A primer on data-driven human gut microbiome research with a focus on colorectal cancer

Georg Zeller, Morgan Essex, Quinten Ducarmon

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

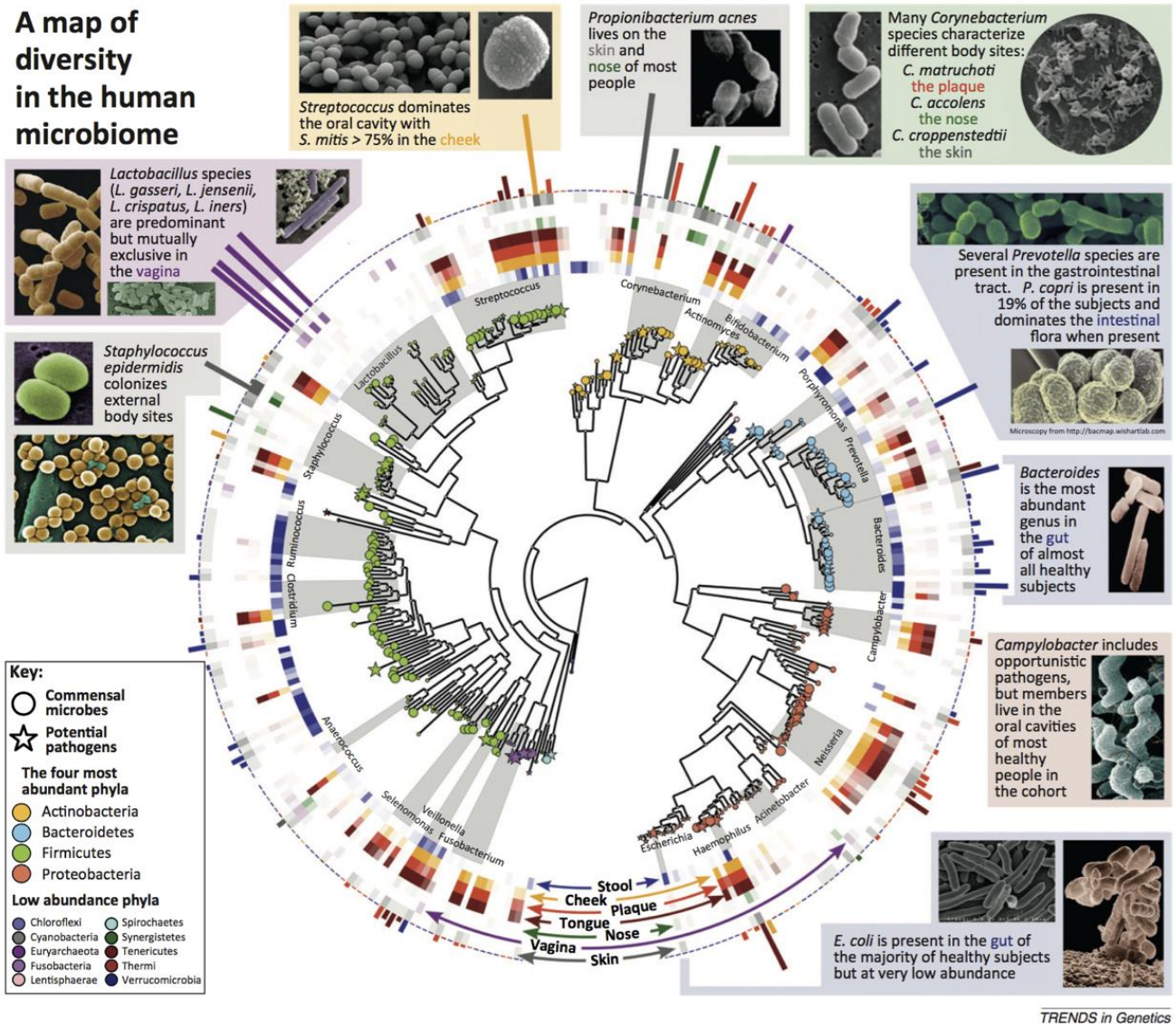
www.zellerlab.org

Learning goals of this lecture – part I

- Recap key features of the **gut microbial ecosystem**
- Recap **microbiome approaches** and have a glance at **microbiome sequencing data**
- Understand how **the gut microbiome can impact host health**, locally and systemically
- Appreciate how microbes could influence **(colorectal) cancer development**

What is the human microbiome?

Diverse microbial ecosystems colonize our body



Morgan et al., *Trends Genet.* 2013

Terminology

Microbiota:
The micro-organisms (bacteria, archaea, viruses and eukaryotes) in a particular environment

Microbiome:
the entire habitat including the microbes, their genes, metabolites and the surrounding milieu
(often used as synonym for microbiota)

Metagenome:
The collective genomes of the microbiota

Marchesi et al., *Microbiome* 2015

The human gut microbiome

Artwork by Charis Tsevis

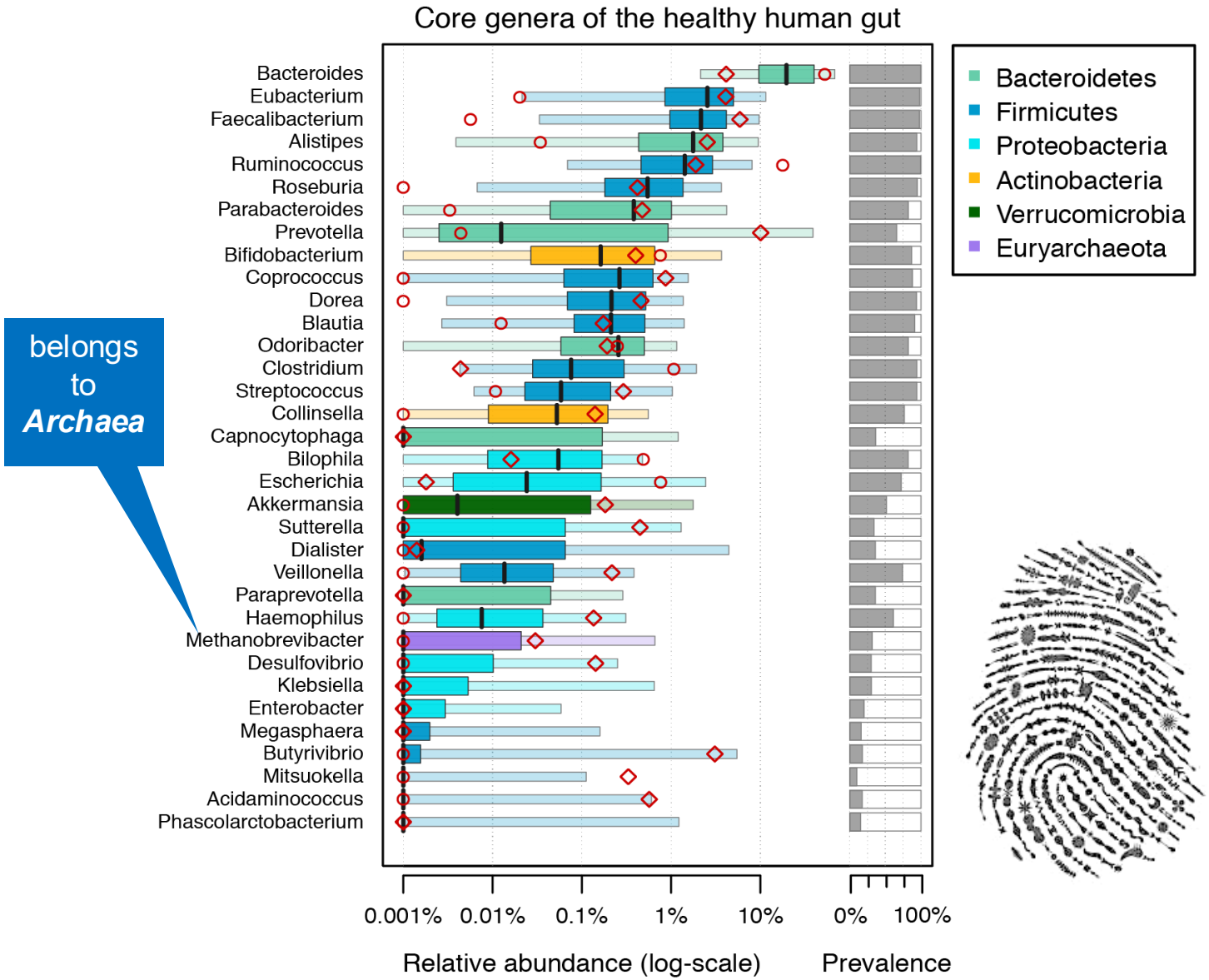
1 human host species

~23,000 genes in the human genome

~250 – 500 prokaryotic species
(plus viruses & fungi)

>1 million prokaryotic genes
in the gut metagenome

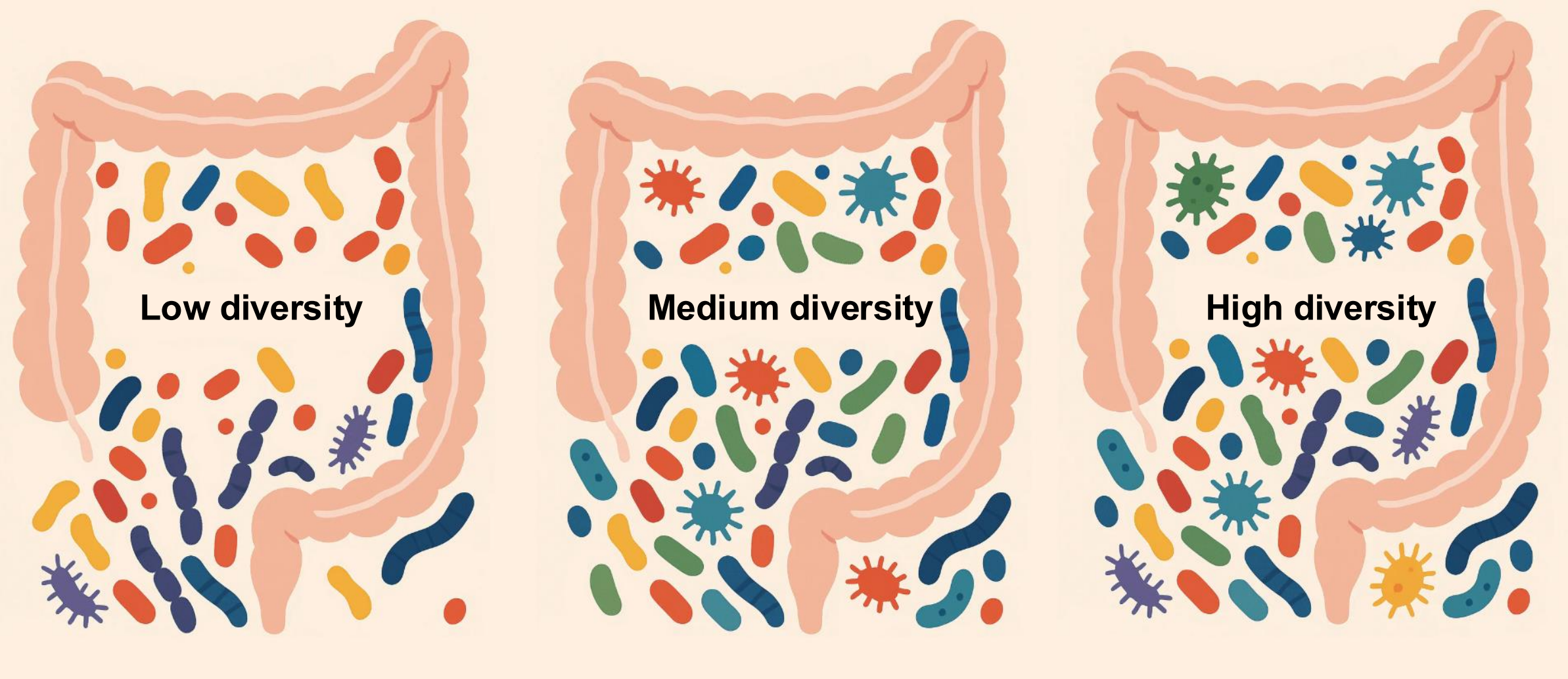
Gut microbiome composition and individuality



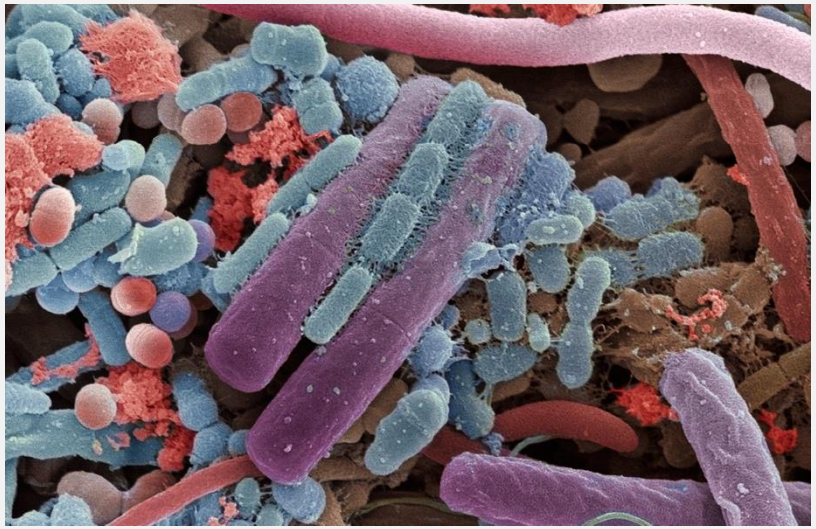
belongs to ***Archaea***

- Human **individuals vary far more in gut microbiome** composition than in their genomes and transcriptomes
- Prokaryotic relative abundance **varies over 5+ orders of magnitude**
- In many individuals ***Bacteroides*** is the dominant genus, in others this can be ***Prevotella*** or others

Diversity of microbial communities



The human gut microbiome – summary of key characteristics



Vast diversity:
each of us carries **hundreds of gut microbial species**

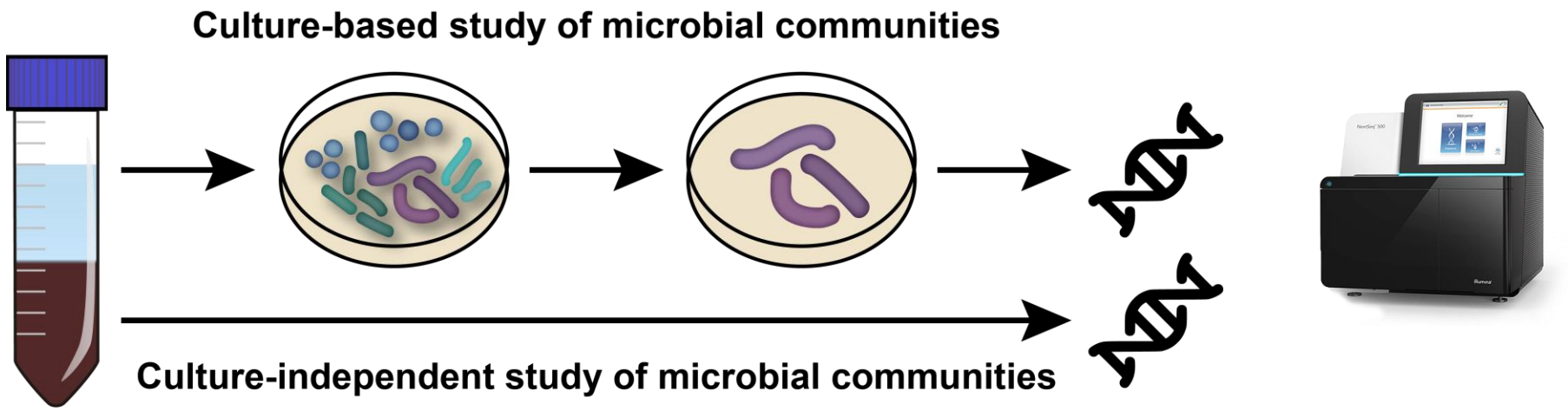
Microbiome composition **varies greatly between individuals**

Microbiome composition **changes** gradually over human **lifetime** (while staying individual-specific)

How to study the human microbiome using sequencing?

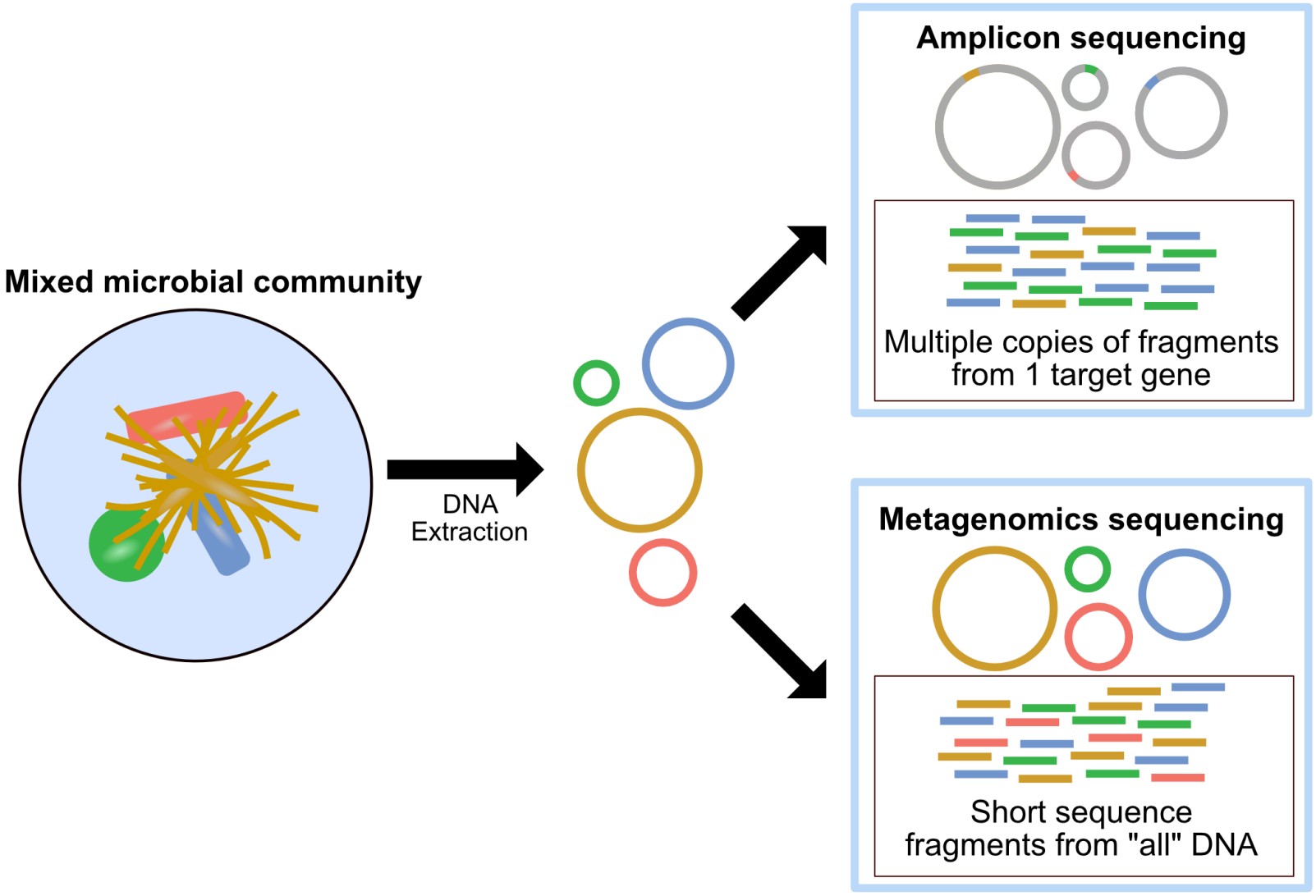
Culture-independent sequencing of microbial communities

Classical microbiology approach



Revolutionized modern microbiome research

Shotgun metagenomics *vs.* 16S rRNA amplicon sequencing



Amplicon sequencing:

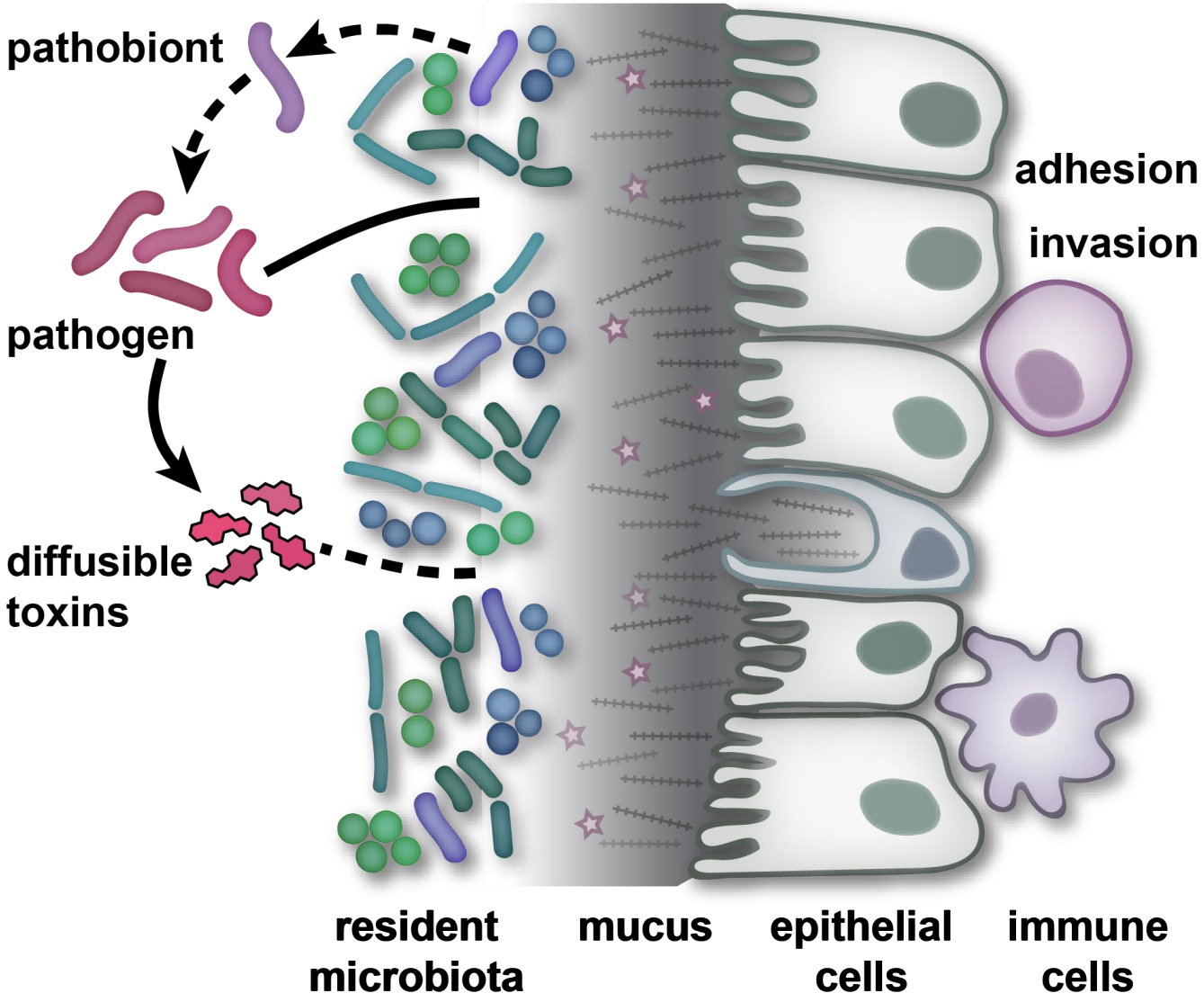
PCR-amplification and sequencing of 16S rRNA gene fragments

Metagenomic sequencing:

sequencing of the whole DNA complement of an environmental sample

How do microbiome-host interactions impact host health?

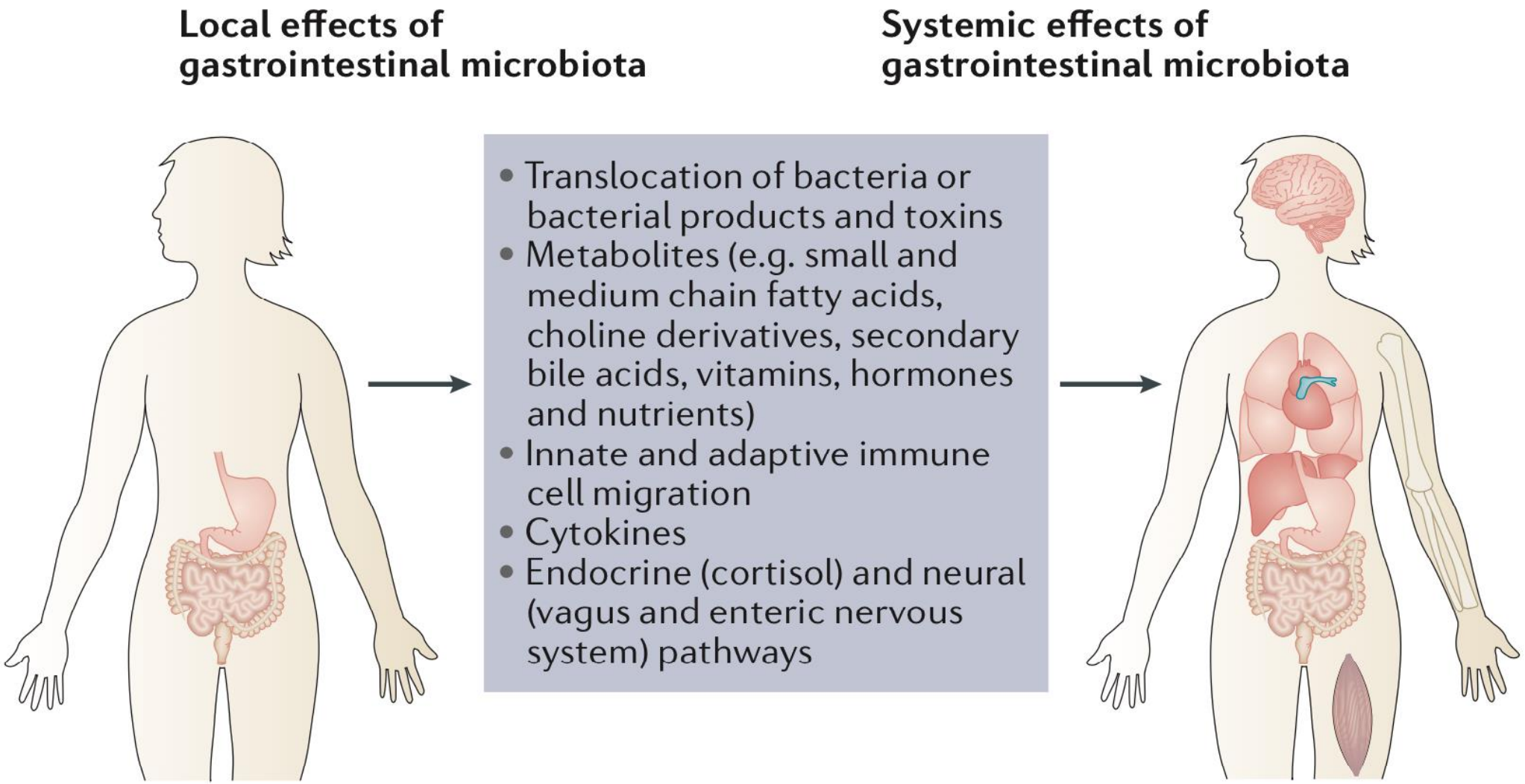
The gut mucosal barrier can be compromised



- **Healthy epithelium is shielded** by mucus, innate and adaptive immunity from luminal bacteria
- Some gut bacteria can **degrade mucus**
- **Invasive pathogens** can penetrate the barrier
- Compromised barrier integrity allows **bacteria and their metabolites to translocate** into circulation
- Translocation can cause **liver** and **systemic inflammation**

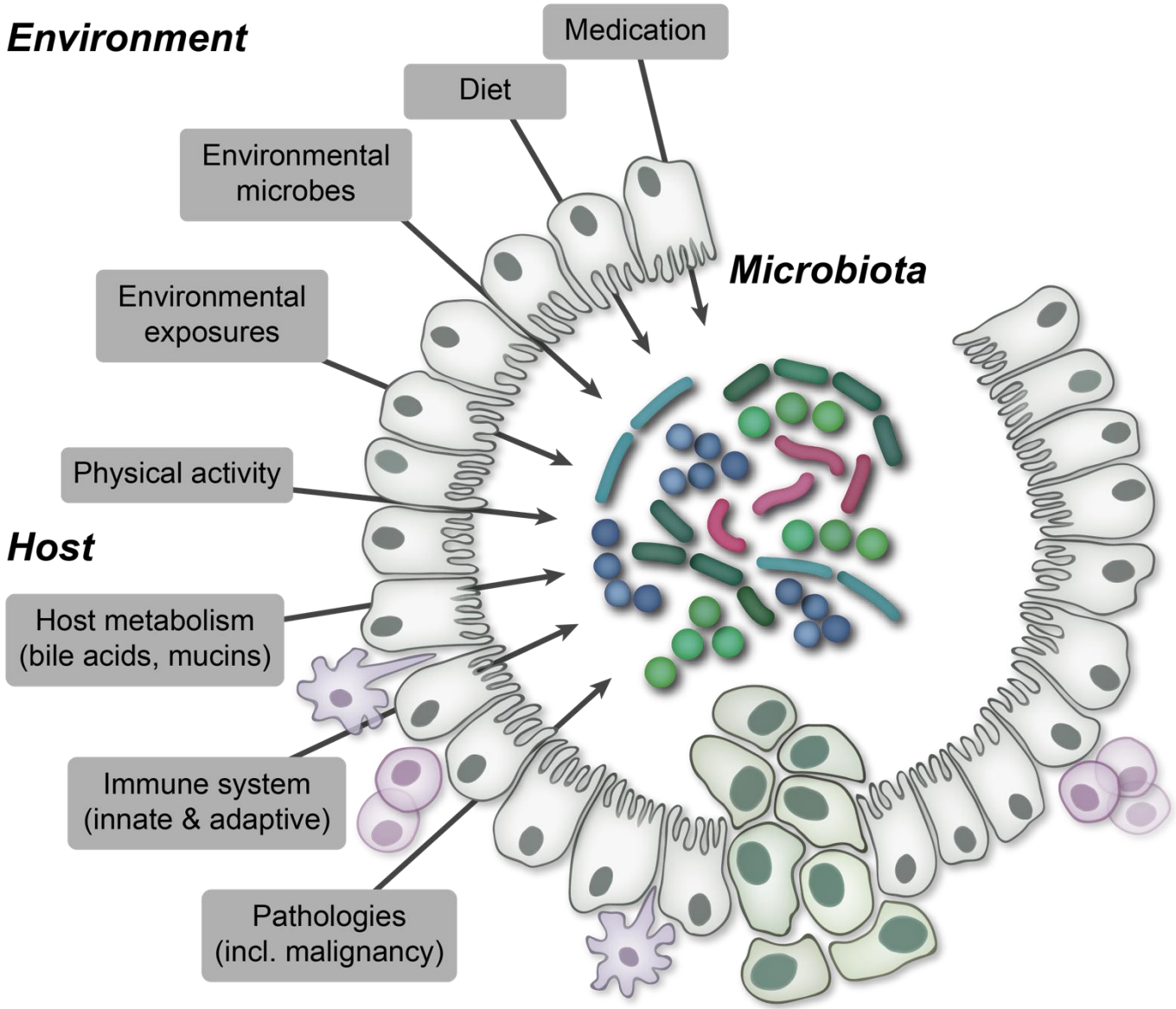
Buylaert et al., *Mitt. Dt. Gesellsch. Nephrol.* 2018

Gut microbiota can have both local and systemic effects

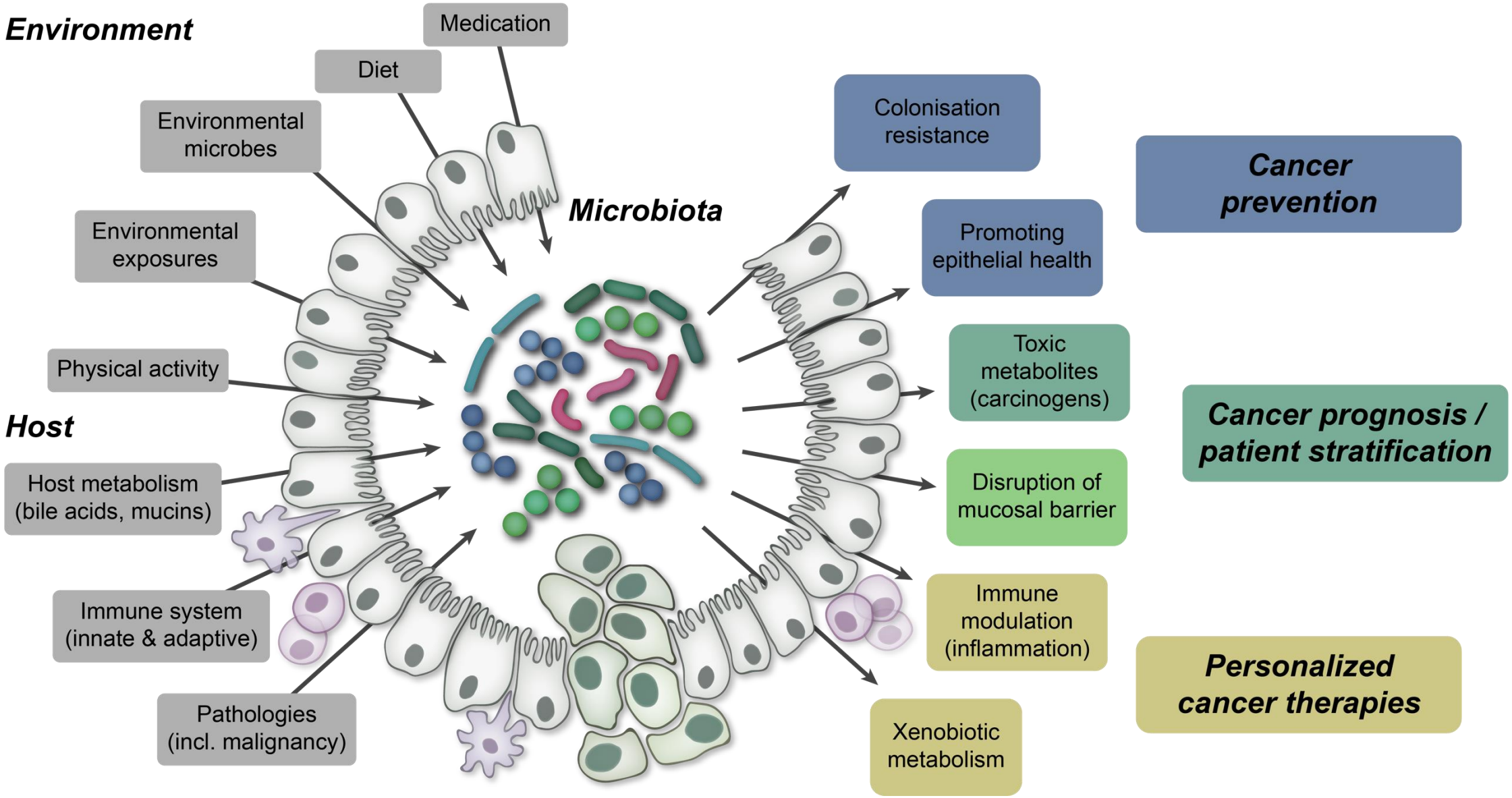


Roy & Trinchieri, *Nat. Rev. Cancer* 2017

Gut microbiome shaped by environmental and host-intrinsic factors



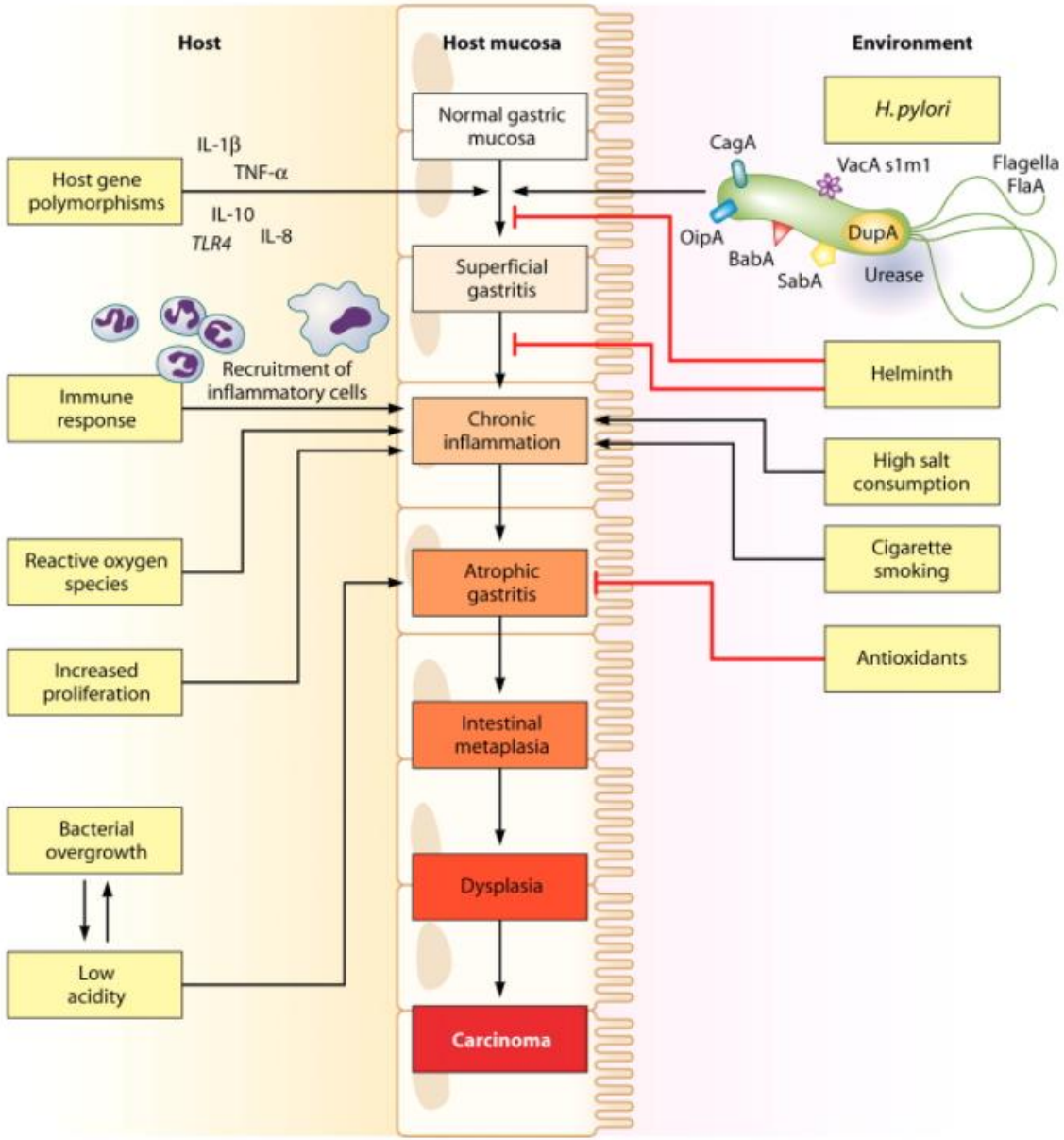
Gut microbiome as individual-specific modulator of environmental risks



Microorganisms and cancer

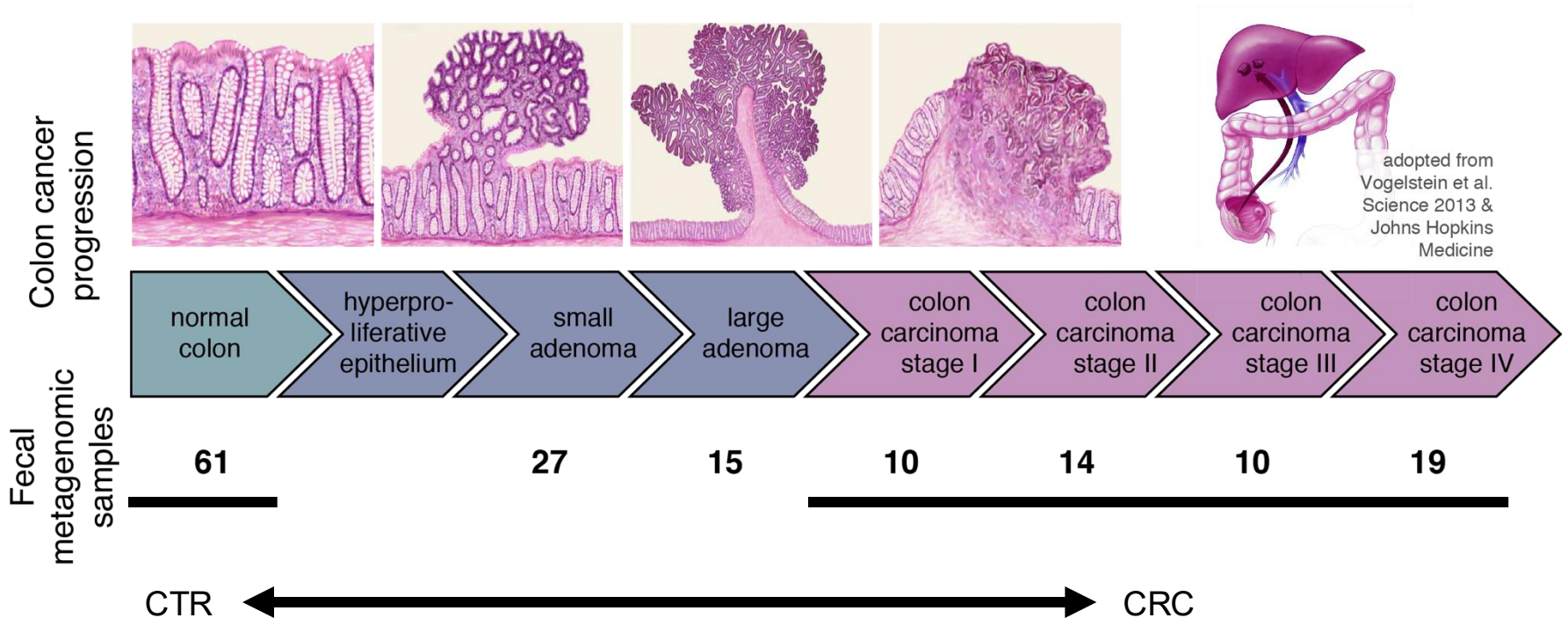
Helicobacter pylori as a strong risk factor for gastric cancer

- For 2018, 2.2 million cancer cases attributable to infection
- 800,000 cases (stomach cancer, MALT lymphoma) attributable to *Helicobacter pylori*
- However, most infected individuals never develop cancer and eradication is not necessarily indicated



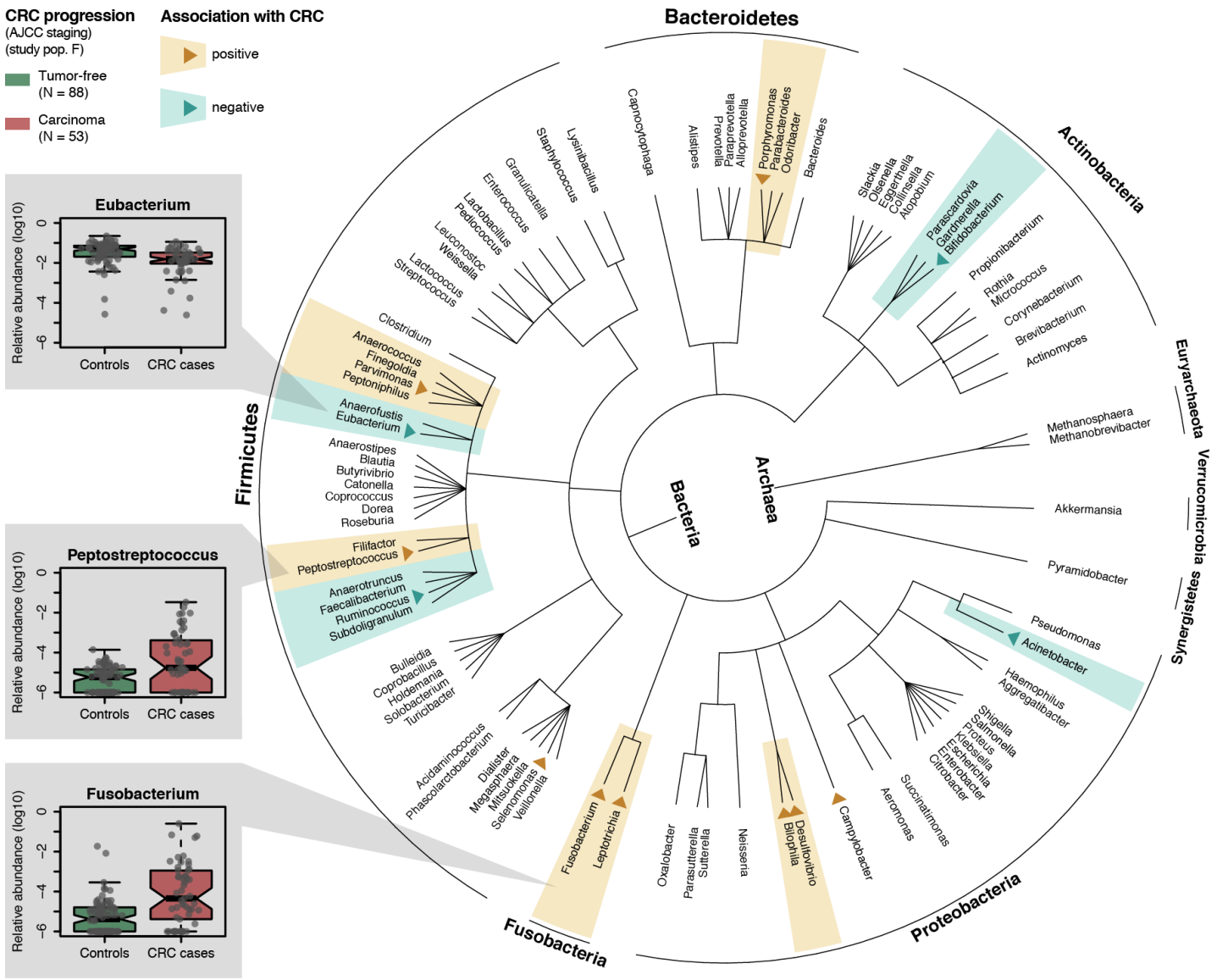
De Martel et al., *Lancet Global Health* 2020
Wroblewski et al., *Clin. Microbiol. Rev.* 2010

Microbiome association study with colorectal cancer (CRC)



- Associated gut microbiome changes with colorectal cancer (CRC)
- Mined fecal metagenome for **diagnostic biomarkers for CRC** towards a non-invasive screening test for reducing CRC mortality
- **Goal:** Explore feasibility of CRC detection from fecal metagenomics

Microbial genera associated with colorectal cancer (CRC)



Which gut microbial genera…

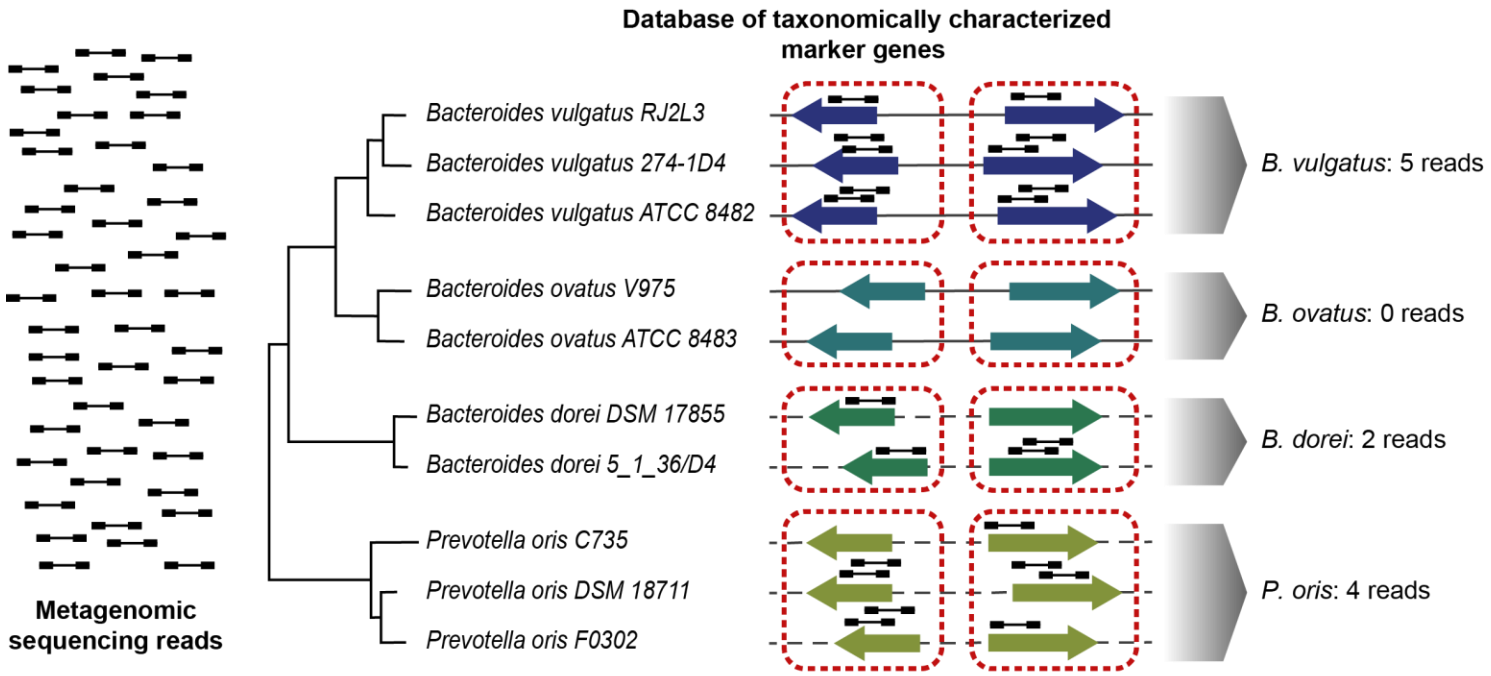
- are significantly enriched in CRC patients (relative to controls)?
- are significantly depleted in CRC patients (relative to controls)?

Microbiome data science

Learning goals of this lecture – part II

- Understand a **sequencing read count table** from a real **microbiome** study
- Start learning how to **formalize research questions as data analysis tasks**
- Familiarize yourselves with basic concepts of **exploratory** (microbiome) **data analysis**
- Recap what **R**, **tidyverse** and **RStudio** are
- **Start exploring data**…

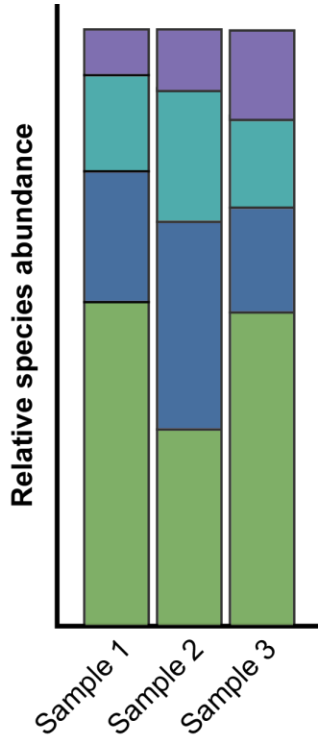
Taxonomic profiling: taking census of “who’s there”



Step 1
Map metagenomic reads against a database to identify their source organism

Step 2
Tabulate read counts across taxa and samples

	Species 1	Species 2	Species 3	Species 4
Sample 1	38	60	77	205
Sample 2	25	55	86	82
Sample 3	18	19	22	65



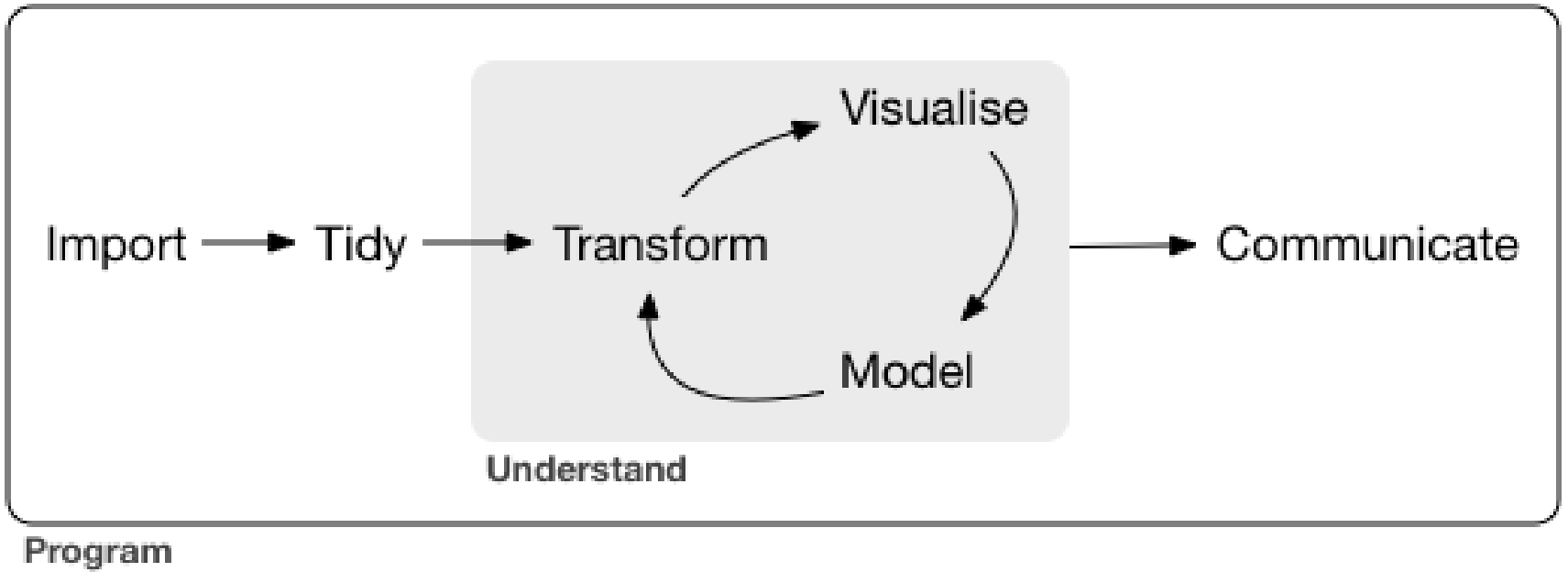
Exploratory data analysis

According to Wikipedia, "**exploratory data analysis (EDA)** is an approach of analyzing data sets to **summarize their main characteristics**, often using […] **data visualization** methods."

- "Primarily EDA is for seeing what the data can tell" thereby **enabling discoveries**
- It is **not** about implementing a pre-defined analysis as in **clinical trials with defined endpoints**
- **Assess assumptions for modeling** to support the selection of statistical tools

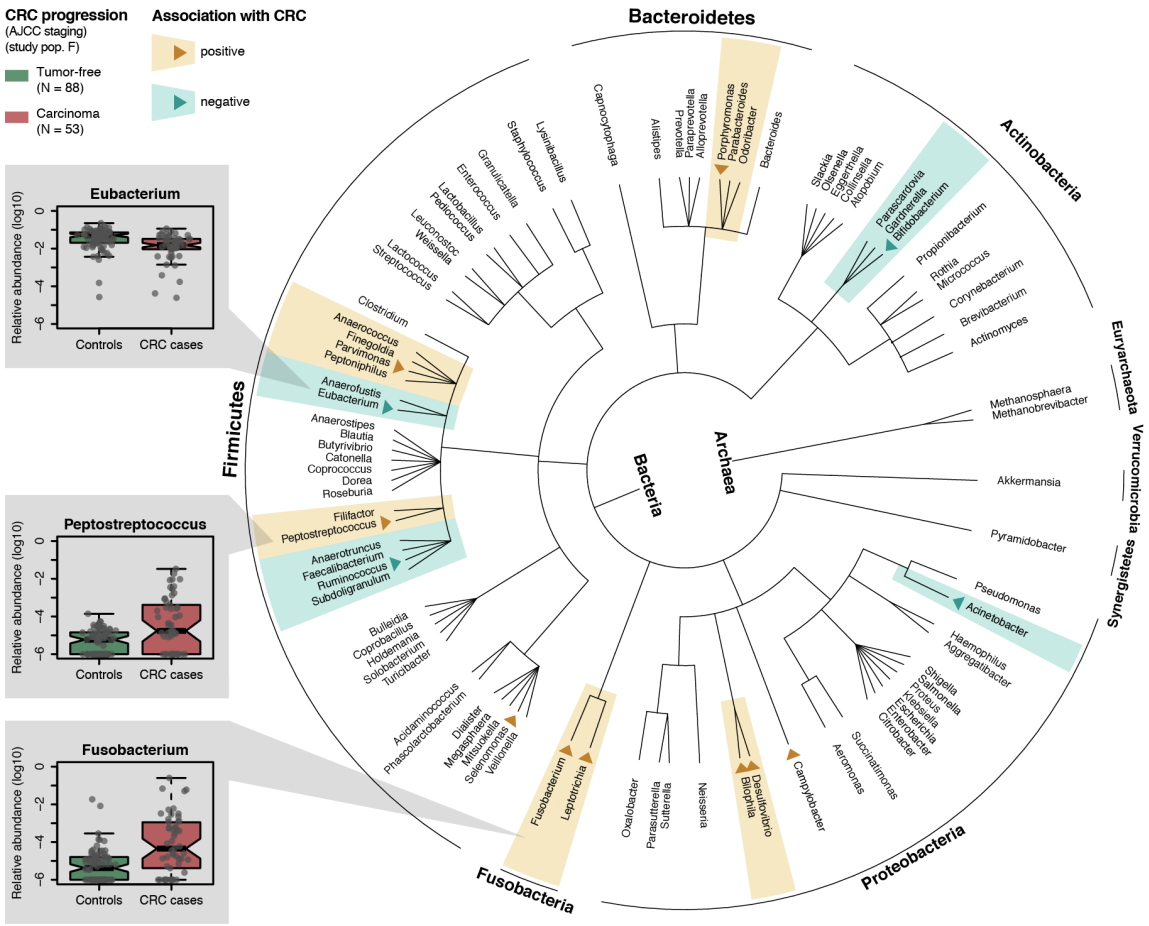
	Species 1	Species 2	Species 3	Species 4
Sample 1	38	60	77	205
Sample 2	25	55	86	82
Sample 3	18	19	22	65

Tabular data



Asking microbiome data science questions

- What is the **relative abundance** of each taxon and how does it **vary across individuals**?
- Are there **significant differences** in the abundances of certain taxa **between groups**?

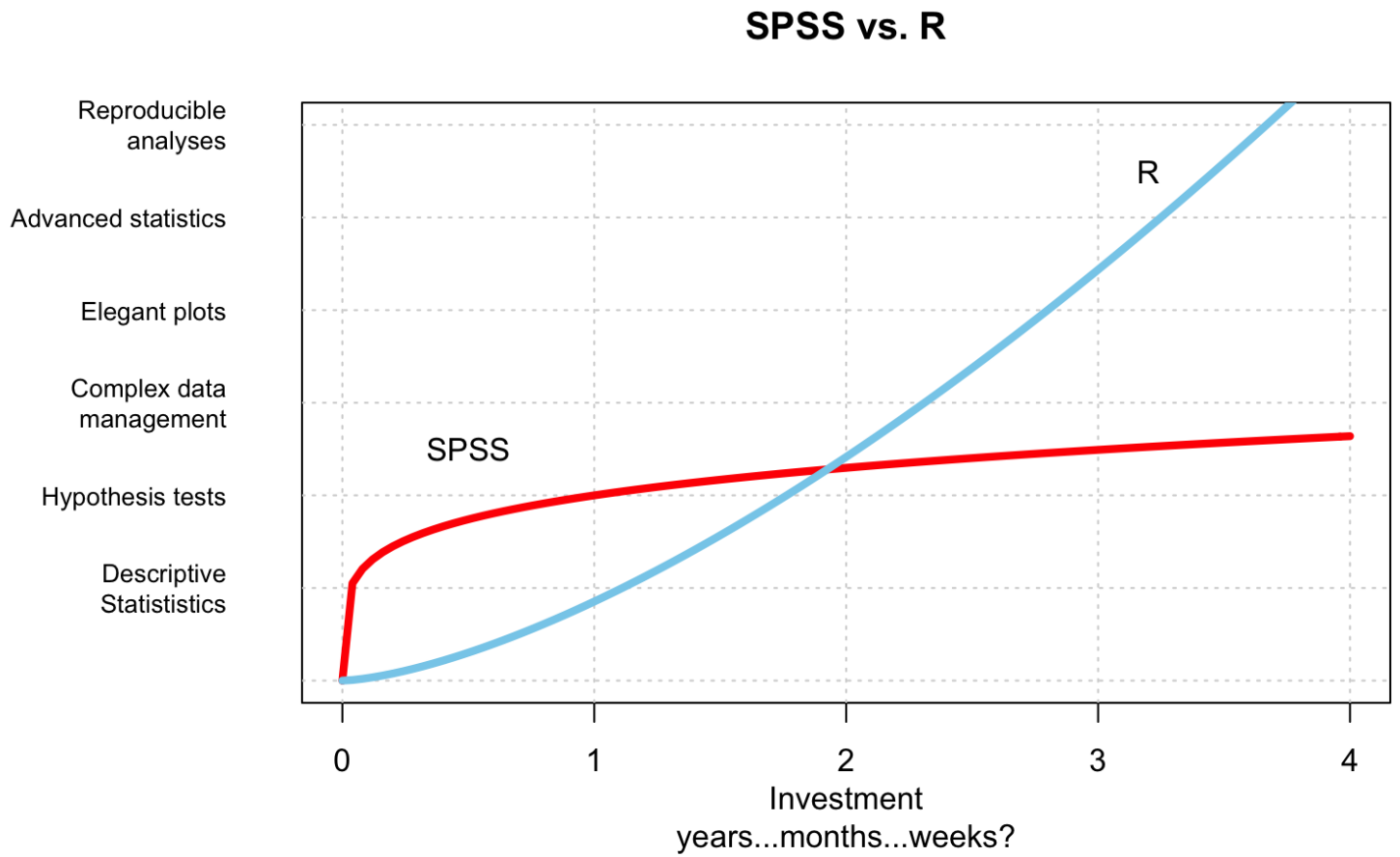


R has a learning curve, but it's worth it

- R is an **open-source statistical programming language** with a **huge developer community**

- **RStudio** is a computer application to write and execute R code

- **SPSS** increasingly obsolete

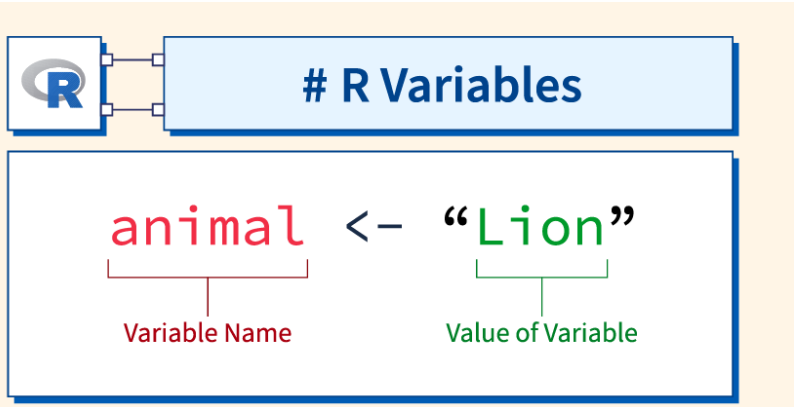


Computer science 101

Script: text file with words written in a programming language (script.R)

Editor / IDE: computer application used to write and execute scripts (Rstudio)

Within an (R) program, information is stored in *variables* and *data structures:* “assigning values to variables”



Variables	Example
integer	100
numeric	0.05
character	“hello”
logical	TRUE
factor	“Green”

Basic data structures in R

1 data type

Vectors

Any number of elements
(of same type)

Matrices

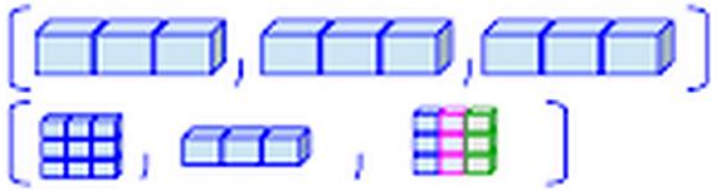
Homogeneous (usually numeric) data organized in rows and columns

"1D" — *"2D"*

Lists

Any number of (heterogeneous) elements

Recursive (list of lists)



Data Frames

List of same-length vectors

1+ data types

Computer science 101 (cont'd)

Syntax: rules that govern a programming language (+ most common type of errors!)

Function: a unit of code designed to perform a specific (defined) task

Argument(s): input(s) to a function (variable name/type)

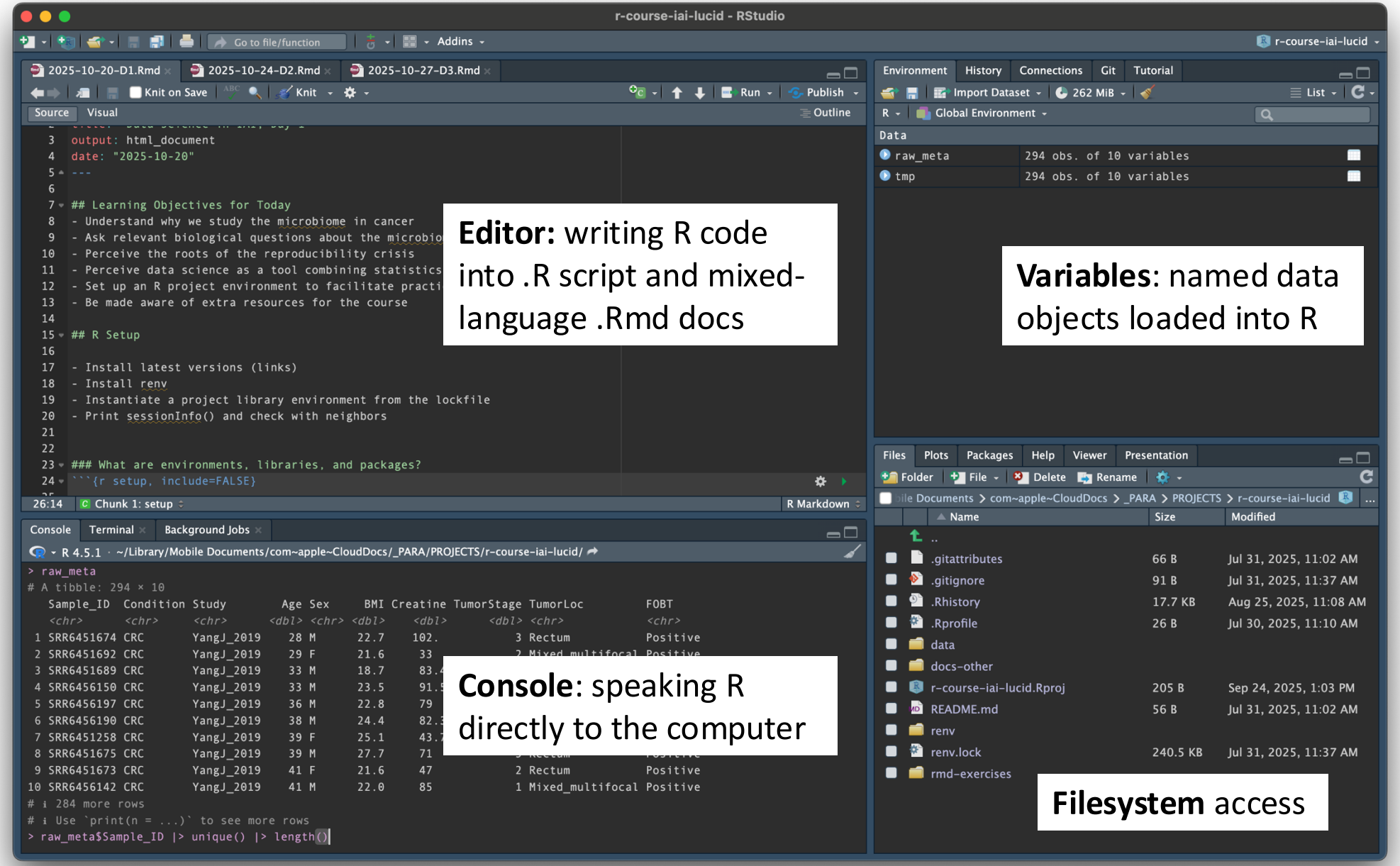
Library/Package: related sets of functions that are stored and work together

```
1  # comments start with a hashtag
2  everything after the # is ignored
3  # (and line 2 would cause an error!)

4  # loading functions into a script
5  library(package_name)
6  source("helper_script.R")

7  getwd() # where am I, computer?
```

Anatomy of RStudio



Summary / take-home

- The human gut microbiome is more variable than our genome and transcriptome, highly individual-specific, so in any given individual many common bacteria are absent
- Microbiome sequencing data is complex, associating microbiome composition with diseases, such as colorectal cancer, is a statistically challenging task
- Exploratory data analysis aims at exploring trends and group differences in complex (microbiome) data sets
- We will explore data analysis techniques and approaches that are usefulf for analysis of microbiome and many other types of high-dimensional omics data

Thank you!