Statistical concepts relevant to microbiome data science

Georg Zeller, Morgan Essex, Quinten Ducarmon

Microbiome Systems Biology Group
Leiden University Center of Infectious Diseases (LUCID)
LUMC

www.zellerlab.org

Learning goals of this lecture

- Recap how to set up **hypothesis tests**
- Recap what a **p-value** is (and what it is not)
- Understand the **multiple hypothesis testing** problem and how to bound the FDR
- Understand the difference between **effect size** and significance
- Get an intuitive understanding of **Principal Component Analysis** and what it's good for

Hypothesis testing

Hypothesis testing in simple terms

Goal: determine if a phenomenon could be **observed by chance**

More precisely **upper-bound the probability** that we **falsely report a chance phenomenon** as interesting ("significant")

1. Decide on a **quantity of interest**,
   pick a data summary function and test statistic
2. Specify a **null hypothesis**
   (a null model to compute the null distribution)
3. Decide how certain you want to be
   (**rejection region**, "alpha" for rejecting the null hypothesis)
4. Collect the data and compute the test statistic
5. Decide if to reject the null hypothsis or not

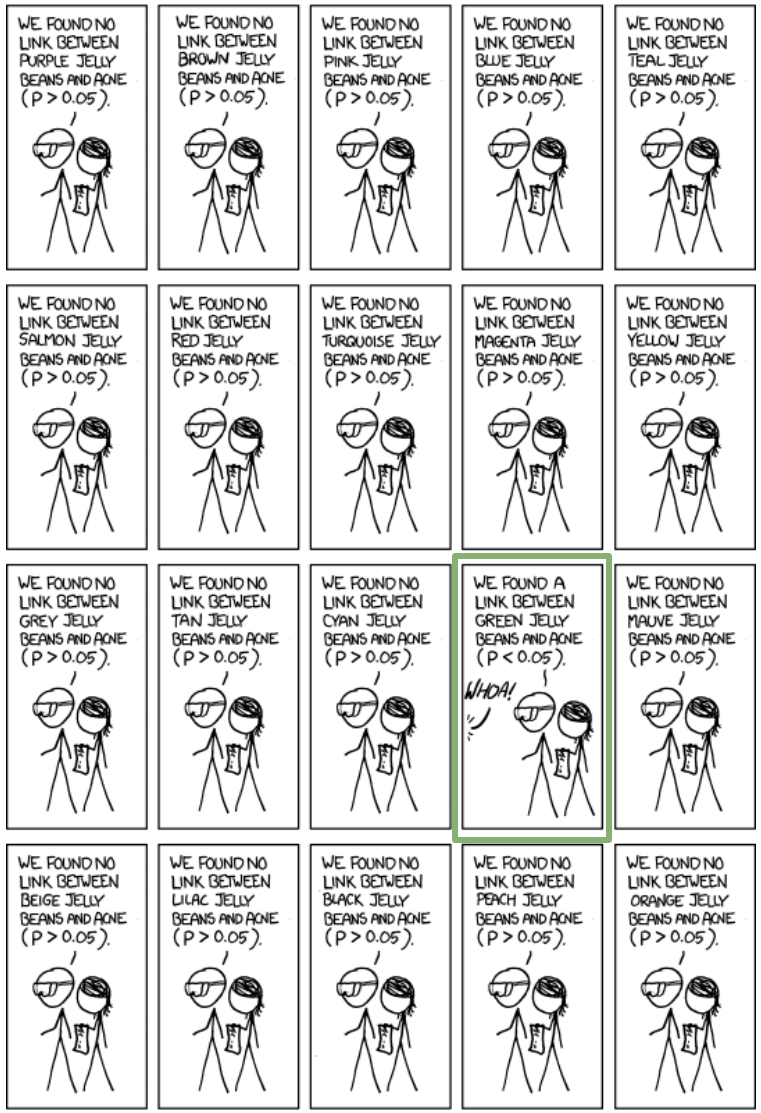
In **microbiome** terms

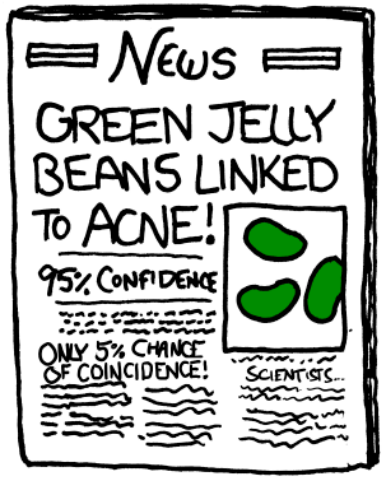
1. Rel. abundance ranks of case and control samples (Wilcoxon rank sum test)
2. Cases have similar ranks (similar relative abundances) as controls
3. Specify alpah < 0.05 (for enriched and depleted CRC-associated microbial taxa)
4. (you'll see later)
5. Is the Wilcoxon p-value < 0.05?

- This procedure (as all hypothesis tests) **controls** the **false-positive rate** (via alpha):
  The proportion of false positive test results among all tests conducted
- It does **not control** the **false negative rate**:
  How often we fail to reject the null hypothesis when the data is nonrandom

The multiple hypothesis testing problem

Multiple testing correction





- Since we test **several hundreds of cases**, some tests will be “significant” by chance
- If the null hypothesis is true in all of 1000 cases tested, how many false positives do you expect?
- If the null hypothesis is false in 50 out of 1000 cases tested (and you can detect all these cases), how many of your significant results do you expect to be wrong?
- The **false discovery rate** (**FDR**): the proportion of wrongly rejected nulls among all rejected nulls (= the proportion of false positives among all positive tests)
- To **control the FDR**, we perform **multiple testing correction** using the **Benjamini-Hochberg procedure**

https://www.xkcd.com/882/

Effect size vs statistical significance

Caveat: significance not to be confused with effect size

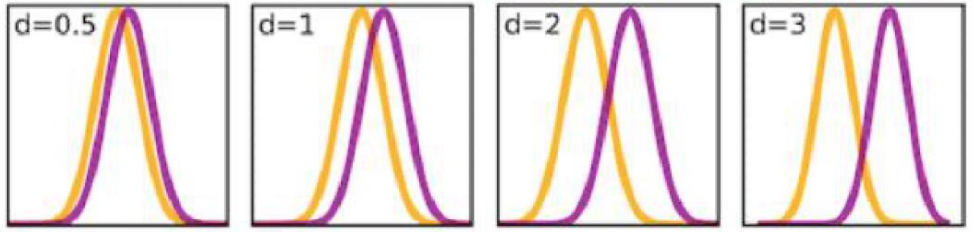
- **Statistical significance** (a small p-value) does not mean that the difference is big, important or biologically significant.

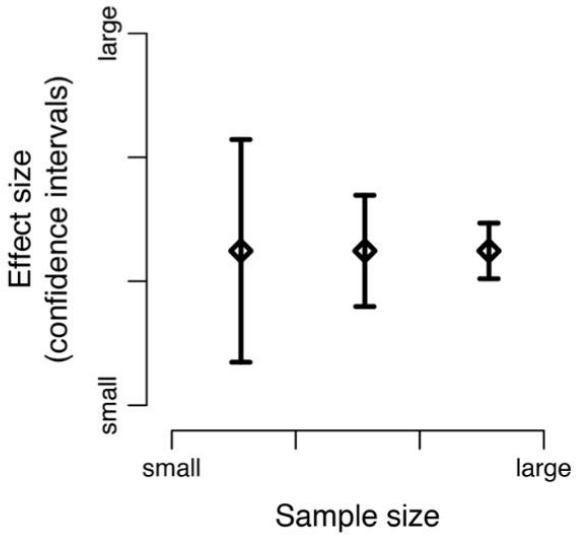
  It simply means you can be confident that the difference is not by chance.

- Any (even a tiny) difference can create a significant results if the **sample size** is large enough

- **Effect size** measures quantify how **large the effect** (e.g. difference between the means) is

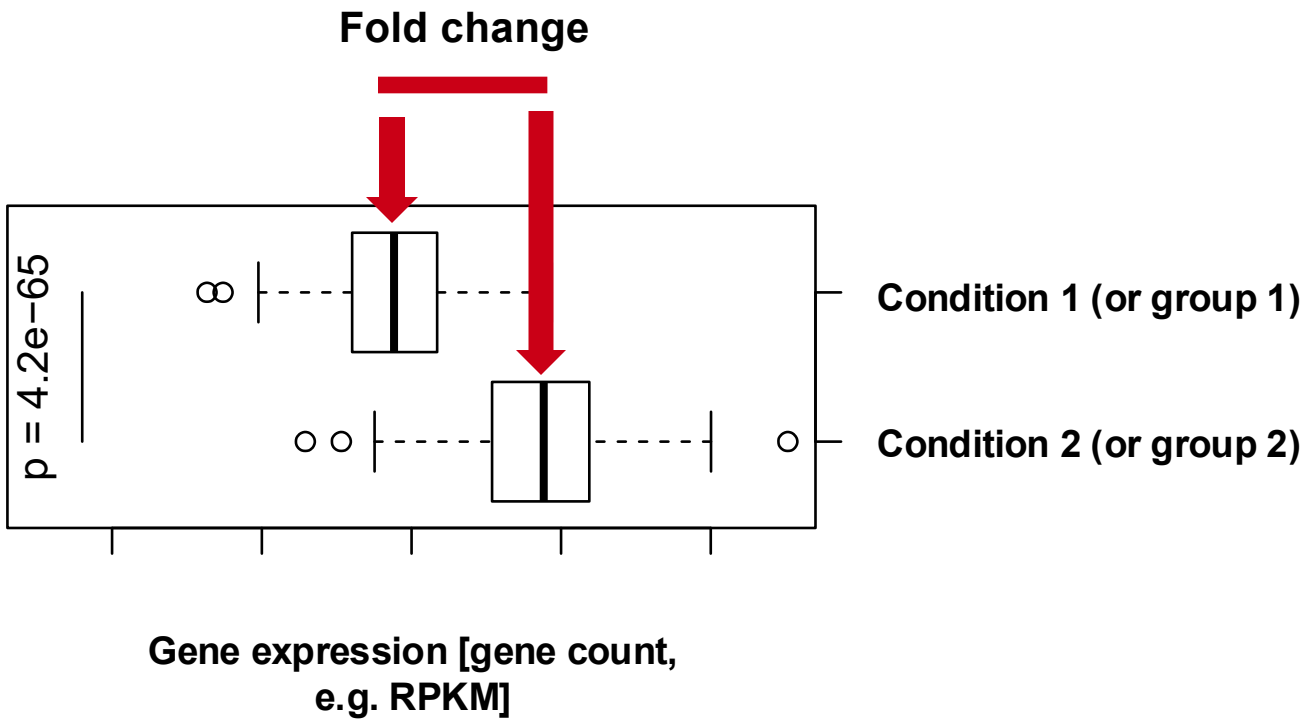
- With increasing sample size, you can **estimate the effect size more precisely**, but it is **not expected to increase**)





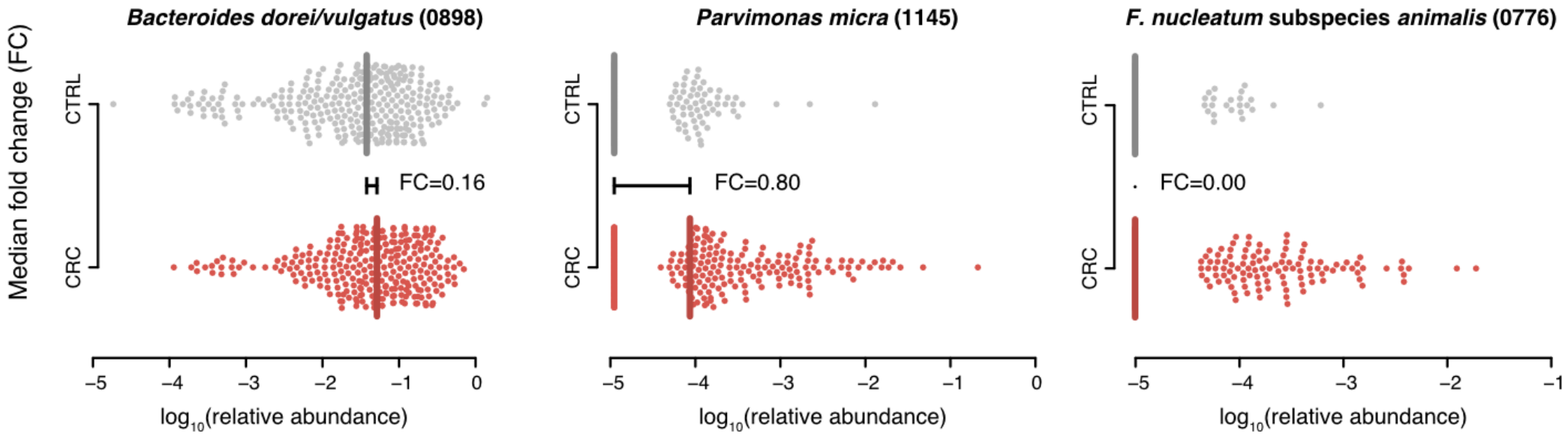
Fold change as a common effect size measure in transcriptomics

- Ratio of the expression level of a gene in one condition to its expression level in another condition
- Usually given on a log-scale
- Expression level per condition as average or median across measurements (replicates)



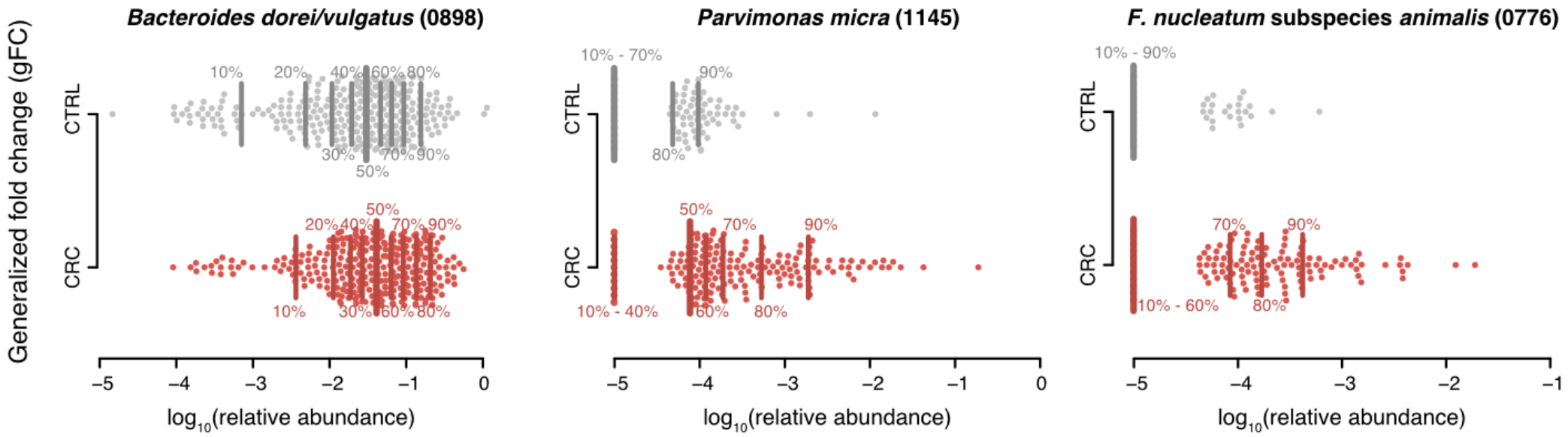
- A more realistic example would have 3-5 dots (replicates) per group

Generalized fold change as an effect size measure for microbiome



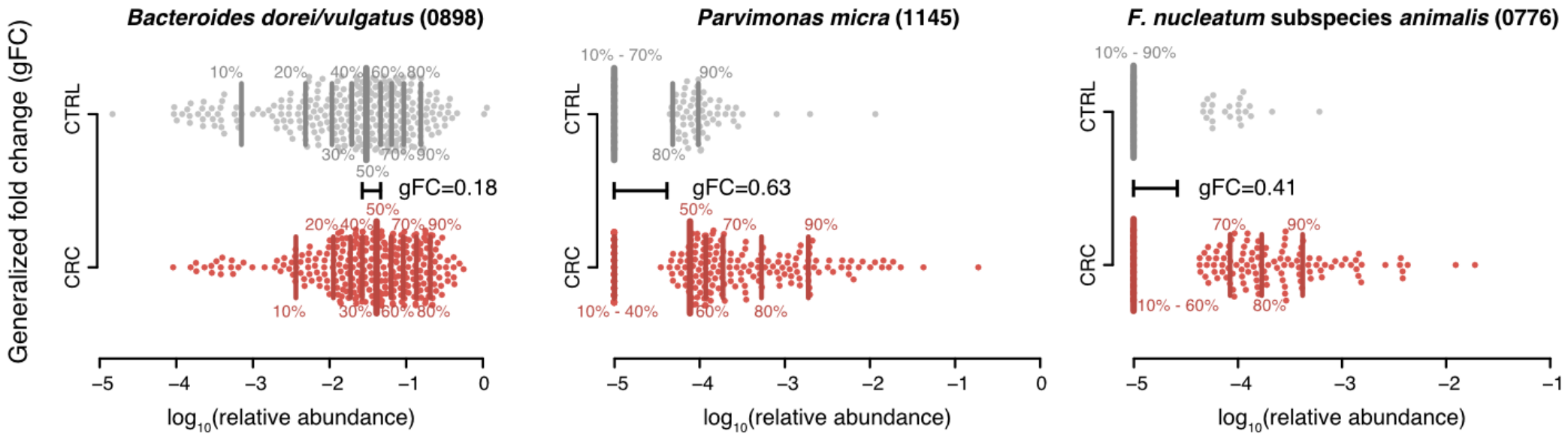
Wirbel*, Pyl* et al., *Nat. Med.* 2019

Generalized fold change as measure for effect size



Wirbel*, Pyl* et al., *Nat. Med.* 2019

Generalized fold change as measure for effect size

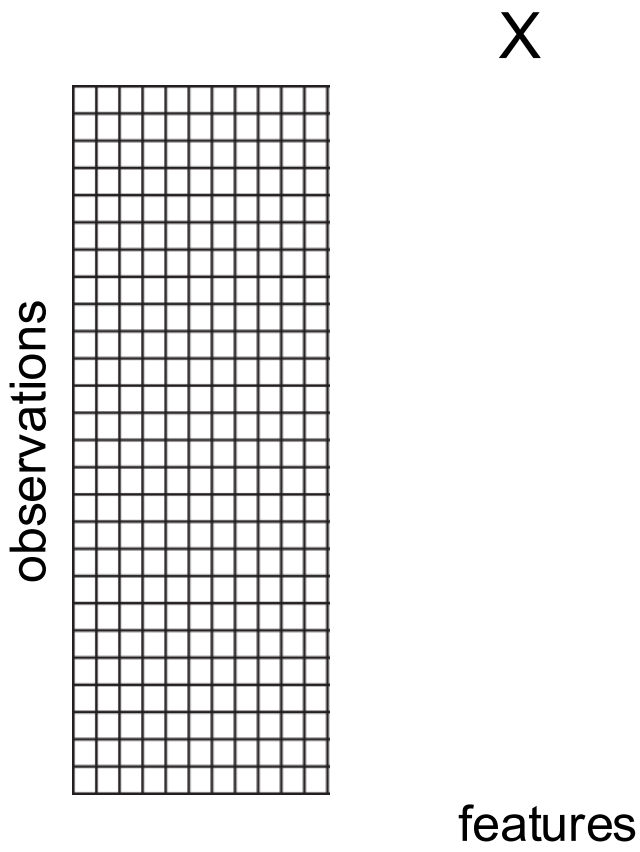


Wirbel*, Pyl* et al., *Nat. Med.* 2019

Principal Component Analysis

What is Principal Component Analysis PCA?

- A Multivariate Analysis (MVA) technique to **analyze more than one variable at a time**
- A tool to **reveal (strong) trends** affecting many (possibly correlated) variables of a high-dimensional data set
- A **dimensionality reduction technique:**
  - tries to **identify the most relevant dimensions** in the data
  - **reduces "noise"** (uninformative dimensions)
  - **enables visualization** of high-dimensional data
  - compactly represents highly correlated measurements / features



Projecting data onto a line

- In general when reducing / summarizing two-dimensional data (a plane) to one (a line) we **lose information**.
- Our goal is to **keep as much information as possible**.
- There are **many ways** to project a point cloud on a line. What is a good (principal) one?

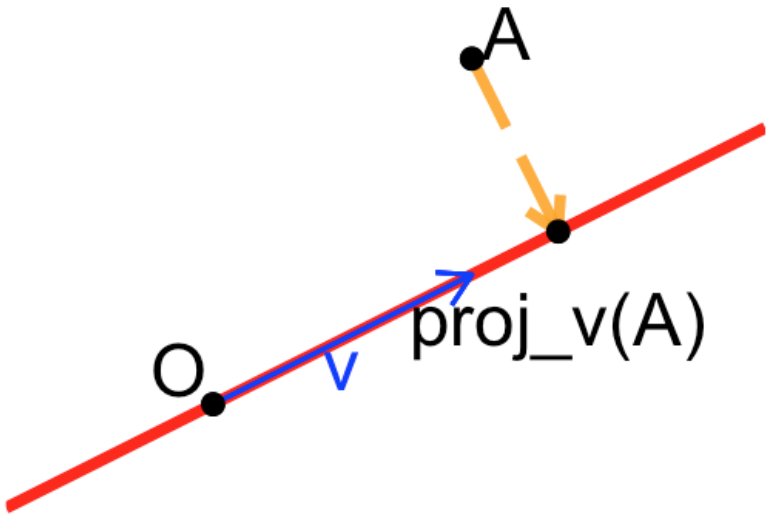
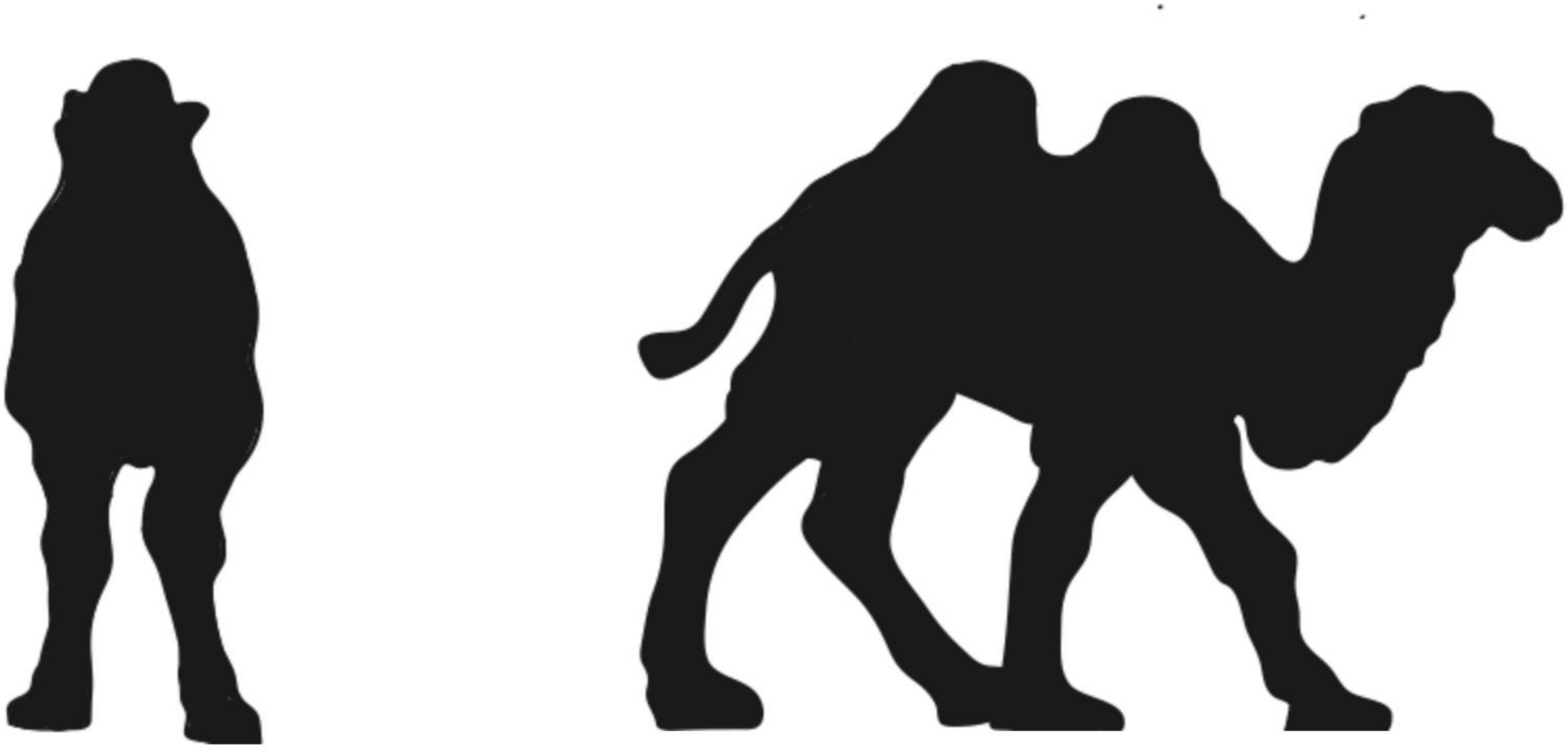


Figure 7.5: Point A is projected onto the red line generated by the vector v. The dashed projection line is perpendicular (or **orthogonal**) to the red line. The intersection point of the projection line and the red line is called the orthogonal projection of A onto the red line generated by the vector v.

Holmes & Huber, Modern Statistics for Modern Biology, 2019

What is a good projection?



Holmes & Huber, Modern Statistics for Modern Biology, 2019

Summary / take-home

- **Statistical hypothesis tests control** the probability of a **false positive** error
- When conducting many tests when true positives are expected to be rare (the usual case with omics data), the (vast) **majority of significant findings will be false positives**, therefore **adjustment** for multiple testing **is not optional**
- **Correcting for multiple testing** with the **false disovery method** controls the proportion of false positives among all significant results
- **Effect size** and **statistical significance** are not to be confused
- **Generalized fold change** is an effect size measure that works well for **sparse** microbiome data
- **Principal Component Analysis** (and generalizations, such as Principal Coordinate Analysis and Multidimensional Scaling) is a great tool to **detect & visualize trends** in omics data and **reduce dimensionality**

Thank you!