

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

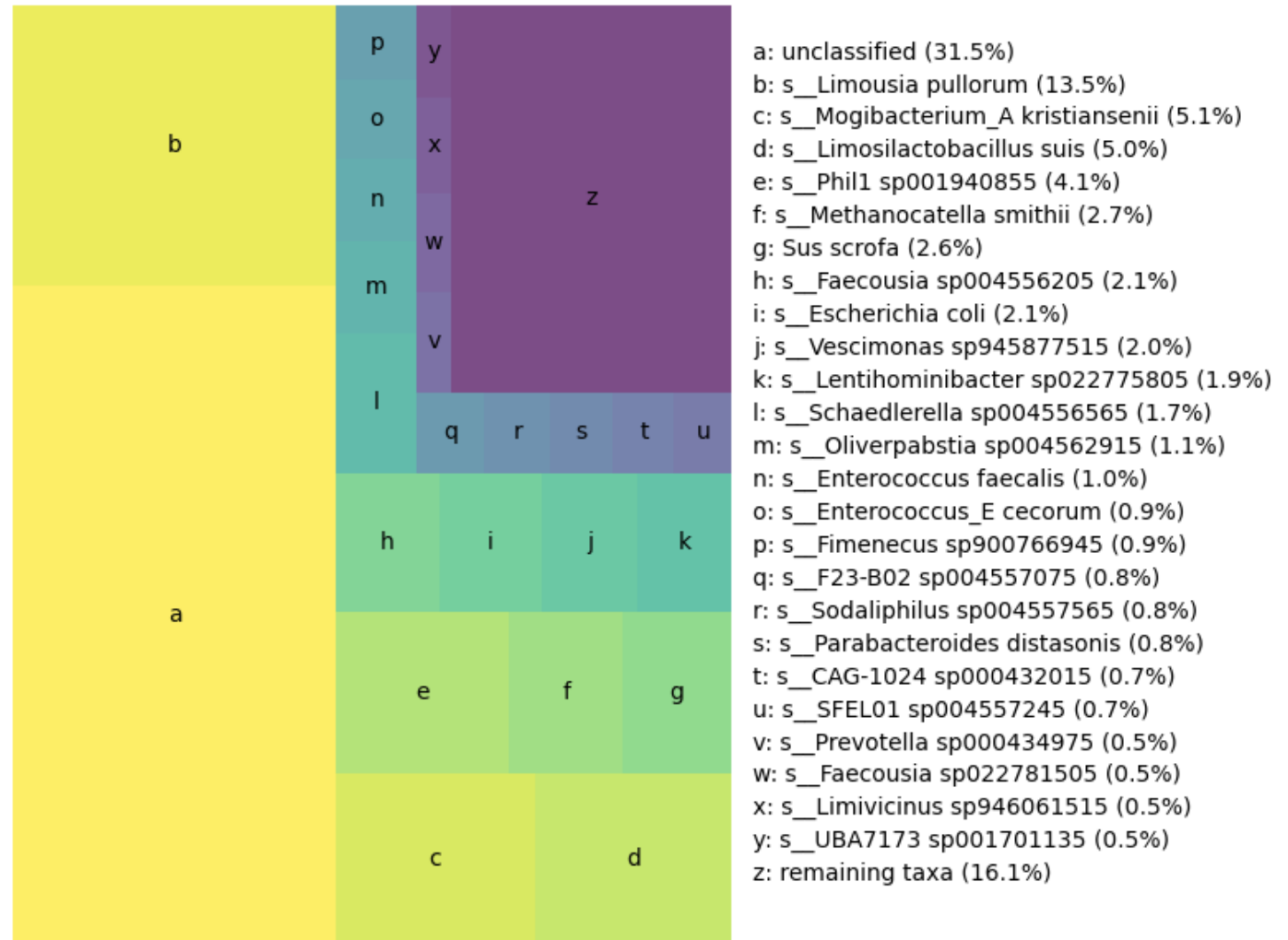
Titus Brown

July 17, 2025

STAMPS 2025

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

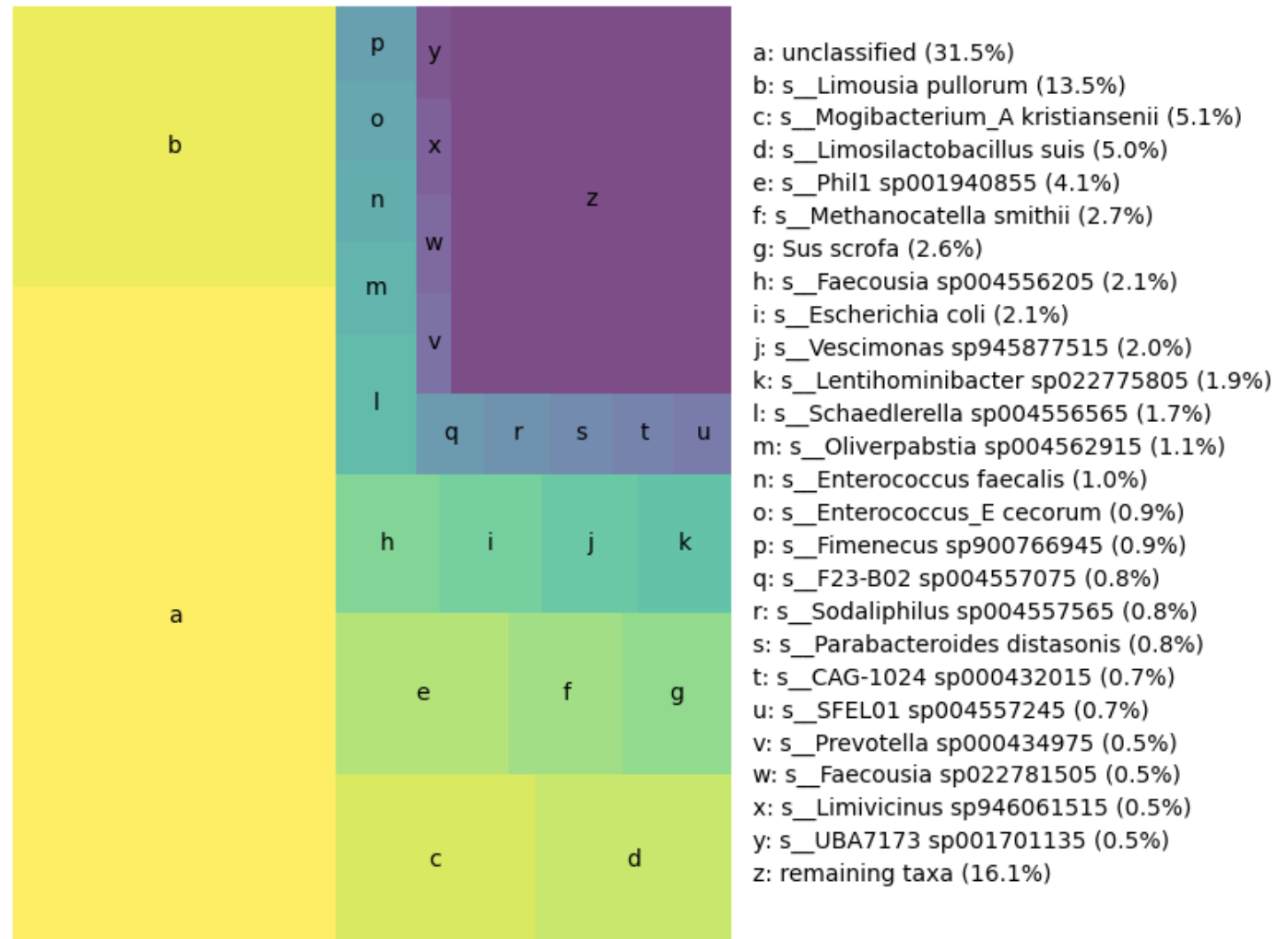
The sizes are  
"accurate" in  
that they are  
visually  
proportional.



The size of  
each box is,  
essentially:

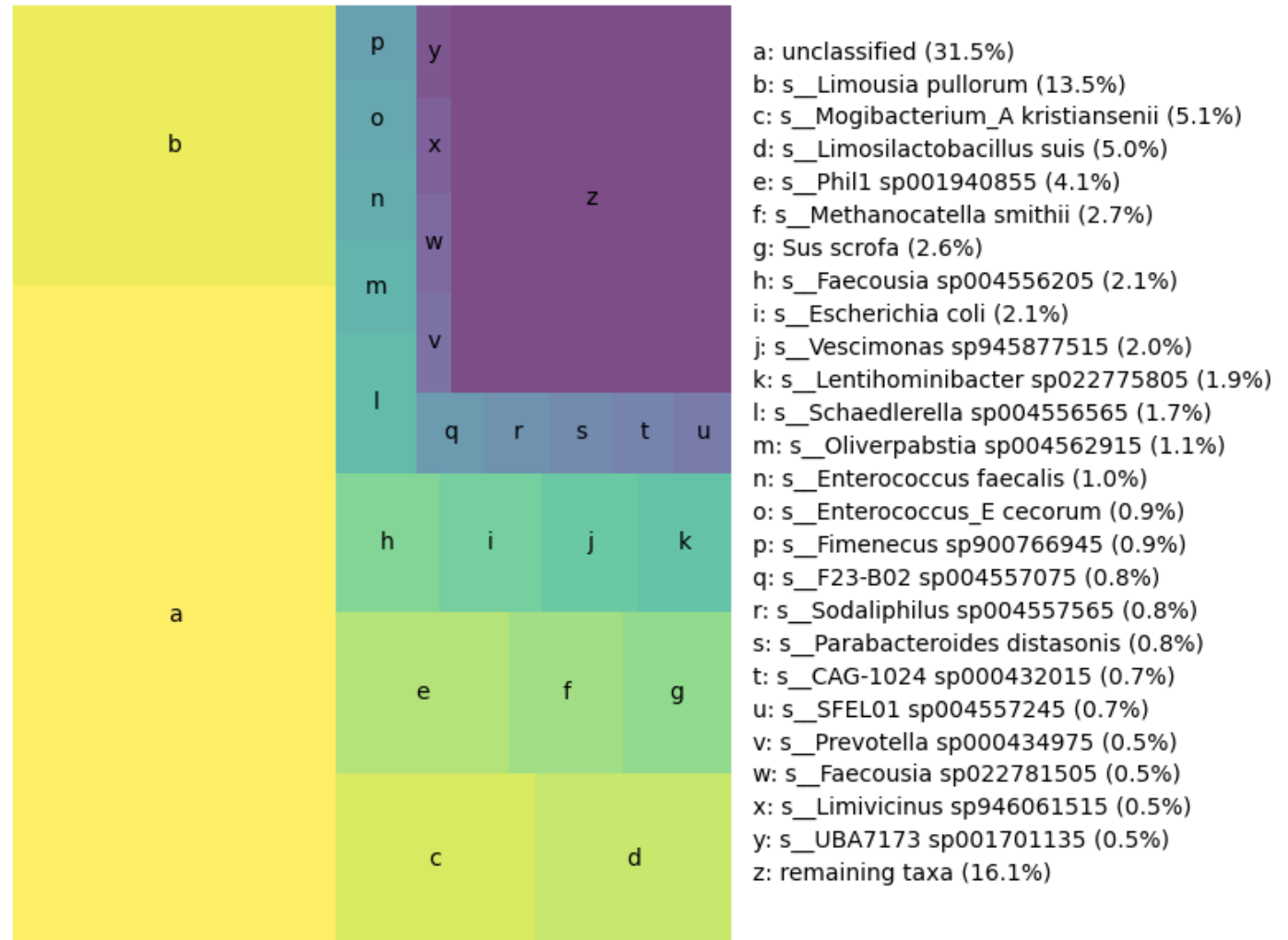
size of genome  
x  
abundance of  
genome

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

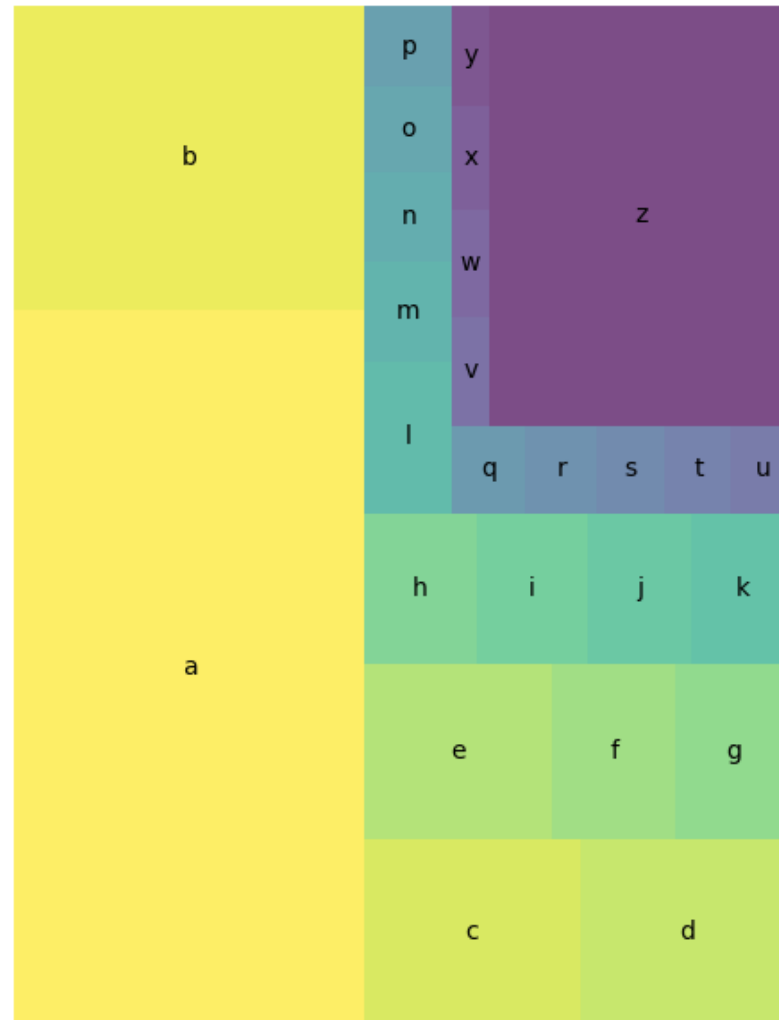


Think of this as a *dartboard*, and of sequencing as throwing a dart at this map.

Throwing more darts is sequencing more deeply.

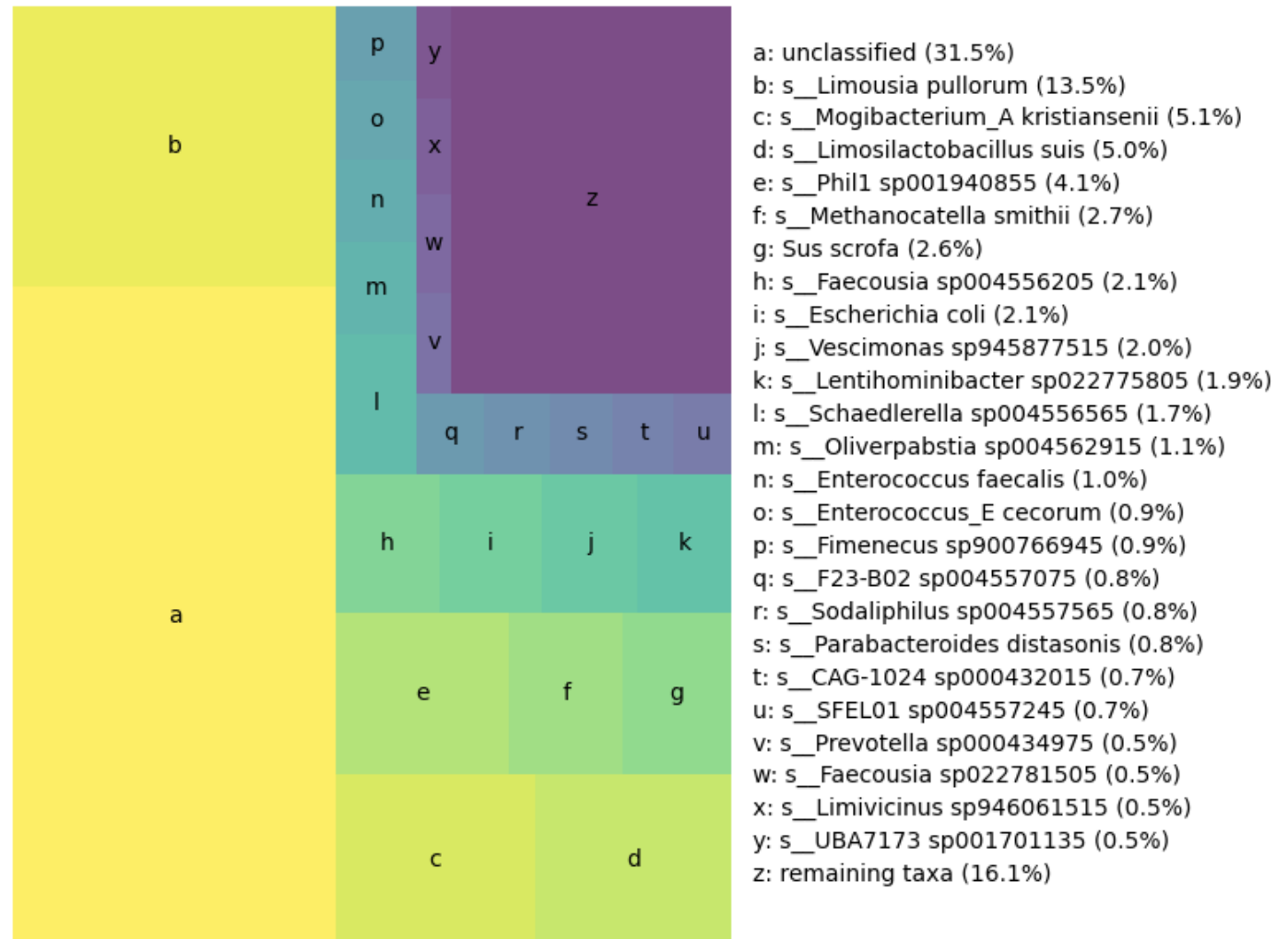


Q1: What will you mostly get if you throw twice as many darts at this dart board?



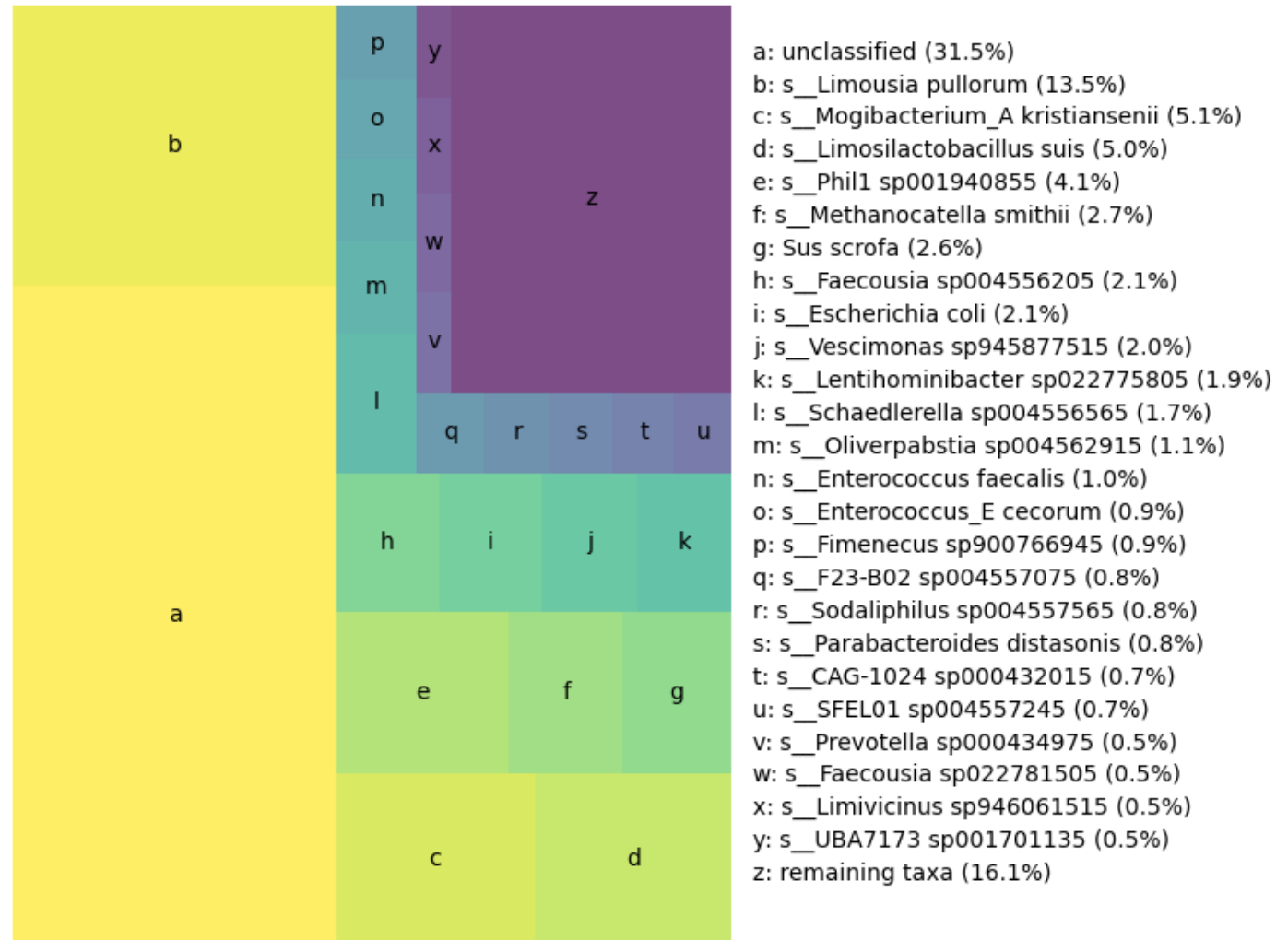
a: unclassified (31.5%)  
b: s\_\_Limousia pullorum (13.5%)  
c: s\_\_Mogibacterium\_A kristiansenii (5.1%)  
d: s\_\_Limosilactobacillus suis (5.0%)  
e: s\_\_Phil1 sp001940855 (4.1%)  
f: s\_\_Methanocatella smithii (2.7%)  
g: Sus scrofa (2.6%)  
h: s\_\_Faecousia sp004556205 (2.1%)  
i: s\_\_Escherichia 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\_\_Limivacinus 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?



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

Q3: How does this change your dart throwing strategy?



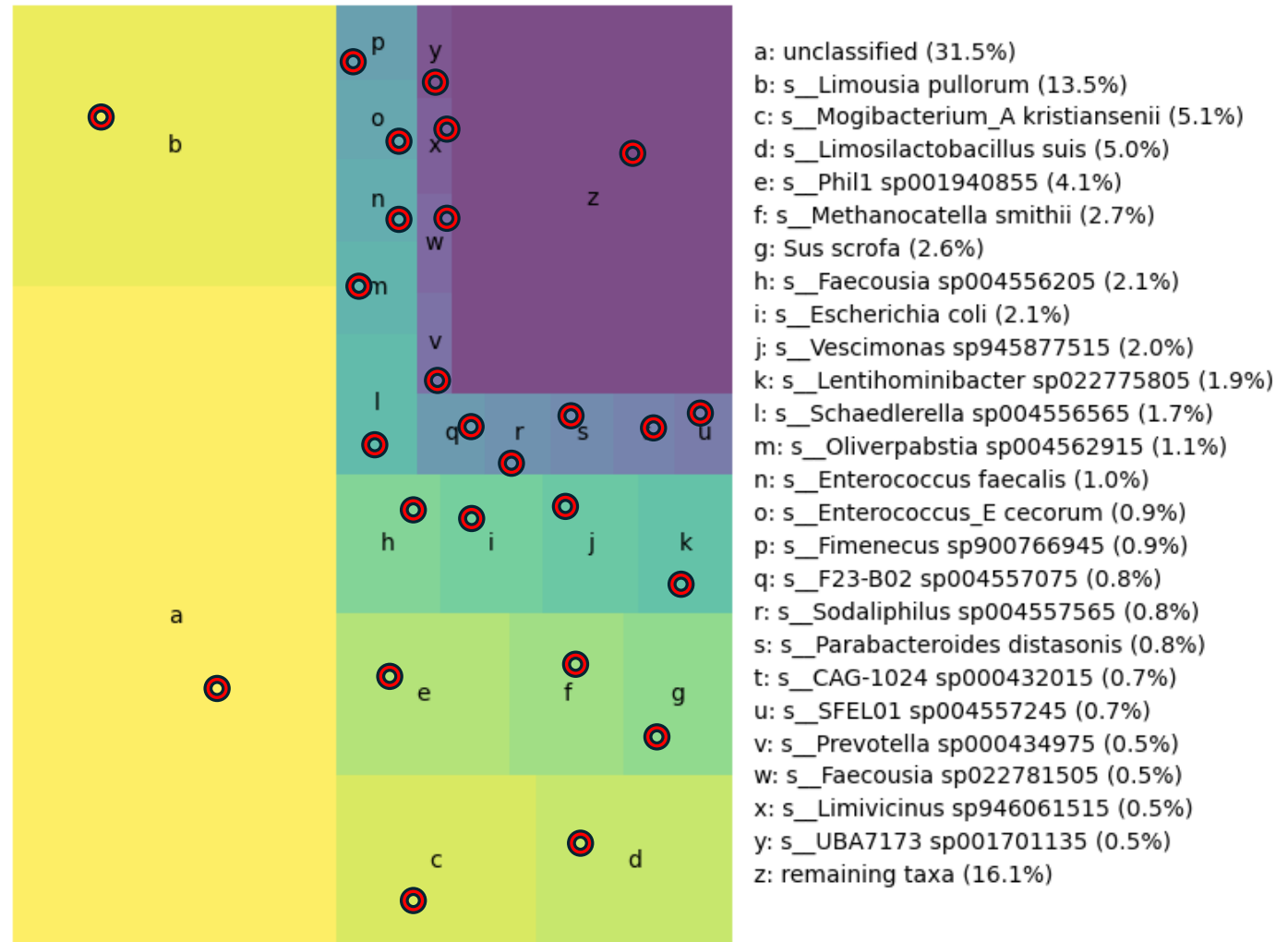
(linkage & sequencing read length)



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

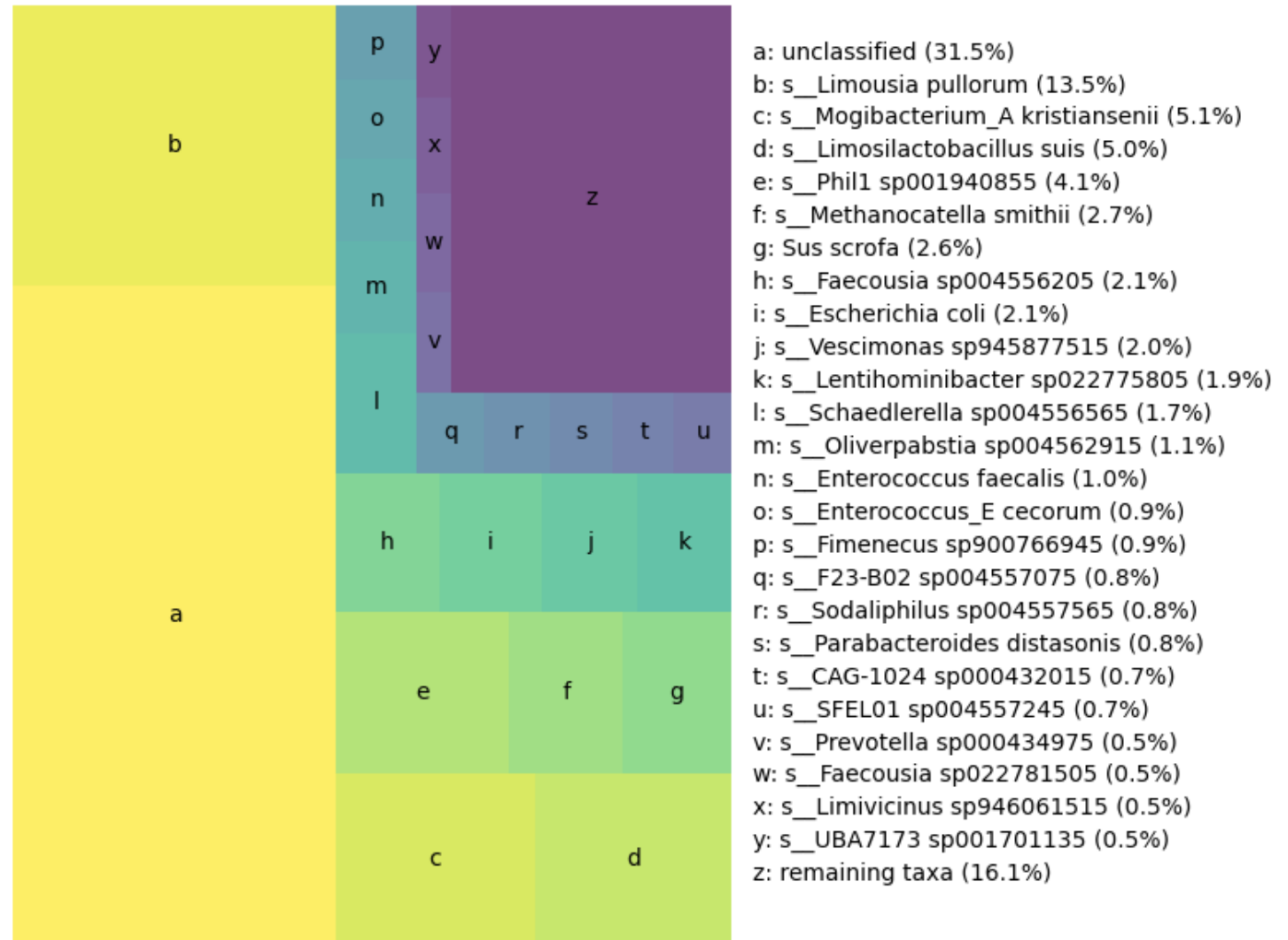
Q4: How does **this** change your dart throwing strategy?

(This is a 16S/amplicon analogy 😊)



Q5: what is the  
unclassified stuff?

(What *could* it be?)



The size of each box is, essentially:

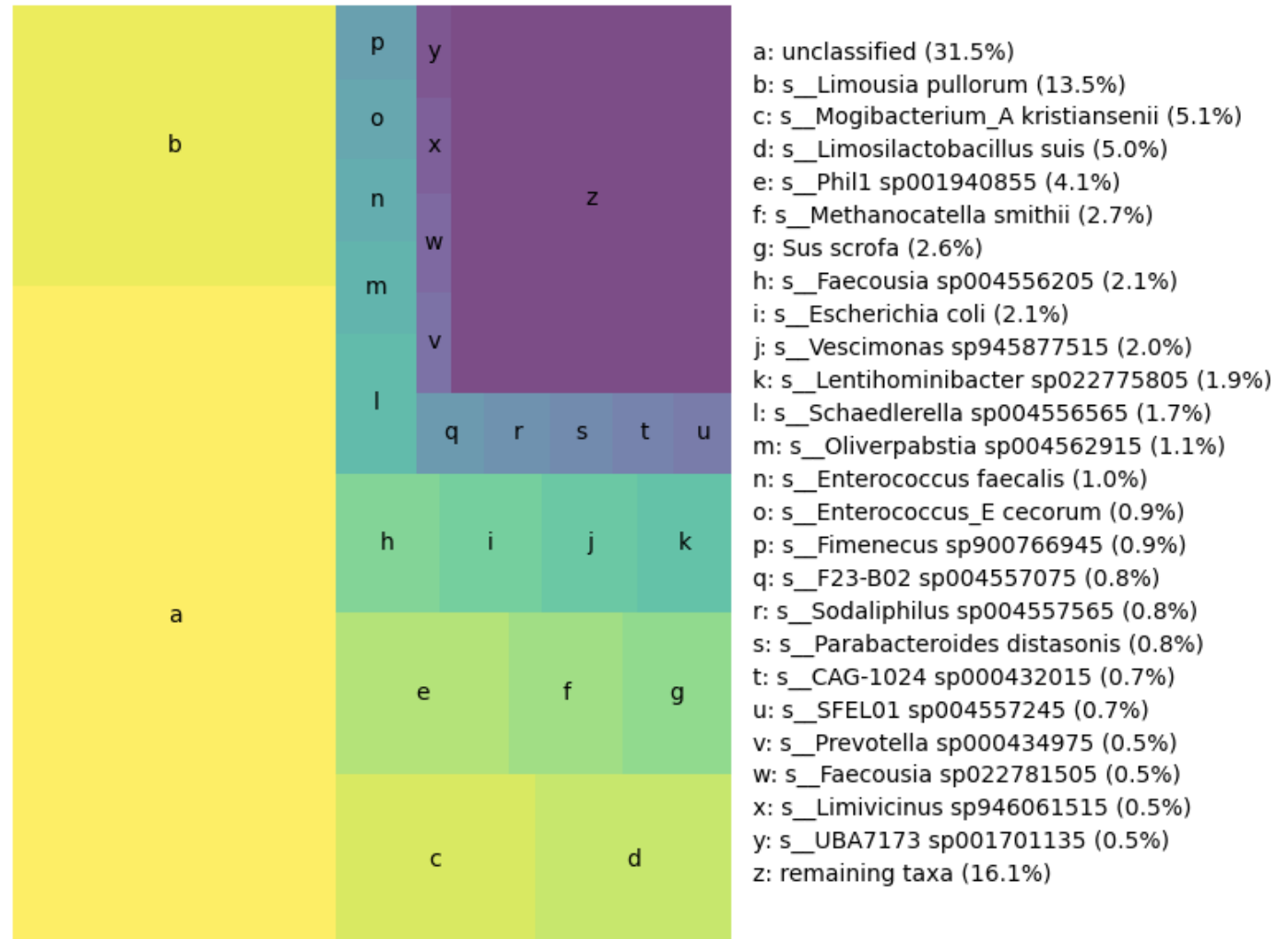
size of genome x  
abundance of genome

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?



Q7: What would rarefaction be doing in this situation?

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

