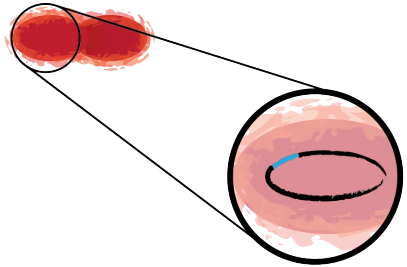

NFDI4
MICROBIOTA
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
jlAB
JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN
From raw data to feature tables

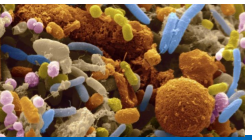
	Tuesday, Oct 8th	Wednesday, Oct 9th	Thursday, Oct 10th
10:00-10:30		QIIME2 and Qiita	Buffer for Hands-On, Coffee break in-between
10:30-11:00		Qiita Hands-On	
11:00-11:30	Welcome & Intro		
11:30-12:00	From raw data to feature tables	Coffee Break (~15min)	Coffee Break (~15min)
12:00-12:30		Qiita Hands-On	Buffer for Hands-On
12:30-13:00		Lunch Break	Closing remarks and Farewell
13:00-13:30	Lunch Break		
13:30-14:00		QIIME2 Hands-On	
14:00-14:30	Sequence quality control		
14:30-15:00	Coffee Break	Coffee Break	
15:00-15:30	QC Hands-on	Diversity Calculation	
15:30-16:00		QIIME2 Hands-On (cont'd.)	
16:00-16:30			

Microbial Community Analysis

https://www.fiosgenomics.com/microbiome-vs-microbiota/

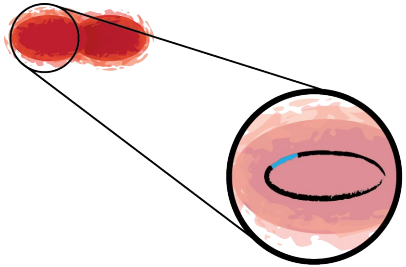
Profile community: meta genome

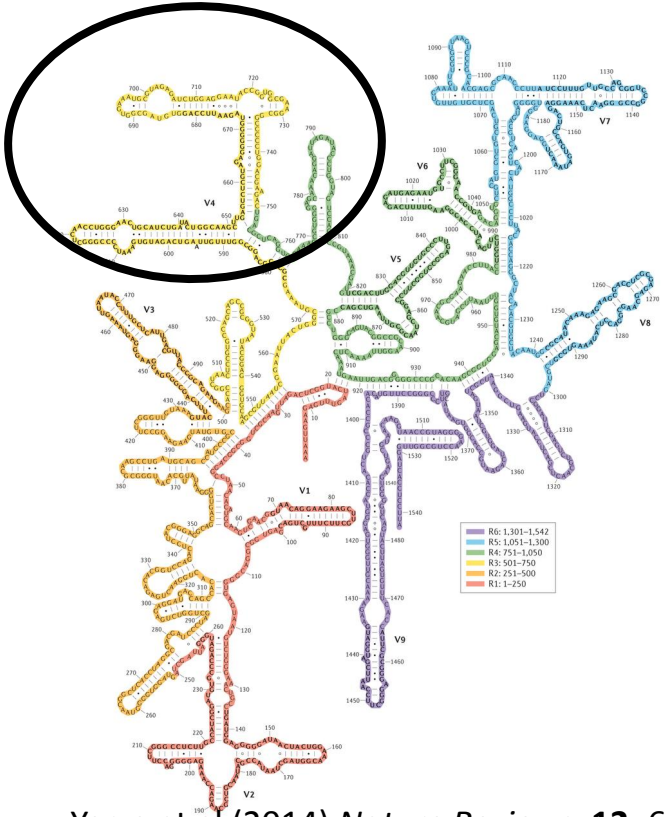




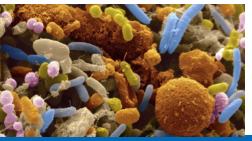
Yarza et al (2014) *Nature Reviews*. **12**: 635-645. Fig 1

Profile community: amplicon

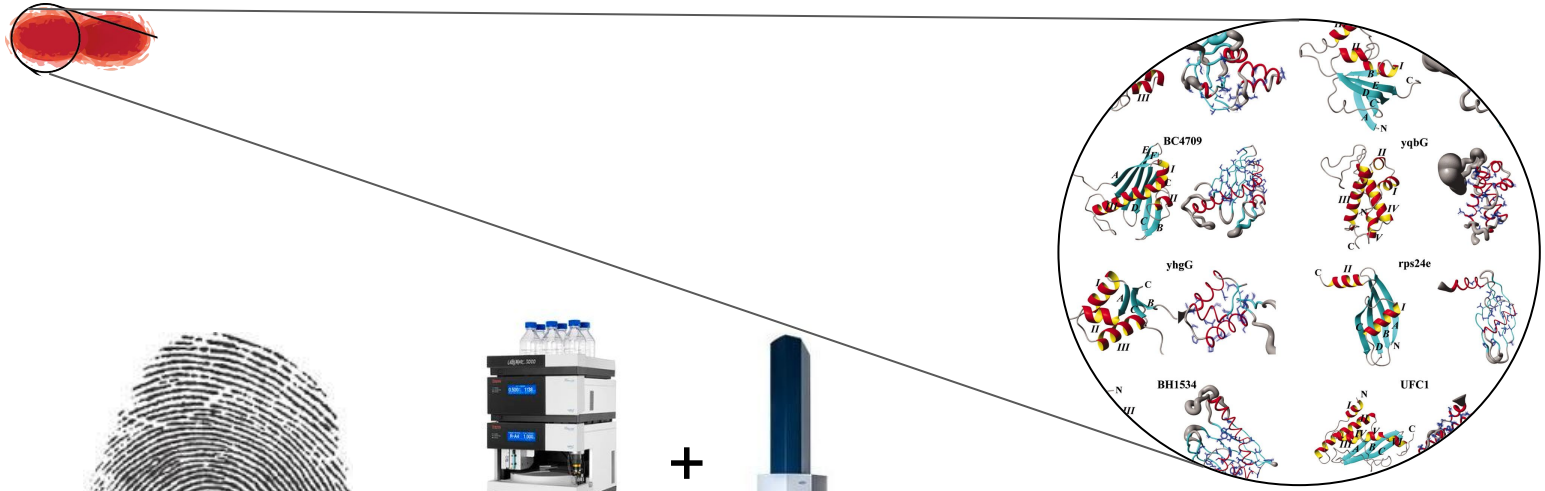




Yarza et al (2014) *Nature Reviews*. **12**: 635-645. Fig 1



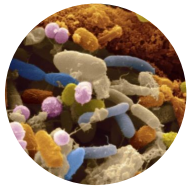
Profile community: metabolome



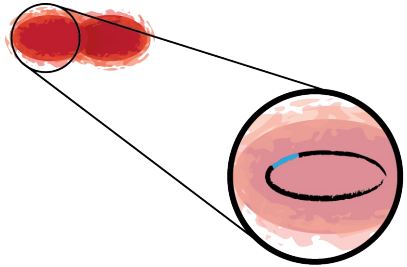
ThermoFisher UltiMate 3000 UPLC

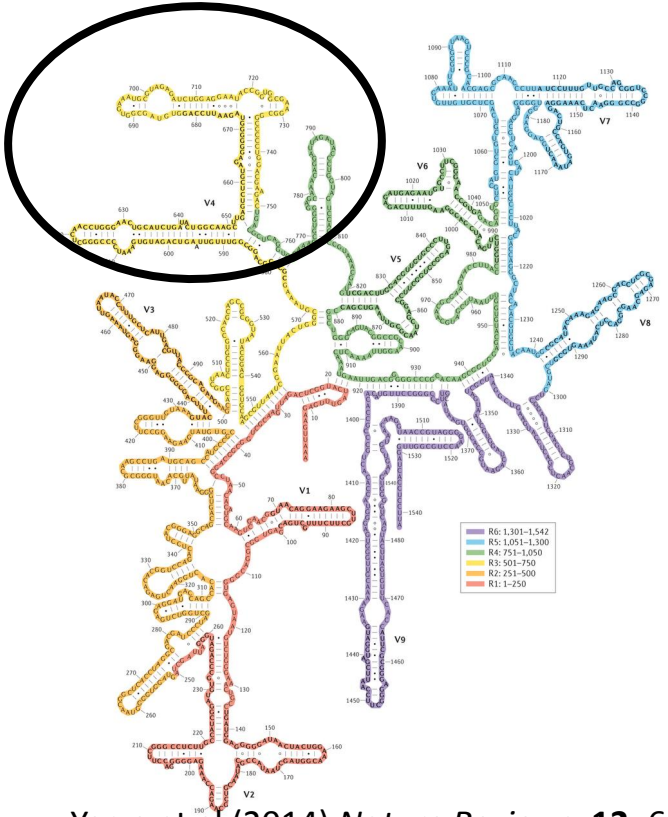
+

Bruker Maxis Q-TOF

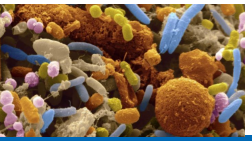


Profile community: amplicon

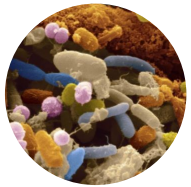




Yarza et al (2014) *Nature Reviews*. **12**: 635-645. Fig 1



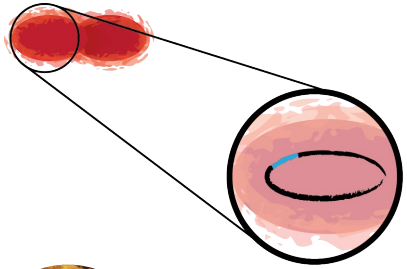
the "Feature" Table

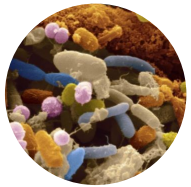


featuretable.xlsx - Excel

	A	B	C	D	E	F	G
1		Sample A	Sample B	Sample C	Sample C	Sample D	Sample E
2	Feature 1	553	236	10	179	465	587
3	Feature 2	442	282	14	496	243	834
4	Feature 3	714	130	13	744	231	436
5	Feature 4	46	19	6	25	203	239
6	Feature 5	876	421	0	376	497	87
7	Feature 6	955	131	12	290	984	705
8	Feature 7	639	232	24	646	398	858
9	Feature 8	280	276	8	559	511	860
10	Feature 9	607	360	3	107	410	341
11	Feature 10	970	141	20	56	823	342
12							
13							
14							
15							

the MAG Table

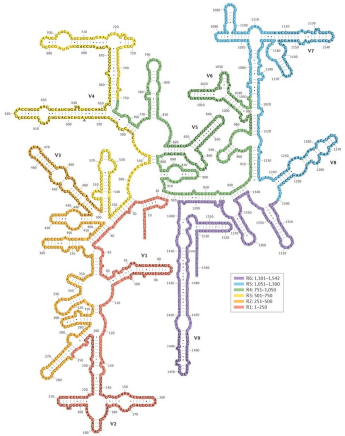


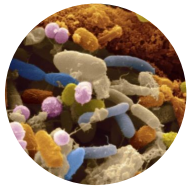


featuretable.xlsx - Excel

	A	B	C	D	E	F	G
1		Sample A	Sample B	Sample C	Sample C	Sample D	Sample E
2	iMGMC-MAG-1171	832	173	9	190	237	775
3	iMGMC-MAG-1164	446	460	4	25	225	72
4	iMGMC-MAG-1109	457	118	16	45	898	796
5	iMGMC-MAG-1072	434	307	9	443	388	224
6	iMGMC-MAG-1060	754	306	20	586	676	813
7	iMGMC-MAG-1058	414	179	26	573	763	31
8	iMGMC-MAG-1053	678	26	26	396	179	373
9	iMGMC-MAG-1052	967	323	23	465	782	614
10	iMGMC-MAG-1047	757	20	23	70	736	324
11	iMGMC-MAG-1035	838	43	6	774	174	818
12							
13							
14							
15							

the OTU Table

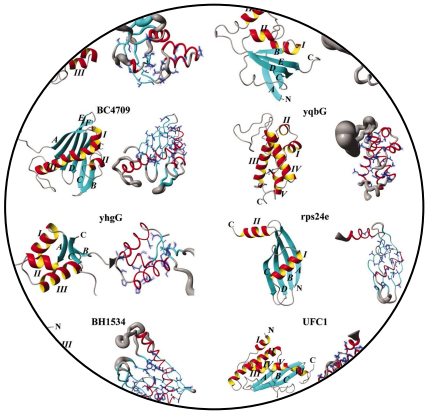


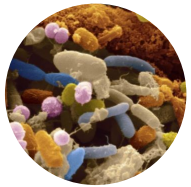


featuretable.xlsx - Excel

	A	B	C	D	E	F	G
1		Sample A	Sample B	Sample C	Sample C	Sample D	Sample E
2	OTU 1	774	65	10	179	50	139
3	OTU 2	453	119	12	389	488	840
4	OTU 3	48	416	2	54	69	54
5	OTU 4	674	416	3	696	422	919
6	OTU 5	910	320	27	50	947	131
7	OTU 6	738	490	23	58	791	849
8	OTU 7	15	83	23	665	249	319
9	OTU 8	369	141	12	632	632	53
10	OTU 9	113	417	15	540	92	232
11	OTU 10	176	443	7	614	82	688
12							
13							
14							
15							

the Proteome Table

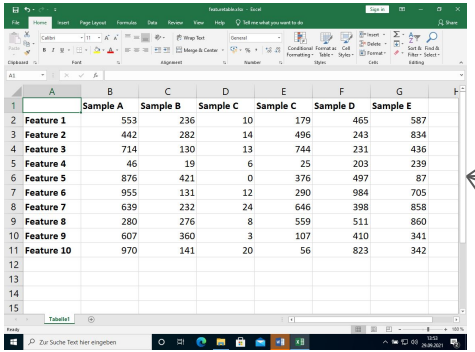




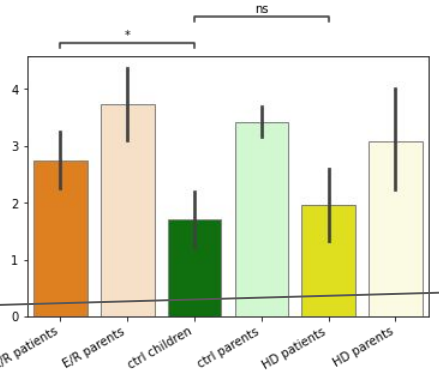
featuretable.xlsx - Excel

	A	B	C	D	E	F	G
1		Sample A	Sample B	Sample C	Sample C	Sample D	Sample E
2	226.951_595	429	498	6	417	454	67
3	278.191_249	246	438	20	21	157	634
4	307.112_169	569	112	12	100	140	829
5	365.136_378	858	337	13	184	993	837
6	337.105_334	104	174	8	641	535	693
7	205.097_85.3	93	425	4	66	545	222
8	343.154_378	831	402	17	526	942	431
9	637.968_563	436	123	7	9	570	200
10	360.181_378	772	189	22	166	331	869
11	666.989_576	581	18	18	732	256	744
12							
13							
14							
15							

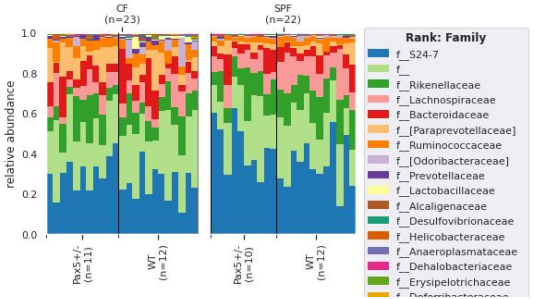
the "Feature" Table



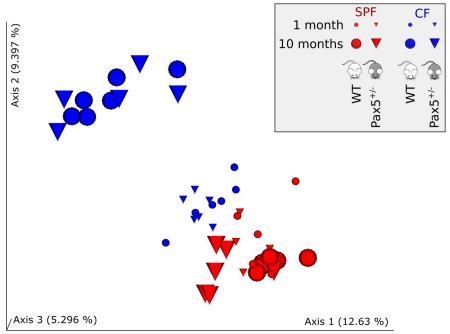
	Sample A	Sample B	Sample C	Sample C	Sample D	Sample E
Feature 1	553	236	10	179	465	587
Feature 2	442	282	14	496	243	834
Feature 3	714	130	13	744	231	436
Feature 4	46	19	6	25	203	239
Feature 5	876	421	0	376	497	87
Feature 6	955	131	12	290	984	705
Feature 7	639	232	24	646	398	858
Feature 8	280	276	8	559	511	860
Feature 9	607	360	3	107	410	341
Feature 10	970	141	20	56	823	342



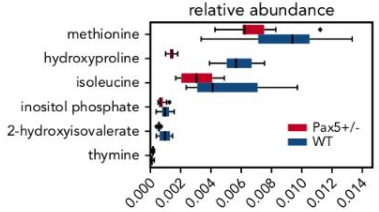
differences in diversity



feature barplot

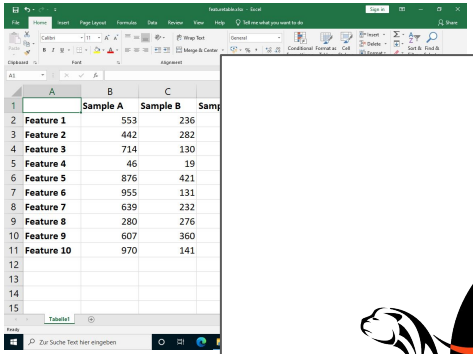


dimensionality reduction

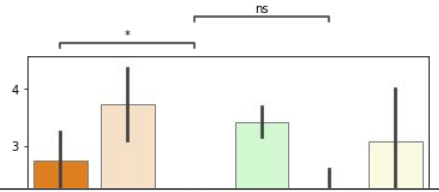


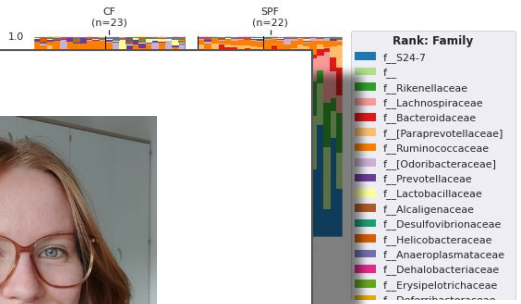
differential abundance

the "Feature" Table



	Sample A	Sample B	Sample C
Feature 1	553	236	
Feature 2	442	282	
Feature 3	714	130	
Feature 4	46	19	
Feature 5	876	421	
Feature 6	955	131	
Feature 7	639	232	
Feature 8	280	276	
Feature 9	607	360	
Feature 10	970	141	



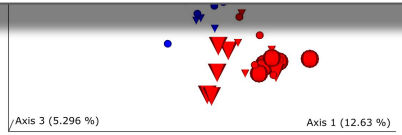


rplot

Qiita

https://qiita.ucsd.edu

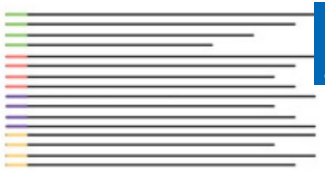
Anna Rehm



dimensionality reduction

differential abundance

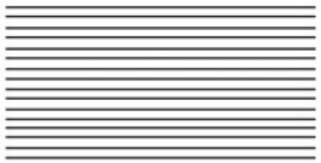
Amplify each sample, introducing barcode into each sequence using tagged PCR primers



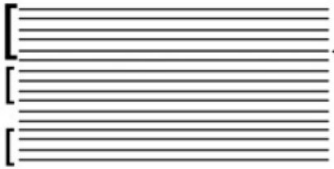
Use barcodes to assign each sequence to the sample it came from, dropping low-quality reads

Quantitative Insights Into Microbial Ecology

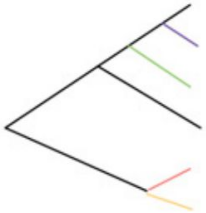
qiime2



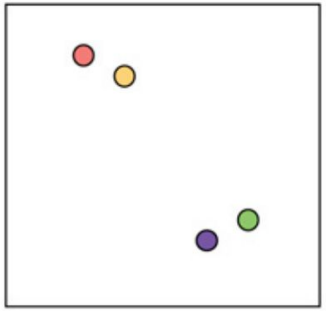
Trim barcodes and build multiple sequence alignment based on reference sequences



Group related sequences into OTUs for downstream analyses



Build phylogenetic tree using one representative of each OTU: track which parts of the tree came from which sample



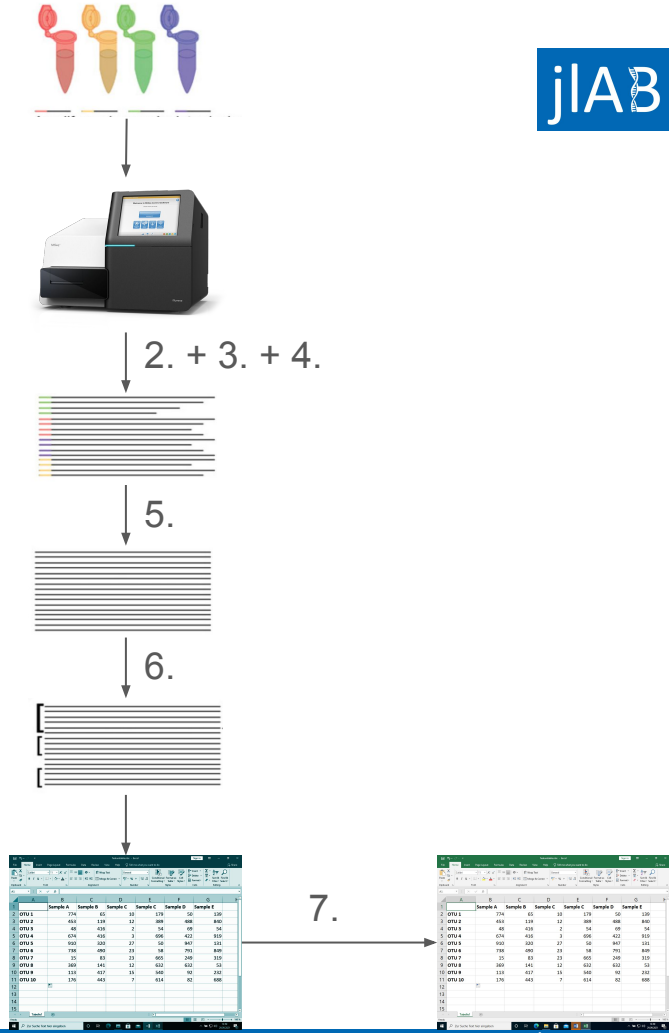
Use community clustering techniques (either OTU-based or tree-based) to relate samples to one another

📄 Bolyen et al. (2019) *Nature biotechnology*

Content

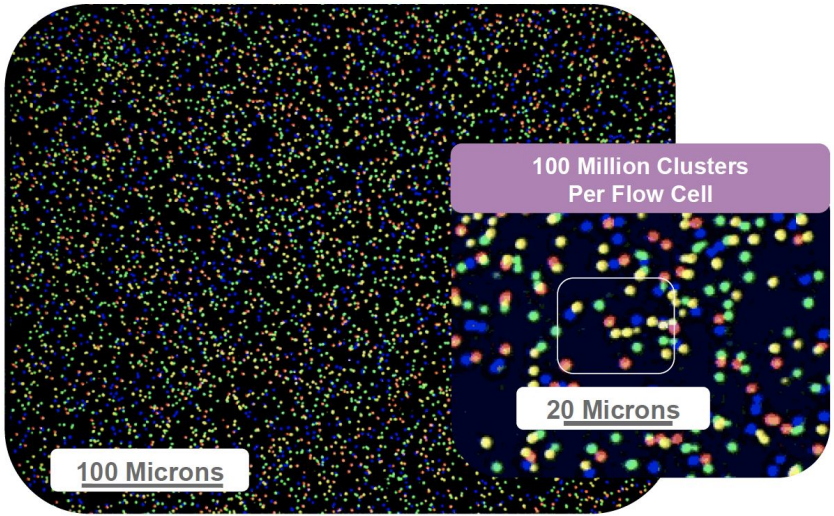
1. Scope ✔
2. Base Calling
3. Demultiplexing
4. Quality Control
5. Adapter Trimming, Clipping
6. "OTU" picking

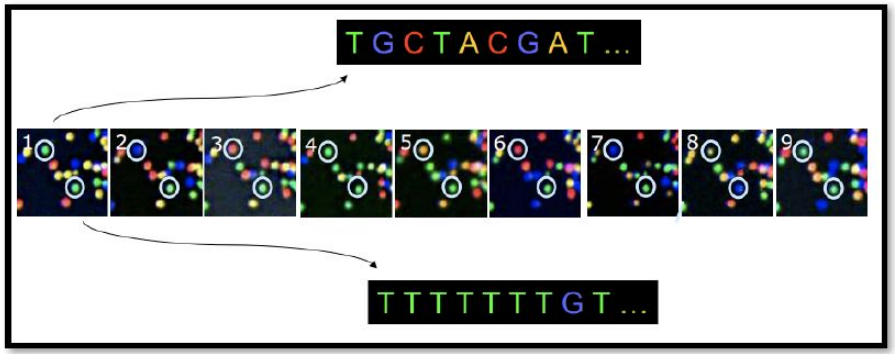
7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction



2. Base Calling

bcl2fastq





https://www.broadinstitute.org/files/shared/illuminavids/sequencingSlides.pdf

download at https://emea.support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html

Sequence Formats: FastQ

- de-facto standard format is FastQ
  - based on FastA + 2 lines for **Q**uality scores

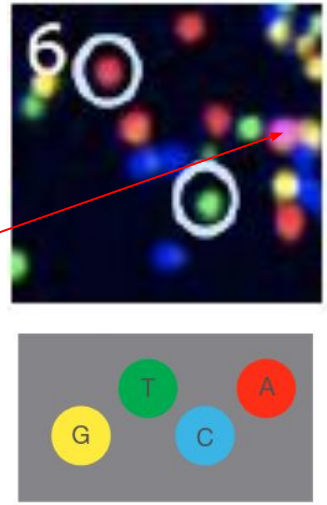
4 lines per read:

1. @Sequence identifier (Illumina: origin and optionally length)
2. Sequence
3. +(Optional: Identifier again)
4. Base quality, mapped to ASCII 33-128 (Illumina)

G: 71
*: 42

```
@M04304:185:000000000-CMBBK:1:1102:12988:2175 1:N:0:AAGAGGCA+TATCCTTT
ACTGACGCTGAGGCACGAAAGCGTGGGTATCGAACAGGATTAGATACCCGTGTAGTCC
+
FGFGGGGGGGGGGGGFGEEFFGGGCEEGF:FD:875*5:CF?F<+<CFEDDGF:4=@F
```

Raw Data Quality

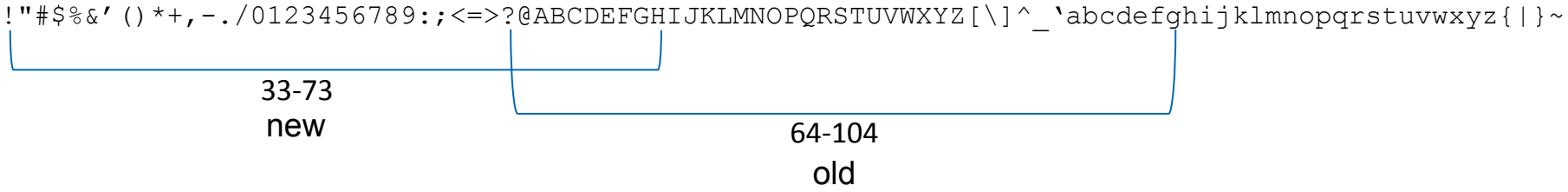


- There are always errors!
- Quality "control"
  - how bad is sequencing quality, i.e. how many reads should be discarded?
- Source of errors: basecalling
  - "Sanger **Phred**" quality score = how certain is this called a G base?
  - how likely is it wrong: P
  - expressed as $Q = -10*\log_{10}(P)$
    - $Q = 10 \Rightarrow P = 0.1$
    - $Q = 40 \Rightarrow P = 0.0001$
- The Phred score is used for
  - assess overall quality
  - discard whole reads
  - truncate reads (clipping)
    - since low quality regions often occur at the start and end
  - SNV determination

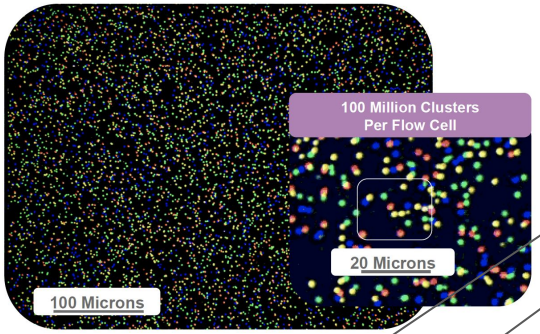
Raw data quality

- encodes Phred Score
- $Q = -10*\log_{10}(P)$
- Q encoded as ASCII character
  - Q+33 for versions ≥ 1.8
  - Q+64 for versions 1.3 to 1.7

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	[END 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]



Sequence Formats: FastQ



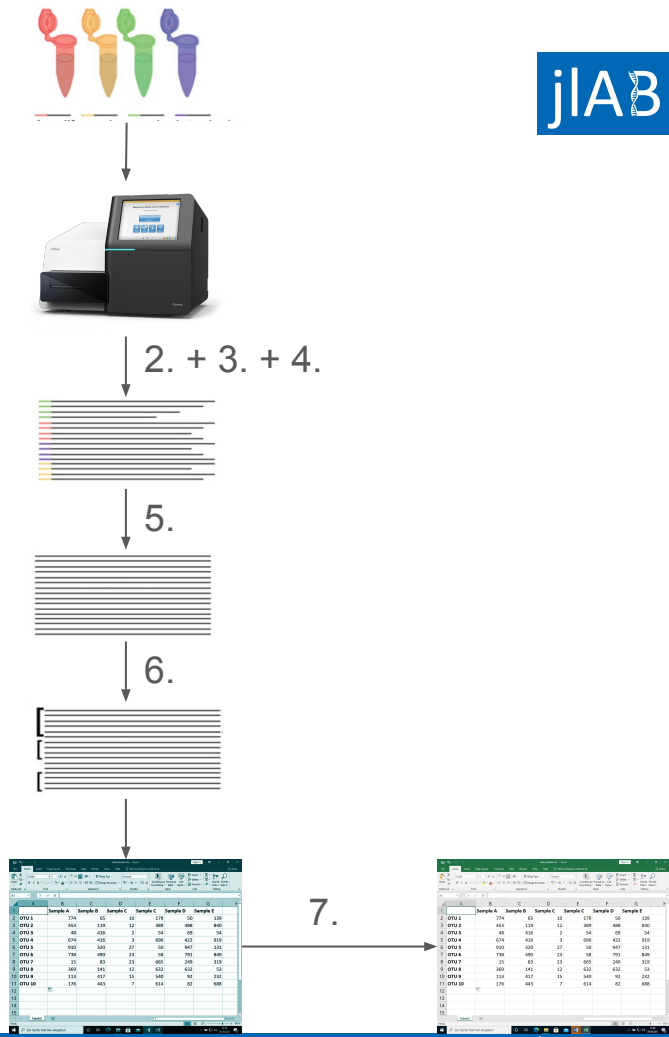
instrument ID
i[th] run on this instrument
flowcell ID
lane
tile number (in flowcell)
x-coord. of cluster (in tile)
y-coord. of cluster (in tile)
partner ID (2 only for PE)
filtered read (N | Y)
control bits: 0 = no errors
index sequences

```
@M04304:185:000000000-CMBBK:1:1102:12988:2175 1:N:0:AAGAGGCA+TATCCTTT
ACTGACGCTGAGGCACGAAAGCGTGGGTATCGAACAGGATTAGATACCCGTGTAGTCC
+
FGFGGGGGGGGGGGGFGEEFFGGGCEEGF:FD:875*5:CF?F<+<CFEDDGF:4=@F
```

Content

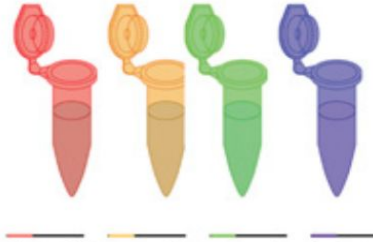
1. Scope ✔
2. Base Calling ✔
3. Demultiplexing
4. Quality Control
5. Adapter Trimming, Clipping
6. "OTU" picking

7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction

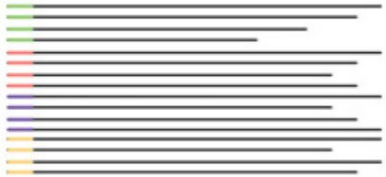


3. Demultiplexing

again: `bcl2fastq`

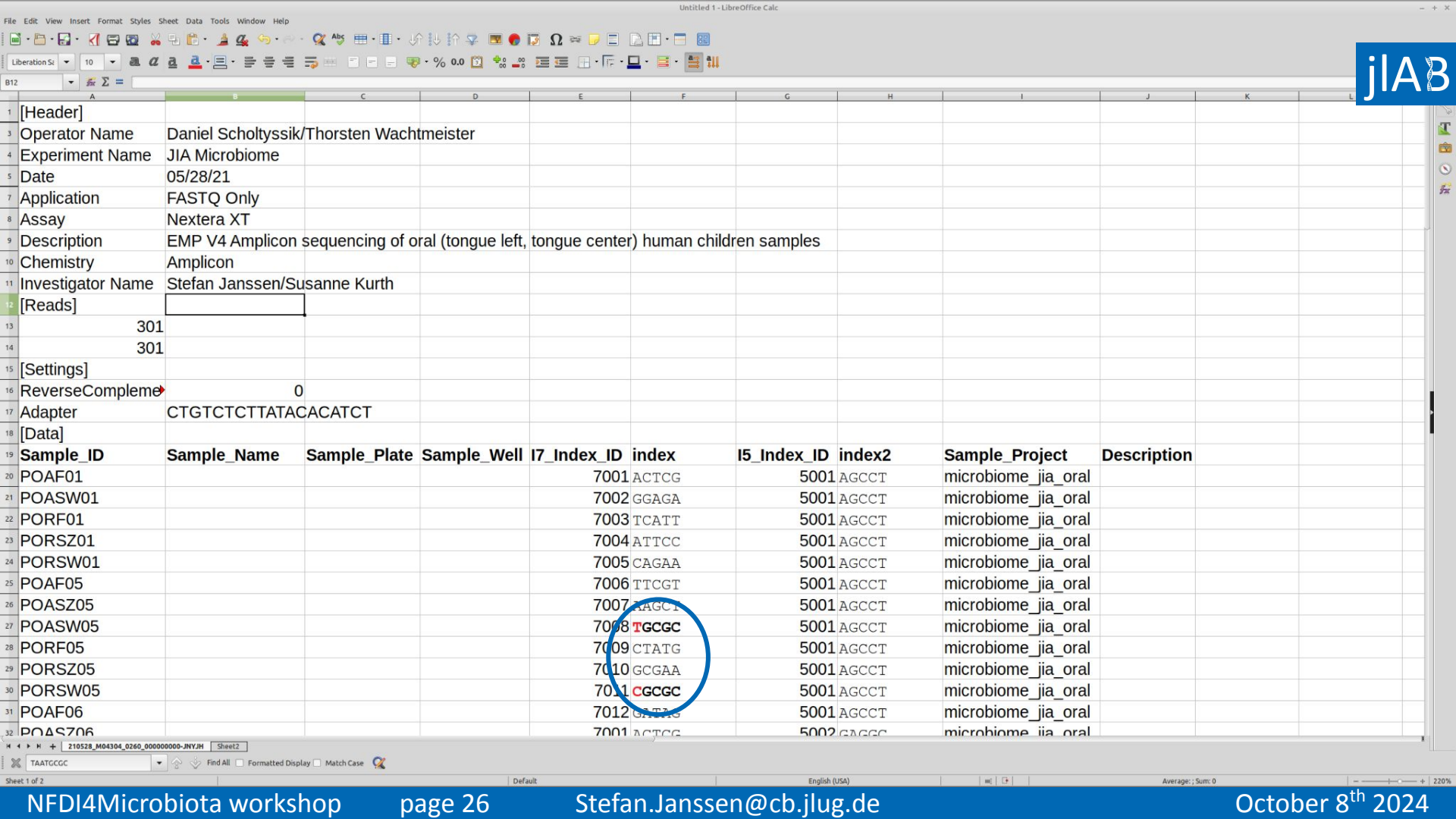


Amplify each sample, introducing barcode into each sequence using tagged PCR primers



Use barcodes to assign each sequence to the sample it came from, dropping low-quality reads

[Header]									
Operator Name	Daniel Scholtyssik/Thorsten Wachtmeister								
Experiment Name	JIA Microbiome								
Date	05/28/21								
Application	FASTQ Only								
Assay	Nextera XT								
Description	EMP V4 Amplicon sequencing of oral (tongue left, tongue center) human children samples								
Chemistry	Amplicon								
Investigator Name	Stefan Janssen/Susanne Kurth								
[Reads]									
301									
301									
[Settings]									
ReverseComplement	0								
Adapter	CTGTCTCTTATACACATCT								
[Data]									
Sample_ID	**Sample_Name**	**Sample_Plate**	**Sample_Well**	**I7_Index_ID**	**index**	**I5_Index_ID**	**index2**	**Sample_Project**	**Description**
POAF01				7001	ATTACTCG	5001	TATAGCCT	microbiome_jia_oral	
POASW01				7002	TCCGGAGA	5001	TATAGCCT	microbiome_jia_oral	
PORF01				7003	CGCTCATT	5001	TATAGCCT	microbiome_jia_oral	
PORSZ01				7004	GAGATTCC	5001	TATAGCCT	microbiome_jia_oral	
PORSW01				7005	ATTCAGAA	5001	TATAGCCT	microbiome_jia_oral	
POAF05				7006	GAATTCGT	5001	TATAGCCT	microbiome_jia_oral	
POASZ05				7007	CTGAAGCT	5001	TATAGCCT	microbiome_jia_oral	
POASW05				7008	TAATGCGC	5001	TATAGCCT	microbiome_jia_oral	
PORF05				7009	CGGCTATG	5001	TATAGCCT	microbiome_jia_oral	
PORSZ05				7010	TCCGCGAA	5001	TATAGCCT	microbiome_jia_oral	
PORSW05				7011	TCTCGCGC	5001	TATAGCCT	microbiome_jia_oral	
POAF06				7012	AGCGATAG	5001	TATAGCCT	microbiome_jia_oral	
POASZ06				7001	ATTACTCG	5002	ATAGAGGC	microbiome_jia_oral	



[Header]										
Operator Name	Daniel Scholtyssik/Thorsten Wachtmeister									
Experiment Name	JIA Microbiome									
Date	05/28/21									
Application	FASTQ Only									
Assay	Nextera XT									
Description	EMP V4 Amplicon sequencing of oral (tongue left, tongue center) human children samples									
Chemistry	Amplicon									
Investigator Name	Stefan Janssen/Susanne Kurth									
[Reads]										
301										
301										
[Settings]										
ReverseComplemen	0									
Adapter	CTGTCTCTTATACACATCT									
[Data]										
Sample_ID	**Sample_Name**	**Sample_Plate**	**Sample_Well**	**I7_Index_ID**	**index**	**I5_Index_ID**	**index2**	**Sample_Project**	**Description**	
POAF01				7001	ACTCG	5001	AGCCT	microbiome_jia_oral		
POASW01				7002	GGAGA	5001	AGCCT	microbiome_jia_oral		
PORF01				7003	TCATT	5001	AGCCT	microbiome_jia_oral		
PORSZ01				7004	ATTCC	5001	AGCCT	microbiome_jia_oral		
PORSW01				7005	CAGAA	5001	AGCCT	microbiome_jia_oral		
POAF05				7006	TTCGT	5001	AGCCT	microbiome_jia_oral		
POASZ05				7007	AAGCT	5001	AGCCT	microbiome_jia_oral		
POASW05				7008	TGCGC	5001	AGCCT	microbiome_jia_oral		
PORF05				7009	CTATG	5001	AGCCT	microbiome_jia_oral		
PORSZ05				7010	GCGAA	5001	AGCCT	microbiome_jia_oral		
PORSW05				7011	CGCGC	5001	AGCCT	microbiome_jia_oral		
POAF06				7012	GATAG	5001	AGCCT	microbiome_jia_oral		
POASZ06				7001	ACTCG	5002	GAGGC	microbiome_jia_oral		

Use Error tolerant barcodes!

```
bcl2fastq --help

--barcode-mismatches arg (=1)                    number of allowed mismatches per index
```

e.g. 📄 Hamady et al. (2008) *Nature methods* "Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex"

[Data]									
Sample_ID	**Sample_Name**	**Sample_Plate**	**Sample_Well**	**I7_Index_ID**	**index**	**I5_Index_ID**	**index2**	**Sample_Project**	**Description**
POAF01				7001	ACTCG	5001	AGCCT	microbiome_jia_oral	
POASW01				7002	GGAGA	5001	AGCCT	microbiome_jia_oral	
PORF01				7003	TCATT	5001	AGCCT	microbiome_jia_oral	
PORSZ01				7004	ATTCC	5001	AGCCT	microbiome_jia_oral	
PORSW01				7005	CAGAA	5001	AGCCT	microbiome_jia_oral	
POAF05				7006	TTCGT	5001	AGCCT	microbiome_jia_oral	
POASZ05				7007	AAGCT	5001	AGCCT	microbiome_jia_oral	
POASW05				7008	TGCGC	5001	AGCCT	microbiome_jia_oral	
PORF05				7009	CTATG	5001	AGCCT	microbiome_jia_oral	
PORSZ05				7010	GCGAA	5001	AGCCT	microbiome_jia_oral	
PORSW05				7011	CGCGC	5001	AGCCT	microbiome_jia_oral	
POAF06				7012	GATAG	5001	AGCCT	microbiome_jia_oral	
POASZ06				7001	ACTCG	5002	GAGGC	microbiome_jia_oral	

Excel is too smart

always prefix sample names with characters!

[Hea									
Ope									
Exp									
Date	05/28/21								
App									
Ass									
Des									
Chemistry	Amplicon								
Investigator Name	Stefan Janssen/Susanne Kurth								
[Reads]									
301									
301									
[Settings]									
ReverseComplem	0								
Adapter	CTGTCTCTTATACACATCT								
[Data]									
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
001	1			7001	ACT				
002	2			7002	GGA				
003	3			7003	TCA				
004	4			7004	ATT				
005	5			7005	CAG				
006	6			7006	TTC				
007	7			7007	AAG				
008	8			7008	TGC				
009	9			7009	CTA				
010	10			7010	GCG				
011	11			7011	CGC				
012	12			7012	GAT				
013	13			7001	ACT				

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg†, Joseph Riss†, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett and John N Weinstein

† Contributed equally

BMC Bioinformatics 2004 5:80 | DOI: 10.1186/1471-2105-5-80 | © Zeeberg et al; licensee BioMed Central Ltd. 2004

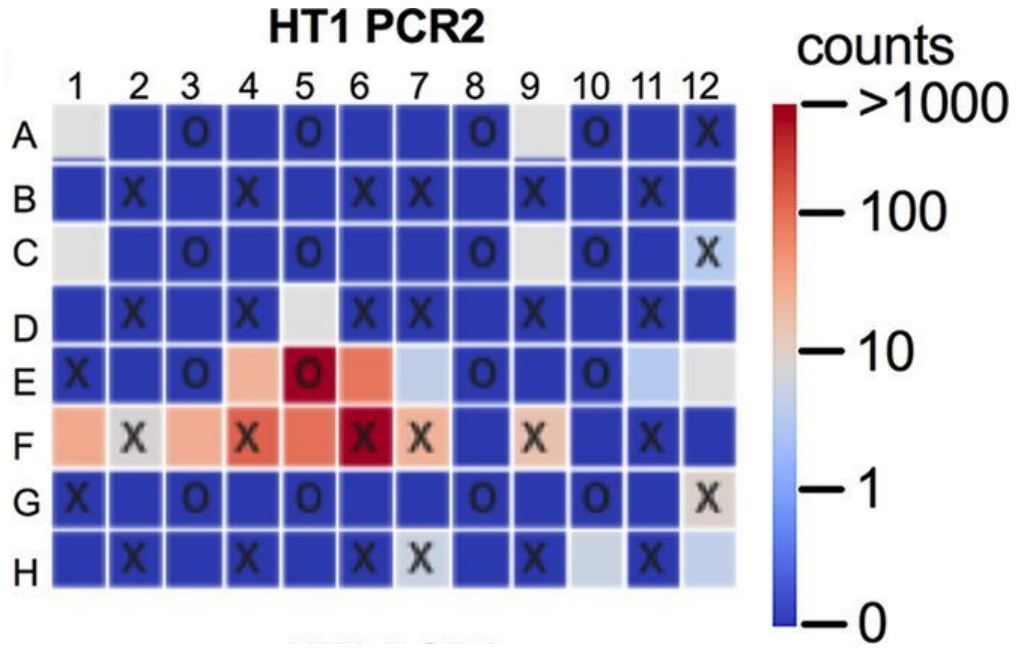
NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

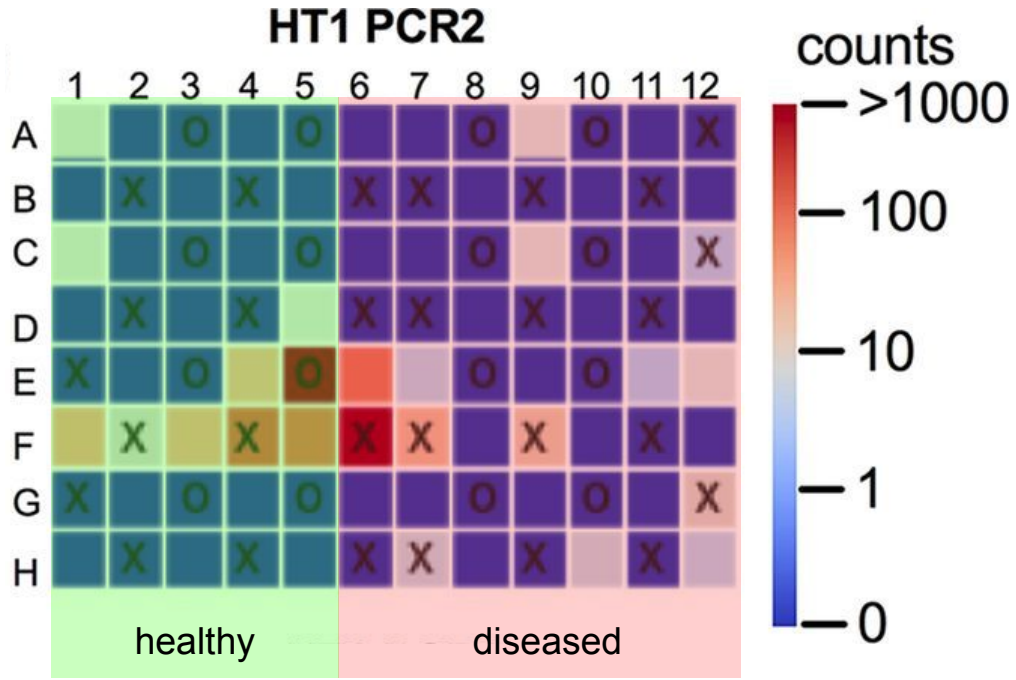
https://www.nature.com/articles/d41586-021-02211-4

Randomize Plate Layout!



Minich et al. (2019) *mSystems* "Quantifying and Understanding Well-to-Well Contamination in Microbiome Research"

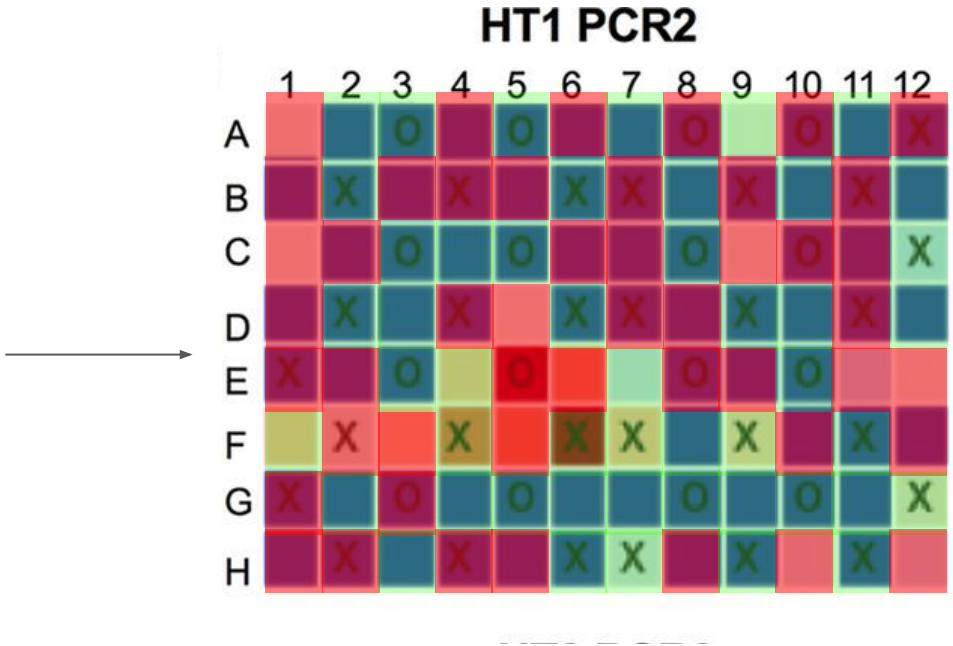
Randomize Plate Layout!



Minich et al. (2019) *mSystems* "Quantifying and Understanding Well-to-Well Contamination in Microbiome Research"

Randomize Plate Layout!

TA: "It feels soooo wrong"



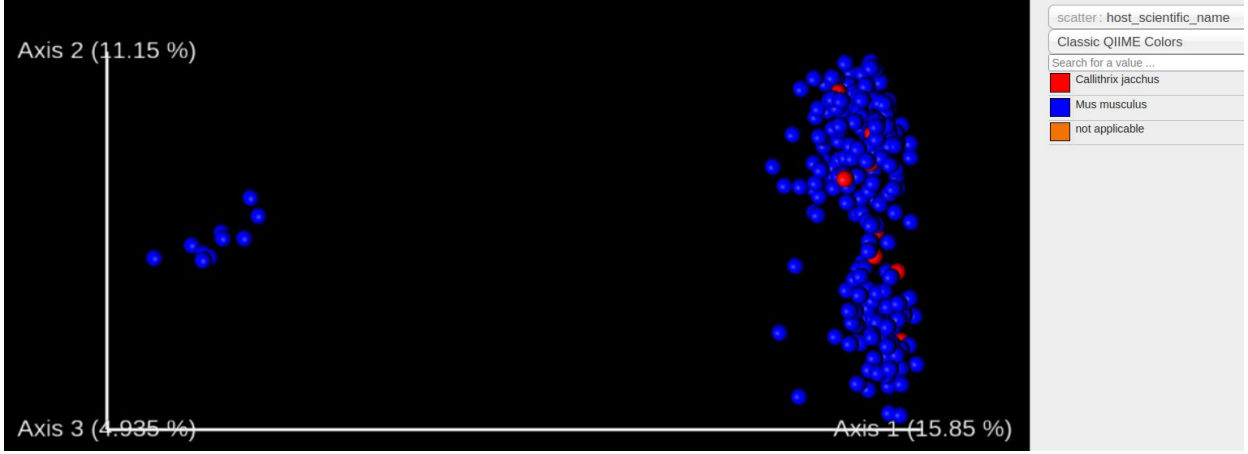
📄 Minich et al. (2019) *mSystems* "Quantifying and Understanding Well-to-Well Contamination in Microbiome Research"

Invest time in your demultiplexing sheet!

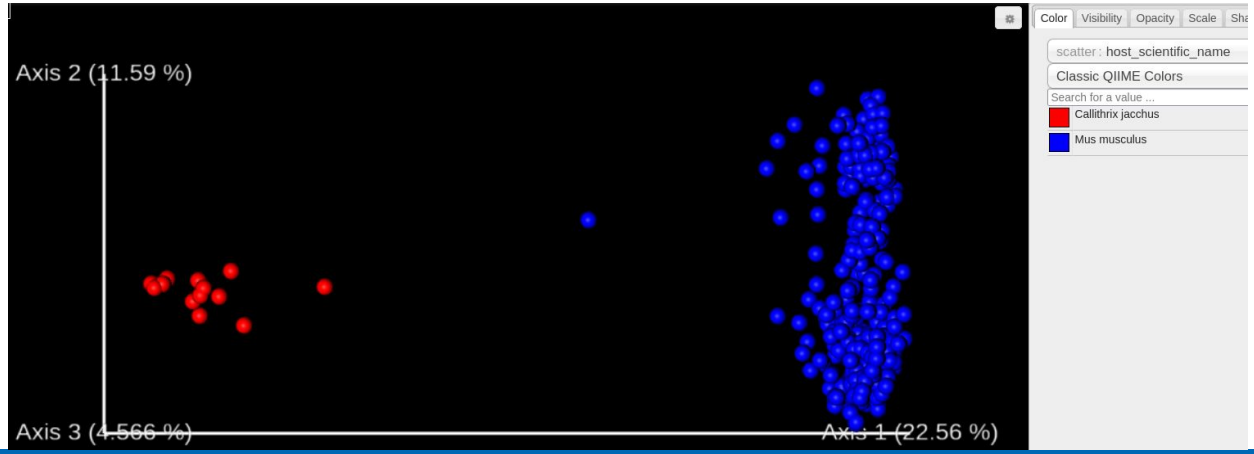
Workflow	GenerateFASTQ	
Application	FASTQ Only	
Instrument Type	MiSeq	
Assay	Nextera XT	
Index Adapters	Nextera XT v2 Index Kit	
Chemistry	Amplicon	
[Reads]		
301		
301		
[Settings]		
ReverseComplement	0	
Adapter	CTGTCTCTTATACACATCT	
[Data]		

Sample_ID	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
1			N701-A	TAAGGCGA	S502-A	CTCTCTAT	148	
2			N701-A	TAAGGCGA	S503-A	TATCCTCT	148	
3			N701-A	TAAGGCGA	S505-A	GTAAGGAG	148	
4			N701-A	TAAGGCGA	S506-A	ACTGCATA	148	
5			N701-A	TAAGGCGA	S507-A	AAGGAGTA	148	
6			N701-A	TAAGGCGA	S508-A	CTAAGCCT	148	
7			N701-A	TAAGGCGA	S510-A	CGTCTAAT	148	
8			N701-A	TAAGGCGA	S511-A	TCTCTCCG	148	
9			N702-A	CGTACTAG	S502-A	CTCTCTAT	148	
10			N702-A	CGTACTAG	S503-A	TATCCTCT	148	
11			N702-A	CGTACTAG	S505-A	GTAAGGAG	148	
12			N702-A	CGTACTAG	S506-A	ACTGCATA	148	
13			N702-A	CGTACTAG	S507-A	AAGGAGTA	148	
14			N702-A	CGTACTAG	S508-A	CTAAGCCT	148	
15			N702-A	CGTACTAG	S510-A	CGTCTAAT	148	
16			N702-A	CGTACTAG	S511-A	TCTCTCCG	148	
17			N703-A	AGGCAGAA	S502-A	CTCTCTAT	148	
18			N703-A	AGGCAGAA	S503-A	TATCCTCT	148	
19			N703-A	AGGCAGAA	S505-A	GTAAGGAG	148	
20			N703-A	AGGCAGAA	S506-A	ACTGCATA	148	
21			N703-A	AGGCAGAA	S507-A	AAGGAGTA	148	
22			N703-A	AGGCAGAA	S508-A	CTAAGCCT	148	
23			N703-A	AGGCAGAA	S510-A	CGTCTAAT	148	
24			N703-A	AGGCAGAA	S511-A	TCTCTCCG	148	
25			N704-A	TCCTGAGC	S502-A	CTCTCTAT	148	
26			N704-A	TCCTGAGC	S503-A	TATCCTCT	148	
27			N704-A	TCCTGAGC	S505-A	GTAAGGAG	148	
28			N704-A	TCCTGAGC	S506-A	ACTGCATA	148	
29			N704-A	TCCTGAGC	S507-A	AAGGAGTA	148	
30			N704-A	TCCTGAGC	S508-A	CTAAGCCT	148	
31			N704-A	TCCTGAGC	S510-A	CGTCTAAT	148	
32			N704-A	TCCTGAGC	S511-A	TCTCTCCG	148	

5	4	2018-08-19	RM	rm2	2019-01-03	recipient mother
6	5	2018-08-19	RM	rm3	2019-01-03	recipient mother
7	6	2018-08-19	RM	rm3	2019-01-03	recipient mother
8	7	2018-11-30	PB	P1-10	2019-03-19	motherP1
9	8	2018-11-30	PB	P1-10	2019-03-19	motherP1
10	9	2018-11-30	PB	P1-11	2019-03-20	motherP1
11	10	2018-11-30	PB	P1-11	2019-03-20	motherP1
12	11	2018-11-30	PB	P1-12	2019-03-19	fatherP1
13	12	2018-11-30	PC	P1-1	2019-03-19	motherP1
14	13	2018-11-30	PC	P1-1	2019-03-19	motherP1
15	14	2018-11-30	PC	P1-2	2019-03-18	fatherP1
16	15	2018-11-30	PC	P1-3	2019-03-20	motherP1
17	16	2018-11-30	PC	P1-3	2019-03-20	motherP1
18	17	2018-11-30	PC	P1-4	2019-03-18	fatherP1
19	18	2018-11-30	PC	P1-5	2019-03-19	motherP1
20	19	2018-11-30	PC	P1-5	2019-03-19	motherP1
21	20	2018-11-30	PC	P1-6	2019-03-18	fatherP1
22	21	2018-11-30	PC	P1-7	2019-03-18	not provided
23	22	2018-11-30	PC	P1-8	2019-03-18	not provided
24	23	2018-11-30	PC	P1-8	2019-03-18	not provided
25	24	2018-11-30	PC	P1-8	2019-03-18	not provided
26	25	2018-11-30	PC	P1-8	2019-03-18	not provided
27	26	2018-11-30	PC	P1-9	2019-03-18	not provided
28	27	2018-11-30	PC	P1-9	2019-03-18	not provided
29	28	2019-02-15	BF1-1	F1-15	2019-05-31	motherF1
30	29	2019-02-15	BF1-1	F1-16	2019-05-31	fatherF1
31	30	2019-02-15	BF1-1	F1-17	2019-05-31	not provided
32	31	2019-02-15	BF1-1	F1-18	2019-05-31	not provided
33	32	2019-02-15	BF1-1	F1-19	2019-05-31	not provided
34	33	2019-02-18	BF1-2	F1-20	2019-06-03	not provided
35	34	2019-02-18	BF1-2	F1-20	2019-06-03	not provided
36	35	2019-02-27	BF1-2	F1-21	2019-06-12	motherF1
37	36	2019-02-27	BF1-2	F1-21	2019-06-12	motherF1
38	37	2019-02-27	BF1-2	F1-22	2019-06-12	fatherF1
39	38	2019-02-27	BF1-2	F1-23	2019-06-12	not provided

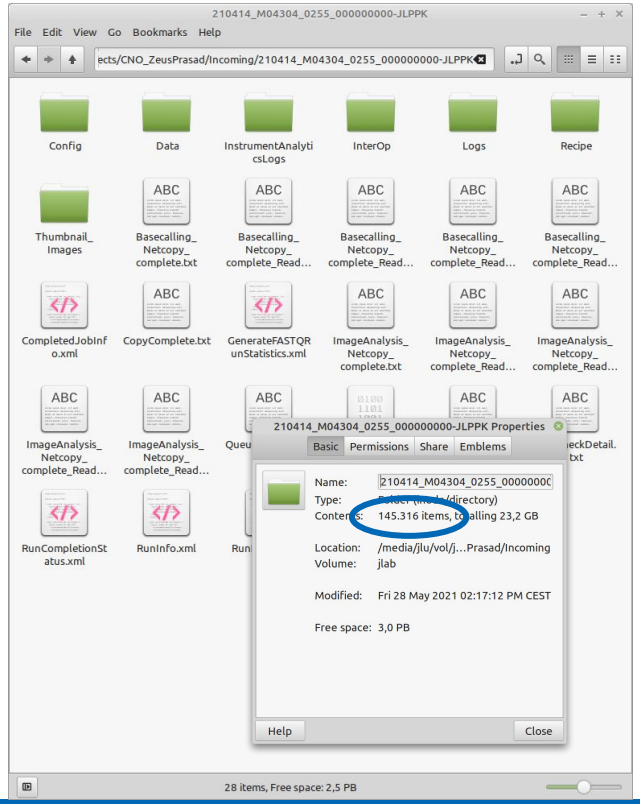


a lot of frustration

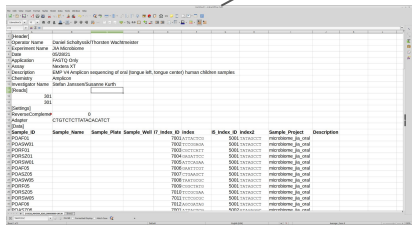


	A	B	C	D	E	F	G	H	I	J	K	L
1	number	well	plate name	sample name	concentration ng/µL	sample volume µL	solvent	sample type	organism	comment	514	
2	1	A01	1	90	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
3	2	B01	1	178	22	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
4	3	C01	1	175	13	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
5	4	D01	1	192	21	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
6	5	E01	1	117	15	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
7	6	F01	1	29	26	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
8	7	G01	1	98	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
9	8	H01	1	99	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
10	9	A02	1	261	14	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
11	10	B02	1	334	1	25	EB (Tris-H	other	other	blank		
12	11	C02	1	82	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
13	12	D02	1	216	28	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
14	13	E02	1	269	33	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
15	14	F02	1	338	1	25	EB (Tris-H	other	other	blank		
16	15	G02	1	252	35	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
17	16	H02	1	158	19	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
18	17	A03	1	185	23	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
19	18	B03	1	80	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
20	19	C03	1	190	14	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
21	20	D03	1	180	22	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
22	21	E03	1	273	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
23	22	F03	1	316	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfeces	monkey	
24	23	G03	1	56	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
25	24	H03	1	179	35	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
26	25	A04	1	23	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
27	26	B04	1	239	19	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
28	27	C04	1	181	10	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
29	28	D04	1	197	12	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
30	29	E04	1	296	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
31	30	F04	1	223	31	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
32	31	G04	1	76	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
33	32	H04	1	103	25	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
34	33	A05	1	120	31	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
35	34	B05	1	215	21	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
36	35	C05	1	335	1	25	EB (Tris-H	other	other	blank		
37	36	D05	1	130	36	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	
38	37	E05	1	303	19	25	EB (Tris-H	genomic DNA (I	prokaryote	DNAfrom colon feces	mouse	

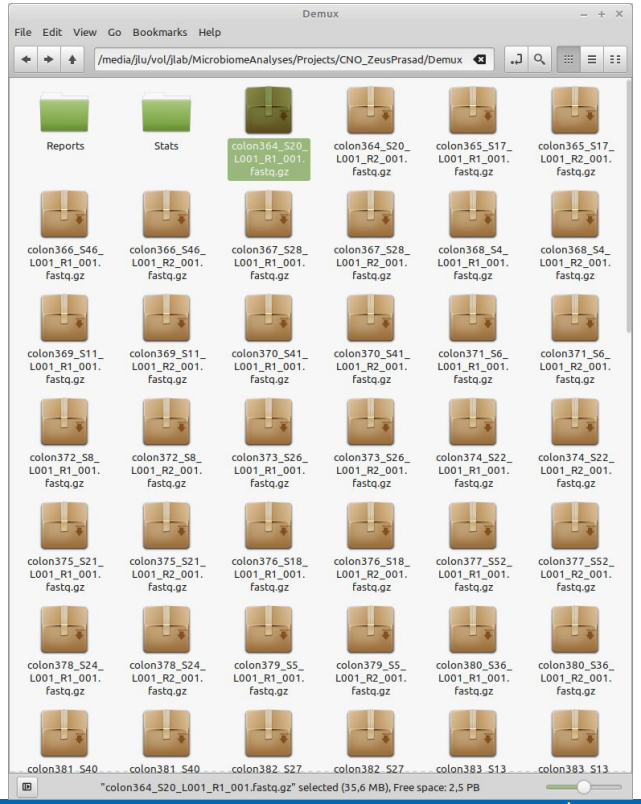
2. Base Calling & 3. Demultiplexing



bcl2fastq



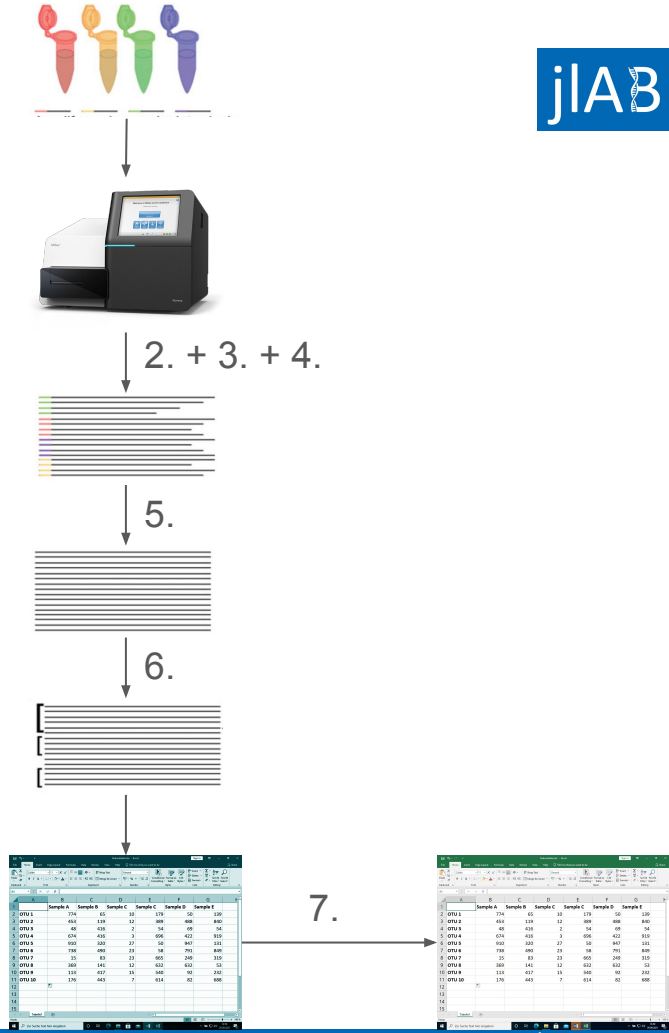
demultiplexing sheet



Content

1. Scope ✔
2. Base Calling ✔
3. Demultiplexing ✔
4. Quality Control
5. Adapter Trimming, Clipping
6. "OTU" picking

7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction



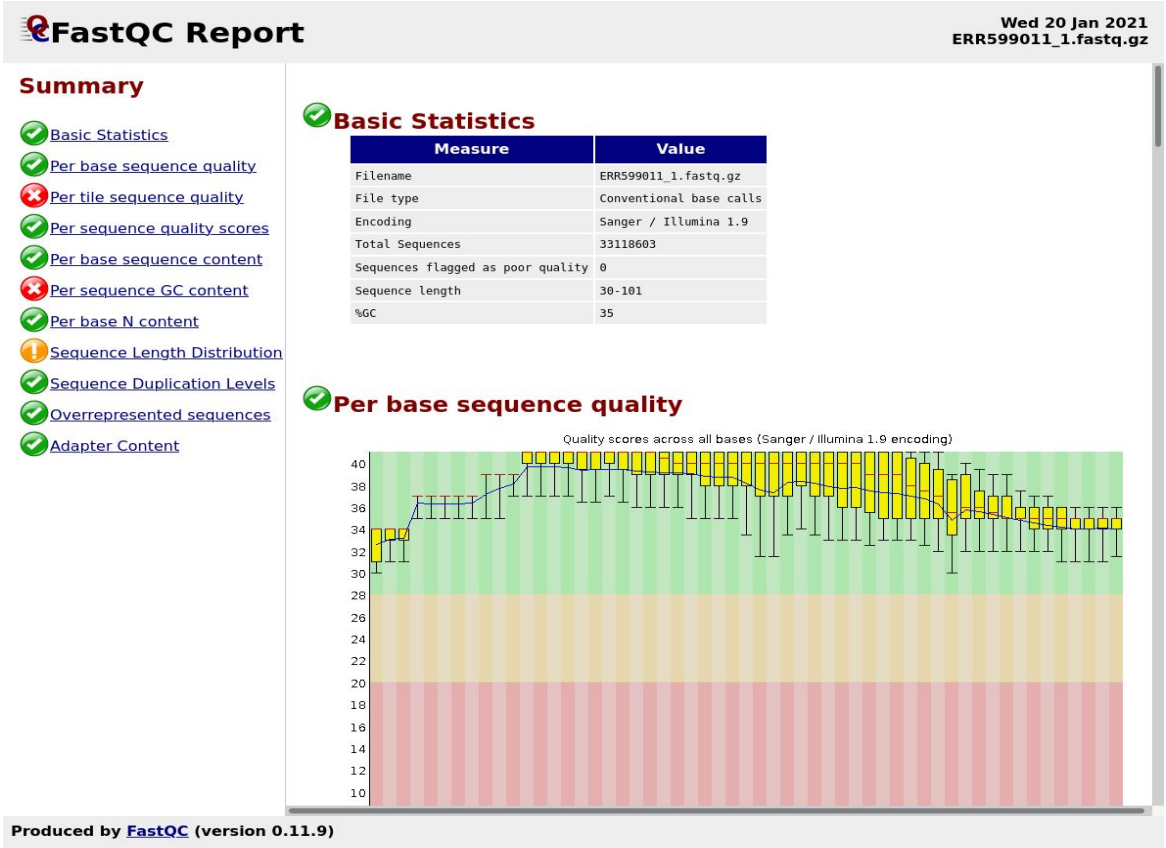
4. Quality Control

Quality assessment of raw sequence data

- How many sequences did we obtain?
- How well did DNA extraction/library prep/sequencing work?
- What preprocessing steps should be performed?

FastQC

- analyzes FASTQ file
- creates graphical report
- does NOT modify the data



FastQC Report

Wed 20 Jan 2021
ERR599011_1.fastq.gz

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	ERR599011_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	33118603
Sequences flagged as poor quality	0
Sequence length	30-101
%GC	35

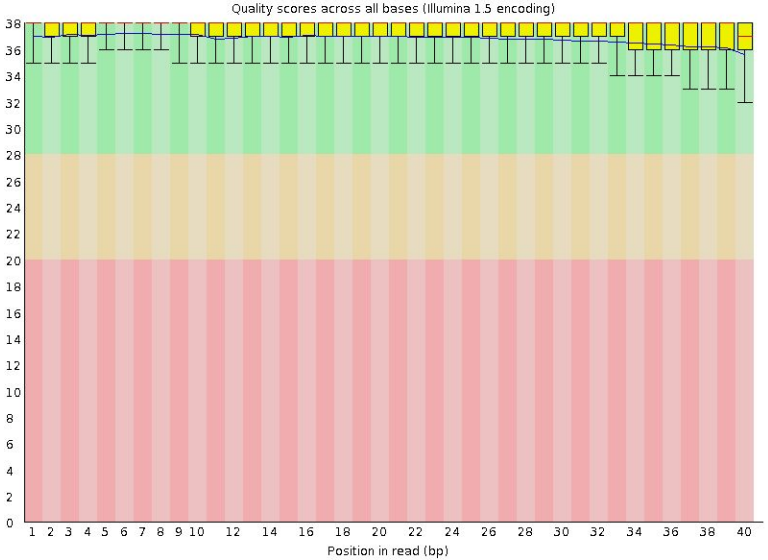
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

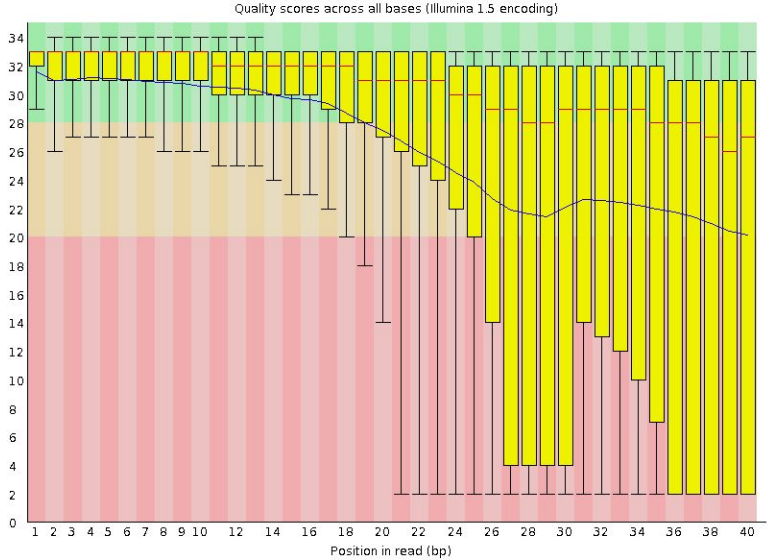
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10

Produced by FastQC (version 0.11.9)

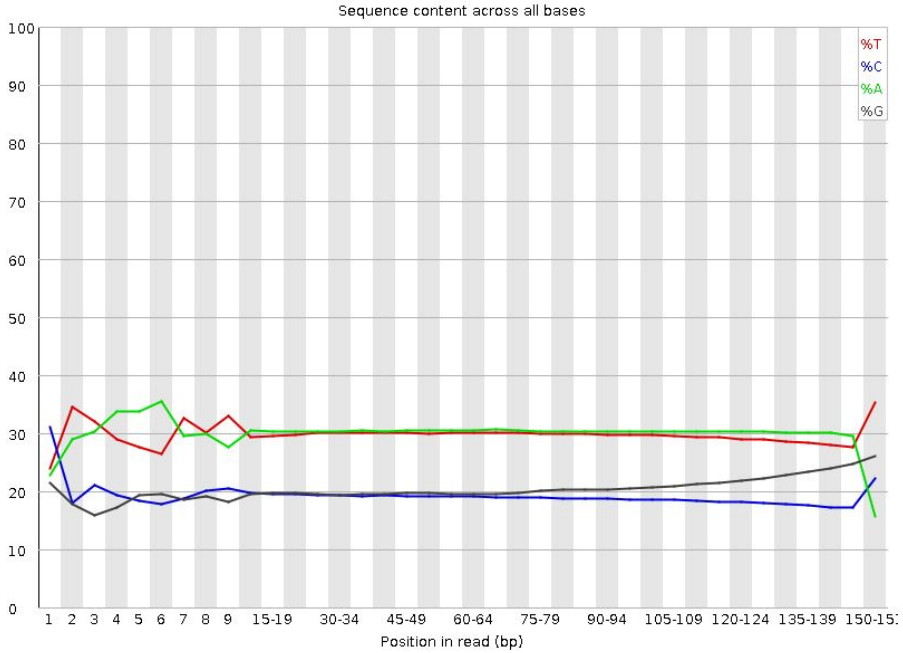
Quality control



good



not so good

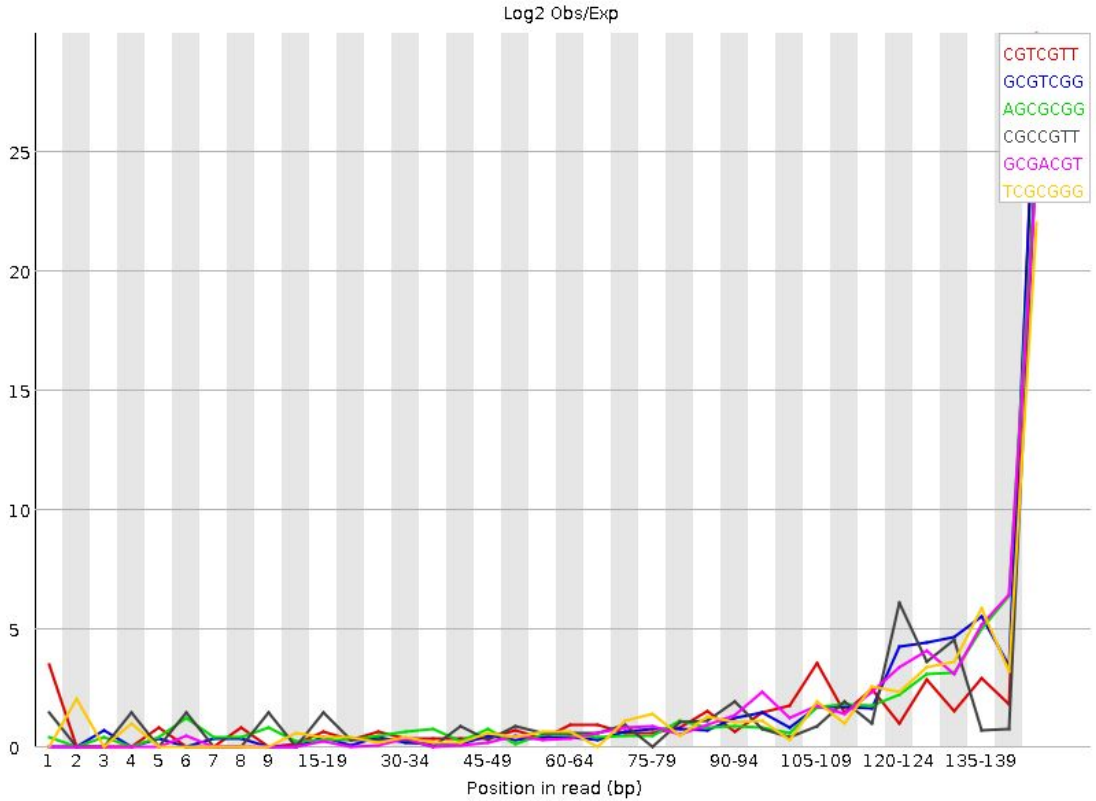


FastQC: Nucleotide distribution across all reads

- expected to be uniform for (meta)genomes
- non-uniform distribution for amplicons, (meta)transcriptomes
- some noise at 5'/3' ends to be expected

FastQC: Overrepresented kmers

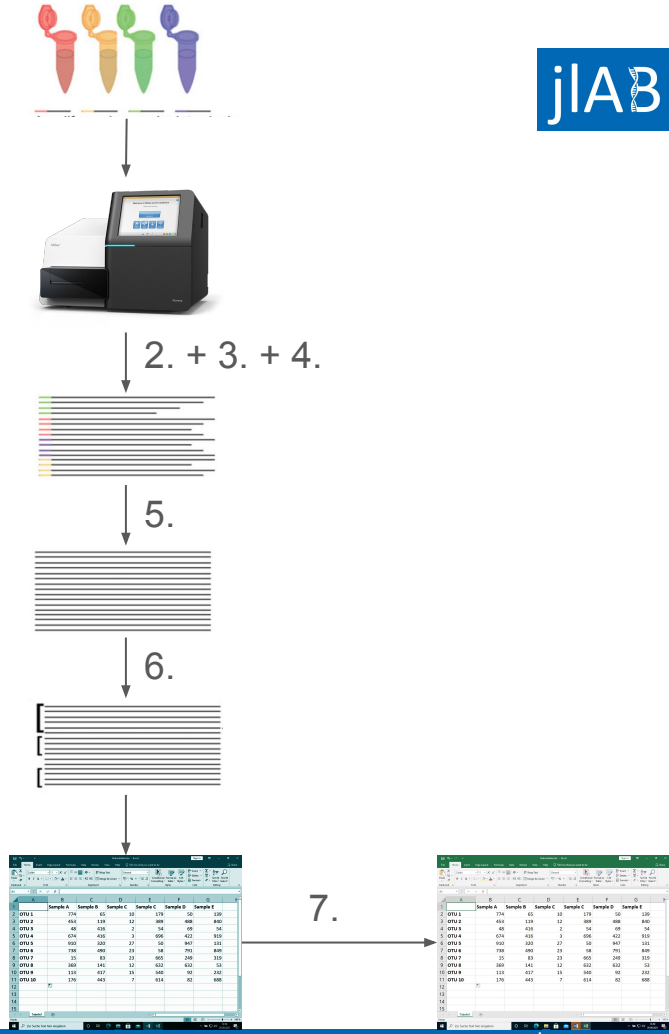
- sequencing adapter?
- barcode?
- ...



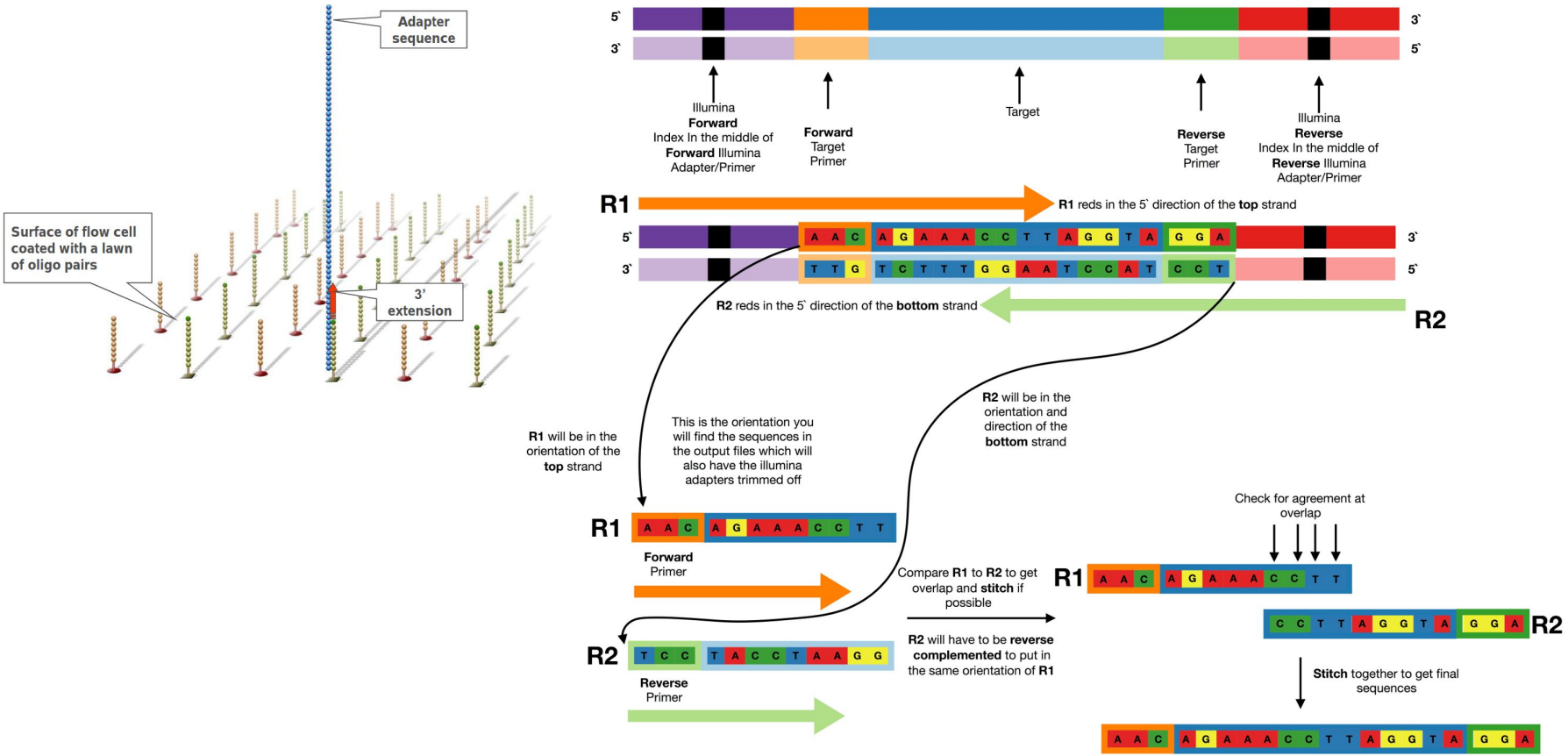
Content

1. Scope ✔
2. Base Calling ✔
3. Demultiplexing ✔
4. Quality Control ✔
5. Adapter Trimming, Clipping
6. "OTU" picking

7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction

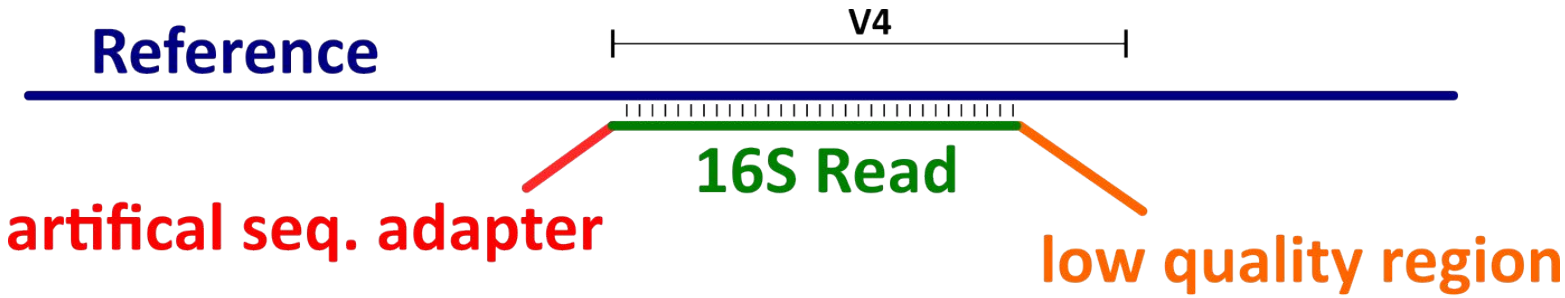


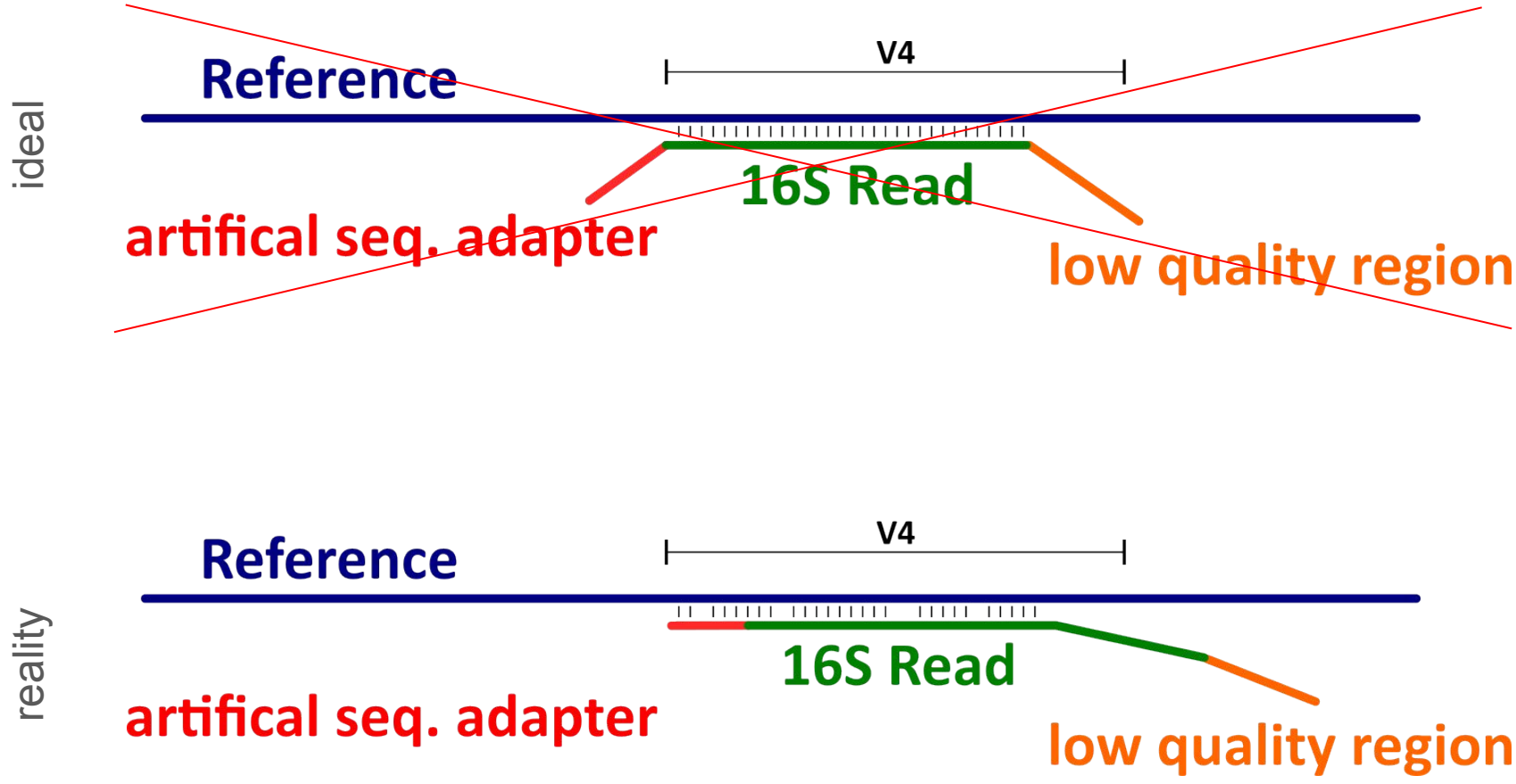
5. Adapter Trimming, Clipping



https://seekdeep.brown.edu/images/Default%20Diagram.jpg

ideal


Reference
V4
16S Read
artifical seq. adapter
low quality region


ideal
Reference
V4
16S Read
artifical seq. adapter
low quality region
reality
Reference
V4
16S Read
artifical seq. adapter
low quality region

Preprocessing: Goals

- remove technical artefacts: sequencing adapters, barcodes, ..
- trim low-quality sequences
- discard short reads
- remove host DNA, e.g. by mapping to corresponding reference genome

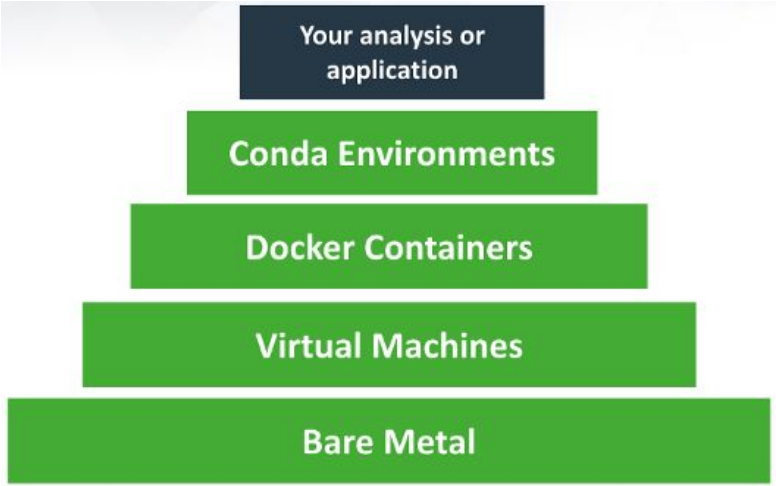
➔ **No need for high-end resources here, typical laptop is sufficient!**

Software installation / distribution not easy!

- how many users?
- central or individual maintenance?
- which operating system**(s)**?
- 1 laptop or supercomputer with 100 servers?
- performance?
- **dependencies!!**
- effort to update software
- ...

https://www.deviantart.com/xxneojadenxx/art/Windows-7-felt-sad-1075901022

https://techcrunch.com/2009/03/13/happy-birthday-linux/



https://towardsdatascience.com/conda-pip-and-docker-ftw-d64fe638dc45

Preprocessing: An iterative process!

1. Assessment: Raw sequencing data
2. **know your primer!**
3. Processing: Trimming, filtering, ..
4. Re-assessment: Data good? How much did we lose?
5. Repeat?

Tools

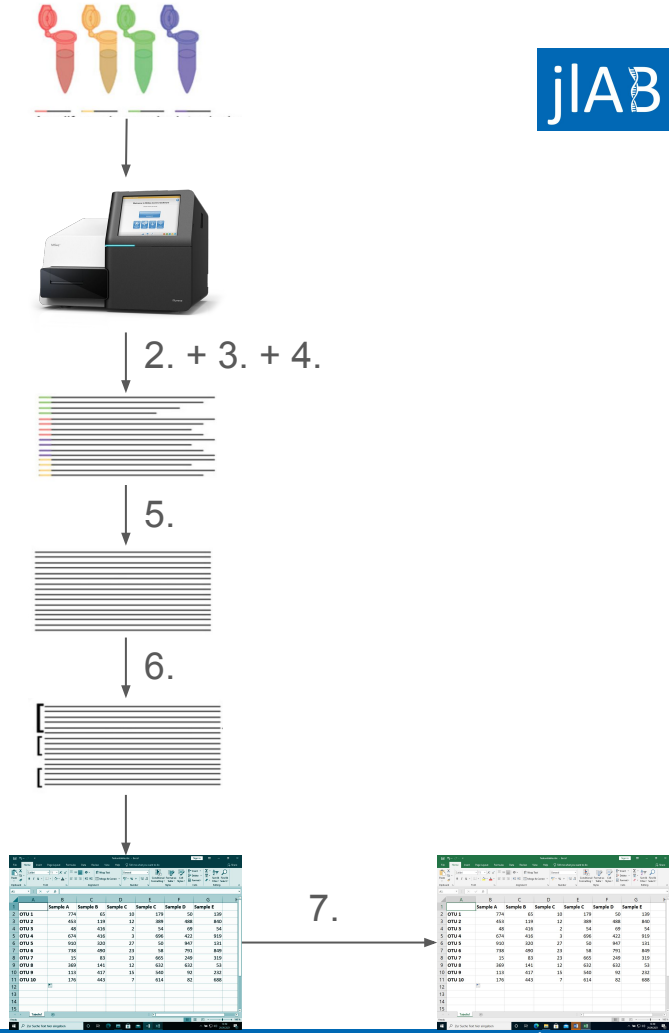
- Lots of different options available: Trimmomatic, **cutadapt**, fastx-toolkit, FastqCleaner, skewer, BbDuk, ..
- Tool performance is pretty similar, just use those you like

```
TrimmomaticSE data.fastq ILLUMINACLIP:adapter.fas:2:40:15
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Content

1. Scope ✔
2. Base Calling ✔
3. Demultiplexing ✔
4. Quality Control ✔
5. Adapter Trimming, Clipping ✔
6. "OTU" picking

7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction



6. "OTU" picking

OTU picking = grouping sequences by similarity

SampleData[SequencesWithQuality]

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

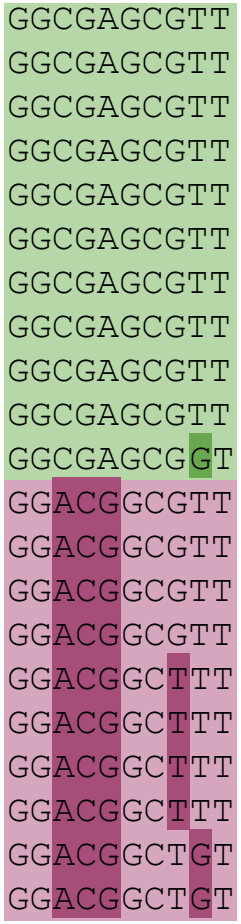
FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG

>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACG
GATGGACAAGTCTGATGTGAAAGGCTGGGGCTCAACCCCGGGACGG

>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTG
GTGGCTTAAGCGCAGGGTTTAAGGCAATGGCTTAACTATTGTTCTC

>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTG
```



90% OTU Clustering

FeatureTable[Frequency]

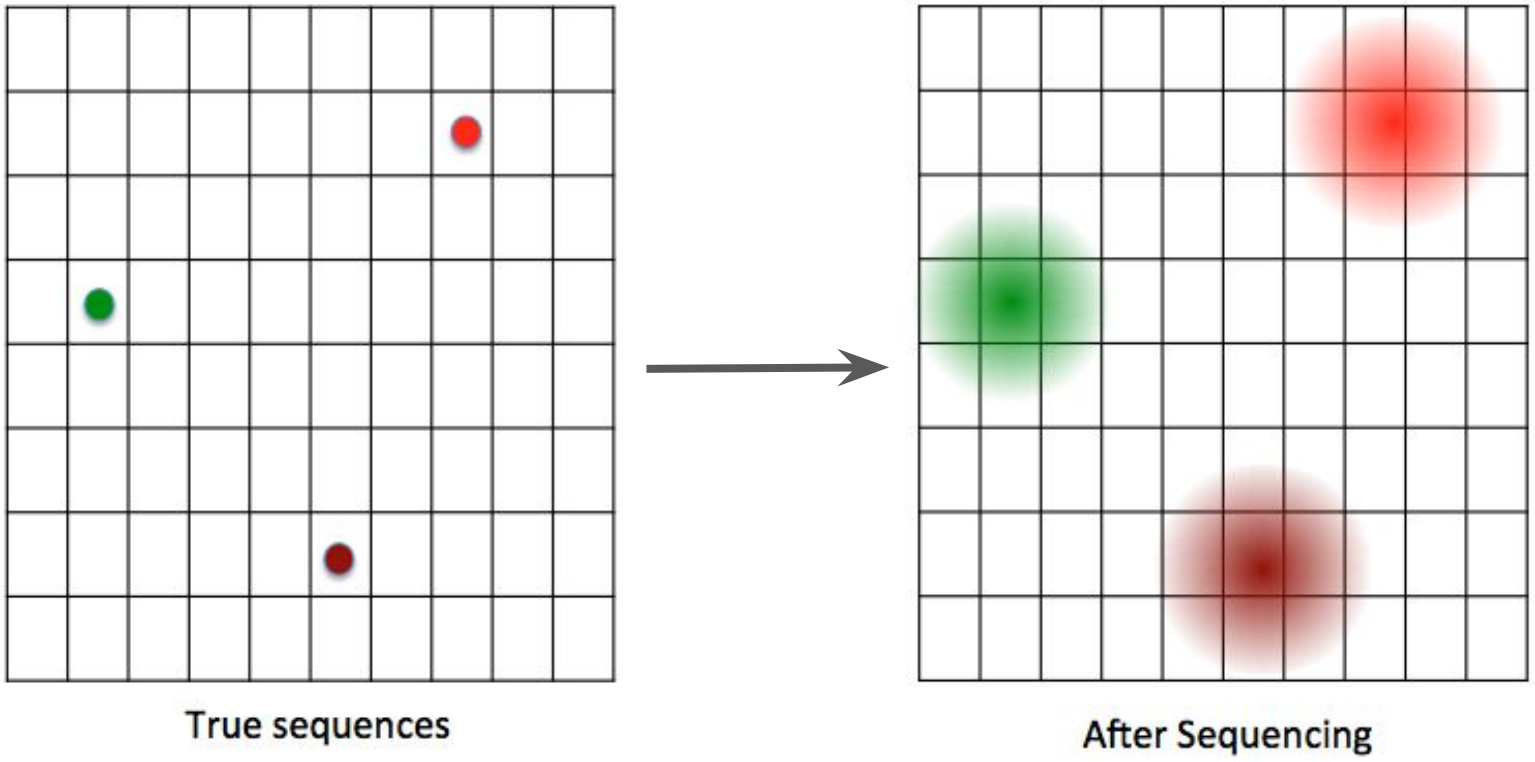
	OTU1	OTU2
4ac2	100	79
e375	88	35
4gd8	86	51
9872	12	87

Denoised Sequence Variants ("Almost like" 100% OTU Clustering)

FeatureTable[Frequency]

	ASV1	ASV2	ASV3	ASV4	ASV5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

What normally happens during sequencing?

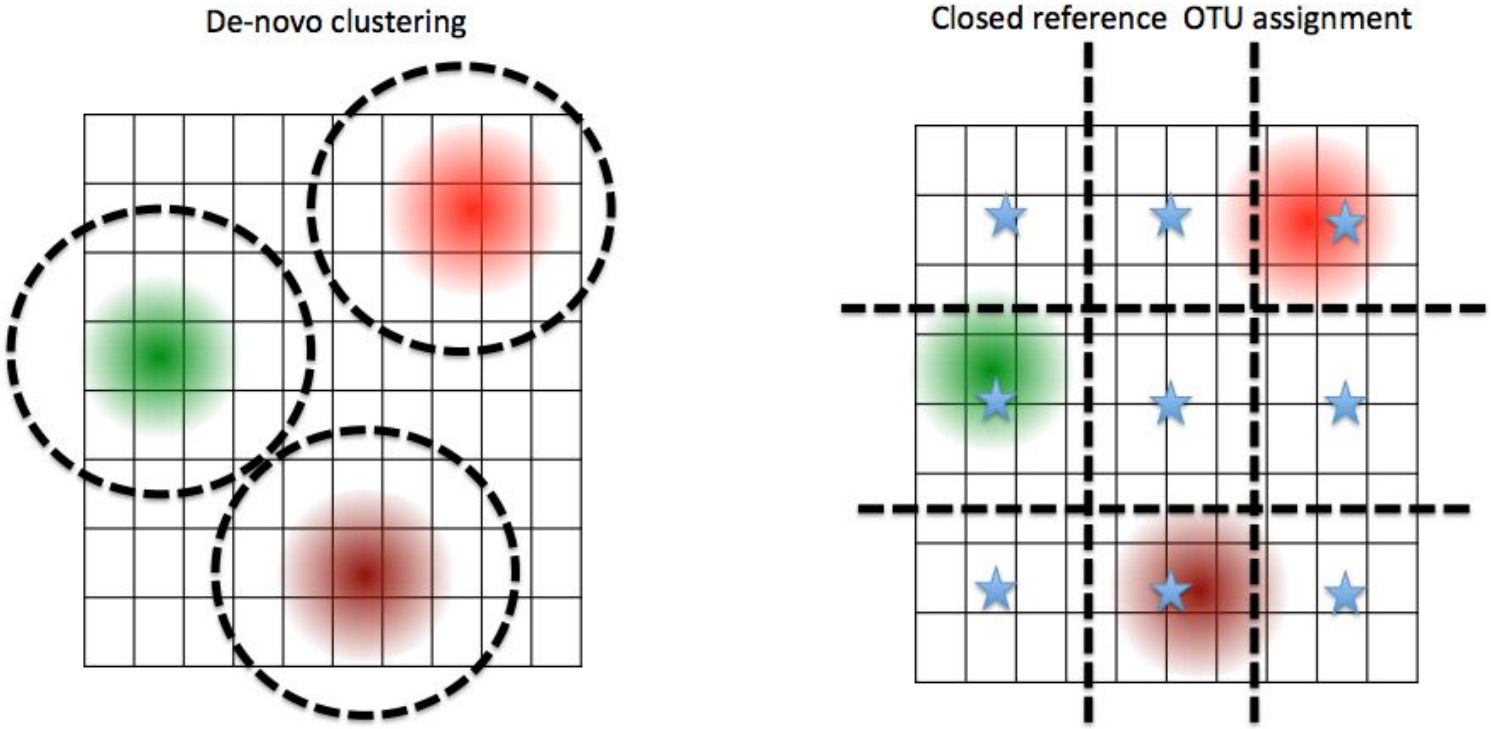


Cleaning and manipulating raw sequences

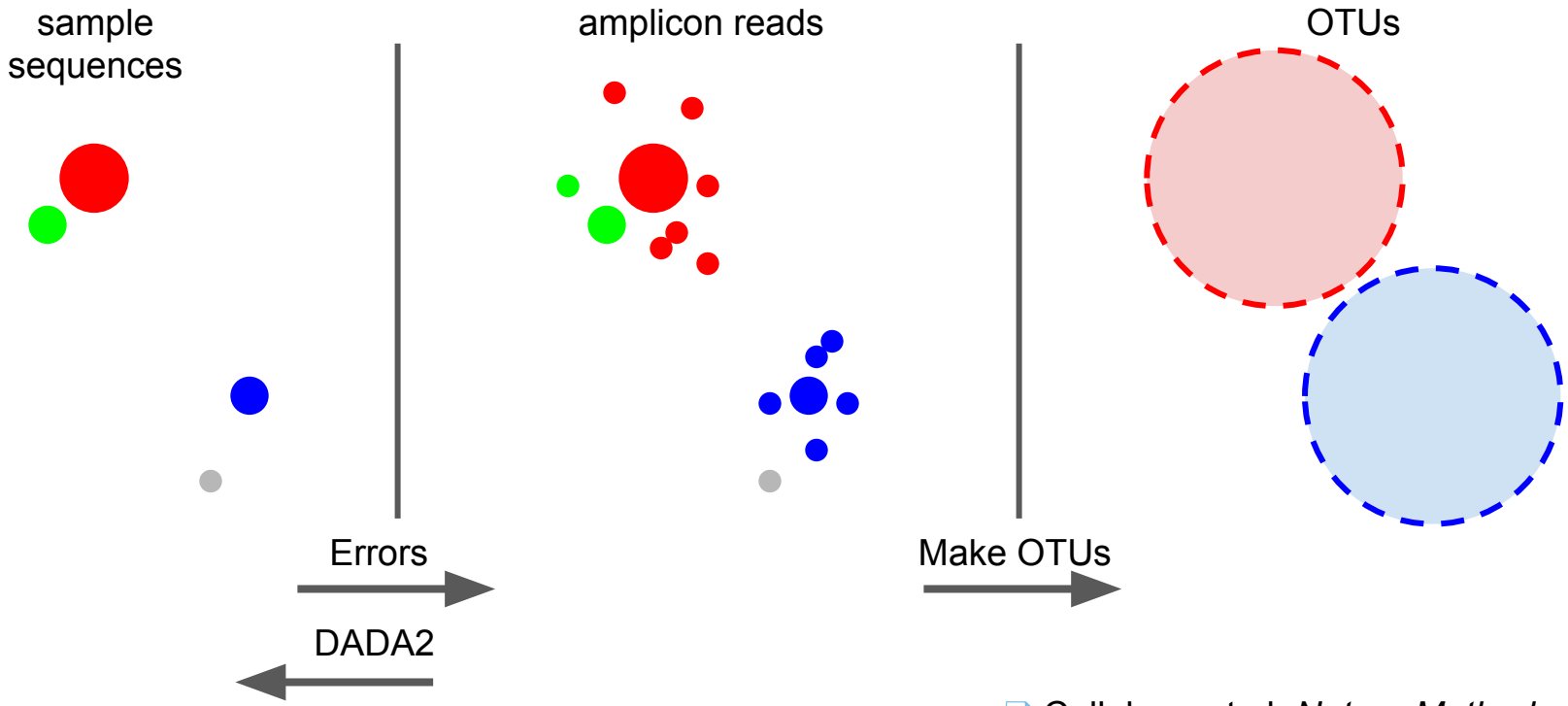
- Clustering
  - Remove noisy sequences and reduce the amount of sequences to process
  - Works based on a given threshold, i.e. 97% similarity but others exist like Oligotyping
  - There are different methods (closed or open reference) and algorithms (`sortmerna`, `vclust`)

- Remove noise
  - Find the cleanest sequence
  - Correct and/or discard super noisy sequences
  - Examples are: DADA2 and Deblur

Clustering methods **ideal** situation

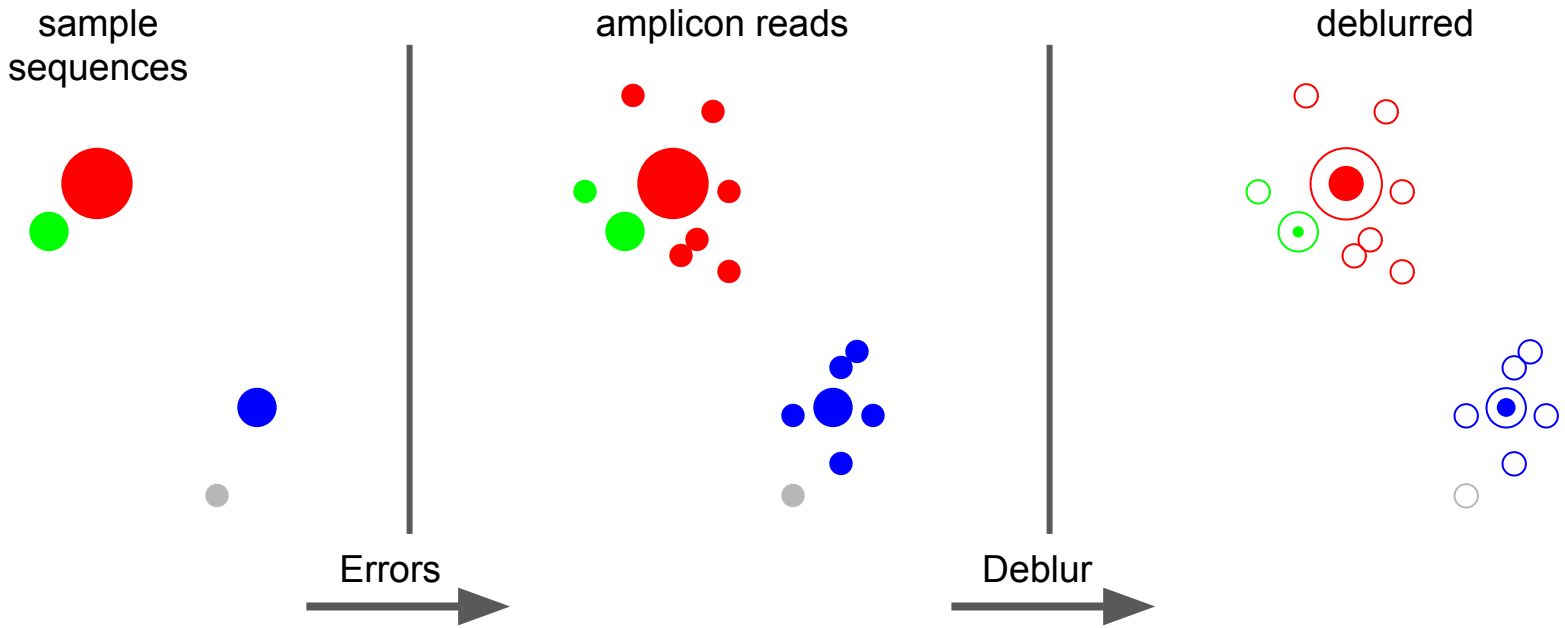


DADA2



Callahan, et al. *Nature Methods*, 2016

Deblur

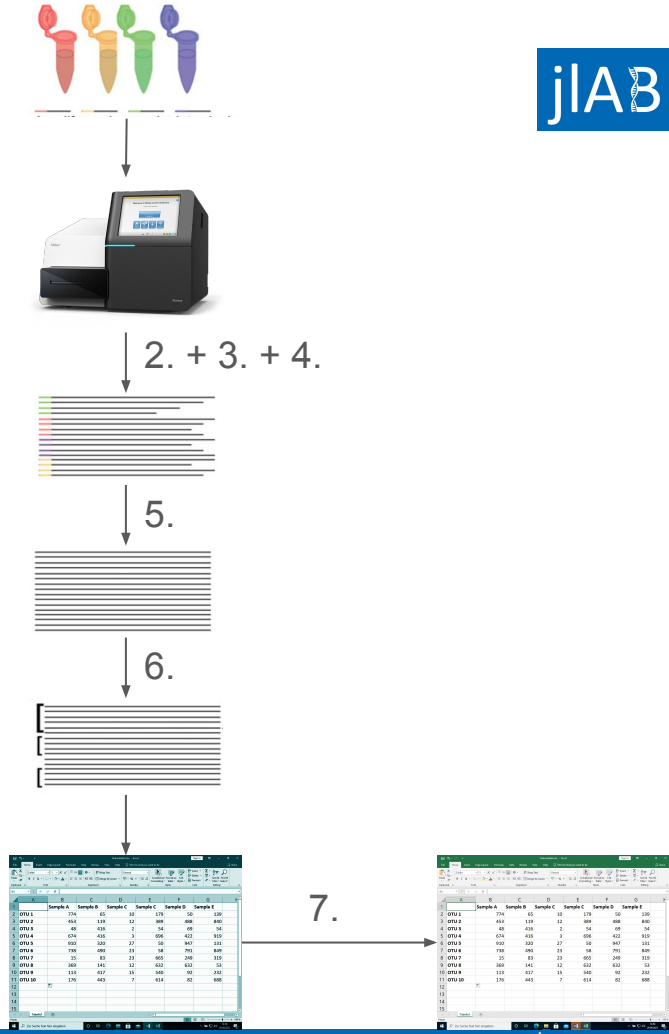


Amir, et al. *mSystems*, 2017

Content

1. Scope ✔
2. Base Calling ✔
3. Demultiplexing ✔
4. Quality Control ✔
5. Adapter Trimming, Clipping ✔
6. "OTU" picking ✔

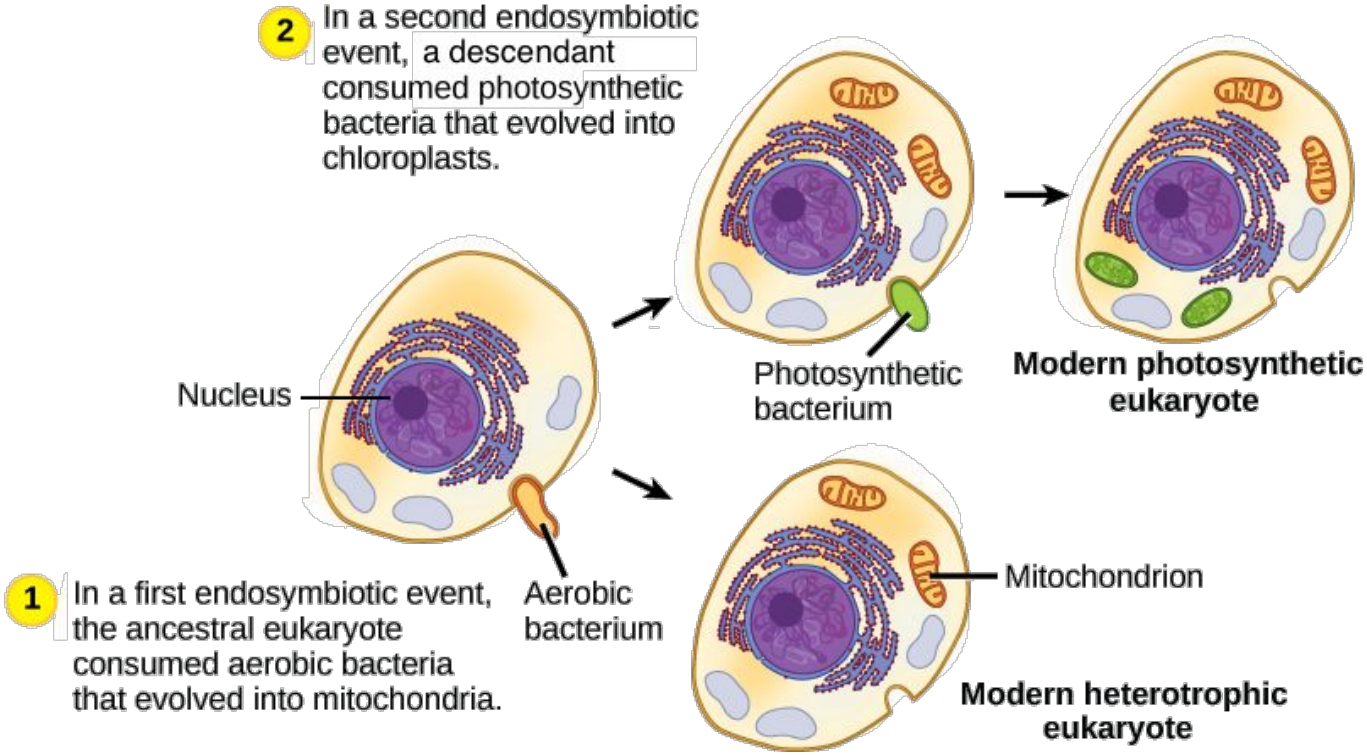
7. Data Normalization
   a. Contamination Removal
   b. very low abundant "OTU" removal
   c. Rarefaction



7. Data Normalization

a. Contamination Removal
b. very low abundant "OTU" removal
c. Rarefaction

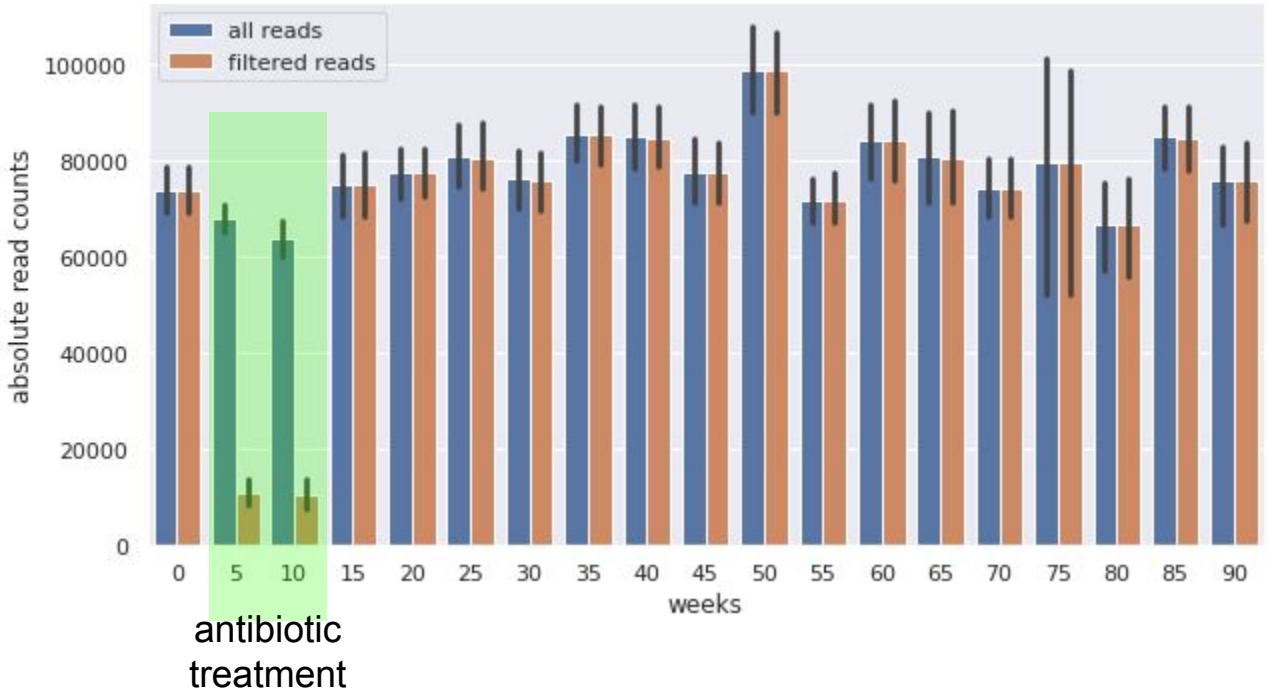
7a. Contamination Removal



https://www.pngfind.com/download/hTbiJox_drawing-cell-endoplasmic-reticulum-endosymbiotic-theory-flowchart-hd/

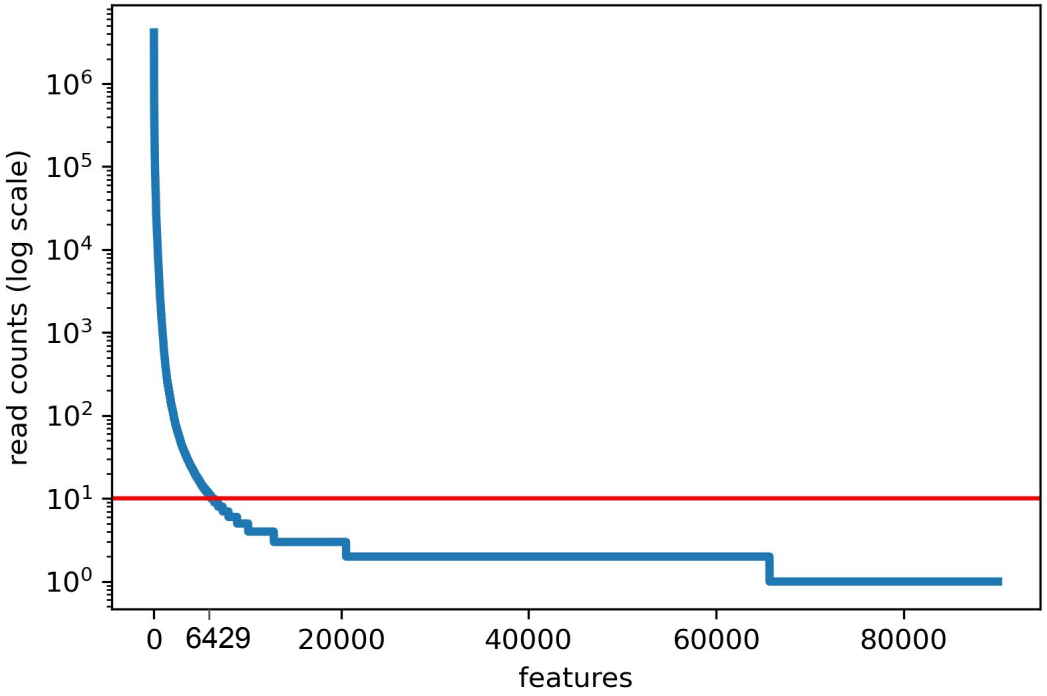
7a. Contamination Removal

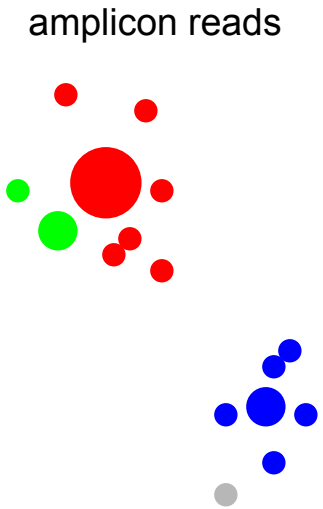
mouse gut microbiome




```
apply(lambda lineage: 'c__Chloroplast' in lineage or 'f__mitochondria' in lineage)
```

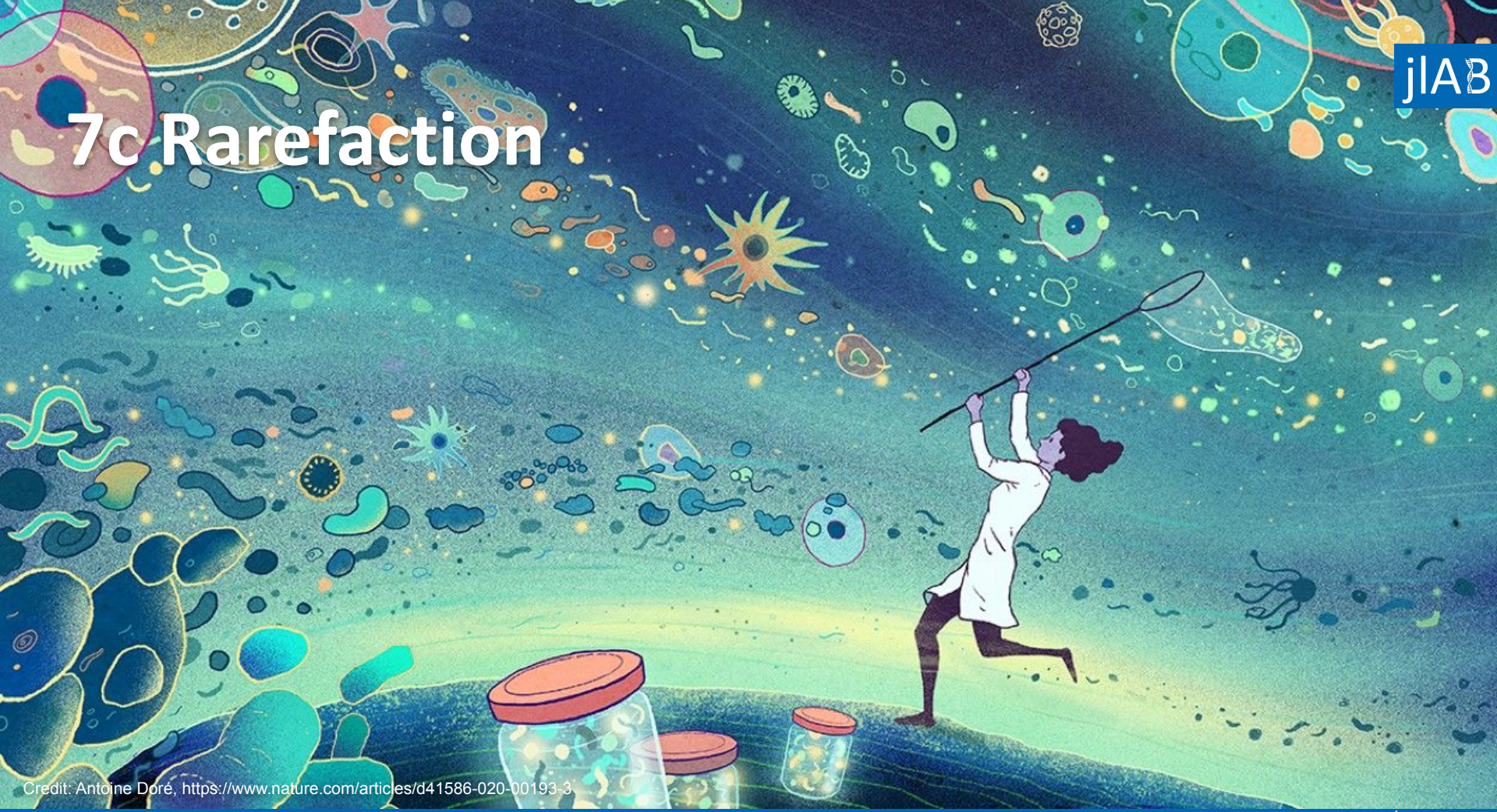
7b. very low abundant "OTU" removal





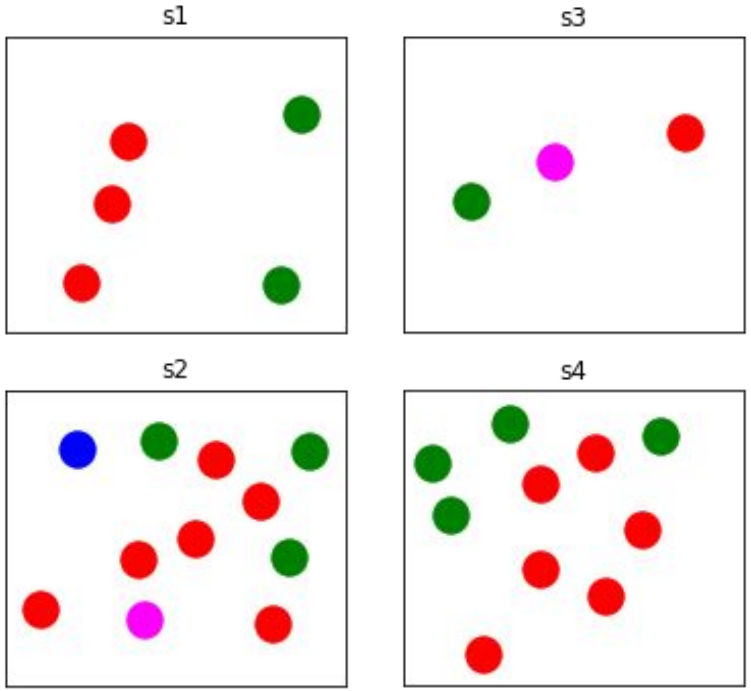

```
counts = counts[counts.sum(axis='features') >= 10]
```

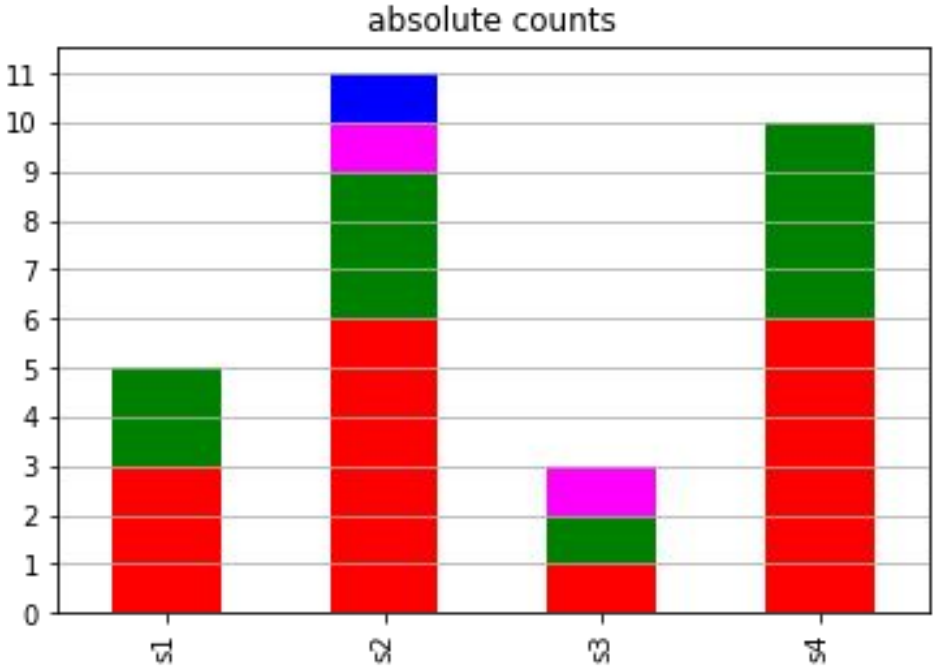
7c Rarefaction



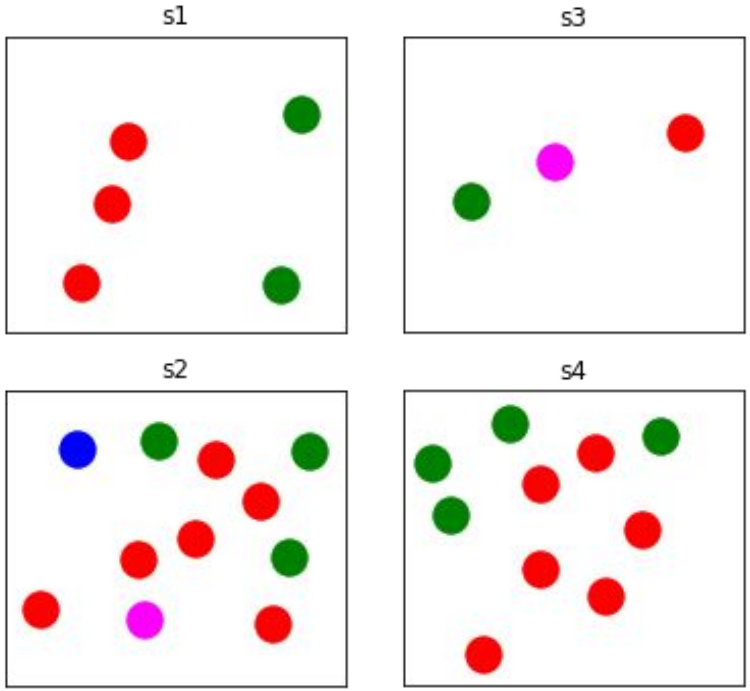
Credit: Antoine Doré, https://www.nature.com/articles/d41586-020-00193-3

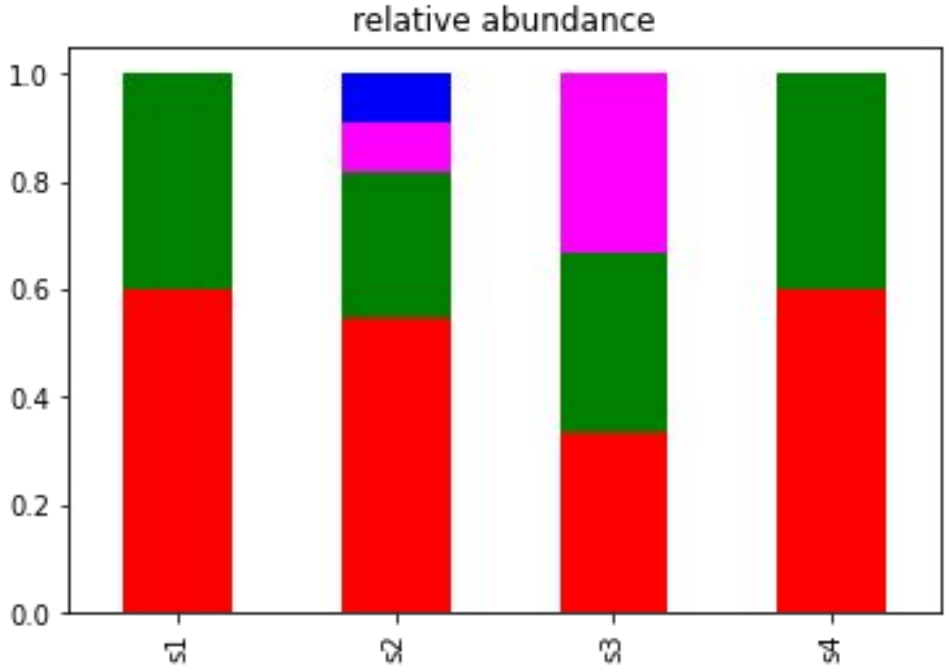
7c Rarefaction



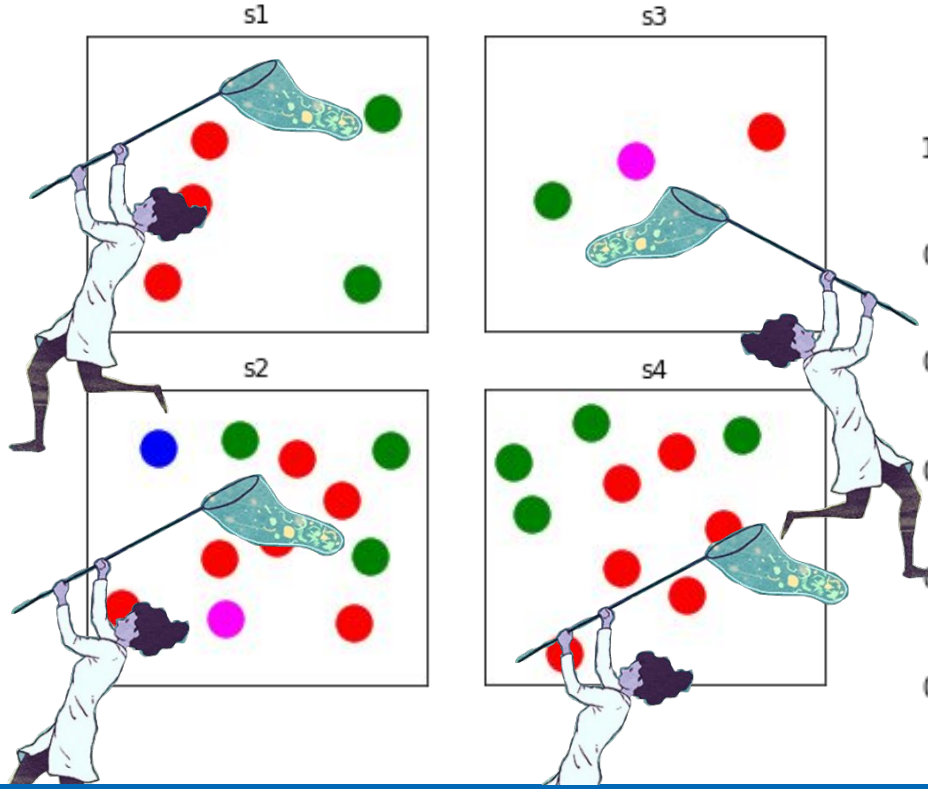


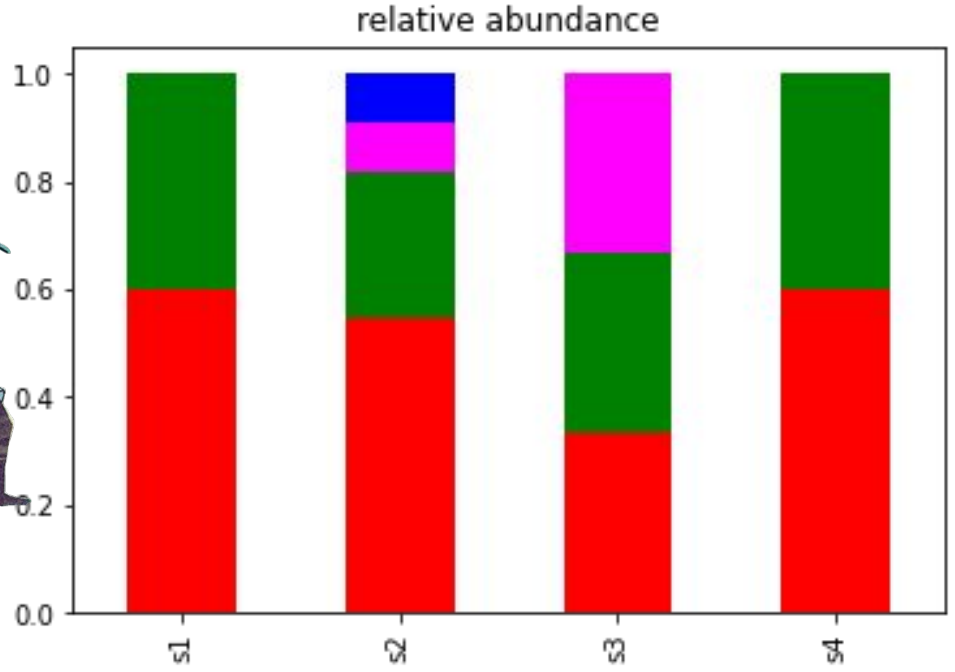
7c Rarefaction



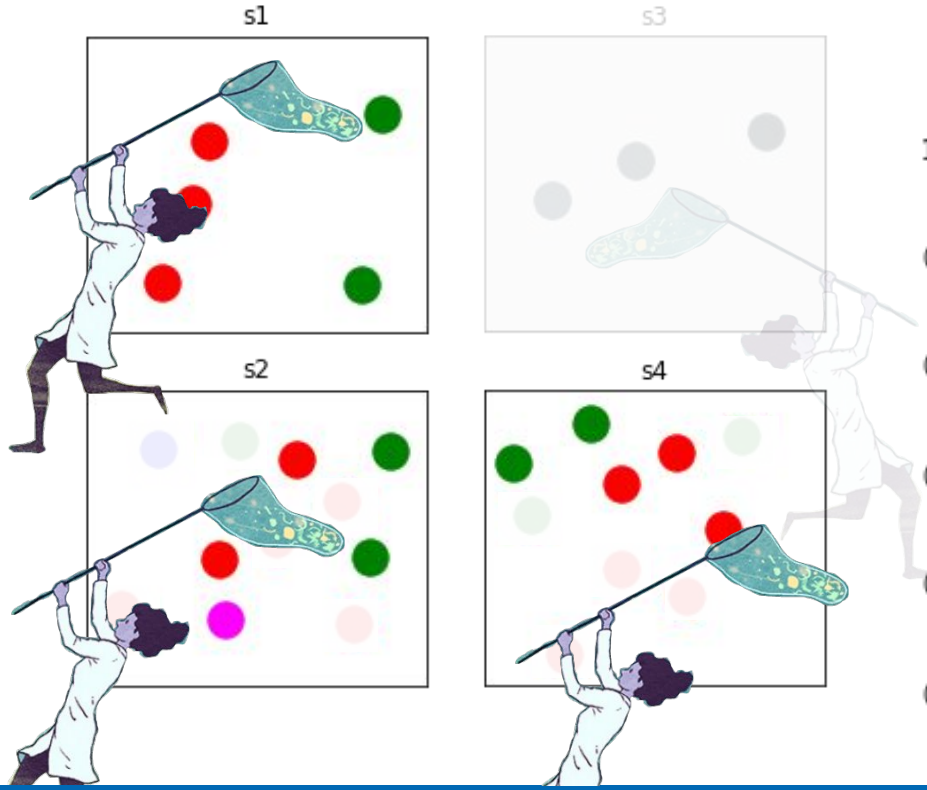


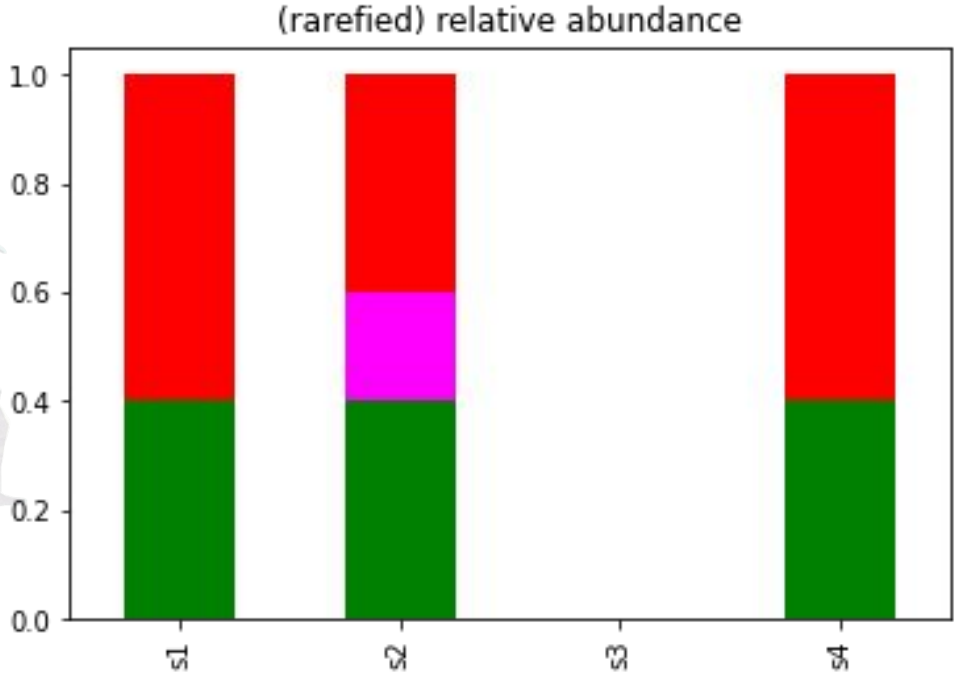
7c Rarefaction: Size of net = rarefaction depth?



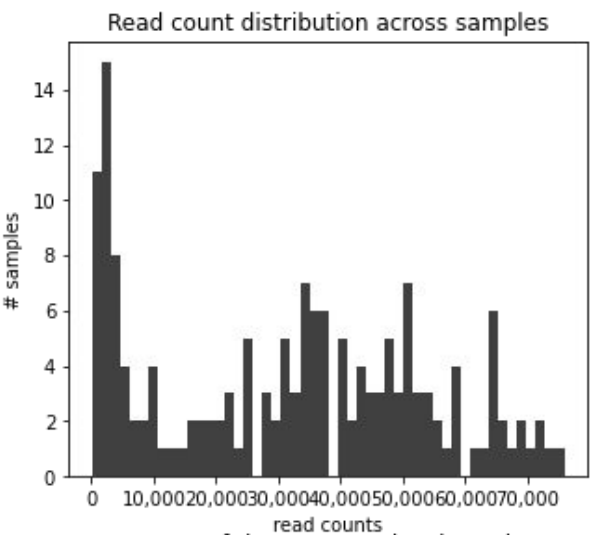


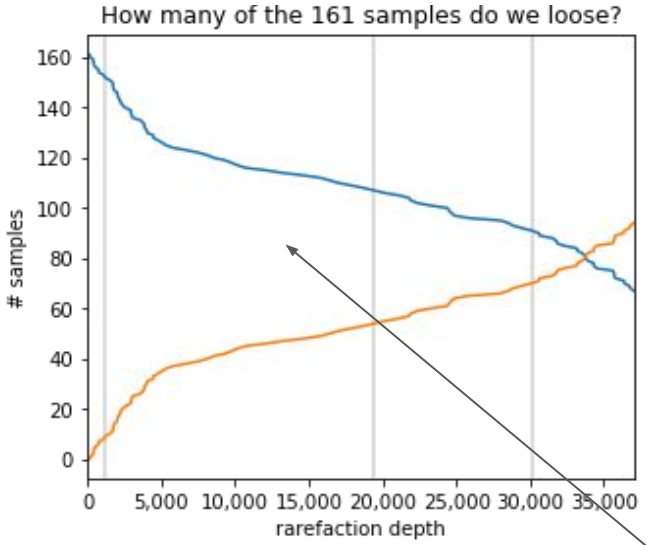
7c Rarefaction: Size of net = rarefaction depth = 5

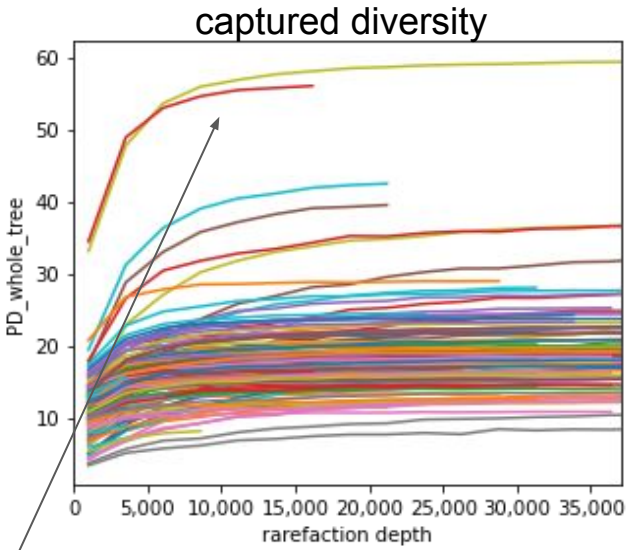




7c Rarefaction: depth

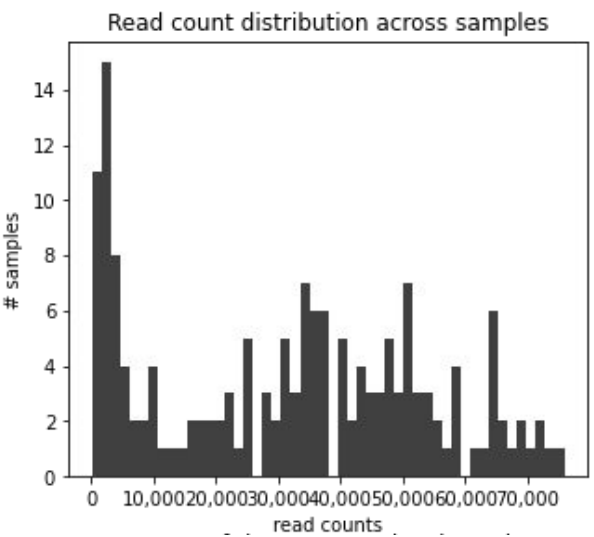


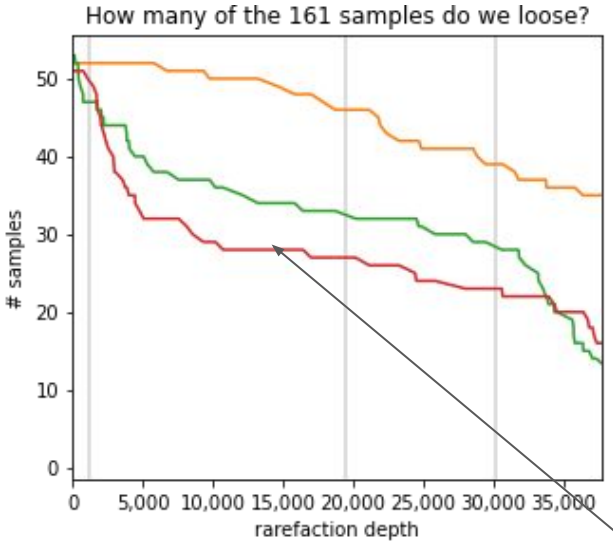


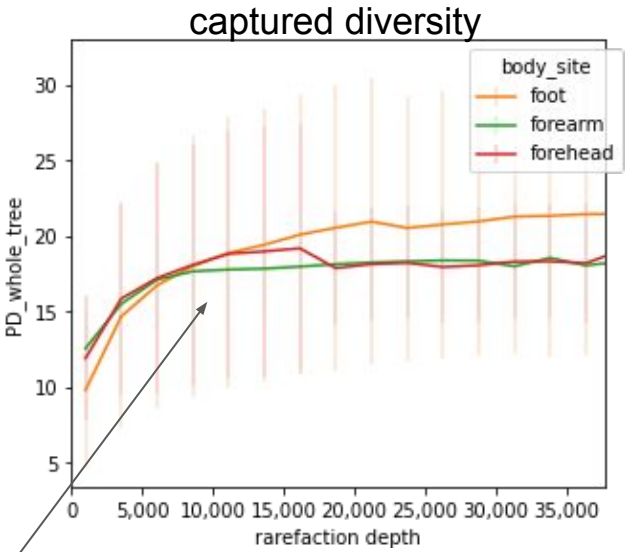


trade off

7c Rarefaction: depth





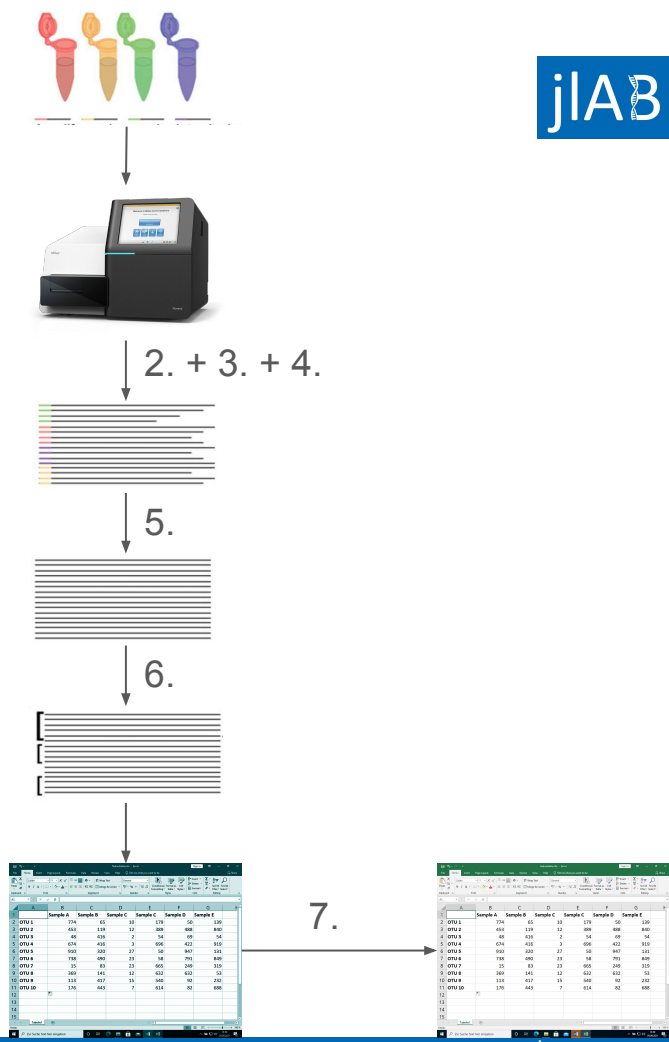


trade off

Summary

1. Scope ✔
2. Base Calling ✔ `bcl2fastq`
3. Demultiplexing ✔ `bcl2fastq`
4. Quality Control ✔ `fastp`
5. Adapter Trimming, Clipping ✔ `fastp`
6. "OTU" picking ✔ `DADA2/Deblur` via `Qiime2.org`

7. Data Normalization ✔
   a. Contamination Removal ✔ taxonomic assignments -> later
   b. very low abundant "OTU" ✔ removal `pandas`
   c. Rarefaction ✔ `Qiime2.org`



	Tuesday, Oct 8th	Wednesday, Oct 9th	Thursday, Oct 10th
10:00-10:30		QIIME2 and Qiita	Buffer for Hands-On, Coffee break in-between
10:30-11:00		Qiita Hands-On	
11:00-11:30	Welcome & Intro		
11:30-12:00	From raw data to feature tables	Coffee Break (~15min)	Coffee Break (~15min)
12:00-12:30		Qiita Hands-On	Buffer for Hands-On
12:30-13:00		Lunch Break	Lunch break and Closing remarks
13:00-13:30	Lunch Break		
13:30-14:00		QIIME2 Hands-On	
14:00-14:30	Sequence quality control		
14:30-15:00	Coffee Break	Coffee Break	
15:00-15:30	QC Hands-on	Diversity Calculation	
15:30-16:00		QIIME2 Hands-On (cont'd.)	
16:00-16:30			

JHaaS

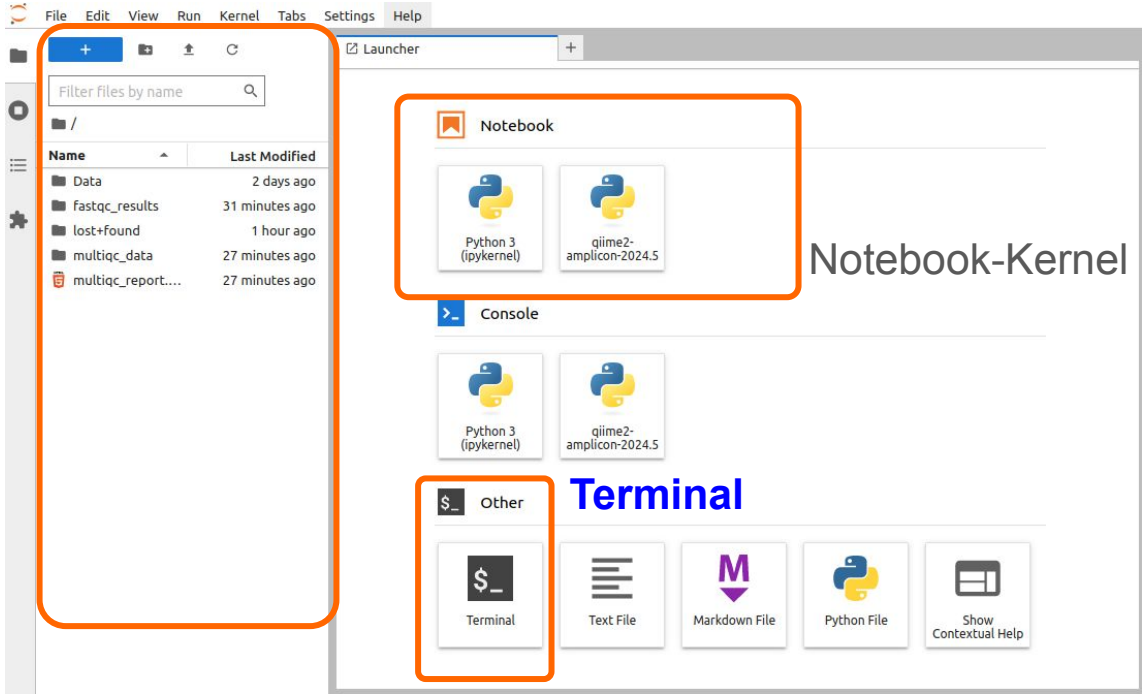
- Please login to https://jhaas.gi.denbi.de/ and request access to https://jhaas.gi.denbi.de/participation/participate/metagenomics2024

- Wait until you are verified!

Hands-On Quality Control

Short JupyterHub Introduction I

File Overview



Notebook-Kernel

Fastqc and Multiqc Hands-On I

1: **Download and unzip the data**

```
wget -qO- http://minio-seed-s3-storage:9000/metagenomics2024/compressed_metagenomics.tar.gz | tar xvz
```

2: **Create Folder for Fastqc Results**

```
mkdir fastqc_results
```

3: **Activate conda environment**

```
source /opt/conda/bin/activate && conda activate qualitycontrol
```

Fastqc and Multiqc Hands-On II

4: **Run fastqc on sequence data**

```
fastqc -o fastqc_results/
Data/sequence_data/pax/study_raw_data_11758_092524-041740/per_sample_FASTQ/6
7800/*.fastq.gz
```

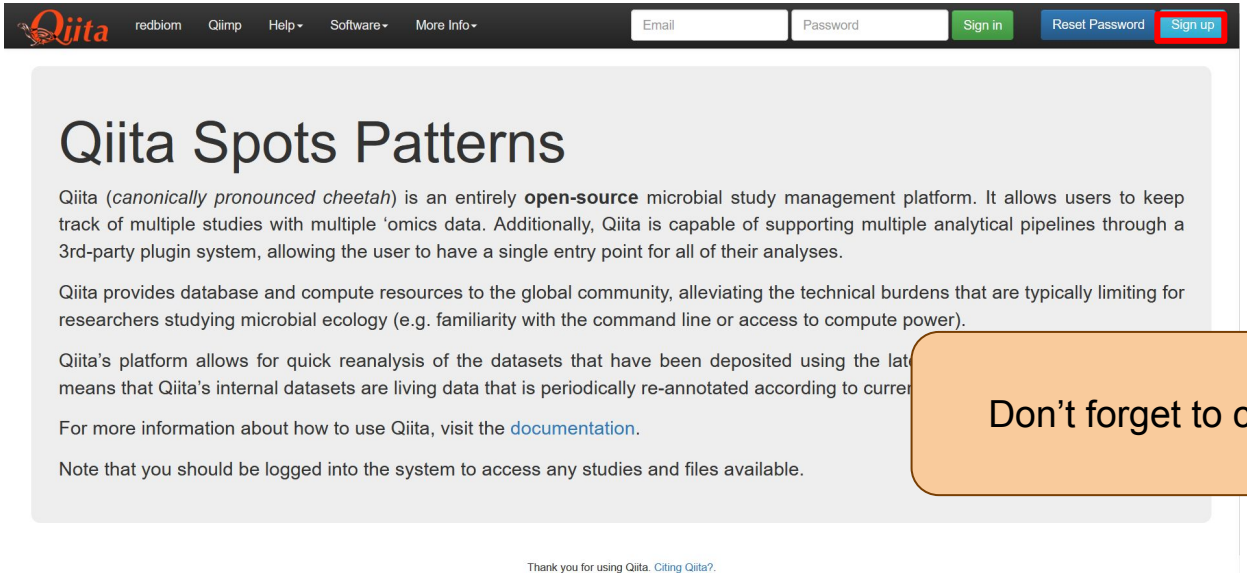
6: **Run Multiqc there**

```
multiqc fastqc_results/.
```

7: **Look for multiqc_report.html file, download it, open in your Browser**

Login and Account for Qiita

For tomorrow's Hands-On session, please create an account at https://qiita.ucsd.edu



Don't forget to confirm your e-mail address!