Markus Johansson

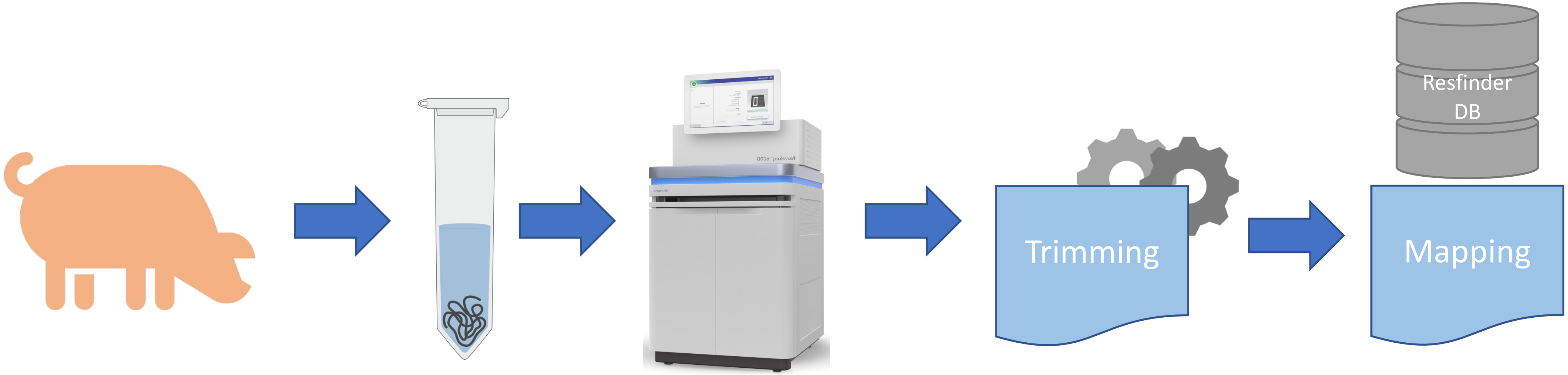
Abundance measurements

Compositional Data Analysis

Today's lecture

- What is compositional data?
- Why are sequence data compositional and why do we need to care?
- Transformations of abundances
- Pros- and cons of the different transformations
- How we handle zeros in our data
- Example of python and R packages for working with compositional data

What has been done so far to your samples



Resfinder DB

Trimming

Mapping

Sample collection

DNA isolation

DNA Sequencing

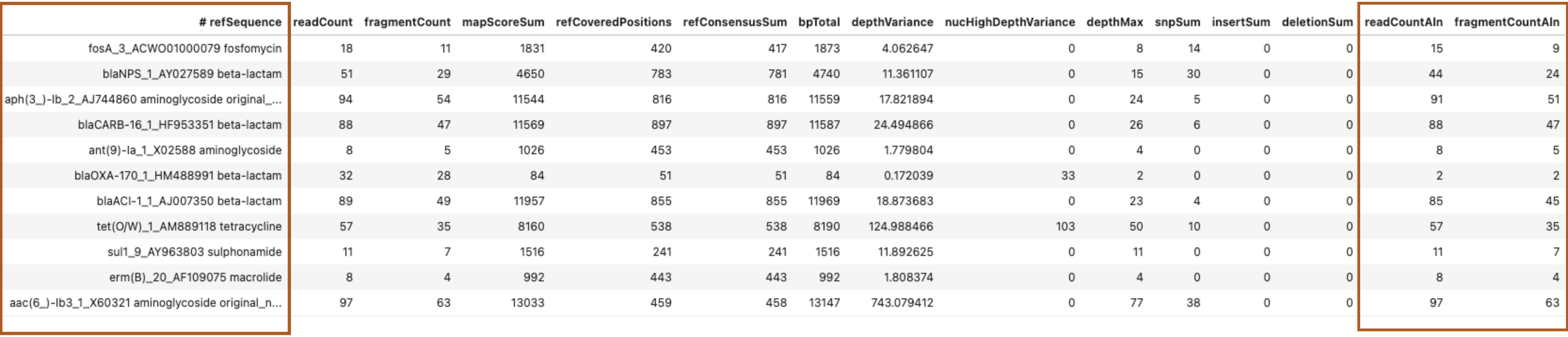
QC control

Abundance estimation

The .mapstat file

Reference

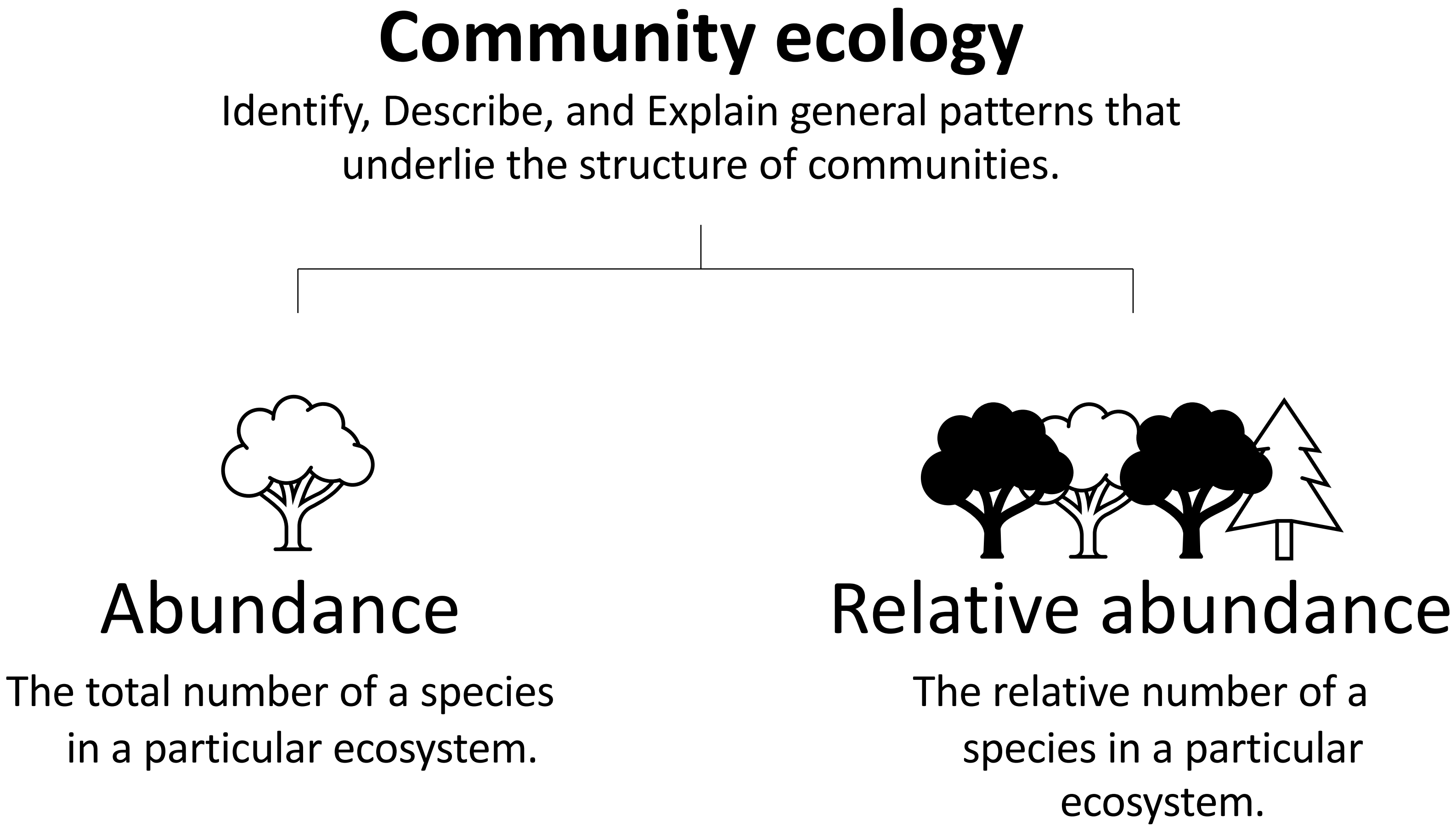
Read counts per feature



# refSequence	readCount	fragmentCount	mapScoreSum	refCoveredPositions	refConsensusSum	bpTotal	depthVariance	nucHighDepthVariance	depthMax	snpSum	insertSum	deletionSum	readCountAln	fragmentCountAln
fosA_3_ACWO01000079 fosfomycin	18	11	1831	420	417	1873	4.062647	0	8	14	0	0	15	9
blaNPS_1_AY027589 beta-lactam	51	29	4650	783	781	4740	11.361107	0	15	30	0	0	44	24
aph(3_)-Ib_2_AJ744860 aminoglycoside original_...	94	54	11544	816	816	11559	17.821894	0	24	5	0	0	91	51
blaCARB-16_1_HF953351 beta-lactam	88	47	11569	897	897	11587	24.494866	0	26	6	0	0	88	47
ant(9)-Ia_1_X02588 aminoglycoside	8	5	1026	453	453	1026	1.779804	0	4	0	0	0	8	5
blaOXA-170_1_HM488991 beta-lactam	32	28	84	51	51	84	0.172039	33	2	0	0	0	2	2
blaACI-1_1_AJ007350 beta-lactam	89	49	11957	855	855	11969	18.873683	0	23	4	0	0	85	45
tet(O/W)_1_AM889118 tetracycline	57	35	8160	538	538	8190	124.988466	103	50	10	0	0	57	35
sul1_9_AY963803 sulphonamide	11	7	1516	241	241	1516	11.892625	0	11	0	0	0	11	7
erm(B)_20_AF109075 macrolide	8	4	992	443	443	992	1.808374	0	4	0	0	0	8	4
aac(6_)-Ib3_1_X60321 aminoglycoside original_n...	97	63	13033	459	458	13147	743.079412	0	77	38	0	0	97	63

First question, what have we measured?

Abu-what?



Community ecology

Identify, Describe, and Explain general patterns that underlie the structure of communities.

Abundance

The total number of a species in a particular ecosystem.

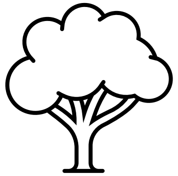
Relative abundance

The relative number of a species in a particular ecosystem.

Abu-what?

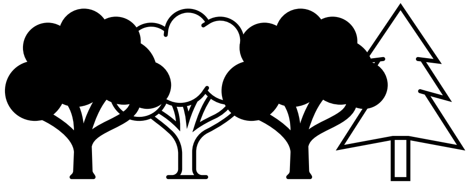
Metagenomics

Identify, Describe, and Explain general patterns that underlie the structure of communities.



Abundance

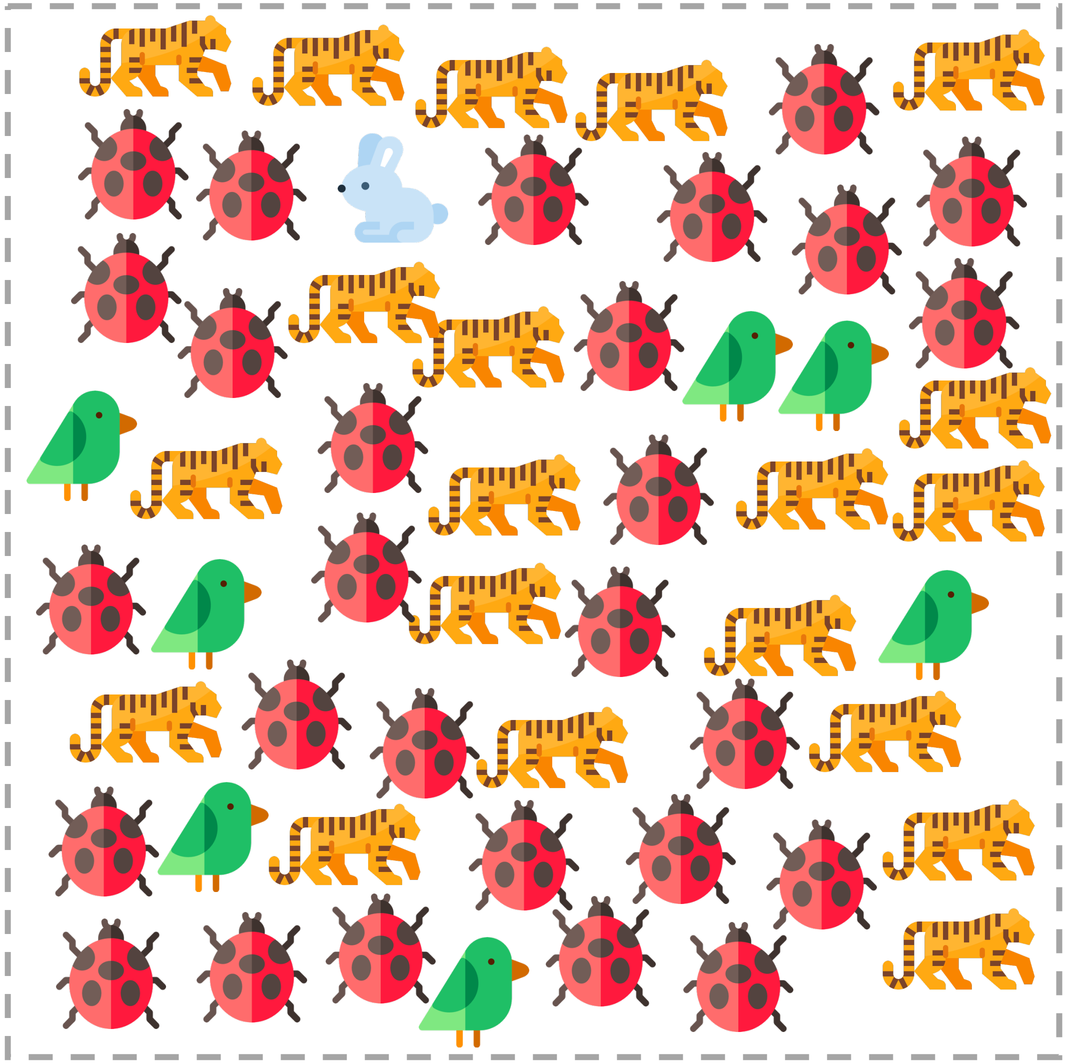
The total number of a reads assigned to a gene in a particular sample.



Relative abundance

The relative read counts of a gene in a particular sample.

What is the abundance?

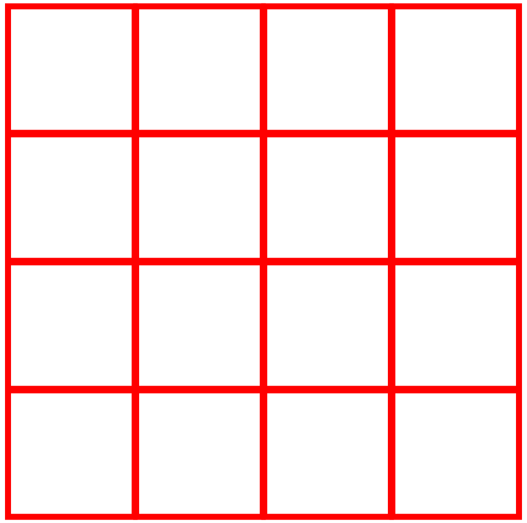


THE ENVIRONMENT

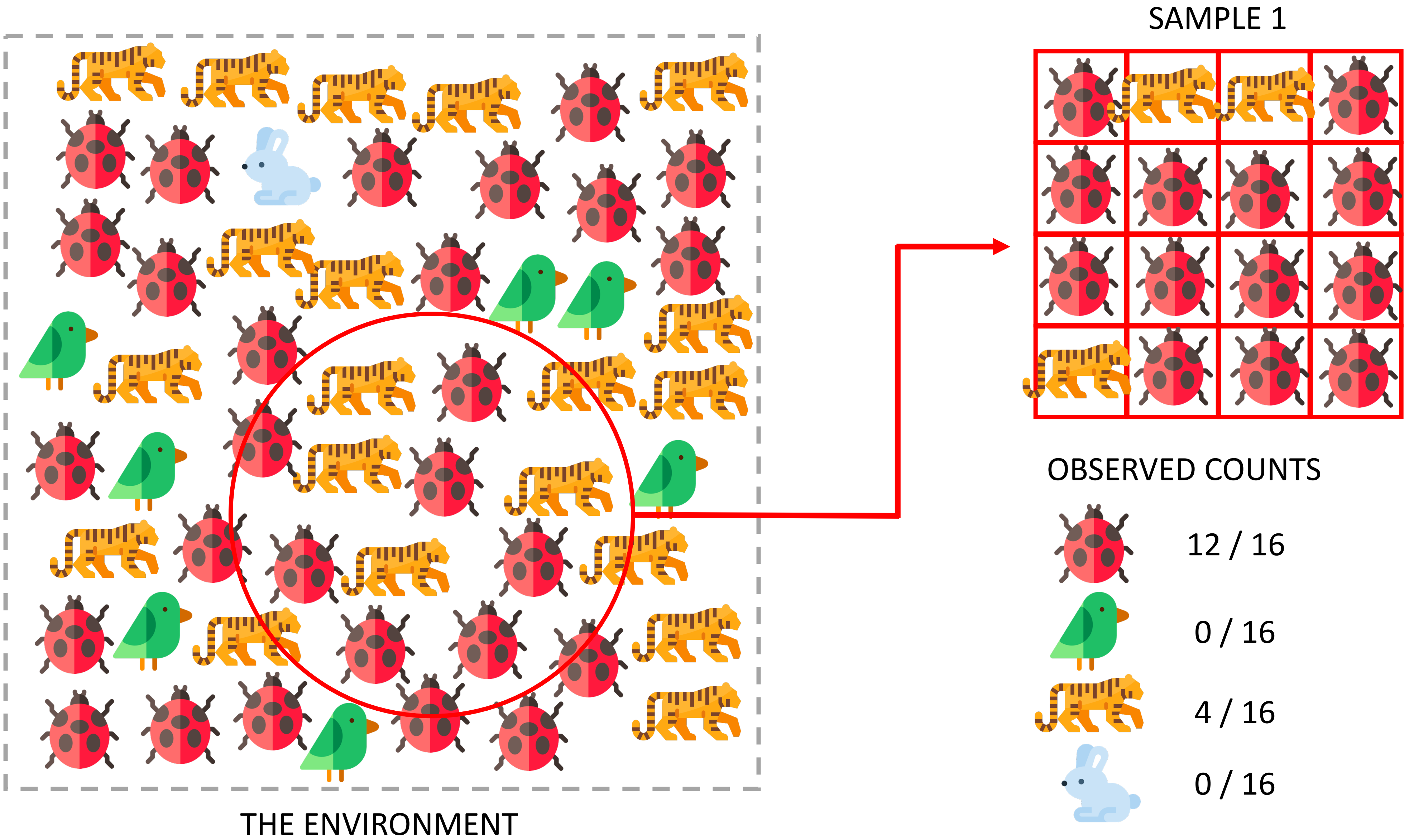
OBSERVED COUNTS

🐞	?
🐦	?
🐅	?
🐇	?

SEQUENCING MACHINE



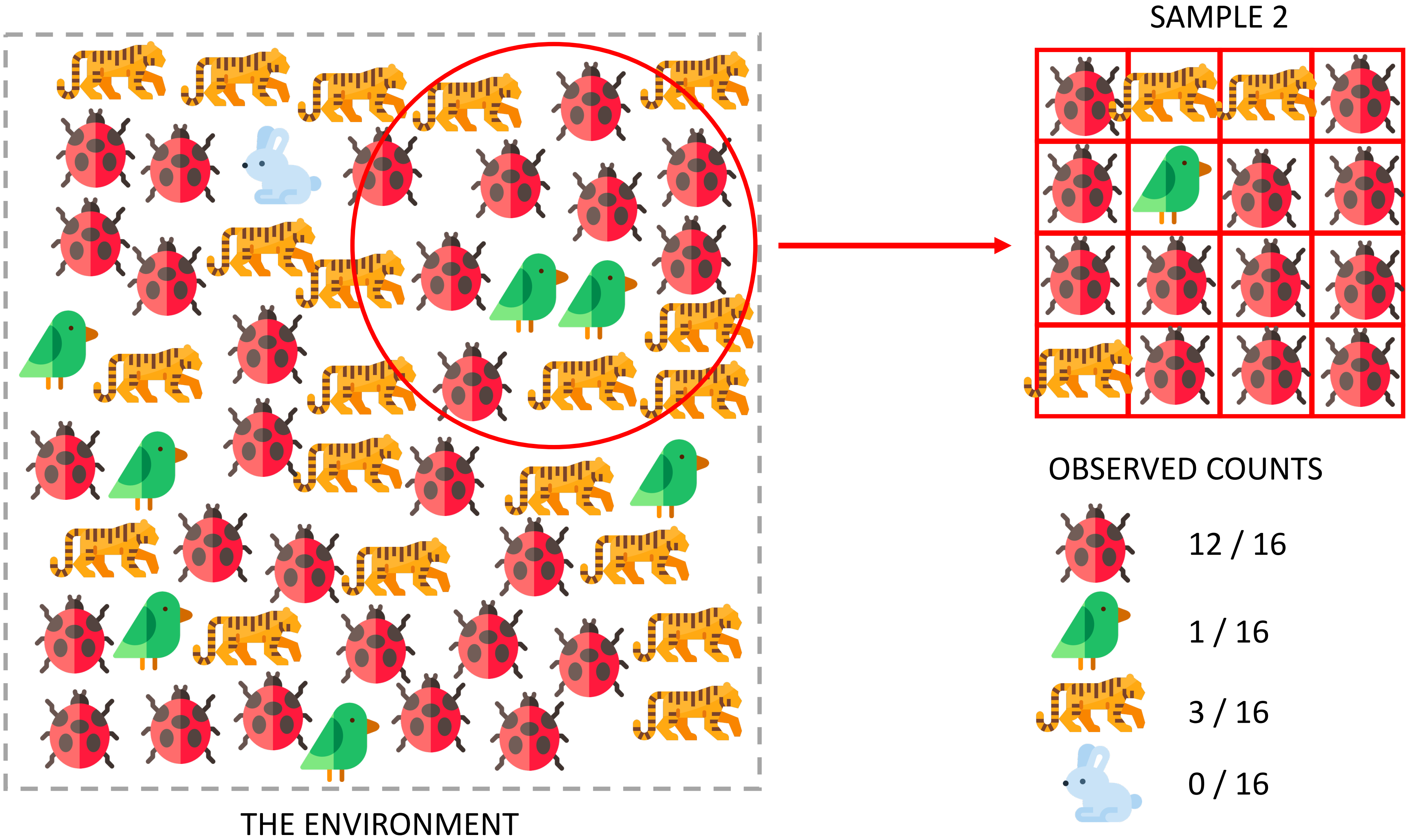
What is the abundance?



OBSERVED COUNTS

🐞	12 / 16
🐦	0 / 16
🐅	4 / 16
🐇	0 / 16

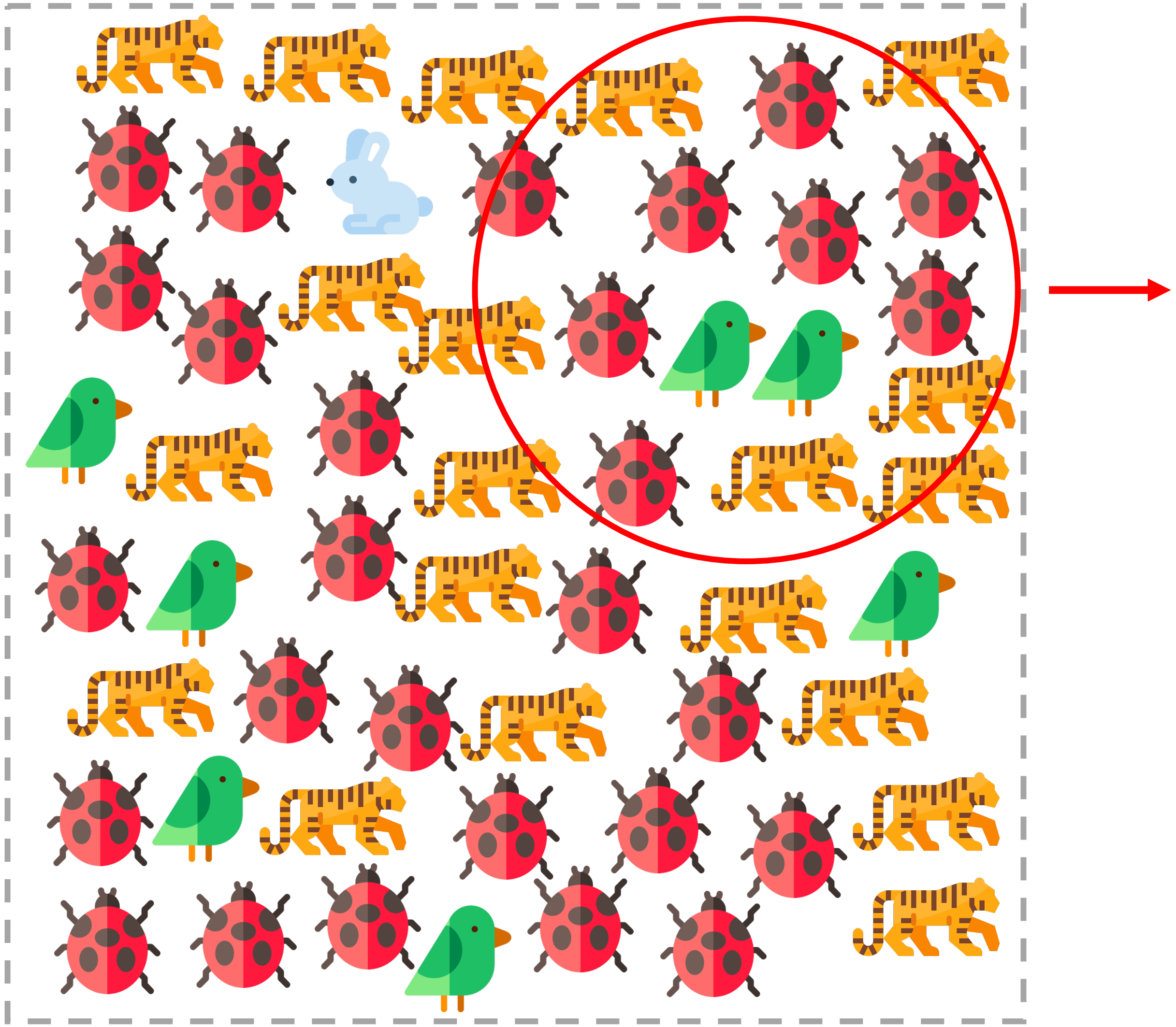
What is the abundance?



OBSERVED COUNTS

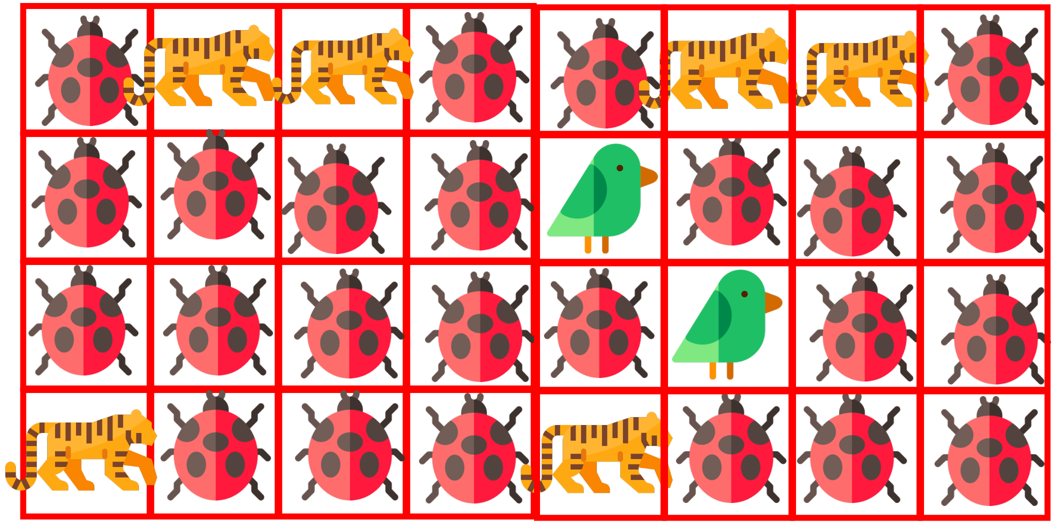
🐞	12 / 16
🐦	1 / 16
🐅	3 / 16
🐇	0 / 16

What is the abundance?



THE ENVIRONMENT

SAMPLE 3

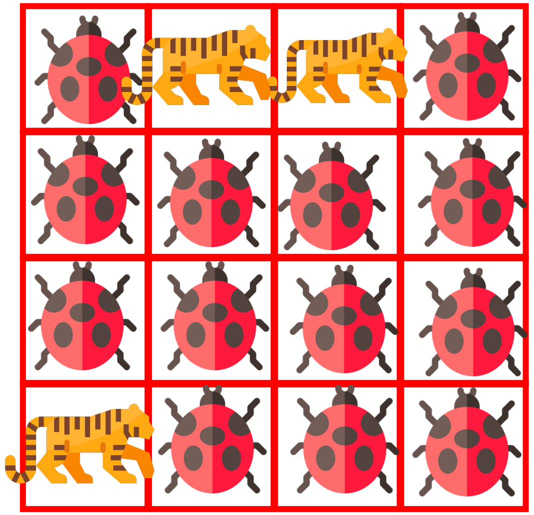


OBSERVED COUNTS

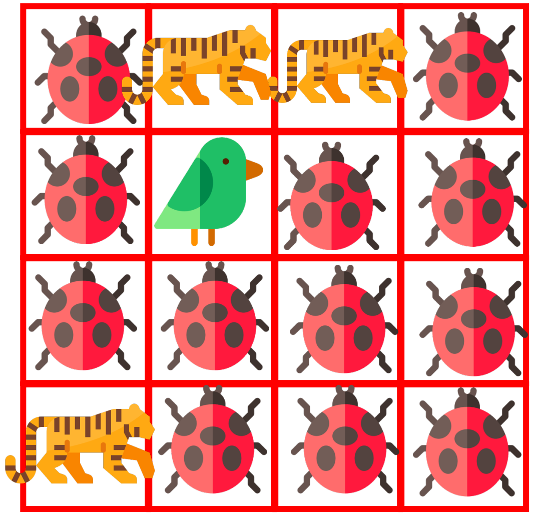
🐞	24 / 32
🐦	2 / 32
🐅	6 / 32
🐇	0 / 32

What is the abundance?

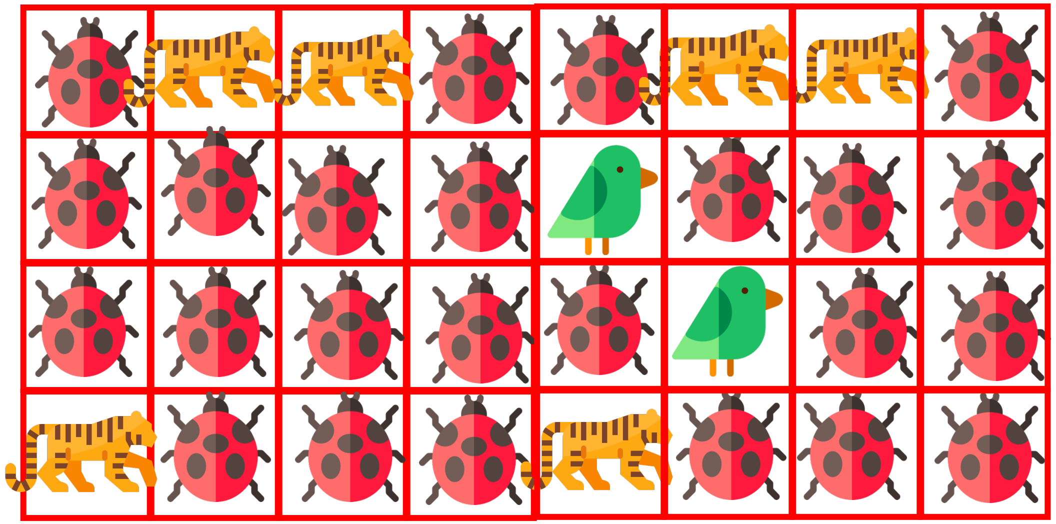
SAMPLE 1



SAMPLE 2

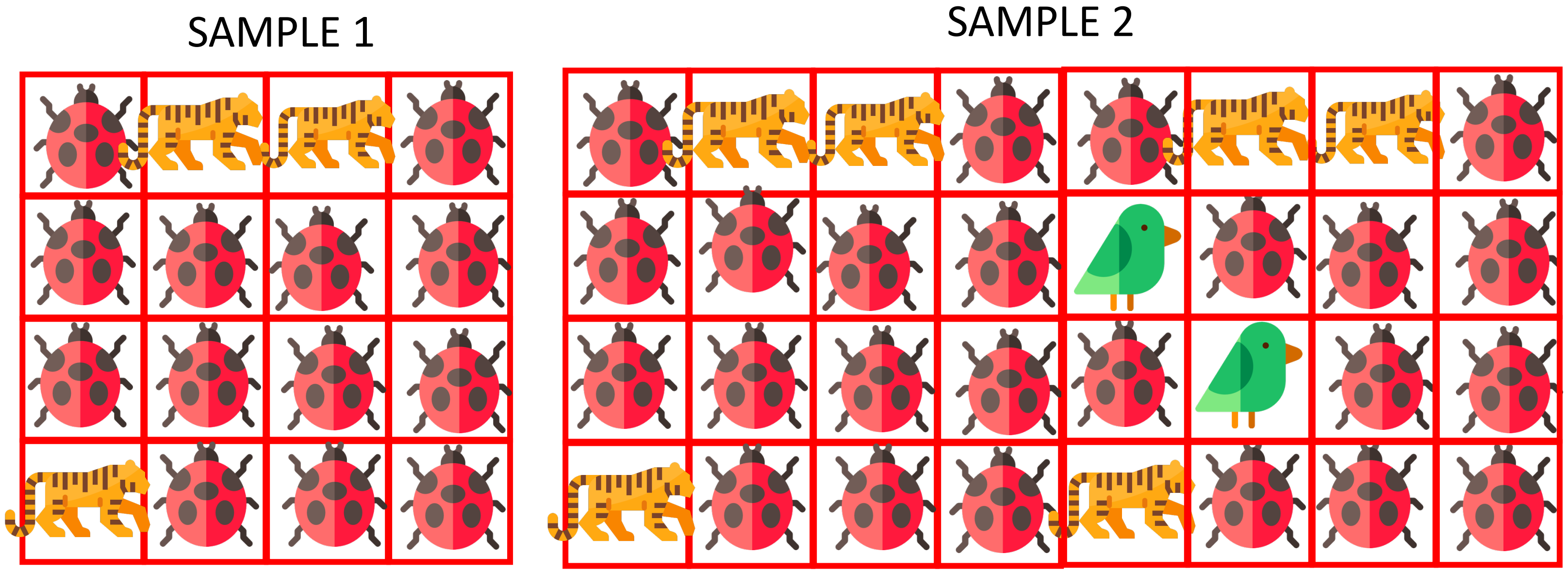


SAMPLE 3



What do the total abundance mean?

If the same samples were sequenced twice, with 16 reads and 32 reads.



Slots	Ladybugs	Birds	Tigers	Rabbits
16	12	0	4	0
32	24	2	6	0

What do the total abundance mean?

If the same samples were sequenced twice, with 16 reads and 32 reads.

Sample (#slots)	Ladybugs	Birds	Tigers	Rabbits
1 (16)	12	0	4	0
3 (32)	24	2	6	0
Actual abundance	26	7	20	1

Sample (#slots)	Ladybugs	Birds	Tigers	Rabbits
1 (16)	75%	0	25%	0
3 (32)	75%	6.25%	18.75%	0
Actual abundance	59.1%	15.9%	22.72%	2.3%

- The total number of observed species are a function of the total number of sequenced reads
- Absolute counts only convey information on the precision, not the abundance
- We can only draw conclusion on the relative difference in species.
- Address variability in counts by normalizing with total number of reads

What is the abundance?

We randomly sampled three times:

Sample	Ladybugs	Birds	Tigers	Rabbits
1	12	0	4	0
2	12	1	3	0
3	24	2	6	0

Sample (#slots)	Ladybugs	Birds	Tigers	Rabbits
1 (16)	75%	0	25%	0
2 (16)	75%	6.25%	18.75%	0
3 (32)	75%	6.25%	18.75%	0

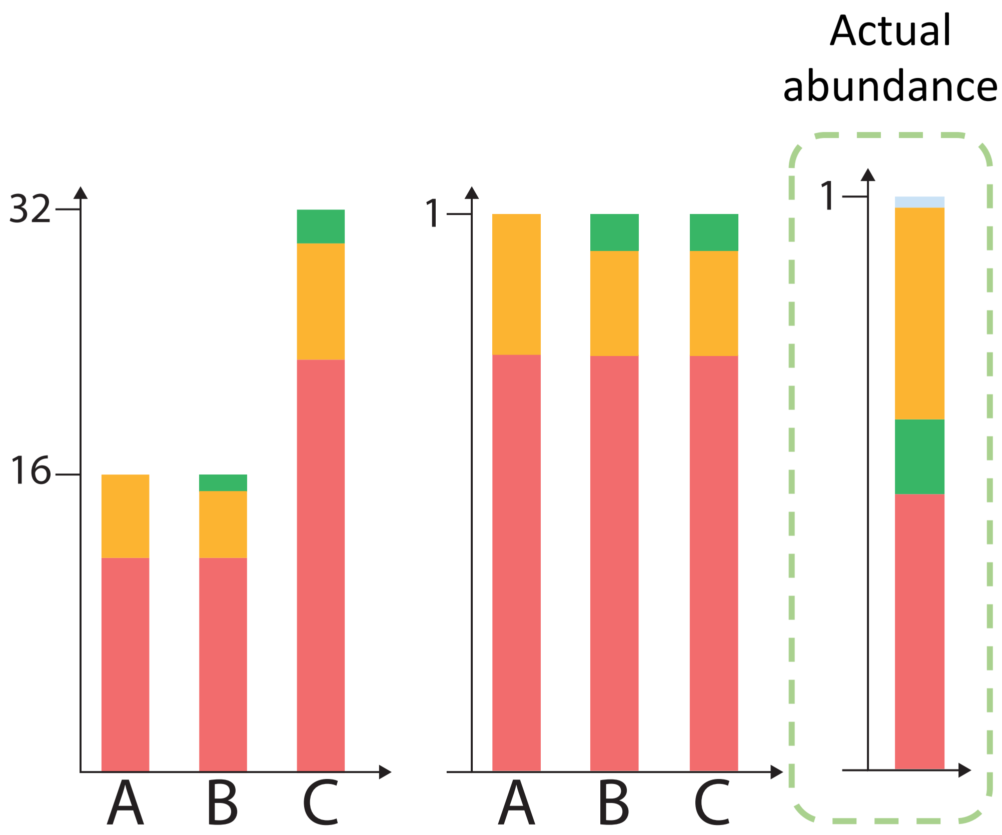
- The total number of observed species are a function of the total number of sequenced reads
- Absolute counts only convey information on the precision, not the abundance
- We can only draw conclusion on the relative difference in species.
- Address variability in counts by normalizing with total number of reads

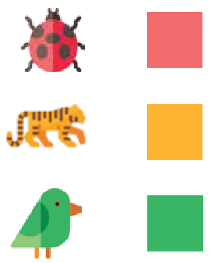
What is the abundance?

We randomly sampled three times:

Sample	Ladybugs	Birds	Tigers	Rabbits
1	12	0	4	0
2	12	1	3	0
3	24	2	6	0

Sample (#slots)	Ladybugs	Birds	Tigers	Rabbits
1 (16)	75%	0	25%	0
2 (16)	75%	6.25%	18.75%	0
3 (32)	75%	6.25%	18.75%	0





Address variability in counts by normalizing with total number of reads

Did we observe the true abundances in any of the samples? Why / Why not?

Discuss with those around you!

- Does a zero mean that the rabbit is not there?
- Why did we not observe a rabbit, despite sampling three times?
- What would happen if we had even more slots to fill (reads)?

Compositional Data Analysis (CoDA)

The total read count is a **fixed-size**, **random sample** of the relative abundance of the molecules in the underlying ecosystem.

- Random sample of the environment
- Fixed capacity of the machine

Causes,

- Observed gene counts can thus not be related to the total read count
- Total number of reads only convey the precision

CoDA focuses on the **relationship between gene counts**

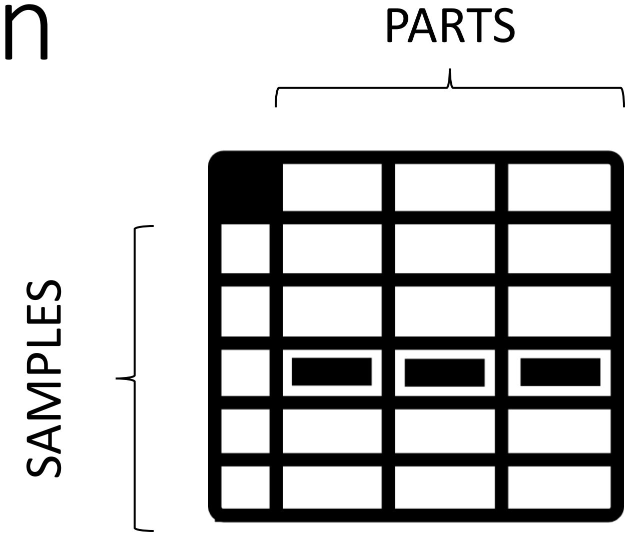
Compositional data are quantitative descriptions of the parts of some whole, conveying relative information. https://en.wikipedia.org/wiki/Compositional_data

The metagenomic composition

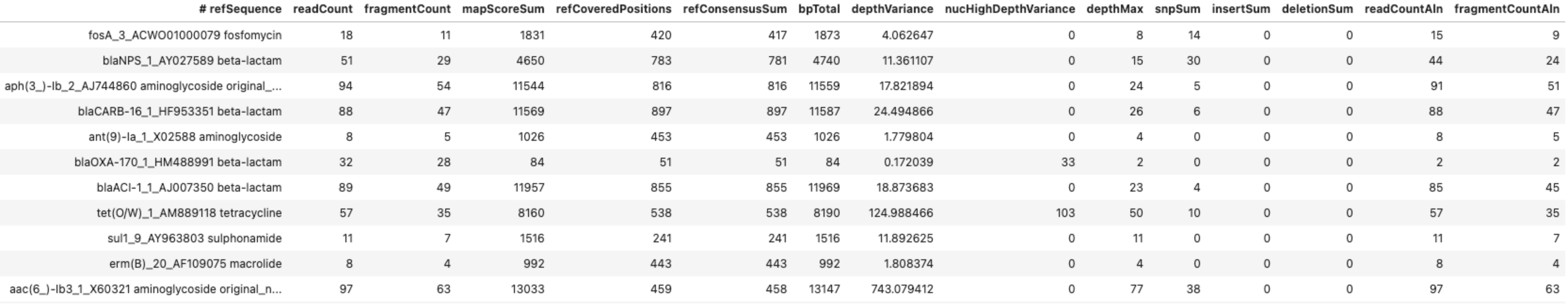
A sample is a composition x of D parts:

$$x = [x_1, x_2, \dots, x_D]$$

where x_i is a count (i.e., read gene count)



The sample composition is the **.mapstat** file from KMA:



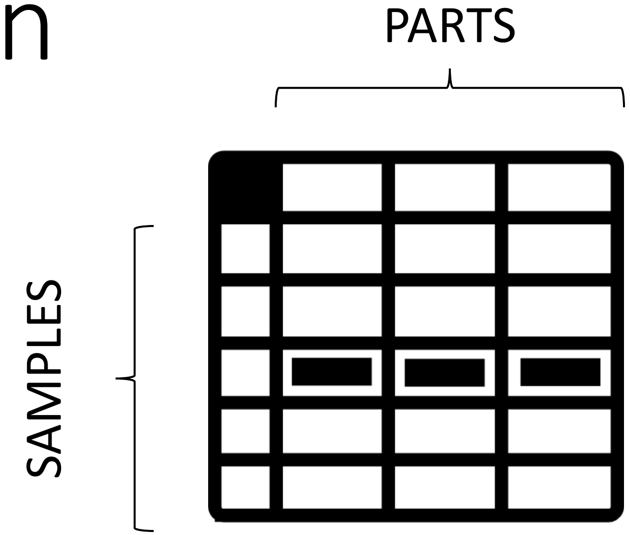
# refSequence	readCount	fragmentCount	mapScoreSum	refCoveredPositions	refConsensusSum	bpTotal	depthVariance	nucHighDepthVariance	depthMax	snpSum	insertSum	deletionSum	readCountAln	fragmentCountAln
fosA_3_ACWO01000079 fosfomycin	18	11	1831	420	417	1873	4.062647	0	8	14	0	0	15	9
blaNPS_1_AY027589 beta-lactam	51	29	4650	783	781	4740	11.361107	0	15	30	0	0	44	24
aph(3_)-Ib_2_AJ744860 aminoglycoside original_...	94	54	11544	816	816	11559	17.821894	0	24	5	0	0	91	51
blaCARB-16_1_HF953351 beta-lactam	88	47	11569	897	897	11587	24.494866	0	26	6	0	0	88	47
ant(9)-Ia_1_X02588 aminoglycoside	8	5	1026	453	453	1026	1.779804	0	4	0	0	0	8	5
blaOXA-170_1_HM488991 beta-lactam	32	28	84	51	51	84	0.172039	33	2	0	0	0	2	2
blaACI-1_1_AJ007350 beta-lactam	89	49	11957	855	855	11969	18.873683	0	23	4	0	0	85	45
tet(O/W)_1_AM889118 tetracycline	57	35	8160	538	538	8190	124.988466	103	50	10	0	0	57	35
sul1_9_AY963803 sulphonamide	11	7	1516	241	241	1516	11.892625	0	11	0	0	0	11	7
erm(B)_20_AF109075 macrolide	8	4	992	443	443	992	1.808374	0	4	0	0	0	8	4
aac(6_)-Ib3_1_X60321 aminoglycoside original_n...	97	63	13033	459	458	13147	743.079412	0	77	38	0	0	97	63

The metagenomic composition

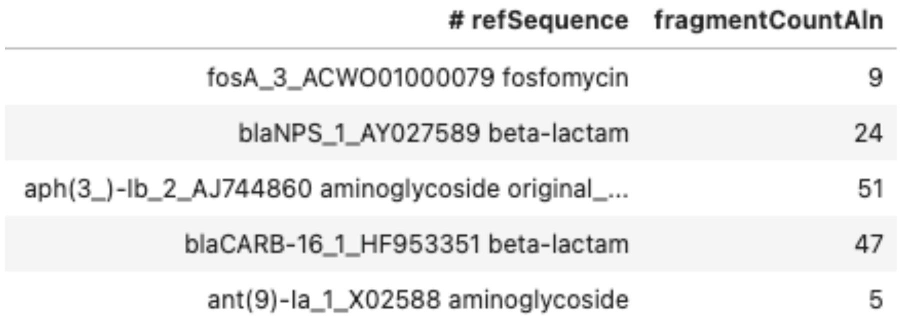
A sample is a composition x of D parts:

$$x = [x_1, x_2, \dots, x_D]$$

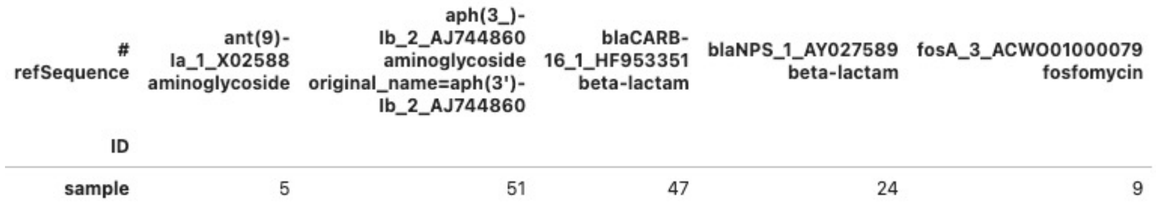
where x_i is a count (i.e., read gene count)



Use fragmentCountAln as counts and pivot:



# refSequence	fragmentCountAln
fosA_3_ACWO01000079 fosfomycin	9
blaNPS_1_AY027589 beta-lactam	24
aph(3_)-Ib_2_AJ744860 aminoglycoside original_...	51
blaCARB-16_1_HF953351 beta-lactam	47
ant(9)-Ia_1_X02588 aminoglycoside	5



# refSequence	ant(9)-Ia_1_X02588 aminoglycoside	aph(3_)-Ib_2_AJ744860 aminoglycoside original_name=aph(3')-Ib_2_AJ744860	blaCARB-16_1_HF953351 beta-lactam	blaNPS_1_AY027589 beta-lactam	fosA_3_ACWO01000079 fosfomycin
ID					
sample	5	51	47	24	9

Aggregating counts

The ResFinder database contains more than 3100 genes – should we look at them all? Maybe we would rather look at resistance classes?

Amalgamation is the summing of parts. Given a set of indices $A = [i_1, i_2, .., i_a]$ to sum, and another set $\tilde{A} = D_i \backslash \mathrm{A}$, the amalgamated composition is:

$$x' = (x_{\tilde{A}}, x_A), \qquad x_A = \sum_{i \in A} x_i$$

We can amalgamate resistance genes that belongs to the same class.

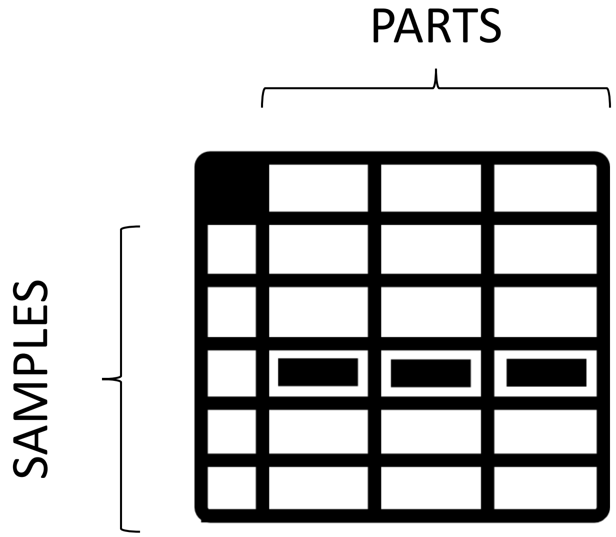
- Reduces the number of columns in the matrix.

Rescaling counts

Applying **closure** to multiple compositions rescales counts to the same total sum:

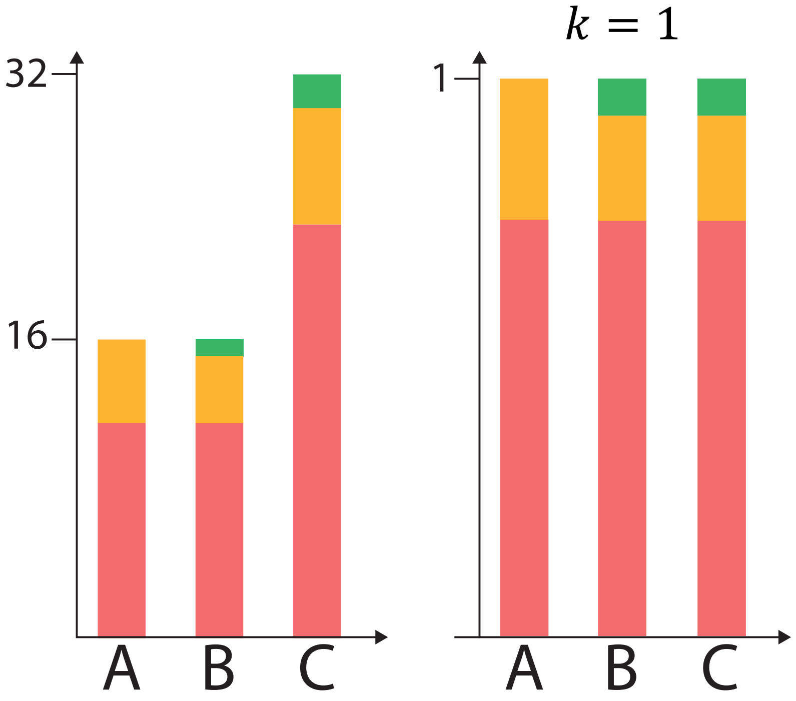
$$C(x) = \frac{k}{\sum_{i=1}^{D} x_i} \cdot x$$

where k is a positive number.



Why not divide with total read/fragment count?

Closure gives the relative abundance of reads that were mappable in the sample.



Implication of NGS data being compositions

Compositional data

- Has negative correlation bias
- Prone to spurious correlations
- Does not have Euclidian distances

Implications

- Common statistical tests are unreliable
- Multivariate analysis doesn’t work, eg clustering

Transforming counts to abundances – ALR

To calculate gene abundances, we can use the additive log-ratio transformation:

Additive log-ratio (ALR) gives parts given as relative to a reference x_D:

$$\mathrm{ALR}(x) = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \dots, \frac{x_{D-1}}{x_D}\right)$$

- The choice of x_D is up to the analyst
- ALR transformation is a within-sample normalization method

Transforming counts to abundances – ALR

Additive log-ratio (ALR) gives parts given as relative to a reference x_D:

$$\mathrm{ALR}(x) = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right)$$

Variations of ALR :

- log(RPKM): Reads Per Kilobase of transcript, per Million mapped reads

$$\log(\mathrm{RPKM}) = \log\left(\frac{[\text{Number of reads mapped to a gene}]\cdot 10^3 \cdot 10^6}{[\text{Total number of mapped reads}]\cdot[\text{gene length in bp}]}\right)$$

- log(FPKM): Fragments Per Kilobase of transcript, per Million mapped reads

$$\log(\mathrm{FPKM}) = \log\left(\frac{[\text{Number of reads mapped to a gene}]\cdot 10^3 \cdot 10^6}{[\text{Total number of read fragments}]\cdot[\text{gene length in bp}]}\right)$$

Transforming counts to abundances – ALR

Additive log-ratio (ALR) gives parts given as relative to a reference x_D:

$$\text{ALR}(x) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)$$

Variations of ALR :

- log(FPKM): Fragments Per Kilobase of transcript, per Million fragments

$$\log(\text{FPKM}) = \log \left(\frac{[\text{Number of fragnents mapped to a gene}] \cdot 10^3 \cdot 10^6}{[\text{Total number of read fragments}] \cdot [\text{gene length in bp}]} \right)$$

Number of read fragments can be found in the header of the mapstat file

```
## method          KMA
## version         1.2.17a
## database        ResFinder_20190905
## fragmentCount   32078691
## date    2019-12-11
```

Transforming counts to abundances – CLR

Instead of choosing which part to compare to all the other parts, we can use the mean of the composition.

But not just any mean: the **geometric mean**.

Geometric mean of a vector x:

$$g_m(x) = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}} = \exp\left(\frac{1}{D}\sum_{i=1}^{D} \ln x_i\right)$$

Transforming counts to abundances – CLR

Instead of choosing which part to compare to all the other parts, we can use the mean of the composition.

Geometric mean of a vector x:

$$g_m(x) = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}} = \exp\left(\frac{1}{D}\sum_{i=1}^{D} \ln x_i\right)$$

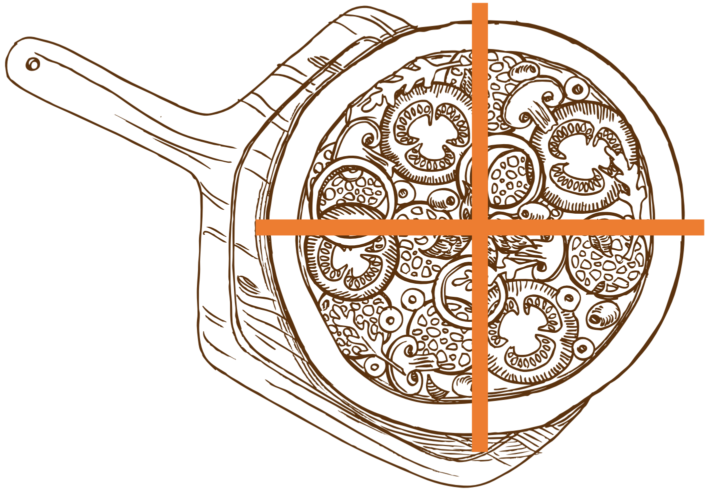
Centered log-ratio (CLR) transformation:

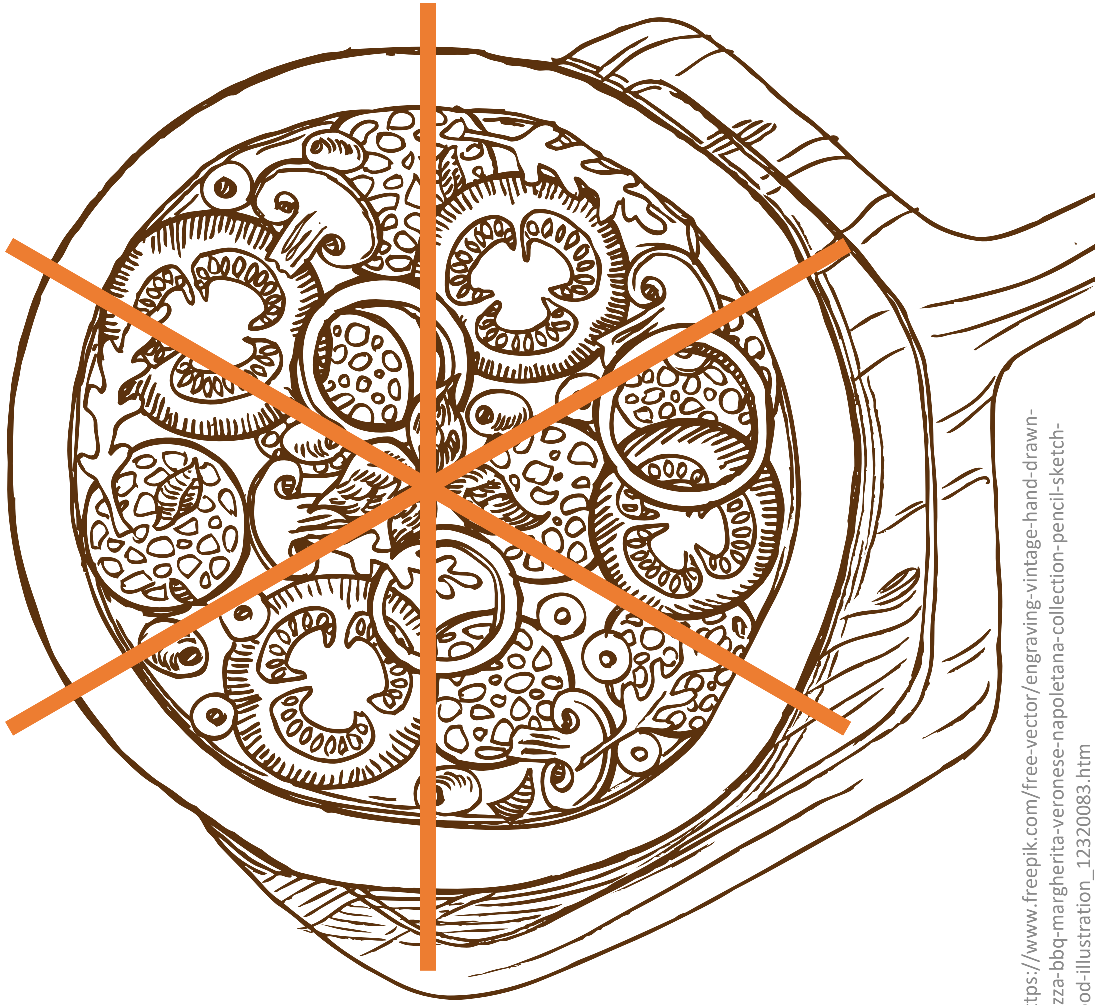
$$\mathrm{CLR}(x) = \left(\ln \frac{x_1}{g_m(x)}, \ln \frac{x_2}{g_m(x)}, \dots, \ln \frac{x_D}{g_m(x)}\right)$$

When to use ALR and CLR

It really depends on which question you want to answer.

- Picking the largest slice in one pizza ALR
- In multiple pizzas CLR





https://www.freepik.com/free-vector/engraving-vintage-hand-drawn-pizza-bbq-margherita-veronese-napoletana-collection-pencil-sketch-food-illustration_12320083.htm

ALR	vs	CLR
Easy to interpret		Hard to interpret
Differs if reference changes		Changes if parts are removed
Only algebraic vector space operations can be used		Standard multivariate analysis techniques can be used

Small example

Sample	Ladybugs	Birds	Tigers	Rabbits
1	12	0	4	0
2	12	1	3	0
3	12	2	6	0

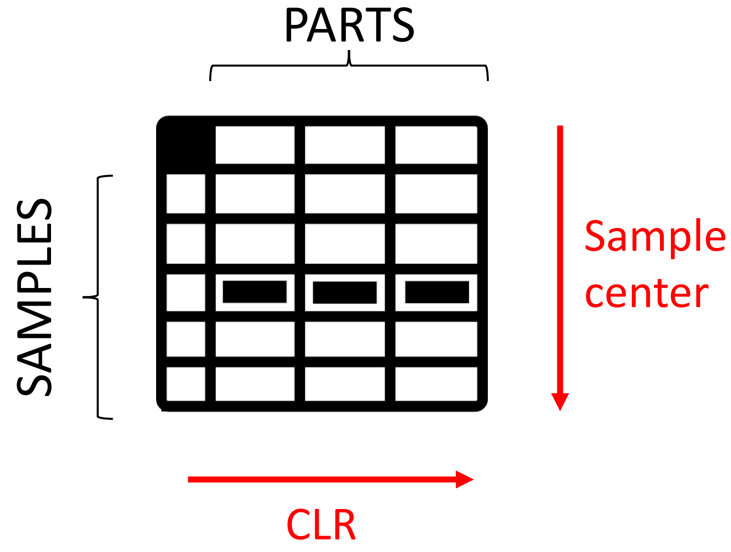
Compositional summary statistics

To describe the central trend and sample dispersion in a compositional dataset, we can calculate the mean and variance.

The **sample center** is the geometric means of parts in a closed composition:

$$\mathrm{Cen}[\mathrm{X}] = \mathrm{C}[\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_D]$$

$$\hat{g}_j = \left(\prod_{i=1}^{n} x_{i,j}\right)^{\frac{1}{n}}, \qquad j = 1,2, \ldots, D$$



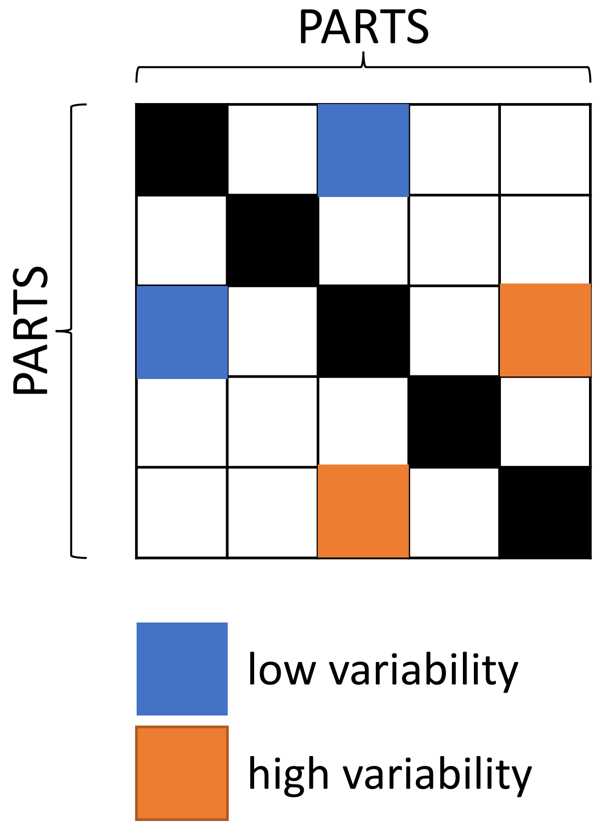
Compositional summary statistics

To describe the central trend and sample dispersion in a compositional dataset, we can calculate the mean and variance.

The **dispersion** in the log-ratio parts is given by the **variation matrix**:

$$T = \left[t_{ij}\right]$$

$$t_{ij} = y = \operatorname{var}\left(\ln \frac{x_i}{x_j}\right), \operatorname{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$



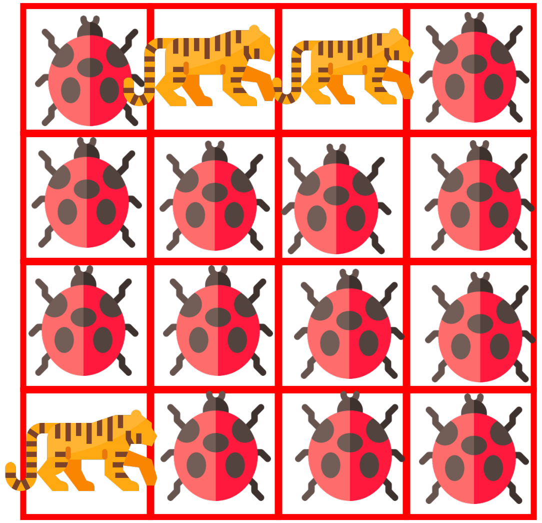
What about zero counts?

$\log(0)$ is not a real number – so what do we do?

- Depends on the type of zero
- Generally, We replace zeroes with a small value

Not all zeroes are the same

Different types of zeroes



OBSERVED COUNTS

🐞	12 / 16
🐦	0 / 16
🐅	4 / 16
🐇	0 / 16

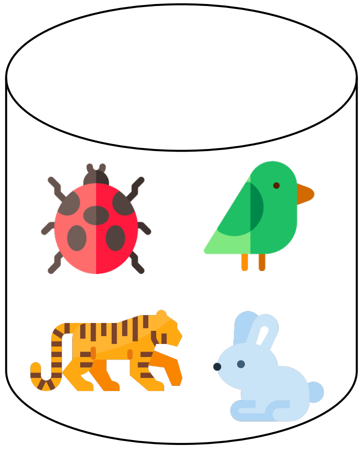
Structural Zeros

- A feature cannot be observed because its not there.
- Could also be caused by methodological problems

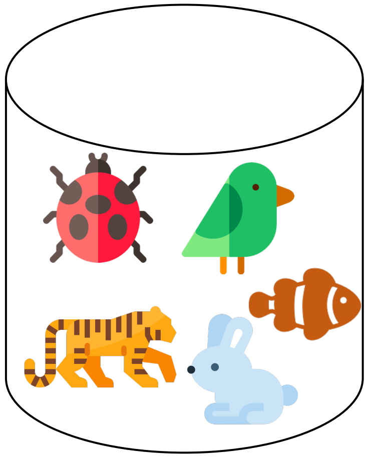
Solution: Better to exclude these features if possible

Different types of zeroes

Database v1 content



Database v2 content

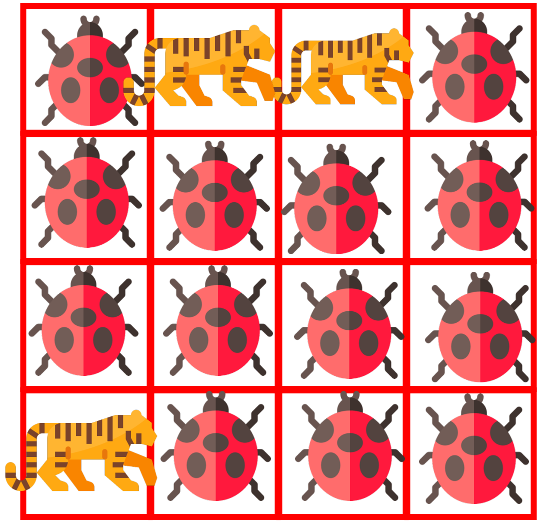


Missing values

- Common for surveys or metadata collection where all fields are not filled in
- Could be missing because of updates to contents in reference databases
- Samples were mapped differently, result were amalgamated. E.g.
  - Sample A – Bacteria and protozoa merged into microorganisms
  - Sample B – Bacteria and protozoa kept seperate

Solution: re-map the data or exclude samples

Different types of zeroes



OBSERVED COUNTS

🐞	12 / 16
🐦	0 / 16
🐅	4 / 16
🐇	0 / 16

Count zeros/ Below detection limit

- By chance a DNA fragment isn't sampled
- Caused by a feature occurring at to low concentration
- With increasing sequence depth, precision, we get a better estimate of the real distribution

Solution: replace the zeros with a small number

What small value should we choose?

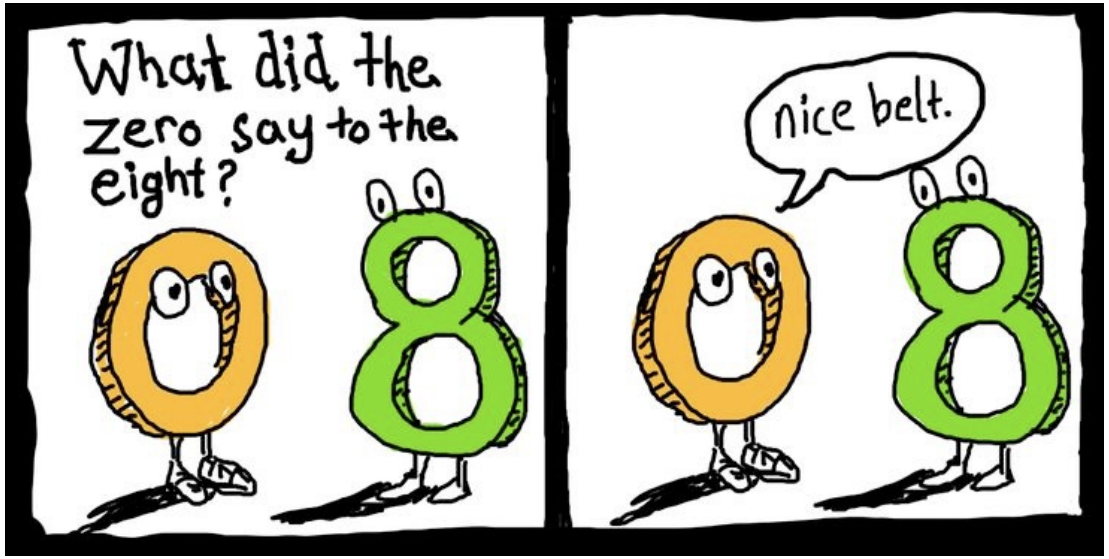
Use a small number that is below the detection limit.

- Doesn't scale with the data

Replace it with a 1/[total read count]

- Scales with the data, but its not informed by it

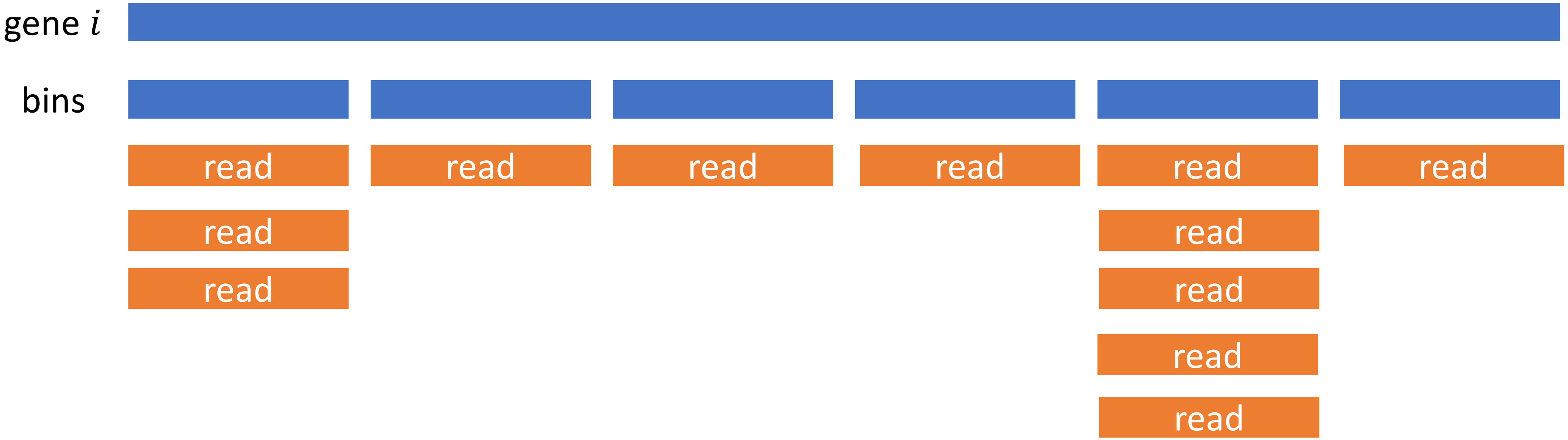
Better to use a Bayesian probability approach, where we model the proportion p of reads instead of the observed count.



A Bayesian probability approach for replacing zeroes

Assumption, reads are randomly sampled

- If we generate enough reads, we should get full coverage of the gene.



n_i : observed read count of gene i

We assume that each n_i was sampled from a Poisson process:

$$n_i \sim Poisson(\lambda_i)$$

A Bayesian probability approach for replacing zeroes

- Model the probability of observing a read given the sequencing depth
- Estimate the underlying proportion of reads by sampling from a Dirichlet distribution
- Observed read count are used as weights
- The estimated proportions are based on the observed abundance.

CoDa in practice

Which programs to use?

Python	R
pandas	tidyverse
matplotlib seaborn python-ternary	ggplot2 ggtern
pyCoDa (https://bitbucket.org/genomicepidemiology/pycoda/src)	compositions zCompositions

https://www.freepik.com/free-vector/data-report-illustration-concept_6195527.htm

CoDa in practice

Preprocessing

Loading	🐍 pandas.read_csv R readr::read_csv
Pivoting	🐍pandas.DataFrame.pivot R tidyr::pivot_wider
Scale counts	Scale fragmentCountsAln with gene lengths in kb
Replace zeroes	🐍 df.coda.zero_replacement R zCompositions::cmultRepl

Statistics & Transformations

Summary statistics	🐍 df.coda.gmean df.coda.varmatrix R compositions::mean compositions::var
Closure	🐍 df.coda.closure R compositions::clo
ALR	🐍 df.coda.alr R compositions::alr
CLR	🐍 df.coda.clr R compositions::clr

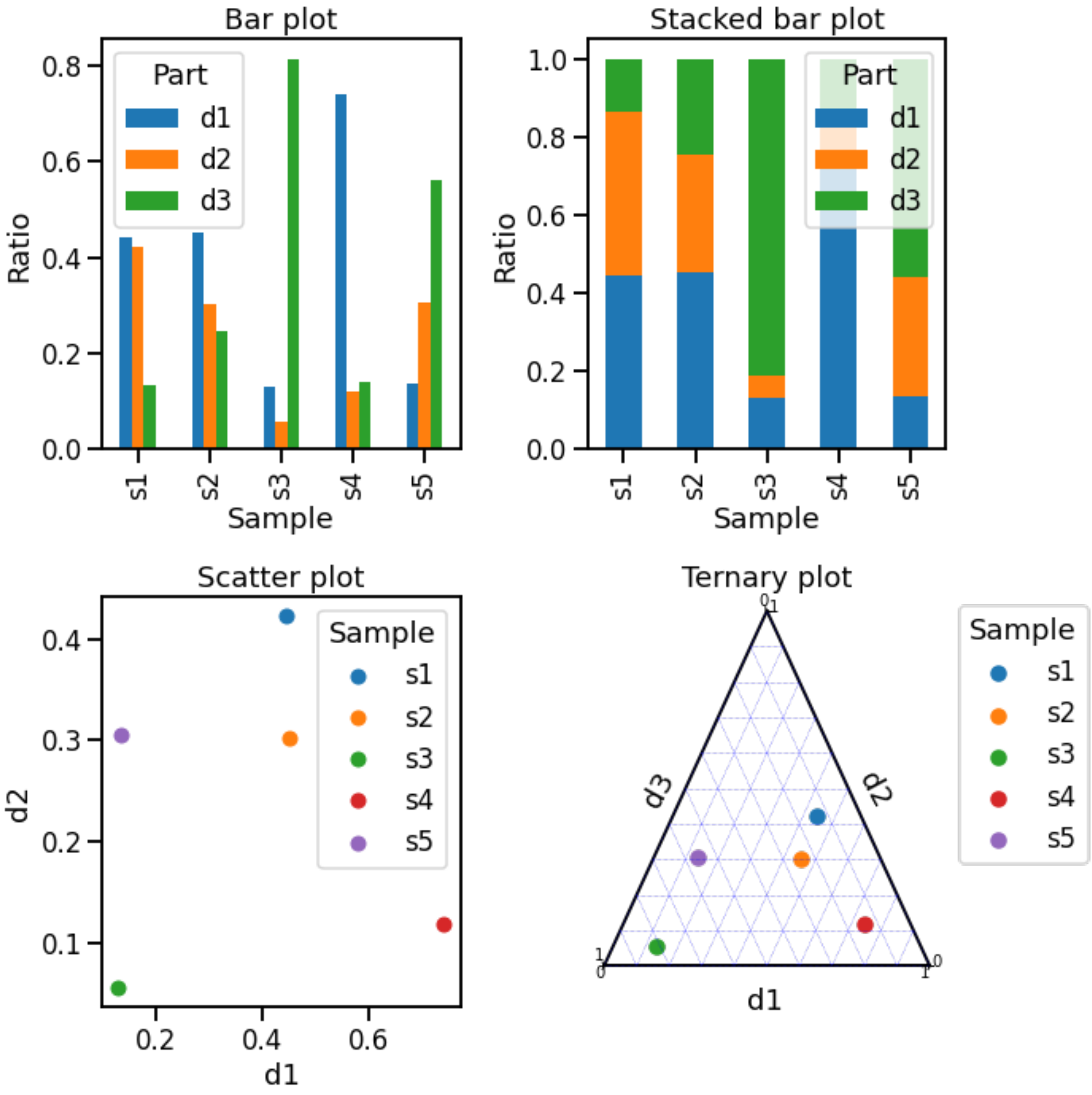
CoDa in practice

Visualizing abundances

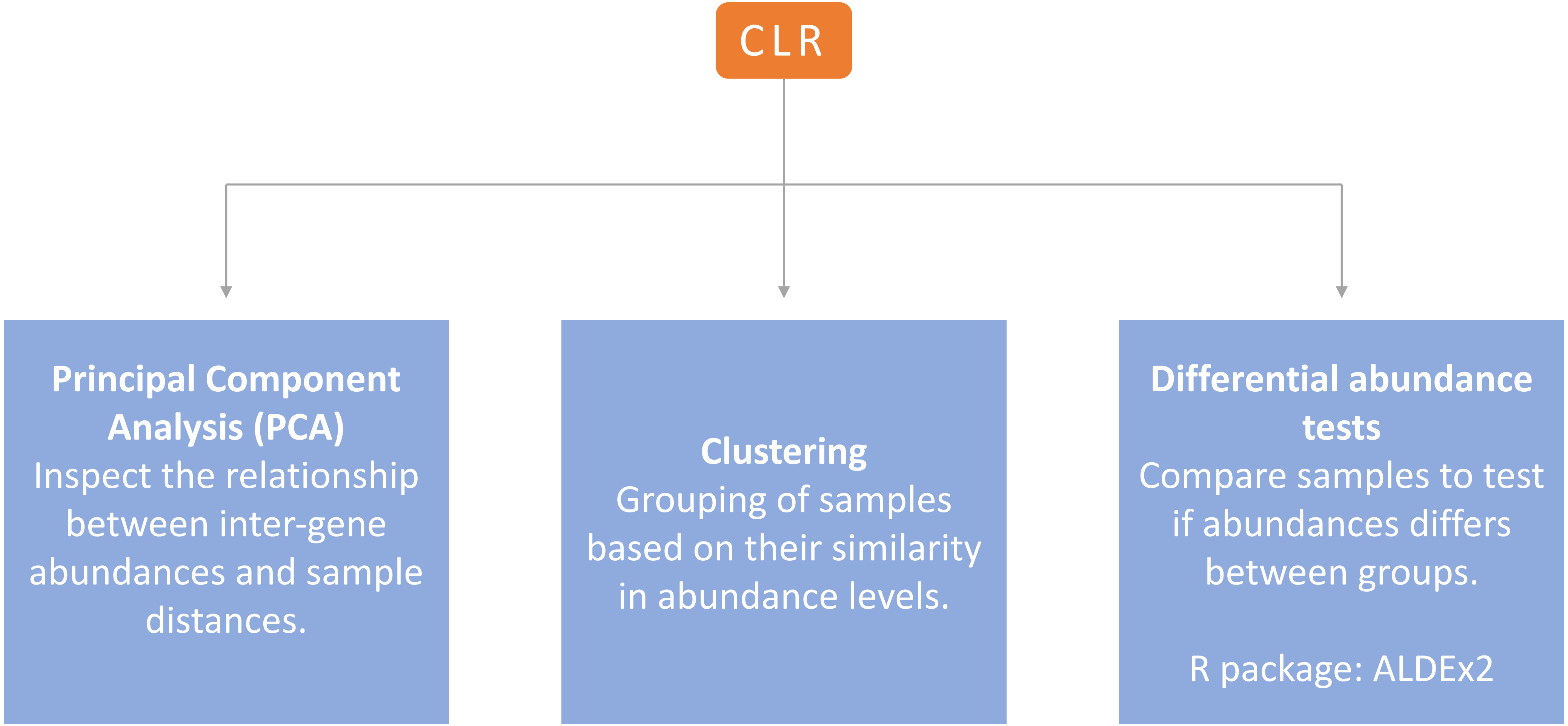
Barplots
matplotlib.pyplot.bar
seaborn.barplot
ggplot2::geom_bar

Scatter-plot
matplotlib.pyplot.scatter
seaborn.scatterplot
ggplot2::geom_point

Ternary plots
ternary.plot
ggtern::ggtern



CoDa in practice – advanced uses



CLR

Principal Component Analysis (PCA)
Inspect the relationship between inter-gene abundances and sample distances.

Clustering
Grouping of samples based on their similarity in abundance levels.

Differential abundance tests
Compare samples to test if abundances differs between groups.

R package: ALDEx2

Recommended reading

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, *8*, 2224.

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics & informatics*, *17*(1).

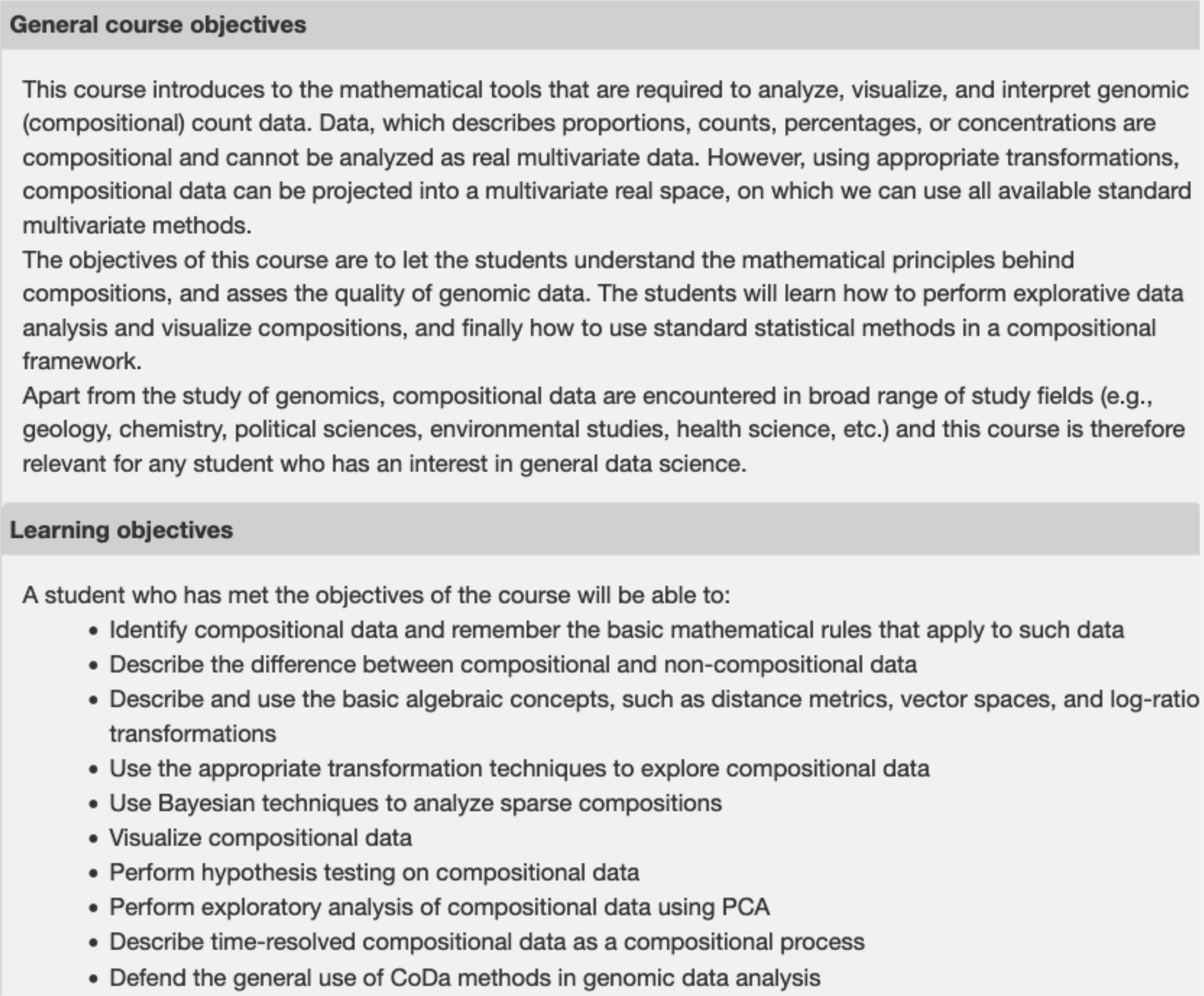
Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Want to know more about CoDa?

23257 Compositional data analysis with applications in genomics

5 ECTS points

F2A



General course objectives

This course introduces to the mathematical tools that are required to analyze, visualize, and interpret genomic (compositional) count data. Data, which describes proportions, counts, percentages, or concentrations are compositional and cannot be analyzed as real multivariate data. However, using appropriate transformations, compositional data can be projected into a multivariate real space, on which we can use all available standard multivariate methods.

The objectives of this course are to let the students understand the mathematical principles behind compositions, and asses the quality of genomic data. The students will learn how to perform explorative data analysis and visualize compositions, and finally how to use standard statistical methods in a compositional framework.

Apart from the study of genomics, compositional data are encountered in broad range of study fields (e.g., geology, chemistry, political sciences, environmental studies, health science, etc.) and this course is therefore relevant for any student who has an interest in general data science.

Learning objectives

A student who has met the objectives of the course will be able to:

- Identify compositional data and remember the basic mathematical rules that apply to such data
- Describe the difference between compositional and non-compositional data
- Describe and use the basic algebraic concepts, such as distance metrics, vector spaces, and log-ratio transformations
- Use the appropriate transformation techniques to explore compositional data
- Use Bayesian techniques to analyze sparse compositions
- Visualize compositional data
- Perform hypothesis testing on compositional data
- Perform exploratory analysis of compositional data using PCA
- Describe time-resolved compositional data as a compositional process
- Defend the general use of CoDa methods in genomic data analysis

Exercises for today

Exercises covering the basis CoDa functions on a small example dataset

Write code in R or Python for abundance analysis of KMA mapping results