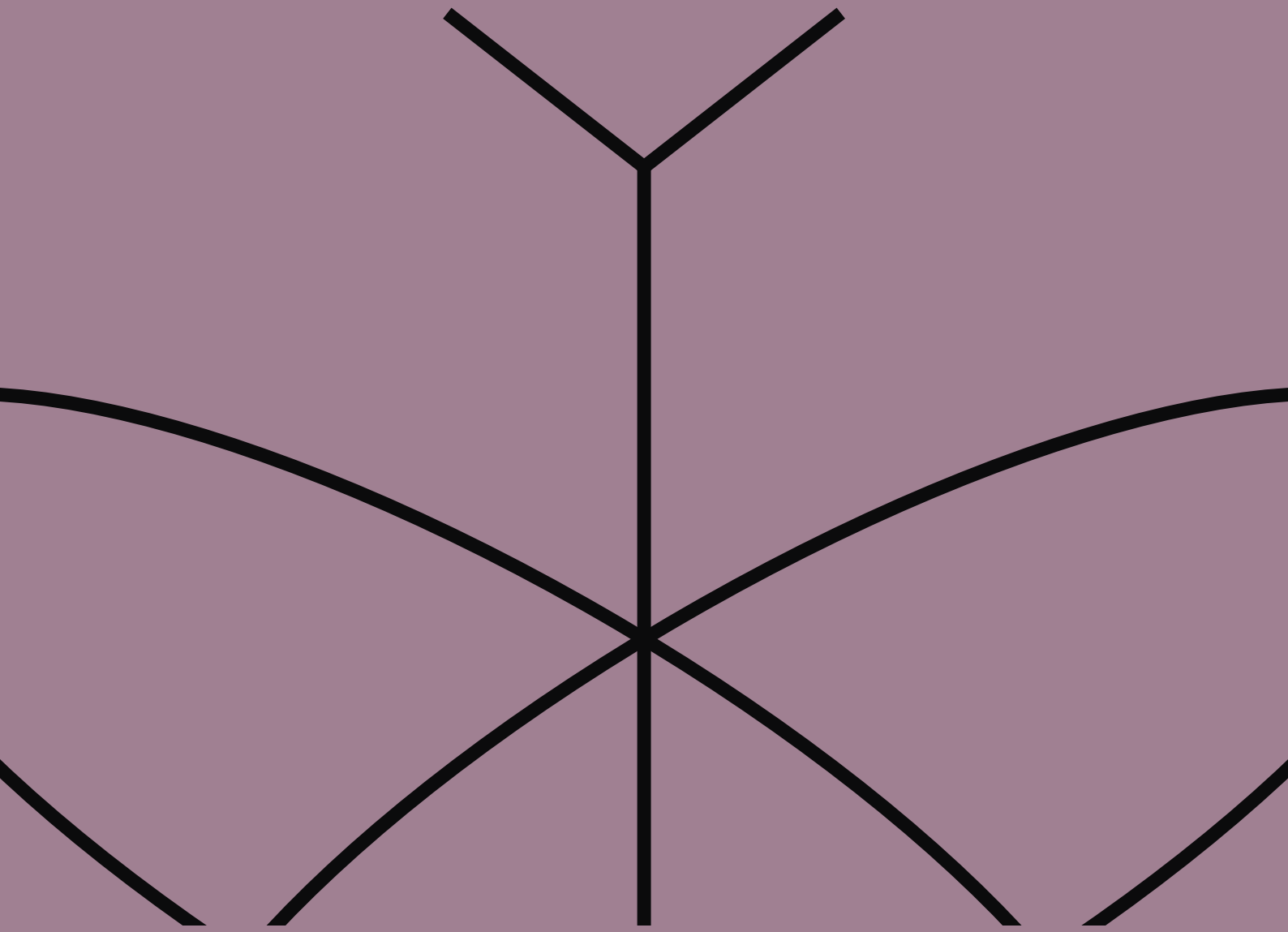


# Bench by Phalaena Automata: 42-run platform validation

Forty-two public datasets run end to end across seven omics pipelines, graded by a question battery and by RAGAS-style judges

PHALAENA
AUTOMATA

phalaena.ai

# 1 Abstract

Bench runs seven omics pipelines (short-read and long-read genomics, bulk, single-cell, and spatial transcriptomics, proteomics, and metabolomics) on a managed cloud backend, and answers questions about each run through a research assistant grounded in a single machine-readable summary contract. This brief reports the platform's real-data validation across two expansion waves. We selected 42 public datasets, most of them from published studies, ran each end to end through the matching pipeline, and confirmed completion from the published summary of each run rather than from job status alone. Wave 2 anchored every pipeline on 21 landmark datasets (closed at 21 of 21 on 2026-05-11). Wave 3 added 21 gap-filling datasets chosen for the organisms, chemistries, and sample types the first corpus did not cover, from East-Asian and Drosophila genomes to microbial and plant long reads, the first real 10x v2 single-cell chemistry, non-brain spatial tissues, and microbial and environmental metabolomes and a compound-standard benchmark (closed at 21 of 21 on 2026-06-11). Every summary records the status success, yet an audit of these runs (Palanikumar 2026, https://doi.org/10.20944/preprints202611.0178.v1) found a confirmed silent failure, traced to code, in 14 of the 42, in five classes. We grade answers two ways: a mechanical question battery (on its final pass over the 21 Wave-2 runs, 87 of the 98 questions it could grade matched the published summaries, 88.8 percent; graded live on the nine runs then complete, the deployed assistant passed 56 of 62, 90.3 percent), and RAGAS-style faithfulness, answer relevancy, and context precision of answers generated from each run summary, scored by Vertex Gemini judges across all three waves. Both measure agreement with the summaries, so neither can show that a summary is right. An earlier HG002 run matched a prior baseline to the digit on eight summary metrics, which shows that the short-read CPU path is reproducible on one human sample, not that its calls are correct. We report the headline scientific outputs each run produced and the limits, the audit's confirmed failures among them: the GPU variant-calling path is wired and fires but has not been validated, three non-human bulk RNA-seq runs (fly, Arabidopsis and yeast) were quantified against the default human index, two non-human short-read runs were annotated with the human gene model, the long-read caller used a haploid model on two diploid human genomes, no proteomics run applied its 1 percent false-discovery filter, metabolomics features were matched across samples on exact keys that replicate injections rarely share, one Wave-2 spatial run processed reads from another study, and some Wave-3 read sets are fixed-prefix subsets suitable for pipeline-shape validation rather than composition claims.

# 2 Platform shape

Bench presents four surfaces on one stack: a submission and results web app, a Nextflow pipeline layer on Google Batch, a research assistant, and a public wiki of every validation run. A submission names a pipeline and a samplesheet; the backend builds and runs the workflow, publishes outputs to cloud storage, and writes one `summary.json` per run against a fixed schema. The assistant is meant to read its numbers from that summary rather than derive them.

## 2.1 Why the summary contract is the platform

The summary contract is what lets one assistant speak for seven pipelines without bespoke per-modality code. Each pipeline emits the same envelope (pipeline identity, inputs, per-sample metrics, aggregate metrics, artifacts, and a short natural-language summary), and the assistant reads that envelope. The same contract is what the validation read. A run counts as complete when it publishes a summary. A summary can be wrong, and the question battery and RAGAS both score answers against it, so neither validates a run. This brief is, in effect, a reading of 42 such summaries.

# 3 The 42-run corpus

We organize the corpus by pipeline. Each pipeline carries three Wave-2 anchor datasets and three Wave-3 gap-filling datasets. Wave 2 favored landmark, truth-set-adjacent data; Wave 3 was selected under a strict runnable-now rule to extend each pipeline into organisms and sample types the anchors missed.

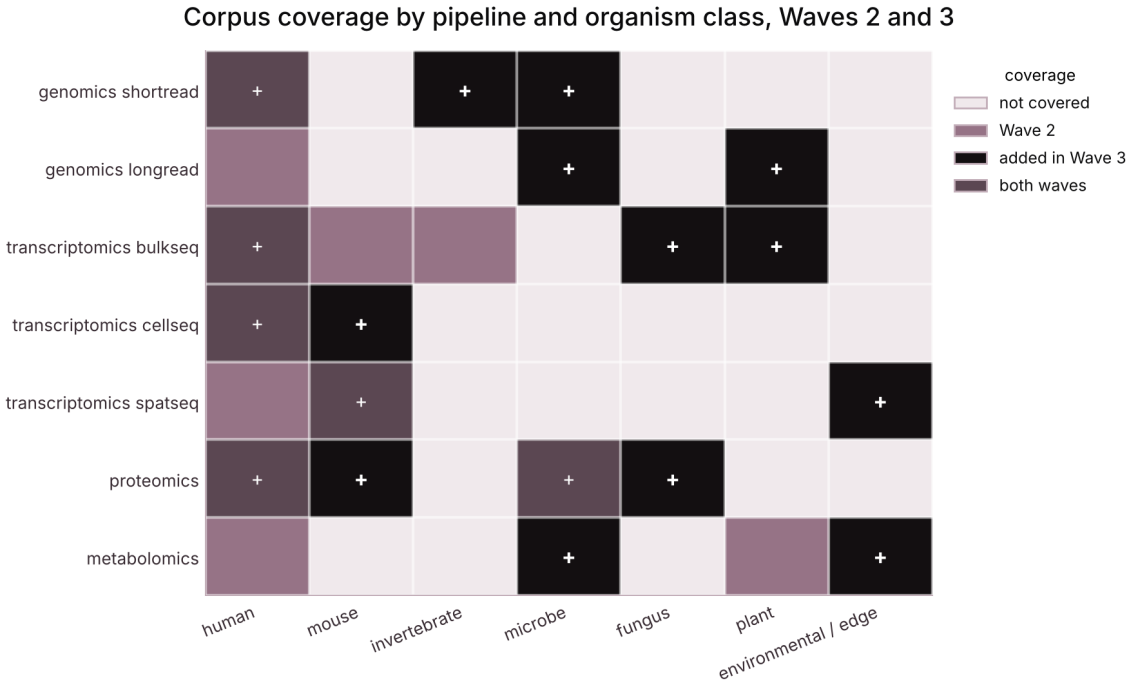



**Figure 1.** Coverage of the validation corpus by pipeline and organism class. Wave 2 anchored each pipeline largely on human and mouse reference data; Wave 3 added at least one new organism class or edge case to every pipeline, from invertebrate and microbial genomes to plant, fungal, and environmental samples. A shaded cell records a dataset of that class, not a result that held: the fly, Arabidopsis and yeast bulk RNA-seq runs, for example, were quantified against a human index (see Scope and limitations).

## 3.1 Short-read genomics

The anchors are the Genome in a Bottle Ashkenazi son HG002 (Zook 2016, PMID 27271295), Platinum Genomes NA12878 (Eberle 2017, PMID 27903644), and the 1000 Genomes high-coverage NA12878 re-sequencing (Byrska-Bishop 2022, PMID 36055201), of which the run used the first 20 million read pairs. Wave 3 adds the Han Chinese son HG005 for a second ancestry (the GIAB panel, PMID 27271295), a Drosophila Genetic Reference Panel line recorded as DGRP-802 for the first invertebrate genome (the panel papers are Mackay 2012, PMID 22318601, and Huang 2014, PMID 24714809; which of them produced these reads is not established), and an E. coli K-12 MG1655 resequencing as a microbial smoke test. The HG005 run produced 1,469,090 variants (1,294,920 single-nucleotide and 174,291 indel) at 99.5 percent mapping; the DGRP line produced 1,212,910 variants at 90.67 percent mapping, annotated with the human SnpEff database and the human ClinVar file, which gave them eight ClinVar records, one of them “pathogenic”; the E. coli run produced 329 variants at 99.46 percent against its 4.6 Mb reference, whose impact counts from the same human SnpEff database sum to zero (see Scope and limitations for both).

## 3.2 Long-read genomics

The anchors are Oxford Nanopore NA12878 (Jain 2018, PMID 29431738), PacBio HiFi HG002 from a Genome in a Bottle data release (one Sequel II SMRT Cell; described here earlier as chromosome 20 only and credited to Wenger 2019, which did not produce these reads; its 15.4 Gb of reads and

3.25 million variants are genome-wide), and one SMRT Cell of CHM13 HiFi reads (Vollger 2020, PMID 31711268), reads later used to assemble T2T-CHM13 (Nurk 2022, PMID 35357919). Wave 3 adds two bacterial Nanopore genomes, E. coli CFT073 on R10.4 (Sanderson 2023, PMID 36748454) and Pseudomonas aeruginosa PAO1 on R10.4.1 (Sanderson 2024, PMID 38713194), and an Arabidopsis thaliana Nanopore genome (Michael 2018, PMID 29416032) that is the first plant-scale genome in the corpus at 135 Mb. The Arabidopsis run produced 689,146 variants (662,447 single-nucleotide and 26,699 indel) at 82.3 percent mapping, with variant calling scattered across 16 reference windows so the work fits preemptible workers.

## 3.3 Bulk transcriptomics

The anchors are untreated Drosophila S2-cell libraries from the pasilla knockdown study (Brooks 2011, PMID 20921232), with no knockdown library among them, the airway smooth-muscle dexamethasone study (Himes 2014, PMID 24926665), and the Bottomly mouse strain comparison (Bottomly 2011, PMID 21455293). Wave 3 adds a clinical polyA-versus-ribodepletion blood library-prep contrast (Zhao 2018, PMID 29556074), an Arabidopsis Col-0 wild-type set (Froussios 2019, PMID 30726870), and the yeast wild-type-versus-snf2 two-condition benchmark (Gierlinski 2015, PMID 26206307). The library-prep contrast produced a large mapping split, but in the direction opposite to the cited study: the libraries labelled polyA mapped at 31.5 and 31.8 percent and those labelled rRNA-depleted at 84.6 and 85.7 percent, whereas Zhao 2018 found higher exonic coverage under polyA selection. The contrast ran the wrong way because the public deposit’s protocol labels appear to be swapped, not because the pipeline erred: the genes Zhao 2018 list as most different between the protocols are all inverted in this run, and in the reads that the European Nucleotide Archive holds for all 16 runs of the deposit, each run carries the markers of the protocol opposite to its library name. The depositors could confirm the swap. The Arabidopsis and yeast runs completed but were quantified against the default human index, as reported in the limitations, and so was the Wave-2 pasilla anchor: its 0.15 to 0.24 percent mapping and human top transcripts identify the same default.

## 3.4 Single-cell transcriptomics

The anchors are the 10x Genomics PBMC 3k v1 public dataset (the platform is described in Zheng 2017, PMID 28091601), a resting-versus-LPS PBMC pair (Derbois 2023, PMID 37414801), and a 10x 1k PBMC v3 reference set. Wave 3 adds a human glioblastoma 10x v3 set for the first solid tumor, a mouse neuron 10x v3 set for the first mouse single-cell, and a mouse heart set on 10x v2 chemistry, the first real v2 data the pipeline has processed; it previously declared the chemistry but had never run it. The glioblastoma run recovered 3,454 cells at 4,278 median genes per cell across 10 clusters; the mouse neuron run recovered 1,087 cells across 12 clusters; the v2 heart run recovered 437 cells across 10 clusters. All three ingested native multi-lane 10x layouts directly, feeding every lane of a library to one alignment.

## 3.5 Spatial transcriptomics

The anchors are the dorsolateral prefrontal cortex Visium reference (Maynard 2021, PMID 33558695) and two 10x mouse-brain Visium sections. The run filed under the dorsolateral prefrontal cortex reference did not process that tissue: the validation’s download fetched Sequence Read Archive run SRR10903401, which belongs to a metatranscriptome study of bronchoalveolar lavage from two pneumonia patients (Chen 2020, PMID 32020836), and the run completed on those reads with a success status, so it validates nothing about the Maynard reference. Wave 3 moves spatial validation

off brain tissue entirely: a mouse spleen malaria model (Du 2025, PMID 40156041), a healing mouse colon after injury (Parigi 2022, PMID 35149721), and a mouse kidney ischemia-reperfusion set on formalin-fixed, probe-based Visium (Xuanyuan 2025, PMID 40813851). These runs exercise the read-intake path, whose barcode profile counts the 16-base prefixes of read 1 as barcodes without comparing them to a barcode list, so it cannot tell Visium reads from other data: the Wave-2 run above passed it on a metatranscriptome. Downstream spatial quantification is out of scope for this mode.

## 3.6 Proteomics

The anchors are the MaxLFQ HeLa-plus-E. coli benchmark (Cox 2014, PMID 24942700), two unrecorded raw files of an eleven-cell-line proteome deposit (PXD002395), and the PXD000584 deposit, catalogued here as a HeLa proteome but, according to its PubMed record, a gastric-cancer biopsy analysed by MudPIT (Aquino 2014, PMID 24283986). Wave 3 adds the first prokaryote, fungus, and mouse proteomes: an E. coli K-12 quantitative proteome (Schmidt 2016, PMID 26641532), a yeast-plus-UPS1 label-free standard (Ramus 2016, PMID 26862574), and a mouse draft proteome (Giansanti 2022, PMID 35710609). The E. coli run reported 4,143 proteins and the mouse run 13,278. Neither is a 1 percent false-discovery result: the filter step aborted on an option OpenMS 3.5.0 rejects, and the fallback passed the list on unfiltered (see Scope and limitations); in the E. coli run's two samples 48.5 and 49.7 percent of peptide-spectrum matches carry q-values above 0.01.

## 3.7 Metabolomics

The anchors are the sacurine human-urine study (Thevenot 2015, PMID 26088811), a Euphorbia dendroides fractionation study (Nothias 2018, PMID 29498278), whose data Nothias 2020 (PMID 32839597) reanalysed, and a NIST plasma standard (Stancliffe 2021, PMID 34239103). Wave 3 adds the first microbial and environmental metabolomes and a compound-standard benchmark: Staphylococcus aureus USA300 culture supernatants, four biological replicates from a study that grew P. aeruginosa and S. aureus strains in monoculture (Dunphy 2021, PMID 34156287), four sections of one deep-sea sediment core from a methane-seep study, on an Agilent QTOF (Redick 2023, PMID 39268414; whether the core comes from a seep is not recorded), and two standard mixtures of the same 1,100 compounds from a software benchmark (Li 2018, PMID 29907290). The deep-sea environmental run reported 52,117 unique features at 13,034 mean features per sample, nearly the sum of its four per-sample counts, because features were not matched across samples (see Scope and limitations).

# 4 Cross-pipeline results

All 42 runs reached a published summary. The summaries were not validated against their JSON schemas when the runs completed, and 20 of the 21 Wave-3 summaries fail their pipeline's schema (16 lack the required schema version and 12 the wall time), so published here does not mean schema-valid. Wall-clock ranges from a few minutes for the lighter quantification pipelines to several hours for whole-genome variant calling, and each of the 42 kept runs completed inside a bounded window on the managed backend; how many other attempts ended in an error is not recorded.

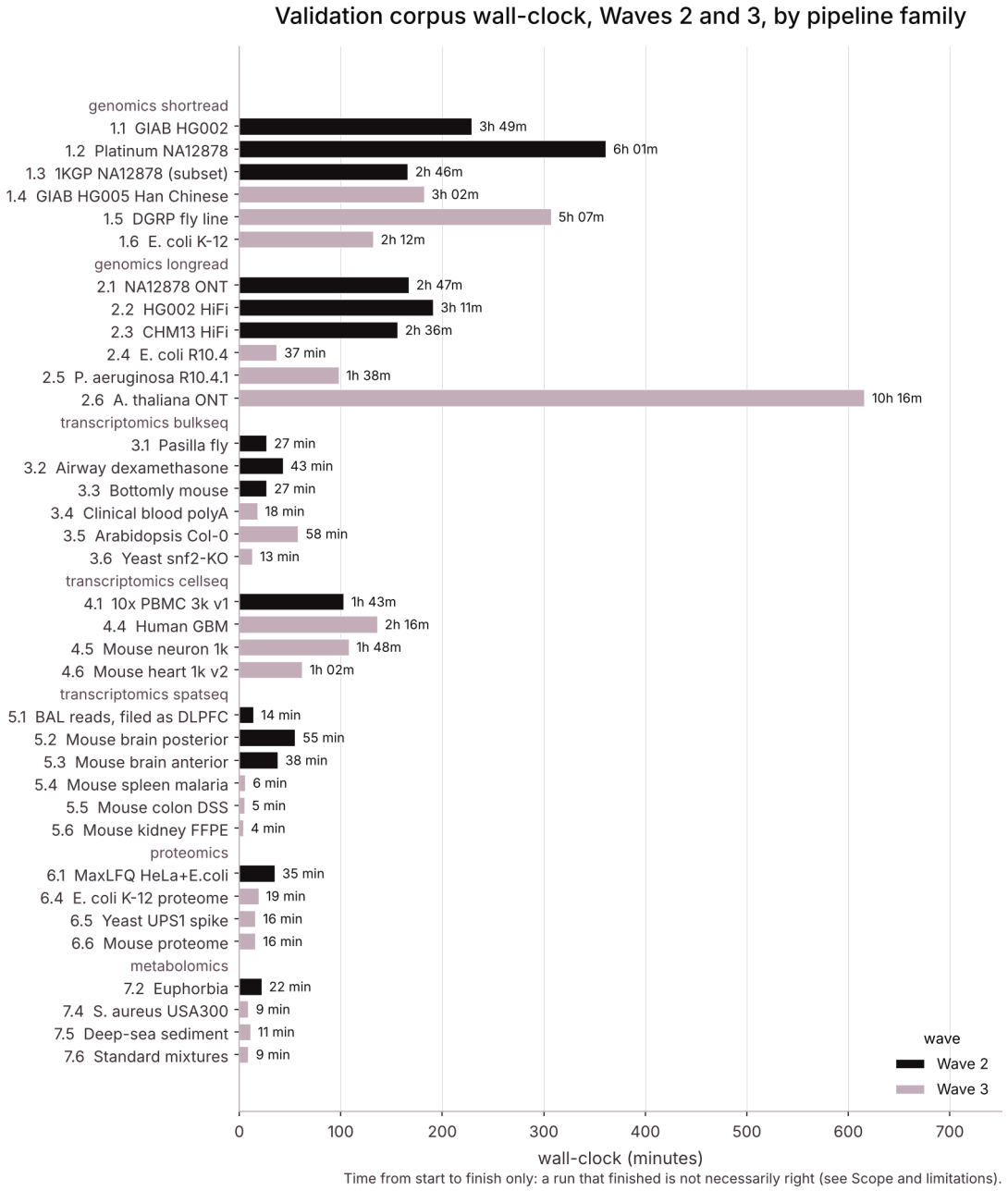



**Figure 2.** Wall-clock for the paper-shaped runs that carry a recorded duration, grouped by pipeline and colored by wave. Genomics variant calling occupies the upper range; quantification pipelines (proteomics, metabolomics, spatial intake) finish in minutes. The point is that each pipeline runs a public dataset end to end in a predictable window; a run that finished is not necessarily right.

The Wave-3 headline outputs span the modalities: variant catalogs from a few hundred calls on a bacterial genome to nearly 1.5 million on one sequencing run of a human genome (HG005, 23,808,174 reads); single-cell maps of 437 to 3,454 cells resolved into 10 to 12 clusters; proteomes of 4,143 to 15,463 proteins; and metabolite tables of tens of thousands of features. The proteome counts come from unfiltered lists. Two of the three Wave-3 proteomics runs (the yeast-plus-UPS1 standard and the mouse proteome) and two of the three Wave-3 metabolomics runs (the S. aureus supernatants and the standard mixtures) carry a confirmed silent failure, and the third of each is a candidate: the E. coli proteome, whose summary medians stay under the threshold although its PSM table shows the unfiltered list, and the deep-sea sediment, whose summary shows the same near-absent matching but whose four sections are not known beforehand to share most of their features (see Scope and limitations). Each of these numbers is read from the run's published summary, the same artifact the assistant answers from.

# 5 Genomics parity gates

The short-read pipeline carries a parity gate: a re-run of the HG002 anchor is compared against a prior baseline on eight summary metrics. A Wave-2 CPU re-run matched the baseline to the digit on every gate. A later Wave-2 run of the same sample, the one the validation's ledger records, reports different variant and annotation counts and is not covered by this comparison.

| Metric | Baseline | Re-run | Delta |
|---|---|---|---|
| Mean mapping rate (percent) | 99.68 | 99.68 | 0.00 |
| Variants | 695,316 | 695,316 | 0 |
| Single-nucleotide variants | 608,123 | 608,123 | 0 |
| Indels | 87,229 | 87,229 | 0 |
| High-impact annotations | 257 | 257 | 0 |
| Moderate-impact annotations | 4,581 | 4,581 | 0 |
| Low-impact annotations | 7,594 | 7,594 | 0 |
| Modifier-impact annotations | 2,598,167 | 2,598,167 | 0 |

This gate shows that the CPU code path (alignment, variant calling, and functional annotation) reproduces the prior baseline on one human sample, not that its calls are correct: neither run's calls were compared with a truth set, and a wrong result reproduces as well as a right one: the Wave-1 and Wave-2 runs of the HeLa-plus-E. coli benchmark, days apart, give the same per-sample PSM counts and the same median q-values to two decimals, and both carry the unfiltered list set out under Scope and limitations. The re-run's summary carries no ClinVar annotation, so the gate does not cover clinical-significance enrichment, and it does not cover the GPU code path, which we cover in the limitations.

# 6 The question battery and the two graders

We grade answers against a battery of 230 questions, parsed per paper and keyed to each run's expected values. Two graders apply the same rules from different angles. The live-API grader sends each question to the deployed assistant and checks the answer; the summary-direct grader applies the same numeric tolerance and substring rules to the flattened summary, which isolates whether the data is present from whether the assistant carries it through. A numeric question passes when the first number in its expected answer lies within 10 percent (at least 0.1) of any number in the answer, which for the summary-direct grader means any number in the summary; a text question requires every expected token.

On the summary-direct grader's final pass over all 21 Wave-2 runs, 87 of the 98 questions it could grade pass (88.8 percent), and 132 of the 230 were left ungraded. Graded live on the nine runs then complete, the deployed assistant passed 56 of 62 (90.3 percent). These rates do not show that the runs were right. Many expected answers were read from the summaries themselves (54 cite the summary as their source, and 73 more defer to it), so a wrong summary sets a wrong answer key; for 22 of the 87 passes, the only numbers within tolerance sat inside text such as a gene name, a run identifier, a timestamp or a version string; and the false-discovery-rate question passed on all three Wave-2 proteomics runs, whose filter step aborted on an option OpenMS 3.5.0 rejects, so that the fallback passed each list on unfiltered. The Wave-1 check of April 2026, whose record is a report the deployment session wrote, scored 28 of its 30 pipeline questions, where its brief reported 30 of 30, and two of the answers it passed repeat failures set out under Scope and limitations: one called the proteomics run's high median q-values expected, and one reported a metabolomics unique-feature count equal to the sum of the samples' counts.

# 7 RAGAS scoring across the corpus

Alongside the mechanical battery, we score answers with RAGAS, an evaluation in which a judge model rates each answer on faithfulness (does the answer stay grounded in the provided context), answer relevancy (does it address the question), and context precision (is the provided context relevant to the question; the Wave-3 prompt defines it differently, as whether the answer could be drawn from the context). The judge is a Vertex Gemini model: in Wave 2, Gemini 2.5 Pro running the RAGAS library; in Wave 3, Gemini 2.5 Flash, which scored three metrics named after RAGAS's in a single prompt that defines them. In Waves 2 and 3 the answers came from a direct Gemini 2.5 Pro call grounded in each run summary rather than from the deployed assistant, so the Wave-2 judge was the same model that wrote the answers and the Wave-3 judge a model of the same family, and the Wave-3 questions were written from the same summaries, so these scores measure faithfulness to the summary, not the deployed assistant and not whether the summary is right. The table gives RAGAS scores for all three validation waves: the 30-question Wave-1 baseline, the 98-question Wave-2 closure, and a 231-question Wave-3 pass spanning all 21 new papers and 11 questions each. The Wave-1 scores are those of the first brief; the record of that check says they came from a one-shot scoring pass that is not kept in the repository.

| Metric | Wave 1 brief (30 questions, 28 answered) | Wave 2 (n=98) | Wave 3 (n=231) |
|---|---|---|---|
| Faithfulness | 1.00 | 0.83 | 0.99 |
| Answer relevancy | 1.00 | 0.63 | 0.99 |
| Context precision | 0.68 | 0.81 | 1.00 |

The Wave-1 column gives the first brief's scores, which it reports for 30 questions. In the record of that check, two of those questions returned server errors (HTTP 500), and the six single-cell questions, asked again of a repeated run, were not rescored, so the column does not describe 30 scored answers.

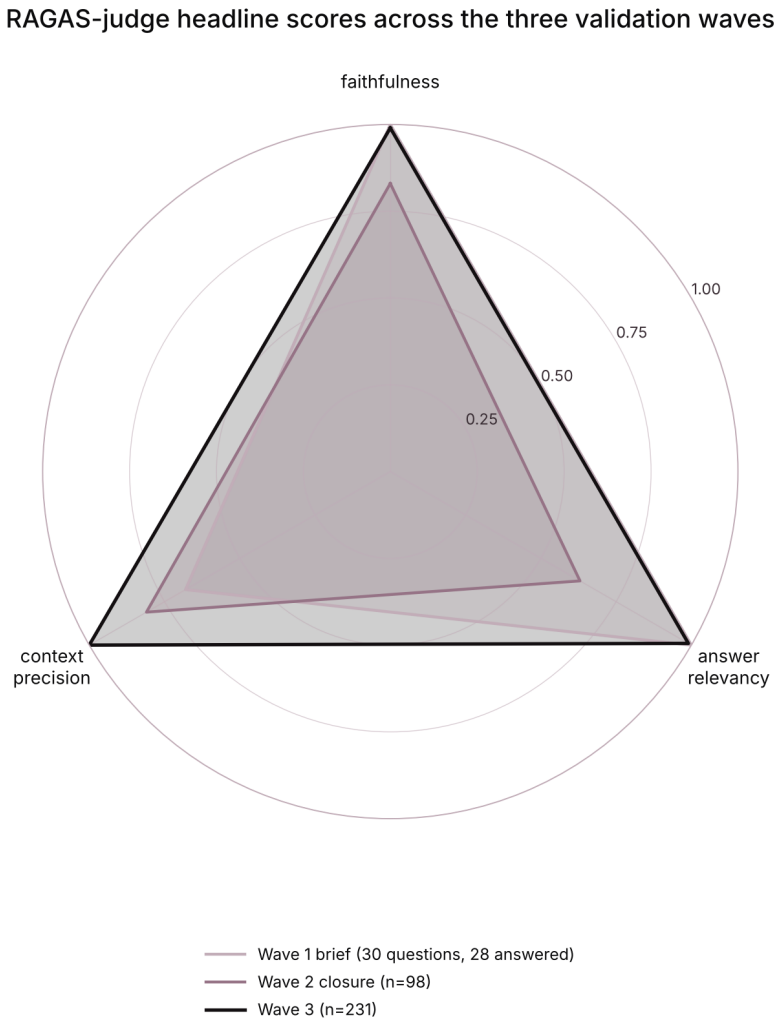



**Figure 3.** RAGAS headline scores across the three waves on the three shared metric names. The Wave-1 series repeats the first brief's scores, with the limits given under the table. Context precision reaches 1.00 in Wave 3, where the judge's prompt defines it as whether the answer could be drawn from the context and every expected answer other than the refusal probes was taken from the summary. Faithfulness measures agreement with the summaries, not whether they are right.

On the Wave-3 corpus, 17 of the 21 papers score at the ceiling on all three metrics, and the refusal

probes all hold: when a question asks for a value the summary does not carry, the answering model says that it is out of scope rather than inventing one. The four departures from the ceiling, on two proteomics and two metabolomics papers, each come from one question, which asked what framework wrapped the workflow: the answer key expected OpenMS and the answers named Nextflow. The ceiling scores include failed runs: five of the eight Wave-3 runs with a confirmed silent failure (see Scope and limitations) averaged the maximum score on every metric, and the answers scored at the maximum include ones that repeat a failure's signature, such as the Arabidopsis run's 0.11 percent mapping rate.

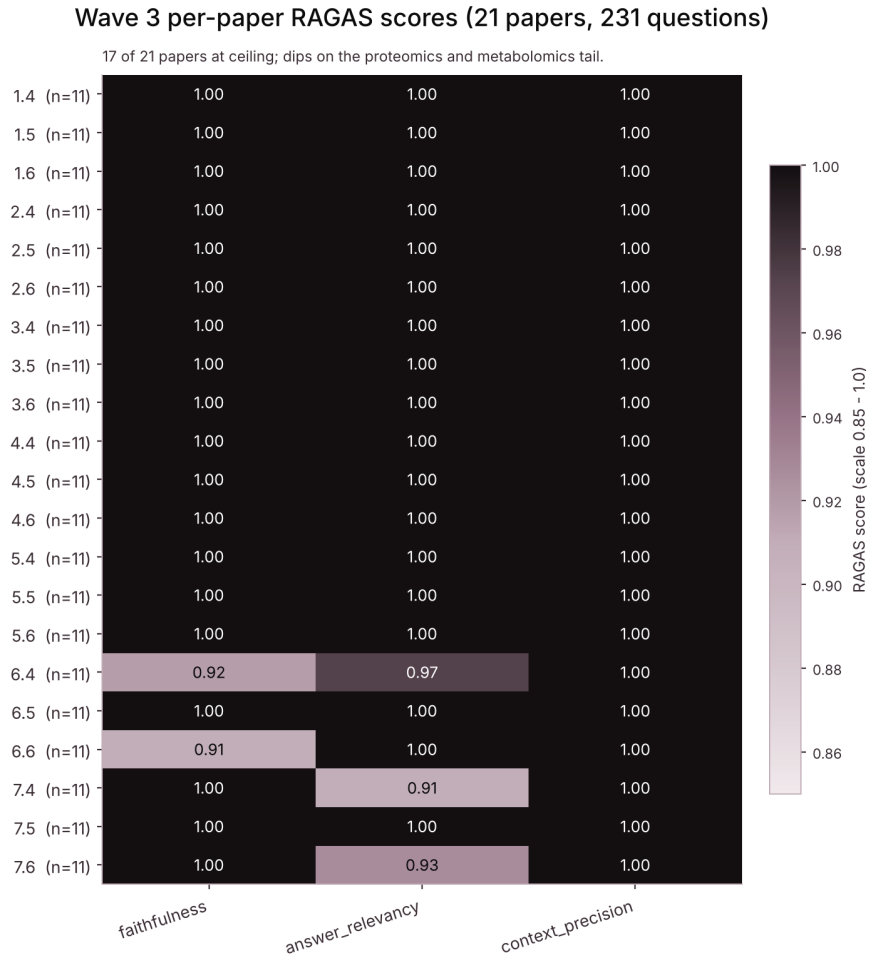



**Figure 4.** Per-paper Wave-3 RAGAS scores, on a 0.85 to 1.0 scale so the departures from the ceiling are visible. Most papers sit at the ceiling on all three metrics; each of the four departures comes from one question about the workflow framework, where the answer key expected OpenMS and the answers named Nextflow.

Two honest caveats frame these numbers. First, the three waves were scored by three judge setups that evolved with the platform, so the comparison across waves is directional rather than controlled; the metric names are shared, but the Wave-3 judge defines the metrics in its own prompt. Second, RAGAS saturates on tightly-grounded extractive question answering, which is exactly what the summary contract produces, so high scores confirm that the answers stay faithful to the summary more than they discriminate between strong and weak answers. Neither these scores nor the mechanical battery tests whether a summary is right: faithfulness to a summary is not correctness, and the failures set out under Scope and limitations passed both. A judge from another model family would address self-preference,

not this limit.

# 8 Scientific signal surfaced during validation

The audit labels nine of the runs controls, clean by its definition rather than shown correct, and leaves six unlabeled, and its provenance, identity and invariant checks flag none of them. None of their results was compared with a truth set, and five of the six unlabeled runs are spatial runs, in whose summaries those checks find nothing to read. Read with that limit, the runs span the modalities. The human genomes yield variant catalogs (the run summaries record no transition-to-transversion ratio), and an earlier run of the short-read anchor matches its baseline on eight summary metrics. The long-read additions run from 5 Mb bacterial genomes to a 135 Mb plant genome. The single-cell additions cluster a solid tumor and two mouse tissues, the first real 10x v2 chemistry among them, though no run checks its barcodes against a whitelist. The proteomics additions move from human cell lines and tissue to a bacterial proteome and a mouse draft proteome, with unfiltered identification counts (see Scope and limitations). The metabolomics additions move from human and plant matrices to microbial culture supernatants, a deep-sea sediment and two compound-standard mixtures, with tens of thousands of features each, not matched across samples (see Scope and limitations). The assistant is designed to answer questions about each of these from the published summary, so that the science is legible to a non-bioinformatician at the bench; this validation put questions to the deployed assistant on nine Wave-2 runs only.

# 9 Paths the wider corpus exercised

Selecting Wave 3 for the gaps in the first corpus exercised paths the anchors never reached. Single-cell ingestion reads native multi-lane 10x layouts directly, merging all lanes of a library into one alignment, which is how 10x mkfastq output arrives. Single-cell chemistry covers 10x v2 as well as v3, in a run that completed on a real v2 dataset; the run’s valid-barcode fraction of 1.0 does not show that the v2 barcodes were read correctly, because the pipeline uses no barcode whitelist. Long-read variant calling on vertebrate and plant-scale genomes scatters across reference windows so the work fits preemptible workers and completes without serializing into a single multi-hour task. Each of these paths ran to completion on a dataset that required it; completion does not show that its results are right.

# 10 Scope and limitations

We hold the corpus to an honest account of what it does and does not establish.

The GPU variant-calling path is wired and fires at runtime, and individual GPU stages succeed, but the path has not been validated: no GPU run has been checked against the eight parity gates. Some runs used GPU stages, including the Wave-2 1000 Genomes NA12878 run and the Wave-3 Drosophila line (recorded as DGRP-802), and their numbers are not covered by the CPU-path parity gate.

Two Wave-3 bulk RNA-seq runs, Arabidopsis Col-0 and yeast snf2, completed end to end but mapped almost no reads (0.11 and 0.009 percent), and their top-quantified transcripts resolved to human identifiers. The cause is a profile default: the pipeline's cloud profiles set the Salmon index to the human GRCh38 / GENCODE v46 index, and a set index takes precedence over the configured FASTA and GTF, so both runs were quantified against human transcripts while their summaries recorded the configured plant and yeast references. We report it as a defect to correct, not as a healthy result; the quantification for these two runs is not usable. The Wave-2 pasilla anchor fell through to the same default, so its quantification is not usable either.

The two non-human short-read additions, Drosophila and E. coli, completed variant calling with plausible mapping (90.67 and 99.46 percent) but annotated functional impact against the default human gene model (SnpEff GRCh38.99), so their impact categories are not biologically meaningful; for E. coli every impact count is zero. The calls themselves were not compared against a truth set, and the E. coli genome was genotyped without a haploid ploidy setting. Organism-matched annotation databases and ploidy settings would correct this, and the live platform has neither. The long-read caller has the opposite problem: it runs with a fixed haploid model, which suits the bacterial genomes, the essentially homozygous CHM13 cell line and the inbred Arabidopsis accession but not the diploid NA12878 and HG002 genomes, whose variant counts come from that model; how much it changed their calls has not been measured.

Two further limitations came to light after publication. In proteomics, the false-discovery filter step aborted on an option that OpenMS renamed in version 3.2.0 and the pipeline's version (3.5.0) rejects, and the step's fallback passed the identifications on unfiltered. Every proteomics run in the corpus was left unfiltered, as its PSM tables show, so no protein, peptide or PSM count in this brief is a 1 percent FDR result. Four runs show it in their per-sample median PSM q-values, against a 1 percent threshold: the Wave-3 yeast-plus-UPS1 standard (0.575 and 0.632) and mouse proteome (0.147 and 0.508), whose protein counts are 15,463 and 13,278, and the Wave-2 HeLa-plus-E. coli benchmark (0.98, 0.90 and 0.98) and PXD000584 set (0.70 and 0.58). The searches of these four runs also identified few PSMs at 1 percent FDR: 0.2 to 27.4 percent of each run's PSMs have q-values at or below 0.01, and in the HeLa-plus-E. coli benchmark 25 of its 12,648 PSMs across three samples do; the cause has not been tested. In the other two, the Wave-2 PXD002395 run and the Wave-3 E. coli proteome, the medians stay under 0.01 because most matches score well, but 19.2 and 20.6 percent, and 48.5 and 49.7 percent, of their matches carry q-values above it. In metabolomics, features are matched across samples on exact rounded m/z and retention-time keys, which run-to-run measurement jitter rarely reproduces: in the Wave-2 Euphorbia run the 8,417 unique features equal the sum of the three per-sample counts, so no feature was matched across samples; in the three Wave-3 runs the unique count is 98.6 to 99.96 percent of the per-sample total; and in the Wave-2 urine run it is 19,768 of 19,820. In the four replicate injections of the NIST plasma standard (Wave 2), the unique count is 99.62 percent of the per-sample total, where identical injections would give 25 percent. Yet a median 47 percent of each injection's features have a partner in another injection within 5 ppm in m/z and 0.2 minutes

in retention time, against 0.6 percent when the m/z window is moved 25 ppm away, which estimates the partners found by chance. The reported feature totals are therefore close to per-sample sums, not aligned feature tables.

The Wave-3 spatial runs exercise the read-intake path only, whose barcode check cannot fail on reads that are not Visium data (see Spatial transcriptomics); per-spot spatial quantification is out of scope for that mode. The single-cell runs use no barcode whitelist either, so every barcode counts as valid and the valid-barcode fraction, 1.0 in every sample, says nothing about whether a run worked. Several Wave-3 read sets (the clinical blood and the healing-colon spatial sets among them) are fixed-prefix subsets of larger files, which is sufficient for pipeline-shape validation but not for composition or coverage-uniformity claims. And the single-cell glioblastoma run is a subset of a much larger tar, so its absolute cell yield is bounded by the subset rather than the full sample.

None of these defects is fixed on the live platform. A guard against wrong references for bulk RNA-seq and short-read runs was merged on 2026-06-15 and has not been deployed: the API revision serving on 2026-10-09 ran an image built on 2026-06-10, before the guard. The platform’s code defaults to the human Salmon index and SnpEff database, makes the same false-discovery filter call, matches metabolomics features on the same exact keys, calls long-read variants with the same haploid model and reads single-cell and spatial barcodes against no barcode list.

# 11 Corrections

Revised 2026-10-06. Four reference PMIDs pointed to unrelated papers and now point to the cited works (Eberle 2017, Jain 2018, Wenger 2019, Brooks 2011). Titles or first authors of eleven further references (9, 11, 12, 15, 17, 19, 20, 23, 24, 27, 28) were corrected against PubMed. The question-battery paragraph now attributes 66 of 69 to the summary-direct grader, where it was measured, and reports the live-API result; the RAGAS section states that its answers did not come from the deployed assistant. The statements on schema validity, the CPU code path, and the cause of the bulk RNA-seq reference mismatch are corrected, and a limitation on proteomics FDR filtering and metabolomics feature matching is added.

Revised 2026-10-07, from the runs’ archived summaries and the deposits’ own metadata. The Wave-2 pasilla run is identified as quantified against the human index. The polyA-versus-rRNA-depletion contrast is no longer called the expected split; its direction is opposite to the cited study. The HG002 HiFi set is genome-wide, not chromosome 20 only. The PXD000584 set is a gastric-cancer biopsy, not a HeLa proteome. The Dunphy 2021 files are S. aureus monoculture replicates, not a co-culture. The transition-to-transversion claim is removed because no run summary records the ratio. The FDR and feature-matching limitation now covers the Wave-2 runs that show the same pattern.

Revised 2026-10-17, from the runs’ identification tables and the Sequence Read Archive. The proteomics false-discovery filter step aborted on every proteomics run, on an option OpenMS 3.5.0 rejects, and the fallback passed the list on unfiltered, so the E. coli protein count is no longer given as a 1 percent FDR result and the limitation covers every proteomics run. The Wave-2 run filed under the dorsolateral prefrontal cortex Visium reference processed bronchoalveolar-lavage reads from another study (Chen 2020) and no longer counts as validating that reference.

Revised 2026-10-18, against the manuscript that audits these runs (Runs that succeed and are wrong: an audit of silent failures in a multi-omics platform’s validation runs) and its table of datasets. Sources: the 30x NA12878 reads are credited to Byrska-Bishop 2022, which produced them (reference 3

describes the earlier low-coverage release); the HG002 HiFi set to a Genome in a Bottle data release rather than Wenger 2019 (reference 5); the CHM13 set to Vollger 2020, as one SMRT Cell rather than a subset; the CFT073 reads to Sanderson 2023 (R10.4) rather than Sanderson 2024; the Euphorbia data to Nothias 2018, which Nothias 2020 (reference 25) reanalysed; and the PBMC 3k set to 10x Genomics as a public dataset. The DGRP line is described as recorded, with its source paper unconfirmed; the PXD002395 run is two unrecorded raw files of an eleven-cell-line deposit, not a confirmed HeLa proteome; and the MTBLS733 run is two standard mixtures of the same 1,100 compounds, not a black-pepper study. References 30 to 38 are added, three of them (34 to 36) for works the text already cited.

Results and methods, same revision: the subtitle and abstract no longer call the datasets reproduced or all published; three non-human bulk RNA-seq runs, not two, were quantified against the human index; the reversed library-prep contrast is traced to labels that appear swapped in the deposit; 20 of the 21 Wave-3 summaries, not only the proteomics ones, fail their schemas; the battery figures come from the grader's final pass over all 21 Wave-2 runs (87 of 98 graded questions), not a ten-run pass, and the battery and the judges are said to measure agreement with the summaries, not correctness; the Wave-3 judge and its four departures from the ceiling are described; the parity gate drops its ClinVar claim and notes the later ledger run; the spatial intake check and the single-cell barcode fraction are described as unable to fail; the haploid long-read caller on diploid genomes, the few PSMs passing 1 percent FDR and the unmatched metabolomics totals are added; the headline ranges are marked Wave 3's. Two figures are redrawn: the coverage grid drops a short-read mouse dataset the corpus lacks and adds the Wave-3 human single-cell and E. coli proteomics data, and the wall-clock chart relabels the HG002 and CHM13 HiFi sets, the run filed as DLPFC, the S. aureus monocultures and the standard mixtures.

Revised 2026-10-19, against the manuscript's matching of the pulled feature tables within tolerances. The NIST plasma injections are no longer said to share at most 0.38 percent of their features: that figure describes the matrix's exact keys, and within 5 ppm and 0.2 minutes a median 47 percent of each injection's features have a partner in another injection.

Revised 2026-11-03, against the final text of the audit (Palanikumar 2026, reference 39, https://doi.org/10.5281/zenodo.23717230). The abstract gives its result, a confirmed silent failure in 14 of the 42 runs, in five classes. A published summary is no longer said to validate a run, and the runs without a failure are described as the audit's controls, clean by its definition rather than shown correct, and its unlabeled runs. The HG002 parity gate is described as an earlier run's match on eight summary metrics, which shows that the CPU path reproduces, not that its calls are correct. The Wave-1 check is reported at the 28 of 30 its record scores, not as graded by hand at 30 of 30, and its RAGAS column as having no record in the repository and as including two questions that returned server errors and six that were not rescored. The proteomics filter step is described as aborting on an option OpenMS 3.5.0 rejects, with a fallback that passed the list on unfiltered, and the Wave-3 headline outputs name the proteomics and metabolomics runs with a confirmed silent failure and the candidates. The pasilla anchor is described as untreated libraries with no knockdown library; the DGRP and E. coli variant counts point to their annotation with human data; the Wave-2 urine run's feature counts are added; the section on paths the wider corpus exercised says that they ran to completion, not that they are established; the wall-clock statement covers only the 42 kept runs; the assistant is described as meant to read its numbers from the summary, and the GPU path as not validated; and the limitations say that none of the defects is fixed on the live platform. The title and the corpus heading count runs rather than papers, and the title page gives the revision date. The earlier notes keep the wording of their dates, except for two changes: the word that named the audit's status in the 2026-10-18 and 2026-10-19 notes now reads "manuscript", and the 2026-10-17 note describes the filter step as this revision does.

Citations updated 2026-11-09. The audit is posted as a preprint, not peer reviewed (Preprints.org, 3 November 2026, version 1), and the abstract and reference 39 give its DOI. A final review on the same day corrected the grounds on which the sediment run is a candidate, the descriptions of the HG005, 1000 Genomes and sediment data, the account of the RAGAS judges and of whose answers were graded, and the Wave-1 labels of the RAGAS table and chart; the wall-clock chart relabels two runs. The link in the 2026-11-03 note leads to the archived sources, not to the preprint.

# 12 References


1. Zook JM, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016. PMID 27271295.
2. Eberle MA, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 2017. PMID 27903644.
3. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015. PMID 26432245.
4. Jain M, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018. PMID 29431738.
5. Wenger AM, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019. PMID 31406327.
6. Nurk S, et al. The complete sequence of a human genome. Science. 2022. PMID 35357919.
7. Mackay TFC, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012. PMID 22318601.
8. Brooks AN, et al. Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. 2011. PMID 20921232.
9. Himes BE, et al. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. PLoS One. 2014. PMID 24926665.
10. Bottomly D, et al. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One. 2011. PMID 21455293.
11. Zhao S, et al. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci Rep. 2018. PMID 29556074.
12. Froussios K, et al. How well do RNA-Seq differential gene expression tools perform in a complex eukaryote? A case study in Arabidopsis thaliana. Bioinformatics. 2019. PMID 30726870.
13. Gierlinski M, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015. PMID 26206307.
14. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017. PMID 28091601.
15. Derbois C, et al. Single cell transcriptome sequencing of stimulated and frozen human peripheral blood mononuclear cells. Sci Data. 2023. PMID 37414801.
16. Maynard KR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021. PMID 33558695.
17. Du MRM, et al. Benchmarking spatial transcriptomics technologies with the multi-sample SpatialBenchVisium dataset. Genome Biol. 2025. PMID 40156041.
18. Parigi SM, et al. The spatial transcriptomic landscape of the healing mouse intestine following damage. Nat Commun. 2022. PMID 35149721.

19. Xuanyuan Q, et al. Multimodal spatial transcriptomic characterization of mouse kidney injury and repair. Nat Commun. 2025. PMID 40813851.
20. Cox J, et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014. PMID 24942700.
21. Schmidt A, et al. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016. PMID 26641532.
22. Ramus C, et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. Data Brief. 2016. PMID 26862574.
23. Giansanti P, et al. Mass spectrometry-based draft of the mouse proteome. Nat Methods. 2022. PMID 35710609.
24. Thevenot EA, et al. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. J Proteome Res. 2015. PMID 26088811.
25. Nothias LF, et al. Feature-based molecular networking in the GNPS analysis environment. Nat Methods. 2020. PMID 32839597.
26. Stancliffe E, et al. DecoID improves identification rates in metabolomics through database-assisted MS/MS deconvolution. Nat Methods. 2021. PMID 34239103.
27. Dunphy LJ, et al. Untargeted metabolomics reveals species-specific metabolite production and shared nutrient consumption by Pseudomonas aeruginosa and Staphylococcus aureus. mSystems. 2021. PMID 34156287.
28. Li Z, et al. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. Anal Chim Acta. 2018. PMID 29907290.
29. Chen L, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020. PMID 32020836.
30. Byrska-Bishop M, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022. PMID 36055201.
31. Huang W, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 2014. PMID 24714809.
32. Vollger MR, et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann Hum Genet. 2020. PMID 31711268.
33. Sanderson ND, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genom. 2023. PMID 36748454.
34. Sanderson ND, et al. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. Microb Genom. 2024. PMID 38713194.
35. Michael TP, et al. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 2018. PMID 29416032.
36. Aquino PF, et al. Exploring the proteomic landscape of a gastric cancer biopsy with the shotgun imaging analyzer. J Proteome Res. 2014. PMID 24283986.
37. Nothias LF, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. J Nat Prod. 2018. PMID 29498278.
38. Redick MA, et al. Integration of untargeted metabolomics and microbial community analyses to characterize distinct deep-sea methane seeps. Front Mar Sci. 2023. PMID 39268414.
39. Palanikumar N. Runs that succeed and are wrong: an audit of silent failures in a multi-omics platform's validation runs. Preprints.org [preprint; not peer reviewed]. 2026. doi:10.20944/preprints202611.0178.v1. Code, data and manuscript sources: https://doi.org/10.5281/zenodo.23717230.