

Microbiome and Metagenomics in Cancer

From Koch's postulates to pan-cancer metagenomics

Sean Davis, MD, PhD

University of Colorado Anschutz School of Medicine

July 22, 2026

PART 01

Establishing causation

How do you prove a microbe causes a disease? The answer has been rewritten twice already — and the microbiome forces a third rewrite.

SPEAKER NOTES

The organizing question for this section is not “what lives in us” but “how would we know if it mattered.”

The formal criteria for microbial causation have been rewritten twice in 150 years, and each rewrite was forced by a new technology: pure culture in the 1880s, molecular genetics in the 1980s, sequencing in the 1990s. The microbiome is forcing a third rewrite that is not finished. This matters for the rest of the talk, because the field’s most prominent recent failure — a retracted pan-cancer diagnostic — was a failure of causal reasoning, not of sequencing.

Koch's postulates

In 1876, Robert Koch traced anthrax to *Bacillus anthracis* by following the organism through its entire life cycle — and in doing so gave medicine its first rigorous standard for **causation**, not mere association.

Die Ätiologie der Milzbrand-Krankheit, begründet auf die Entwicklungsgeschichte des Bacillus Anthracis.¹⁾

Von

Dr. R. Koch,
Kreisphysikus in Wollstein.

Hierzu Tafel I.

I. *Einleitung.* Seit dem Auffinden der stäbchenförmigen Körper im Blute der an Milzbrand gestorbenen Tiere hat man sich vielfach Mühe gegeben, dieselben als die Ursache für die direkte Übertragbarkeit dieser Krankheit ebenso wie für das sporadische Auftreten derselben, also als das eigentliche Kontagium des Milzbrands nachzuweisen. In neuerer Zeit hatte sich hauptsächlich D a v a i n e mit dieser Aufgabe beschäftigt und, gestützt auf zahlreiche Impfversuche mit frischem oder getrocknetem stäbchenhaltigen Blute, mit aller Entschiedenheit dahin ausgesprochen, daß die Stäbchen Bakterien seien und nur beim Vorhandensein dieser Bakterien das Milzbrandblut die Krankheit von neuem zu erzeugen vermöge. Die ohne nachweisbare direkte Übertragung entstande-

SPEAKER NOTES

This is the original 1876 paper. Koch was a district physician working from a home laboratory, not a university appointment.

Others had already seen rod-shaped bodies in the blood of animals dead of anthrax, and Davaine had argued they were causal. Koch's addition was following the organism through its complete life cycle, including the spore stage — which explained why pastures remained infectious for years after an outbreak. He converted an observation anyone could make into a mechanism.

The standard being set is worth stating precisely. Not “the microbe is present in disease,” but “the microbe, and nothing else, produces the disease when introduced into a healthy host.” That is a much higher bar than association, and it is the bar most microbiome research cannot clear.

Koch's postulates

1. The microorganism is found in abundance in **diseased** but not in healthy individuals
2. The microorganism can be **isolated** from the diseased host and grown in pure culture
3. The cultured microorganism **causes disease** when inoculated into a healthy host
4. The microorganism can be **re-isolated** from the inoculated host and is the same as the original
5. **Elimination** of the microbe from the host alleviates disease

The load-bearing assumption

Every postulate assumes **one organism**, growable in **pure culture**, **sufficient on its own**. All three assumptions fail for the microbiome.

SPEAKER NOTES

The callout is the substance. Each assumption fails in a specific way.

One organism. Microbiome effects are typically community properties. Remove or add any single taxon and the effect usually does not reproduce, so the unit of causation is not an organism.

Pure culture. Most gut organisms historically resisted culture. Postulate 2 therefore excluded the majority of the interesting biology for a century, and is precisely why sequencing-based approaches took over.

Sufficient on its own. This fails even for the clean cases. Roughly half the world carries *Helicobacter pylori*; a small minority develop gastric cancer. The microbe is necessary but not sufficient — host genetics, bacterial strain genotype, and diet all condition the outcome.

Koch knew the framework leaked. He was aware of asymptomatic carriers of cholera and typhoid, which violate the first postulate outright. These criteria were a standard of evidence, never a law of nature.

Koch's postulates: the molecular age

Falkow recast causation in terms of **genes** rather than organisms — by the 1980s the interesting variation was between strains of the same species (Falkow, 1988).

1. The virulence gene is found in **pathogenic** but not nonpathogenic strains
2. **Deletion or inactivation** of the gene leads to loss of pathogenicity
3. **Reactivation or allelic replacement** of the gene restores pathogenicity

Falkow revisited these criteria fifteen years later, noting how often they bend in practice (Falkow, 2004).

SPEAKER NOTES

The motivating problem: *E. coli* is a harmless commensal in most people and lethal in others. “Is *E. coli* a pathogen?” is unanswerable, because the species is the wrong unit. Pathogenicity lives in genes that some strains carry and others do not.

The structure is knockout and complementation — the standard of proof from molecular genetics, imported into infectious disease. Find the gene, remove it, lose virulence; put it back, recover virulence.

Falkow’s own retrospective fifteen years later is a useful counterweight: he observed that the criteria bend constantly in practice, and that applying them rigidly would have blocked genuine discoveries. Some virulence genes are essential for viability and cannot be deleted; some phenotypes require animal models that do not exist. Criteria of this kind are frameworks for argument, not checklists to be satisfied mechanically.

Koch's postulates: the sequencing age

Fredricks and Relman rewrote the postulates again once sequence, not culture, became the primary evidence — most pathogens of interest had never been cultured at all (Fredericks & Relman, 1996).

- Focus shifts to **nucleic acid sequences** rather than culture or whole genes
- **Individuals or communities**, as identified by sequencing, differ in abundance, organization, and/or function in diseased vs. healthy hosts
- Community virulence **may or may not** depend on specific, well-defined virulence factors
- **Modifying the community** — or removing specific members — alleviates disease

SPEAKER NOTES

The cases driving this in 1996 were organisms nobody could grow: *Tropheryma whipplei* in Whipple's disease, hepatitis C, the hantaviruses, and the herpesvirus behind Kaposi sarcoma. All were identified by sequence before anyone cultured them, so a framework requiring pure culture could not accommodate them.

The last criterion is the one that survives as the field's working standard. If modifying a community changes the disease, you have something causal. That is the logic behind fecal microbiota transplant for recurrent *Clostridioides difficile* infection, which remains the field's most convincing intervention.

There is a cost to this move that becomes important later. Once sequence alone is the evidence, a new failure mode appears: sequence can originate from a contaminated reagent, a mislabeled entry in a reference database, or a batch effect. Koch could see his bacillus under a microscope and grow it on a plate. Sequence-based inference has no equivalent independent check unless one is built in deliberately.

Why study the microbiome?

Humans are full of microorganisms — skin, gut, oral cavity, nasal cavity, eyes — and they affect health, drug metabolism, and treatment response.

3.8×10^{13}

bacterial cells in a reference 70 kg adult

$\approx 1.3 : 1$

bacteria-to-human-cell ratio

$\approx 150 \times$

more genes than the human complement — 3.3 M microbial genes

! Correction to a familiar number

The often-repeated “**10× more microbial cells than human cells**” traces to a 1972 back-of-envelope estimate. Sender, Fuchs and Milo’s re-derivation puts the ratio near **1:1** — close enough that one defecation event shifts it. Most of the human-cell count is red blood cells, which is where the old estimate went wrong. The **gene** claim survived; the **cell** claim did not.

SPEAKER NOTES

The 10:1 ratio appeared in essentially every microbiome review for forty years. It traces to a 1972 estimate by Luckey that was explicitly an order-of-magnitude calculation, repeated until citation made it fact.

Sender, Fuchs and Milo did the arithmetic properly in 2016: about 3.8×10^{13} bacteria, the overwhelming majority in the colon, against about 3.0×10^{13} human cells. Roughly 90% of human cells are red blood cells — small, numerous, and easy to underestimate when reasoning from tissue volume, which is where the old figure went wrong. The corrected ratio is near 1:1 and varies enough that a single bowel movement measurably shifts it.

The gene claim held up. Qin and colleagues catalogued 3.3 million non-redundant microbial genes from 124 individuals, against roughly 20,000 human genes. So the functional-potential argument for studying the microbiome survives intact; it was only the headline cell-count statistic that failed.

The transferable point: a number everyone repeats and nobody has re-derived is a liability, not a fact.

Measuring the microbiome

Every step between the swab and the count matrix adds bias. Knowing which step added which bias is most of the skill.

SPEAKER NOTES

This section moves from motivation to measurement.

The governing idea: the numbers in a microbiome table are not a measurement of the community. They are a measurement of the community after it has passed through sampling, extraction, amplification, sequencing, and a reference database — each of which distorts it in a characteristic way. A large share of the irreproducibility in this literature comes from comparing studies that made different choices at those steps.

Workflow



! The thing to remember

Every step can add bias or noise — swab site, extraction chemistry, primer choice, PCR cycle number, sequencing platform, and reference database each leave a fingerprint on the final abundance table.

Each step has a characteristic failure.

Sampling. Body site, time of day, recent food, recent antibiotics. Stool is not “the gut” — it samples the lumen’s output, not the mucosa, and the two communities differ substantially.

Extraction. The most underestimated step. Gram-positive cell walls are tough; insufficient bead-beating systematically under-recovers Firmicutes. A reported Bacteroidetes-to-Firmicutes ratio can be largely an artifact of lysis protocol.

Amplification. Primer choice determines which organisms are visible at all. PCR cycle number drives chimera formation and amplification bias.

Sequencing. Platform error profiles differ, which affects how reads are resolved into features.

Bioinformatics. Reference database and version, covered in more detail later.

The practical consequence: when two labs report different abundances for the same organism, protocol differences are a more likely explanation than biology.

Two sequencing approaches

SHOTGUN

- Sequence **all** DNA in the sample
- More information — taxonomy *and* function
- Higher complexity
- Higher cost

AMPLICON

- Sequence **one** marker gene
- No functional information
- Less complex to analyze
- Cheaper

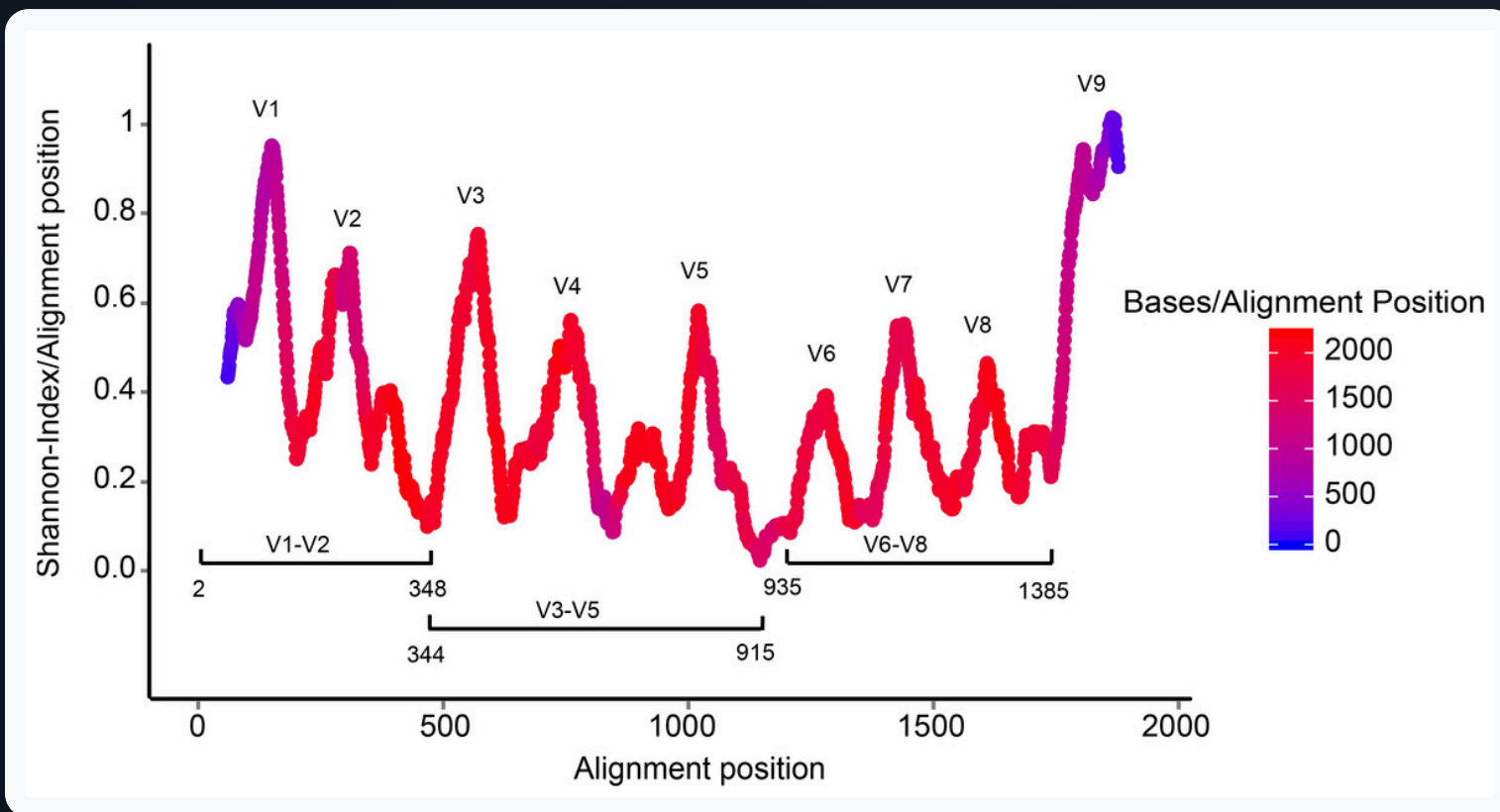
SPEAKER NOTES

The intuition: amplicon sequencing asks “who is here?” by reading one barcode gene from every organism. Shotgun sequencing asks “what DNA is here?” and reconstructs both membership and functional capability from the fragments.

Amplicon gives you a census. Shotgun gives you a census plus a gene inventory, at roughly ten to twenty times the cost per sample.

That cost difference, more than any scientific argument, explains why most of the existing literature is 16S amplicon data: it is cheap enough to run hundreds of samples on a typical grant, and shotgun generally is not. Detailed trade-offs follow once both workflows have been shown.

The 16S rRNA gene



Nine **variable regions** (V1–V9) sit between highly conserved stretches. The conserved parts give you universal primer binding sites; the variable parts give you taxonomic signal.

Which region you amplify determines what you can resolve — and what you systematically miss.

Shannon index per alignment position (50-bp moving average); per-base coverage shown by the heatmap; numbering follows the *E. coli* 16S rRNA gene. Figure 2 of the RIM-DB paper — computed over a **methanogen** alignment, so read the shape of the profile rather than exact peak heights (Seedorf et al., 2014).

SPEAKER NOTES

Why this gene: 16S rRNA is present in every bacterium and archaeon, it is functionally constrained so it evolves slowly, and it alternates conserved and variable stretches along its length.

That alternation is the entire trick. Conserved blocks provide primer binding sites that work across the whole domain; the variable blocks between them carry enough sequence difference to distinguish taxa. In the figure, peaks are variable positions and troughs are conserved; the brackets mark commonly amplified windows.

The practical consequence: short-read platforms cover only a few hundred bases, so you must choose a window. V4 is the most common choice, V3-V4 also popular. Different windows give different answers — some genera are indistinguishable across V4 but separable across V1-V2, and the reverse also happens. “We did 16S” is therefore not a method description; “we amplified V4 with the 515F/806R primer pair” is.

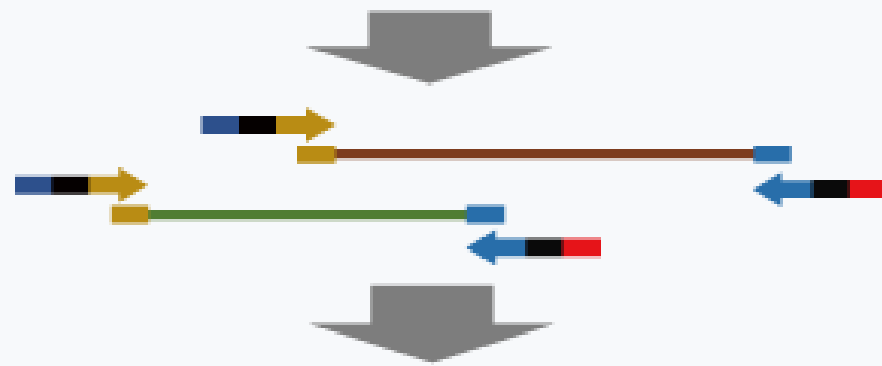
Note on the figure itself: it comes from RIM-DB, a methanogen database, so the exact peak heights reflect archaeal sequences. The structure of the argument generalizes; the specific values are not universal.

16S amplicon sequencing

16S rRNA genes



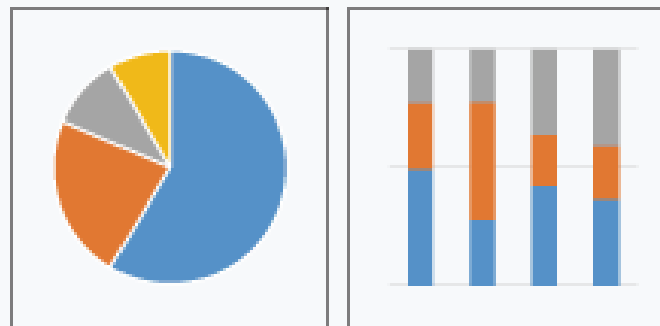
① Amplification with a gene specific primer with an adapter (tag)



② Added barcode and sequence with Index PCR



③ Sequencing with Miseq



④ Data analysis with dedicated analysis software

SPEAKER NOTES

Four steps. First, amplify the chosen variable region with primers carrying adapter tails. Second, a short index PCR adds a sample-specific barcode, which is what allows hundreds of samples to be pooled in one sequencing run. Third, sequence the pool. Fourth, demultiplex by barcode and process.

Two artifacts to know. Barcode hopping — index misassignment between samples, particularly on patterned flow cells — introduces low-level cross-contamination into every sample on a run. This is tolerable for high-biomass samples and serious for low-biomass ones.

Second, the processing step in panel 4 has changed substantially. The field formerly clustered reads into operational taxonomic units at 97% similarity; current practice resolves exact amplicon sequence variants using DADA2 or Deblur. The difference matters when reading older literature: an OTU is defined relative to the dataset that produced it, so OTUs are not comparable across studies, whereas ASVs are exact sequences and therefore are.

Shotgun metagenomics

Genomes of the
microorganisms included in the
sample



DNA
Extraction



Fragmentation



Sequencing



Unknown genome

Assembly and alignment with genomes referenced in
databases

Assembly and new
annotation

SPEAKER NOTES

No marker gene is amplified. Total DNA is extracted, fragmented, and sequenced, then reads are handled along two paths.

Reads matching entries in a reference database are assigned taxonomically and functionally. Reads that match nothing can still be assembled *de novo* into contigs and binned into metagenome-assembled genomes — MAGs — which is how the field discovers organisms that have never been cultured. That second path is the main scientific advantage of shotgun over amplicon.

The dominant practical problem is host DNA. In a human sample, the great majority of DNA is human. For stool this is manageable. For tumor tissue, blood, or bronchoalveolar lavage, 99% or more of sequencing effort can go to host DNA to recover a small number of microbial reads. This is the low-biomass regime, and it is where the contamination problems discussed later become severe.

Choosing between them

SHOTGUN

Pros

- Not biased by amplicon primer set
- Not limited by conservation of one gene
- Provides functional information

Cons

- Environmental and host contamination
- Expensive (\$1000+/sample)
- Complex analysis; needs HPC, high memory

AMPLICON

Pros

- Well established, large body of prior data
- Inexpensive (\$50–\$100/sample)

Cons

- V-region choice biases results
- Built on a very well-conserved gene — hard to resolve species and strains

SPEAKER NOTES

A workable decision rule.

If the question is whether community composition differs between groups, and statistical power matters, 16S is appropriate and the cost difference is the decisive argument — the same budget buys ten to twenty times as many samples.

If the question involves function, specific pathways, or strain identity, 16S cannot answer it. Strain resolution matters more often than expected: pathogenic and commensal strains of the same species are frequently identical across the amplified 16S window, so a 16S study cannot distinguish them.

Framed generally, this is a trade between statistical power and biological resolution. Neither option is more modern or more correct; the right answer depends on the question.

One caveat on the cost figures: they are order-of-magnitude and have fallen steadily. Analysis cost — compute and expert time — has fallen much more slowly than sequencing cost, and is now frequently the binding constraint on shotgun studies.

16S reference databases

- Berkeley lab
- August 2013
- 202,421 entries



- Max Planck Institute
- July 2015
- 172,418 entries



⚠ These snapshots are stale

Greengenes 13_8 and SILVA 128 are the versions many published pipelines still pin. Taxonomy has moved substantially since — results are **not comparable across database versions**, and “which database” is a real methods choice, not a detail (DeSantis et al., 2006; Quast et al., 2012).

SPEAKER NOTES

The dates are the point. Greengenes 13_8 dates from 2013 and was effectively unmaintained for roughly a decade, yet remained the default in widely used tutorials and therefore in a large body of published analysis.

Bacterial taxonomy has been substantially revised in that interval, much of it under the Genome Taxonomy Database, which reassigns groups on the basis of whole-genome phylogeny rather than historical convention. Firmicutes became Bacillota; Bacteroidetes became Bacteroidota. The renaming is superficial; the regrouping underneath it is not. Two studies reporting “Firmicutes abundance” against different database versions are not measuring the same quantity.

Practical rules that follow. Report the database *and its version* in methods. Do not compare abundances across studies that used different references without reprocessing from raw reads. This is among the most common silent errors in microbiome meta-analysis, and it is the direct motivation for the uniformly processed compendium resources at the end of the talk.

Greengenes2 unifies the two worlds

nature biotechnology



Brief Communication

<https://doi.org/10.1038/s41587-023-01845-1>

Greengenes2 unifies microbial data in a single reference tree

Received: 16 December 2022

Accepted: 25 May 2023

Published online: 27 July 2023

Check for updates

Daniel McDonald¹, Yueyu Jiang², Metin Balaban³, Kalen Cantrell⁴, Qiyun Zhu^{5,6}, Antonio Gonzalez¹, James T. Morton⁷, Giorgia Nicolaou⁸, Donovan H. Parks⁹, Søren M. Karst¹⁰, Mads Albertsen¹¹, Philip Hugenholtz⁹, Todd DeSantis¹², Se Jin Song¹³, Andrew Bartko¹³, Aki S. Havulinna^{14,15}, Pekka Jousilahti¹⁴, Susan Cheng^{16,17}, Michael Inouye^{18,19}, Teemu Niiranen^{14,20}, Mohit Jain²¹, Veikko Salomaa¹⁴, Leo Lahti²², Siavash Mirarab² & Rob Knight^{1,4,13,23} ✉

Studies using 16S rRNA and shotgun metagenomics typically yield different results, usually attributed to PCR amplification biases. We introduce Greengenes2, a reference tree that unifies genomic and 16S rRNA databases in a consistent, integrated resource. By inserting sequences into a whole-genome phylogeny, we show that 16S rRNA and shotgun metagenomic data generated from the same samples agree in principal coordinates space, taxonomy and phenotype effect size when analyzed with the same tree.

16S and shotgun studies of the *same samples* have historically disagreed — usually blamed on PCR amplification bias.

Greengenes2 inserts both data types into a **single whole-genome phylogeny**. Analyzed against the same tree, 16S and shotgun data agree in principal coordinates space, taxonomy, and phenotype effect size (McDonald et al., 2023).

SPEAKER NOTES

The puzzle: running 16S and shotgun sequencing on the same physical samples has historically produced different answers. The field's standard explanation was PCR amplification bias in the amplicon arm.

McDonald and colleagues showed that a large part of the discordance was instead an artifact of comparing results obtained through two different reference systems — a 16S taxonomy on one side and a genome-based taxonomy on the other. Greengenes2 places both data types in one whole-genome phylogeny, inserting 16S fragments into that tree rather than matching them against a separate 16S reference. Analyzed this way, the two technologies agree on ordination, taxonomy, and phenotype effect size.

The generalizable instinct: when two measurement technologies disagree, check whether they are being compared through a common reference frame before concluding that the technologies themselves differ.

Profiling shotgun data: MetaPhlAn 4

Metagenomic assembly finds novel organisms but recovers only the abundant ones. MetaPhlAn 4 combines **metagenome assemblies with isolate genomes** — built from a curated collection of ~1.01 M prokaryotic reference and metagenome-assembled genomes. The practical consequence: a large fraction of reads that previously went unclassified now map to *uncharacterized* species (Blanco-Míguez et al., 2023).

Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4

Aitor Blanco-Míguez¹, Francesco Beghini¹, Fabio Cumbo¹, Lauren J. McIver^{2,3}, Kelsey N. Thompson^{2,3}, Moreno Zolfo¹, Paolo Manghi¹, Leonard Dubois¹, Kun D. Huang¹, Andrew Maltez Thomas¹, Gianmarco Piccinno¹, Elisa Piperni^{1,4}, Michal Punčochář¹, Mireia Valles-Colomer¹, Adrian Tett^{1,5}, Francesca Giordano⁶, Richard Davies⁶, Jonathan Wolf⁶, Sarah E. Berry⁷, Tim D. Spector⁸, Eric A. Franzosa^{2,3}, Edoardo Pasolli⁹, Francesco Asnicar¹, Curtis Huttenhower^{2,3}, Nicola Segata^{1,4,#}

¹ Department CIBIO, University of Trento, Trento, Italy

² Harvard T.H. Chan School of Public Health, Boston, MA, USA

³ The Broad Institute of MIT and Harvard, Cambridge, MA, USA

⁴ Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy

⁵ Centre for Microbiology and Environmental Systems Science, University of Vienna, Austria

⁶ Zoe Global, London, UK

⁷ Department of Nutritional Sciences, King's College London, London, UK

⁸ Department of Twin Research, King's College London, London, UK

⁹ Department of Agricultural Sciences, University of Naples, Naples, Italy

Corresponding author: nicola.segata@unitn.it

Keywords: Metagenomics, microbiome taxonomic profiling, metagenomic assembly

Abstract

Metagenomic assembly enables novel organism discovery from microbial communities, but from most metagenomes it can only capture few abundant organisms. Here, we present a method - MetaPhlAn 4 - to integrate information from both metagenome assemblies and microbial isolate genomes for improved and more comprehensive metagenomic taxonomic profiling. From a curated collection of 1.01M prokaryotic reference and

SPEAKER NOTES

MetaPhlAn works by clade-specific marker genes — sequences unique to a given taxon. It quantifies a community by counting reads that hit those markers, which is far faster than assembling every genome and yields relative abundance directly. Note this is a different sense of “marker gene” than 16S: these are many taxon-specific sequences, not one universal gene.

What changed in version 4: earlier versions could only profile organisms with a cultured reference genome, leaving uncultured organisms invisible. Version 4 incorporates roughly a million metagenome-assembled genomes alongside isolate genomes, so a large share of previously unassignable reads now map to defined species-level genome bins.

The trade is real rather than a free improvement. Coverage increases, but many of the newly quantified organisms have never been cultured or characterized, so biological interpretability of the resulting abundances is weaker. MetaPhlAn is also the profiler underlying curatedMetagenomicData, discussed later.

What comes out: the count matrix

	CL3	CC1	SV1	M31Fcsw	M11Fcsw	M31Plmr	M11Plmr	F21Plmr	M31Tong	M11Tong	LMEpi24M	SLEpi20M	AQC1cm
549322	0	0	0	0	0	0	0	0	0	0	0	1	27
522457	0	0	0	0	0	0	0	0	0	0	0	0	0
951	0	0	0	0	0	0	1	0	0	0	0	0	0
244423	0	0	0	0	0	0	0	0	0	0	0	0	0
586076	0	0	0	0	0	0	0	0	0	0	0	0	0
246140	0	0	0	0	0	0	0	0	0	0	0	0	0
143239	7	1	0	0	0	0	3	0	0	0	0	0	0
244960	0	0	0	0	0	0	0	0	0	0	0	0	0
255340	153	194	0	0	0	0	0	0	0	0	0	0	1
144887	3	5	0	0	0	0	0	0	0	0	0	0	0
141782	9	35	0	0	0	0	0	0	0	0	0	0	0
215972	0	3	0	0	0	0	0	0	0	0	0	0	0
31759	0	1	0	0	0	0	0	0	0	0	0	0	0
30678	0	3	0	0	0	0	0	0	0	0	0	1	0
138353	0	0	19	0	0	0	0	0	0	0	0	0	0
406058	1	0	0	0	0	0	0	0	0	0	0	0	0
30405	808	11	3	0	0	0	0	0	0	0	0	0	0
1126	0	0	0	0	0	0	0	0	0	0	0	0	0
46043	0	0	0	0	0	0	0	0	4	0	0	0	0
512519	0	0	0	0	0	0	0	0	0	0	0	0	0
554594	0	0	0	0	0	0	0	0	0	0	0	0	0

i Read the shape, not the numbers

Taxa in rows, samples in columns. **Sparse** — mostly zeros. **Compositional** — column sums are an artifact of sequencing depth, not biology. **Over-dispersed** — non-zero counts span orders of magnitude.

SPEAKER NOTES

This object is the hinge of the talk. Everything upstream exists to produce it; everything downstream is a transformation of it. Three structural properties matter, and each breaks a standard statistical method.

Sparse. Most entries are zero, and a zero is ambiguous: the organism may be genuinely absent, or present below the detection limit set by sequencing depth. Those are different claims and the matrix cannot distinguish them.

Compositional. Each column sums to whatever depth that sample was sequenced to, which is a technical choice, not biology. If one taxon genuinely increases, every other taxon's relative abundance must fall — so apparent decreases can be entirely artifactual. This is the most common inferential error in the field.

Over-dispersed. Non-zero counts span orders of magnitude, far exceeding what a Poisson distribution predicts, which is why count models applied here need an explicit dispersion parameter.

This slide reappears after the cancer material; the repetition is deliberate.

The microbiome in cancer

Some microbes cause cancer outright. Far more shape how it grows, how it is detected, and how it responds to treatment.

SPEAKER NOTES

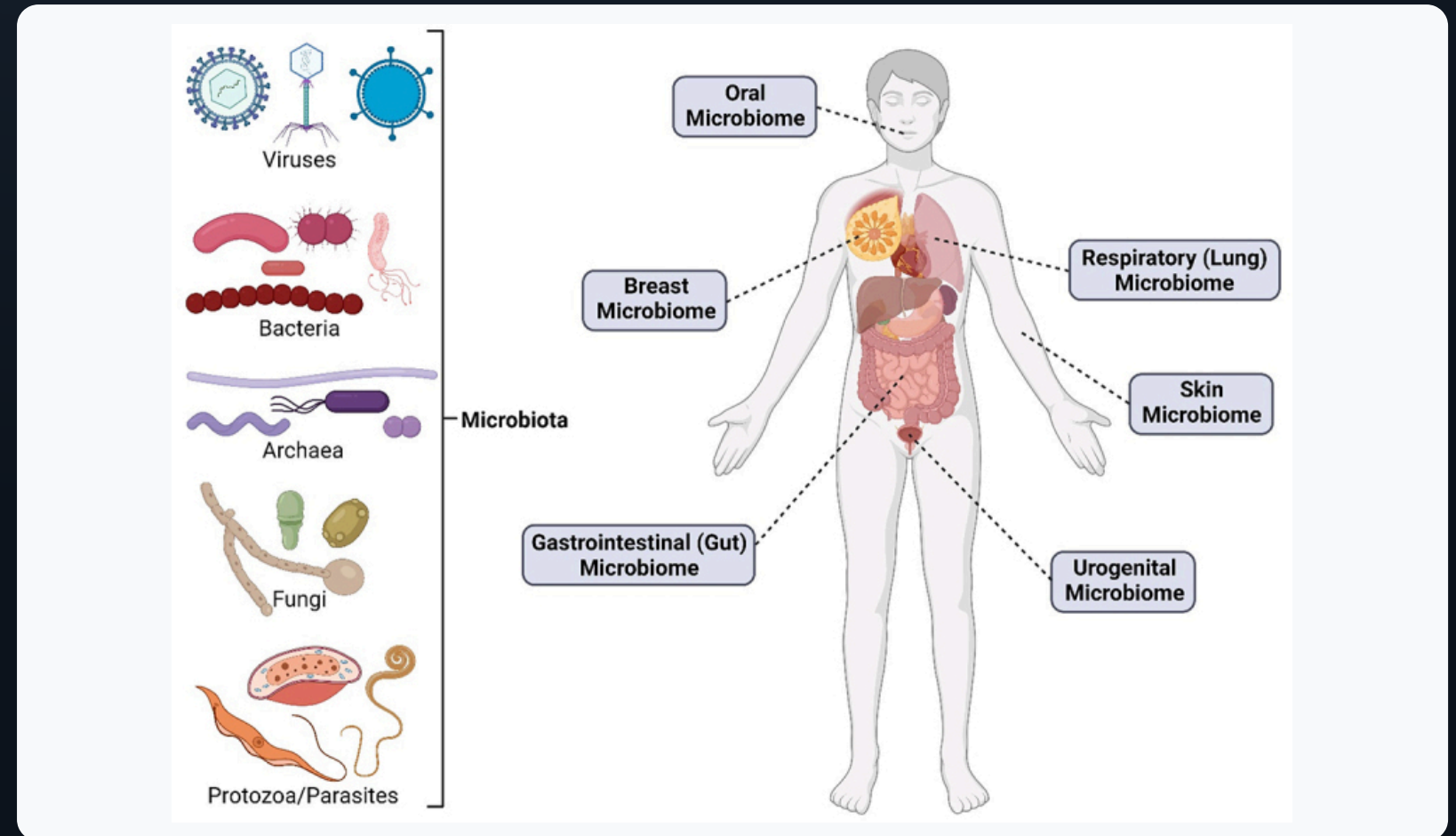
The section runs in four parts: the small set of microbes that genuinely cause cancer; the much larger set that modulate it; diagnostics, including an extended cautionary tale; and therapy.

The organizing distinction is between causing and modulating. Conflating the two is the source of a great deal of overclaiming in this literature.

The human microbiota

The microbiota spans **viruses, bacteria, archaea, fungi, and protozoa/parasites**, organized into distinct site-specific communities — oral, respiratory, breast, gastrointestinal, skin, and urogenital.

Site matters more than almost anything else: two gut samples from different people resemble each other far more than a gut and a skin sample from the same person (Kandalai et al., 2023).



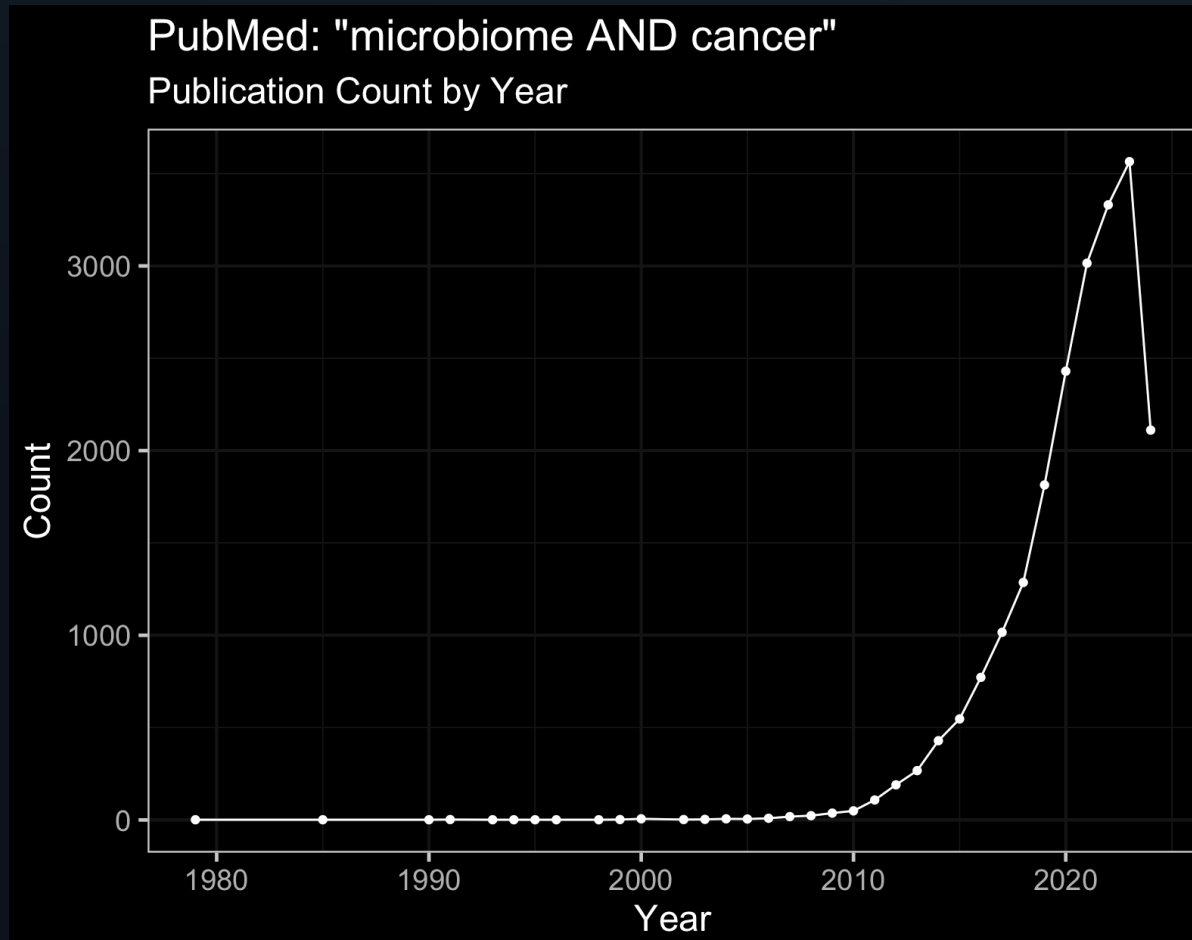
SPEAKER NOTES

In casual usage “microbiome” means gut bacteria. The left column of the figure is the honest scope: viruses, archaea, fungi, and eukaryotic parasites are all present. They are systematically under-studied largely for methodological reasons — 16S primers do not amplify them — so the virome and mycobiome are far less well mapped than the bacteriome. Absence of data is not evidence of unimportance.

The dominant source of variation is body site. Skin and gut communities within one person differ more from each other than either differs from the same site in a stranger. The consequence for study design is that site must be controlled before anything else; comparing tumor tissue against a stool sample measures anatomy, not disease.

Several sites shown here — breast, lung — were considered sterile until relatively recently. Those claims remain contested, and precisely because the biomass is so low, they are the sites where contamination artifacts are most dangerous. That thread continues into the diagnostics section.

A field that grew very fast



Essentially nothing before 2010;
roughly **3,600 papers in 2023 alone.**

That growth is the opportunity and the problem. A field this young, moving this fast, accumulates findings faster than it validates them — as the next few slides show.

SPEAKER NOTES

Flat until about 2010, then near-exponential. The inflection tracks the availability of cheap high-throughput sequencing and the Human Microbiome Project — a change in method and funding, not a biological discovery.

The implication is worth stating plainly rather than cynically: in a field operating at scale for barely fifteen years, a large fraction of published findings have never been independently replicated. Replication takes years, and there has not been time.

The final point on the curve drops because indexing for that year was incomplete when the query was run. It is an artifact, not a decline — a small reminder to read one's own plots skeptically before interpreting them.

Single microbes that cause cancer

These are the cases where something close to Koch’s postulates actually holds — a single agent, an established mechanism, and in several cases a vaccine or eradication therapy that measurably reduces incidence.

Cancer type	Causative microbe
Gastric cancer	<i>Helicobacter pylori</i>
Liver cancer	Hepatitis B virus, Hepatitis C virus
Biliary tree cancer	<i>Clonorchis sinensis</i> , <i>Opisthorchis viverrini</i>
Cervical cancer	Human papillomavirus (HPV)
Head and neck cancer	HPV
Urinary bladder cancer	<i>Schistosoma haematobium</i>
Lymphoma	Epstein–Barr virus
Merkel cell carcinoma	Merkel cell polyomavirus
Kaposi sarcoma	Kaposi sarcoma–associated herpesvirus

H. pylori and gastric cancer: Uemura et al. (2001). *Fusobacterium nucleatum* in colorectal carcinoma sits one rung below this table — robustly replicated, mechanism still open (Castellarin et al., 2011; Kostic et al., 2011).

SPEAKER NOTES

Roughly 15–20% of the global cancer burden is attributable to infection, and this table accounts for most of it. Note the composition: mostly viruses, one bacterium, a few parasites.

The strongest evidence here is interventional rather than observational. HPV vaccination has produced measurable reductions in cervical precancer, and *H. pylori* eradication reduces gastric cancer incidence. That is Koch's fifth postulate satisfied in humans — remove the microbe, reduce the disease — and it is a far stronger form of evidence than any association study.

The canonical *H. pylori* cohort is Uemura and colleagues: 1,526 patients followed a mean of 7.8 years, with gastric cancer developing in about 2.9% of infected patients and in none of the uninfected. The appropriate caveats are a Japanese cohort, a high-incidence population, and largely CagA-positive strains, so it generalizes imperfectly — but the contrast is stark. It also illustrates necessity without sufficiency, since the large majority of infected patients did not develop cancer.

Fusobacterium nucleatum is the informative boundary case, which is why it appears in the aside rather than the table. Two back-to-back 2012 *Genome Research* papers found it enriched in colorectal carcinoma; it has replicated repeatedly and plausible mechanisms exist, including FadA-mediated β -catenin signalling. It has not cleared the interventional bar that the table sets.

Indirect effects of the microbiome

Adapted from Kandalai et al. ([Kandalai et al., 2023](#)). Intratumoral bacteria were confirmed by microscopy and culture, not sequencing alone ([Nejman et al., 2020](#)).

This is the larger and less tidy category: microbes that shape cancer without causing it. One concrete example per route.

Metabolites. Butyrate, produced by fibre-fermenting gut bacteria, is a histone deacetylase inhibitor and is broadly protective in colonic epithelium. The same class of mechanism runs the other way too — colibactin, produced by certain *E. coli* strains, alkylates DNA and leaves a characteristic mutational signature in colorectal tumors.

Direct interactions. *F. nucleatum*'s FadA adhesin binds E-cadherin and activates β -catenin signalling, driving proliferation.

Inflammation. Chronic inflammatory signalling is the classic route, and is also the principal mechanism by which gut community composition modulates response to checkpoint immunotherapy.

Barrier disruption. Compromised epithelial or vascular barriers facilitate both bacterial translocation and metastatic spread.

The Nejman citation deserves emphasis. That group demonstrated intracellular bacteria in over a thousand tumors across seven cancer types using FISH, immunohistochemistry, electron microscopy, and culture — orthogonal methods, not sequencing alone. This is why those findings survived the reanalyses that invalidated sequencing-only work, and it is the direct contrast to the next two slides.

Microbiome-based diagnostics

Method	Cancer type	Marker(s)	AUROC
Salivary microbiome	Pancreatic	<i>N. elongata, S. mitis</i>	0.90
Salivary microbiome	Lung squamous cell	<i>Capnocytophaga, Veillonella</i>	0.86
Fecal microbiome	Colorectal	<i>C. symbiosum</i>	0.73
Fecal microbiome	Colorectal	<i>F. nucleatum</i>	0.86
Fecal microbiome	Lung	Various bacteria	0.76
Plasma cell-free DNA †	Various types	Various bacteria	0.90
Plasma cell-free DNA †	Various types	Various fungi	0.80
Plasma cell-free DNA †	Various types	Bacteria + fungi	0.92

! † Do not quote the last three rows

The source of those numbers has been **retracted**. Next slide.

Table adapted from Kandalai et al. (Kandalai et al., 2023).

SPEAKER NOTES

This is a table that would have been presented uncritically three years ago. It divides into two very different halves.

The top five rows are single-site, relatively high-biomass samples — saliva and stool — testing for one cancer type each. The bottom three are plasma cell-free DNA, claiming pan-cancer discrimination. The harder problem reports the higher AUROCs, which should itself prompt suspicion.

The top rows are not beyond criticism either. The pancreatic salivary result rests on a small validation cohort of roughly 28 cases and 28 controls. Several of these figures are discovery-cohort performance that falls substantially on validation: the fecal lung result was about 0.98 in discovery and 0.76 in independent validation. The validation number is the one worth quoting, and the gap between the two is itself a useful lesson about optimism in model development.

AUROC as a metric also deserves scrutiny for screening applications. A test with excellent AUROC can still have poor positive predictive value when the disease prevalence is low, which is the situation in any general-population screen.

A cautionary tale

In 2020, Poore *et al.* reported that microbial DNA read out of blood and tissue could discriminate dozens of cancer types with near-perfect accuracy (Poore *et al.*, 2020). The result launched a company and an FDA breakthrough-device designation.

! It was wrong

Reanalysis found the signal came from **contaminated reference genomes**, human reads **misclassified as microbial**, and a **batch-correction step that manufactured cancer-type structure**. The classifier was leaning on, among other things, a **shrimp virus** as a human cancer biomarker (Gihawi *et al.*, 2023). *Nature* retracted the paper on **26 June 2024**, with all authors agreeing (Poore *et al.*, 2024).

Two independent 2025 reanalyses — one reprocessing all of TCGA, one using a separate 8,908-patient cohort with contamination controls — found that when contamination is properly handled, **only colorectal cancer** carries a robust distinct microbial signature (Ge *et al.*, 2025; Gihawi *et al.*, 2025).

SPEAKER NOTES

This is a methods lesson in the form of a story, and the most transferable material in the talk.

In 2020 Poore and colleagues published in *Nature* that microbial DNA in blood and tissue could discriminate 33 cancer types, including at stage I and including from plasma. The implication was a universal blood-based cancer screen. It was cited hundreds of times, a company was founded on it, and it attracted an FDA breakthrough-device designation.

Gihawi, Salzberg and colleagues reanalyzed it and identified three distinct failures, each of which generalizes beyond this study.

First, reference database contamination: human sequence present within bacterial genome entries meant human reads were classified as microbial. Second, permissive classification generating large numbers of false-positive microbial assignments. Third — the subtle and most instructive one — a batch-correction step applied across batches that were confounded with cancer type, which wrote cancer-type structure into the features themselves. The model then recovered the structure that the correction had inserted.

The memorable detail: among the features the classifier treated as informative was a *Hepandensovirus*, a shrimp virus. Nothing biological explains a shrimp virus discriminating human cancers; it is a database artifact correlated with batch.

The authors published a substantial rebuttal in *Oncogene* in early 2024 defending the signatures. *Nature* retracted the paper in June 2024 after independent expert review, with all authors agreeing. In 2025 a downstream paper built on the same input data was itself retracted, because its results could not be reproduced once the source data was withdrawn — a clean example of citation cascade, where an error propagates into work that never touched the original analysis.

The constructive endpoint is the 2025 replication work: two independent efforts, one reprocessing all of TCGA and one using an entirely separate 8,908-patient Genomics England cohort with proper contamination controls. No pan-cancer microbial signature survived. Colorectal cancer was the lone robust exception, along with a viral association in oral cancer.

What to take from that

i The tumor microbiome is real

H. pylori in gastric cancer, *F. nucleatum* in colorectal cancer, and Nejman's intracellular bacteria — confirmed by FISH, immunohistochemistry, electron microscopy, and culture, not sequencing alone — all stand (Nejman et al., 2020).

! What collapsed was pan-cancer diagnostics from low-biomass sequencing

At these DNA concentrations, contamination from reagents, kits, and reference databases can **exceed the true signal**. Batch correction applied across confounded batches then converts that contamination into apparent biology. This failure mode generalizes well beyond microbiome work.

The practical lesson: when a machine-learning classifier reports an AUROC of 0.95 on low-biomass data, the first question is not “which taxa?” but “**what else differs between my batches?**”

SPEAKER NOTES

The correct conclusion is narrower than “the tumor microbiome was a mirage.”

What stands is the work with orthogonal validation — findings where bacteria can be visualized, stained, and in some cases cultured, with sequencing as corroboration rather than sole evidence. What collapsed is a specific and much stronger claim: that a pan-cancer diagnostic signature can be read out of low-biomass sequencing data.

The generalizable mechanism has two stages. At very low DNA concentrations, the true biological signal can be smaller than contamination introduced by extraction kits, laboratory reagents, and errors in the reference database itself — the “kitome” is the standard term for reagent-derived contamination. Batch correction then, if applied across batches that are confounded with the outcome variable, converts that contamination into structure that looks biological. Neither step involves misconduct. Both are things careful researchers do routinely.

The controls that guard against it are specific and worth memorizing: negative extraction controls carried through the complete protocol; positive mock communities of known composition; randomization of samples across processing batches so batch is not confounded with the outcome; and validation in a cohort processed independently.

The closing question generalizes past microbiome work to any high-dimensional biological classifier. Given surprisingly strong performance on noisy, low-signal data, ask what differs technically between the batches before asking which features are responsible.

Microbiome and cancer therapy

Therapy	Model	Microbe (site)	Finding
Radiotherapy	Melanoma, lung, cervical	Gram-positives (gut)	Depleting them with vancomycin <i>improves</i> RT response
Radiotherapy	Breast	Fungi vs. bacteria (gut)	Opposite directions: depleting fungi helps, depleting bacteria hurts
Cyclophosphamide	Melanoma, sarcoma	<i>L. johnsonii</i> , <i>E. hirae</i> (gut)	Translocation drives the Th17/Th1 response the drug needs
Gemcitabine	Pancreatic	Gammaproteobacteria (tumor)	Bacterial cytidine deaminase inactivates the drug
Dacarbazine	Melanoma lung mets	<i>L. rhamnosus</i> (aerosolized)	Pulmonary probiotic promotes anti-metastatic immunity

Rows in order: Uribe-Herranz et al. (2019); Shiao et al. (2021); Viaud et al. (2013); Geller et al. (2017); Le Noci et al. (2018). Note the direction of the first two — the benefit comes from **removing** microbes, and bacteria and fungi push opposite ways.

SPEAKER NOTES

The directions in this table are counterintuitive, and versions of it circulating in review articles state two of them backwards.

Row 1 (Uribe-Herranz). Oral vancomycin depletes gut Gram-positive bacteria, and this *improves* radiotherapy response, through enhanced tumor-antigen cross-presentation to CD8 T cells. Butyrate from vancomycin-sensitive bacteria abrogates the effect. The claim is not that these bacteria assist radiotherapy.

Row 2 (Shiao). Bacteria and fungi act in opposite directions in the same system. Depleting fungi improves radiotherapy response; depleting bacteria worsens it, partly via compensatory fungal overgrowth. A useful correction to the assumption that more diversity is always better.

Row 3 (Viaud). Cyclophosphamide's antitumor efficacy depends on gut bacteria translocating to secondary lymphoid organs and driving Th17 and memory Th1 responses. Germ-free and Gram-positive-depleted mice respond poorly — the drug requires an intact microbiome to work fully.

Row 4 (Geller) is the only true mechanism on the slide. Intratumoral Gammaproteobacteria express a long isoform of cytidine deaminase that metabolizes gemcitabine to an inactive form inside the tumor; ciprofloxacin reverses resistance in the model. Be precise about the evidence: resistance experiments were performed in a colon cancer model, while the pancreatic evidence is a survey finding bacteria in 76% of human pancreatic adenocarcinomas. Mechanism demonstrated; clinical relevance inferred.

Row 5 (Le Noci). The route is aerosolized into the lung, not oral. It is frequently miscategorized as a gut-microbiome study.

Overall: these are predominantly mouse models. The best-established human translation is in immunotherapy, where gut microbiome composition predicts checkpoint-inhibitor response, and FMT trials in that setting are ongoing.

Analysis and data resources

The count matrix is where the biology stops
and the statistics start.

SPEAKER NOTES

The final section is the most practical. Two ideas carry it: the statistical difficulty of microbiome data is a property of the data structure rather than a failure of effort, and curated public resources now exist so that reprocessing everything from raw reads is not a prerequisite to doing analysis.

Back to the count matrix

The same figure as earlier, revisited now for consequences rather than description.

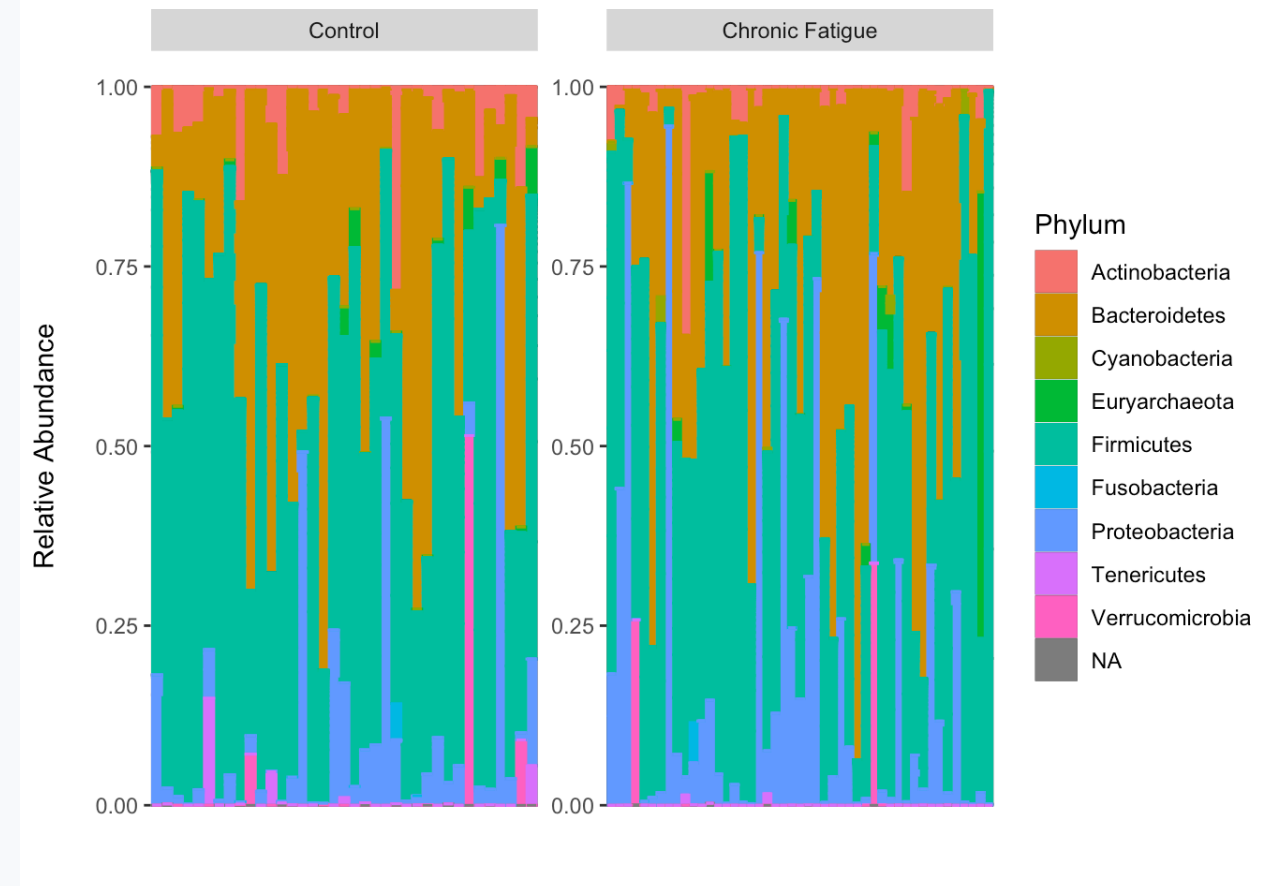
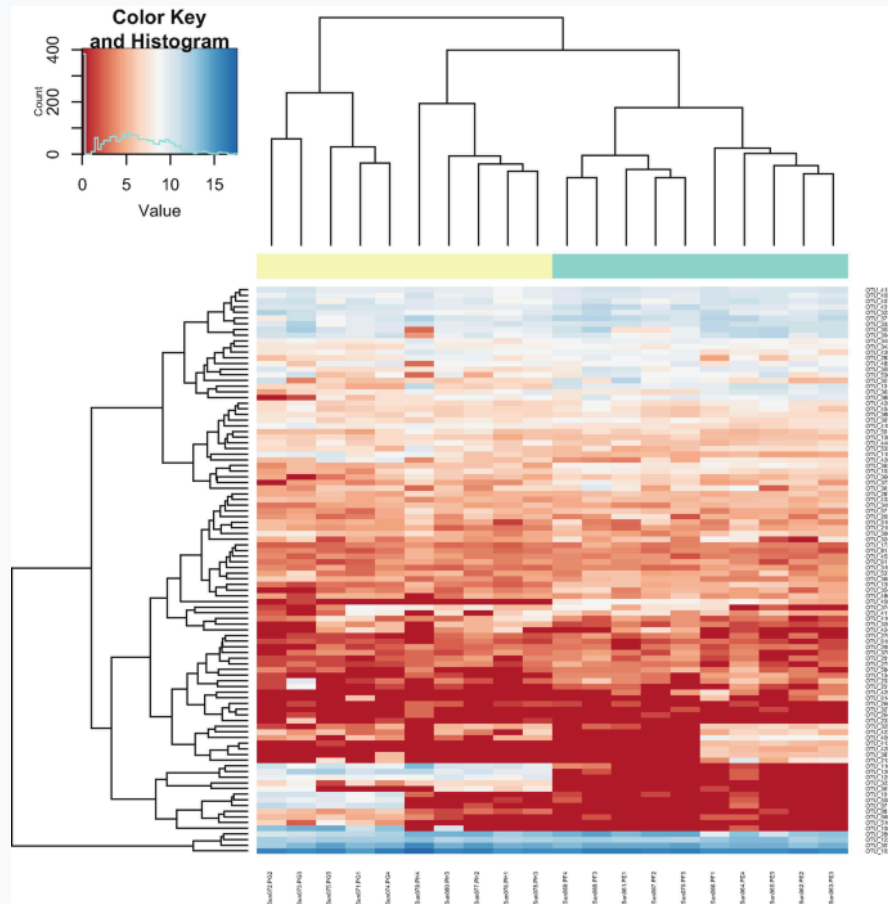
Sparsity makes zero-handling a modelling decision rather than a preprocessing detail. Whether a zero is treated as a structural absence or a sampling artifact changes the result, and different tools make different choices.

Compositionality is the deep problem. Only ratios between components carry information, which is why the field draws on Aitchison's compositional data analysis and transforms such as the centered log ratio. Rarefying — subsampling all samples to equal depth — remains widely used and genuinely contested: it discards data, but it does control the depth confound directly. This is an unsettled methodological question rather than one with a consensus answer.

Over-dispersion explains why methods developed for bulk RNA-seq — negative binomial models such as DESeq2 and edgeR — are commonly applied here, though the fit is imperfect because microbiome data is sparser and more extreme.

For practical tool selection: ANCOM-BC, ALDEx2, and LinDA are the compositionally-aware options in common use. Benchmarking studies show uncomfortably low agreement between methods on the same dataset, so running two and comparing is reasonable practice rather than excessive caution.

Differential abundance analysis



Naive *t*-tests on relative abundance are badly behaved on compositional, sparse, over-dispersed counts — hence a whole literature of dedicated methods.

SPEAKER NOTES

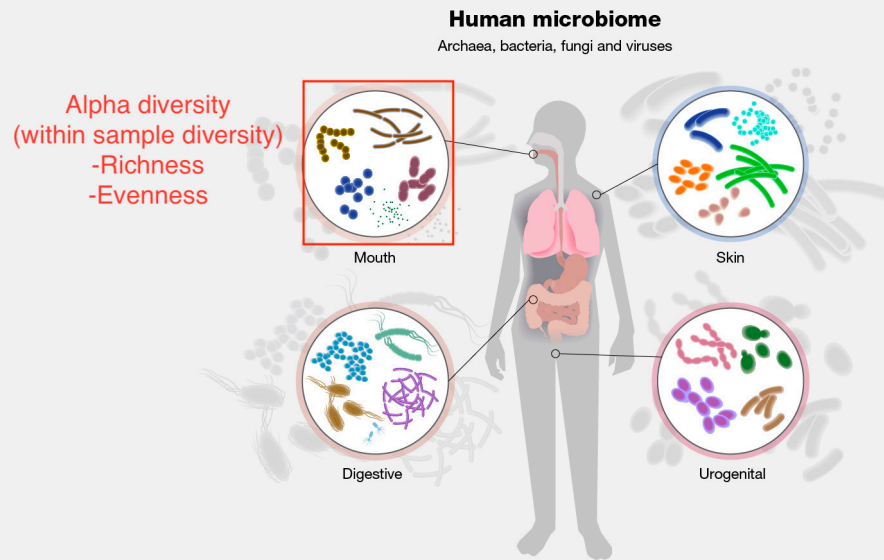
Two standard visualizations, each with a characteristic trap.

The **clustered heatmap** is easy to over-interpret. Hierarchical clustering always produces clusters, whether or not structure exists. The group colour bars above the heatmap are what indicate whether the recovered structure corresponds to the variable of interest — and frequently it corresponds to sequencing batch or extraction date instead.

The **stacked bar chart** is the field's most common plot and the clearest illustration of compositionality. Every column sums to one by construction, so when one phylum increases, all others must appear to decrease regardless of what happened to those organisms in absolute terms. A statement such as “Bacteroidetes decreased” cannot be supported by this plot alone.

Both panels also show something worth noticing: within-group sample-to-sample variability is large relative to any between-group difference. That is typical of microbiome data, and it is why these studies routinely require far larger sample sizes than investigators anticipate.

Alpha diversity



One number per sample. *How diverse is this one community?*

Built from **richness** (how many taxa) and **evenness** (how equally abundance is spread across them).

Questions it answers:

- Does gut diversity fall after antibiotics — and how fast does it recover?
- Does low pre-treatment diversity predict poor checkpoint-inhibitor response? (Gopalakrishnan et al., 2018)

Indices weight the two components differently: Shannon balances them, Simpson favours dominant taxa, Faith's PD counts phylogenetic branch length. Richness rises with sequencing depth, so depth must be controlled before comparing.

SPEAKER NOTES

Richness and evenness are separable, and the distinction matters. A community with 100 taxa in which one organism is 99% of the biomass is rich but not even. Most single indices blend the two, which is why two samples can share an index value for quite different reasons.

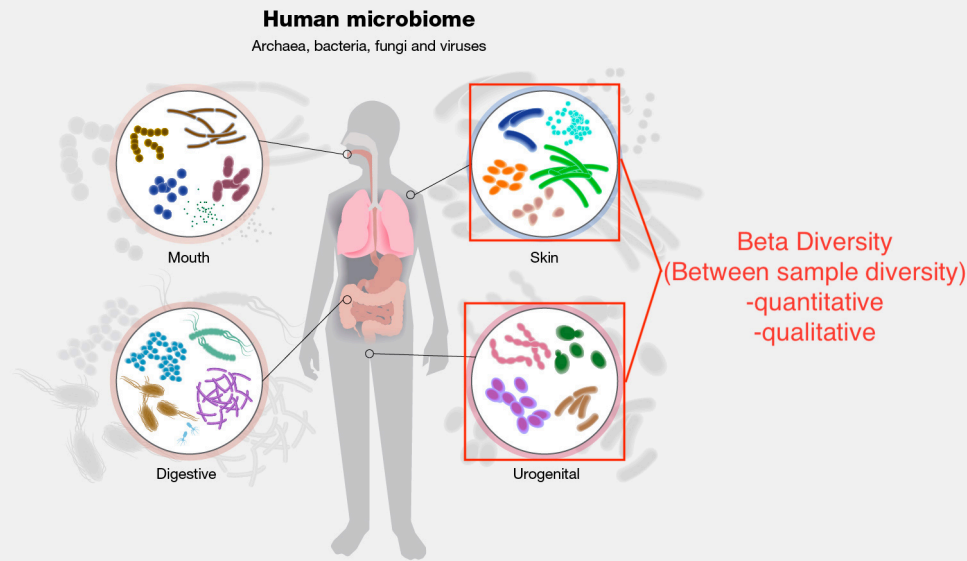
Common indices differ in what they weight. Observed richness counts taxa and ignores abundance entirely. Shannon balances richness and evenness. Simpson weights dominant taxa heavily and is relatively insensitive to rare ones. Faith's phylogenetic diversity sums branch length on the tree, rewarding taxa that are evolutionarily distinct rather than merely numerous. Report which index you used; "diversity decreased" is not interpretable without it.

The critical methodological point is depth dependence. Deeper sequencing detects more rare taxa, so richness rises with sequencing effort. Depth must be controlled — by rarefaction or by an explicit model — before alpha diversity is compared across samples, or the comparison is measuring the sequencing run.

On the immunotherapy example: Gopalakrishnan and colleagues found higher gut diversity in melanoma patients who responded to anti-PD-1, with Routy and colleagues reporting a related finding in epithelial tumors the same week. These are associations in modest cohorts, and the causal claim rests on the accompanying germ-free mouse transfer experiments rather than the human correlation alone.

Finally, resist the assumption that higher diversity means healthier. It is a reasonable generalization for the gut and false elsewhere: a healthy vaginal microbiome is characteristically low in diversity and dominated by lactobacilli.

Beta diversity



One number per *pair* of samples. *How different are these two communities?*

Questions it answers:

- Do responders and non-responders differ in overall composition?
- Do samples cluster by disease — or by sequencing batch?

i The distinction in one line

Alpha summarizes **one sample**; beta compares **two**. Samples can have identical alpha diversity and still separate perfectly on beta — the same *number* of taxa, entirely different taxa.

Qualitative metrics use presence/absence (Jaccard, unweighted UniFrac); **quantitative** ones use abundance (Bray-Curtis, weighted UniFrac). UniFrac scores closely related taxa as more similar by using the phylogeny.

SPEAKER NOTES

The callout is the point to land. Alpha collapses a community to one number; beta is a distance defined only between a pair. An n -sample study yields n alpha values but an $n \times n$ distance matrix, which is why the two are visualized completely differently — box plots by group for alpha, ordination scatter plots for beta.

Note the practical consequence of the last sentence: alpha and beta can disagree, and neither is wrong when they do. Identical richness with entirely different membership gives you flat alpha and strong beta separation. Reporting only alpha would miss the finding entirely.

On metrics: qualitative measures use presence and absence only, so they are more sensitive to rare taxa — and therefore more sensitive to contamination, which connects back to the low-biomass discussion.

UniFrac is the family worth understanding, because it is where the phylogeny earns its place. Rather than treating taxa as independent labels, it measures the fraction of tree branch length unique to each community, so two samples holding different but closely related organisms score as more similar than two holding distant ones. That is also why the tree must stay attached to the data — the motivation for the TreeSummarizedExperiment slide coming up.

Standard workflow: compute the distance matrix, ordinate with principal coordinates analysis, test group differences with PERMANOVA. One caution — PERMANOVA is sensitive to differences in dispersion, so a significant result can reflect one group simply being more variable rather than differing in mean composition. Check with betadisper before claiming a compositional shift.

Accessing microbiome data: MicroBioMap

MicroBioMap 0.99.13 Reference Articles ▾

Search for

Microbiome Compendium

Our dataset includes over 170,000 samples of publicly available 16S rRNA amplicon sequencing data, all processed using the same pipeline and reference database.

The goal of the MicroBioMap package is simply to expose these data to the broad community of Bioconductor and R users with the smallest fuss possible.

See the [vignette](#) for more details.

Installation

For the smoothest installation experience, use the [BiocManager Bioconductor package](#).

```
BiocManager::install('seandavi/MicroBioMap')
```

Usage

Load the compendium using:

```
library(MicroBioMap)
cpd <- getCompendium()
```

License

[Full license](#)

[MIT](#) + file [LICENSE](#)

Citation

[Citing MicroBioMap](#)

Developers

Sean Davis
Author, maintainer 

Richard Abdill
Author 

Ran Blehman
Author 

Samantha Graham
Author

Casey Greene
Author

Over **170,000 publicly available 16S amplicon samples**, all processed through the same pipeline against the same reference database — so cross-study comparison is not confounded by pipeline choice.

```
1 BiocManager::install('seandavi/MicroBioMap')
2 library(MicroBioMap)
3 cpd <- getCompendium()
```

SPEAKER NOTES

Disclosure: this is our package and I am an author.

The motivation follows directly from the reference-database slide. If every study processes its data with a different pipeline against a different reference version, results cannot be pooled, and the field's large accumulated public data becomes effectively unusable for meta-analysis. MicroBioMap reprocesses more than 170,000 public 16S samples through a single uniform pipeline against a single reference, so cross-study comparison becomes tractable.

Three lines of R return the entire compendium as an analysis-ready object.

Typical uses: testing whether a finding from one cohort holds across many; establishing a reference distribution for what is typical at a given body site; and estimating variance components for power calculations from real data rather than assumption.

An honest limitation: uniform processing eliminates pipeline heterogeneity, but does nothing about study heterogeneity. Differences in population, sampling protocol, and DNA extraction remain baked into the underlying data, and still require modelling as batch or study effects.

Accessing microbiome data: `curatedMetagenomicData`

SPEAKER NOTES

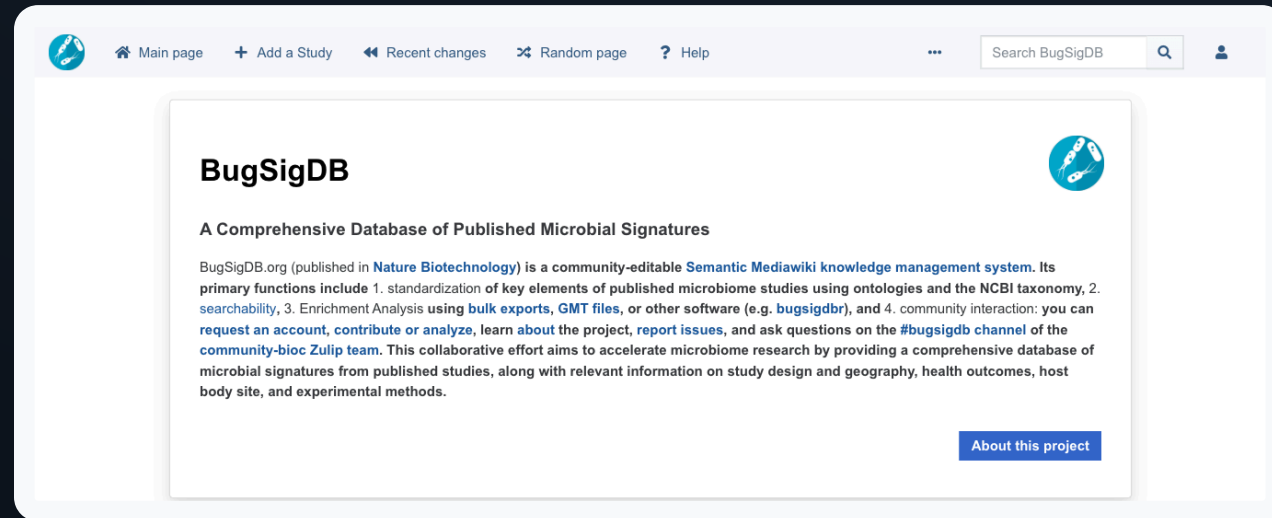
The shotgun counterpart, from the Waldron lab — the same philosophy applied to a different data type.

The operative word is *curated*. Sample metadata is manually standardized, and that work is consistently underestimated. Public metadata is largely unstructured free text: age recorded in years or months, a control group labelled variously “control,” “healthy,” or “HC,” inconsistent or absent disease coding. Harmonizing this is unglamorous and is what makes cross-study analysis possible at all. Without it, a pooled analysis silently compares incomparable groups.

Both taxonomic profiles from MetaPhlAn and functional profiles from HUMAnN are provided, so “who is present” and “what they are capable of” can be asked of the same object.

Data are returned as SummarizedExperiment objects, which leads directly to the final slide.

Signatures, not samples: BugSigDB



The other two resources give you **samples**. BugSigDB gives you **published results** — curated signatures of differentially abundant taxa, with study design, body site, outcome, and statistical method in controlled vocabulary.

>2,500 signatures from >600 studies at release; community-editable since (Geistlinger et al., 2023).

i Gene set enrichment analysis, for microbes

Signatures export as **GMT files**, the same format as MSigDB gene sets — so your taxa become a query set and you ask which published signatures they overlap more than chance. `bugsigdbr` does it in R.

Disclosure: I am an author on this one as well.

The distinction from the previous two slides is the unit of curation. MicroBioMap and curatedMetagenomicData curate *samples* — you get abundance matrices and re-analyze them. BugSigDB curates *published results*: for each differential abundance study, the list of taxa reported as enriched or depleted, together with structured metadata on host species, body site, condition, geography, and the statistical method used. That metadata is recorded against ontologies and NCBI taxonomy rather than free text, which is what makes it queryable.

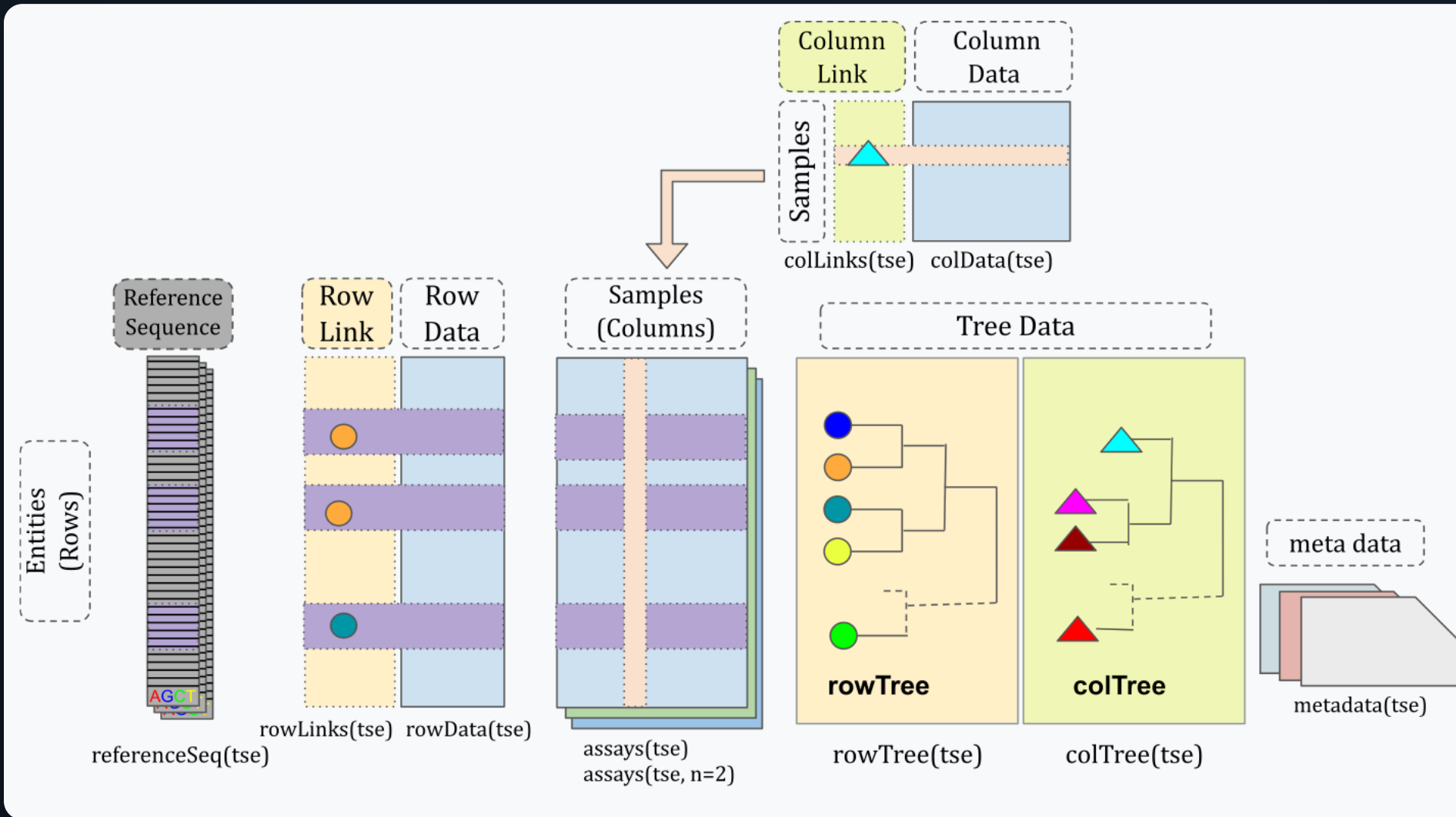
The initial release held more than 2,500 signatures from over 600 studies across three host species; it is a community-editable semantic wiki, so it has grown since and anyone can contribute a curation.

The analogy worth making explicit is to gene set enrichment analysis. In transcriptomics you take your differentially expressed genes and test them against curated gene sets in MSigDB. BugSigDB supplies the microbial equivalent: signatures export as GMT files, the same format, so you can take the taxa that came out of your own differential abundance analysis and ask which published signatures they overlap more than chance would predict. The *bugsigdbr* Bioconductor package handles the retrieval and formatting.

Why this matters practically. A single microbiome study gives a list of taxa with p-values and little context — you cannot tell whether your hits are specific to your condition or are the taxa that show up in nearly every study. Enrichment against BugSigDB answers that directly. It also supports the database-wide questions the paper asks: which conditions produce consistent signatures across independent studies, and which taxa co-occur or mutually exclude. One finding from that analysis worth repeating is the frequent introgression of oral pathobionts into the gut across many disease-associated signatures.

The caveat: BugSigDB inherits the biases of the literature it curates. Taxa are as they were reported, against whatever reference database and pipeline each original study used, and publication bias toward positive findings is baked in. It is a map of what has been published, not of what is true.

One object to hold it all



Counts, sample metadata, taxon metadata, reference sequences **and the phylogeny** in one object — so the tree travels with the data through subsetting and filtering rather than drifting out of sync (Huang et al., 2021).

SPEAKER NOTES

The failure this structure prevents is mundane and extremely common. Counts sit in one file, sample metadata in a spreadsheet, and the phylogenetic tree in a third. Low-abundance taxa are filtered from the counts, but the tree still contains tips for organisms no longer in the matrix — and every UniFrac distance computed afterwards is quietly wrong, with no error raised.

TreeSummarizedExperiment holds the count assay, sample metadata, taxon metadata, reference sequences, and both row and column trees in a single object. Subsetting propagates across all components simultaneously: filter the object and the tree is filtered with it.

This is the general Bioconductor pattern — the same structure underlies SingleCellExperiment for single-cell data — so the concept transfers well beyond microbiome analysis. The single most useful software habit to adopt from this talk is keeping related data in one object rather than in parallel files that can drift apart.

Where to go next: the OMA book

Orchestrating Microbiome Analysis

Overview

Package: OMA
Authors:
- Tuomas Borman [aut, cre]
- Leo Lahti [aut]
- Felix GM Ernst [aut]
- and others (see the full list of contributors) [ctb]
Compiled: 2026-03-15
Package version: 0.99.1
R version: R Under development (unstable) (2026-03-12 r89607)
BioC version: 3.23
License: CC BY-NC-SA 4.0

Welcome

Important!
This is a development version of the [Orchestrating Microbiome Analysis with Bioconductor](#) (Tuomas Borman et al. 2026) book from the the [miaverse](#).

You are reading the online book, [Orchestrating Microbiome Analysis with Bioconductor](#) (Tuomas Borman et al. 2026), where we walk through common strategies and workflows in microbiome data science.

The book shows through concrete examples how you can take advantage of the latest developments in R/Bioconductor for the manipulation, analysis, and reproducible reporting of hierarchical and heterogeneous microbiome profiling data sets. The book was borne out of necessity, while updating microbiome analysis tools to work with Bioconductor classes that provide support for multi-modal data collections. Many of these techniques are generic and widely applicable in other contexts as well.

Tuomas Borman, Leo Lahti, Sudarshan Shetty, Felix M Ernst, et al. 2026. *Orchestrating Microbiome Analysis with Bioconductor* [Beta Version]. [microbiome.github.io/oma/](#).

Amezquita, Robert, Aaron Lun, Stephanie

Orchestrating Microbiome Analysis with Bioconductor — free, online, and maintained by the [miaverse](#) project. It is built on exactly the object from the previous slide.

Importing and containers, QC, diversity, differential abundance, multi-omics, machine learning — each chapter is **runnable R against example data**, so it works as a course as much as a reference.

If you take one link away from this talk, take this one.

The book is the practical companion to everything in Part 04. It is written around `mia` and `TreeSummarizedExperiment` — the object from the previous slide — so the data structure introduced there is the one you actually work in throughout. That coherence is the point: you are not stitching together tutorials that each assume a different container.

The scope maps closely onto this section. There are chapters on importing data and choosing containers, on quality control and preprocessing, on alpha and beta diversity, on association and differential abundance testing, on multi-omics integration, and on machine learning and statistical modelling. It is deliberately opinionated about method choice, which is more useful than a neutral survey when you are starting out.

Every chapter is executable. The code runs against packaged example datasets, so you can work through it without your own data in hand, then swap your data in once you have it. There are also training materials and course notes linked from the site for anyone who prefers a structured path.

Two practical notes. It is a living document — the version you read is built against the current Bioconductor release, so it stays in step with the packages rather than rotting the way a printed methods chapter does. And it is community-maintained on GitHub, so if something is unclear or wrong, an issue or a pull request is a genuinely reasonable response.

The honest caveat: it teaches you the tooling and the mechanics well, but it will not tell you whether your study design is sound or your sample size adequate. Points 5 and 6 on the next slide still apply.

Takeaways

1. **Causation is the hard part.** Koch's postulates have been rewritten twice; community-level causation is still an open methodological problem.
2. **Amplicon vs. shotgun is a real trade-off** — cost and simplicity against resolution and function. Neither is the default right answer.
3. **The reference database is a methods choice**, not a detail. Results are not comparable across database versions.
4. **A handful of microbes cause cancer; many more modulate it** — through metabolites, inflammation, barrier disruption, and drug metabolism.
5. **Low-biomass diagnostics deserve skepticism.** Contamination and batch effects can manufacture an excellent-looking AUROC out of nothing.
6. **The count matrix is sparse, compositional, and over-dispersed** — and every downstream method lives or dies on handling that.

SPEAKER NOTES

Points 5 and 6 have the longest shelf life. They are transferable methodological instincts rather than microbiome facts, and they apply to any high-dimensional biological measurement analyzed with machine learning.

A closing thought worth holding onto: this is a young field in which much of the most valuable current work consists of separating real signal from technical artifact. That is not a sign of a field in trouble. Developing error-correcting machinery — reanalysis, replication, retraction when warranted — is what a discipline does as it matures.

QUESTIONS

Thank you

Sean Davis, MD, PhD · University of Colorado
Anschutz School of Medicine

Common questions and short answers.

16S or shotgun? Depends on the question and the budget. Composition across many samples: 16S. Function or strain-level resolution: shotgun.

Is the tumor microbiome real? Yes, where orthogonal validation exists — microscopy, culture, immunohistochemistry. What is not established is pan-cancer diagnostics derived from low-biomass sequencing.

Can I improve my own microbiome? Diet, particularly dietary fibre, has the best evidence. Commercial probiotics have considerably weaker evidence than their marketing suggests, and direct-to-consumer microbiome testing is not clinically actionable.

How many samples do I need? More than expected. Effect sizes are small relative to between-person variability; estimate variance from a compendium resource rather than assuming.

What about fecal microbiota transplant? Recurrent *C. difficile* infection is the one clearly established indication. Other applications remain investigational.

References

- Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., ... Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1633–1644. <https://doi.org/10.1038/s41587-023-01688-w>
- Castellari, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R. A., & Holt, R. A. (2011). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Research*, *22*(2), 299–306. <https://doi.org/10.1101/gr.126516.111>
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. <https://doi.org/10.1128/aem.03006-05>
- Falkow, S. (1988). Molecular Koch's Postulates Applied to Microbial Pathogenicity. *Clinical Infectious Diseases*, *10*(Supplement 2), S274–S276. https://doi.org/10.1093/cid/10.supplement_2.s274
- Falkow, S. (2004). Molecular Koch's postulates applied to bacterial pathogenicity — a personal recollection 15 years later. *Nature Reviews Microbiology*, *2*(1), 67–72. <https://doi.org/10.1038/nrmicro799>
- Fredericks, D. N., & Relman, D. A. (1996). Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. *Clinical Microbiology Reviews*, *9*(1), 18–33. <https://doi.org/10.1128/cmr.9.1.18>
- Ge, Y., Lu, J., Puiu, D., Revsine, M., & Salzberg, S. L. (2025). Comprehensive analysis of microbial content in whole-genome sequencing samples from The Cancer Genome Atlas project. *Science Translational Medicine*, *17*(814). <https://doi.org/10.1126/scitranslmed.ads6335>
- Geistlinger, L., Mirzayi, C., Zohra, F., Azhar, R., Elsafoury, S., Grieve, C., Wokaty, J., Gamboa-Tuz, S. D., Sengupta, P., Hecht, I., Ravikrishnan, A., Gonçalves, R. S., Franzosa, E., Raman, K., Carey, V., Dowd, J. B., Jones, H. E., Davis, S., Segata, N., ... Waldron, L. (2023). BugSigDB captures patterns of differential abundance across a broad range of host-associated microbial signatures. *Nature Biotechnology*, *42*(5), 790–802. <https://doi.org/10.1038/s41587-023-01872-y>
- Geller, L. T., Barzily-Rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z. A., Shee, K., Thaiss, C. A., Reuben, A., Livny, J., Avraham, R., Frederick, D. T., Ligorio, M., Chatman, K., Johnston, S. E., Mosher, C. M., ... Straussman, R. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, *357*(6356), 1156–1160. <https://doi.org/10.1126/science.aah5043>
- Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C. S., Brewer, D. S., Perte, M., & Salzberg, S. L. (2023). Major data analysis errors invalidate cancer microbiome findings. *mBio*, *14*(5). <https://doi.org/10.1128/mbio.01607-23>
- Gihawi, A., Wood, H. M., Clark, J., O'Grady, J., Eeles, R. A., Wedge, D. C., Jakobsdottir, G. M., Magiorkinis, G., Schache, A. G., Masterson, L., Lechner, M., Fenton, T. R., Jones, T. M., Flanagan, A. M., De Noon, S., Rubinsteyn, A., Hurst, R., Cooper, C. S., & Brewer, D. S. (2025). The landscape of microbial associations in human cancer. *Science Translational Medicine*, *17*(814). <https://doi.org/10.1126/scitranslmed.ads6166>
- Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., Vicente, D., Hoffman, K., Wei, S. C., Cogdill, A. P., Zhao, L., Hudgens, C. W., Hutchinson, D. S., Manzo, T., Petaccia de Macedo, M., Cotechini, T., Kumar, T., Chen, W. S., ... Wargo, J. A. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science*, *359*(6371), 97–103. <https://doi.org/10.1126/science.aan4236>

- Huang, R., Soneson, C., Ernst, F. G. M., Rue-Albrecht, K. C., Yu, G., Hicks, S. C., & Robinson, M. D. (2021). TreeSummarizedExperiment: A S4 class for data with hierarchical structure. *F1000Research*, *9*, 1246. <https://doi.org/10.12688/f1000research.26669.2>
- Kandalai, S., Li, H., Zhang, N., Peng, H., & Zheng, Q. (2023). The human microbiome and cancer: A diagnostic and therapeutic perspective. *Cancer Biology & Therapy*, *24*(1). <https://doi.org/10.1080/15384047.2023.2240084>
- Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R. A., Ogino, S., Birren, B. W., Huttenhower, C., Garrett, W. S., & Meyerson, M. (2011). Genomic analysis identifies association of *fusobacterium* with colorectal carcinoma. *Genome Research*, *22*(2), 292–298. <https://doi.org/10.1101/gr.126573.111>
- Le Noci, V., Guglielmetti, S., Arioli, S., Camisaschi, C., Bianchi, F., Sommariva, M., Storti, C., Triulzi, T., Castelli, C., Balsari, A., Tagliabue, E., & Sfondrini, L. (2018). Modulation of Pulmonary Microbiota by Antibiotic or Probiotic Aerosol Therapy: A Strategy to Promote Immunosurveillance against Lung Metastases. *Cell Reports*, *24*(13), 3528–3538. <https://doi.org/10.1016/j.celrep.2018.08.090>
- McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J. T., Nicolaou, G., Parks, D. H., Karst, S. M., Albertsen, M., Hugenholtz, P., DeSantis, T., Song, S. J., Bartko, A., Havulinna, A. S., Jousilahti, P., Cheng, S., Inouye, M., ... Knight, R. (2023). Greengenes2 unifies microbial data in a single reference tree. *Nature Biotechnology*, *42*(5), 715–718. <https://doi.org/10.1038/s41587-023-01845-1>
- Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., ... Straussman, R. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science*, *368*(6494), 973–980. <https://doi.org/10.1126/science.aay9189>
- Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11), 1023–1024. <https://doi.org/10.1038/nmeth.4468>
- Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolk, T., Janssen, S., Metcalf, J., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., McKay, R., Patel, S. P., Swafford, A. D., & Knight, R. (2020). RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, *579*(7800), 567–574. <https://doi.org/10.1038/s41586-020-2095-1>
- Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolk, T., Janssen, S., Metcalf, J., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., McKay, R., Patel, S. P., Swafford, A. D., & Knight, R. (2024). Retraction Note: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, *631*(8021), 694–694. <https://doi.org/10.1038/s41586-024-07656-x>
- Qin, J. and, Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., ... Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. <https://doi.org/10.1038/nature08821>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Seedorf, H., Kittelmann, S., Henderson, G., & Janssen, P. H. (2014). RIM-DB: A taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. *PeerJ*, *2*, e494. <https://doi.org/10.7717/peerj.494>
- Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, *14*(8), e1002533. <https://doi.org/10.1371/journal.pbio.1002533>
- Shiao, S. L., Kershaw, K. M., Limon, J. J., You, S., Yoon, J., Ko, E. Y., Guarnerio, J., Potdar, A. A., McGovern, D. P. B., Bose, S., Dar, T. B., Noe, P., Lee, J., Kubota, Y., Maymi, V. I., Davis, M. J., Henson, R. M., Choi, R. Y., Yang, W., ... Underhill, D. M. (2021). Commensal bacteria and fungi differentially regulate tumor responses to radiation therapy. *Cancer Cell*, *39*(9), 1202–1213.e6. <https://doi.org/10.1016/j.ccell.2021.07.002>

- Uemura, N., Okamoto, S., Yamamoto, S., Matsumura, N., Yamaguchi, S., Yamakido, M., Taniyama, K., Sasaki, N., & Schlemper, R. J. (2001). *Helicobacter pylori* Infection and the Development of Gastric Cancer. *New England Journal of Medicine*, *345*(11), 784–789. <https://doi.org/10.1056/nejmoa001999>
- Uribe-Herranz, M., Rafail, S., Beghi, S., Gil-de-Gómez, L., Verginadis, I., Bittinger, K., Pustynnikov, S., Pierini, S., Perales-Linares, R., Blair, I. A., Mesaros, C. A., Snyder, N. W., Bushman, F., Koumenis, C., & Facciabene, A. (2019). Gut microbiota modulate dendritic cell antigen presentation and radiotherapy-induced antitumor immune response. *Journal of Clinical Investigation*, *130*(1), 466–479. <https://doi.org/10.1172/jci124332>
- Viaud, S., Saccheri, F., Mignot, G., Yamazaki, T., Daillère, R., Hannani, D., Enot, D. P., Pfirschke, C., Engblom, C., Pittet, M. J., Schlitzer, A., Ginhoux, F., Apetoh, L., Chachaty, E., Woerther, P.-L., Eberl, G., Bérard, M., Ecobichon, C., Clermont, D., ... Zitvogel, L. (2013). The Intestinal Microbiota Modulates the Anticancer Immune Effects of Cyclophosphamide. *Science*, *342*(6161), 971–976. <https://doi.org/10.1126/science.1240537>