Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

glycoRIP: a dedicated pipeline for statistical differential analysis of glycoRNA sequencing data

Wei Li

lw@cloud-seq.com.cn

cloud-seq Biotech Inc., Shanghai https://orcid.org/0009-0002-6525-6413

Method Article

Keywords: glycoRNA, small RNA, enrichment sequencing, dsRIP-Seq, reproducible workflow

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11864092/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: W.L. is an employee of Cloud-seq Biotech Co., Ltd. (Shanghai, China), a company that provides glycoRNA and other RNA sequencing services on a commercial basis and that licenses the glycoRIP pipeline described in this preprint for commercial use. All benchmark data analysed here are public (GSE136967; rPAL SRR35543218/SRR35543220), and all code, scripts and raw measurement outputs are openly available; no proprietary customer data were used. The author declares no other competing interests.

Abstract

Background

GlycoRNAs are small RNAs bearing N-glycans displayed on the cell surface. They are studied by metabolic labelling, enrichment and small-RNA sequencing, but the analysis of the resulting libraries remains ad hoc and hard to reproduce: published studies release project-specific code rather than installable software, curated resources apply per-dataset enrichment tests, and small-RNA quantifiers resolve multi-mapping without an enriched-versus-input contrast. No packaged implementation currently casts glycoRNA enrichment as a formal differential test with explicit handling of the cases in which that test is not identifiable.

Results

We present glycoRIP, a source-available Nextflow pipeline that turns enriched-channel FASTQ files (immunoprecipitation, IP; input, INPUT; optional mock controls) into an auditable candidate table and a self-contained HTML report in one command. It combines reference-free adapter discovery, small-RNA-aware alignment against a compact tRNA-plus-transcript index, streaming expectation-maximisation (EM) assignment of multi-mapping reads, and a single edgeR quasi-likelihood GLM implementing the three dsRIP-Seq enrichment criteria, degrading explicitly to descriptive statistics when replication is insufficient. We benchmarked the pipeline exclusively on public data. Against the Flynn 2021 gold standard, 83.5% (106/127) of mappable known positives in HeLa and 81.5% (101/124) in H9 showed concordant enrichment direction, versus null expectations of 57.9% and 54.4% under expression-matched permutation (10,000 draws, empirical $p = 1 \times 10^{-4}$); on synthetic data with known truth, sensitivity and specificity were both 1.000. In rPAL data, which use a different chemistry, 52.9% (55/104) of the same positives were enriched, above the 34.4% null but below the same-chemistry value. On GSE136967 the pipeline runs within 1,334 (HeLa) and 2,068 (H9) 8-core CPU-seconds per sample, peaking at 1.5/1.7 GB; a high-complexity rPAL library fits the memory target but remains 1.8-fold over the runtime target.

Conclusions

glycoRIP provides a reproducible, auditable path from raw glycoRNA sequencing reads to ranked candidate features, with documented statistical assumptions and explicit failure modes. It is distributed with Docker and Singularity definitions, a one-command self-test and reference-build scripts, and is free for academic and non-commercial use. The benchmark is deliberately reported as it stands, including where the pipeline falls short of its own performance target.

Background

Small RNAs bearing N-glycans are displayed on the surface of living cells [1], and their glycan content can now be profiled by labelling, enrichment and sequencing. Current workflows metabolically label sialic acids (for example with Ac_4ManNAz), attach an affinity handle by click chemistry, enrich the modified species, and sequence the recovered small RNAs alongside an unenriched input. Related chemistries, such as rPAL, broaden the accessible glycan classes [2]. The computational task is similar across these variants: given an enriched channel and a reference channel, decide which RNA features are enriched rather than merely abundant, multi-mapping, or poorly modelled.

The statistical foundation for this question was established for dsRIP-Seq, which formalised three criteria: enrichment in IP relative to INPUT, differential interaction between conditions, and a third, non-test flag that retains features whose change between conditions is larger in absolute value in the IP channel than in the input channel, so that apparent enrichment driven by a larger input-side expression change is down-weighted [3]. Those criteria have never been carried into a reusable implementation. Published glycoRNA studies release their data—the Flynn 2021 libraries are public [1]—but the accompanying analysis is project-specific code rather than an installable tool; curated resources such as GlycoRNAdb compile enriched sequences while applying per-dataset enrichment tests [4]; and quantifiers built for tRNAs resolve multi-mapping without an IP/INPUT contrast [5]. No open, end-to-end pipeline connects raw reads to an auditable candidate table, and the obstacles are specific rather than generic: adapter layouts vary, multi-mapping among near-identical tRNAs and small ncRNA families is the norm, modified bases violate aligner assumptions, and enrichment data violate the normalisation assumptions of standard RNA-seq tools (Table 1).

We therefore built glycoRIP as a reproducible workflow; its statistical core is inherited from dsRIP-Seq. To our knowledge, no packaged, containerised pipeline dedicated to statistical differential analysis of glycoRNA sequencing data currently exists, and no released implementation casts glycoRNA enrichment as a formal three-criterion differential test rather than an ad hoc fold-change threshold. The contributions are: (i) a single installable pipeline that takes enriched and input FASTQ files to an auditable candidate table and a self-contained report; (ii) a small-RNA-first quantification path that keeps multi-mapping reads and resolves them by streaming, signature-level EM with byte-reproducible outputs; (iii) a single GLM implementation of the dsRIP-Seq criteria with explicit, automatic degradation to descriptive statistics when biological replication is insufficient; (iv) a reference-building procedure and container images that make the entire analysis reproducible from public sources; and (v) a public-data benchmark that reports what the pipeline reproduces and where it falls short.

Table 1
Prior implementations for glycoRNA and related enrichment data.

Implementation	Type	Code released	Three criteria	Multi-mapping model	Installable	glycoRNA
Flynn 2021 [1]	Discovery study	No†	Partial (IP vs INPUT)	n.r.	No	Yes
dsRIP-Seq [3]	Protocol	No	**Defined**	No	No	No (dsRNA)
rPAL-seq [2]	Chemistry + companion scripts	Yes (GPL-3.0)‡	No (DESeq2 enrichment)	Yes (UMI-aware EM)	No (loose scripts)	Yes
GlycoRNAdb [4]	Database	No (reuses published pipelines)	No (per-dataset test)	No	No (web only)	Yes
mim-tRNAseq [5]	tRNA quantifier	Yes	No	Yes	Yes	No
Salmon / RSEM / kallisto [6–8]	Generic EM quantifiers	Yes	No	Yes	Yes	No
glycoRIP (this work)	Dedicated pipeline	**Yes (PolyForm NC)**	**Yes**	**Yes (signature EM)**	**Yes**	**Yes**

Capability entries were determined from each publication's Methods and Data/Code availability statements, checked 19–20 November 2026. n.r., not reported in the publication or its availability statement. † No analysis repository was located; data are public in GEO (GSE136967). ‡ The released repository was inspected on 20 November 2026 (FlynnLab/rPAL-seq, GPL-3.0-or-later; Zenodo concept DOI 10.5281/zenodo.21969177): workflow and analysis scripts including a UMI-aware EM module for multi-mapping assignment, distributed as loose scripts without packaging, containers or a workflow engine.

Implementation

Input, sample sheet and workflow architecture

The pipeline accepts single- or paired-end FASTQ files and a five-column sample sheet: sample_id, fastq, fraction (IP, INPUT, or mock), condition, and replicate_id. Mock channels are detected automatically and switch the report to capture-specific wording. When paired IP–INPUT samples share a replicate_id, the design blocks on the pair.

glycoRIP is implemented in Nextflow (DSL2) [9] and consists of eight processes (Fig. 1). **P0** discovers the 3' adapter from soft-clipped bases in a subsample of reads, without a reference. **P1** trims adapters and low-quality ends with cutadapt [10] and, when a UMI is present, extracts it with umi_tools [11]; three UMI modes are supported (umi, none, collapse_identical). **P2** aligns reads with bowtie2 [12] in local mode against a compact index built from mature tRNAs and RefSeq transcripts, retaining up to 20 alignments per read. **P3** parses the SAM stream and assigns multi-mapping reads by EM. **P4** merges per-sample counts. **P5** runs the edgeR quasi-likelihood GLM. **P6** annotates features and matches curated known glycoRNAs. **P7** renders a self-contained HTML report with inline SVG figures and control-aware wording.

Two design choices distinguish the workflow from a generic small-RNA pipeline. First, alignment is not followed by coordinate sorting: sorting separates the alignments of one read in the file, which turns a multi-mapping read into several apparently independent single-mapping reads. Second, EM operates on *signatures*—reads sharing an identical set of alignment coordinates are aggregated before iteration—which is mathematically equivalent to per-read EM but bounds memory by the number of distinct signatures rather than by read depth.

Adapter discovery, trimming and alignment

The 3' adapter was inferred from soft-clipped bases in a subsample of reads, by consensus over the first 100,000 reads, without a reference. Reads were trimmed with cutadapt 4.9 [10] (minimum length 15 nt). Where a UMI was declared, the eight-base barcode was moved into the read name by umi_tools 1.1.5 [11] before alignment, and duplicates were collapsed during EM on the key (full alignment-coordinate set, UMI). In collapse_identical mode no UMI is required and duplicates are keyed on (coordinate set, read sequence); with none no deduplication is applied. Keying on the coordinate set rather than on bowtie2's randomly chosen primary alignment keeps tie-breaking from splitting one molecule into several.

Reads were aligned with bowtie2 2.5.4 [12] in --local mode, -k 20, 16 threads per task. Local mode tolerates the mismatches produced by modified bases. The SAM stream was consumed directly; coordinate-sorted input was rejected because sorting destroys the grouping of a read's alignments.

Reference construction and multi-mapping quantification

The human reference contains 201,812 sequences: 260 collapsed mature tRNAs from GtRNAdb (3' CCA added after collapsing identical sequences) and 201,552 RefSeq transcripts (mRNA, ncRNA, misc_RNA, rRNA and precursor RNA) [13, 14]. The index is approximately 1.2 GB and is rebuilt by a single script from public sources. Keeping the full transcript set, rather than pruning to short ncRNAs, preserves the expression background required by the third criterion (change dominance between the IP and input channels) and avoids discarding long transcripts that can carry captured small RNAs; a pruned ≤ 500 nt index is available as an optional preset but is not the default.

Alignments are produced with --local -k 20, which tolerates the mismatches introduced by modified bases [15] and retains alternative loci. Multi-mapping reads are distributed by EM over their alignment sets, with fractional counts written to the count matrix. Identical alignment sets are collapsed into signatures before iteration, and convergence is assessed by the maximum per-feature relative change in the posterior weights (tolerance 1e-4, maximum 200 iterations). Deduplication, when enabled, keys on the full alignment-coordinate set rather than on bowtie2's randomly chosen primary alignment, so that tie-breaking cannot split one molecule into several.

Formally, for each read, the set of alignment coordinates defines a signature. Reads with identical signatures are aggregated, and EM is run over signatures with a uniform initial distribution over alignments. The E-step assigns each signature's count to its alignments in proportion to current weights; the M-step re-estimates feature abundances. Fractional counts are emitted per feature. θ represents read-level abundance rather than molecule abundance, and no effective-length correction is applied: the target quantity is the IP/INPUT read ratio, in which feature length cancels, and multi-mapping reads are allocated in proportion to the read counts already observed for each candidate feature.

Features are addressed by integer id during parsing. Once parsing completes, the signature table is written to two temporary files (an int32 element stream and a per-signature count/length tuple) and released, so memory no longer scales with signature diversity. EM then runs on dense float64 numpy arrays in chunks of at most 3.2×10^7 elements, cut on signature boundaries; θ is accumulated across chunks rather than re-estimated per chunk, so the model is identical to a single pass and only the grouping of floating-point additions differs. On the six benchmark libraries this reproduces the previous counts exactly (identical md5) while cutting the EM task 3.2–5.6 fold in runtime and 5.0–6.8 fold in peak memory (see Performance).

In the HeLa IP library, EM-assigned reads were distributed as 59.0% tRNA, 13.7% mitochondrial tRNA, 10.0% rRNA, 7.0% lncRNA, 6.3% mRNA and 4.1% other classes (Fig. 3). The rRNA fraction was lower in IP than in INPUT (10.0% vs 15.3%), whereas lncRNA was higher (7.0% vs 4.2%), consistent with depletion of ribosomal RNA and retention of specific non-coding species by capture.

Differential enrichment statistics

All samples are modelled together in one edgeR quasi-likelihood GLM [16, 17] (R 4.3, edgeR 4.0) with empirical Bayes dispersion. The design includes fraction (IP versus INPUT) and, when present, condition, with paired IP–INPUT blocking; the interaction term provides the differential-interaction criterion. Enrichment is called when the IP–INPUT log2 fold change is positive and the Benjamini–Hochberg FDR is below 0.05. A separate, non-test flag implements the third dsRIP-Seq criterion: it marks features whose change between conditions is larger in absolute value in IP than in the unenriched channel, flagging candidates whose signal could be driven by an input-side expression change; it does not filter features and, as in the original protocol, carries no p value [3]. No TMM normalisation is applied [18]: on

enrichment data, TMM factors absorbed the capture signal in our hands (a 60-fold distortion in a controlled test), so libraries are compared on raw library-size scaling instead.

The GLM requires more samples than design parameters. The benchmark analysed one library per channel, so the design reduced to ~ fraction: two coefficients estimated from two libraries, leaving zero residual degrees of freedom. The module therefore degraded explicitly, reporting descriptive enrichment directions, writing NA for FDR, and stating in the HTML report that no statistical test was performed. Adding the second sequencing run of the same library would give four libraries and two residual degrees of freedom; that is enough to fit the model, but the variance it estimates is sequencing variance, not biological variance, so we did not use it to produce FDR-level claims. HeLa, H9 and the rPAL pair were run separately rather than pooled: paired blocking is on condition × replicate, and pooling two cell lines in one run would let features unique to one cell line enter the other's EM denominator with near-zero counts.

Annotation and reporting

Features were classified from reference headers (tRNA, rRNA, snRNA, snoRNA, scaRNA, Y RNA, 7SL, vault, miRNA, mRNA, lncRNA, other) and matched against a curated list of known glycoRNAs. Cross-links to RNAcentral [22] and Ensembl were included where available. The report is a single self-contained HTML file with inline SVG; wording adapts to the controls present, distinguishing capture enrichment from glycan-dependent enrichment.

Reproducibility, pre-specified targets and availability of the exact environment

The workflow is implemented in Nextflow (DSL2) with Python helper scripts, R/edgeR, bowtie2, cutadapt, samtools and numpy. The exact environment used for every number reported here was Nextflow 26.10.5, Python 3.12, numpy 2.5.5, R 4.3.3 with edgeR 4.0.16, bowtie2 2.5.4, cutadapt 4.9 and samtools 1.25; the Docker and Singularity definitions [23, 24] pin the same set. UMI-aware extraction additionally relies on umi_tools 1.1.5 [11], which bioconda publishes only for Python ≤ 3.11; it is therefore installed in a separate Python 3.11 environment when a UMI is declared, and no result reported here used UMI mode. A self-test of four suites (25 pass/fail assertions) generates synthetic references and reads and verifies every output layer end to end—count-matrix columns, three-criterion statistics tables, per-feature annotation and the self-contained HTML report (Additional file 3)—so the installation can be verified without data or a reference. The reference index is rebuilt from GtRNAdb Hsapi38 mature tRNAs [25] (retrieved 14 November 2026) and NCBI RefSeq GRCh38.p14 (GCF_000001405.40, rna_from_genomic) by a documented script that fixes both accessions and yields 201,812 sequences (260 collapsed tRNAs and 201,552 transcripts); species presets cover human, mouse, rat, *Arabidopsis thaliana* and *Oryza sativa*.

Resource use was read from Nextflow's execution trace. Cost is reported on a pre-specified 8-core basis: 8-core-equivalent seconds per sample = Σ(realtime × %CPU/100) / n_samples / 8, where the sum runs over all tasks. Peak memory is the largest resident-set size reported by any task. Wall-clock figures are from paired runs in which both libraries of a dataset ran concurrently on a 64-logical-core host with 16 alignment threads per task.

The performance envelope (≤ 3,600 s per sample on an 8-core-equivalent basis, ≤ 16 GB peak), the ≥ 80% direction-concordance target for the public-data benchmark, and the four statistical test-bench gates were fixed before the corresponding runs and are archived with their dates in the project decision log (statistical bench accepted 15 November 2026, performance re-run recorded 16 November 2026). No target was changed after results were known.

Results

Gold standard and direction concordance

We used GSE136967 (Ac_4ManNAz) and the 193-candidate Table S2 from Flynn et al. as the gold standard [1]. Each cell line was analysed as an independent run, with one INPUT and one IP library. The public archive contains one BioSample per channel, each sequenced twice; the two biological replicates described in the original study are not separately available in the archive, so the real-data analyses are descriptive. The two cell lines were deliberately not merged: EM aggregates alignment signatures across samples, so features unique to one cell line would enter the other's denominator with near-zero counts and noise-driven directions (in a merged run the HeLa concordance fell from 0.835 to 0.805).

Of 148 HeLa gold-standard positives with positive log2 fold change, 127 could be mapped to features in our index and detected in this dataset; direction concordance (the fraction of these with a positive IP/INPUT log2 fold change) was **0.835 (106/127)**. For H9, 124 of 145 positives were mappable and detected, with concordance **0.815 (101/124)** (Table 2; Fig. 2a). Counting every literature positive in the denominator, including the 21 entries that could not be mapped to the current RefSeq release or were not detected, gives conservative lower bounds of 71.6% (106/148) in HeLa and 69.7% (101/145) in H9; both remain well above the corresponding expression-matched nulls. The per-entry mapping of every gold-standard positive to index features (exact, alias or multi-copy family), with match method and inclusion in the concordance denominator, is given in Additional file 1: Table S1.

Agreement with the published analysis

Ranking agreement with the published DESeq2 [19] analysis was lower and of similar magnitude in both cell lines: the Spearman correlation between our enrichment log2 fold change and the published log2 fold change was **0.543** over 419 comparable RNAs in HeLa and 0.465 over 455 in H9; top-N set overlap was limited (Jaccard 0.042, 0.053 and 0.078 in HeLa; 0.099, 0.099 and 0.115 in H9 for N = 50, 100 and 200; Fig. 2d). These values are a consistency check rather than a ranking of methods: the published

analysis used biological replicates, independent alignment and DESeq2 size factors, whereas our runs used one library per channel, no normalisation, and a family-representative rule for multi-copy ncRNAs (below). We separated these factors on our own count matrices. Replacing library-size scaling with DESeq2's median-of-ratios size factors left the correlation unchanged (0.525 versus 0.525 in HeLa, 0.456 versus 0.456 in H9), and an unbiased family-level aggregation—summing member counts before the fold change—gave 0.517 and 0.456, within 0.008 of the representative rule. The residual therefore sits in the counts and the feature mapping (EM assignment and annotation differences), not in normalisation or in how families are collapsed; forcing DESeq2's normalisation does not improve agreement. Because the datasets lack biological replication, the statistical layer ran in descriptive mode; the concordance metric depends only on the sign of the fold change and is therefore unaffected.

Null models and background

A high concordance could in principle reflect a global tendency of the IP library to over-count most features rather than enrichment of the gold-standard set. Two checks address this, applied separately to each cell line. Among detected features outside the gold set (IP + INPUT ≥ 10 reads), 56.3% (536/952) were positive in HeLa and 53.6% (621/1158) in H9, so a positive sign is common even for features the gold standard does not list; features with fewer than 10 combined reads were excluded because their fold-change direction is dominated by counting noise. Gold positives were nonetheless 3.9-fold (HeLa) and 4.0-fold (H9) more likely than background features to be positive (two-sided Fisher exact test, $p = 1.2 \times 10^{-9}$ and 3.9×10^{-10}). Permuting gold membership while matching the expression distribution of the gold set—10,000 draws stratified by decile of log10(IP + INPUT + 1)—gave null concordances of 0.579 (95% interval 0.504–0.654) in HeLa and 0.544 (0.463–0.626) in H9, against observed values of 0.835 and 0.815 (both on the entry basis; the null script records the H9 observation as 0.821 on its de-duplicated feature basis, see below). The empirical p value is 1×10^{-4} in both, the smallest attainable with 10,000 draws. Because the gold standard is tRNA-rich, we repeated the permutation with joint matching on RNA class and expression decile, in case classes differ in their overall positivity. This moved the null to 0.491 in HeLa and 0.583 in H9—opposite directions, so class composition has no consistent effect—while the observed values stayed outside both null distributions (empirical $p = 1 \times 10^{-4}$). The direction result is therefore not an artefact of class composition. The null-model script counts distinct family-representative features rather than literature entries: two H9 gold positives (snoU13 and SNORD13, an alias pair) map to the same SNORD13 feature, so this check is computed over 123 distinct features, whereas the primary concordance above is 101/124 = 0.815 over entries (per-entry mapping in Additional file 1: Table S1).

Cross-chemistry check

To ask whether the direction signal survives a change of chemistry, we applied the same pipeline to the rPAL protocol, which enriches glycans through vicinal diols rather than through an azide handle [2]. We used one HeLa replicate pair from the public rPAL series (SRR35543218, IP; SRR35543220, INPUT). Of the 104 HeLa gold positives that were mappable and detected in this library, **52.9% (55/104)** had a

positive IP/INPUT log2 fold change. Background positivity was lower than in the click-chemistry libraries (29.9%, 570/1909), and the expression-matched null was 0.344 (95% interval 0.260–0.423; empirical $p = 1 \times 10^{-4}$; Fisher odds ratio 2.64, $p = 2.1 \times 10^{-6}$). Cross-chemistry recovery is therefore real but partial: it is far above its null and far below the same-chemistry value, and it is not a precision estimate for either chemistry.

Synthetic validation

On synthetic data with known truth (60 reference sequences, six positives boosted six-fold in IP, three replicates × two channels, 8,000 reads per sample), the full path from FASTQ to GLM recovered **6/6** positives and reported **0/54** false positives (sensitivity and specificity both 1.000; Fig. 2e). This test exercises the entire chain rather than the statistical module alone. A separate module-level bench, starting from a 2,000-feature count matrix with 12 paired samples, 100 true enrichments (40 up-regulated, 20 down-regulated, 40 constant), 40 input-driven decoys and 1,860 background features, recovered **100%** of criterion-34 positives, **85%** of criterion-36 differential features, **0%** of input-driven decoys and **100%** of the criterion-37 truth set, with an empirical background FDR of 0.000 (Additional file 2). The two tests are complementary: the whole-chain test asks whether the pipeline is assembled correctly, and the bench asks whether the statistics are calibrated.

Table 2
Benchmark metrics.

Metric	Definition	Result
Direction concordance	Known positives with positive IP/INPUT log2FC	**0.835** (106/127) HeLa; **0.815** (101/124) H9
Conservative bound	Unmapped or undetected known positives counted as discordant	71.6% (106/148) HeLa; 69.7% (101/145) H9
Null-model check	Observed concordance vs. expression-matched permutation null (10,000 draws)	**0.835** vs. 0.579 (0.504–0.654) HeLa; **0.815**□ vs. 0.544 (0.463–0.626) H9
Cross-chemistry recovery	rPAL HeLa: known positives with positive IP/INPUT log2FC	**0.529** (55/104); null 0.344 (0.260–0.423)
Rank correlation	Spearman ρ vs. published DESeq2 log2FC	**0.543** (419 RNAs, HeLa); 0.465 (455 RNAs, H9)
Strong-change subset	ρ / direction agreement for |log2FC| ≥ 1	0.525 / 76.1% (HeLa); 0.479 / 70.8% (H9)
Top-N overlap	Jaccard vs. the published DESeq2-significant set (N = 50/100/200)	0.042 / 0.053 / 0.078 (HeLa); 0.099 / 0.099 / 0.115 (H9)
Synthetic sensitivity	True positives recovered (6 positives)	**1.000** (6/6)
Synthetic specificity	False positives among 54 negatives	**1.000** (0/54)
Runtime	8-core-equivalent CPU seconds per sample	**1,334 s** (HeLa), **2,068 s** (H9), 6,329 s (rPAL)
Peak memory	Maximum resident set size	**1.5 GB** (HeLa), **1.7 GB** (H9), 9.2 GB (rPAL)

□ All observed values are on the literature-entry basis used throughout the text (0.835 = 106/127; 0.815 = 101/124). The null-model script counts distinct family-representative features: two H9 gold positives (snoU13 and SNORD13, an alias pair) map to the same SNORD13 feature, so the same 101 concordant entries correspond to 123 distinct features and the script records 0.821 (101/123) internally. The two bases differ only in the denominator (Additional file 1: Table S1).

Worked example: the Y RNA family across cell lines

To show how the pipeline reads on a known target without any new processing, we revisited the four Y RNA genes (RNY1, RNY3, RNY4, RNY5)—the prominent glycoRNA family of the Flynn et al. study [1]—directly in the benchmark count matrices. In H9, glycoRIP reports positive IP/INPUT enrichment for all four Y RNAs (log2 fold change + 2.3 to + 4.2), matching the direction of the published DESeq2 analysis, in which all four are significant (padj ≤ 8.6×10^{-7}). In HeLa, three of the four reproduce the published direction (Y1 + 1.5, Y3 + 0.9, Y5 + 1.1), while Y4 shows no enrichment (log2 fold change ≈ 0) against a

published + 1.65—one of the minority of discordant directions behind the 83.5% concordance above. Published magnitudes are larger than our fold changes (Y5: +4.70 versus + 1.1), consistent with the moderate global rank agreement (ρ = 0.543 and 0.465; Fig. 2b,c) and with the differences in normalisation and replication between the two analyses. The example reuses counts already produced by the benchmark run, is limited by one library per condition, and makes no claim of statistical significance (Fig. 4; per-feature values in the archived supplement).

Performance

On a 64-logical-core host, the two libraries of each dataset were processed concurrently. For HeLa the longest library chain (adapter discovery, trimming, alignment and EM assignment) finished in 1,177 s wall clock with a peak of 1.5 GB; for H9 the corresponding figures were 2,292 s and 1.7 GB. Expressed on the pre-specified 8-core basis—total CPU time divided by 8 and by the number of samples—this is **1,334 (HeLa) and 2,068 (H9) CPU-seconds per sample**, below the 3,600 s target, with memory margins of 10.7-fold and 9.4-fold under the 16 GB target (Table 3; Fig. 2f,g). The two chains were not equal in length: for HeLa, alignment took 8m51s at most and EM assignment took 9m50s; the larger H9 libraries roughly doubled both. Wall clock for a paired run is therefore set by the slower chain, not by the sum of the two. Stage times are measured maxima from the same libraries; alignment is unaffected by the quantifier rewrite described below.

The rPAL pair is far more complex. Its IP and INPUT libraries yielded 14.2 and 35.1 million distinct signatures, against 1.86 million for HeLa INPUT, because near-identical tRNA and small ncRNA families generate many distinct alignment sets. The first version of the pipeline kept every signature in memory and paid for it: 22,255 s wall clock, 18,433 s in a single EM task and a 64.4 GB peak, exceeding both targets (2.1-fold on time, 4-fold on memory). We therefore rewrote the quantifier rather than relaxing the statistic. Signatures are spilled to disk once parsing is complete, and the EM runs in numpy over fixed-size chunks that never split a signature, with θ accumulated globally across chunks. The prefix-sum accumulation order is preserved, so all six benchmark libraries reproduce the previous counts byte-for-byte. The EM task falls from 18,433 s to 3,303 s and its peak from 62.9 to 9.2 GB, which brings the pair inside the 16 GB memory target (1.7-fold margin). Runtime remains 6,329 s per sample on the 8-core basis (1.8-fold over the target), and 95% of that residual is bowtie2 aligning 51 million reads rather than EM. We report this as it stands: the 8C16G envelope is met for GSE136967-class libraries and for rPAL memory, but not for rPAL runtime at this sequencing depth.

Thread count affected both runtime and CPU cost. On this host, 32 alignment threads sit on the hyperthread boundary of 32 physical cores and cost 2.4-fold more CPU than 16 threads while running 30% slower; switching the default to 16 threads reduced the 8-core-equivalent cost from 3,951 to 1,460 s with bit-identical statistics (1,334 s after the EM rewrite above). The workflow therefore defaults to 16 threads per alignment task and documents that users should recalibrate on their own hardware.

Table 3
Runtime and memory.

Process	Wall time, longest task (HeLa / H9)	Peak RSS (HeLa / H9)
Adapter discovery	7 s / 7 s	1.4 GB
Trimming	< 1 min	0.6 GB
Alignment (16 threads)	8m51s / 17m36s	1.5 GB
EM assignment	9m50s / 19m34s	1.2 / 1.7 GB
Statistics, annotation, report	< 1 s each	< 64 MB
Full workflow (both libraries concurrent)	**1,177 s / 2,292 s**	**1.5 / 1.7 GB**
8-core equivalent	**1,334 / 2,068 s per sample**	—

Discussion

Glycan typing, glycosylation-site inference and replacement of orthogonal validation are out of scope. The goal is a transparent path from reads to candidates, with documented statistical assumptions and known failure modes. The benchmark supports four conclusions. First, enrichment direction, the least assumption-dependent signal, agrees with the published gold standard in both cell lines (83.5% and 81.5%) and lies far outside the range produced by expression-matched random feature sets. Second, when truth is known, the full computational chain neither loses true positives nor invents false ones. Third, the direction signal partially survives a change of chemistry: in rPAL data 52.9% of the same known positives were enriched, well above the matched null of 34.4% but well below the same-chemistry value. Fourth, the resource targets hold for the GSE136967 libraries and for the memory budget of a higher-complexity rPAL library, whose runtime remains 1.8-fold over the target because alignment, not quantification, dominates at that read depth.

The benchmark has clear limits. The public GSE136967 archive provides one BioSample per channel (two sequencing runs each); the biological replicates described in the original study are not separately available, so the real-data analyses are descriptive, and FDR-level performance on biological data remains to be established. Two cell lines were analysed, each with a single replicate pair, so cross-cell-line consistency rests on two independent runs rather than on a variance estimate. The cross-chemistry check rests on one HeLa rPAL pair; it cannot separate chemistry from library or batch effects, and the public rPAL series (~ 260 GB) was not transferred in full. One high-complexity library sits 1.8-fold over the runtime target while staying inside the memory budget, so the 8-core/16 GB envelope should be read as conditioned on library complexity. The reference is a current RefSeq release, whereas the published study used an earlier build; six snoRNA naming variants per cell line (nine distinct symbols across the two) therefore have no direct symbol match. Finally, EM resolves how reads are distributed within a

family of near-identical tRNAs, but it cannot distinguish individual members; single-member fold changes should not be interpreted biologically, and the report uses family-level language accordingly.

The near-term priorities follow from these limits: biological-replicate data (newly generated, or several rPAL replicates) to exercise the full GLM and FDR path on real samples; and a long-RNA mode (genome alignment with HISAT2/STAR [20, 21], GENCODE annotation, and pathway reporting) for libraries that target longer transcripts. The first extends the same statistical core, the second extends the same workflow and reporting layer. Memory on high-signature-diversity libraries has since been addressed by spilling signatures to disk and running EM in chunks (see Performance), which leaves alignment as the dominant cost.

Conclusions

glycoRIP converts glycoRNA enrichment sequencing data into an auditable, ranked candidate table through a single installable workflow, and makes its statistical assumptions explicit rather than implicit. Its validation is deliberately limited: it rests entirely on public data, the real-data analyses are descriptive, and one chemistry sits outside the runtime envelope. Those limits are reported rather than smoothed over, because the intended use is a reproducible starting point for glycoRNA studies whose key findings are confirmed by independent methods. Replicate data and a long-RNA mode are the next extensions, and both reuse the existing statistical core.

Availability and requirements

Item	Specification
Project name	glycoRIP
Project home page	https://github.com/cloud-seq-biotech/glycoRIP
Archived version	v0.1.3, https://doi.org/10.5281/zenodo.24464455 (concept DOI; resolves to the latest release)
Benchmark and reproducibility supplement	https://doi.org/10.5281/zenodo.24458512
Operating system(s)	Linux (x86_64); platform-independent through Docker or Singularity images
Programming language	Nextflow (DSL2), Python 3.12, R 4.3
Other requirements	Docker ≥ 20 or Singularity ≥ 3.8; 16 GB RAM recommended (8-core/16 GB target); all third-party tools pinned in the container images (bowtie2 2.5.4, cutadapt 4.9, samtools 1.25, edgeR 4.0.16)
License	PolyForm Noncommercial 1.0.0
Any restrictions to use by non-academics	Yes. Free for academic and other non-commercial use; commercial use requires a separate licence from the authors

Abbreviations

Ac_4ManNAz
N-azidoacetylmannosamine-tetraacylated
CPU
central processing unit
dsRIP-Seq
double-stranded RNA immunoprecipitation sequencing
EM
expectation–maximisation
FDR
false discovery rate
GLM
generalised linear model
glycoRNA
glycosylated RNA
GtRNAdb
genomic tRNA database
INPUT
unenriched input channel

IP
immunoprecipitation (enriched) channel
ncRNA
non-coding RNA
RefSeq
NCBI Reference Sequence database
rRNA
ribosomal RNA
snoRNA
small nucleolar RNA
tRNA
transfer RNA
TMM
trimmed mean of M-values
UMI
unique molecular identifier.

Declarations

Authors' contributions. W.L.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. The author approved the final manuscript.

Ethics approval and consent to participate.

Not applicable. This study reanalyses publicly available sequencing data (GSE136967; rPAL SRR35543218/SRR35543220) and generates no new human or animal data.

Consent for publication.

Not applicable.

Competing interests.

W.L. is an employee of Cloud-seq Biotech Co., Ltd. (Shanghai, China), a company that provides glycoRNA and other RNA sequencing services on a commercial basis. All benchmark data analysed here are public (GSE136967; rPAL SRR35543218/SRR35543220), and all code, scripts and raw measurement outputs are openly available; no proprietary customer data were used.

Funding.

This work received no external funding; it was supported by internal resources of Cloud-seq Biotech Co., Ltd.

Acknowledgements.

We thank the authors of the public datasets used in the benchmark (GSE136967 and the rPAL series), the maintainers of GtRNAdb and RefSeq, and the open-source bioinformatics community.

Availability of data and materials.

Public sequencing data: GSE136967 (HeLa and H9) and, for the cross-chemistry check, rPAL runs SRR35543218 (IP) and SRR35543220 (INPUT). Gold-standard table: Table S2 of Flynn et al. [1]. Code: https://github.com/cloud-seq-biotech/glycoRIP (v0.1.3, PolyForm Noncommercial 1.0.0; source archive at https://doi.org/10.5281/zenodo.24464455). Container images and reference-build scripts are distributed with the repository. The benchmark and reproducibility supplement package (metrics originals, six-library EM re-test reports and md5 manifest, traces, gold-standard package with an independent recomputation script, and Y RNA worked-example data) is archived at https://doi.org/10.5281/zenodo.24458512.

References

1. Flynn RA, Pedram K, Malaker SA et al (2021) Small RNAs are modified with N-glycans and displayed on the surface of living cells. Cell 184(12):3109–3124e22. 10.1016/j.cell.2021.04.023
2. Ge RQ, Jeppesen DK, Rai SK, Zhang Q, Higginbotham J, Coffey RJ, Flynn RA (2025) Catch and release of sialoglycoRNAs enables sequencing-based profiling across cellular and extracellular material. *bioRxiv*. 10.1101/2025.10.04.680438. Analysis code: github.com/FlynnLab/rPAL-seq (GPL-3.0-or-later); doi:10.5281/zenodo.21969177
3. Gao Y, Chen S, Halene S, Tebaldi T (2021) Transcriptome-wide quantification of double-stranded RNAs in live mouse tissues by dsRIP-Seq. STAR Protoc 2(1):100366. 10.1016/j.xpro.2021.100366
4. Zheng C, Ao Y, Zhang Z, Mei S et al (2026) GlycoRNAdb: a database of glycoRNA sequences, structures, abundance, and glycan information across tissues and cell lines. Nucleic Acids Res 54(D1):D158–D167. 10.1093/nar/gkaf1246
5. Behrens A, Rodschinka G, Nedialkova DD (2021) High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. Mol Cell 81(8):1802–1815e7
6. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419
7. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323

8. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527
9. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. Nat Biotechnol 35(4):316–319
10. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–12
11. Smith T, Heger A, Sudbery I (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 27(3):491–499
12. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
13. Chan PP, Lowe TM (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 44(D1):D184–D189
14. O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745
15. Zheng G, Qin Y, Clark WC et al (2015) Efficient and quantitative high-throughput tRNA sequencing. Nat Methods 12(9):835–837
16. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140
17. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40(10):4288–4297
18. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(3):R25
19. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550
20. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37(8):907–915
22. RNAcentral C (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res 49(D1):D212–D220
23. Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: scientific containers for mobility of compute. PLoS ONE 12(5):e0177459
24. Merkel D (2014) Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014(239):2
25. Chan PP, Lowe TM (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol 1962:1–14

Figures

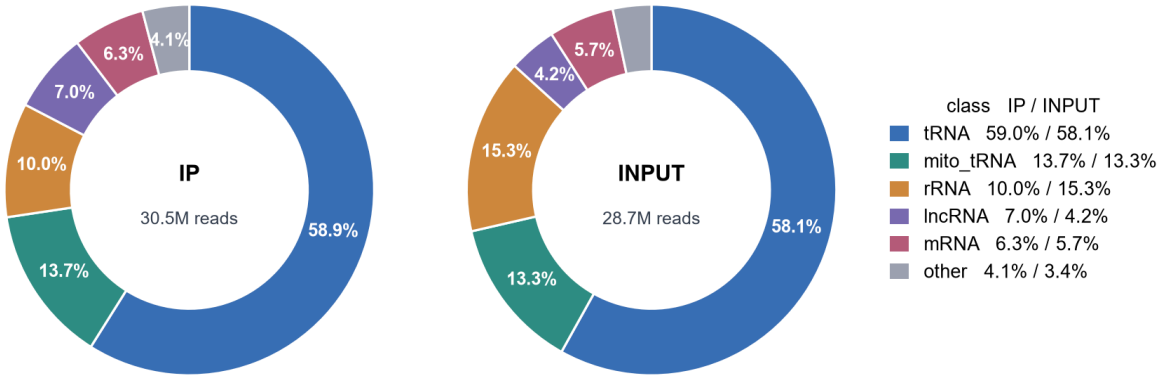
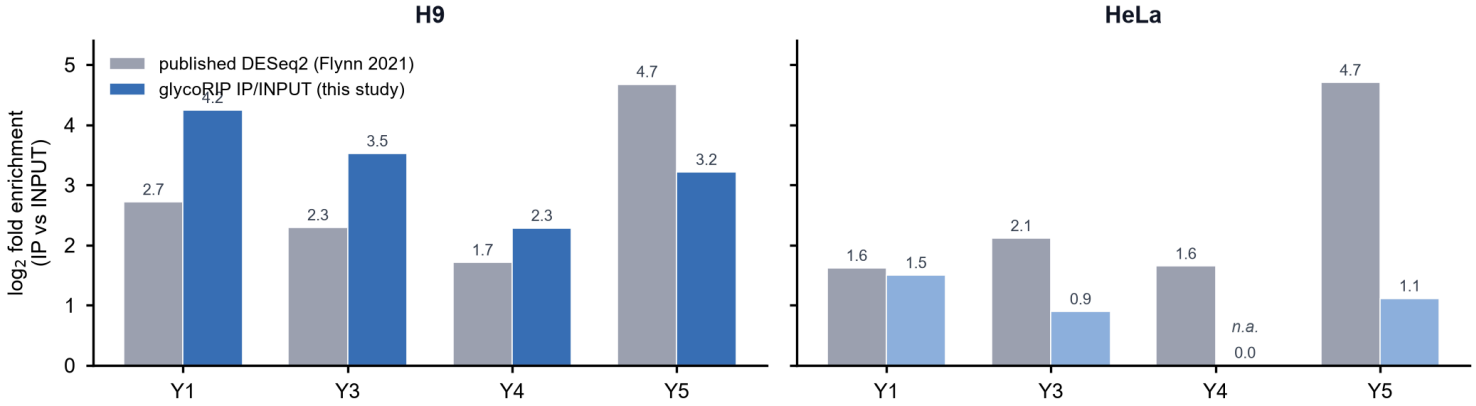


Figure 3

RNA-class composition of the HeLa benchmark libraries. Donut charts show the fraction of EM-assigned read equivalents in IP and INPUT falling into each annotated class; centre labels give total assigned reads. Legend values are IP / INPUT percentages. The lower rRNA fraction in IP (10.0% vs 15.3%) is consistent with rRNA depletion by capture.



GlycoRIP $\log_2$(IP+1)/(INPUT+1) per gene, single library per condition; published values from the DESeq2 analysis of Flynn 2021. n.a. marks the one direction-discordant feature (HeLa Y4: glycoRIP $\log_2$FC $\approx$ 0). Descriptive comparison — no statistical significance is claimed.

Figure 4

Worked example: Y RNA enrichment reproduced against the published analysis. (a) H9 and (b) HeLa. Grey bars: published DESeq2 log2 fold change from Flynn et al. [1]; coloured bars: glycoRIP IP/INPUT enrichment (log2((IP+1)/(INPUT+1))) computed from the benchmark count matrices, one library per condition. Direction agreement with the published analysis is 4/4 in H9 and 3/4 in HeLa; n.a. marks the discordant HeLa Y4 (glycoRIP log2FC approximately 0 versus published +1.65). The comparison is

descriptive; no statistical significance is claimed. Per-feature values and the generation scripts are in the repository and the archived supplement.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2NoteS1.docx
- Additionalfile3TableS2.xlsx