Metagenomics, Day 3, Morning: Sampling & Presence/Absence

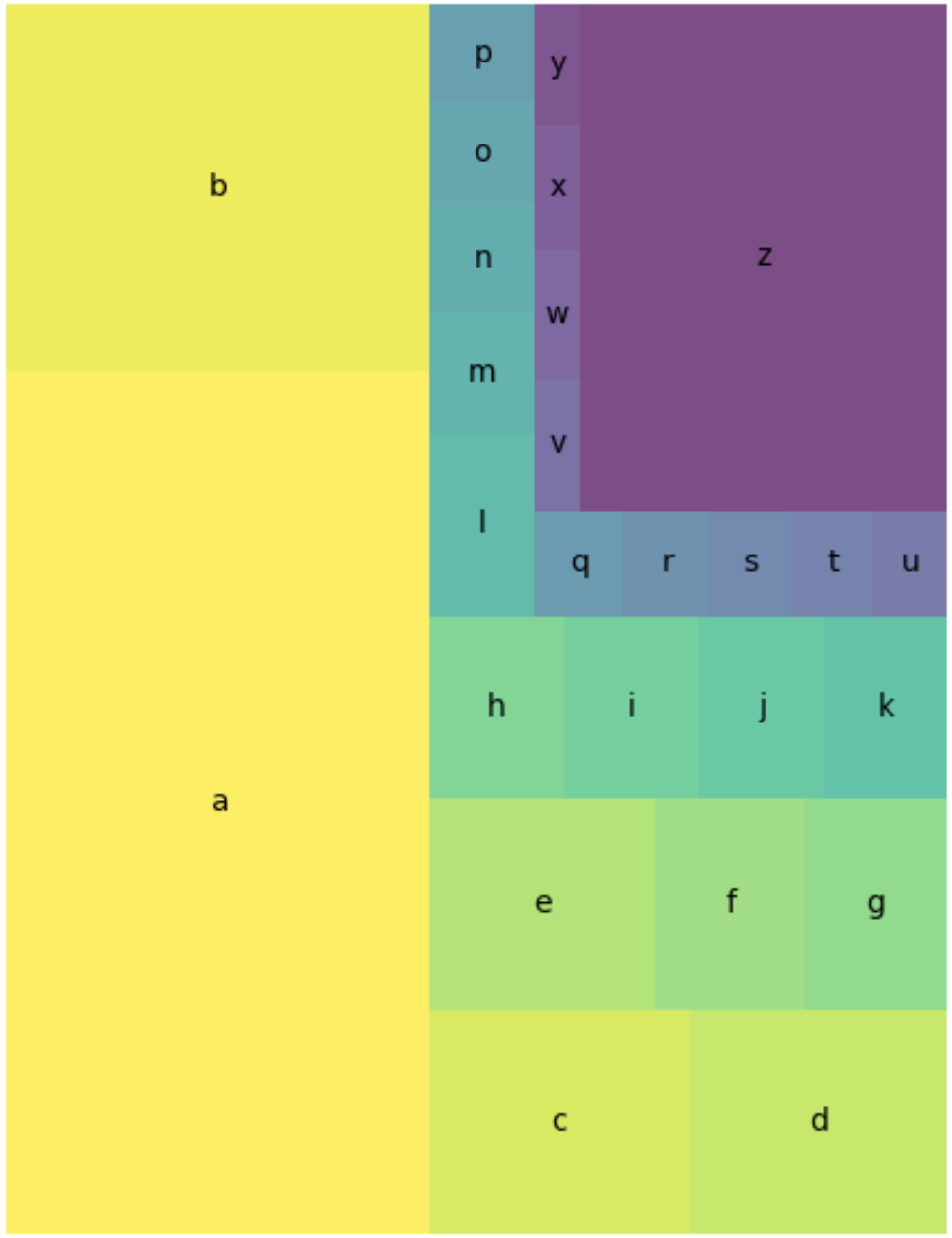
Titus Brown

July 17, 2025

STAMPS 2025

This is an proportional representation of the content of a metagenome, as estimated by sourmash.

The sizes are ”accurate” in that they are visually proportional.

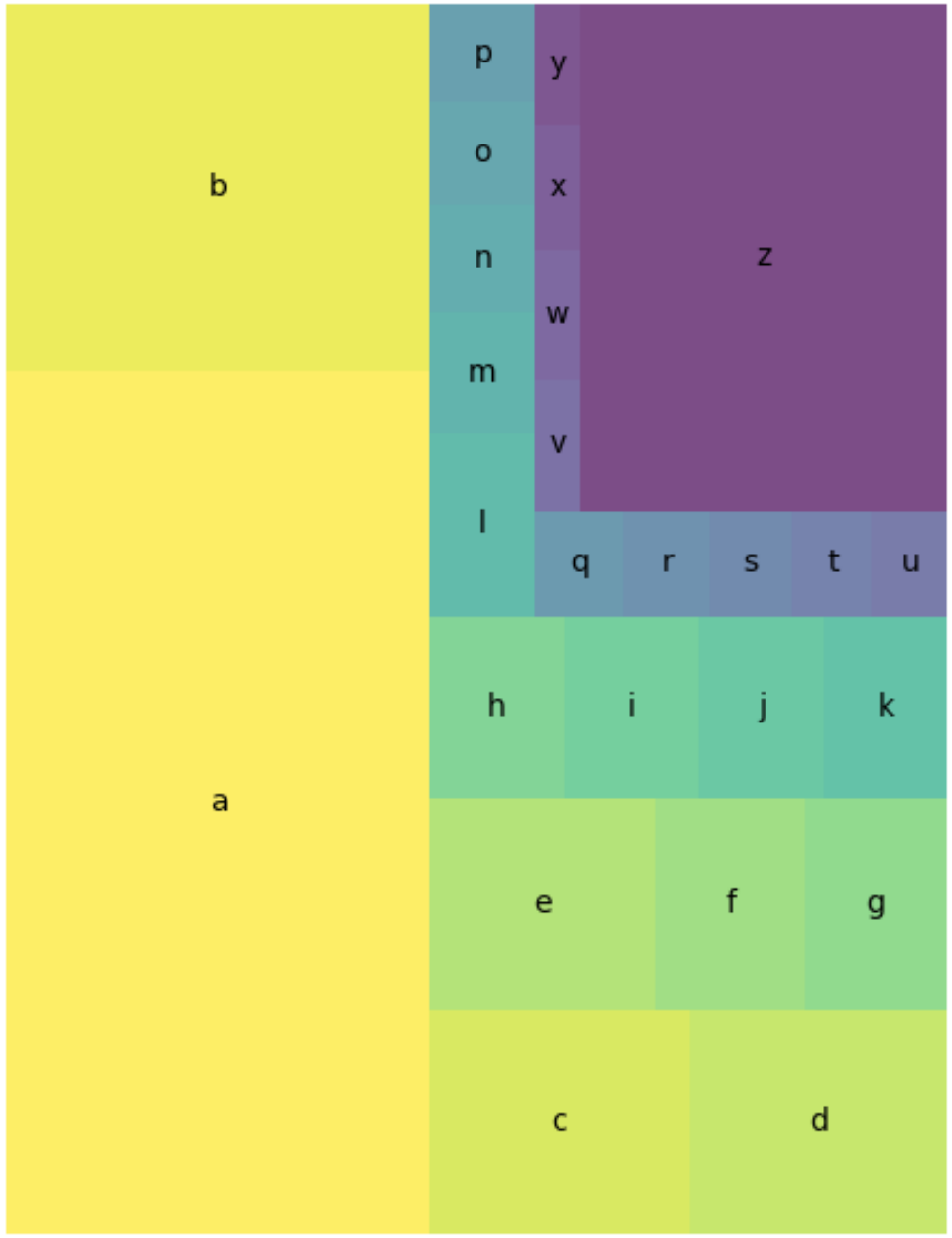


a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

The size of each box is, essentially:

size of genome
x
abundance of genome

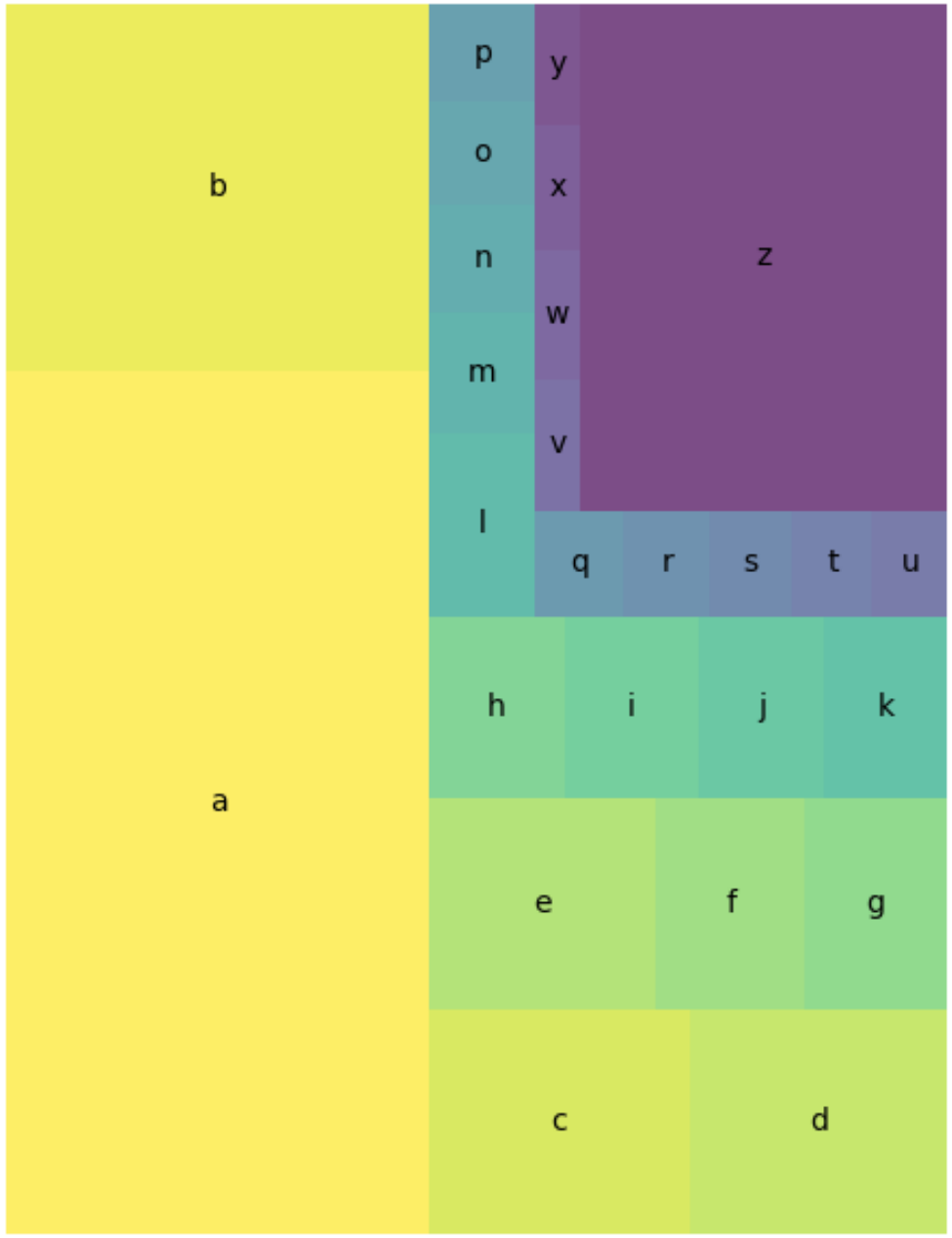
(The actual bp numbers are available, too; ask me when we get to the command line!)



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

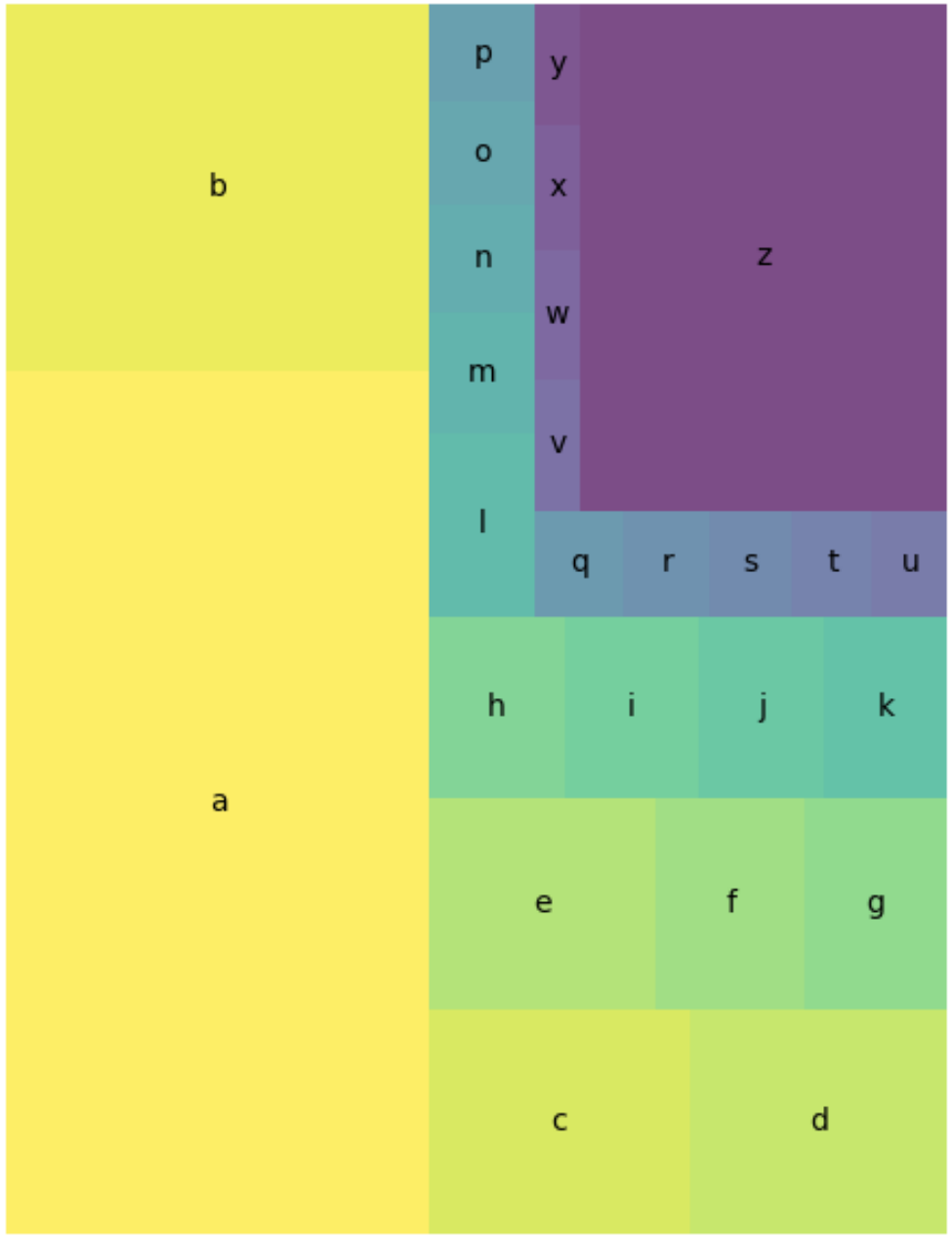
Think of this as a *dartboard*, and of sequencing as throwing a dart at this map.

Throwing more darts is sequencing more deeply.



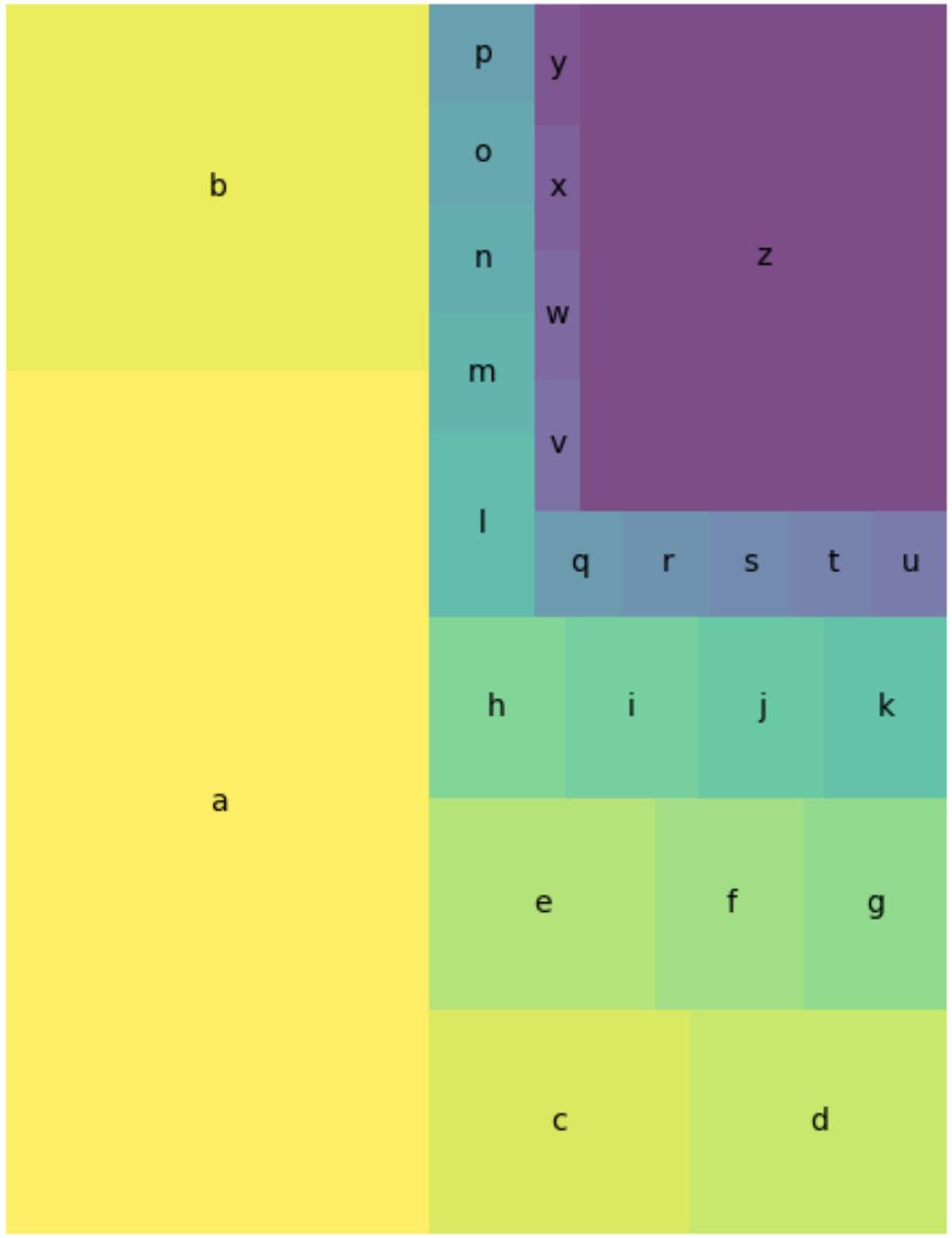
a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

Q1: What will you mostly get if you throw twice as many darts at this dart board?



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

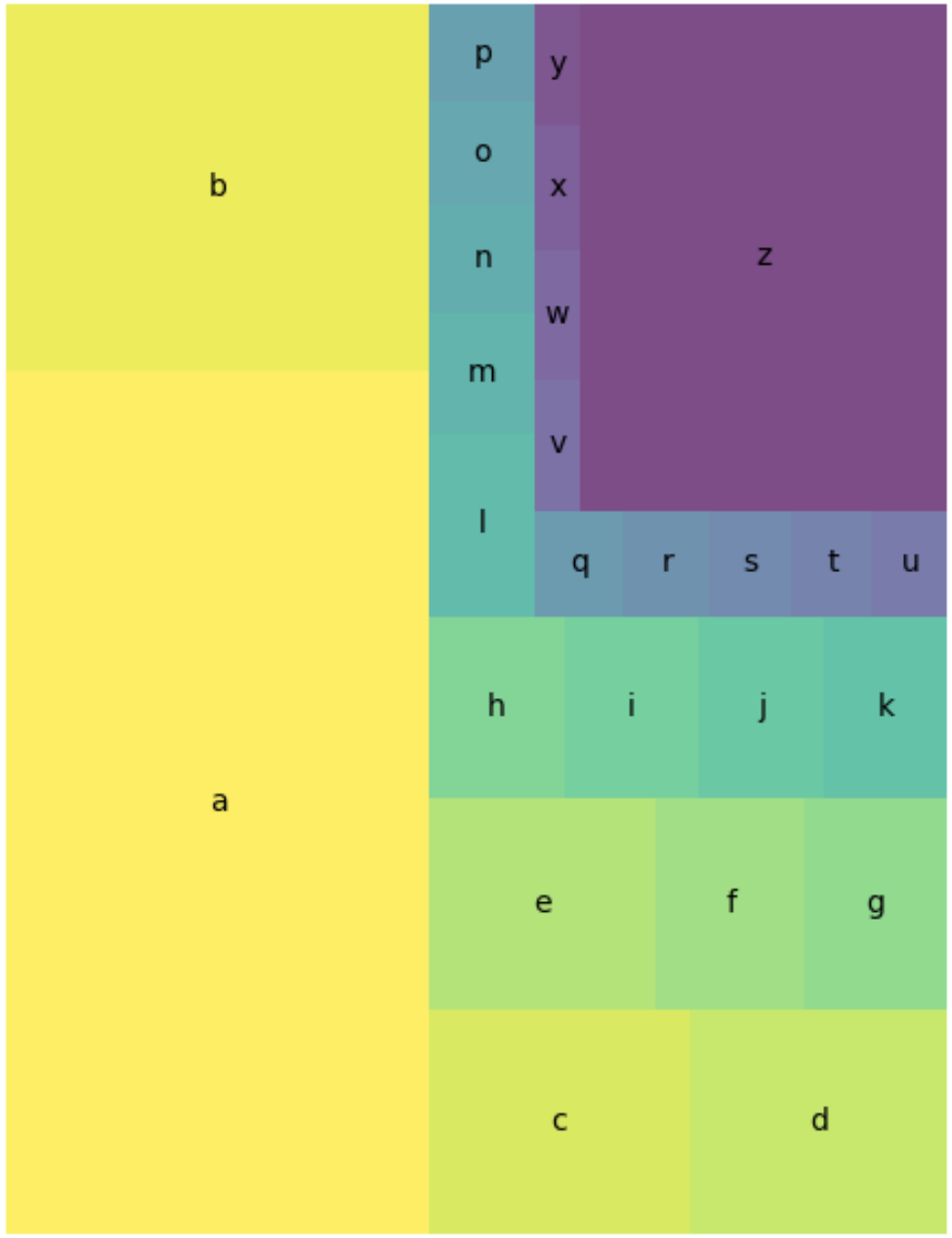
Q2: What happens to the abundances (not displayed ;) if you throw twice as many darts at this dart board?



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

Let’s suppose you can select darts with different size tips.

Q3: How does this change your dart throwing strategy?



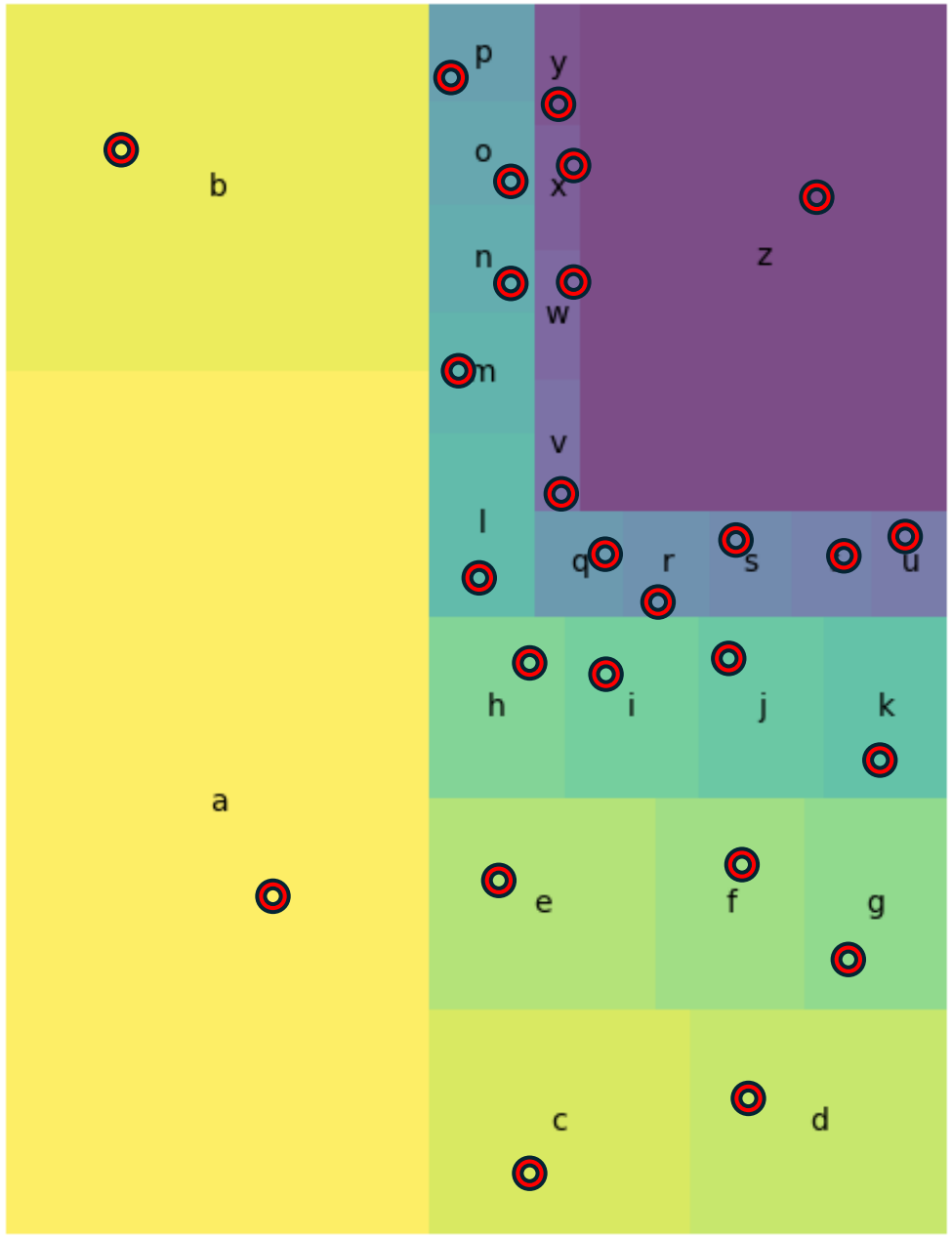
a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

(linkage & sequencing read length)

Now let's change the darts to be strongly magnetic, so that they ONLY hit the red circles.

Q4: How does ***this*** change your dart throwing strategy?

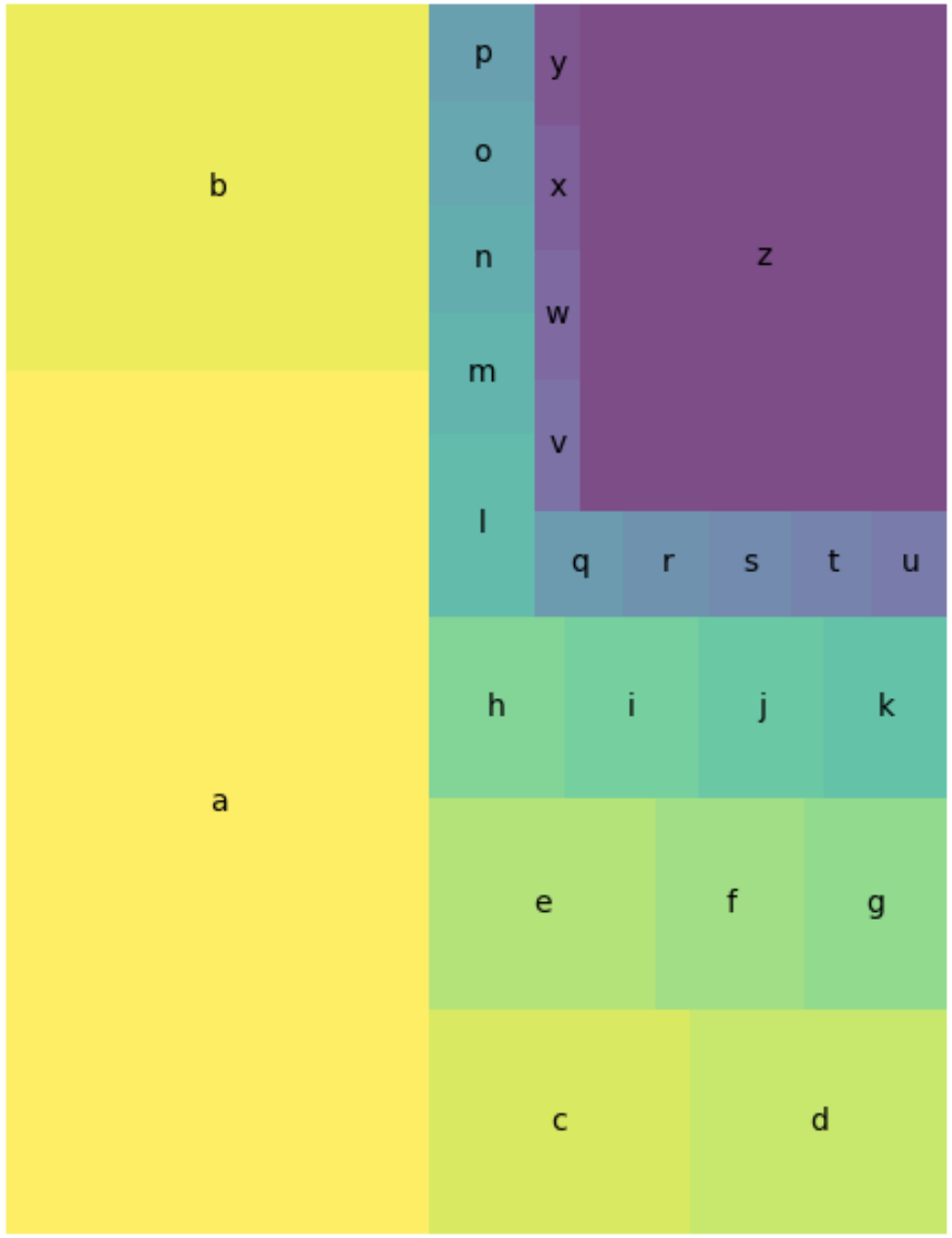
(This is a 16S/amplicon analogy ☺)



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

Q5: what is the unclassified stuff?

(What *could* it be?)



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

The size of each box is, essentially:

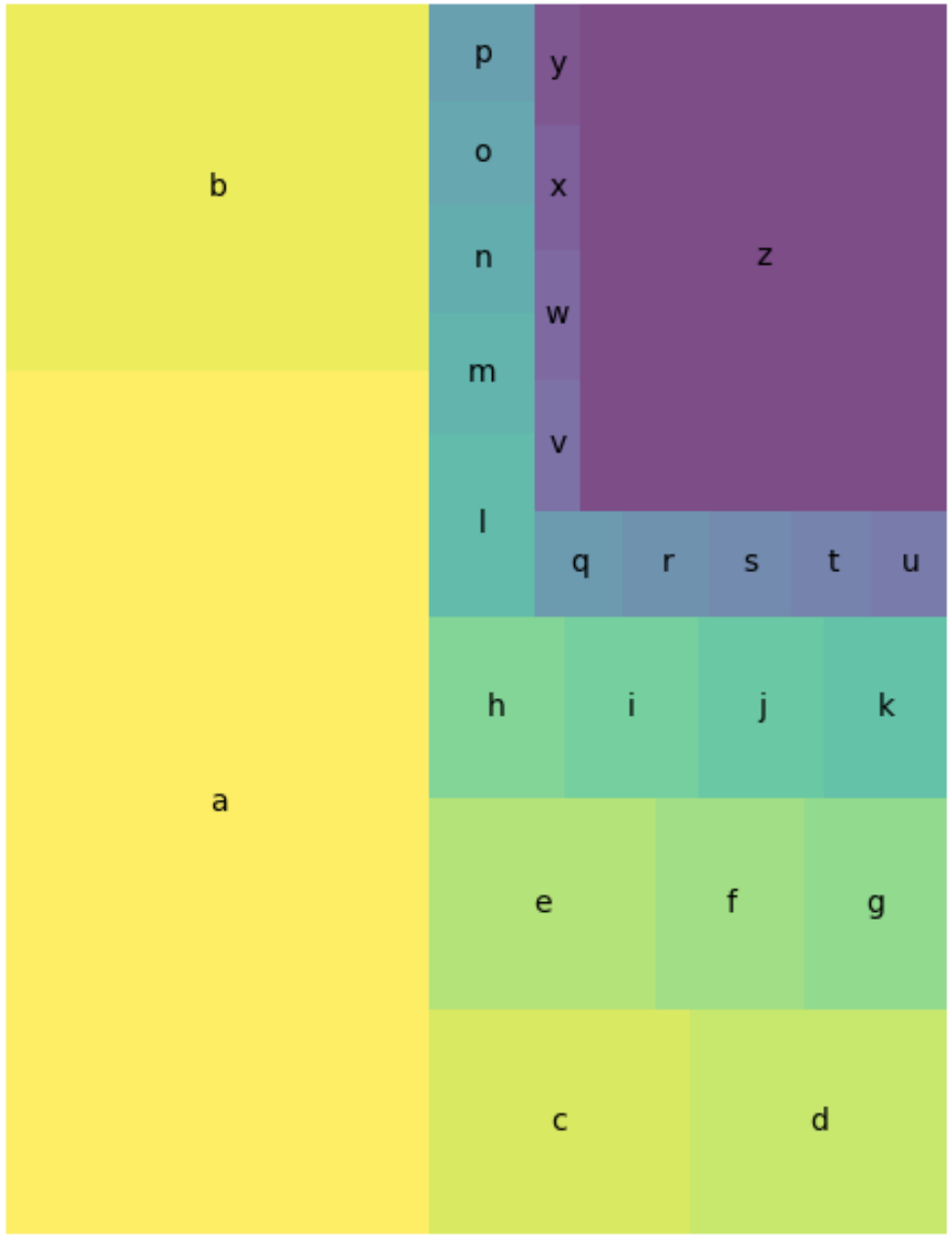
size of genome x abundance of genome

Bacteria: ~5 million base pairs.

Eukaryotes: 100 million-5 billion base pairs (20-1000x bigger)

Viruses: 10,000-50,000 base pairs. (100x *smaller)*

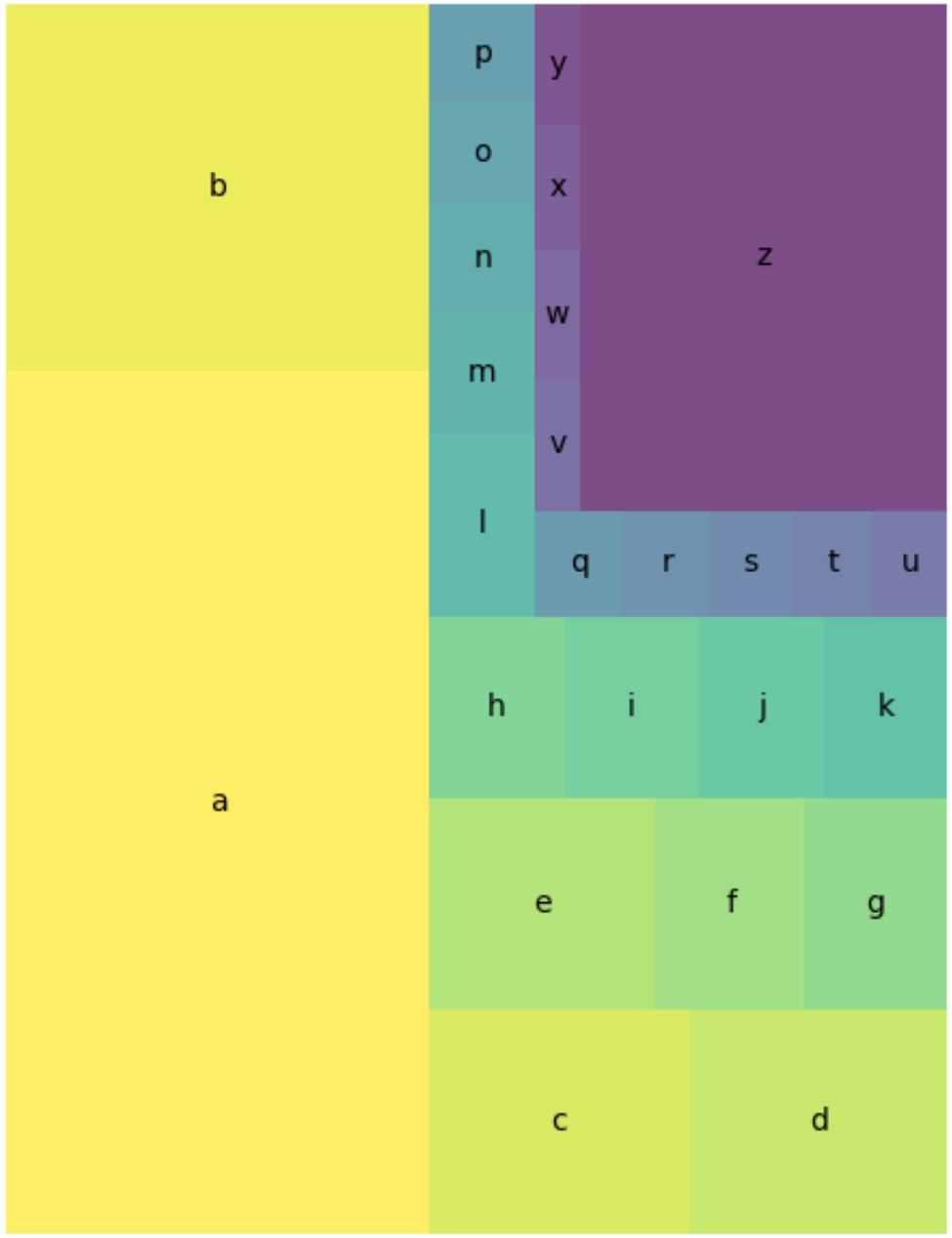
Q6: what would viruses look like on here?



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)

Q7: What would rarefaction be doing in this situation?

(We'll talk about this more on Monday… stats day!)



a: unclassified (31.5%)
b: s__Limousia pullorum (13.5%)
c: s__Mogibacterium_A kristiansenii (5.1%)
d: s__Limosilactobacillus suis (5.0%)
e: s__Phil1 sp001940855 (4.1%)
f: s__Methanocatella smithii (2.7%)
g: Sus scrofa (2.6%)
h: s__Faecousia sp004556205 (2.1%)
i: s__Escherichia coli (2.1%)
j: s__Vescimonas sp945877515 (2.0%)
k: s__Lentihominibacter sp022775805 (1.9%)
l: s__Schaedlerella sp004556565 (1.7%)
m: s__Oliverpabstia sp004562915 (1.1%)
n: s__Enterococcus faecalis (1.0%)
o: s__Enterococcus_E cecorum (0.9%)
p: s__Fimenecus sp900766945 (0.9%)
q: s__F23-B02 sp004557075 (0.8%)
r: s__Sodaliphilus sp004557565 (0.8%)
s: s__Parabacteroides distasonis (0.8%)
t: s__CAG-1024 sp000432015 (0.7%)
u: s__SFEL01 sp004557245 (0.7%)
v: s__Prevotella sp000434975 (0.5%)
w: s__Faecousia sp022781505 (0.5%)
x: s__Limivicinus sp946061515 (0.5%)
y: s__UBA7173 sp001701135 (0.5%)
z: remaining taxa (16.1%)