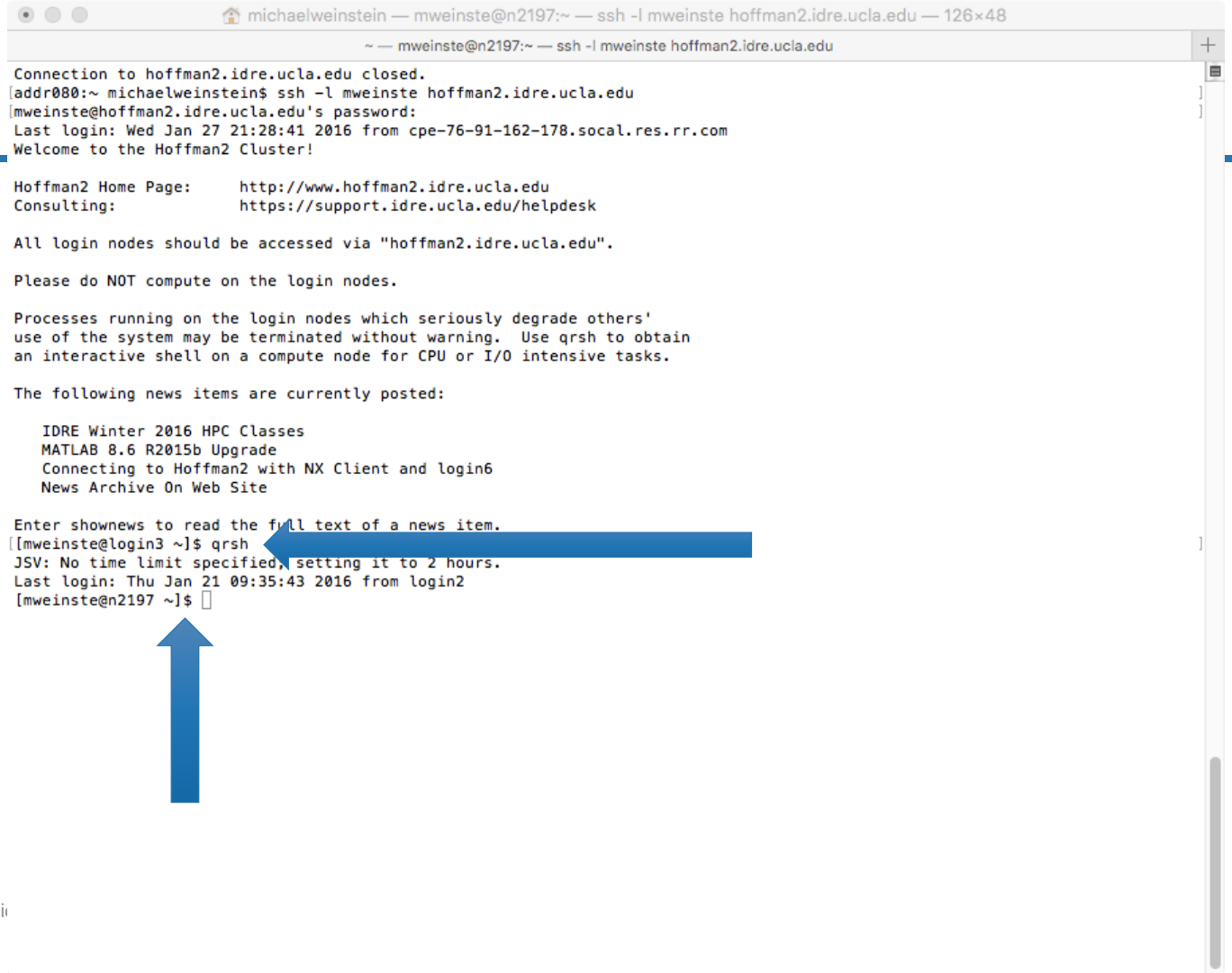


ssh username@hoffman2.idre.ucla.edu

Windows




michaelweinstein — mweinste@login3:~ — ssh -l mweinste hoffman2.idre.ucla.edu — 126×48
~ — mweinste@login3:~ — ssh -l mweinste hoffman2.idre.ucla.edu
Connection to hoffman2.idre.ucla.edu closed.
addr080:~ michaelweinstein$ ssh -l mweinste hoffman2.idre.ucla.edu
mweinste@hoffman2.idre.ucla.edu's password:
Last login: Wed Jan 27 21:28:41 2016 from cpe-76-91-162-178.socal.res.rr.com
Welcome to the Hoffman2 Cluster!
Hoffman2 Home Page: http://www.hoffman2.idre.ucla.edu
Consulting: https://support.idre.ucla.edu/helpdesk
All login nodes should be accessed via "hoffman2.idre.ucla.edu".
Please do NOT compute on the login nodes.
Processes running on the login nodes which seriously degrade others'
use of the system may be terminated without warning. Use qrsh to obtain
an interactive shell on a compute node for CPU or I/O intensive tasks.
The following news items are currently posted:
IDRE Winter 2016 HPC Classes
MATLAB 8.6 R2015b Upgrade
Connecting to Hoffman2 with NX Client and login6
News Archive On Web Site
Enter shownews to read the full text of a news item.
[mweinste@login3 ~]$


michaelweinstein — mweinste@n2197:~ — ssh -l mweinste hoffman2.idre.ucla.edu — 126×48
~ — mweinste@n2197:~ — ssh -l mweinste hoffman2.idre.ucla.edu
Connection to hoffman2.idre.ucla.edu closed.
addr080:~ michaelweinstein$ ssh -l mweinste hoffman2.idre.ucla.edu
mweinste@hoffman2.idre.ucla.edu's password:
Last login: Wed Jan 27 21:28:41 2016 from cpe-76-91-162-178.socal.res.rr.com
Welcome to the Hoffman2 Cluster!
Hoffman2 Home Page: http://www.hoffman2.idre.ucla.edu
Consulting: https://support.idre.ucla.edu/helpdesk
All login nodes should be accessed via "hoffman2.idre.ucla.edu".
Please do NOT compute on the login nodes.
Processes running on the login nodes which seriously degrade others'
use of the system may be terminated without warning. Use qrsh to obtain
an interactive shell on a compute node for CPU or I/O intensive tasks.
The following news items are currently posted:
IDRE Winter 2016 HPC Classes
MATLAB 8.6 R2015b Upgrade
Connecting to Hoffman2 with NX Client and login6
News Archive On Web Site
Enter shownews to read the full text of a news item.
[mweinste@login3 ~]$ qrsh
JSV: No time limit specified, setting it to 2 hours.
Last login: Thu Jan 21 09:35:43 2016 from login2
[mweinste@n2197 ~]$

Intro to Hoffman2 & Unix Command Line

To run software/jobs/programmes on hoffman2:

1 – Start an interactive node and run programs from the command line.

2 – Submit a script to run the program to the job queue

Which ever way you use, you must specify the resources you need i.e. (time/memory/cores).
If you exceed this – your job will be killed!

Other qrsh options on Hoffman2

This requests an interactive node with default settings (2 hours and 1G memory)

```
$ qrsh
```

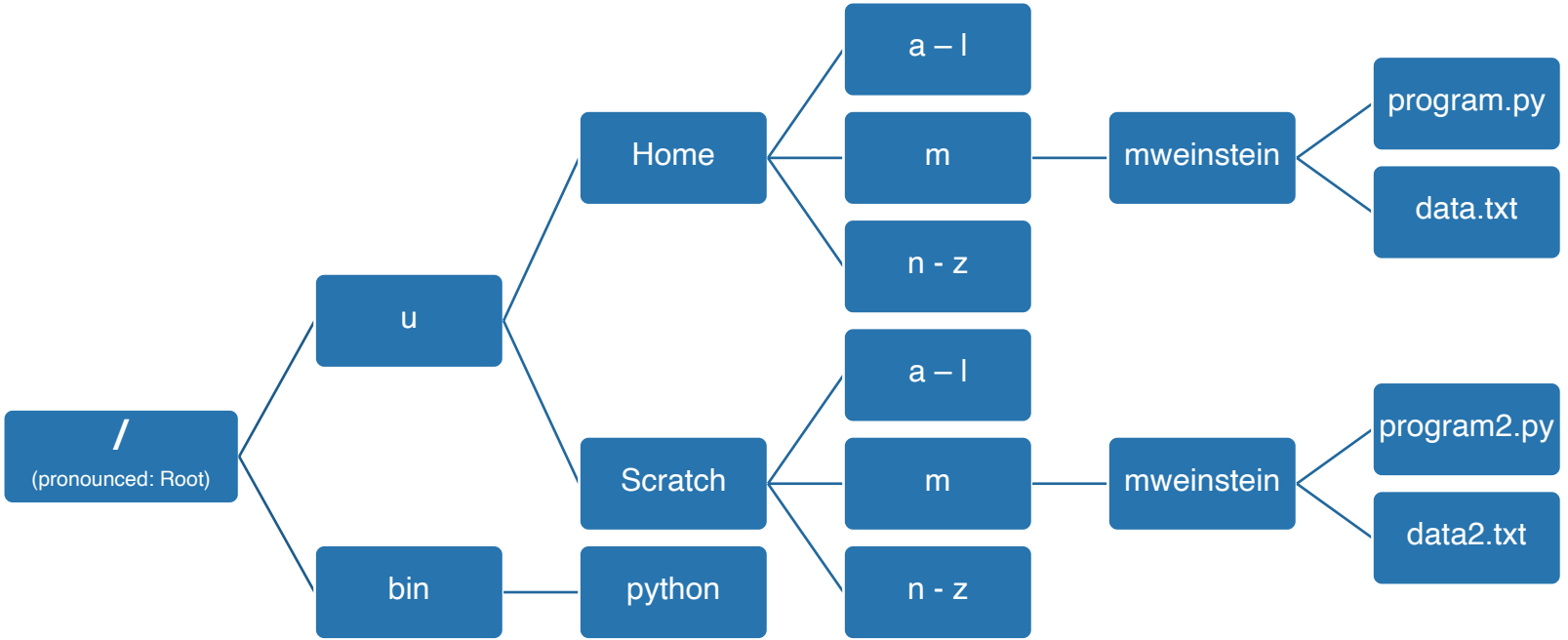
This requests 3G data. As time is not specified, the default of 2hrs is given.

```
$ qrsh –l h_data=3G
```

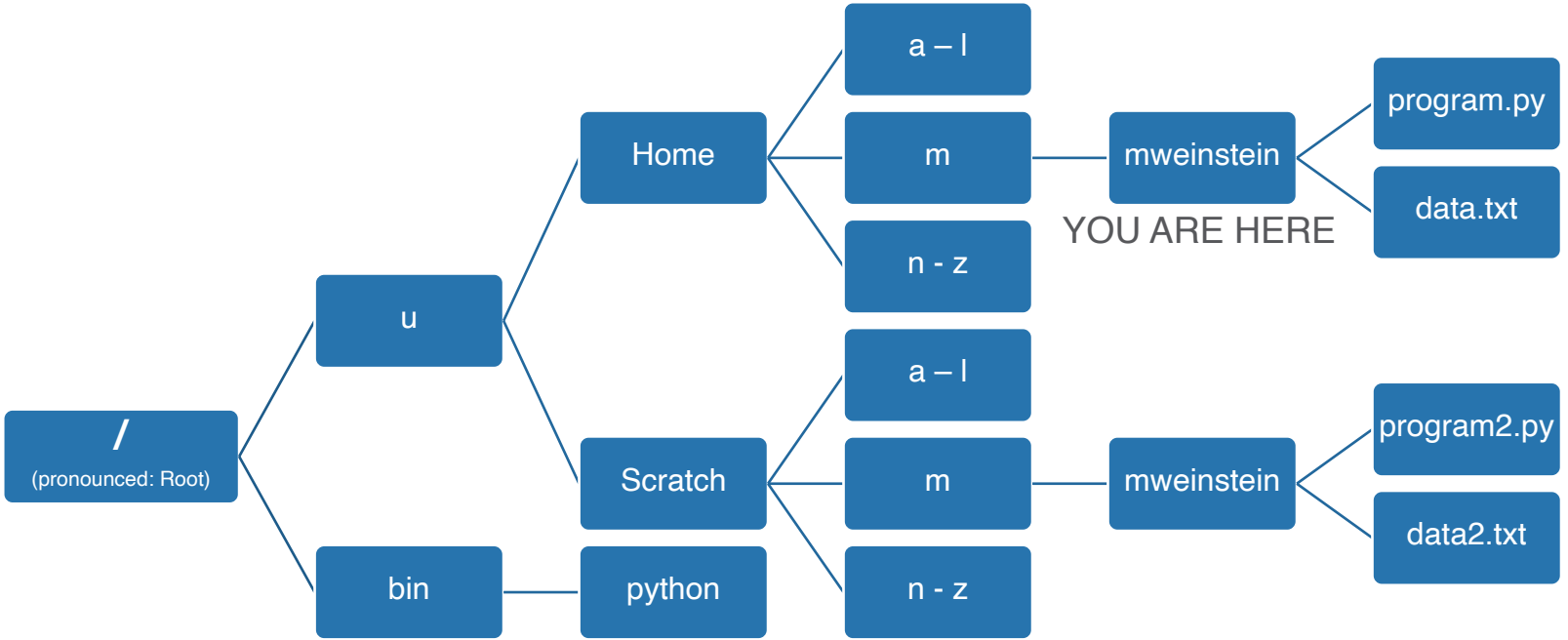
If you have a sponsor with nodes on hoffman2, you can request to use one of these as your interactive node. This will generally be quicker.

```
$ qrsh –l highp –l h_data=4G
```

The Linux File System

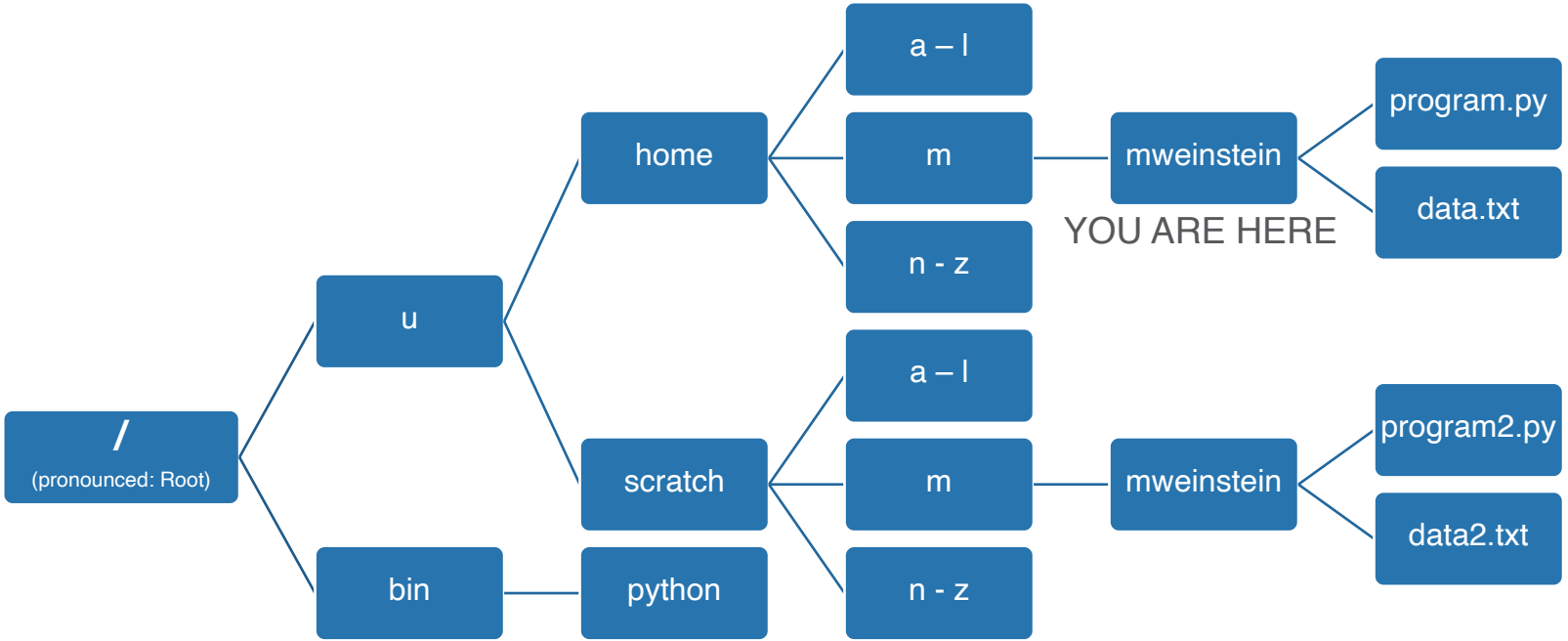


The Linux File System



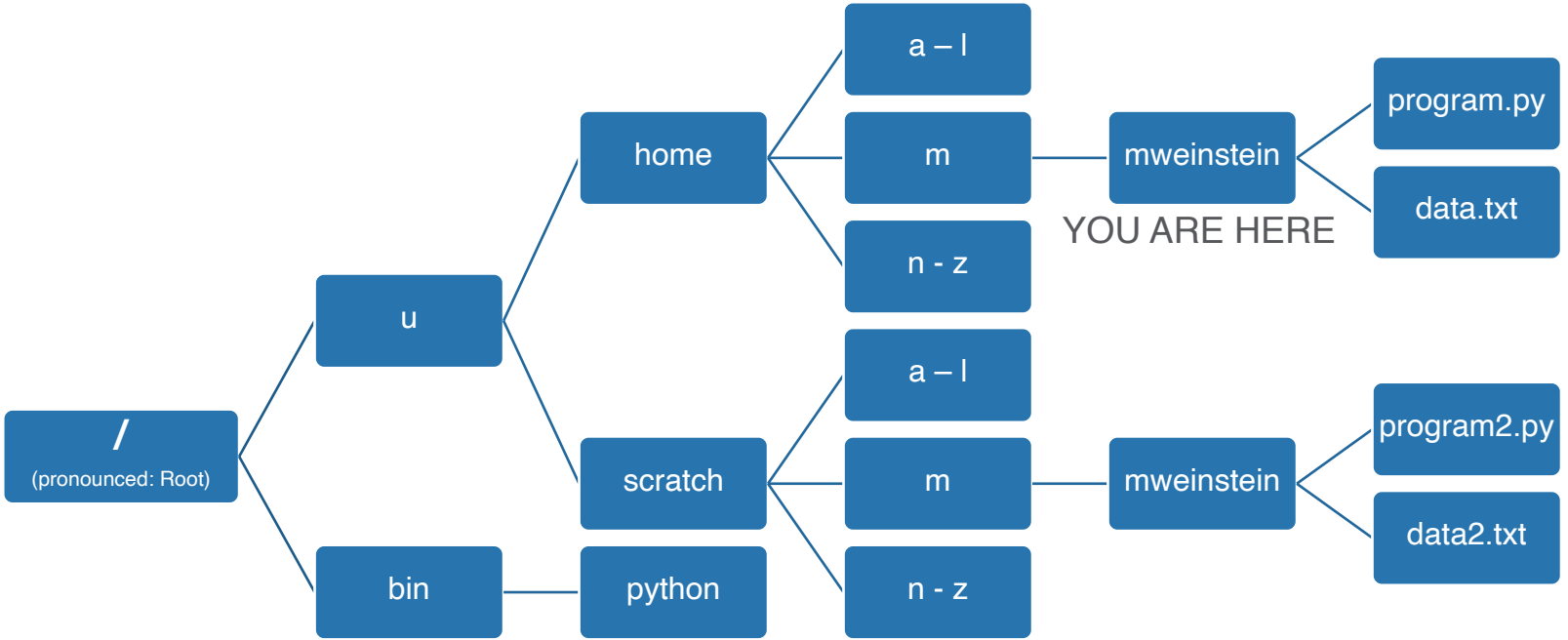

```
[mweinstein@computer ~]$
```

The Linux File System



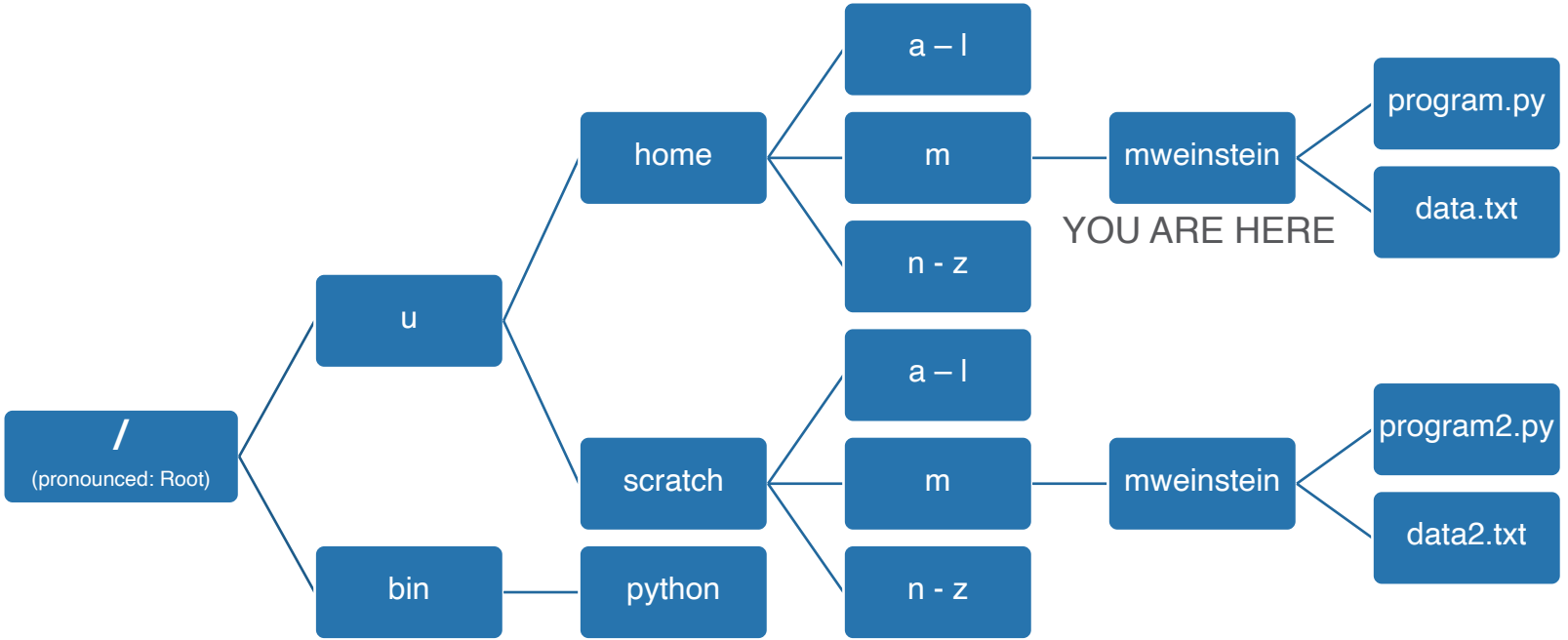

```
[mweinstein@computer ~]$ pwd
```

The Linux File System



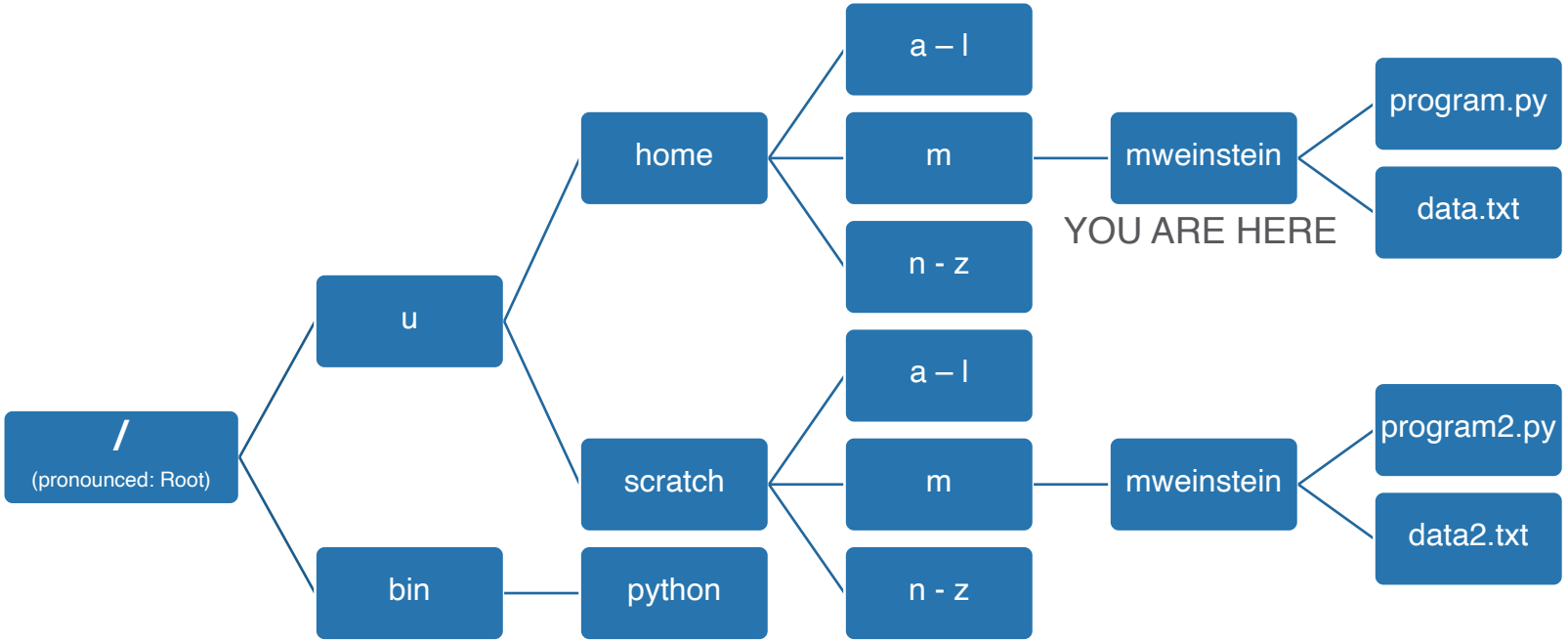

```
[mweinstein@computer ~]$ pwd
/u/home/scratch/m/mweinstein
```

The Linux File System



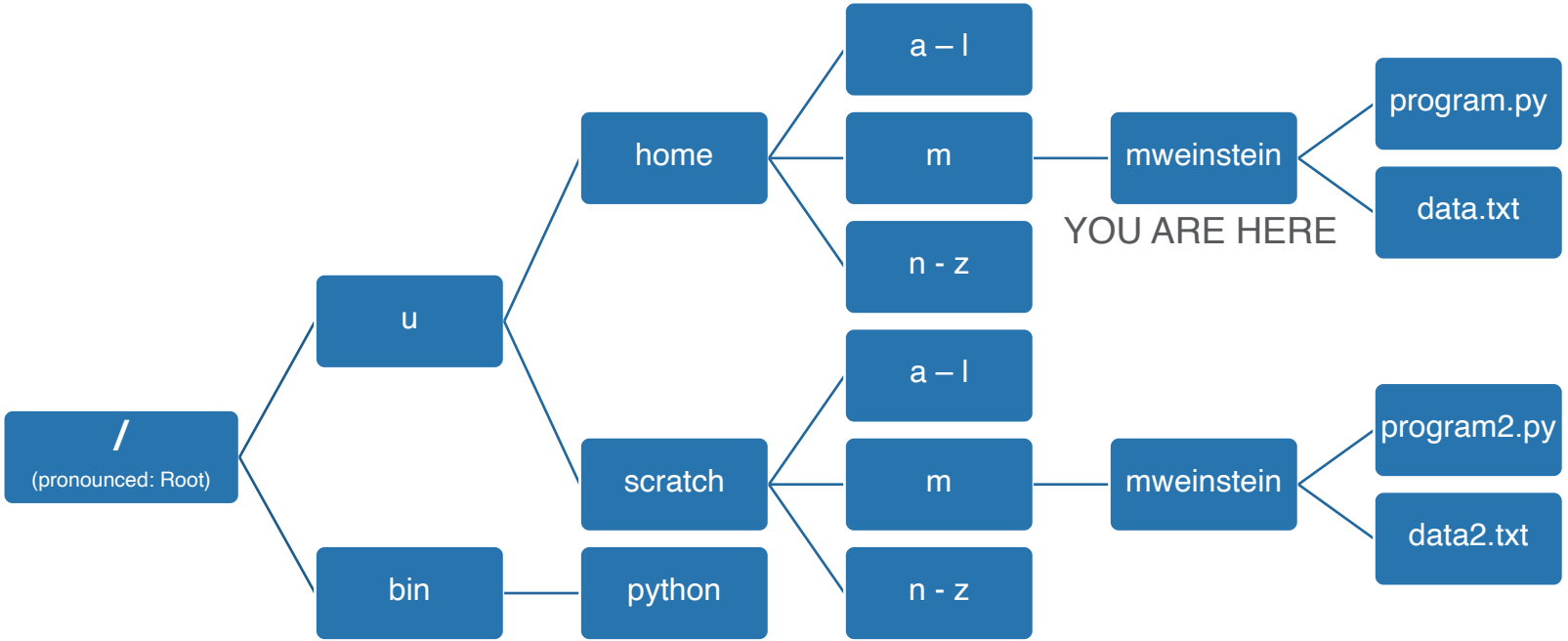

```
[mweinstein@computer ~]$ ls
```

The Linux File System



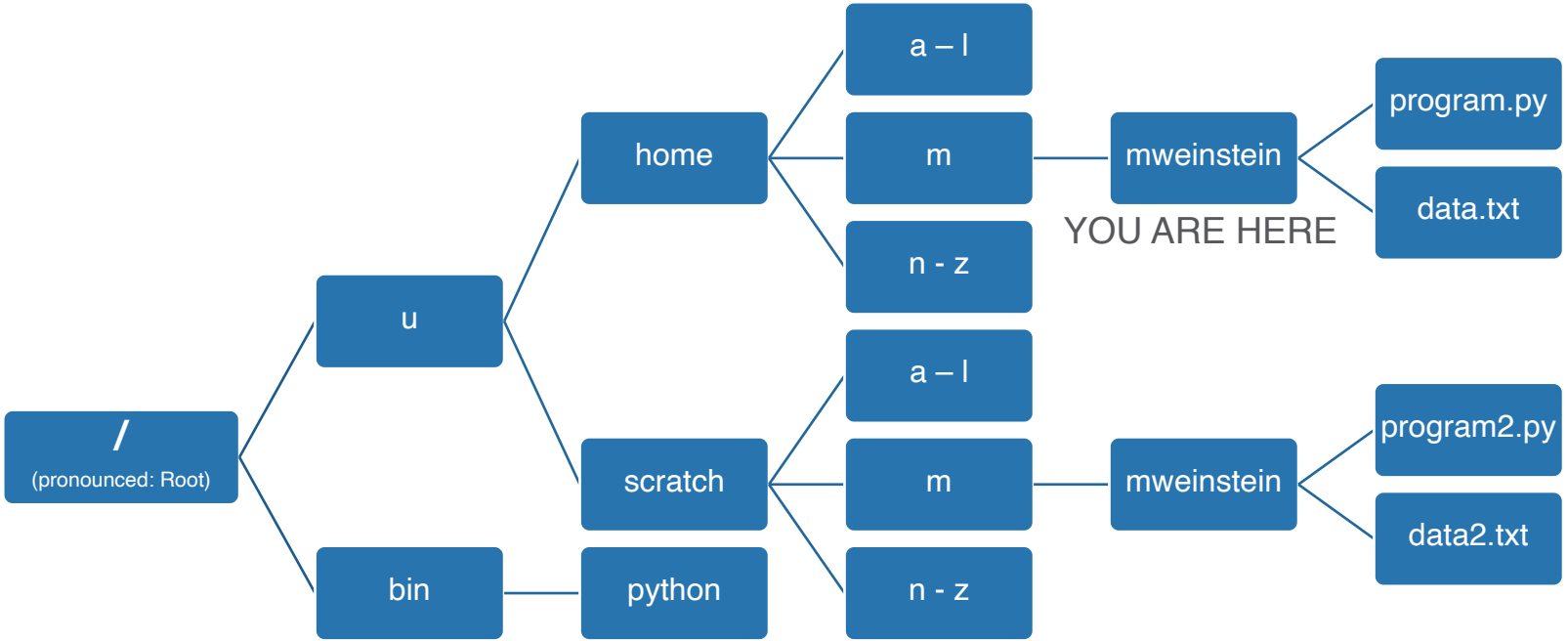

```
[mweinstein@computer ~]$ ls
program.py    data.txt
```

The Linux File System



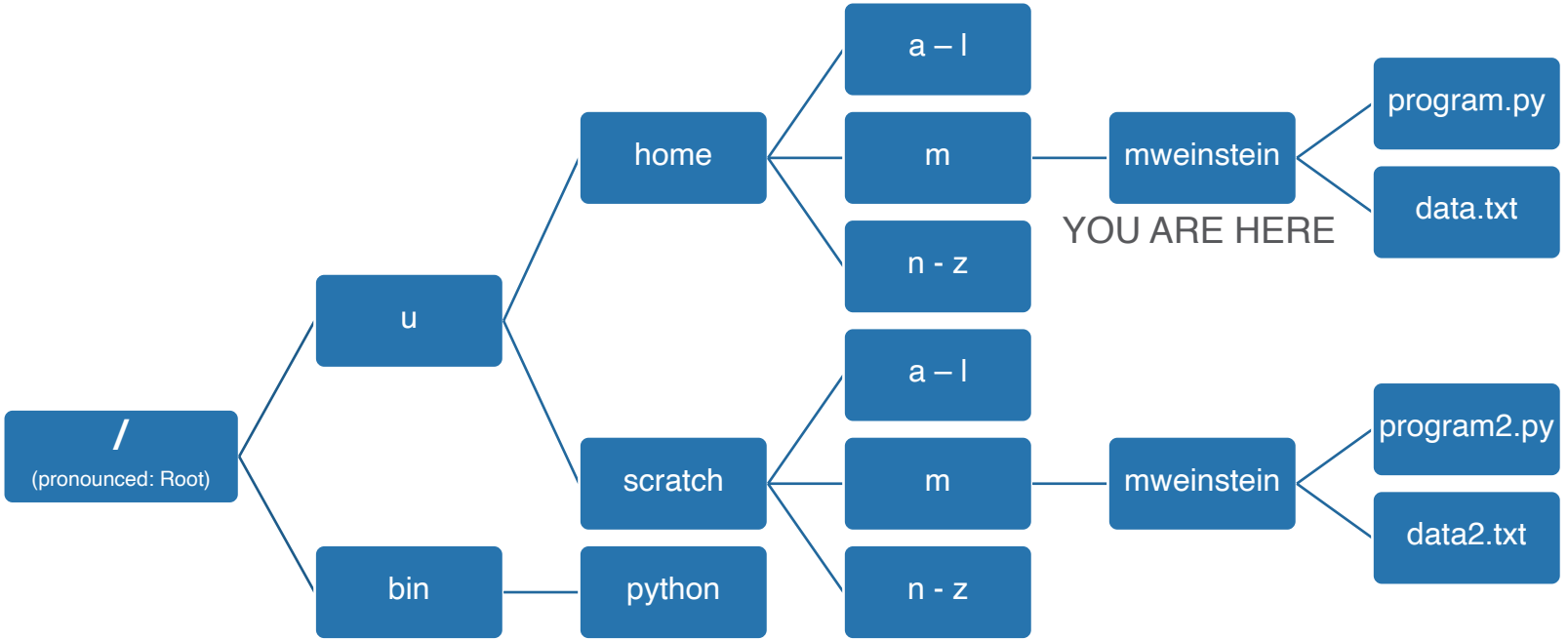

```
[mweinstein@computer ~]$ ls  ..
```

The Linux File System



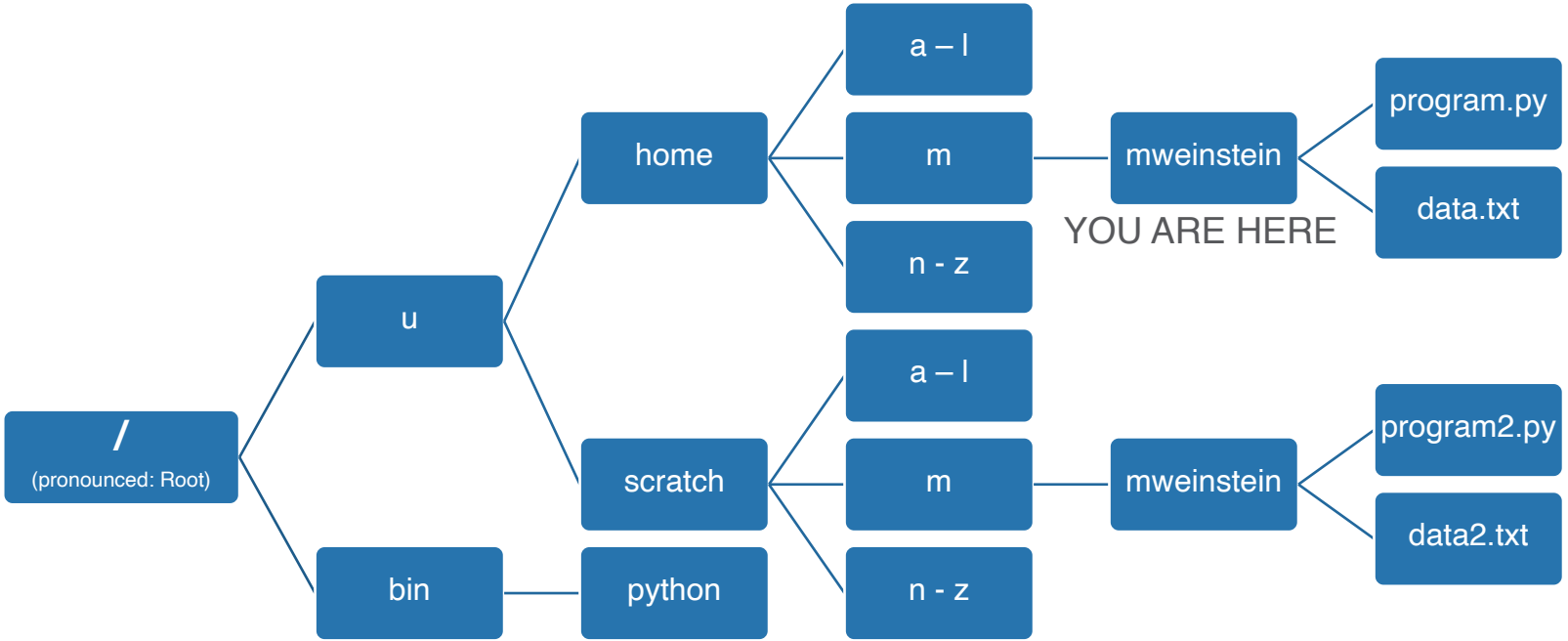

```
[mweinstein@computer ~]$ ls  ..
mweinstein
```

The Linux File System



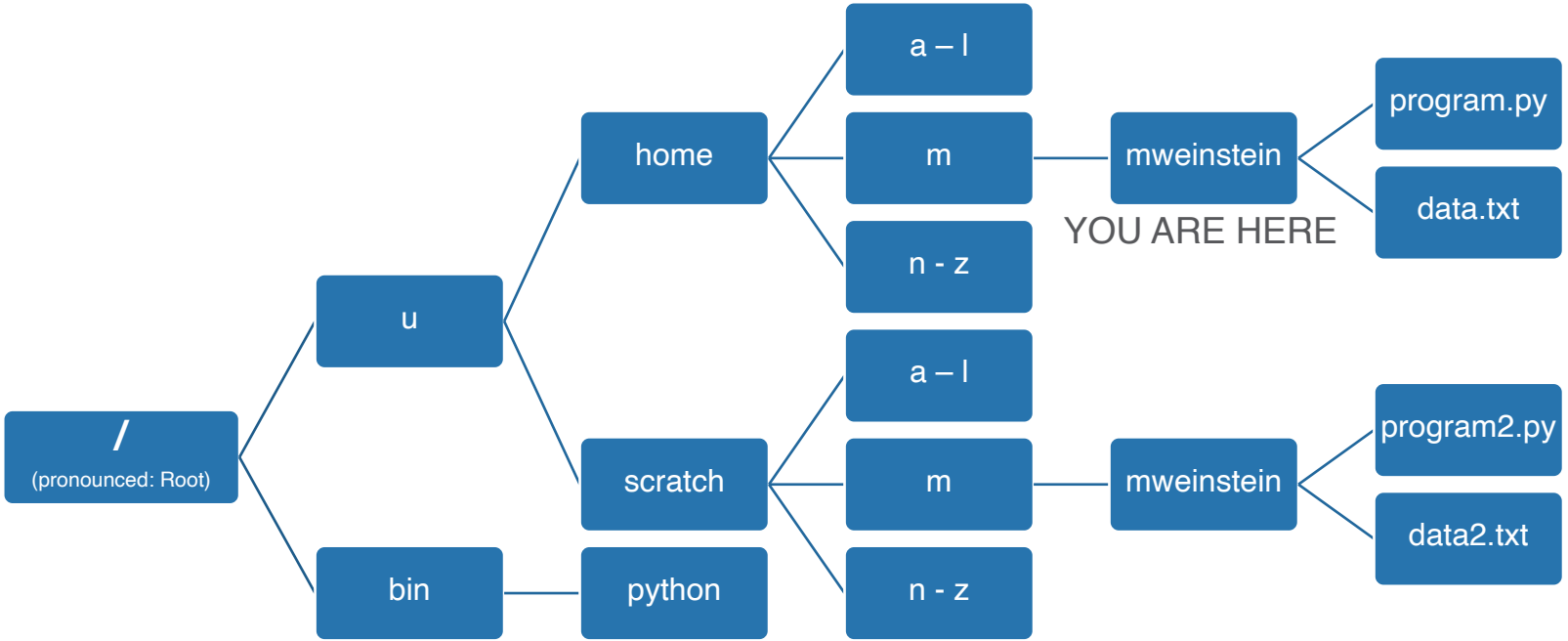

```
[mweinstein@computer ~]$ program.py  data.txt
```

The Linux File System



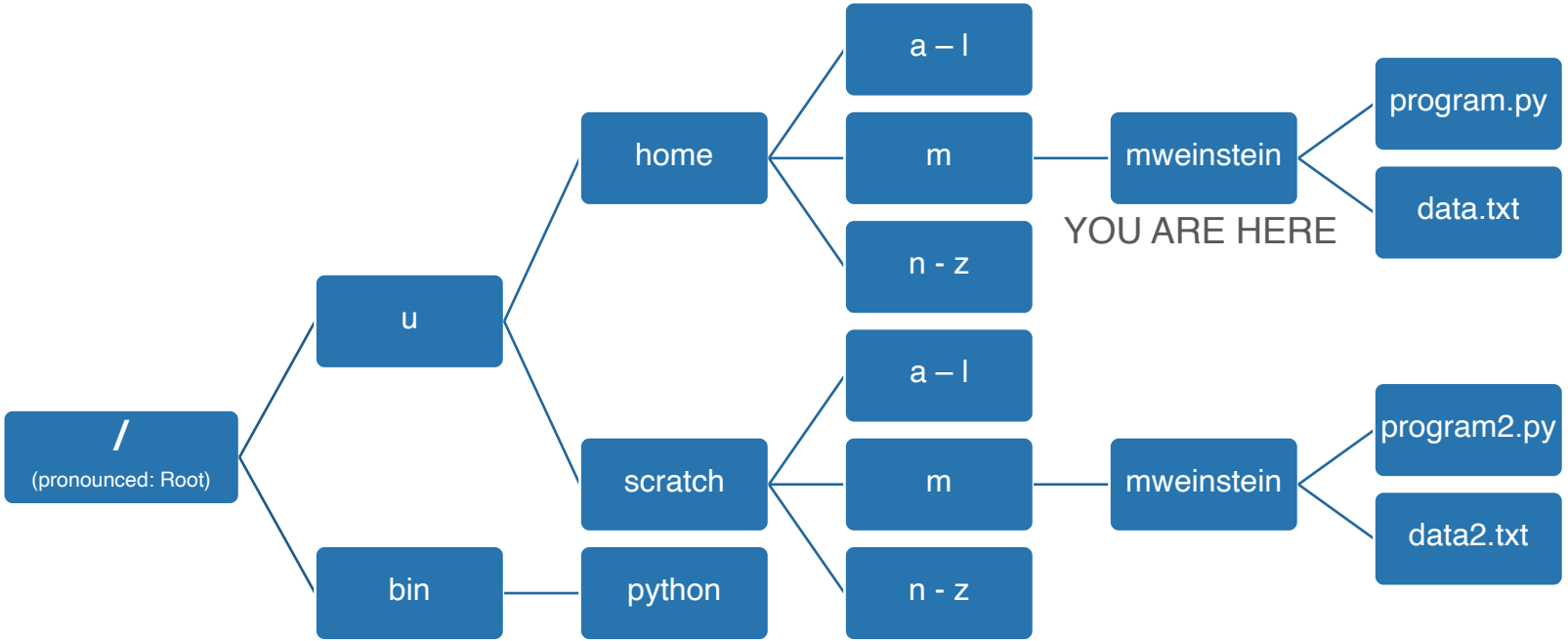

```
[mweinstein@computer ~]$ program.py  data.txt
ERROR! Executable not found
```

The Linux File System



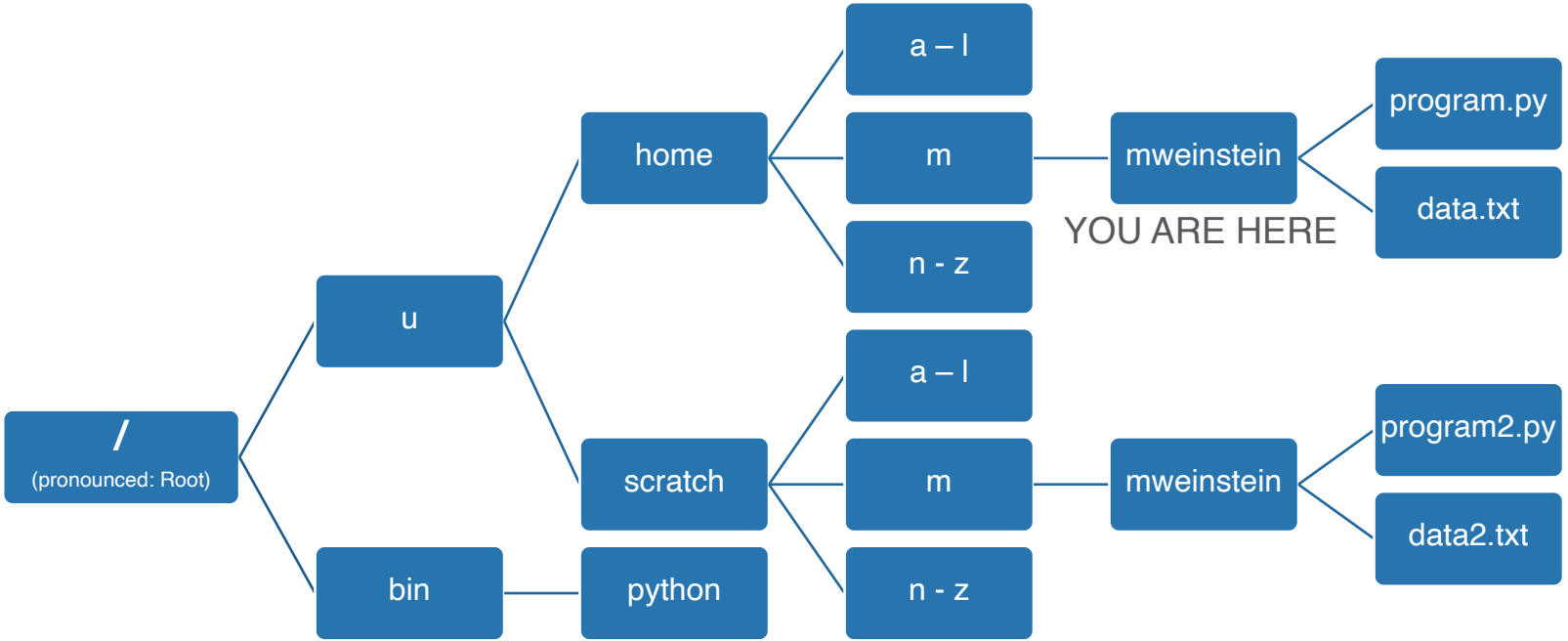

```
[mweinstein@computer ~]$ ./program.py  data.txt
```

The Linux File System



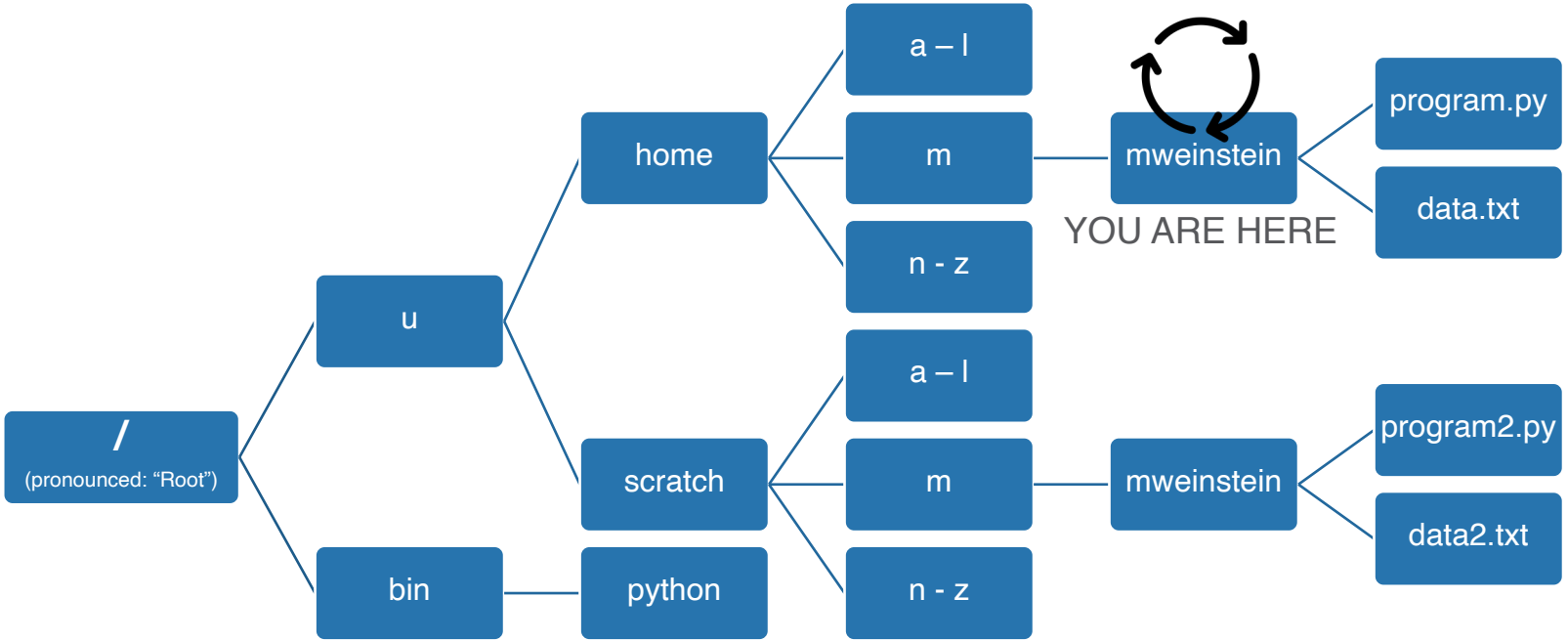

```
[mweinstein@computer ~]$ ./program.py  data.txt
```
[Executes program.py with data.txt as input]

The Linux File System



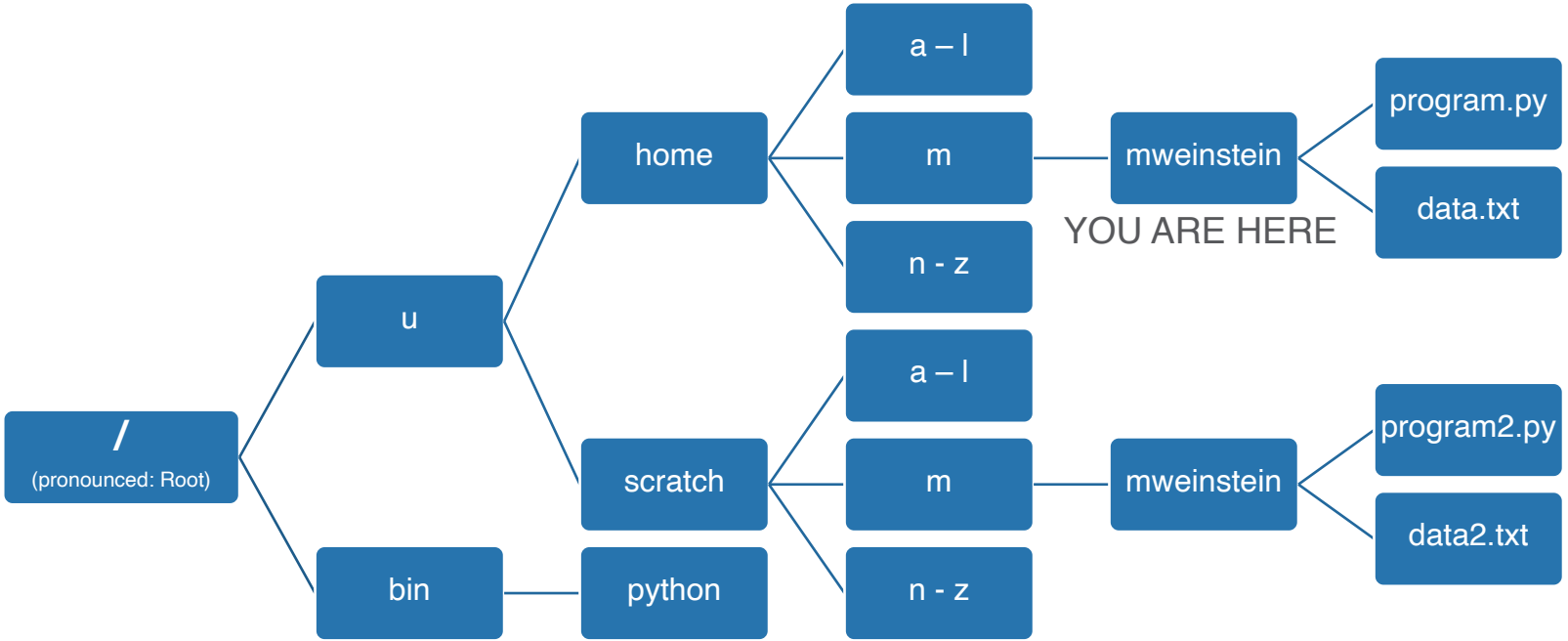

```
[mweinstein@computer ~]$ cd  .
```

The Linux File System



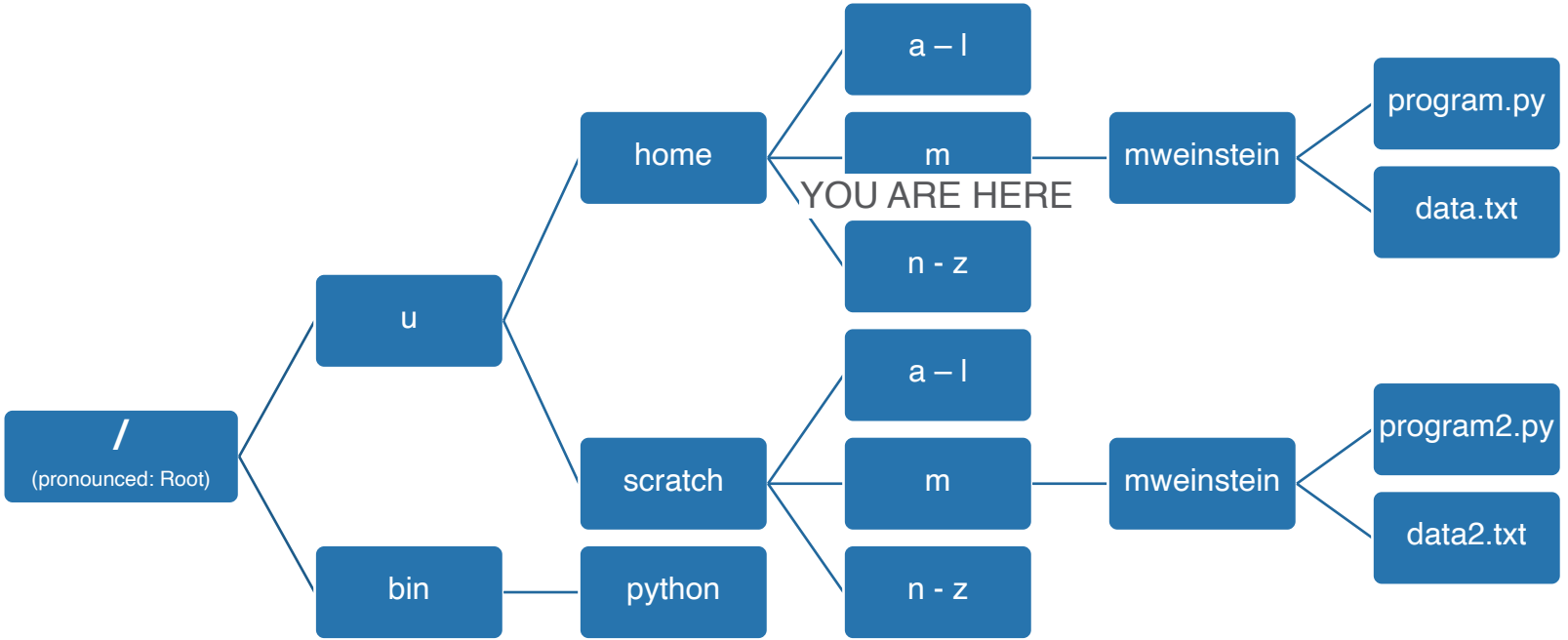

```
[mweinstein@computer ~]$ cd  .      [Does nothing, useless command]
[mweinstein@computer ~]$
```

The Linux File System



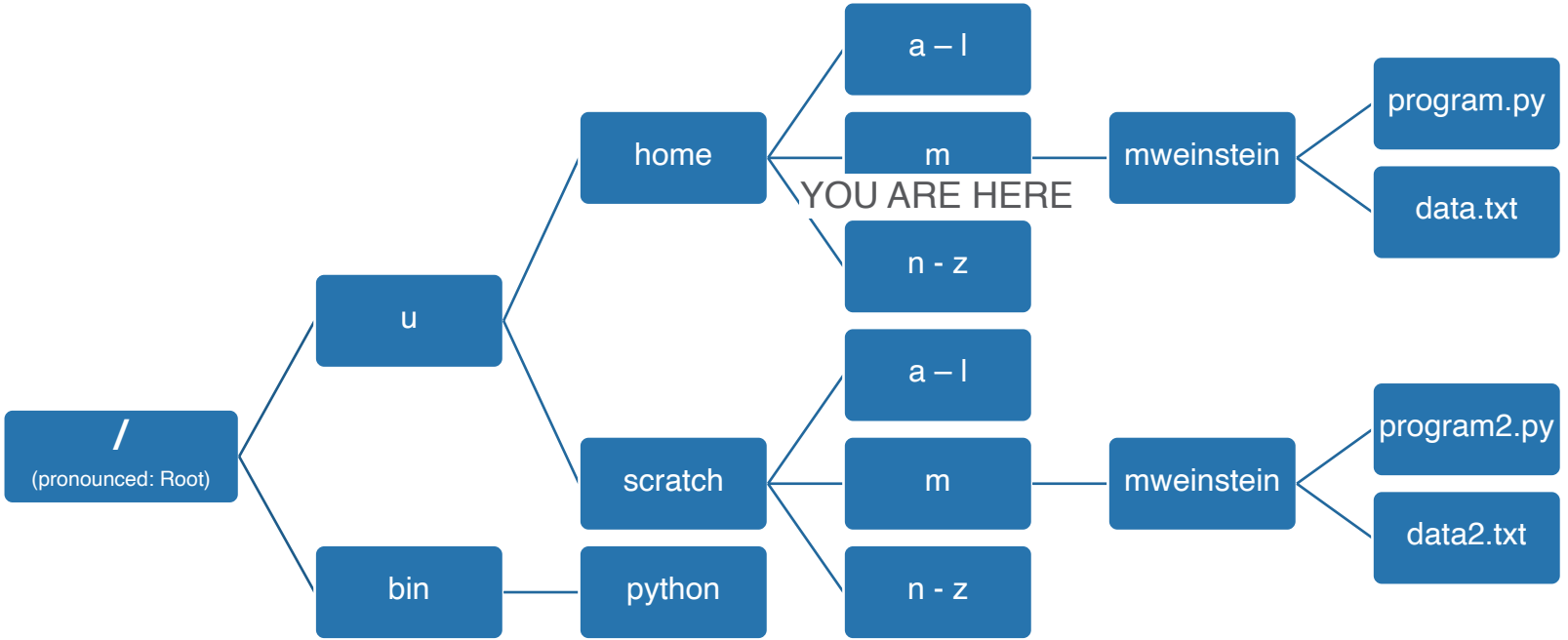

```
[mweinstein@computer ~]$ cd  ..
```

The Linux File System



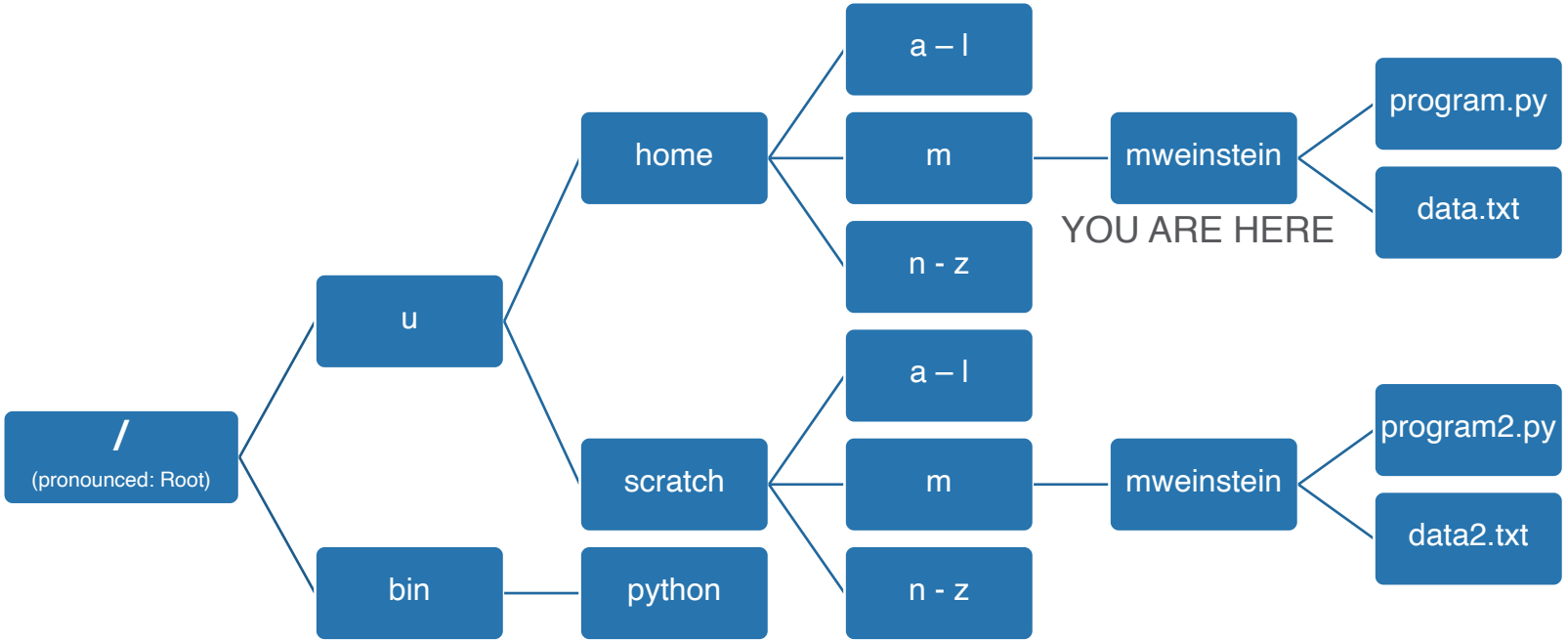

```
[mweinstein@computer ~]$ cd  ..
[mweinstein@computer m]$
```

The Linux File System




```
[mweinstein@computer m]$ cd  mweinstein
```

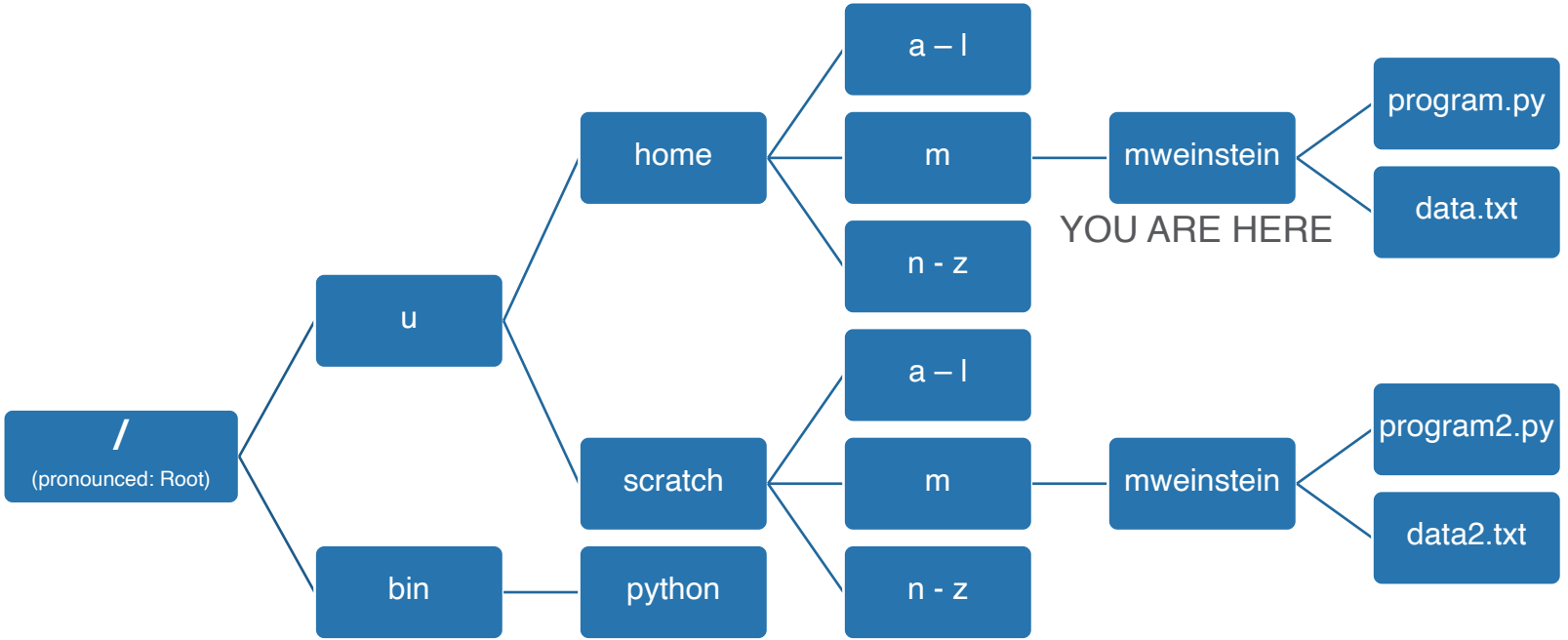
The Linux File System




```
[mweinstein@computer m]$ cd  mweinstein
[mweinstein@computer ~]$
```

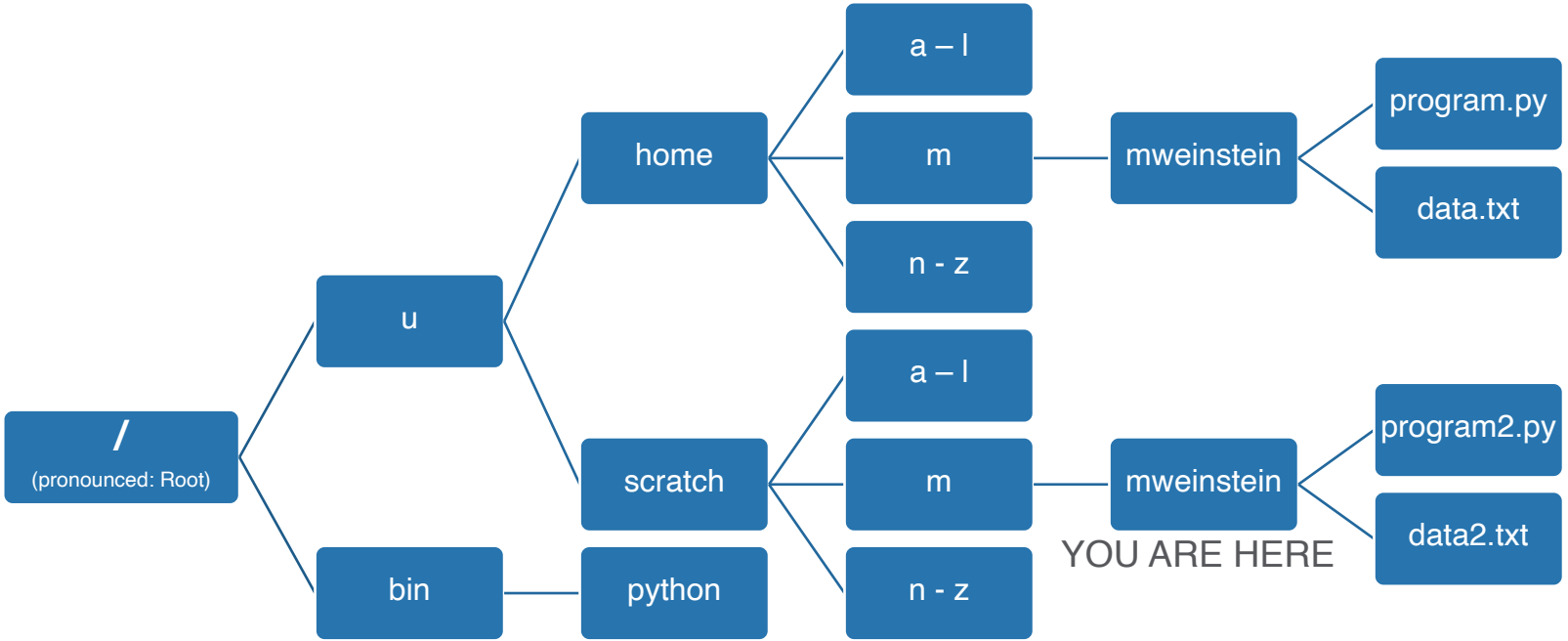
[Relative path]

The Linux File System



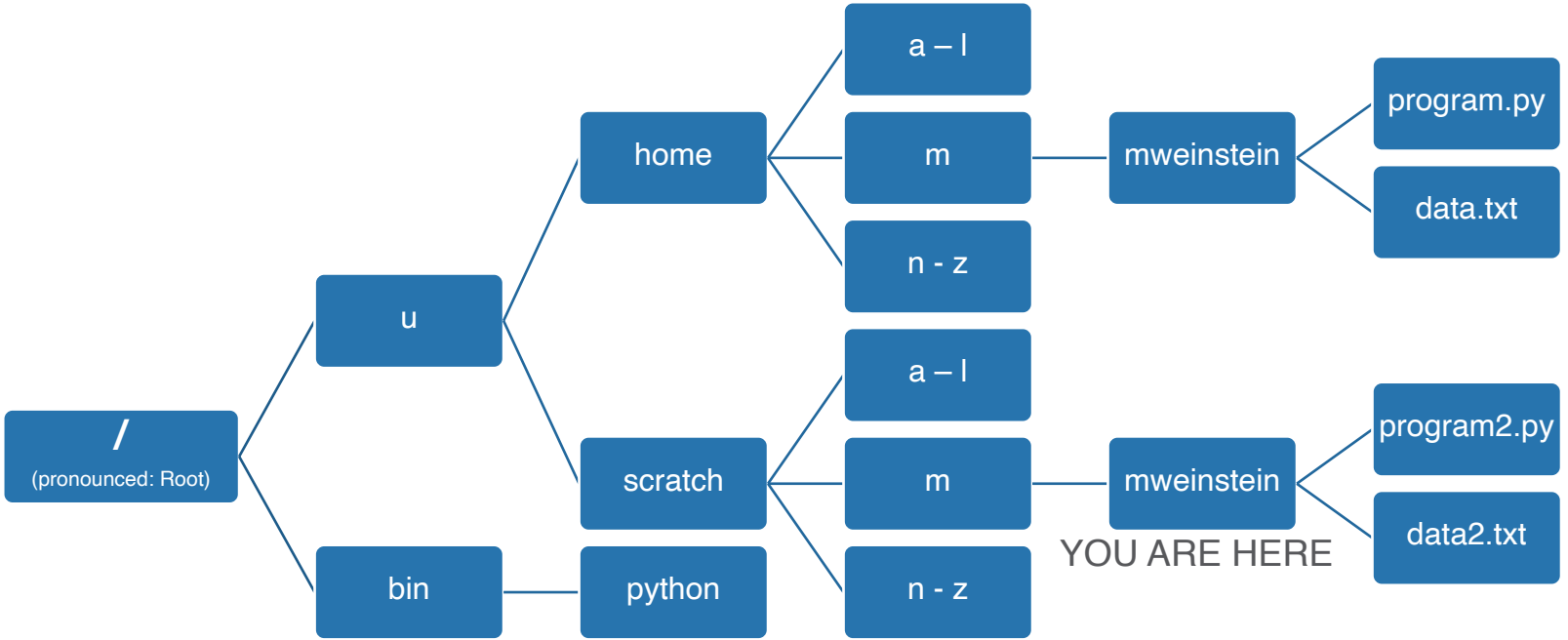

```
[mweinstein@computer ~]$ cd  /u/scratch/m/mweinstein
```

The Linux File System



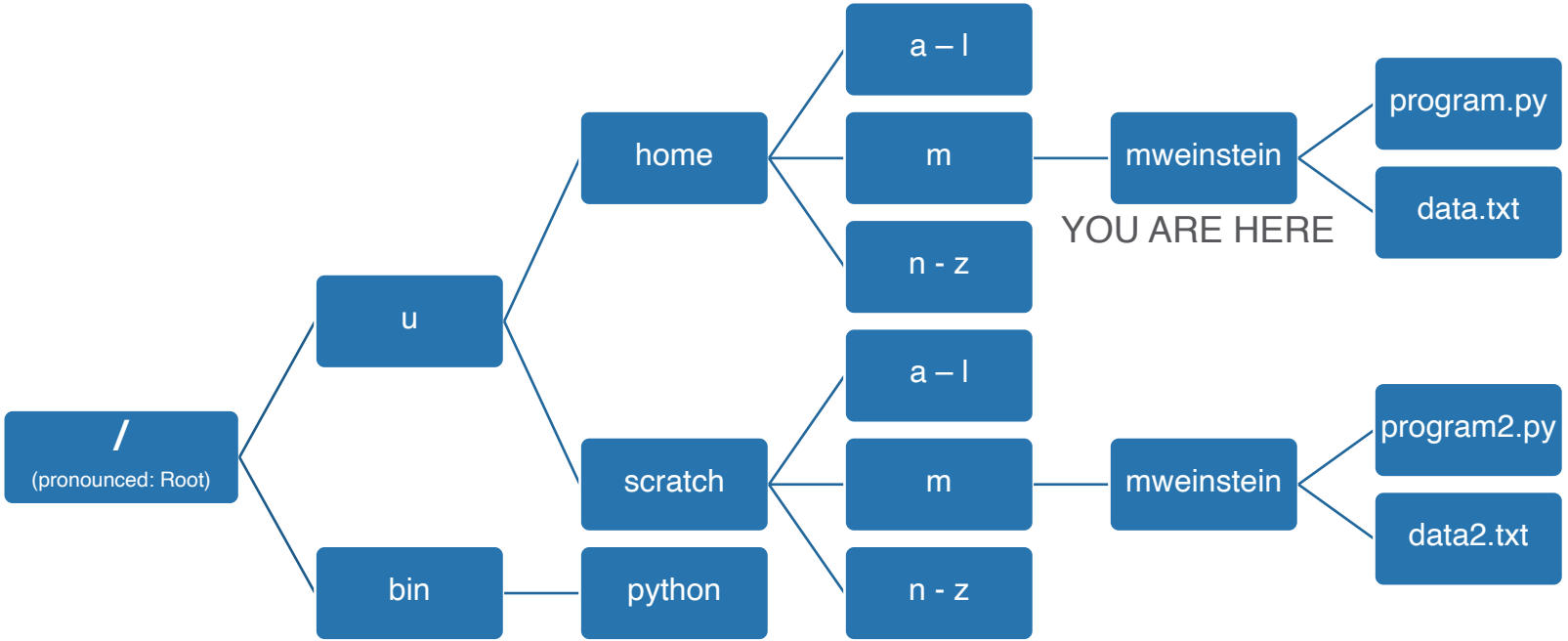

```
[mweinstein@computer ~]$ cd  /u/scratch/m/mweinstein     [Absolute Path]
[mweinstein@computer mweinstein]$
```

The Linux File System



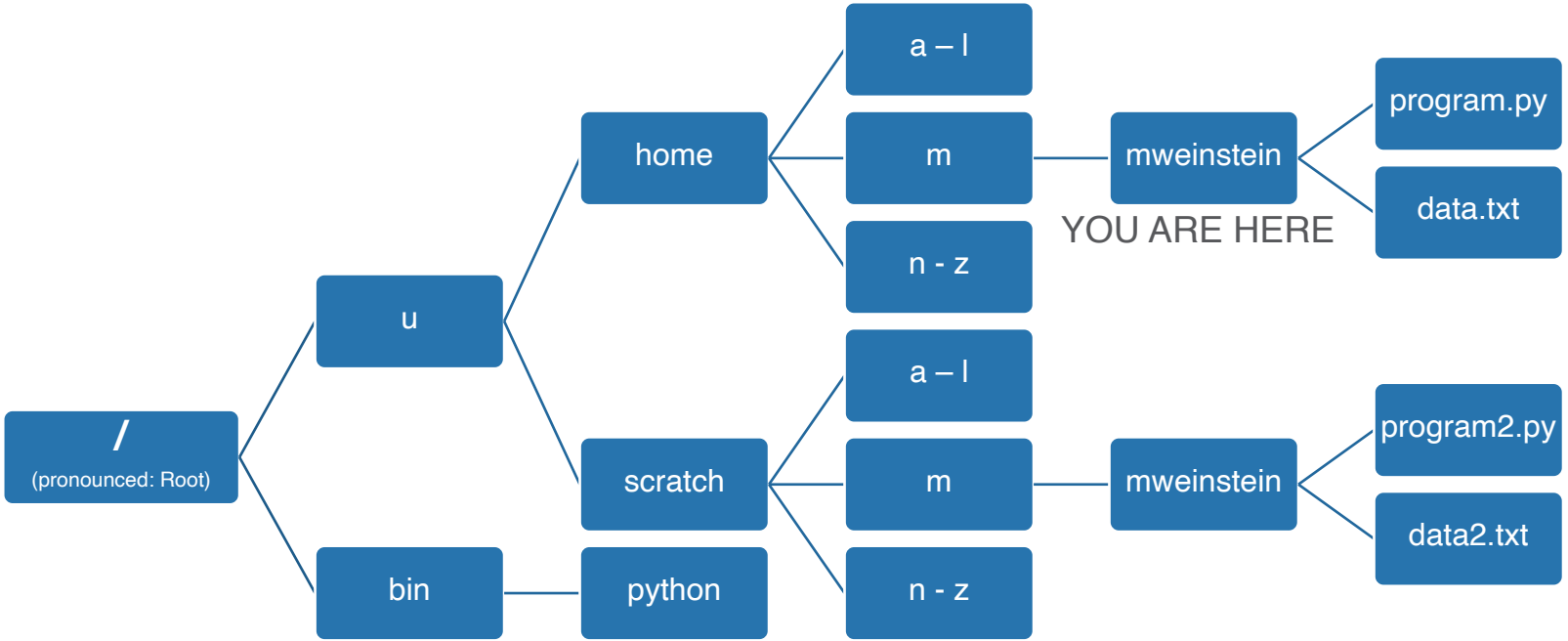

```
[mweinstein@computer mweinstein]$ cd  ~
```

The Linux File System



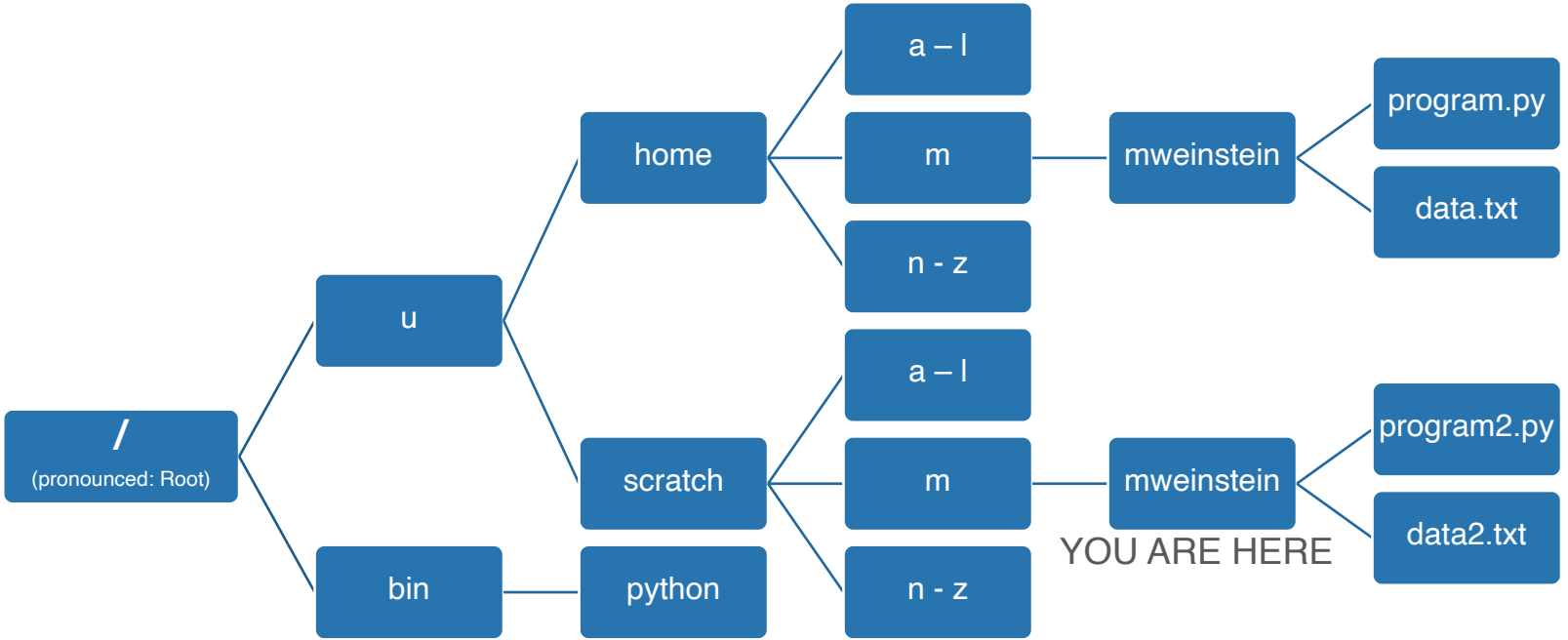

```
[mweinstein@computer mweinstein]$ cd  ~     [Shortcut home, most Linux]
[mweinstein@computer ~]$
```

The Linux File System




```
[mweinstein@computer ~]$ cd  $SCRATCH
```

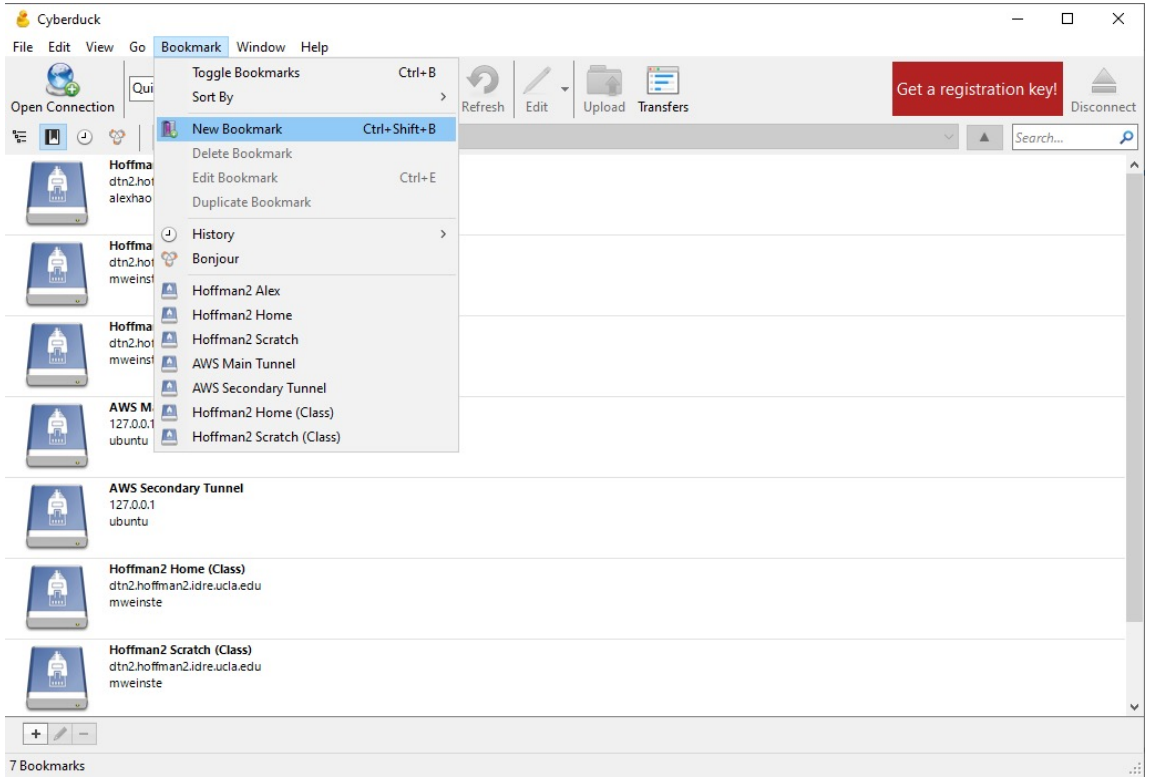
The Linux File System



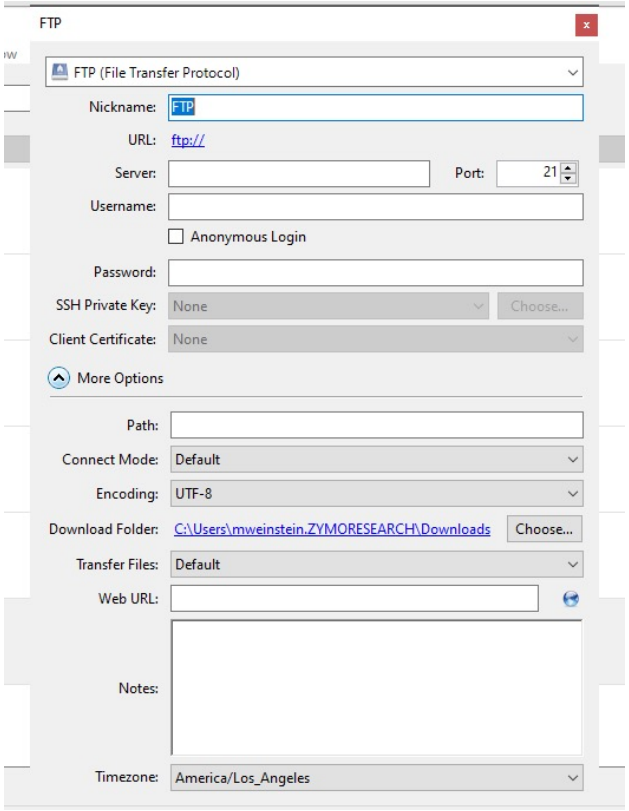

```
[mweinstein@computer ~]$ cd  $SCRATCH
[mweinstein@computer mweinstein]$
```

[An environment variable on Hoffman2]

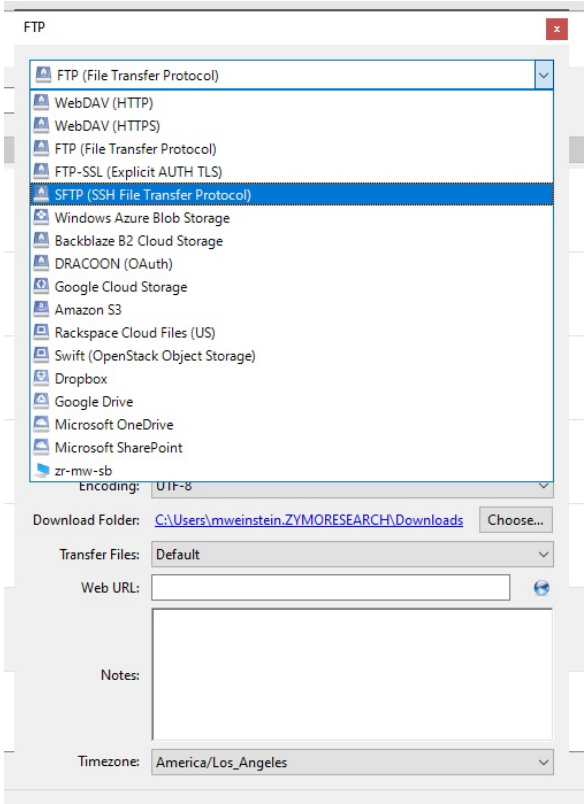
Cyberduck Setup Demo



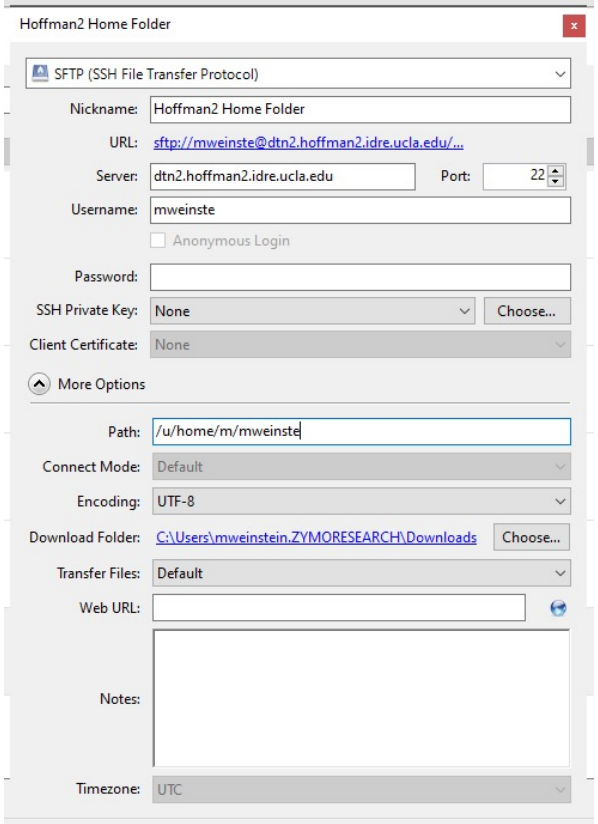
Cyberduck Setup Demo



Cyberduck Setup Demo



Cyberduck Setup Demo



Server:

`dtn2.hoffman2.idre.ucla.edu`

Username: Your username (like when you SSH in)

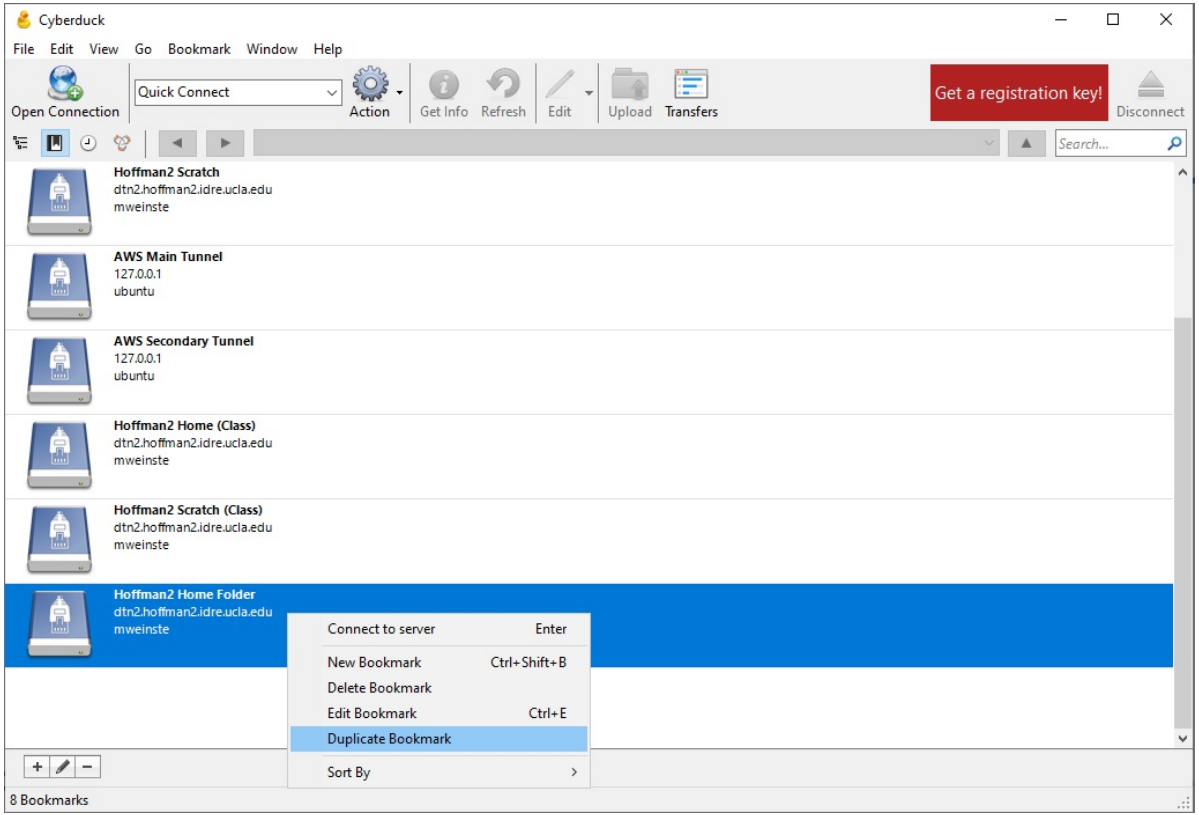
Password: Your SSH passowrd

Path: /u/home/m/mweinste (my home folder)

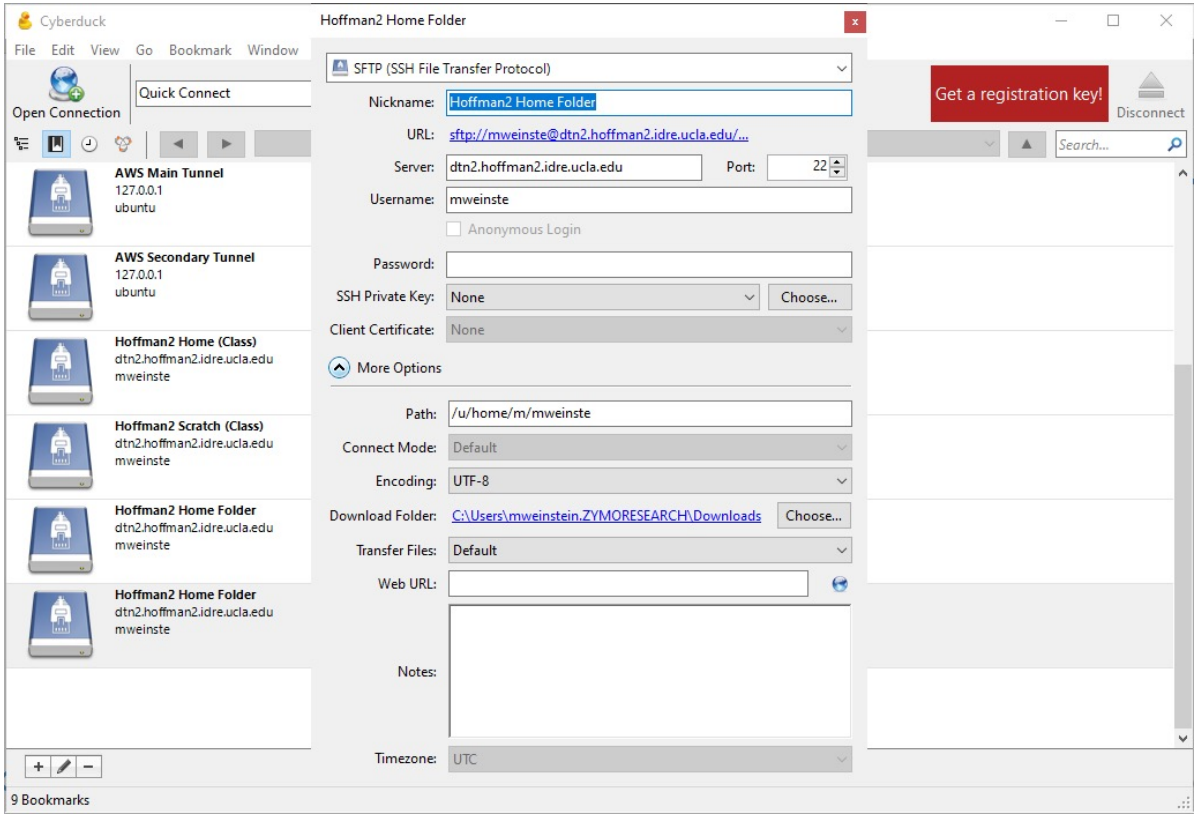
Your username

First letter of your username

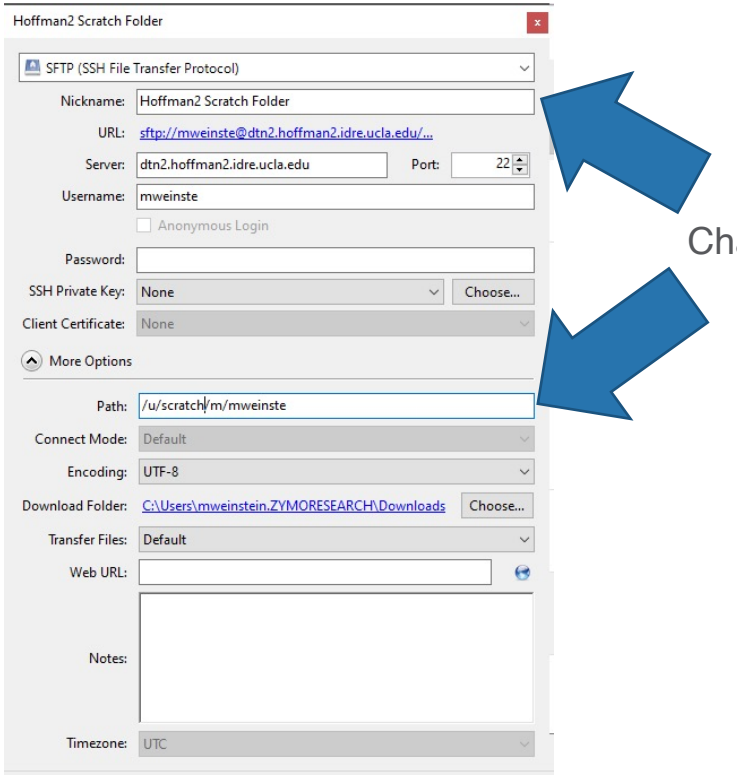
Cyberduck Setup Demo



Cyberduck Setup Demo

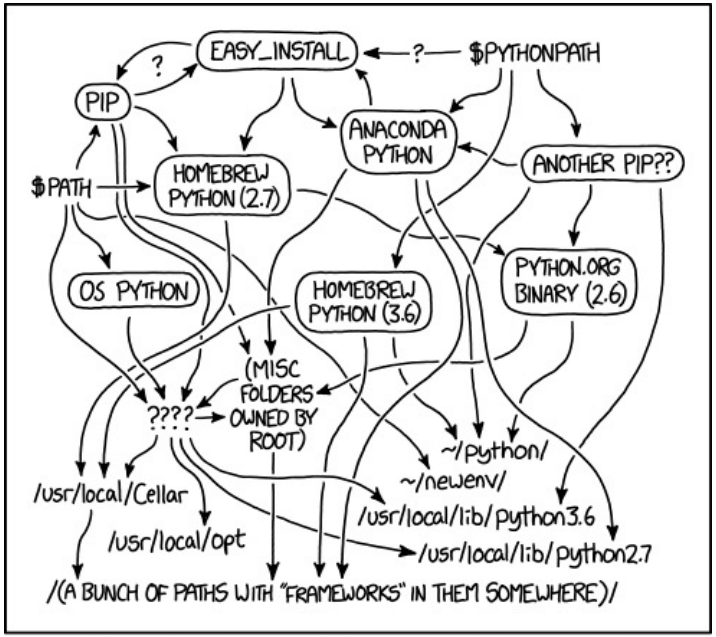


Cyberduck Setup Demo



Change “home” to “scratch”

Dependency Hell (from XKCD)



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Avoiding Dependency Hell

COMPARTMENTALIZE COMPARTMENTALIZE COMPARTMENTALIZE

Isolate and abstract your computing environment to make dealing with dependencies fast, easy, and reproducible across systems

- Virtual Machines
  - Well isolated
  - Heavy for storage and resources
- Containers (Docker or Singularity)
  - Pretty well isolated
  - Lighter for resources
- Environment Managers (VirtualEnv, Anaconda, Module Load)
  - Super light
  - Available on Hoffman2
  - Poorly isolated, can sometimes use external dependencies you don't realize

Topic

COURSE GOALS

Course Goals

WHAT YOU SHOULD BE ABLE TO UNDERSTAND AND DO AFTER THIS CLASS

Microbiome informatics is a relatively young field with the software and analysis landscape changing every day. We are aiming for depth on fundamentals over breadth of techniques.

- Understand the basics of sequencing and operating in a Linux/Unix environment
- Understand the differences between methods of organism identification
- Understand the differences between targeted and shotgun seqeuncing
- Understand and execute organism identification using a database
  - Understand the origins, biases, and limitations of identification and databases
- Understand and execute a comparison of different microbiome samples

Topic

TYPES OF ANALYSIS

What Do You Want To Know?

TO GET THE DESIRED ANSWER, WE MUST ASK THE RIGHT QUESTION

Microbiomics/metagenomics gives us a few options for what we examine, what we look at determines what we see.

- DNA sequencing
  - Whole genome shotgun sequencing (sequence everything)
    - Picks up functional genes, we know what this microbiome is capable of
    - Cannot distinguish host from microbe
    - Larger genomespace – harder informatics, bigger computers
    - Some genes are notoriously mobile
    - Worse, others are mobile, but not notoriously
  - Targeted sequencing
    - Can target a gene common to species of interest (often domain-level)
    - Variability in the gene can provide a “barcode” of identity
    - Function can be inferred at best

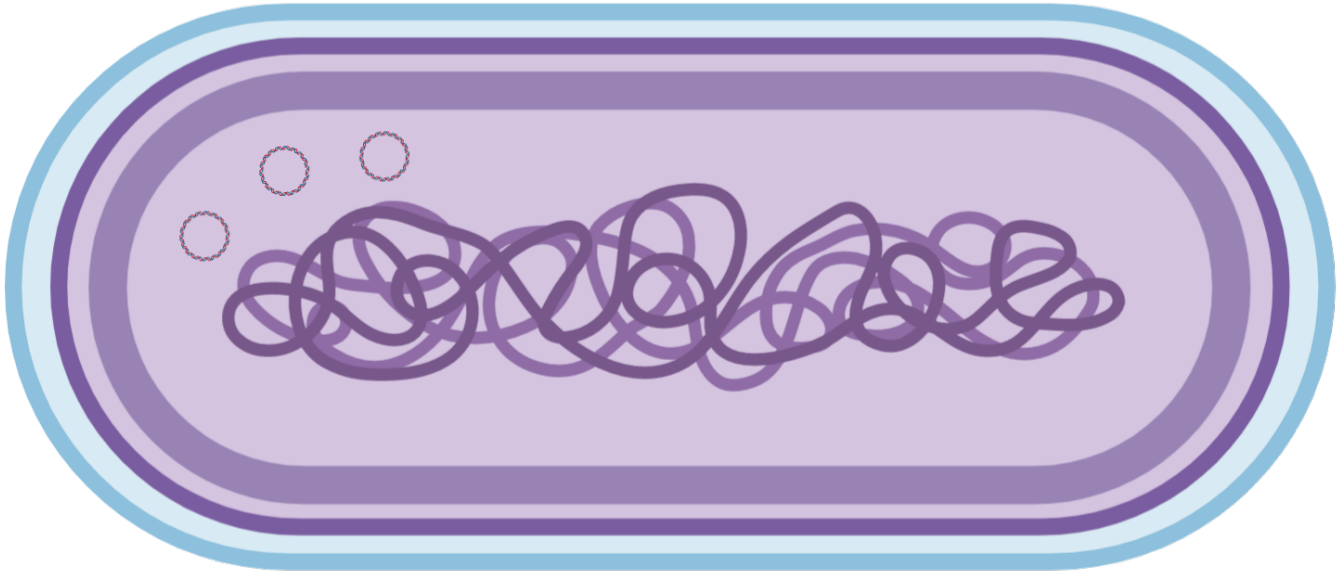
What Do You Want To Know?

TO GET THE DESIRED ANSWER, WE MUST ASK THE RIGHT QUESTION

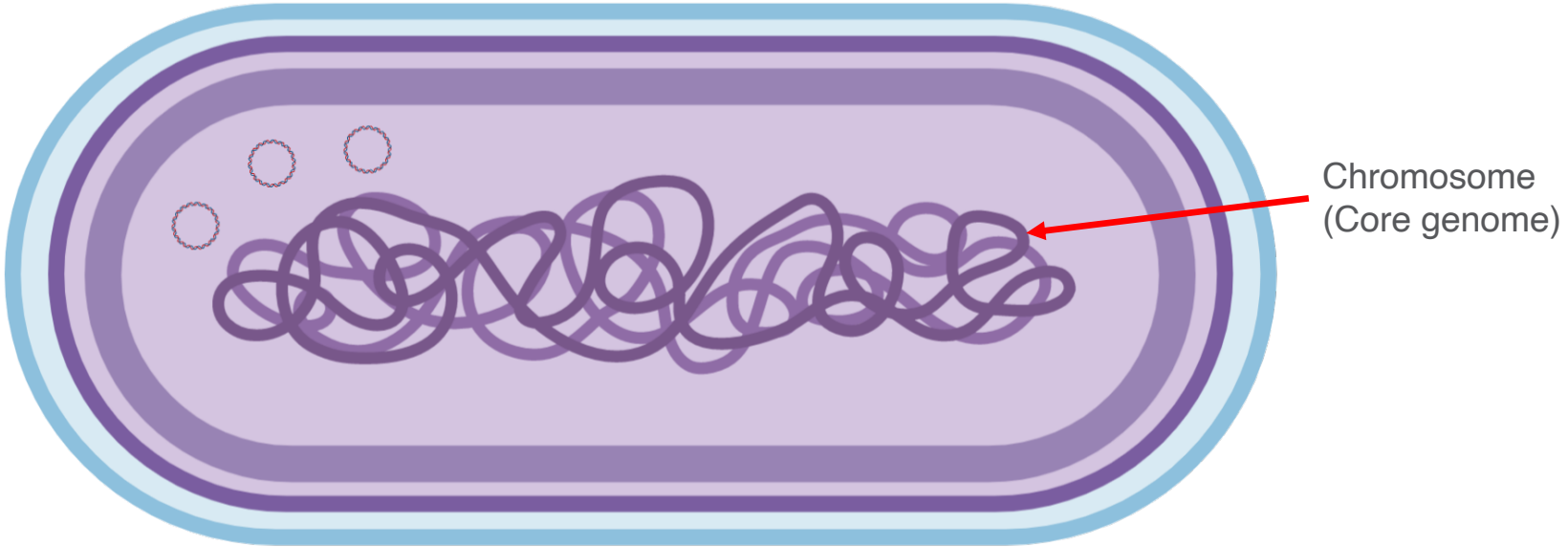
Microbiomics/metagenomics gives us a few options for what we examine, what we look at determines what we see.

- RNA sequencing/Metatranscriptomics
  - Transcripts can be tied back to origin by sequence to a decent extent
  - You only get decent coverage of expressed genome portions
  - Need to remove ribosomal RNA
  - Quantitative gene expression – what are these microbes doing?
- Metabolomics
  - Chemicalspace can be a nightmare if not careful
    - Lipids, nucleic acids, carbohydrates, amino acids, proteins, etc.
    - Some chemicals are volatile or unstable, sample process can be a confounding
    - What are these microbes making and breaking?

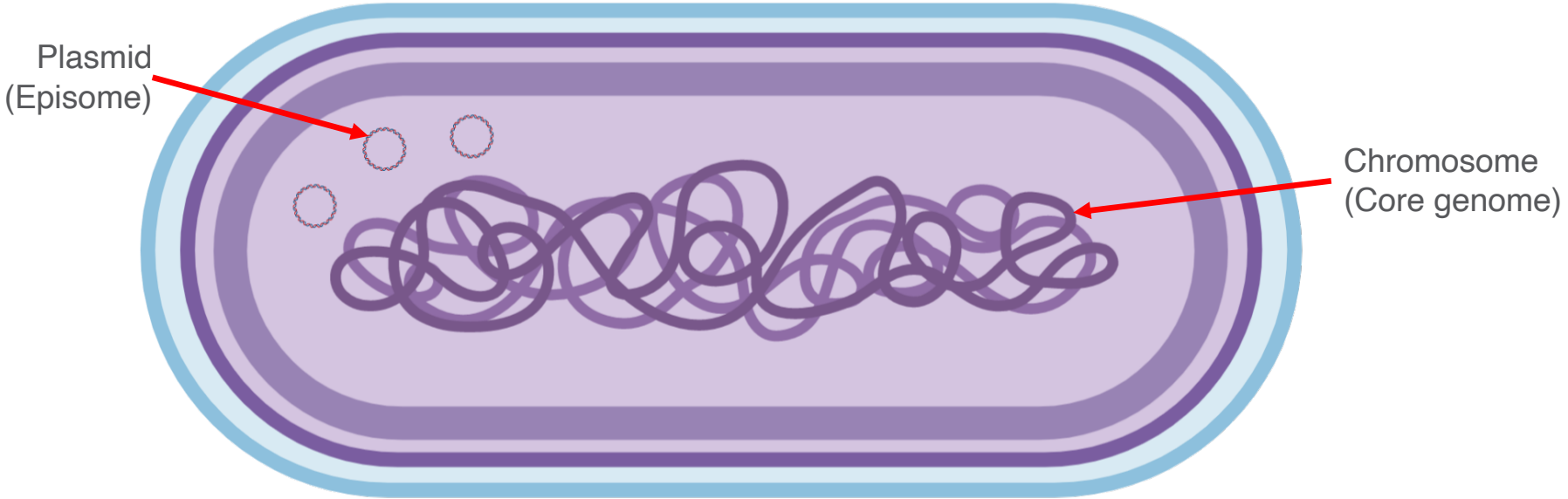
DNA Provides a Few Options



DNA Provides a Few Options

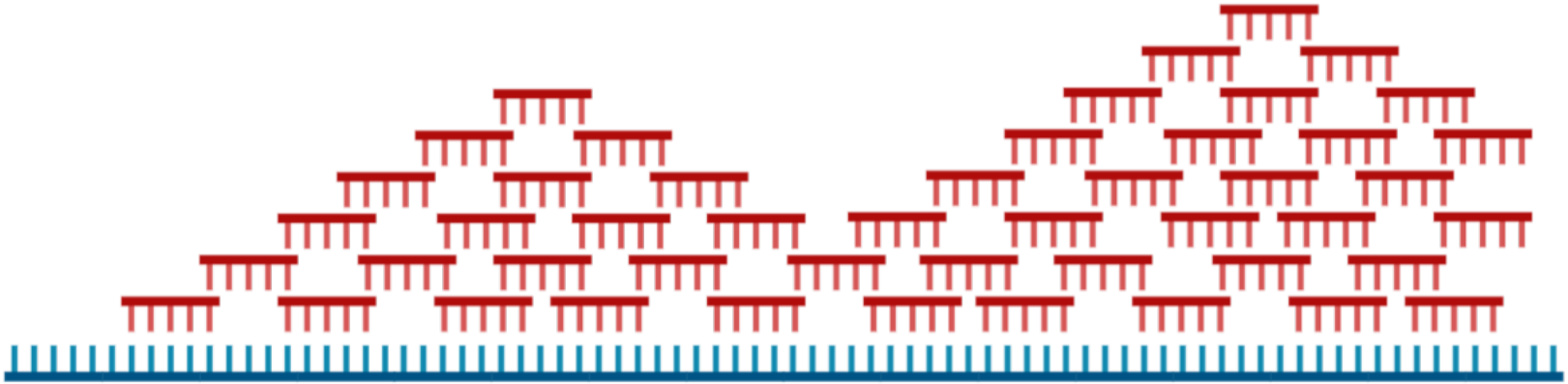


DNA Provides a Few Options



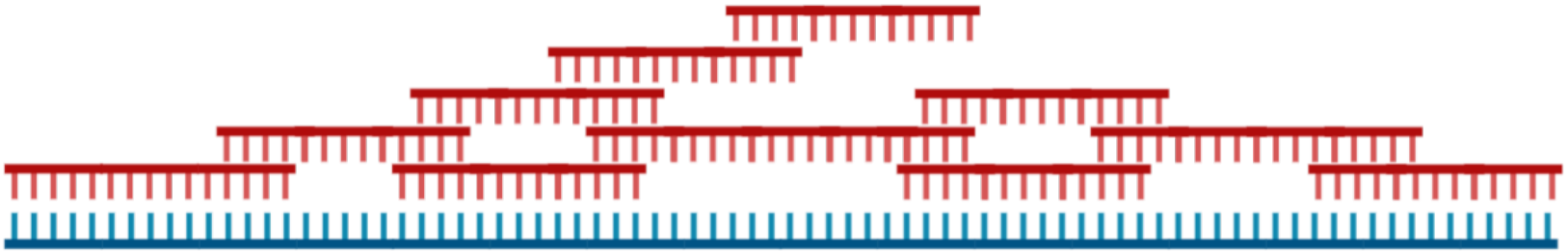
DNA Provides a Few Options

Genome Assembly Is Possible…



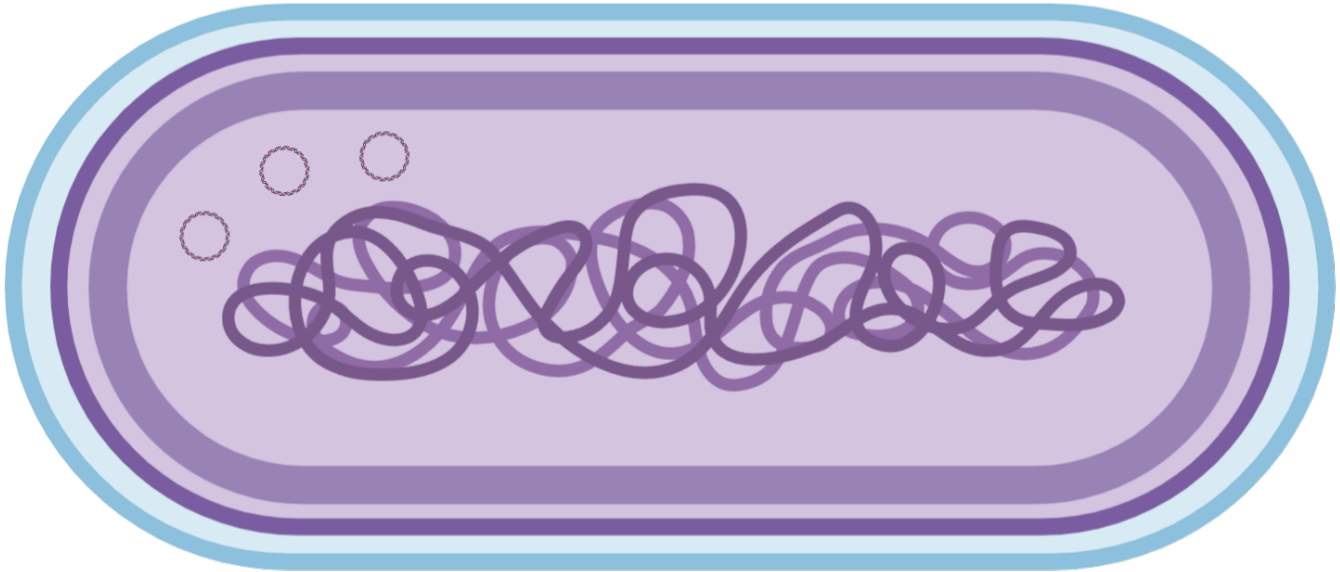
…with enough reads

Genome Assembly Is Possible…



…or with longer reads

Some Species Were Differentiated on Their Plasmids

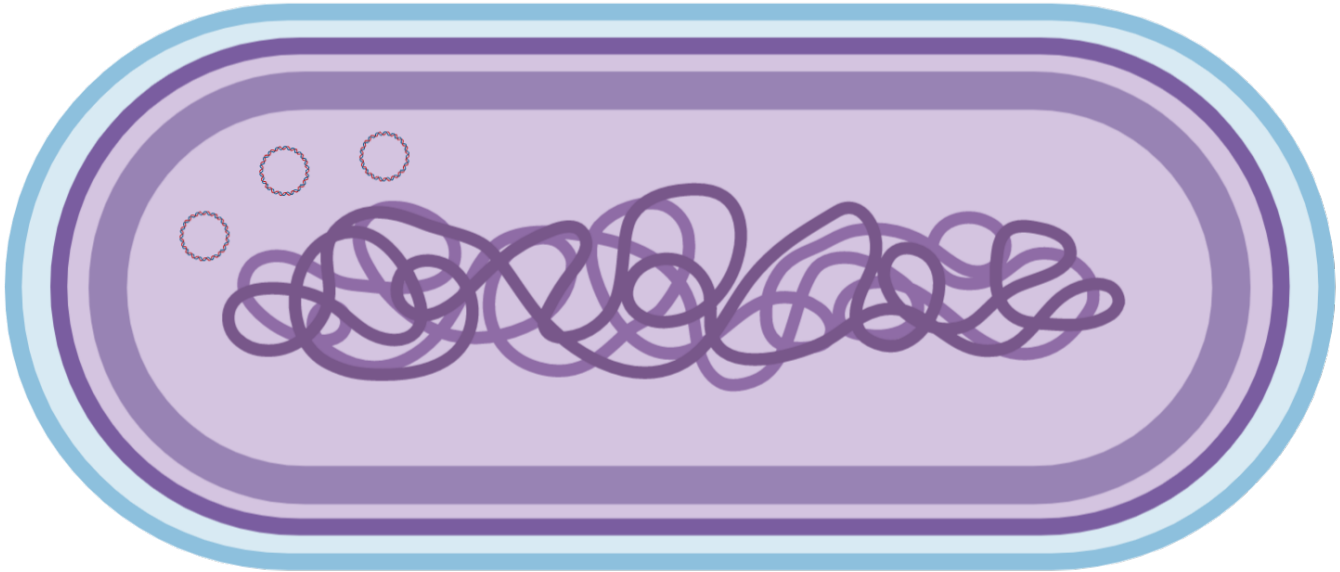


Bacillus cereus

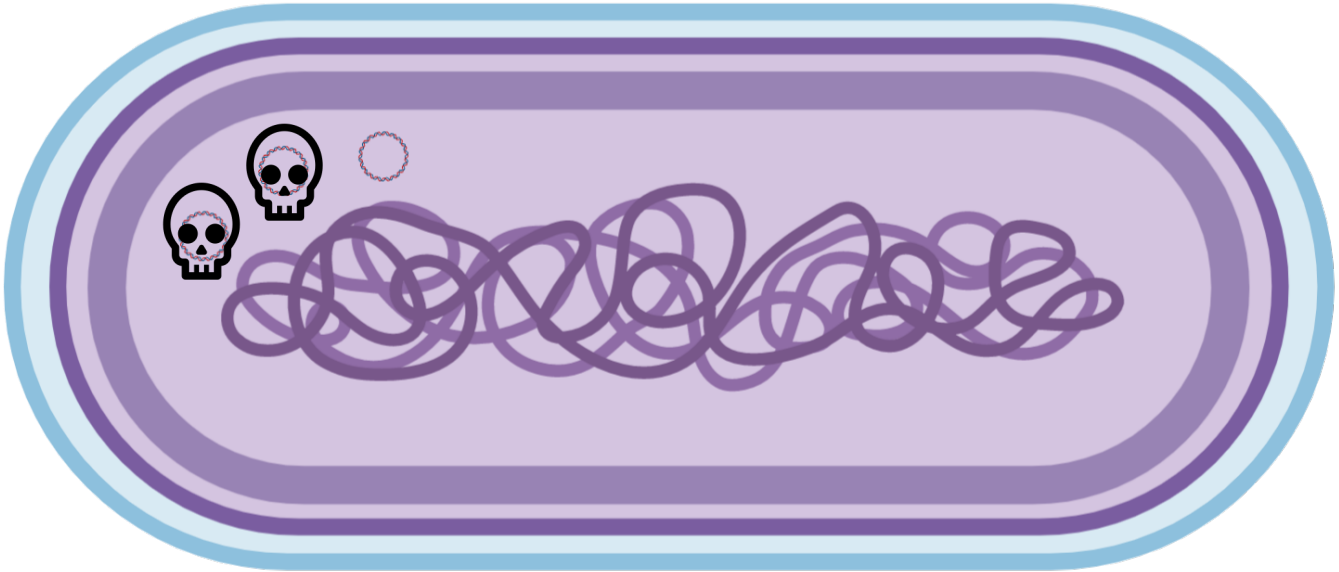
Some Species Were Differentiated on Their Plasmids

Bacillus cereus

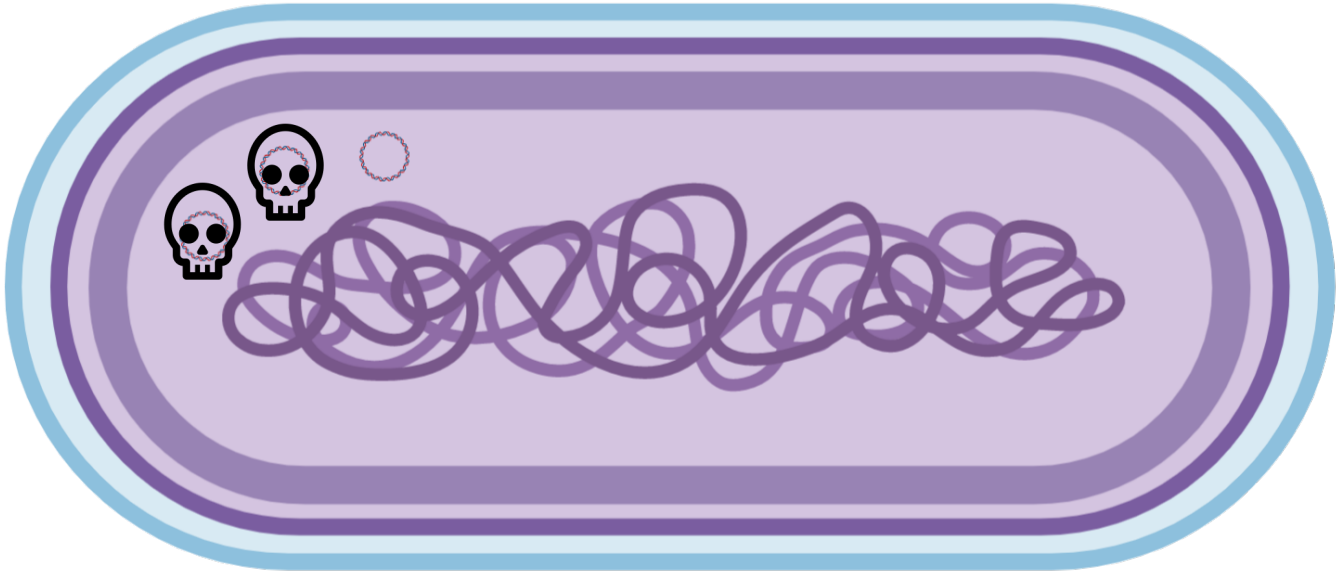
Some Species Were Differentiated on Their Plasmids



Some Species Were Differentiated on Their Plasmids



Some Species Were Differentiated on Their Plasmids

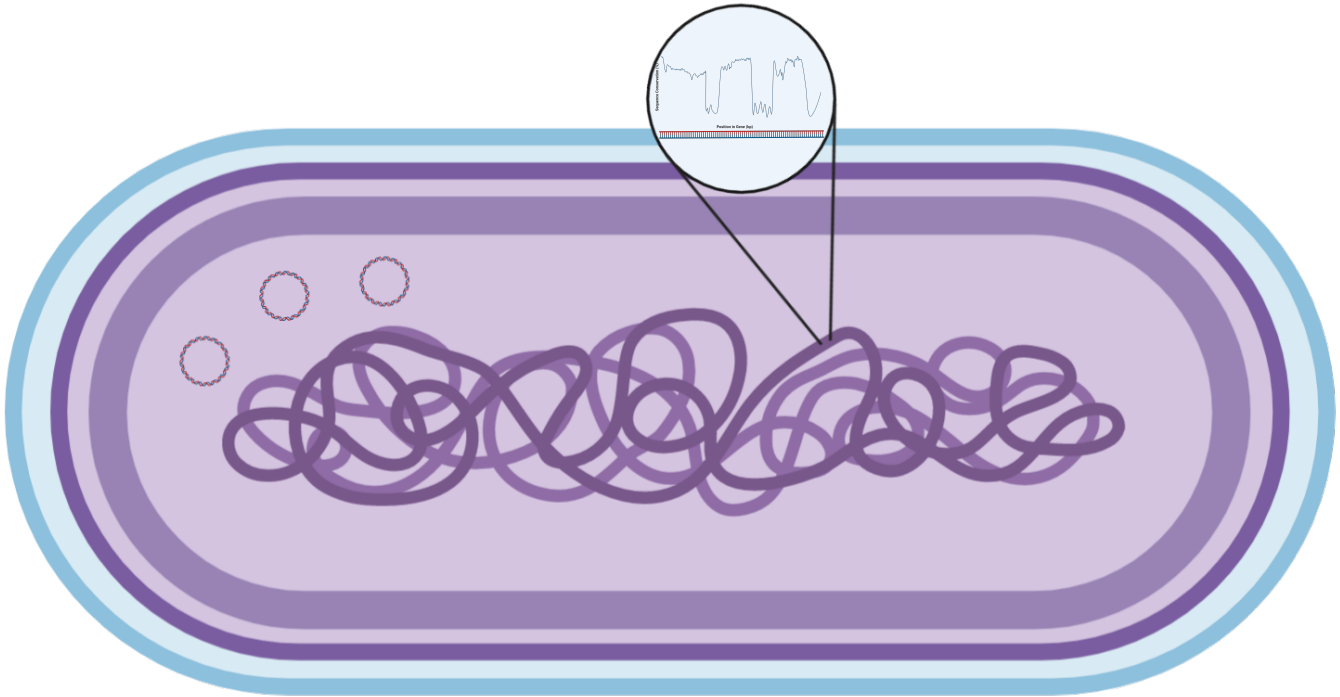


Bacillus anthracis

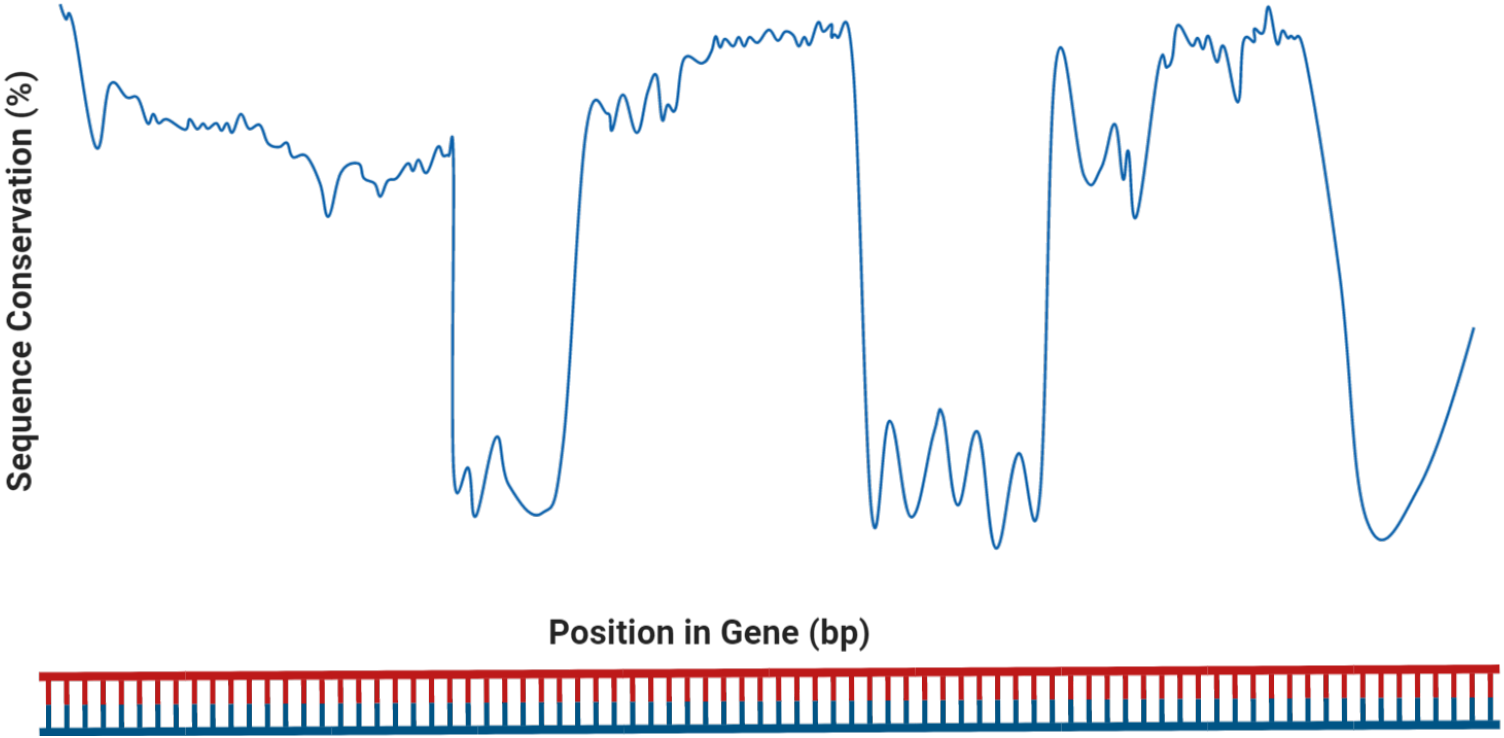
Some Species Were Differentiated on Their Plasmids

Some Species Were Differentiated on Their Plasmids

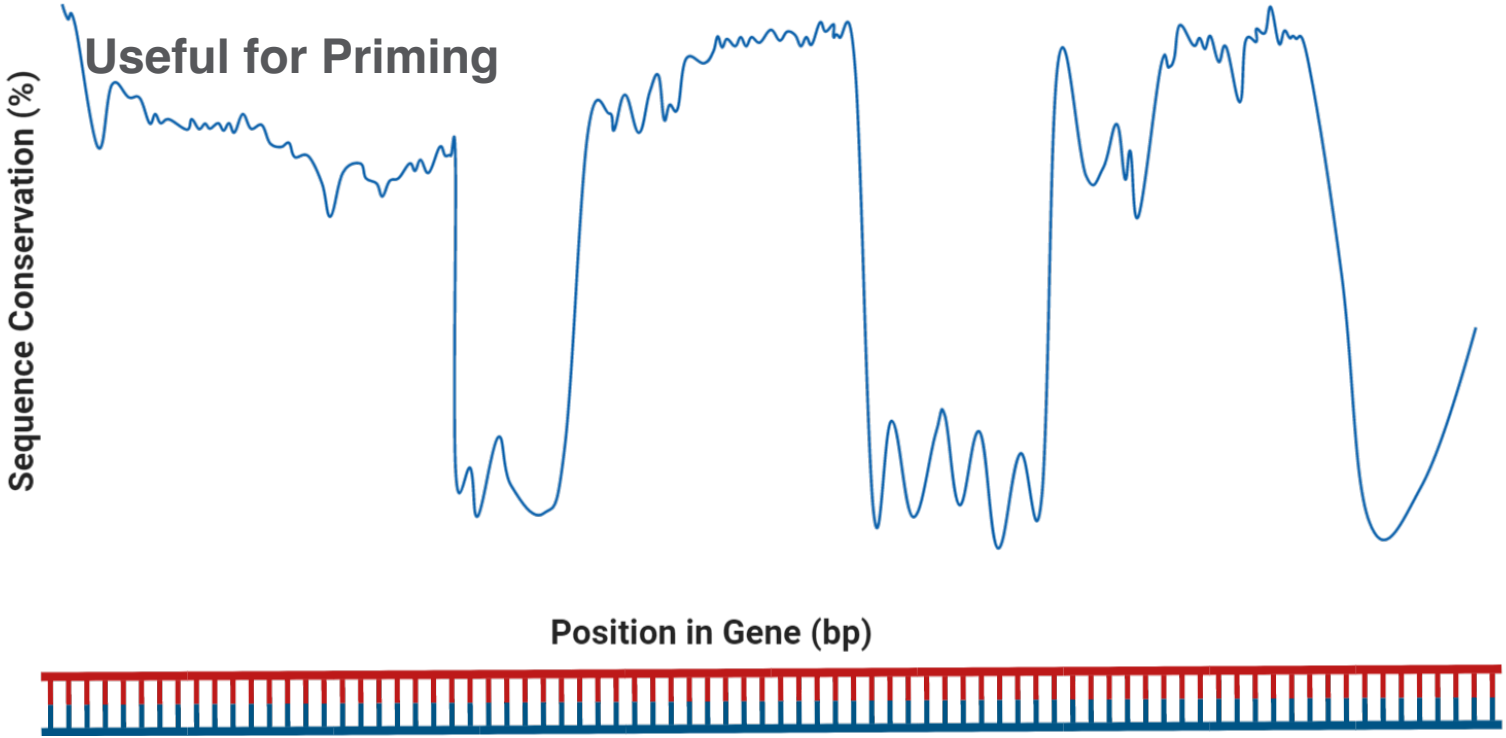
DNA Provides a Few Options



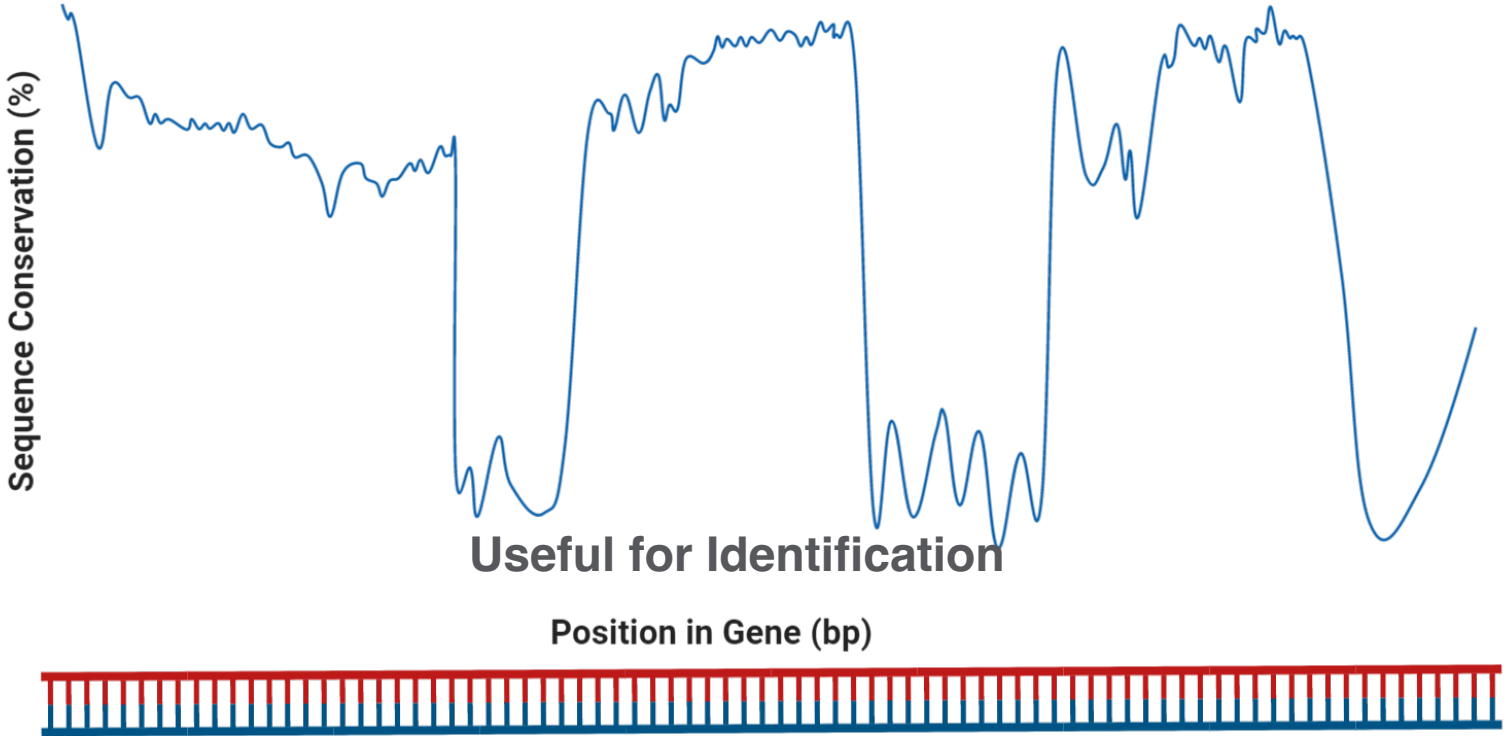
DNA Provides a Few Options



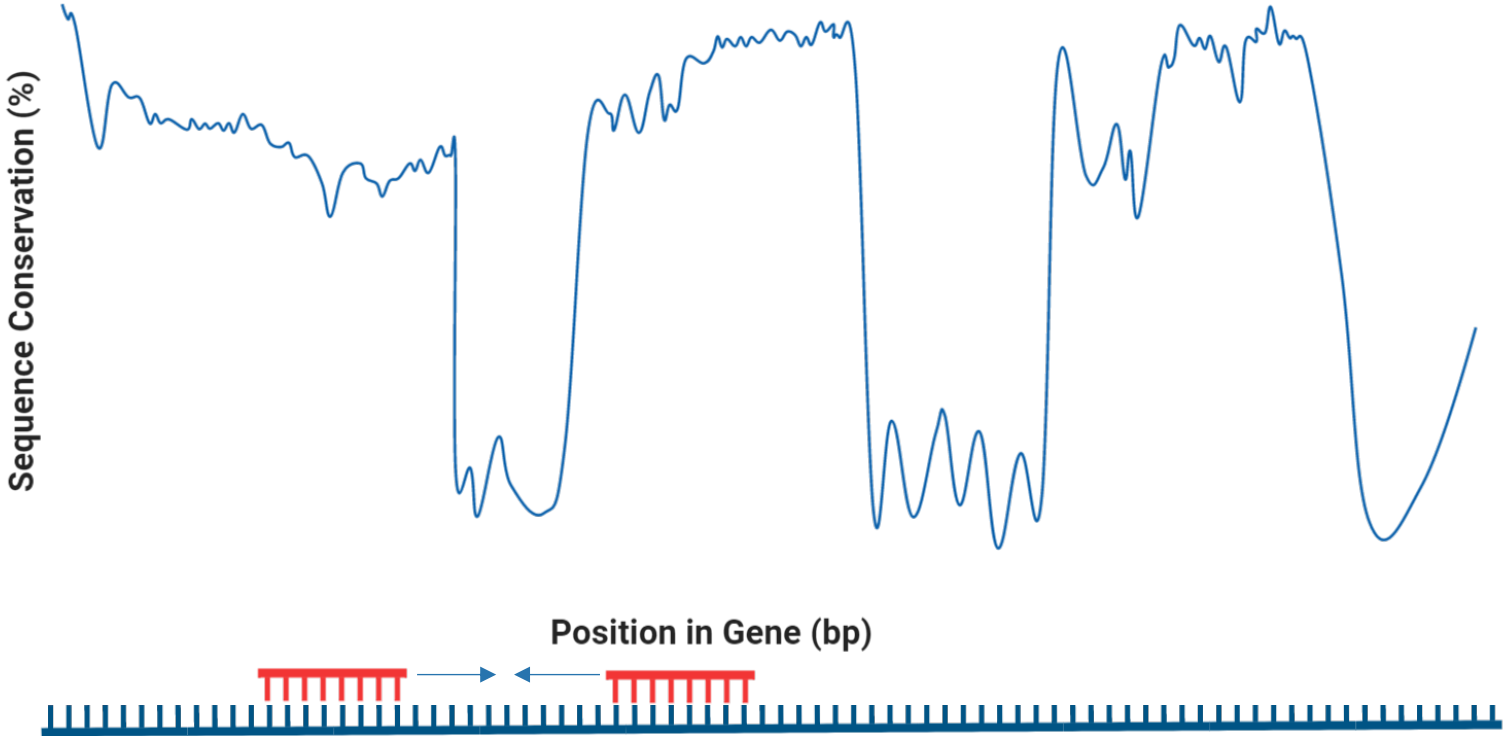
DNA Provides a Few Options



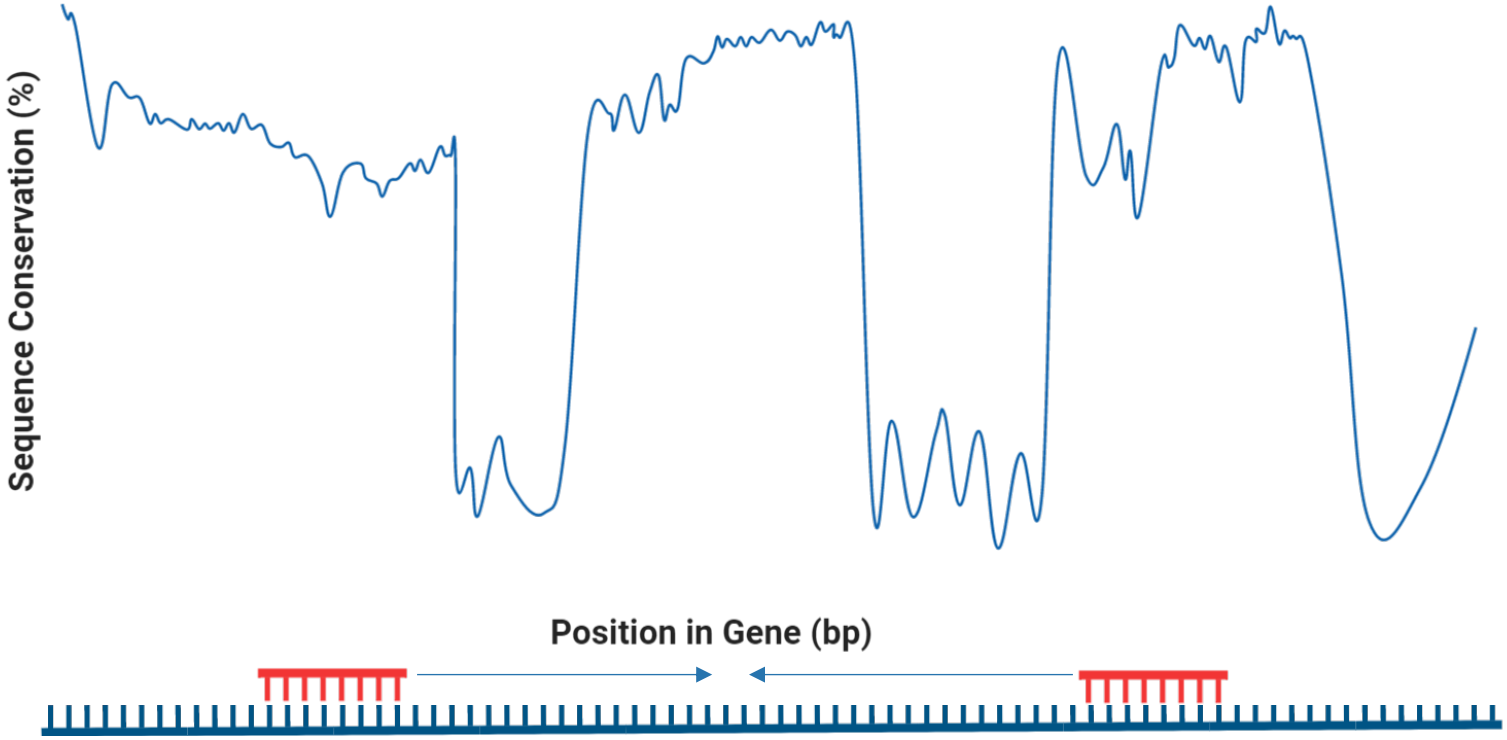
DNA Provides a Few Options



DNA Provides a Few Options



DNA Provides a Few Options



Ideal Target Properties

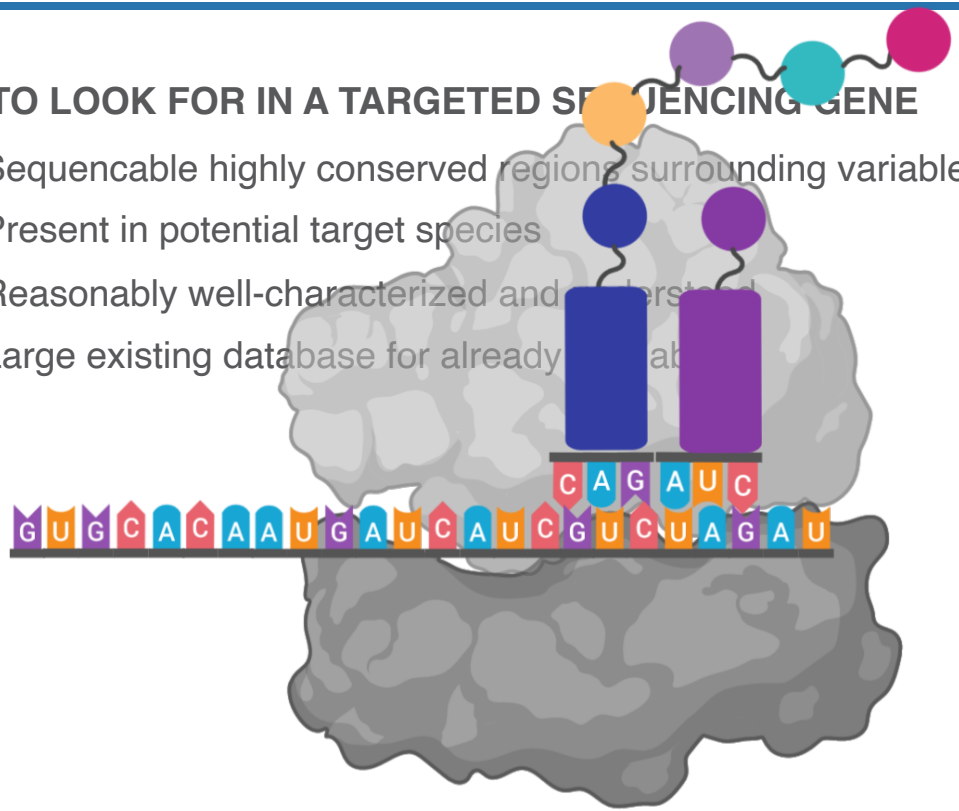
WHAT TO LOOK FOR IN A TARGETED SEQUENCING GENE

- Sequencable highly conserved regions surrounding variable regions
- Present in potential target species
- Reasonably well-characterized and understood
- Large existing database for already available

Ideal Target Properties

WHAT TO LOOK FOR IN A TARGETED SEQUENCING GENE

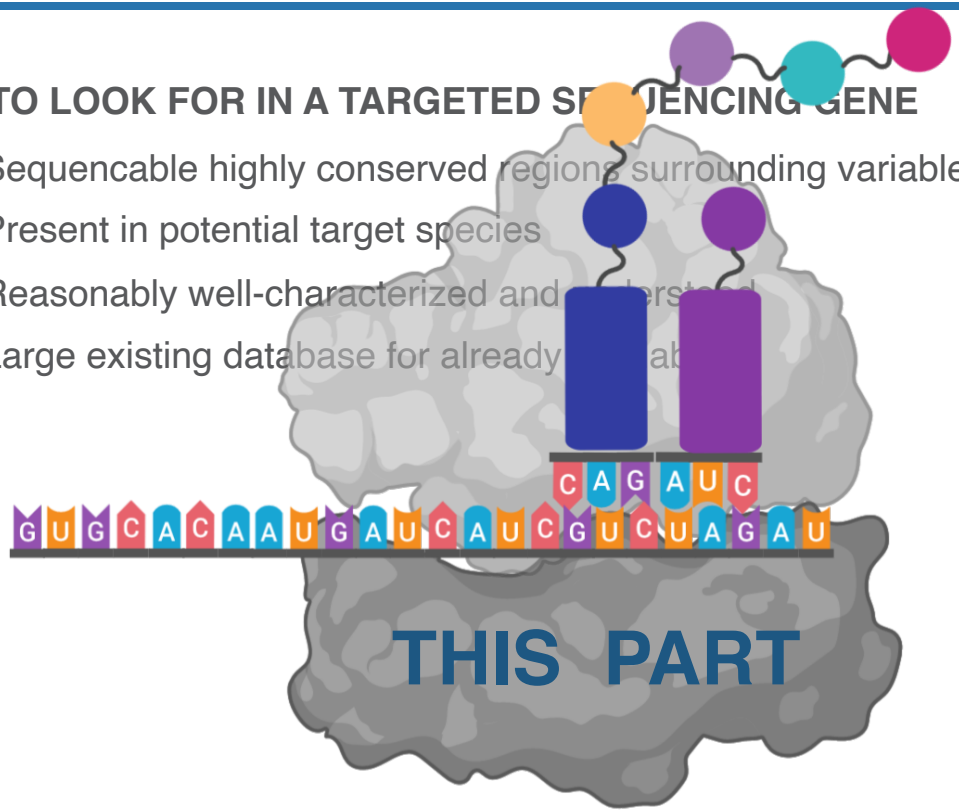
- Sequencable highly conserved regions surrounding variable regions
- Present in potential target species
- Reasonably well-characterized and understood
- Large existing database for already available



Ideal Target Properties

WHAT TO LOOK FOR IN A TARGETED SEQUENCING GENE

- Sequencable highly conserved regions surrounding variable regions
- Present in potential target species
- Reasonably well-characterized and understood
- Large existing database for already [illegible]



Topic

16S ANALYSIS

16S Sequencing Challenges

TARGETED SEQUENCE WITH A FEW BASES DIFFERENTIATING SPECIES

- Sequencing is imperfect
  - Illumina usually makes about 2% base call errors
  - Nanopore makes more
  - Errors are not necessarily evenly-distributed
- We do not want errors to be confused with real diversity/new species

The True Image

The Addition of Noise

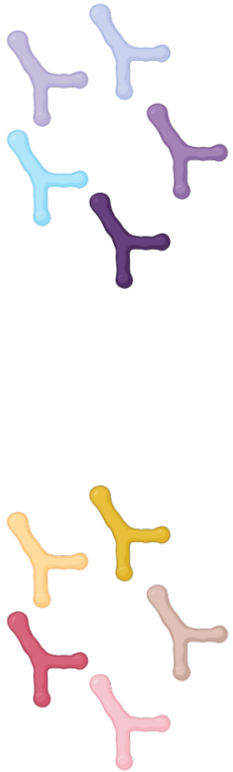
Blur: Loss of Detail and Noise

Blur *vs.* Denoise

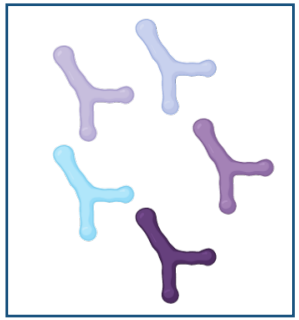
If the Noise Could Be Predicted

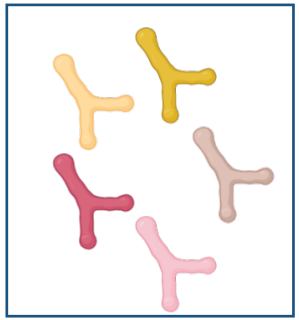
Could We Return to Something

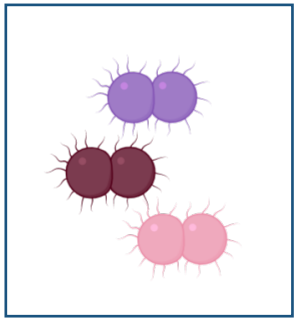
Clustering (Blurring)

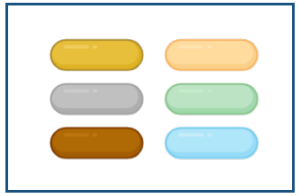


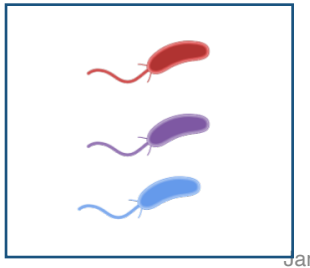
Clustering



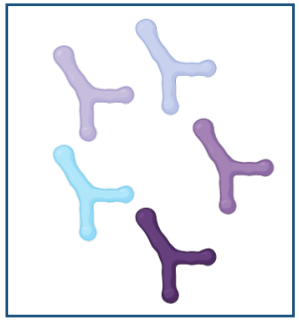


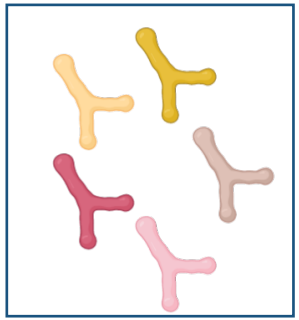


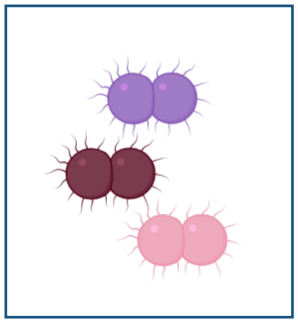


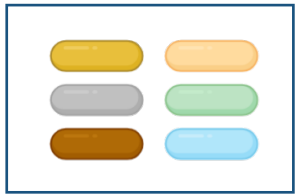


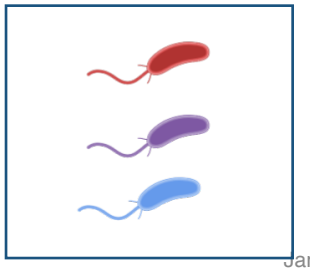
Clustering



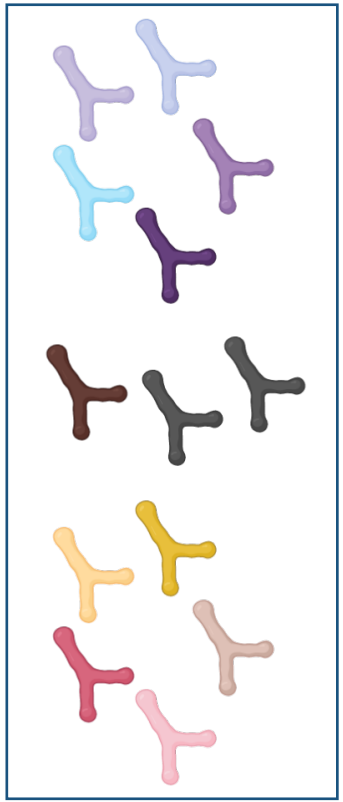




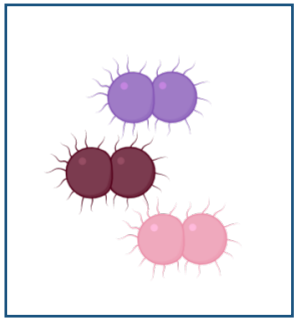


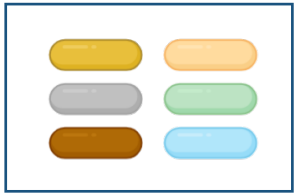


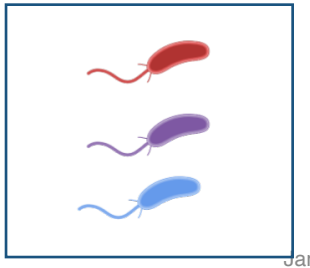
Clustering



Operational
Taxonomic
Units







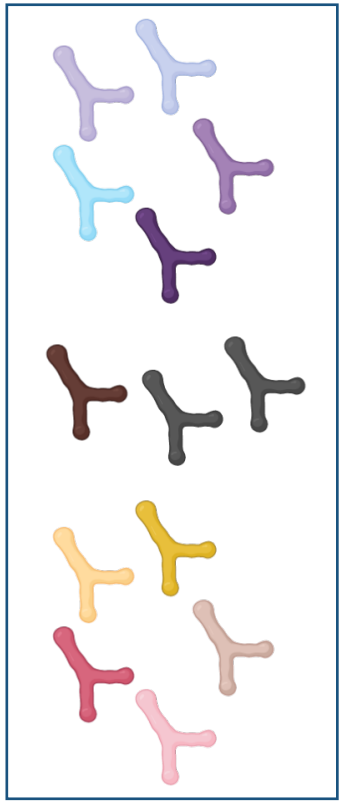
Operational Taxonomic Unit (OTU) Approach

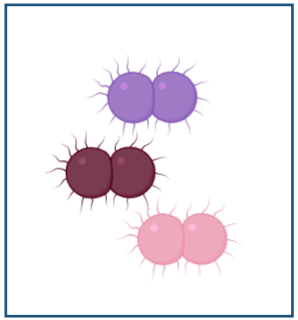
WE KNOW SOME OF THESE SEQUENCES AROSE FROM ERROR/ARTIFACT.

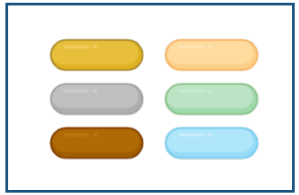
Combine extremely similar sequences (usually 97% or more identity) to minimize the effects of observed errors. Then treat each OTU as a representative sequence.

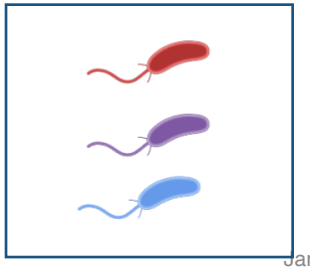
- Perform any necessary preprocessing (demultiplex, etc.)
- Pick OTUs
  - Fast an efficient if guided by a reference set, but limited to reference set.
  - Slower if using reference set as a non-restrictive guide
  - Very computationally expensive if going reference-free
- Extract representative sequence from each OTU
- Conduct further analysis with representative sequences that can represent one or more organisms from your sample

But…

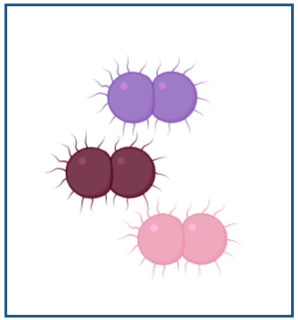




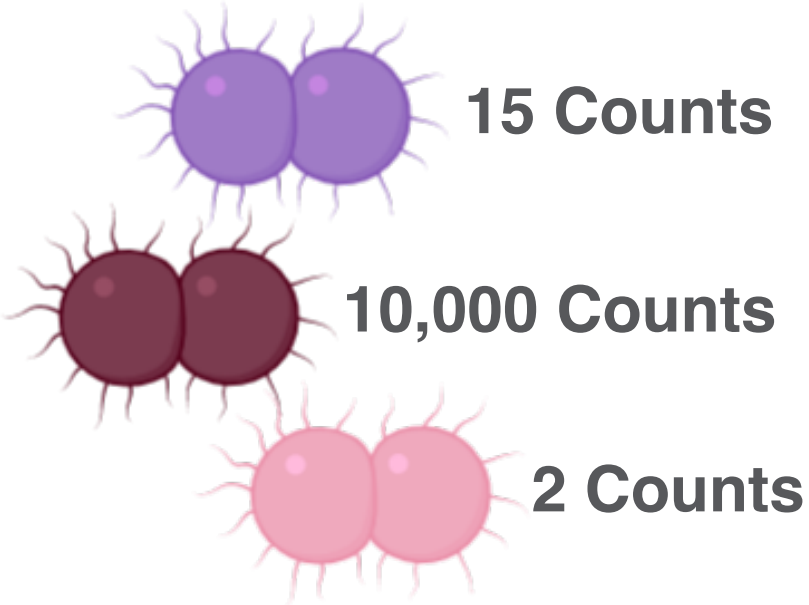




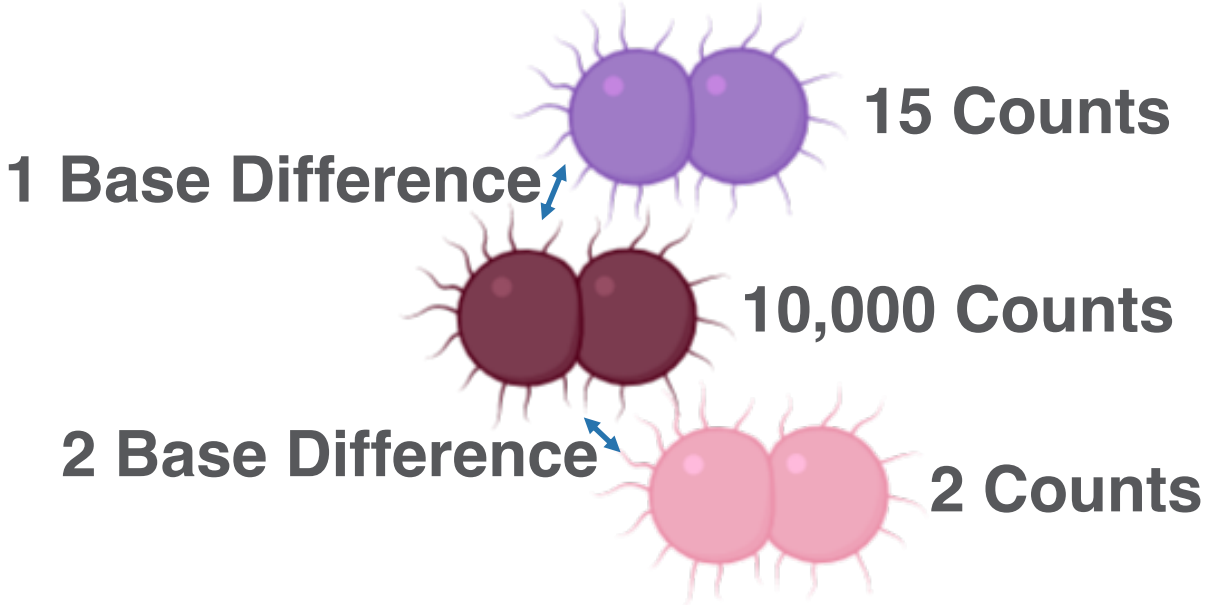
What if…



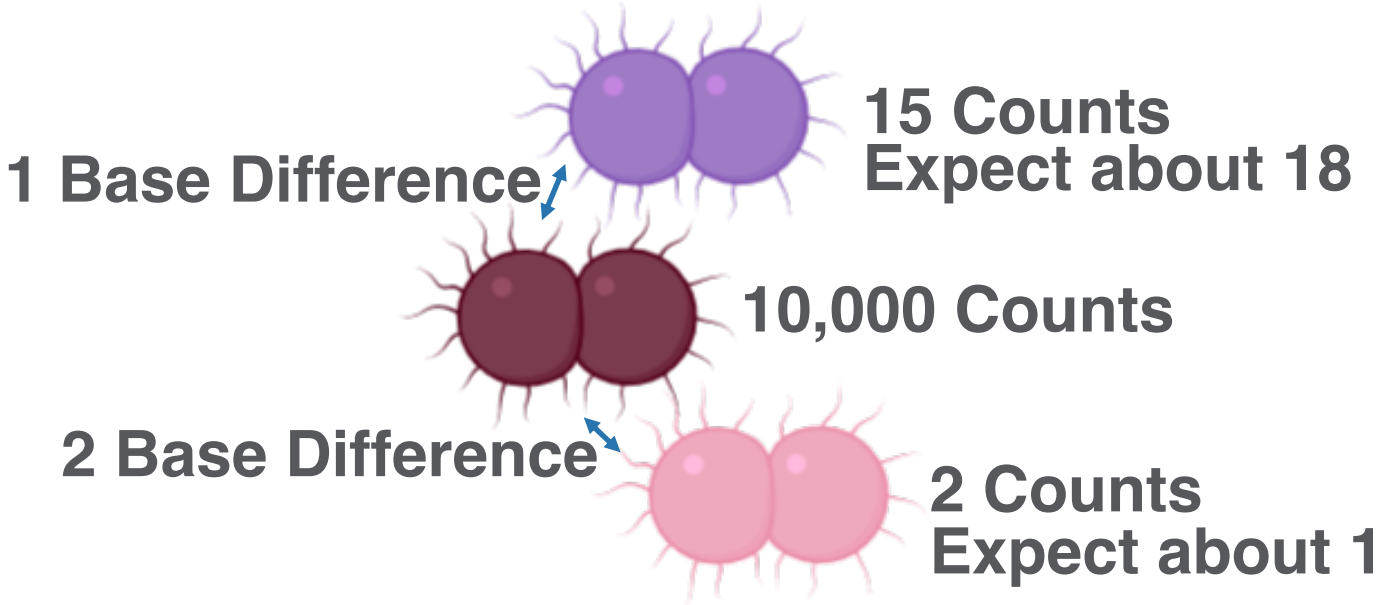
We Considered Frequency…



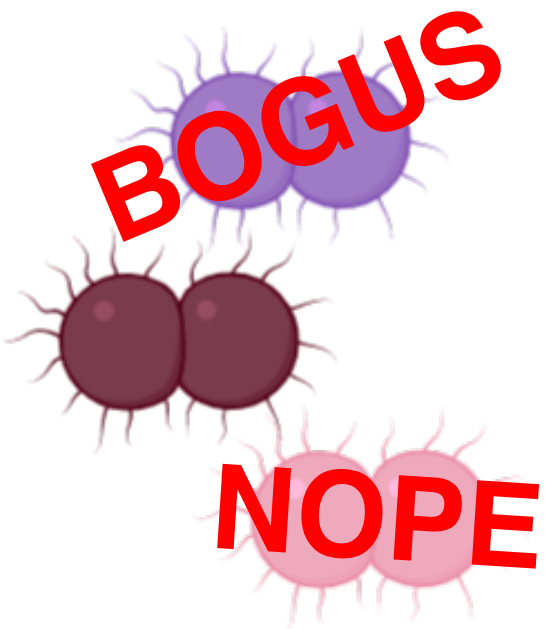
And Differences…



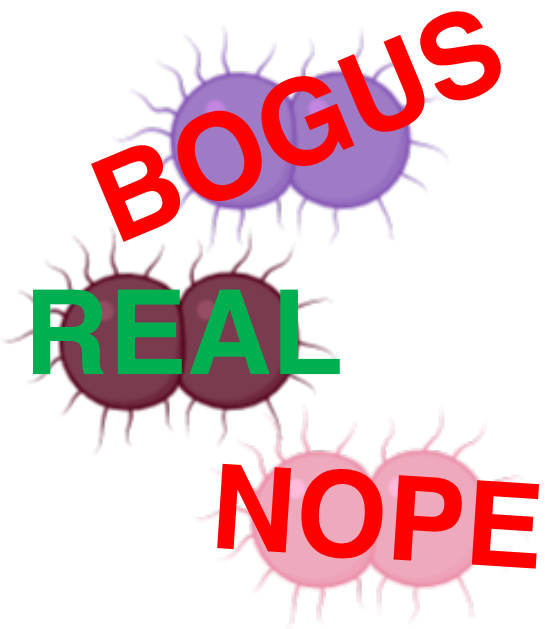
And an Error Model for the Run…



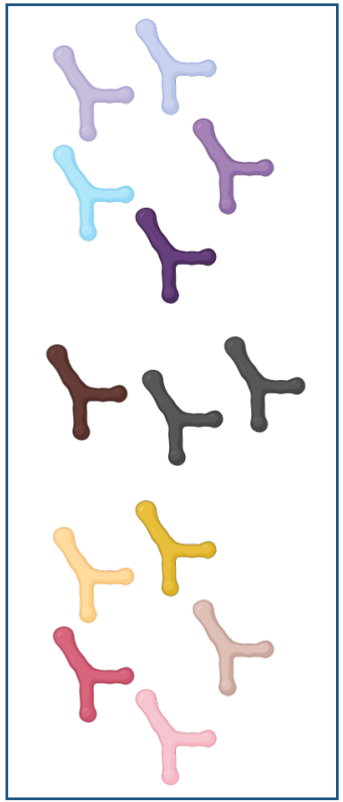
We Can Remove What Likely Is Not There…



And Be More Confident In What We Keep…



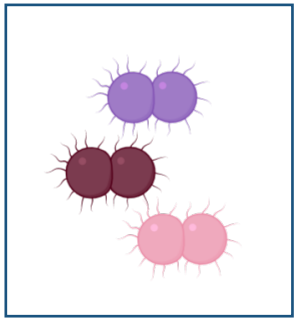
The Outcome Looks Like Our Clustering…

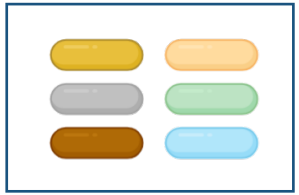


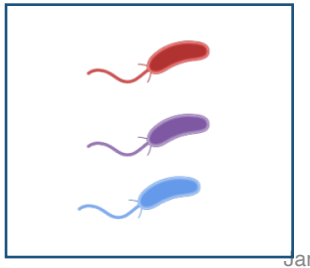
Operating

Taxonomic

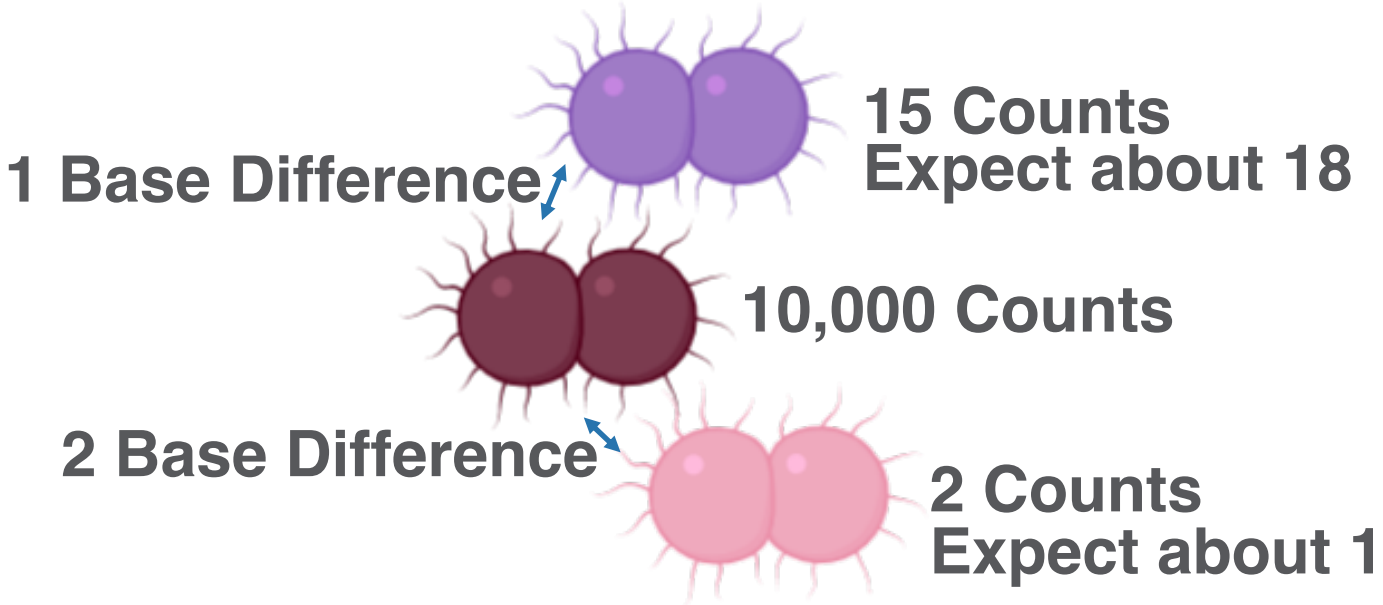
Units



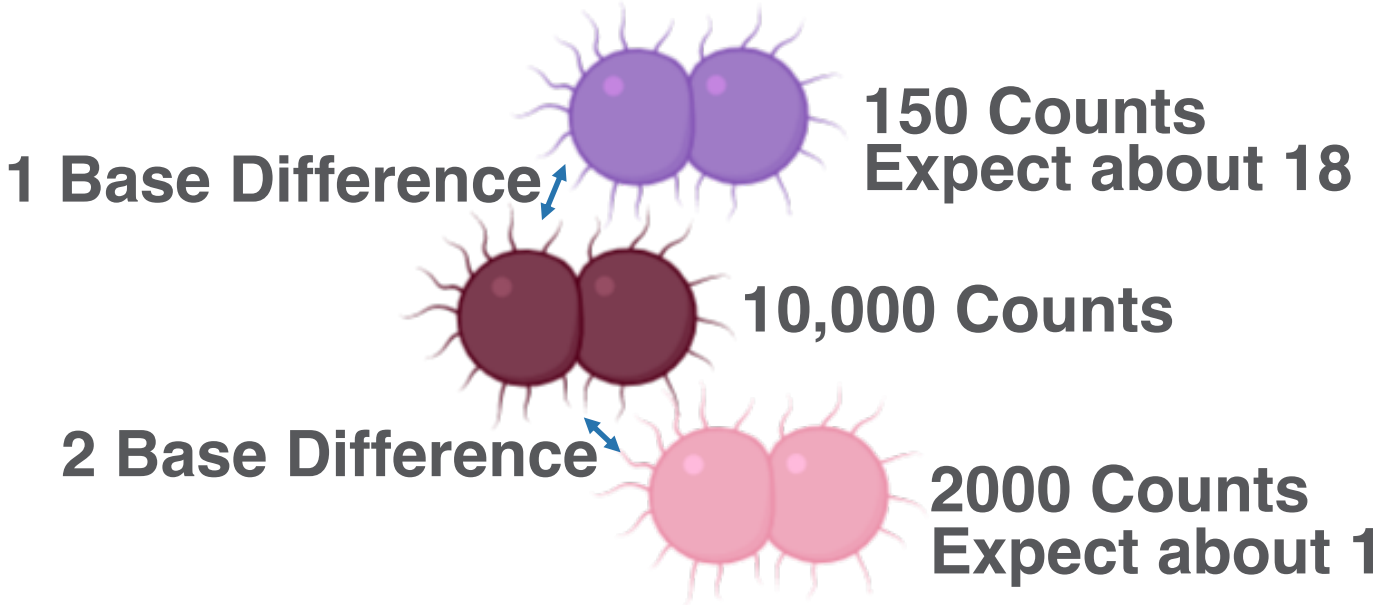




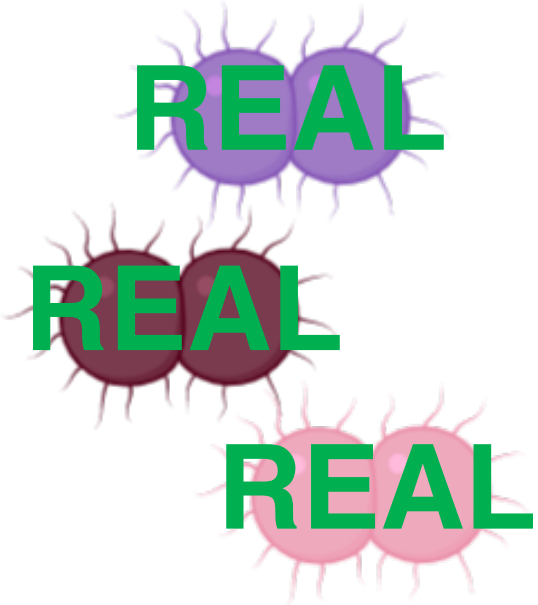
Except…


1 Base Difference
15 Counts
Expect about 18
10,000 Counts
2 Base Difference
2 Counts
Expect about 1

One OTU Can Hide Multiple Species…



And We Do Not Want To Miss Them.



Amplicon Sequence Variant (ASV) Approach

WHAT IS THE STATISTICAL SUPPORT FOR EACH SEQUENCE'S EXISTENCE?

Throw out amplicon sequences that lack strong statistical support for not being artifacts of sequencing. Cost: potential loss of real sequence that was present at very low levels.

- Preprocess reads (demultiplex, etc.)
- Trim and filter reads to retain highest possible quality sequence
- Build an error model for forward and reverse reads
- Merge forward and reverse reads into a single amplicon at an overlap region
- Count occurrences of each exact amplicon
- Use error model to determine support for rare amplicons based on the frequency
- Eliminate poorly-supported amplicons
- Conduct further analysis with exact amplicon sequences whose existence is robustly supported

Topic

GETTING OUR SAMPLE DATA

Getting a Started

```
Last login: Mon May  4 10:23:31 2020 from
Welcome to the Hoffman2 Cluster!

Hoffman2 Home Page:     http://www.hoffman2.idre.ucla.edu
Consulting:             https://support.idre.ucla.edu/helpdesk

All login nodes should be accessed via "hoffman2.idre.ucla.edu".

Please do NOT compute on the login nodes.

Processes running on the login nodes which seriously degrade others'
use of the system may be terminated without warning.  Use qrsh to obtain
an interactive shell on a compute node for CPU or I/O intensive tasks.

The following news items are currently posted:

   IDRE Workshops and Training Sessions
   News Archive On Web Site

Enter shownews to read the full text of a news item.
[mweinste@login4 ~]$ qrsh -l h_data=16G,h_rt=10:00:00
[mweinste@n7282 ~]$ cd $SCRATCH
```

Getting a Started

```
[mweinste@n7282 mweinste]$ cp /u/scratch/m/mweinste/ws11.tar.gz $SCRATCH

[mweinste@n7282 mweinste]$ tar -xvf ws11.tar.gz
ws11/
ws11/data/
ws11/data/zbStandard_R2.fastq.gz
ws11/data/zbStandard_R1.fastq.gz
ws11/data/fecal_R2.fastq.gz
ws11/data/fecal_R1.fastq.gz
ws11/data/gg_13_5_taxonomy.txt.gz
ws11/data/gg_13_5.fasta.gz
[mweinste@n7282 mweinste]$
```

Install BioConductor

```
[mweinste@n7361 ~]$ module load R/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

        Unloading the conflicting module 'intel/18.0.4'

[mweinste@n7361 ~]$ R

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("BiocManager")
Installing package into '/u/home/m/mweinste/R/x86_64-pc-linux-gnu-library/3.6'
(as 'lib' is unspecified)
```

Answer “yes” to questions about installation location if asked

Install DADA2

```
55: USA (MI 2) [https]
56: USA (OH) [https]
57: USA (OR) [https]
58: USA (TN) [https]
59: USA (TX 1) [https]
60: Uruguay [https]
61: (other mirrors)

Selection: 57
trying URL 'https://ftp.osuosl.org/pub/cran/src/contrib/BiocManager_1.30.10.tar.gz'
Content type 'application/x-gzip' length 40205 bytes (39 KB)
==================================================
downloaded 39 KB

* installing *source* package 'BiocManager' ...
** package 'BiocManager' successfully unpacked and MD5 sums checked
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (BiocManager)

The downloaded source packages are in
        '/work/tmp/RtmpqzTNfa/downloaded_packages'
> BiocManager::install("dada2", version = "3.10")
Bioconductor version 3.10 (BiocManager 1.30.10), R 3.6.1 (2019-07-05)
Installing package(s) 'BiocVersion', 'dada2'
also installing the dependencies 'ps', 'processx', 'callr', 'prettyunits', 'desc', '
```

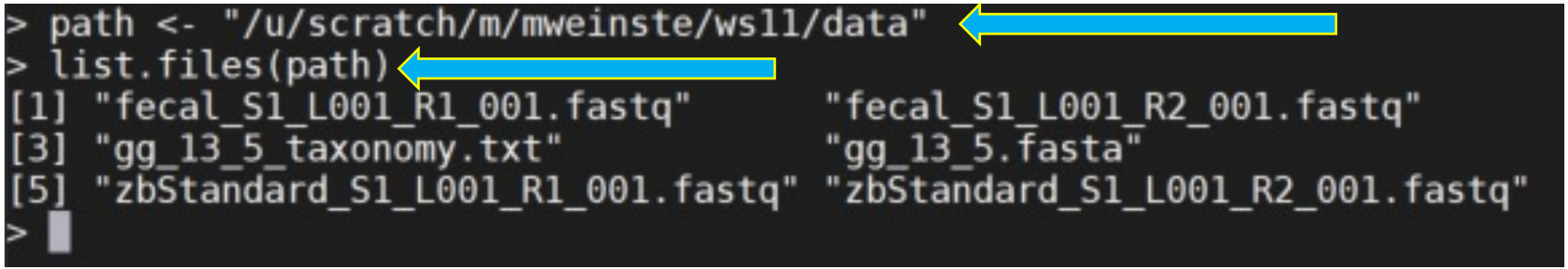
…it's gonna take a while

```
** R
** data
*** moving datasets to lazyload DB
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** checking absolute paths in shared objects and dynamic libraries
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (dada2)

The downloaded source packages are in
        '/work/tmp/RtmpqzTNfa/downloaded_packages'
Installation path not writeable, unable to update packages: backports, boot,
  class, cli, digest, glue, jsonlite, KernSmooth, lattice, MASS, Matrix, mgcv,
  nlme, nnet, pillar, Rcpp, repr, rlang, spatial, survival, uuid, vctrs
>
```

Set “path” to Where the Sequences Are

Use your own scratch folder here

```
> path <- "/u/scratch/m/mweinste/ws11/data"
> list.files(path)
[1] "fecal_S1_L001_R1_001.fastq"       "fecal_S1_L001_R2_001.fastq"
[3] "gg_13_5_taxonomy.txt"             "gg_13_5.fasta"
[5] "zbStandard_S1_L001_R1_001.fastq" "zbStandard_S1_L001_R2_001.fastq"
>
```


> path <- "/u/scratch/m/mweinste/ws11/data"
> list.files(path)
[1] "fecal_S1_L001_R1_001.fastq" "fecal_S1_L001_R2_001.fastq"
[3] "gg_13_5_taxonomy.txt" "gg_13_5.fasta"
[5] "zbStandard_S1_L001_R1_001.fastq" "zbStandard_S1_L001_R2_001.fastq"
>

Have R Find and Sort Your Sequencing Files

```
> fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
> fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))
> fnFs
[1] "/u/scratch/m/mweinste/ws11/data/fecal_S1_L001_R1_001.fastq"
[2] "/u/scratch/m/mweinste/ws11/data/zbStandard_S1_L001_R1_001.fastq"
> fnRs
[1] "/u/scratch/m/mweinste/ws11/data/fecal_S1_L001_R2_001.fastq"
[2] "/u/scratch/m/mweinste/ws11/data/zbStandard_S1_L001_R2_001.fastq"
>
```

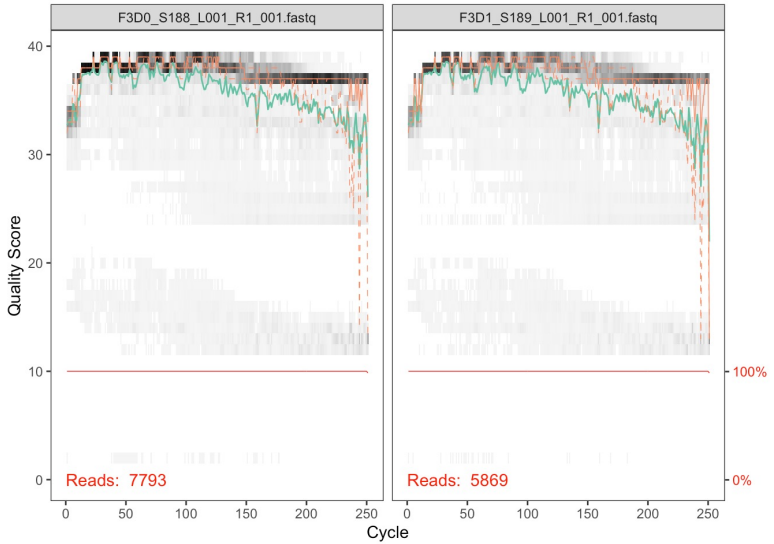
The “Official” Method

Inspect read quality profiles

We start by visualizing the quality profiles of the forward reads:

```
plotQualityProfile(fnFs[1:2])
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```



Open a New Terminal Tab/Window

```
The following news items are currently posted:

   IDRE Workshops and Training Sessions
   News Archive On Web Site

Enter shownews to read the full text of a news item.
[mweinste@login3 ~]$ qrsh -l h_data=16G,h_rt=10:00:00
Last login: Mon May 11 21:59:41 2020 from login4
[mweinste@n7282 ~]$ cd $SCRATCH
[mweinste@n7282 mweinste]$ cd ws11
[mweinste@n7282 ws11]$ ls
data  figaro
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24 --inputDirectory data/
```

Amplicon length and primer lengths will be determined by your method of library prep and target region

Open a New Terminal Tab/Window

```
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24
{"trimPosition": [305, 225], "maxExpectedError": [5, 4], "readRetentionPercent": 76.9, "score": 51.896132561338405}
{"trimPosition": [304, 226], "maxExpectedError": [5, 4], "readRetentionPercent": 76.28, "score": 51.28354819103676}
{"trimPosition": [303, 227], "maxExpectedError": [5, 4], "readRetentionPercent": 75.9, "score": 50.90448162246889}
{"trimPosition": [302, 228], "maxExpectedError": [5, 5], "readRetentionPercent": 80.83, "score": 48.8339464508493}
{"trimPosition": [301, 229], "maxExpectedError": [5, 5], "readRetentionPercent": 80.68, "score": 48.6835993730207}
{"trimPosition": [300, 230], "maxExpectedError": [5, 5], "readRetentionPercent": 80.6, "score": 48.59563033812098}
{"trimPosition": [299, 231], "maxExpectedError": [5, 5], "readRetentionPercent": 80.52, "score": 48.52045679920667}
{"trimPosition": [298, 232], "maxExpectedError": [5, 5], "readRetentionPercent": 80.3, "score": 48.30133393045648}
{"trimPosition": [297, 233], "maxExpectedError": [5, 5], "readRetentionPercent": 80.25, "score": 48.25335082051119}
```

You may have to scroll back up a bit to find the first set of values.

Return to Your Terminal Running R

```
> sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
> sample.names
[1] "fecal"      "zbStandard"
>
```

Prepare File Names For Filtered Reads

```
> sample.names
[1] "fecal"       "zbStandard"
> filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
> filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
> filtFs
[1] "/u/scratch/m/mweinste/ws11/data/filtered/fecal_F_filt.fastq.gz"
[2] "/u/scratch/m/mweinste/ws11/data/filtered/zbStandard_F_filt.fastq.gz"
> filtRs
[1] "/u/scratch/m/mweinste/ws11/data/filtered/fecal_R_filt.fastq.gz"
[2] "/u/scratch/m/mweinste/ws11/data/filtered/zbStandard_R_filt.fastq.gz"
>
```

More File Name Management

```
> names(filtFs) <- sample.names
> names(filtRs) <- sample.names
> names
function (x)  .Primitive("names")
```

Trim and Filter

```
> out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(305,225), maxN=0, maxEE=c(5,4), truncQ=2, rm.phix=TRUE, compress=TRUE, multithread=FALSE)
> out
                                reads.in reads.out
fecal_S1_L001_R1_001.fastq         62708     48592
zbStandard_S1_L001_R1_001.fastq    62335     47014
>
```

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(305,225), maxN=0, maxEE=c(5,4), truncQ=2 rm.phix=TRUE, compress=TRUE, multithread=FALSE)

```
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24
{"trimPosition": [305, 225], "maxExpectedError": [5, 4], "readRetentionPercent": 76.9, "score": 51.896132561338405}
{"trimPosition": [304, 226], "maxExpectedError": [5, 4], "readRetentionPercent": 76.28, "score": 51.28354819103676}
{"trimPosition": [303, 227], "maxExpectedError": [5, 4], "readRetentionPercent": 75.9, "score": 50.90448162246889}
{"trimPosition": [302, 228], "maxExpectedError": [5, 5], "readRetentionPercent": 80.83, "score": 48.8339464508493}
{"trimPosition": [301, 229], "maxExpectedError": [5, 5], "readRetentionPercent": 80.68, "score": 48.6835993730207}
{"trimPosition": [300, 230], "maxExpectedError": [5, 5], "readRetentionPercent": 80.6, "score": 48.59563033812098}
{"trimPosition": [299, 231], "maxExpectedError": [5, 5], "readRetentionPercent": 80.52, "score": 48.52045679920667}
{"trimPosition": [298, 232], "maxExpectedError": [5, 5], "readRetentionPercent": 80.3, "score": 48.30133393045648}
{"trimPosition": [297, 233], "maxExpectedError": [5, 5], "readRetentionPercent": 80.25, "score": 48.25335082051119}
```

Learn Forward Errors

```
> errF <- learnErrors(filtFs, multithread=TRUE)
29159830 total bases in 95606 reads from 2 samples will be used for learning the error rates.
> errF
$err_out
            0          1          2          3          4          5
A2A 0.61495589 0.61495589 0.61495589 0.61495589 0.61495589 0.61495589
A2C 0.22121132 0.22121132 0.22121132 0.22121132 0.22121132 0.22121132
A2G 0.13953688 0.13953688 0.13953688 0.13953688 0.13953688 0.13953688
A2T 0.02429590 0.02429590 0.02429590 0.02429590 0.02429590 0.02429590
C2A 0.17850849 0.17850849 0.17850849 0.17850849 0.17850849 0.17850849
C2C 0.73241539 0.73241539 0.73241539 0.73241539 0.73241539 0.73241539
```

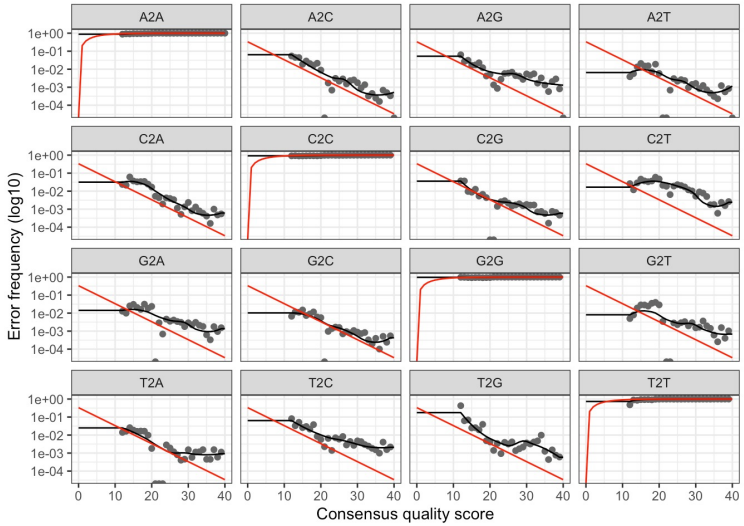
Learn Reverse Errors

```
> errR <- learnErrors(filtRs, multithread=TRUE)
21511350 total bases in 95606 reads from 2 samples will be used for learning the error rates.
>
```

Visualizing the Error Models

It is always worthwhile, as a sanity check if nothing else, to visualize the estimated error rates:

```
plotErrors(errF, nominalQ=TRUE)
```



The error rates for each possible transition (A→C, A→G, …) are shown. Points are the observed error rates for each consensus quality score. The black line shows the estimated error rates after convergence of the machine-learning algorithm. The red line shows the error rates expected under the nominal definition of the Q-score. Here the estimated error rates (black line) are a good fit to the observed rates (points), and the error rates drop with increased quality as expected. Everything looks reasonable and we proceed with confidence.

Apply the Denoising

```
> dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
Sample 1 - 48592 reads in 29560 unique sequences.
Sample 2 - 47014 reads in 30901 unique sequences.
> dadaRs <- dada(filtRs, err=errR, multithread=TRUE)
Sample 1 - 48592 reads in 40879 unique sequences.
Sample 2 - 47014 reads in 41538 unique sequences.
>
```

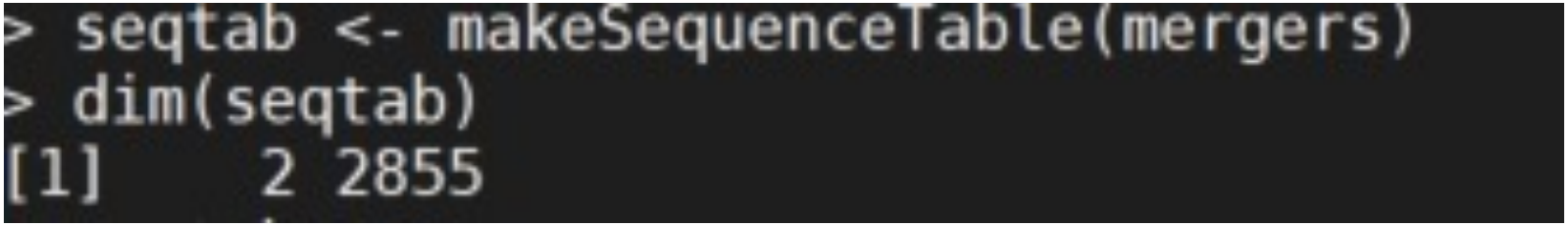
And See How We Did

```
> dadaFs
$fecal
dada-class: object describing DADA2 denoising results
414 sequence variants were inferred from 29560 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16

$zbStandard
dada-class: object describing DADA2 denoising results
138 sequence variants were inferred from 30901 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16

> dadaRs
$fecal
dada-class: object describing DADA2 denoising results
131 sequence variants were inferred from 40879 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16

$zbStandard
dada-class: object describing DADA2 denoising results
59 sequence variants were inferred from 41538 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16

>
```

Attempt to Merge Read Pairs

```
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
44743 paired-reads (in 1800 unique pairings) successfully merged out of 46562 (in 2295 pairings) input
.
45668 paired-reads (in 1055 unique pairings) successfully merged out of 46732 (in 1735 pairings) input
```

And See How We Did


> seqtab <- makeSequenceTable(mergers)
> dim(seqtab)
[1]   2 2855

Check Sequence Length Distribution

```
> table(nchar(getSequences(seqtab)))

 439  440  441  442  443  459  460  464  465  466
   2  461   10    7   16   12 1094  166  881  206
>
```

Sequences should be close to the expected amplicon size minus the forward and reverse primers.

Anything that is extremely long or extremely short is potentially non-specific priming garbage and should be inspected for deletion.

Remove Chimeras

```
> seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRUE)
Identified 2341 bimeras out of 2855 input sequences.
> dim(seqtab.nochim)
[1]   2 514
> sum(seqtab.nochim)/sum(seqtab)
[1] 0.4627534
> table(nchar(getSequences(seqtab.nochim)))

440 441 442 443 459 460 464 465 466
 90   5   4   7   5 125  28 199  51
>
```

Track Our Reads

```
> getN <- function(x) sum(getUniques(x))
> track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.nochim))
> colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
> rownames(track) <- sample.names
> head(track)
           input filtered denoisedF denoisedR merged nonchim
fecal      62708    48592     47219     47742  44743   20033
zbStandard 62335    47014     46799     46932  45668   21805
```

Assign Taxa

```
> taxa <- assignTaxonomy(seqtab.nochim, "/u/scratch/m/mweinste/ws11/data/silva_nr_v138_train_set.fa.gz", multithread=TRUE)
> taxa <- addSpecies(taxa, "/u/scratch/m/mweinste/ws11/data/silva_species_assignment_v138.fa.gz")
> taxa.print <- taxa
```

Examine Taxa

```
> taxa.print <- taxa
> rownames(taxa.print) <- NULL
> head(taxa.print)
     Kingdom    Phylum           Class                Order
[1,] "Bacteria" "Firmicutes"     "Clostridia"         "Lachnospirales"
[2,] "Bacteria" "Firmicutes"     "Clostridia"         "Lachnospirales"
[3,] "Bacteria" "Firmicutes"     "Clostridia"         "Lachnospirales"
[4,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Pseudomonadales"
[5,] "Bacteria" "Firmicutes"     "Clostridia"         "Lachnospirales"
[6,] "Bacteria" "Firmicutes"     "Clostridia"         "Lachnospirales"
     Family             Genus              Species
[1,] "Lachnospiraceae"  "CAG-56"           NA
[2,] "Lachnospiraceae"  "Anaerostipes"     NA
[3,] "Lachnospiraceae"  "Blautia"          NA
[4,] "Pseudomonadaceae" "Pseudomonas"      NA
[5,] "Lachnospiraceae"  "Fusicatenibacter" NA
[6,] "Lachnospiraceae"  "Dorea"            NA
> write.table(taxa, "taxaResults.txt", sep = "\t", row.names = TRUE, col.names = TRUE)
```

Taxa table is now in your ws11 folder