WHAT ONE DIGESTATE SAMPLE CAN TELL YOU

How a genome-resolved digester analysis is built, what it measures, and where it stops

Taavi Päll

2026-10-30

ABSTRACT

Anaerobic digestion is managed with numbers that describe the outcome (gas yield, FOS/TAC, volatile fatty acids) and not the organisms that produce them. DNA from a single digestate sample can name those organisms and the metabolic routes they carry. This article describes how that analysis is built, what it returns, and what it cannot answer.

THE QUESTION THIS ANSWERS

A digester operator has good instrumentation for what came out and almost none for who did the work. Gas yield falls, FOS/TAC drifts, and the plant answers by changing feed or load: a decision about microorganisms, based on numbers that describe none of them.

The decisions are concrete. Whether to raise the organic loading rate, and how fast. Whether to add more protein-rich material to the feed, or less. Whether to buy trace elements. Whether a drop in gas output is due to the feed, the temperature, or the community. Each of these is a question about organisms, and the instruments answer a different question: they say what the reactor produced yesterday.

A plant that runs steadily can live with that gap. It cannot when something changes: a new feedstock, a higher load, a start-up, a recovery after an upset. At

exactly those moments, the operator has the least information about the part of the process that is actually doing the work.

A digestate sample carries the DNA of that part. Reading it does not replace the instruments and does not predict tomorrow's gas curve. It answers a narrower question the instruments cannot: which organisms are present, which steps of the process they can carry, and how much reserve remains if the load changes.

WHAT IS MEASURED

A digestate sample carries the community's DNA, which sequencing turns into short reads. The analysis uses those reads in two ways at once, because the two answer different questions.

Assembly reconstructs the genomes present in that sample from the sequencing reads. These are the client's own organisms, not references: each genome is a specific strain in that reactor, and each is scored for the metabolic steps it carries: hydrolysis, acid formation, and the three routes to methane.

Mapping compares the same reads against a reference collection from public genome databases without assembling anything. It finds organisms that are present but too rare to reconstruct, and it gives their abundance.

Both readings use the same panel: 86 marker genes grouped into 20 modules, covering hydrolysis of cellulose and protein, acid formation, syntrophic acetate oxidation, and each route to methane. A genome is credited with a route only when it carries the marker genes for the steps of that route and its lineage is one where the route is established. The lineage rules exist because genes alone mislead: anaerobic methane oxidisers carry the methanogenesis genes and run them backwards, and several methylotrophic genera carry markers that would otherwise read as a hydrogen route.

The metabolic marker-gene panel is built only on sources that are free for commercial use: NCBIfam, Pfam and Rhea. The usual reference databases for this work, KEGG and MetaCyc, require a paid licence for commercial analysis. That constraint is why this panel exists and why a fixed-price report is possible.

Every rule in that policy carries a literature citation, and the panel, the scorer and the benchmarks are public (10.5281/zenodo.22913615) and described in a preprint (10.64898/2026.09.28.755039), so a disputed call can be checked rather than argued.

WHAT ONE SAMPLE RETURNS

The first question was how deeply a sample needs to be sequenced to describe the methanogenic community in the reactor. It matters because sequencing cost follows depth directly. The measurements in this article come from one reactor: plant BP3 in North Rhine-Westphalia, a 40 °C stirred-tank digester fed maize silage and pig manure, sampled twice at different time points and published as public data. The depth tests use the second sample, which was sequenced very deeply, to 87 Gb, and then read at depths relevant to monitoring, 2.5, 6 and 12 Gb, by drawing subsets of it.

Assembly keeps paying for depth, but less for each step. Going from 2.5 to 6 Gb doubled the good-quality genomes from the same sample, from 23 to 47; doubling again to 12 Gb added another 19, to 66 (Table 1). Per gigabase sequenced, the second step returned about half as many new genomes as the first. Mapping behaves differently, and two counts have to be kept apart. What mapping *sees* keeps rising with depth, because progressively rarer organisms leave a few reads behind. What a report *names* does not: an organism is named only at 0.1 % abundance or more with at least a quarter of its genome covered, and at that floor the same six methanogens are named at every depth, 2.5 Gb included (Figure 1). Across all genomes the named count moves by two (117, 115, 115) only because a handful of genomes sit right at 0.1 % and cross the floor in one direction or the other as the estimate firms up. Depth does not change which methanogens are named. It changes how firmly they are established: the acetate-route genome in the next section goes from under half of its genome covered at 2.5 Gb to nine tenths at 12 Gb,

	GENOMES:		
DEPTH	GOOD GENOMES	SEEN / NAMED	METHANOGEI SEEN / NAM
2.5 Gb	23	374 / 117	14
6 Gb	47	482 / 115	18
12 Gb	66	574 / 115	19

Table 1: The second BP3 sample at three sequencing depths. Good genomes: assembled at ≥ 50 % completeness and ≤ 10 % contamination. Seen: any genome mapping finds. Named: seen at ≥ 0.1 % abundance with at least a quarter of its genome covered, the floor below which a report does not name an organism.

and an organism right at the floor clears the quarter-genome test comfortably at 6 Gb and with less room at 2.5 Gb.

The archaea that assembly recovers say the same thing from the other side. Every depth assembled the hydrogen-route *Methanoculleus*, two genomes at 2.5 Gb and three from 6 Gb on. At 6 Gb an unnamed DPANN archaeon, which is not a methanogen, was also assembled, and 12 Gb added no further archaea. No depth assembled the acetate route. So the practical reading is this: on this reactor, 6 Gb names the methane side of the community as completely as 12 Gb and with a safe margin at the floor; the extra depth buys bacterial genomes, which carry hydrolysis and acid formation, and a stronger base for the genome set a series is built on. Which depth the service runs at is not fixed yet. It rests on two things this article cannot supply: the price difference between depths, and whether a second reactor behaves like this one, which the first client samples will show.

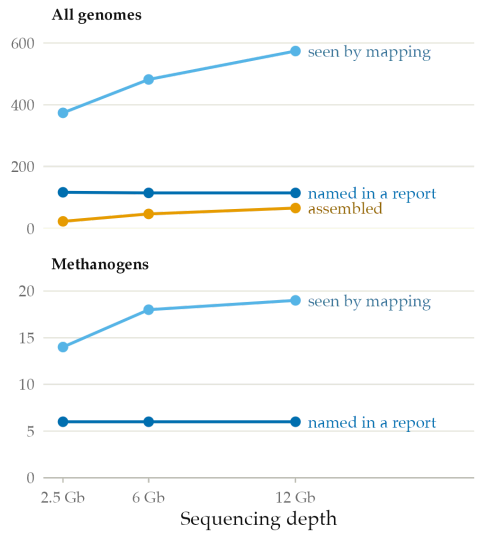


Figure 1: The second BP3 sample at three sequencing depths. All genomes: what mapping sees, what a report names, and good genomes recovered by assembly. Methanogens: seen and named. What mapping sees keeps rising with depth; what a report names does not, and for methanogens it is the same six genomes at every depth.

THE MEASUREMENT THAT DECIDES THE DESIGN

The reactor in these tests carries its acetate route on a single *Methanosarcina flavescens* genome at roughly 0.13 % relative abundance of the genomes mapping detected. The route matters out of proportion to its abundance: it is one of the two ways acetate becomes methane, and the genus that carries it determines how the reactor behaves when ammonia rises.

At the sequencing depths a service can sell, that genome is never assembled. It is always detected (Table 2).

The right-hand column is the useful check. At a coverage of c, a genome's k-mers should be seen with probability $1 - e^{-c}$, and the column computes that from each row's own coverage. The measured share runs a

few points below it at every depth, and by about the amount it should: this strain differs from the reference genome by about 0.1–0.3 % (sylph puts the identity at 99.7–99.9 %), and any 31-mer spanning a difference cannot match, which removes roughly one k-mer in twenty. An organism that is present, close to its reference, and sampled at random behaves like this; a pile of reads from one conserved gene does not.

The same data show where the check stops. A *Methanothrix* appears at 6 and 12 Gb at about 0.003 %, with 3–6 % of its genome covered. That is about what its coverage predicts, so the data cannot tell a rare *Methanothrix* from reads of a relative landing on conserved regions: both would cover a few percent of the genome. The report does not name what it cannot confirm, and requiring a quarter of the genome to be covered is what enforces that. Since *Methanothrix* and *Methanosarcina* respond to ammonia in opposite ways, naming one on this evidence would not have been a harmless mistake.

The design follows from this. A report that only assembles would tell this plant it has no acetate route. A report that only maps would never name the client's own organisms. The analysis does both, and says which is which.

DEPTH	ABUNDANCE	COVERED	EXPECTED
2.5 Gb	0.132 %	46 %	48 % at 0.66×
6 Gb	0.126 %	75 %	78 % at 1.52×
12 Gb	0.137 %	91 %	96 % at 3.27×

Table 2: *Methanosarcina flavescens*, the genome carrying the acetate route in this reactor. Abundance: taxonomic abundance from sylph, the mapping tool: a coverage-weighted share of the genomes detected. Covered: share of the genome's sampled 31-mers found in the reads. Expected: $1 - e^{-c}$ at that row's coverage c. Never assembled at these depths, always detected.

HOW A SUGGESTION IS PRODUCED

A report that stopped at tables would leave the reader to do the translation. The report goes one step further and produces suggestions, but they are not written per client, and they are not opinions. Each one is a row in a table, with a condition that either holds for a sample or does not:

Same tables in, same suggestions out. The rules are ordered so that the cheapest action comes first: measuring something you already measure, before sampling again, before touching the load or the feed.

Each line in the report carries the finding that produced it and, where one exists, the literature behind it.

What the suggestions deliberately are not is instructions. The report says what the evidence supports, and the operator, who has the process data and the plant, decides. For the levers an operator actually pulls (loading, feed, retention time, temperature, additives, inoculum), a suggestion may not use the imperative at all, and the renderer refuses to build a report where one does.

FIELD	VALUE
rule	`LOAD-THIN`
fires when	`acetate_status in note,attn & reactor_type ≠ dry_batch`
lever	loading
text	The genes do not support a higher loading rate before the next sample.
why	The acetate route is thin, filled in with that sample's numbers

WHAT CHANGES BETWEEN TWO SAMPLES

One sample describes a community. Two describe a change, and a change is harder to report than a state.

The first difficulty is that shares are compositional. If bacteria bloom, every methanogen share falls without anything happening to the methanogens. In the reactor used here, the methanogens went from 8.0 % to 4.2 % of mapped reads between samples, which reads as a collapse until you look within the methanogens, where the hydrogen route went from about 99 % to about 100 %. Nothing moved except the denominator. Route fractions within the methane producers are therefore reported alongside shares of the whole community, and the denominator is stated every time.

The second is that "the same genome" has to mean something across samples. Genomes assembled separately from two samples are two different sequences, so a genome that looks absent the second time may be the same organism binned differently. A series is therefore built against one genome set: genomes are clustered at 95 % identity into a versioned set, every sample is mapped back to it, and a genome keeps its identifier for the life of the series. When a later sample adds a genome, the earlier samples are re-mapped so the whole series speaks about the same objects.

The third is that “changed” needs a threshold, and one sample per quarter cannot supply one. The practical answer is to extract one sample of the series twice and sequence both. That single extra library measures how much the method itself varies on that client’s material, and every later change is judged against it.

There is also an arithmetic reason a series returns more than the sum of its samples. Genomes too rare to assemble from one sample can be assembled when sequencing reads from several samples are pooled, both because coverage accumulates and because genomes are easier to tell apart when they rise and fall differently between samples. Four quarterly samples at 12 Gb accumulate 48 Gb, at which point an organism at 0.1 % abundance reaches roughly tenfold coverage: enough to assemble. The first report describes the abundant community and detects the rare organisms by mapping. Each further sample moves some of them from the second category into the first.

LIMITS

Everything above describes genetic capacity. It is what the organisms present can do, not what they did last week, and no DNA method measures activity.

The depths in this article were produced by subsampling two deep public datasets rather than by sequencing the same material shallowly. For the detection question that is a fair model: reading one library less deeply is a random draw from the same pool of molecules, and the measured coverage breadth matched the theoretical prediction at all three depths, which is what a fair draw looks like. For the assembly counts it is an upper bound. A library prepared from less input DNA, or with more amplification, carries more duplicate reads and fewer distinct molecules, so a real run at the same nominal depth can be worth slightly less than the numbers here. The first client

sample sequenced at a chosen depth settles that directly.

Communities change over weeks and months, while acidification shows up in FOS/TAC within days. For anything fast, the chemistry is the earlier warning, and nothing here replaces it.

There are two detection floors, and they are far apart. Mapping sees organisms below 0.1 %, but a report names organisms only down to about 0.1 %, with a quarter of the genome covered, because below that a detection cannot be confirmed. Assembly, at the depths in this article, needs roughly ten times that. The counts also depend on the reference: the monitoring example maps the same reactor against its own published genomes rather than against the Genome Taxonomy Database (GTDB), and so lists a somewhat different set of methanogens. A report therefore also names organisms that could not be assembled from the sample itself.

Five of the 20 panel modules have no reliable public-domain marker and are reported as not assessed rather than as zero: butyrate oxidation, ethanol fermentation, lactate fermentation, direct interspecies electron transfer, and energy conservation. A module that cannot be scored is stated as such in every report, because a zero and a blank are different claims.

Taxonomy is a classification against a specific database release, not a species identification, and the release number is printed in every report. Mapping sees only organisms that a reference collection contains, which for agricultural digesters in Europe is good coverage and for an unusual feedstock may not be.

Glycolysis is not scored for a different reason, being assumed present in every genome.

The measurements in this article come from one reactor, sampled twice, from public data, which makes this a demonstration of the method and not a validation across plants. The panel itself has been benchmarked more broadly: 1,401 genomes from 134 digesters, and 3,043 reference genomes spanning 203 phyla.

Finally, there is no laboratory accreditation behind this and no ring-test participation. Sequencing is subcontracted and the analysis is one person's work. What stands in place of a certificate is that the method is public and the benchmarks can be re-run by anyone who doubts them.

METHOD AND PROVENANCE

Genomes are assembled with MEGAHIT after normalisation, binned with four binners and refined with Binette, and quality is scored with CheckM2. Taxonomy comes from GTDB-Tk against GTDB release 232. Mapping for detection uses sylph against GTDB, and abundances against a client's own genome set use CoverM. Panel scoring uses HMMER against the marker set described above. Every report prints the version of each tool and the release of each database it used, together with a checksum of the result tables it was built from.

The panel, the scorer, the lineage policy and the benchmark scripts are public and archived (10.5281/zenodo.22913615), and the method and both benchmarks are described in a bioRxiv preprint (10.64898/2026.09.28.755039). The report renderer is not public. Nothing in the analysis depends on it: the tables it formats are delivered with every report.