bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Verifiable Agentic Science on Public Infrastructure

Dannon Baker[1,7], Scott Cain[2,7], Marius van den Beek[2,7], John Chilton[2,7], Danielle Callan[3], Steven Weaver[3], Nate Coraor[2,7], Kelsey Beavers[4], Jeremy Goecks[5,7], Michael C. Schatz[1,7], Björn Grüning[6,7,8], Sergei L. Kosakovsky Pond[3,*], and Anton Nekrutenko[2,7,*]

[1]Department of Biology, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA, USA
[3]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Philadelphia, PA, USA
[4]Texas Advanced Computing Center, The University of Texas at Austin, Austin, TX, USA
[5]Department of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[6]Department of Computer Science, University of Freiburg, Freiburg, Germany
[7]galaxyproject.org
[8]usegalaxy.eu

[*]Correspondence: spond@temple.edu, anton@nekrut.org

Abstract

Autonomous artificial intelligence agents are transforming the practice of science. Across every discipline, researchers are handing data analysis to coding agents that can design and execute complex studies from a single prompt. Yet this revolution carries a threat. When agents improvise in local environments, they leave no verifiable record, ultimately risking a wave of irreproducible science. Solving this crisis requires moving agentic science onto public infrastructure where every action leaves an examinable trail. Here we connect autonomous coding agents to shared computing resources, turning ephemeral prompts into auditable scientific records. After showing that 30 unguided local agents splinter into 18 incompatible pipelines on the same dataset, we conduct an agentic reproduction and discovery study on a large scale—reproducing a large, diverse array of studies with every workflow, history, and intermediate dataset publicly accessible. This complete provenance trail exposes unstated modeling choices, reference mismatches, and subsampled tutorial inputs in published studies, and establishes a new foundation where human scientists can understand, steer, and trust AI-driven discovery.

Introduction

Across all scientific domains, language-model agents now formulate hypotheses, build pipelines, write code, and draft manuscripts [1–6]. Although several groups have built specialized research agents [7], domain-specific wrappers cannot keep pace with general-purpose command-line coding agents. At our own institutions, students, postdocs, and principal investigators—including long-time AI skeptics—now use command-line coding agents as their primary working environment.

Yet running autonomous agents on personal machines is equally thrilling and terrifying. Researchers can launch complex analyses beyond their training in minutes, but three flaws threaten reliability. Unsupervised local agents leave no verifiable provenance, generating disposable scripts and tables that vanish when the session closes. Without curated domain guidance, agents improvise fragile pipelines from unvetted snippets, and local execution hits a hardware and security

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

wall whenever an analysis requires high-memory alignment, large cohorts, or GPU inference on shared clusters. Without workflow, tool, and dataset provenance on shared public infrastructure, autonomous analyses are scientifically unverifiable.

For over twenty years, the Galaxy platform has provided open, reproducible infrastructure to life scientists worldwide [8], offering a natural public plane for agentic science. Agents have been connected to workflow platforms before—such as ClawBio's Galaxy Bridge over BioBlend and comparable tooling for Nextflow and AWS HealthOmics—yet what has been missing is a harness that keeps a checkable record of that work on free public compute. Any coding agent—or Orbit, our desktop workbench—can connect to Galaxy to draw on thousands of vetted containers, peer-reviewed workflows, and community training tutorials [9–14] on national supercomputers without requiring cluster credentials. Every analytical step, tool version, parameter setting, and intermediate dataset is recorded in a durable, shareable history and visual workflow linked to a research notebook, allowing human scientists to understand, steer, and verify what the agent is doing.

Here we show why autonomous agents need shared public infrastructure and what they can achieve once connected to it. After illustrating how unassisted local agents splinter into conflicting pipelines and disposable scripts on a single dataset, we conduct a large-scale validation and reanalysis of published studies spanning bacterial, fungal, vector, mammalian, parasitic, and viral systems on UseGalaxy.org and national supercomputers. Across spatial transcriptomics, single-cell atlases, bulk RNA-seq, chromatin accessibility, protein structure phylogenomics, and epidemic dating, every reanalysis checks out and leaves a complete, publicly auditable trail of histories, workflows, and datasets. That end-to-end traceability not only reproduces published biological signals and scales to global pathogen discovery, but also resolves unstated sample pooling, annotation mismatches, custom table filtering, and tutorial-subset reference data across the original publications.

Results

Advanced models propose vastly different analysis plans ...

When handed an RNA-seq dataset, a coding agent returns a plausibly looking workflow in seconds, leaving no easy way to judge whether the plan is sound. We asked 30 independent agents—10 each from Anthropic (Claude Code, Claude Opus 5.5), OpenAI (Codex, GPT-6 Astra), and Google (Antigravity, Gemini 3.8 Flash)—to design a replication plan for a transcriptomic study of the foodborne pathogen *Bacillus cereus* (PRJEB86181), which causes diarrheal illness by secreting pore-forming toxins in the gut [15]. Vyklická et al. tested how 0.5% dietary glucose regulates toxin expression in the dairy-associated strain ATCC 10987 ($n = 2$ glucose vs. $n = 3$ control replicates from a 19-library sugar screen, see Methods). Two pipeline details were left unstated in the methods text: the paper reported locus tags from *Bacillus pacificus* anQ-h4 without naming the reference assembly (while RefSeq had suppressed ATCC 10987 during taxonomic revision), and it did not state whether DESeq2 modeled all 19 libraries together or only the 5 focal samples.

Faced with these ambiguities, agreement among the 30 agents collapsed. Twenty-nine of 30 plans diverged from the published workflow, splintering into 18 distinct eight-step pipelines and 11 incompatible designs across the five choices that alter final gene counts and p-values: reference genome, rRNA filtering, aligner, gene annotations, and sample pooling (Figure 1, Supplementary Table S1).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

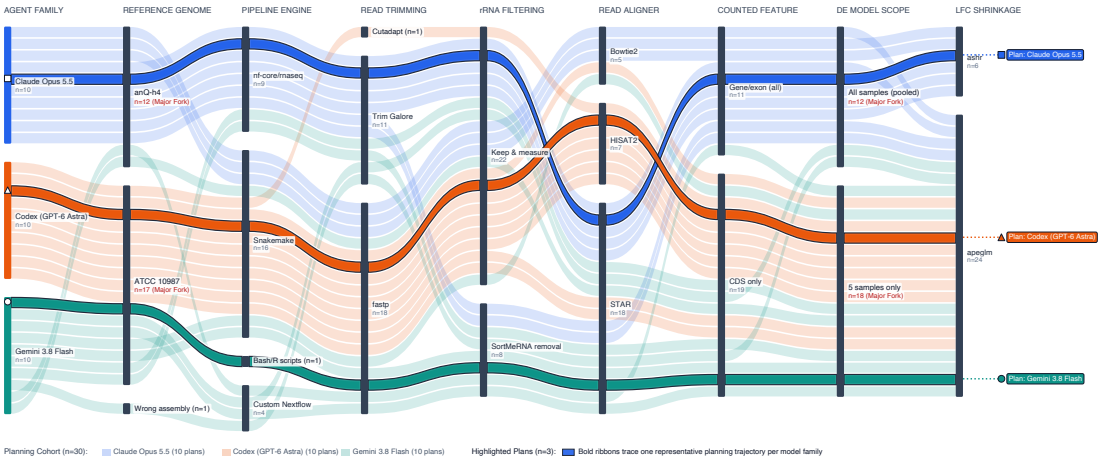


Figure 1: Thirty autonomous agents propose 18 different pipelines for the same bacterial RNA-seq study. Alluvial diagram tracing 30 independent reanalysis plans for a published *Bacillus cereus* carbohydrate-supplementation study (PMC12804731) [15] without access to Galaxy infrastructure. Plans were generated by 10 Anthropic Claude Code agents running Claude Opus 5.5 (blue), 10 OpenAI Codex agents running GPT-6 Astra (orange), and 10 Google Antigravity agents running Gemini 3.8 Flash (teal). Columns represent eight sequential pipeline decisions from reference assembly selection to log-fold-change shrinkage, with node heights proportional to the number of agents choosing each option ($n = 30$ total). Transparent ribbons show individual plans, while bold ribbons with geometric markers trace one executed trajectory per model family (square for Claude, triangle for Codex, and circle for Gemini). Across all eight steps, the 30 plans split into 18 distinct workflows (11 distinct designs across the five result-changing steps), and 29 of 30 differed from the published workflow.

... and leave a dusty trail of intermediates

We then asked one agent per provider to actually execute its plan locally. Once execution began, each run drifted even further from its own plan as it improvised around local software and archive metadata (Supplementary Table S1). All three runs recovered headline repression of *nheB* and induction of *entFM* (Supplementary Figure S1A), yet diverging reference genomes silently altered the detected toxin repertoire. Aligning to *B. pacificus* anQ-h4 missed cytotoxin K (*cytK2*) entirely because that assembly lacks the locus, whereas aligning to the sequenced isolate *B. cereus* ATCC 10987 recovered active *cytK2* transcription across every library (Supplementary Figure S1B). Compute footprints varied by orders of magnitude, and the three runs left behind 26 unversioned glue scripts totaling over 1,200 lines of non-portable code (Supplementary Tables S1 and S2). Without shared infrastructure, local agents leave a debris field of disposable scripts in place of a verifiable record.

Agentic science on public infrastructure: The Galactic plane

Eliminating pipeline fragmentation, disposable glue code, and local hardware limits requires pairing domain skills with a shared execution engine (Supplementary Figure S2). Any coding agent—or Orbit, our desktop workbench—connects to public Galaxy infrastructure [8] through modular skills and the Galaxy MCP server [16, 17]. Instead of improvising shell commands, the agent invokes containerized tools and peer-reviewed workflows via typed API calls across national supercomputers and cloud nodes (Supplementary Figure S2, bottom), running everything from memory-intensive aligners like STAR [18] to foundation models and petabyte-scale index searches—such as AlphaGenome [19], Logan [20], and LexicMap [21]—that no personal workstation can host.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Shared infrastructure makes complete provenance mandatory by default (Supplementary Figure S2, top right). Narrative model summaries omit silent parameter changes and failed retries, and few biologists possess version control discipline over interactive shell sessions. Pushing execution onto Galaxy and Orbit records every input file, tool version, parameter, and intermediate dataset in a durable, shareable Galaxy history and workflow graph. The research notebook then links each narrative claim directly to the corresponding Galaxy datasets and job invocations, making the entire analysis human-readable and immediately verifiable. When an analysis needs custom code, the agent registers that script as a versioned, containerized User-Defined Tool[1] inside the workflow rather than leaving an untracked local file.

Reproducing a paper using a generic agent on the Galactic plane

To test how shared infrastructure changes interactive analysis, we connected Google's Antigravity CLI (Gemini 3.8 Flash) to UseGalaxy.org via the Galaxy MCP server [16] to reproduce the *B. cereus* RNA-seq study [15] in a human-in-the-loop session. Unlike local runs that generated 26 unversioned scripts or exhausted laptop disk space, the entire replication ran on free public supercomputers (Supplementary Figure S3).

Shared history and workflow provenance proved essential for human steering. Both the local GPT-6 Astra run and the MCP-connected Gemini 3.8 Flash agent initially ran HISAT2, a eukaryotic spliced aligner that is inefficient for an intronless bacterium (see Methods). Inspecting the shared Galaxy history allowed us to spot the HISAT2 run, steer the agent to assemble a reusable Kallisto workflow [22], and keep the initial HISAT2 output as a direct cross-check ($r = 0.9622$). Quantifying reads against bacterial coding sequences with Kallisto avoids spliced alignment overhead and finishes each library in 38–90 seconds (median 59 seconds) instead of 3.6–6.3 minutes for HISAT2, provided duplicate transcript identifiers on frameshifted transposase features are deduplicated (with both raw and deduplicated references preserved in the Galaxy history, see Methods).

Formal workflow validation also prevented silent scripting workarounds. When the agent wired an incompatible standalone transcript-import step between Kallisto's output collection and DESeq2—and after GTF-based gene mapping failed at runtime on the bacterial annotation—Galaxy's workflow editor flagged the collection mismatch in orange so the agent supplied an explicit transcript-to-gene table to DESeq2's built-in import mode (Supplementary Figure S3A,B). The validated workflow reproduced fivefold repression of *nheB* ($p_{\text{adj}} = 2.99 \times 10^{-3}$), nominally significant nearly threefold repression of *nheA* ($p = 0.0274$, $p_{\text{adj}} = 0.106$) and induction of *entFM* ($p = 0.0158$, $p_{\text{adj}} = 0.0730$) in the five-library contrast, and unchanged expression of component *nheC* (Supplementary Figure S3D, see Methods). Auditing the resulting Galaxy Page notebook (Supplementary Figure S3C) against the history confirmed that every embedded dataset card and fold-change summary resolves to verified history outputs (Supplementary Table S3).

An opinionated agent

Domain skills and the Galaxy MCP server guide generic command-line agents effectively, yet terminal agents require command-line fluency and nothing in a skill file stops an agent from falling back on untracked local shell commands when a step fails. To give bench biologists a friendly desktop application whose harness can block and record what the agent does, we built Orbit on top of Pi.dev (Supplementary Figure S2). Orbit is Galaxy-first, steering tool execution to Galaxy while keeping

[1]Because executing arbitrary user-supplied tool definitions on public supercomputing infrastructure carries security and resource risks, User-Defined Tool registration currently requires account approval on UseGalaxy.org or routes through an automated packaging and review pipeline on UseGalaxy.eu (see Discussion).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

guarded local execution for lightweight work, tracking live cluster jobs, registering custom scripts as versioned User-Defined Tools [23], and syncing the analysis record into a version-controlled project notebook across commercial, public API, and local open-weight models.

Controlling an agent operates across three levels: domain skills that advise how to design an analysis, the MCP server that exposes typed Galaxy tools, and harness logic that mediates every tool call in code. Harness logic is essential because prompt-level instructions can be ignored. Because Galaxy provides a server-side record of typed tool runs that the agent cannot edit, Orbit's harness logic can verify directly against Galaxy whether a step's job has actually finished before allowing the plan step to be marked complete (see Methods). These guarantees apply to the record rather than the science: Orbit does not make an agent's analytical choices correct, but its harness logic ensures that what the agent did is checkable. We used Orbit on UseGalaxy.org to execute all remaining reproductions and discovery analyses summarized in Table 1 and detailed in the sections that follow.

Uncovering a reference discrepancy in spatial transcriptomics study

Because imaging-based spatial transcriptomics measures only a few dozen genes across a tissue section, TISSUE [25] uses paired single-cell RNA-seq reference atlases to impute unmeasured transcripts and calibrate uncertainty intervals. Recently, Miao et al. [24] showcased Paper2Agent by wrapping TISSUE into an agent to predict expression and uncertainty for the vascular smooth muscle marker *Acta2* across 3,405 mouse somatosensory cortex cells (33-gene osmFISH assay). Wrapping code into a standalone agent, however, leaves execution unverified unless input datasets and workflow steps are recorded on shared infrastructure.

Tasked with reproducing the Paper2Agent spatial example, Orbit uncovered an ironic discrepancy: a demonstration designed to showcase automated agent reproducibility had run on a 1,000-cell tutorial subset rather than the full 5,613-cell reference atlas it cited. In Miao et al. [24], human researchers prompted Paper2Agent to follow the TISSUE GitHub tutorial to verify that the agent reproduced a human running the tutorial, while their Methods and Data Availability statement cited the full 5,613-cell mouse cortex reference archive (Zenodo Dataset 15) without noting that the tutorial uses a 96 KB, 1,000-cell subset bundled in the software repository. Because the 1,000-cell tutorial reference omits 82% of the 5,613 reference cells in Dataset 15, it mislocalizes predicted *Acta2* expression into a ventral band and leaves predictions nearly uncorrelated with measured ground truth ($r = 0.0861$, Figure 2A,B,D,E and Methods). Orbit registered two TISSUE User-Defined Tools on UseGalaxy.org, reproduced the 1,000-cell tutorial predictions to floating-point precision ($r = 1.0000$, $\Delta_{\max} < 10^{-15}$, 86.46% Galaxy vs. 86.49% local coverage at nominal 77%, Figure 2A,B), and then ran the same workflow on all 5,613 reference cells in Dataset 15. Using the full reference archive recovered dorsal pial smooth muscle localization, tripled *Acta2* prediction correlation to $r = 0.2553$ ($r = 0.2845$ in 10-fold cross-validation, 85.61% coverage, Figure 2C,F), and improved accuracy across 27 of 33 cortical genes (mean $r = 0.177$ to 0.273, Figure 2G and Supplementary Table S4). Whereas wrapping code into a standalone agent confirms that a tutorial script can be rerun, recording input datasets and workflows in a shared Galaxy history makes explicit what data were actually analyzed.

Uncovering analysis inconsistencies and defunct tools

To test multi-step functional genomics, we combined Orbit with Workflow Foundry [14]—which compiles published methods prose into validated Galaxy workflows—and BRC-Analytics [39], which stages curated pathogen reference assemblies and sequencing archives (see Methods). We reconstructed the 30-sample RNA-seq study of Hokken et al. [26] in *Aspergillus fumigatus*, an

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Analysis & System	Assay & Scale	Agent & Stack	Highlight
***Bacillus cereus* enterotoxin regulation** Vyklická et al. [15]	Bacterial RNA-seq (5 libraries, 5,472 coding loci)	Generic CLI agent + Galaxy MCP + Skills	Replaces 26 unversioned local glue scripts from unconstrained runs with a validated Galaxy workflow, preserving raw and deduplicated bacterial CDS references in the history and switching from HISAT2 to $\sim 5\times$ faster Kallisto pseudoalignment (38–90 s per library, median 59 s).
Mouse cortex spatial transcriptomics Miao et al. [24], Sun et al. [25]	Mammalian scRNA-seq + spatial prediction (3,405 spatial & 5,613 reference cells)	Orbit + 2 UDTs on UseGalaxy.org	Reconciles the 1,000-cell tutorial reference used in Extended Data Fig. 1c against the 5,613-cell Dataset 15 archive cited in Data Availability, reproduces the tutorial subset on Galaxy, and scales to all 5,613 reference cells to triple prediction accuracy ($r = 0.0861$ to 0.2553).
***Aspergillus fumigatus* azole time-course** Hokken et al. [26]	Fungal RNA-seq (30 libraries, 3 isolates $\times$ 5 timepoints)	Orbit + Workflow Foundry + BRC-Analytics + 3 UDTs	Compiles a 60-step workflow from published methods text, shows that the authors fitted one pooled 30-sample model instead of pairwise contrasts, and replaces a decommissioned web service (AspGD) and two manual web steps (Venn overlap and REVIGO) with automated UDTs.
***Anopheles gambiae* infection chromatin** Ruiz et al. [27]	Vector ATAC-seq (midguts & salivary glands, 4 libraries)	Orbit + 4 UDTs on UseGalaxy.org	Recovers 90.7–93.2% of the 111,586 replicate-shared chromatin regions in the deposited table (81.8–92.2% of all 193,005 deposited peaks), while showing that the paper's headline counts reflect unmerged per-replicate table rows and tissue-specific overlap filtering.
***Aedes aegypti* midgut single-cell atlas** Wang et al. [28]	Vector 10x 3' scRNA-seq (2 replicates, 6,707 post-QC cells)	Orbit + BRC-Analytics on UseGalaxy.org	Stages non-model vector reference data into a public single-cell pipeline, resolves a mitochondrial gene prefix mismatch that breaks generic prefix-based Scanpy QC, and recovers three broad midgut populations and known marker genes.
***Plasmodium vivax* VIR phylogenomics** Carlton et al. [29], Auburn et al. [30], Minassian et al. [31], Demeulenaere et al. [32], Harrison et al. [33]	Structure-aware phylogenomics (3 genomes, 2,592 proteins)	Orbit + 10 UDTs (1 GPU, 9 CPU) on UseGalaxy.org	Orchestrates GPU protein structure models (ProstT5, Foldseek, and FoldMason) via unprivileged UDTs, collapsing 203 sequence clusters (146 subfamilies with ≥ 3 members) into 73–76 structural classes (24–30 with ≥ 3 members) and showing that 20–21 pan-continental classes hold 96.5–97.3% of all VIR proteins.
Pathogen emergence dating across 42 cohorts 42 BEAST alignments across 37 studies [34–38]	Genomic epidemiology (42 viral & bacterial epidemic alignments, 41 unique)	Orbit + 2 UDTs on UseGalaxy.org	Dates 42 pathogen epidemic alignments in seconds per cohort instead of hours to days of Bayesian MCMC sampling, matching reference Bayesian emergence dates across 61.5–64.1% of scored cohorts.

Table 1: Summary of empirical validations and discovery analyses executed on the Galactic plane. Across seven biological investigations spanning bacteria, mammalian cortex, fungi, mosquito vectors, protozoan parasites, and 42 viral and bacterial epidemic alignments (41 distinct outbreaks across 37 publications), autonomous agents executed reproducible workflows on free public UseGalaxy.org and ACCESS-CI supercomputing infrastructure. Every analysis produced inspectable Galaxy histories, reusable workflows, and shareable research notebooks (complete accession identifiers in Supplementary Tables S3–S6).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

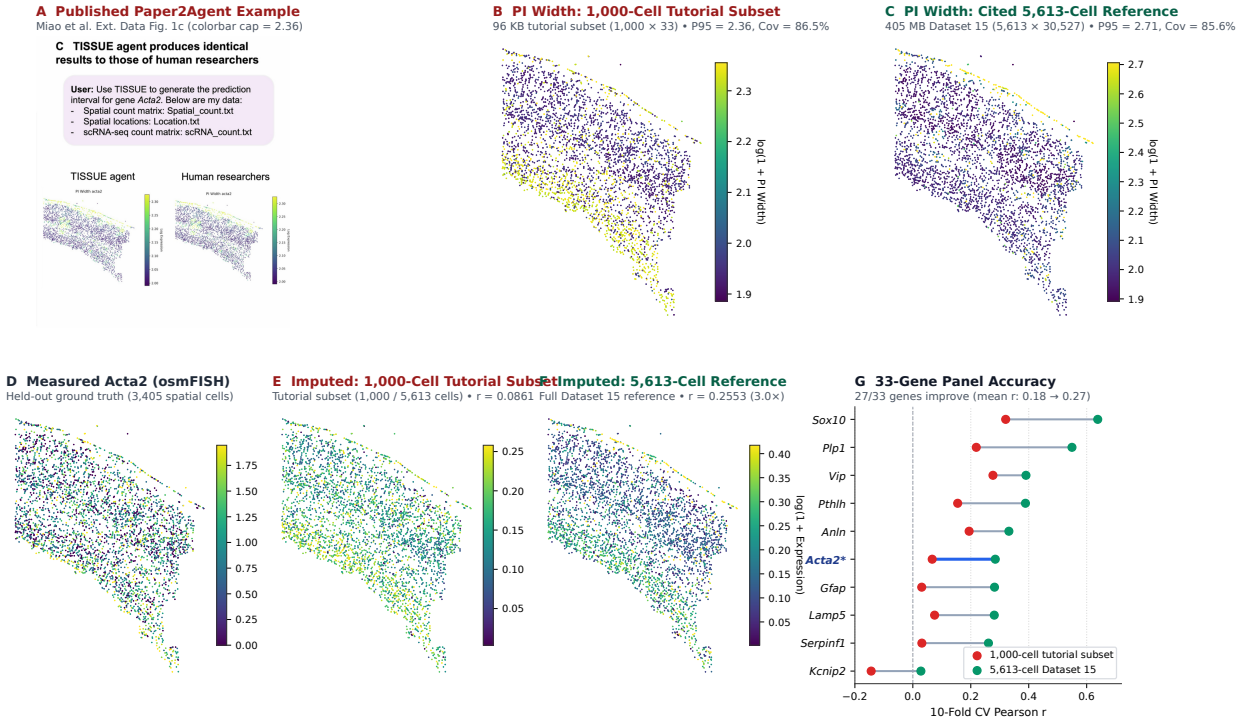


Figure 2: Comparing the 1,000-cell tutorial subset and full 5,613-cell reference atlas for spatial transcriptomics imputation on public infrastructure. A, Published Paper2Agent example from Extended Data Figure 1c of Miao et al. [24] showing conformal prediction interval (PI) widths ($\log(1 + \text{PI Width})$) for imputed *Acta2* expression across 3,405 mouse somatosensory cortex cells at nominal 77% coverage ($\alpha = 0.23$). The 95th-percentile colorbar upper bound (≈ 2.36, ticks from 2.0 to 2.3) matches the 96 KB, 1,000-cell tutorial subset rather than the 5,613-cell Dataset 15 reference cited in Data Availability. **B**, Orbit reproduction on UseGalaxy.org using the 1,000-cell tutorial subset (33-gene panel, 31 calibration genes retained after a 1% minimum expression filter drops *Flt1* from 32 unfiltered genes), matching the published 95th-percentile colorbar cap (P95 = 2.3552) and local execution to floating-point precision ($r = 1.0000$, $\Delta_{\max} < 10^{-15}$ under identical cross-platform splits, 86.46% Galaxy vs. 86.49% local coverage, mean $\log(1 + \text{PI Width})$ of 2.0621 vs. 2.0634 locally). **C**, Scaling the same Galaxy workflow to the cited 405 MB Dataset 15 reference (5,613 reference cells × 30,527 genes, retaining all 32 calibration genes) shifts the 95th-percentile PI width cap to P95 = 2.7062 (mean $\log(1 + \text{PI Width}) = 2.0651$, 85.61% coverage) and localizes uncertainty to the pial vasculature. **D**, Measured osmFISH ground-truth expression ($\log(1 + \text{Counts})$) for held-out *Acta2* from Dataset 15, showing true vascular smooth muscle localization along the dorsal pial surface (see Methods). **E**, Imputed *Acta2* expression using the 1,000-cell tutorial subset ($r = 0.0861$), which mislocalizes predicted expression into a diffuse ventral band. **F**, Imputed *Acta2* expression using the full 5,613-cell Dataset 15 reference ($r = 0.2553$, 3.0× higher), recovering the dorsal pial vascular layer. **G**, Leak-free 10-fold cross-validated prediction accuracy (Pearson r) across the 33-gene cortical panel comparing the 1,000-cell tutorial subset (red) against the full 5,613-cell Dataset 15 reference (green). Full reference data improves imputation for 27 of 33 genes (mean $r = 0.177$ to 0.273), including *Acta2* (0.067 to 0.285), *Gfap* (0.031 to 0.282), *Serpinf1* (0.031 to 0.261), *Lamp5* (0.075 to 0.281), *Plp1* (0.219 to 0.549), and *Sox10* (0.320 to 0.638).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

airborne mold causing lethal invasive infections where frontline triazole resistance is rising. Hokken et al. profiled transcriptional responses to itraconazole across three clinical isolates and five exposure timepoints in duplicate (PRJNA482512). BRC-Analytics staged the Af293 genome and all 30 libraries, and Workflow Foundry compiled a 60-step Galaxy workflow (86.88% mean unique alignment vs. 86.5% published, see Methods).

Comparing our initial differential expression outputs against the authors' supplementary spreadsheet uncovered an unstated statistical modeling choice and three manual downstream web steps. Four separate 12-sample pairwise DESeq2 models reproduced the published fold changes ($r \geq 0.998$) but inflated standard errors by a median of 20–26% and called 31% and 39% fewer differentially expressed genes at 120 and 240 minutes (missing 34% and 42% of the published DEG sets). Inspecting the supplementary spreadsheet revealed the signature: every gene carried a single identical `baseMean` value across all four contrasts, showing that the authors had fitted one pooled 30-sample model across all five timepoints simultaneously. With only two replicates per group, pooling all 30 libraries provides 25 residual degrees of freedom to stabilize dispersion and applies a uniform expression threshold (baseMean ≥ 30). Switching our Galaxy workflow to a pooled 30-sample timepoint model matched the published standard error distribution (median ratio 1.000, interquartile range -0.4% to $+0.3\%$) and reproduced DEG counts within 0.3–1.3% at every timepoint ($r = 0.9996$–0.9999, Supplementary Figure S4A), while three User-Defined Tools replaced a decommissioned fungal Gene Ontology server (AspGD) and automated Venn overlap and REVIGO summarization inside Galaxy to recover the shift from cell-wall stress to drug-efflux induction (Supplementary Table S3).

Uncovering undocumented analysis details

End-to-end provenance also separates reproducible chromatin biology from table-formatting conventions. In the ATAC-seq study of Ruiz et al. [27] across midguts and salivary glands of the malaria vector *Anopheles gambiae*—the two epithelial barriers traversed by *Plasmodium falciparum* parasites during infection (GSE152924)—calling sub-nucleosomal open-chromatin peaks (< 130 bp) with the cited MACS2 parameters recovered 90.7–93.2% of the 111,586 replicate-shared regions in the authors' deposited table (81.8–92.2% across all 193,005 deposited peaks), yet standard consensus peak merging produced less than half of the paper's headline counts (43,010 midgut and 68,576 salivary-gland regions, which the paper termed unique regions). Comparing the deposited table against our Galaxy outputs showed that the table lists each replicate's peak as a separate unmerged row and that the stated two-sided 51% overlap rule matches midguts (-3.3%) but under-calls salivary glands (-16.1%, Supplementary Figure S4B, see Methods). Fitting an empirical relaxed overlap threshold reproduced the headline region totals ($+4.1\%$ and -1.4%) and matched downstream gene, promoter, and enhancer counts derived from the deposited table within 0.4–2.9% (Supplementary Table S3), preventing unmerged table rows and tissue-specific overlap thresholds from being mistaken for failed peak calling.

Dealing with disparate annotation formats

Reanalyzing the adult female midgut single-cell RNA-seq atlas of Wang et al. [28] in sugar-fed *Aedes aegypti*—the primary vector of dengue, Zika, chikungunya, and yellow fever viruses—illustrated an annotation trap that affects generic single-cell workflows on non-model organisms. Standard Scanpy pipelines filter dying cells by matching mitochondrial gene prefixes such as `MT-` or `ND`, whereas VectorBase labels *A. aegypti* mitochondrial genes with standard `AAEL` identifiers (which Wang et al. hard-coded as a 38-gene list in their local script). In our Galaxy history, querying a generic prefix matched zero genes and terminated Scanpy quality control, while an empty prefix flagged

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

every gene as mitochondrial. Prefixing the 38 mitochondrial genes in the annotation GTF resolved the failure, retained 6,707 post-QC cells across both replicates (compared with 7,147 published cells after doublet removal), and resolved three broad midgut populations (enterocyte/digestive, hemocyte-like, and visceral muscle) whose sample-blocked pseudobulk profiles recovered known midgut markers and digestive proteases (FDR ≤ 0.002, see Methods and Supplementary Table S3).

Heavy-Weight Model Orchestration and Structure-Aware Phylogenomics

We next tested whether Orbit can orchestrate GPU protein structure models on free public supercomputers to study variant surface antigen diversity in *Plasmodium vivax*, the most geographically widespread human malaria parasite. Each *P. vivax* genome encodes nearly 1,000 VIR variant surface proteins proposed to remodel infected red blood cells and mediate spleen cytoadherence [29–32,40,41]. Because diversifying selection drives VIR sequences into the twilight zone ($\leq$ 30% amino acid identity), sequence clustering across Indonesia (PvP01 [30]), Thailand (PvW1 [31]), and Peru (PvPAM [32]) splits the 2,664 PF05795-positive VIR proteins (2,592 after length filtering) into 203 sequence clusters (146 subfamilies with ≥ 3 members, of which 132 span all three genomes), obscuring family-level relationships across divergent sequences even though structure is conserved in related PIR proteins (where Harrison et al. [33] homology-modeled ten *P. vivax* VIRs onto the *P. chabaudi* CIR fold). Orbit registered ten User-Defined Tools on UseGalaxy.org (one GPU and nine CPU tools) chaining ProstT5, Foldseek 3Di clustering, FoldMason alignment, synteny analysis, and IQ-TREE 3 phylogenetics across ACCESS-CI NVIDIA A100 GPUs and multi-core CPUs (see Methods, Supplementary Table S5).

Structure-aware clustering compressed repertoire diversity across all execution branches, collapsing the 203 sequence clusters (146 subfamilies with ≥ 3 members) into 73 structural classes (24 with ≥ 3 members) on the local CPU reference branch and 76 classes on UseGalaxy.org GPU (25 with ≥ 3 members) and CPU nodes (30 with ≥ 3 members, Supplementary Figure S5A, see Methods). Across every branch, the biological conclusion held: 20 to 21 pan-continental structural classes contained members from all three genomes and accounted for 96.5–97.3% of all VIR proteins (96.8% on Galaxy GPU across 20 classes, 96.5% on Galaxy CPU across 21 classes, and 97.3% locally across 21 classes, or 97.6% of proteins in Galaxy GPU classes with ≥ 3 members), whereas single-genome classes held only 2.0–2.7%. On the 338-tip local reference alignment of the dominant superfamily (FS_001), maximum-likelihood structural phylogenetics and tip-label permutation testing showed no detectable geographic clustering across Indonesia, Thailand, and Peru (parsimony score = 160 vs. null median 160, $p = 0.55$, Supplementary Figure S5B). Although 64.5–65.7% of the 1,672 VIR genes placed on main chromosomes lie within the subtelomeric outer 10% of chromosome arms (57–90% per assembly) and recombination scrambles flanking gene order in all but 11 of 194 local (3 of 141 GPU) reciprocal best-hit pairs, these results are consistent with VIR structural folds diversifying before *P. vivax* spread across continents.

Dating emergence of key pathogens across 42 benchmark alignments

Finally, we tested high-throughput epidemic dating across 42 viral and bacterial benchmark alignments spanning Ebola, SARS-CoV-2, Zika, Dengue, Chikungunya, Mpox, Influenza, *Mycobacterium abscessus*, and *Yersinia pestis*. We executed this study in Orbit across two separate stages driven by dedicated prompts. First, in a dedicated harvesting and preparation stage, we prompted Orbit to collect, verify, and archive 42 self-contained author BEAST XML files from public repositories and data archives (GitHub, Zenodo, Dryad, and Figshare) across 37 publications (41 outbreaks plus a paired 14.5 kb Ebola coding partition) and unpack their inlined sequence alignments, tip

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

sampling dates, and clock specifications into synchronized Galaxy collections inside a preparation history on UseGalaxy.org (see Methods). Second, because estimating the time to the most recent common ancestor (t_{MRCA}) from these configurations in BEAST [42] typically requires hours to days of Bayesian MCMC sampling per alignment, a companion analysis prompt directed Orbit to register a User-Defined Tool wrapping ChronAeon (`https://github.com/veg/chronaeon`) [38] on UseGalaxy.org and date all 42 alignments concurrently in seconds of manifold computation per dataset inside an analysis history (Figure 3, see Methods).

Across major recent outbreaks—including 2014 West African Ebola [34–36] and SARS-CoV-2 Gamma [37], where ChronAeon's estimate (2020.71) fell within two months of the published BEAST root (2020.87)—tree-free pairwise-distance regression matched reference dates across 61.5% of the 39 scored cohorts (24 of 39), rising to 64.1% (25 of 39, median error 0.40 years across 19 post-1980 outbreaks) with HyphAeon Attention PGLS on Mpox Clade Ib (Supplementary Table S6). Preserving every alignment and diagnostic output in a public Galaxy collection also explained the 14 discordant cohorts: unrooted distance regression breaks down on ancient roots compressed by purifying selection and mutational saturation (such as Lassa [43, 44] and Measles/Rinderpest [45]) and on multi-lineage alignments where whole-alignment regression dates a deep species split rather than a regional introduction (see Methods).

Putting this all together

Anyone can use this architecture today (`https://galaxyproject.org/agents/`) with a free account and API key on UseGalaxy.org, UseGalaxy.eu, or UseGalaxy.org.au. Command-line users can install `agentic-plugins` (`https://galaxyproject.org/agents/plugins/`) to register `galaxy-mcp` with curated Galaxy and Workflow Foundry skills, while desktop users can download Orbit for macOS, Linux, or Windows (`https://galaxyproject.org/agents/orbit/`) using commercial APIs, ChatGPT subscriptions, or free local Ollama models. Both paths execute every step as a verifiable Galaxy job, and we are embedding a native open-model agent directly inside the Galaxy web interface.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

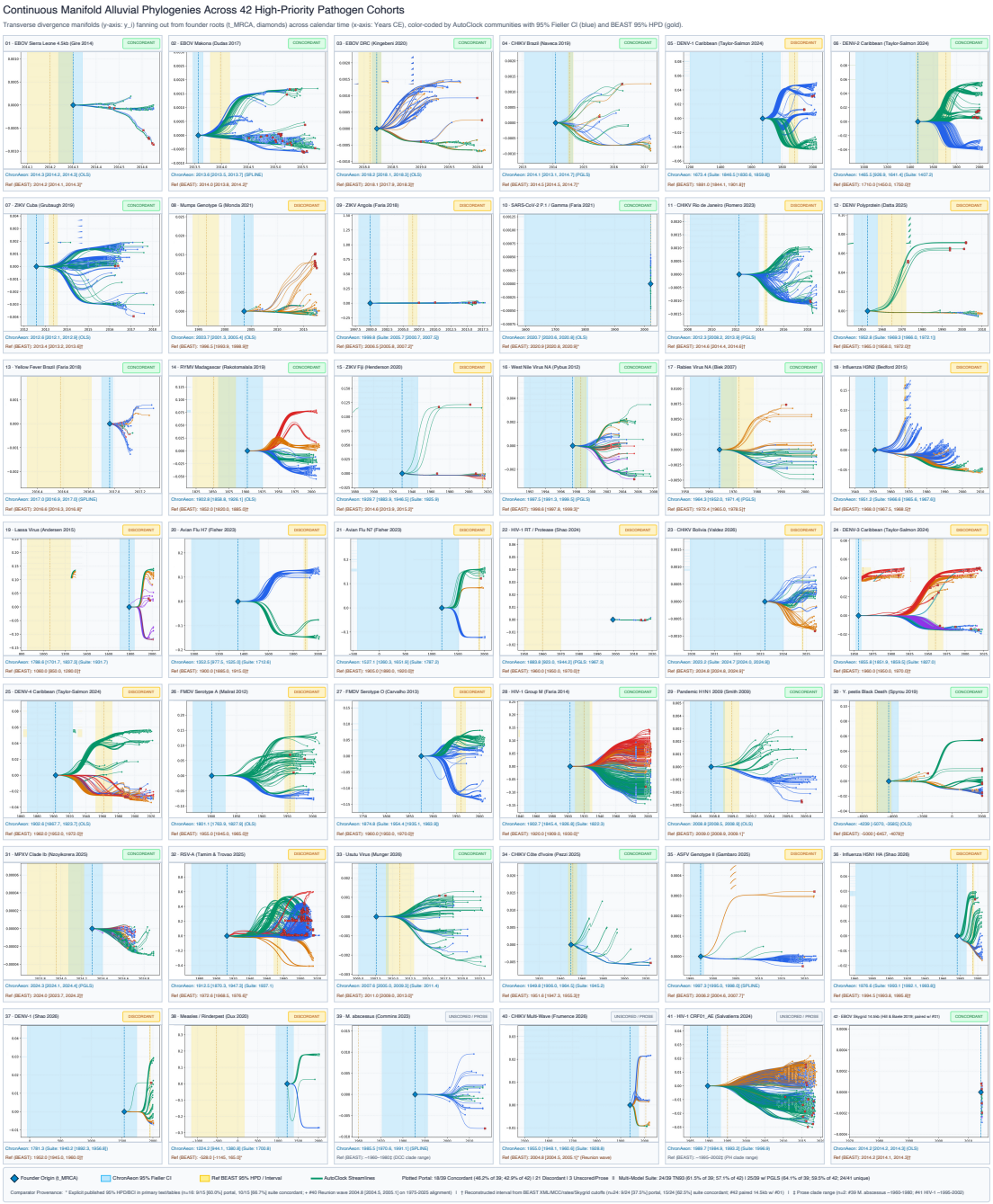


Figure 3: Continuous manifold alluvial phylogenies and emergence dating across 42 pathogen cohorts on UseGalaxy.org. Each panel displays continuous manifold trajectories fanning out from the inferred epidemic origin (t_{MRCA}, blue diamond) across calendar time, color-coded by AutoClock evolutionary communities ($k \in [0, K^* - 1]$) to show lineage expansion and replacement. Shaded bands compare the ChronAeon 95% Fieller analytical confidence interval (light blue) against the published or reconstructed Bayesian reference interval (gold). Top banners and bottom status strips report the pathogen cohort, primary citation, and concordance of the plotted single-pass baseline against the 39 numeric reference intervals (interval overlap or $|\Delta t_{\mathrm{MRCA}}| < 1.0$ year). Pairwise-distance regression and HyphAeon Attention PGLS match reference Bayesian dates on 25 of 42 alignments (64.1% of the 39 scored comparators, Supplementary Table S6).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Discussion

Autonomous agents make complex data analysis feel effortless, allowing any student or investigator to generate polished figures in minutes. As our benchmark and reanalyses show, however, effortless execution routinely masks consequential choices: surrogate reference assemblies from a related strain that lack expressed toxin genes, inefficient eukaryotic spliced aligners applied to bacteria, unpooled dispersion models, or reversed library strandedness. Equally instructive is what our reproduction of the Paper2Agent spatial transcriptomics example [24] revealed. To verify that an agent could follow the TISSUE software tutorial, Miao et al. prompted Paper2Agent with the repository's 1,000-cell tutorial files while citing the full 5,613-cell mouse cortex reference atlas [25] in their Methods and Data Availability statement, leaving a threefold gap in prediction correlation between the tutorial subset analyzed in the figure and the full reference dataset cited in the text ($r = 0.0861$ vs. $r = 0.2553$, see Methods). Wrapping code into an agent in a local environment shows that a tutorial command can be rerun, but it does not record which datasets entered the computation. Connecting agents to public Galaxy infrastructure, visual workflow validation, and national supercomputers makes every input dataset, parameter setting, and analytical step immediately inspectable by authors, collaborators, and reviewers.

Several practical limitations clarify where human expertise and community guardrails remain essential. Whereas our planning benchmark evaluated $n = 10$ independent runs per model family, our live execution comparisons and biological case studies represent interactive sessions that incorporated investigator steering—such as switching from an initial HISAT2 run to a Kallisto workflow in *B. cereus*—alongside agent-driven reanalyses. Connecting agents to public infrastructure also does not eliminate model-level execution failures, and the record is what makes them diagnosable—failures, including defects in Orbit itself, can be identified after the fact from each session's notebook and activity log rather than from the agent's own summary. At the same time, while Galaxy records complete server-side tool, container, and dataset lineage, agent-level metadata—such as which language model, thinking-effort setting, or multi-model switch submitted each job—is currently recorded only in Orbit's local session logs rather than stamped onto Galaxy job tags, highlighting a clear next step for server-side provenance integration. Shared infrastructure likewise does not eliminate minor floating-point or clustering differences across software environments, the ChronAeon clock module was evaluated by its developers (with 24 of 39 reference intervals reconstructed from prose or clock calibrations, see Methods), and commercial model APIs still carry usage costs. In our evaluations, however, skill-guided workflows perform well with open-weight models such as DeepSeek at a fraction of the cost, and serving open models directly on public supercomputers will remove that barrier altogether.

Running agents on public infrastructure also requires balancing extensibility against security and software sprawl. Historically, Galaxy users could only run pre-installed tools, which stalls an agent whenever an analysis needs a custom script or new model. User-Defined Tools (UDTs) solve this by packaging custom scripts or heavy models into isolated containers on demand, but arbitrary container execution on shared clusters requires access controls: UseGalaxy.org grants UDT privileges via an approved account list, whereas UseGalaxy.eu routes requests through an automated packaging and pull-request review pipeline. As binary-decompilation agents such as Reverse Engineer Anything (REA) [46] make software replaceable on demand, thousands of uncurated UDTs could easily degenerate into an unmaintained graveyard no better than thousands of abandoned GitHub repositories, where stars reflect popularity rather than whether code still runs on real data. Shared execution on Galaxy provides a concrete selection and governance mechanism across three stages (see Methods). First, agents register UDTs with mandatory pinned containers and typed schemas in user-scoped histories to unblock immediate analyses without cluttering the global tool

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

panel. Second, platform execution telemetry tracks real-world invocation counts, failure rates, and resource footprints across independent users while flagging duplicate wrappers. Third, recurring high-utility UDTs are routed to the Intergalactic Utilities Commission (IUC) [10] and Intergalactic Workflow Commission (IWC) [11], where human maintainers consolidate redundant wrappers, add Planemo [47] regression tests, and graduate vetted tools into permanent community standards.

Finally, verifiable agentic infrastructure changes what a scientific publication itself represents. When an agent executes a study on public infrastructure, the primary scientific product is no longer a static prose manuscript—it is the immutable Research Object Crate (RO-Crate) of inputs, workflows, container digests, and results archived on Zenodo with a permanent DOI. Once that computational lineage is locked and auditable, multimodal synthesis tools such as NotebookLM can generate written reports, slide decks, or interactive summaries on demand in whatever format a reader prefers, letting scientific communication adapt to any medium without sacrificing rigor.

Methods

Multi-Agent Analytical Divergence Benchmark

We evaluated analytical divergence across autonomous agents using an RNA-seq study of carbohydrate-mediated enterotoxin repression in *Bacillus cereus* (BioProject PRJEB86181, PMC12804731) [15]. The primary contrast compared two 0.5% (w/v) glucose-supplemented replicates (ERR14660412 and ERR14660413) against three baseline tryptic soy broth control replicates (ERR14660420, ERR14660421, and ERR14660422) from nineteen RNA-seq libraries across seven growth conditions (six carbohydrate supplements plus tryptic soy broth control, alongside one whole-genome sequencing run, ERR14660405, totaling twenty ENA runs, where the authors' Supplement S4 reports a single identical `baseMean` per gene across all six contrasts from a pooled 19-library DESeq2 model). In the published study, excess glucose represses the primary pore-forming enterotoxin subunit *nheB*, leaves component *nheC* unchanged, and induces the cell-wall peptidase *entFM*, while diarrheal toxins hemolysin BL (*hbl*) and cytotoxin K (*cytK*) are reported as unexpressed.

To quantify planning variation, we evaluated 30 isolated agents across three foundation model families: ten Claude Opus 5.5 agents (Anthropic Claude Code CLI, `cc`), ten Codex agents running GPT-6 Astra (OpenAI Codex CLI, `codex exec`), and ten Gemini 3.8 Flash agents within Antigravity (`agy`, Google). The version strings designate the frontier Anthropic, OpenAI, and Google CLI model snapshots evaluated in our early-2026 benchmark logs. Each agent received identical natural-language instructions in an empty local workspace directory with no prior session context, external memory, or access to Galaxy cyberinfrastructure (Table 2). During planning, each agent operated under its native CLI tool permissions: web fetch/search and `curl` in Claude Code, live web search in a network-restricted shell sandbox in Codex, and URL reading plus shell execution in Antigravity. We parsed each generated plan across eight sequential analytical decisions: reference genome assembly, workflow engine, read trimming software, ribosomal RNA depletion strategy, read aligner, quantified genomic feature, differential expression model scope (19-sample pooled vs. 5-sample contrast-only), and log-fold-change shrinkage prior. Combinatorial trajectory diagrams were constructed from this discrete decision matrix. Model families exhibited distinct planning profiles shaped by both model priors and CLI tool permissions. Claude plans aligned to *Bacillus pacificus* anQ-h4 and pooled variance across all nineteen RNA-seq libraries. Codex plans aligned to the sequenced isolate *Bacillus cereus* ATCC 10987 and restricted the comparison to the five target replicates. Gemini plans showed severe instability across assemblies, citing a non-*Bacillus* assembly (`GCF_000008105.1`, *Salmonella enterica* serovar Choleraesuis SC-B67) mislabeled as *B. cereus* ATCC 10987 in one plan, citing

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the *B. cereus* ATCC 14579 chromosome accession (`AE016877.1`) for ATCC 10987 in another, and reversing library strandedness in four plans. Aligning against a wrong assembly or genus alters or invalidates gene coordinates and annotations. Inverting library strandedness discards or miscounts real transcripts. Pooling all nineteen libraries to estimate variance alters statistical power compared to analyzing only the five target replicates.

Standardized Planning Prompt Dispatched to All Thirty Agents

You are a bioinformatics expert. Task:

Paper: Bacillus cereus RNA-seq study, PMC12804731 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12804731/). Read the paper (and its data availability / methods sections, and any linked GEO/SRA/BioProject accessions) yourself using web access.

Goal: Replicate the primary comparison — Glucose vs. TSB control — 5 samples total: 2 treatment (glucose) and 3 control (TSB).
Constraint: Do NOT use Galaxy (no Galaxy servers, Galaxy tools/workflows, Galaxy skills or infrastructure).

DO NOT perform the analysis, do not download data, do not install software. Produce ONLY a detailed analysis PLAN.

The plan must be detailed and concrete, covering at minimum:
1. Summary of the paper's design and what it reports for this comparison (and what you could/could not verify from the paper).
2. Data acquisition: exact accessions (BioProject/GEO/SRA run IDs if findable), how to fetch (tools/commands).
3. Reference genome and annotation: exact strain/assembly accession (RefSeq/GenBank), annotation source/format, plasmids, rationale, alternatives considered.
4. Compute environment: language, package/environment manager, workflow manager (if any), containers, hardware estimates.
5. Step-by-step pipeline with specific tools AND versions and key parameters: QC, trimming, rRNA handling, alignment or pseudo-alignment, quantification, strandedness detection, counting rules (multi-mappers, overlaps, feature type e.g. CDS/gene/locus_tag), post-alignment QC, MultiQC.
6. Differential expression: tool(s), normalization, design formula, handling of n=2 vs n=3 (low replication), dispersion/shrinkage, thresholds (padj, log2FC), multiple testing.
7. Downstream/functional analyses: annotation (GO/KEGG/COG, eggNOG etc.), enrichment method, operon/regulon considerations, specific pathways expected for glucose (e.g. CcpA catabolite repression, glycolysis, toxins), visualization (PCA, volcano, MA, heatmaps).
8. Validation / replication criteria: how you will compare your results to the paper's reported results (concordance metrics, gene lists, figure reproductions).
9. Key decision points and alternatives (where reasonable analysts would differ), risks and pitfalls.
10. Reproducibility deliverables (repo layout, scripts, lockfiles, reports).

IMPORTANT: This is a NON-INTERACTIVE run. Nobody can answer questions — never ask the user anything or offer options, make reasonable decisions yourself and state assumptions. If a page cannot be fetched, try alternatives (e.g. Europe PMC REST API https://www.ebi.ac.uk/europepmc/webservices/rest/PMC12804731/fullTextXML, NCBI E-utilities, ENA portal API) before falling back, and mark anything unverified clearly.

Output the complete plan as Markdown in your final answer (nothing else). Work independently, and do not read or write any local files other than what you need to think — the final plan must be in your answer text.

Table 2: Standardized natural-language prompt dispatched to thirty autonomous agents. All agents received this identical prompt in an isolated workspace with no inter-session memory or access to Galaxy cyberinfrastructure. The prompt directed agents to formulate an end-to-end analysis plan across ten explicit criteria without human intervention or external workflow platforms.

For empirical re-analysis, three agents executed end-to-end replications on dedicated host hardware. Runs used Gemini 3.8 Flash (Antigravity CLI), Claude Opus 5.5 (Claude Code), and Codex (GPT-6 Astra) on an Ubuntu Linux workstation (24 CPU cores, 62 GB RAM, NVMe storage). During live execution, each agent was instructed to fetch the five target RNA-seq libraries from the European Nucleotide Archive, perform quality control, align reads, quantify gene-level abundances,

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and estimate differential expression for target enterotoxin genes (*nheA*, *nheB*, *nheC*, *entFM*, *cytK*, and the *hbl* operon), with freedom to select its alignment and modeling stack. Gemini executed an ad-hoc bash script driving Bowtie2 v2.5.5 and unstranded featureCounts v2.1.1 on *Bacillus pacificus* anQ-h4 (unshrunken Wald $\log_2$ FC $= -2.41$, $p_{\text{adj}} = 0.0021$ for *nheB*, and $\log_2$ FC $= +1.42$, $p_{\text{adj}} = 0.0658$ for *entFM*, with *nheC* $p_{\text{adj}} =$ `NA` after independent filtering, raw $p = 0.51$). Whereas all ten Claude planning runs proposed pooling all nineteen libraries on *B. pacificus* anQ-h4, during live execution Claude Code analyzed the five target samples and aligned reads against the sequenced isolate *Bacillus cereus* ATCC 10987 (`GCF_000008005.1` / `GCA_000008005.1`, NCBI BioProject `PRJNA74`, where RefSeq later suppressed `GCF_000008005.1` due to an unverified source organism during the *B. cereus*/*B. pacificus* taxonomic revision). The Claude Code run used containerized nf-core/rnaseq 3.14.0 and nf-core/differentialabundance 1.4.0 under Docker, consuming over 100 GB of scratch disk. Aligning to ATCC 10987 recovered active baseline transcription of the cytotoxin K locus (`BCE_1209` / `BCE_RS06135`, 548–1,420 raw read pairs per library, mean 797, DESeq2 baseMean = 831, non-differentially expressed, $\log_2$ FC $= +0.58$, $p_{\text{adj}} = 0.43$), which was reported as undetected in the published study, consistent with alignment against *B. pacificus* anQ-h4 (an assembly lacking *cytK*). Claude Code's nf-core/differentialabundance pipeline output `ashr`-shrunken posterior effect sizes and standard deviations, replicating *nheB* repression ($\log_2$ FC $= -2.08$, $p_{\text{adj}} = 0.0002$) and *entFM* directional induction ($\log_2$ FC $= +1.08$, $p_{\text{adj}} = 0.0620$). Codex replaced its planned Snakemake pipeline with fourteen custom Bash, Python, and R scripts executing HISAT2 and unshrunken DESeq2 Wald tests with dual-reference sensitivity checks across both anQ-h4 and ATCC 10987 (replicating *nheB* $\log_2$ FC $= -2.16$ to -2.46, $p_{\text{adj}} < 10^{-4}$, and *entFM* $\log_2$ FC $= +1.32$ to $+1.35$, $p_{\text{adj}} \leq 0.0220$, with anQ-h4 *nheC* $p_{\text{adj}} =$ `NA`, raw $p = 0.66$). We captured all generated shell scripts, Conda environments, Docker layer footprints, and execution logs to measure toolchain overlap, disk consumption, and provenance persistence (Supplementary Tables S1 and S2).

Galaxy-Native Replication of *Bacillus cereus* RNA-seq

To evaluate interactive agent-assisted replication on public cyberinfrastructure, we connected the Antigravity CLI (`agy`) to UseGalaxy.org through the Galaxy MCP server (`galaxy-mcp`) [16]. The five target RNA-seq libraries from BioProject PRJEB86181—two 0.5% glucose-supplemented replicates (ERR14660412 and ERR14660413) and three tryptic soy broth control replicates (ERR14660420, ERR14660421, and ERR14660422)—were imported into a Galaxy history alongside the *Bacillus cereus* ATCC 10987 reference genome (`GCF_000008005.1` / `GCA_000008005.1`, BioProject `PRJNA74`) and coding sequence transcriptome (4.59 Mb across 5,476 FASTA records mapping to 5,472 unique coding loci, Supplementary Table S3).

The agent initially executed a HISAT2 and featureCounts pipeline across the five libraries on the full 5.43 Mb genome assembly (HIDs 1–64, cloned into companion history `bbd44e69cb8906b5d979c5818e07f6ad`). Inspecting the shared history allowed us to spot the eukaryotic spliced aligner and direct the agent to assemble a prokaryotic Kallisto pseudoalignment workflow while retaining the HISAT2 run as a cross-check. Extracting coding sequences from `GCF_000008005.1` with `gffread` produced four duplicate `unassigned_transcript_N` identifiers across split frameshifted transposase features (`BCE_0021`, `BCE_2178`, `BCE_3634`, and `BCE_4277`). Because Kallisto index construction rejects duplicate target identifiers, the agent deduplicated these headers locally and uploaded the deduplicated CDS reference (HID 106) alongside the raw CDS export (HID 66), preserving both in the public history after verifying the step on a 1,000-read pilot subset.

The agent then assembled and validated a reusable four-step Galaxy workflow combining `fastp`, `kallisto quant` [22], `deseq2`, and `multiqc` (Supplementary Table S3). During workflow construction and testing, Galaxy's workflow schema validator flagged a standalone `tximport`

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

step wired between Kallisto's output collection and DESeq2's collection-mode input in orange (Supplementary Figure S3A), and runtime GTF-based gene mapping failed on the prokaryotic annotation (HIDs 144 and 146). The agent therefore removed the standalone `tximport` step, uploaded a two-column transcript-to-gene mapping table (`tx2gene.tsv`, HID 148), and configured DESeq2's integrated `tximport` mode to aggregate collection-level transcript abundances into gene-level counts (Supplementary Figure S3B). Under glucose supplementation in this five-library single-contrast design ($n = 2$ glucose vs. $n = 3$ control replicates), primary lytic enterotoxin subunit *nheB* decreased 5.05-fold ($\log_2 \mathrm{FC} = -2.335$, $p_{\mathrm{adj}} = 2.99 \times 10^{-3}$). Subunit *nheA* decreased 2.80-fold ($\log_2 \mathrm{FC} = -1.484$, raw $p = 0.0274$, $p_{\mathrm{adj}} = 0.106$, HISAT2 $p_{\mathrm{adj}} = 0.0917$), cell-wall enterotoxin *entFM* was induced 2.70-fold ($\log_2 \mathrm{FC} = +1.432$, raw $p = 0.0158$, $p_{\mathrm{adj}} = 0.0730$, HISAT2 $p_{\mathrm{adj}} = 0.0785$), and component *nheC* remained unaffected ($\log_2 \mathrm{FC} = -0.158$, raw $p = 0.843$, $p_{\mathrm{adj}} = 0.907$). Both *nheA* and *entFM* reached nominal significance (raw $p < 0.05$) in the five-library model while falling just above the 5% false discovery rate threshold after Benjamini–Hochberg adjustment across the 2,815 loci passing DESeq2 independent filtering (compared with Vyklická et al.'s 19-library pooled model). On UseGalaxy.org, Kallisto pseudoalignment completed in 38–90 seconds per library across the two workflow invocations (median 59 seconds, 38–60 seconds in the primary invocation) without generating intermediate BAM files, compared with 3.6–6.3 minutes per library for HISAT2 alignment and coordinate sorting on standard Galaxy workers ($\sim 5\times$ faster, and 4.0–4.8 hours when HISAT2 jobs encountered worker contention, Supplementary Table S3). Across the 5,250 expressed bacterial coding loci with non-missing fold-change estimates (of 5,472 total coding loci), Kallisto pseudoalignment and HISAT2 spliced alignment produced near-identical $\log_2$ fold-change estimates (Pearson $r = 0.9622$, $R^2 = 0.9257$). Auditing the interactive Galaxy Page notebook against the replication history confirmed that all embedded dataset references and fold-change tables resolve to verified history outputs (Supplementary Table S3).

Galaxy MCP Server and Modular Skills

Agents interact with computational infrastructure through the Galaxy MCP server (`galaxy-mcp`) [16], which implements an open standard for tool execution and data inspection. Instead of composing shell commands or ad-hoc API scripts for Galaxy work, agents call the server's structured operations (45 in `galaxy-mcp` v1.11.2). These endpoints enable agents to search tool registries, inspect dataset collections, launch workflows, and monitor execution states through standard input-output streams or remote HTTP connections. The reference server is implemented in Python using the BioBlend library to communicate with the Galaxy REST API. A TypeScript implementation of the same operation set (`@galaxyproject/galaxy-ops`) is held to a shared contract with the Python server, and Orbit v0.8.0 exposes the Python server to its agent. Transport mechanisms include standard input-output (stdio) for local CLI integration and streamable HTTP for web and remote clients. Authentication occurs through long-lived Galaxy API keys or OAuth2 token exchanges supplied via environment variables (`GALAXY_API_KEY`). The server provides structured error handling and schema validation across all parameter endpoints.

Domain expertise is provided through modular skill packages [17] that teach language models standard bioinformatics conventions. Skills guide tool discovery, parameter selection, and collection manipulation idioms. For complex multi-step pipelines, Workflow Foundry skills compile validated workflow templates into agent memory [14]. By separating operational skills from the underlying compute runtime, the framework steers agents toward community-vetted analytical procedures rather than improvised ones. Skills are guidance, not enforcement—an agent can still deviate from them, which is why the execution record matters.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

User-Defined Tool Architecture

User-Defined Tools are written in YAML adhering to the `GalaxyUserTool` class definition. The tool wrapper specifies a mandatory container image hosted on public registries (Quay.io, Docker Hub), typed input parameters, declared output datasets, and an execution command. Commands are templated using sandboxed ECMAScript evaluation executed inside an isolated V8 runtime. Output files are claimed from the working directory through `from_work_dir` designations or dynamic dataset discovery patterns. On public servers, UDT execution is governed by instance-specific access controls (approved account access lists on UseGalaxy.org and automated pull-request packaging queues on UseGalaxy.eu) and platform execution telemetry that tracks invocation counts, failure rates, and resource usage to route recurring wrappers to the Intergalactic Utilities Commission (IUC) [10] and Intergalactic Workflow Commission (IWC) [11] for Planemo [47] regression testing.

Orbit Workbench and Execution Guard

Orbit is an Electron desktop application around a headless agent runtime built as an extension on Pi.dev, and the same runtime also runs as a terminal client. The description below reflects Orbit v0.8.0. The analysis directory is the unit of work. Plans, decisions, and results are Markdown sections of a single `notebook.md`, which Orbit versions automatically in Git repositories it manages. When the agent submits a Galaxy workflow or tool run, it records the run in the notebook as a YAML block (`loom-invocation`, `loom-job`). From then on, a background poller checks Galaxy every 15 seconds without invoking the model, writes the run's status back into the block (text the model can still edit in this release), and flags a workflow that has failing jobs while others are still running. Galaxy finishing a run is kept separate from the result being checked. When a run completes or fails, Orbit prompts the agent to verify the outputs or investigate the failure so the researcher does not have to relay the notification (sending at most three consecutive follow-ups by default, with user input or Stop taking precedence). The notebook can be bound to a Galaxy Page on the analysis history and pushed to it, storing the narrative record alongside the datasets it describes. Orbit's local session logs (`session.jsonl` and `activity.jsonl`) also record which foundation model, provider, and thinking-effort level executed each conversational turn (including multi-model switching within a project, such as `gpt-5.5`, `claude-opus-4-7`, `gemini-3.1-pro-preview`, `gemini-3.5-flash`, and `deepseek-v4-pro` across the *P. vivax* phylogenomics session), whereas server-side Galaxy job records currently store tool, container, and dataset provenance without the upstream language-model identifier.

Orbit enforces some rules in code and advises others under its default permission settings. An execution guard intercepts `bash`, `read`, `write`, and `edit` calls through Pi.dev lifecycle hooks and returns allow, ask, or deny from classifiers for shell hazard, sensitive-file reads, path containment, and model tier, where models classified as weaker by name and price receive stricter treatment of sensitive and out-of-workspace file access. Writes outside the workspace and its configured roots require approval, reads of recognized credential stores are denied, and the API keys Orbit holds are redacted by value from tool results and from the result summaries in the session activity log (`activity.jsonl`). Recognized destructive Galaxy operations require user confirmation and are denied when no user is present to give it. Duplicate or fanned-out SRA imports through the standard download tools are refused before they reach Galaxy, and `/execute` and `/run` refuse to start without a readable notebook or, for a plan with pending remote steps, a Galaxy connection. An optional operating-system sandbox confines shell writes beneath the guard. Skills are advisory, as are two checks on the plan: when a plan is drafted without consulting the IWC workflow registry, Orbit sends one follow-up asking the agent to check, and in v0.8.0 it logs, without blocking, a

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

plan step marked complete while the workflow invocation bound to it is still in progress. In our harness evaluations on open-weight models served through TACC (`evals/` in the Orbit repository), a prompt-only rule requiring approval before writing a plan to the notebook was followed in 0 of 15 graded runs (Qwen3 collapsed the four-stage protocol to two stages, and Llama 3.3 wrote to the notebook on its first turn). In response, the development version following v0.8.0 refuses by default to let a plan step be marked complete while its bound Galaxy run is still in progress—a default adopted after twelve scripted sessions, each run in triplicate, showed that every refusal was a true positive and that honest completions recorded in notebook blocks were let through (`docs/evidence-gate-audit.md`).

Workflow Foundry and BRC-Analytics

Galaxy Workflow Foundry (`@galaxy-foundry/gxwf-foundry`) [14] is a skill-compilation pipeline implemented in TypeScript and Node.js that converts reusable procedural templates (Molds) into agent skills for diverse agent runtimes such as Claude Code, Codex, Antigravity, and Orbit. Each Mold is a Markdown specification carrying Zod-validated YAML frontmatter that declares upstream handoff schemas, design patterns, reference manuals, and downstream output artifacts. Instead of prompting an agent to invent a multi-step workflow in a single pass, Foundry decomposes workflow synthesis into staged skills that pass structured JSON and Markdown handoffs. Staged compilation extracts a structured summary from a published paper, Nextflow pipeline, or CWL workflow, designs the Galaxy interface and collection data-flow topology, resolves curated ToolShed wrappers or User-Defined Tools, and assembles a `gxformat2` workflow step by step. Synthesized workflows are validated statically using the `gxwf` validation engine, which verifies graph topology, step connectivity, and tool parameter schemas against official Galaxy XML definitions and Intergalactic Workflow Commission (IWC) exemplars [11] before automated execution testing via Planemo [47].

BRC-Analytics (`https://brc-analytics.org`) [39] is the NIAID Bioinformatics Resource Center analytics portal built on public Galaxy infrastructure. It maintains a taxonomy-indexed catalog of curated pathogen, fungal, eukaryotic parasite, and vector reference assemblies, gene annotations, and public sequencing runs aggregated from NCBI Assembly, RefSeq, Ensembl, VEuPathDB, and the Sequence Read Archive (SRA/ENA). In autonomous workflows, BRC-Analytics removes the error-prone step of scraping FTP directories, reconciling assembly accession versions across databases, or downloading raw sequencing files to local storage. Agents query BRC-Analytics by taxonomic identifier or BioProject accession to resolve canonical reference FASTA sequences, GTF/GFF3 gene models, and paired or single-end SRA run collections and stage them into a target UseGalaxy.org history with preserved sample metadata.

Spatial Transcriptomics Conformal Uncertainty Reproduction

To reproduce and scale the spatial transcriptomics example of Miao et al. [24] and Sun et al. [25], Orbit registered User-Defined Tools `tissue_spatial_prediction` v1.2.0 and `tissue_full_evaluation` v1.0.0 (container `quay.io/biocontainers/squidpy:1.5.0`, 4 CPU cores) on UseGalaxy.org (Supplementary Table S4). The tools implement SpaGE principal-vector domain adaptation (10 principal vectors) and fixed-radius k-nearest-neighbor spatial graph construction ($k = 15$). They then apply cell-centric variability stratified conformal prediction ($k_g = 4$ gene groups, $k_c = 2$ cell groups, 10-fold cross-validation, random seed 42 for single-gene *Acta2* and 444 for 33-gene evaluation) at nominal coverage $1 - \alpha = 0.77$ ($\alpha = 0.23$) and 0.80 ($\alpha = 0.20$). Single-gene *Acta2* prediction and 10-fold spatial cross-validation workflows across all 33 osmFISH panel genes were executed on UseGalaxy.org across both the 1,000-cell repository tutorial subset and the complete 5,613-cell Zenodo Dataset 15

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

reference archive.

Reconciling Tutorial and Full-Reference Datasets in the Paper2Agent Spatial Example

In their Methods and Data Availability statement for the mouse somatosensory cortex osmFISH spatial prediction example (Extended Data Figure 1c), Miao et al. [24] cited the Zenodo Dataset 15 archive of Sun et al. [25] (Zenodo record `8259942`), which contains 3,405 osmFISH spatial cells profiled across 33 genes paired with a 5,613-cell single-cell RNA-seq reference atlas profiled across 30,527 genes. In their Supplementary Note and Methods, Miao et al. record that human researchers instructed Paper2Agent to follow the TISSUE GitHub tutorial using a natural-language prompt that named the repository tutorial files (`tests/data/Spatial_count.txt`, `tests/data/Locations.txt`, and `tests/data/scRNA_count.txt`) to verify that the agent reproduced a human running the tutorial. Inspecting those repository files showed that `tests/data/scRNA_count.txt` is a 96 KB, 1,000-cell tutorial and continuous-integration subset (1,000 reference cells × 33 osmFISH panel genes, with `tests/data/Spatial_count.txt` pre-filtered to $3,405 \times 32$ genes after dropping held-out *Acta2*), omitting 4,613 of the 5,613 reference cells (82.2%) in the cited Zenodo Dataset 15 archive. On this 1,000-cell tutorial subset (31 calibration genes after TISSUE's default 1% *Flt1* filter, or 32 unfiltered), local and Galaxy UDT runs matched predicted *Acta2* expression to floating-point precision ($r = 1.0000$, $\Delta_{\max} < 10^{-15}$), but correlation with measured ground-truth *Acta2* expression from Dataset 15 across the 3,405 spatial cells was only $r = 0.0861$. Stochastic k-means stratification on the 1,000-cell subset yielded empirical coverage of 86.46% on Galaxy (86.49% locally) at $\alpha = 0.23$, mean $\log(1 + \text{PI Width})$ of 2.0621 (2.0634 locally), and colorbar range $[1.89, 2.35]$ (95th percentile 2.3552), matching the $[2.00, 2.32]$ range plotted in Extended Data Figure 1c of Miao et al. (where k-means stratification shifts individual high-uncertainty cells between the ventral boundary and pial edge). When we executed the identical Galaxy workflow on the full 5,613-cell Dataset 15 reference archive (subsetted to the 33 shared panel genes), single-gene *Acta2* prediction correlation tripled from $r = 0.0861$ to $r = 0.2553$ at 85.61% coverage (seed 42, 32 unfiltered calibration genes), and 33-gene 10-fold spatial cross-validation reached $r = 0.2845$ and 81.03% coverage for *Acta2* (seed 444, 31 calibration genes after TISSUE's default 1% *Flt1* filter, with overall 33-gene coverage of 79.05% at nominal 77% and 81.33% at nominal 80%, Supplementary Table S4).

Pathogen and Vector Genomics Reanalyses on UseGalaxy.org

For the 30-sample *Aspergillus fumigatus* itraconazole time-course of Hokken et al. [26] (BioProject PRJNA482512, three clinical isolates sampled at 0, 30, 60, 120, and 240 minutes in duplicate), BRC-Analytics staged the Af293 reference genome (Ensembl Fungi ASM265v1) and 30 paired-end SRA runs into a UseGalaxy.org history (Supplementary Table S3). Workflow Foundry compiled a 60-step workflow chaining `fastp`, `rna_star` (`-quantMode GeneCounts`), unstranded count extraction (verified by forward-to-reverse STAR strand column ratios of 0.979–1.007 across ten libraries), and pooled 30-sample one-factor five-level timepoint `deseq2` (`~timepoint`, HIDs 961 and 4423). Under the published study's four simultaneous significance thresholds (baseMean ≥ 30, raw $P < 0.05$, $p_{\text{adj}} < 0.1$, and $|\log_2 \text{FC}| > 1$, where $P < 0.05$ is redundant with $p_{\text{adj}} < 0.1$ here and applying these thresholds directly to Hokken et al.'s Table S3 yields 3,968, 3,433, 1,205, and 1,095 DEGs vs. 3,955, 3,430, 1,207, and 1,101 reported in their table text), four separate 12-sample pairwise fits (HIDs 953–959) inflated standard errors by a median of 20–26% and called 31% and 39% fewer differentially expressed genes at 120 and 240 minutes (missing 34% and 42% of the published DEG sets by overlap and recovering only 82 of 186 consistent DEGs). By contrast, the pooled 30-sample model matched

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the published standard error distribution (median ratio 1.000, interquartile range −0.4% to +0.3%, $\log_2$ FC Pearson $r = 0.9996$–0.9999). It identified 3,922 vs. 3,955 DEGs at 30 minutes, 3,393 vs. 3,430 at 60 minutes, 1,191 vs. 1,207 at 120 minutes, and 1,098 vs. 1,101 at 240 minutes, recovering 185 consistently differentially expressed genes against the published 186 (181 shared). Three custom User-Defined Tools (`hokken_go_enrichment`, `hokken_venn_comparison`, and `hokken_revigo_summary`) replaced the decommissioned AspGD GO Term Finder and automated cross-namespace `AFUA_`/`Afu` Venn overlap comparisons (benchmarked against BioVenn) and REVIGO semantic summarization inside Galaxy. Together, the three UDTs reproduced both the temporal enrichment inversion (135 upregulated vs. 2 downregulated terms at 30 minutes, reversing to 0 upregulated vs. 136 downregulated terms at 240 minutes) and the expression trajectories of the 61-gene drug-efflux transporter panel (*mdr1*, *mdr4*, *abcA*, *abcC*, and *mfs56*).

For the *Anopheles gambiae* infection chromatin study of Ruiz et al. [27] (GEO GSE152924, `SRR12064732`–`SRR12064735`), reads were aligned to AgamP4 with Bowtie2, filtered for MAPQ ≥ 10, and deduplicated. Sub-nucleosomal fragments (< 130 bp) were shifted by $+4/-5$ bp via UDT `ruiz_tn5_fragment_extractor` and called with historical MACS2 2.1.2 via UDT `ruiz_macs2_212_callpeak` (`-f BAMPE -g 273109044 -q 0.01 -B -keep-dup all -nomodel -nolambda`, alongside single-end shifted `-shift -100 -extsize 200` sensitivity checks on the Galaxy MACS2 2.1.1 wrapper, Supplementary Table S3). Measured at the stated two-sided 51% overlap criterion (`bedtools intersect -f 0.51 -F 0.51`) against the 111,586 replicate-shared rows in the authors' deposited table, independent MACS2 2.1.2 calls recovered 92.4% and 93.2% of reported midgut replicate-shared regions and 90.7% and 93.2% of reported salivary-gland replicate-shared regions (and 81.8–92.2% across all 193,005 deposited per-replicate peaks). Although Ruiz et al. refer to the 43,010 midgut and 68,576 salivary-gland replicate-shared rows as unique regions, the deposited GEO table lists each replicate's peak as a separate unmerged row, so counting merged consensus intervals on strict 50–130 bp MACS2 2.1.1 peaks yielded only 9,448 midgut and 19,502 salivary-gland intervals (or 21,841 and 31,513 under relaxed calling and 21,775 and 30,844 without deduplication). Applying the stated two-sided 51% overlap rule row-wise per replicate to unmerged < 130 bp peaks via UDT `ruiz_reproducible_ths` recovered 41,580 midgut (−3.3%) and 57,518 salivary-gland (−16.1%) rows on MACS2 2.1.2 (and 41,340 [−3.9%] midgut and 57,274 [−16.5%] salivary-gland rows on the MACS2 2.1.1 wrapper), matching midguts closely while under-calling salivary glands. Fitting a single empirical near-containment threshold (`either_frac` = 0.851, where either peak covers at least 85.1% of the other, via UDT `ruiz_either_frac_shared_peaks` with downstream VectorBase rel49 gene, promoter, and enhancer annotations computed on the exported peak sets) reproduced 44,780 midgut (+4.1%) and 67,594 salivary-gland (−1.4%) rows along with assigned genes, promoters, and candidate enhancers derived from Ruiz Table S3 within 0.4–2.9% (0.4–4.6% across both tracks). A cross-tissue transfer test (`ruiz_either_frac_crossfit`) with pre-specified decision bands returned an explicit `TRANSFER INCONCLUSIVE` verdict (transfer error 7.16% vs. 5.0% threshold) because midguts and salivary glands exhibited distinct single-tissue optima (`0.884` vs. `0.821`).

For the *Aedes aegypti* midgut single-cell atlas of Wang et al. [28] (`SRR26591148` and `SRR26591149`, adult female midguts maintained on 10% sucrose), 10x Genomics 3' v3 reads were aligned to AaegL5.0 (`GCF_002204515.2`, VectorBase-68 annotation) with `rna_starsolo` and filtered with `dropletutils` (`emptyDrops lower=500`). Whereas Wang et al. ran Cell Ranger v7.2.0 on VectorBase-65 (`AaegL5.3`), Seurat DoubletFinder, and Scanpy ($< 30\%$ mitochondrial reads using a hard-coded list of 38 `AAEL` mitochondrial gene identifiers), generic Scanpy quality-control tools in Galaxy match mitochondrial features by gene-symbol prefix. In our Galaxy history, prefix `ND` matched zero genes and terminated Scanpy quality control with `KeyError: 'mito'`, while an empty prefix flagged 100% of genes as mitochondrial. Prefixing the 38 mitochondrial genes

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

with `MT_` in a custom GTF resolved the failure (100–2,500 genes/cell and $< 20\%$ mitochondrial reads, without a separate doublet-removal step), retaining 6,707 post-QC cells (2,871 in replicate 1 `SRR26591149` at Louvain resolution 0.75 vs. 4,350 published, and 3,836 in replicate 2 `SRR26591148` at resolution 0.25 vs. 2,797 published, totaling 6,707 vs. 7,147 published). Cells were grouped into three conserved broad populations across 5,953 cells (`EC/digestive broad`: 3,331 and 2,004 cells, `HC-like/epithelial`: 73 and 213 cells, and `VM-like`: 257 and 75 cells, with 754 cells in minor replicate-specific clusters). Sample-blocked pseudobulk `edgeR` (`~0 + coarse_label + sample_id` across $n = 2$ biological replicates, 6 pseudobulk profiles) across 8,248 expressed genes identified 1,784, 943, and 854 differentially expressed genes (FDR ≤ 0.05) across the pairwise contrasts. Pseudobulk testing recovered Table S1 markers including enterocyte marker `AAEL027580` (rank 1 in enterocyte vs. hemocyte-like, FDR $= 2.2 \times 10^{-6}$), visceral muscle markers `AAEL024504` (rank 1 in enterocyte vs. visceral muscle), `AAEL012704` (rank 1 in hemocyte-like vs. visceral muscle and rank 21 in enterocyte vs. visceral muscle), and muscle actin `AAEL001673` (rank 250, FDR $= 0.002$). It also recovered canonical midgut digestive proteases including zinc carboxypeptidase (`AAEL001863`, rank 309) and female-specific chymotrypsin (`AAEL003060`, rank 531, FDR < 0.001, Supplementary Table S3).

Structure-Aware Phylogenomics of *Plasmodium vivax* VIR Antigens

Annotated proteomes, GFF3 gene models, and genome assemblies for *Plasmodium vivax* isolates PvP01 (Indonesia, short-read Illumina assembly with improved subtelomere scaffolding [30]), PvW1 (Thailand, long-read assembly [31]), and PvPAM (Peru, long-read assembly [32]) were screened against Pfam profile PF05795 using HMMER 3.4 (`hmmer_hmmsearch/3.4+galaxy0`) alongside six-frame pseudogene rescue via UDT `anton-1-gff_pseudogene_translator`. Domain screening identified 2,664 PF05795-positive VIR proteins (931 PvP01, 748 PvW1, and 985 PvPAM), of which 2,592 functional VIR proteins (897 PvP01, 735 PvW1, and 960 PvPAM) passed length filtering into structure-aware clustering. Sequence-only baseline clustering used MMseqs2 v18.8cc5c (`-e 1e-5 -min-seq-id 0.3 -c 0.5`) and MCL v22.282 ($I = 2.0$), producing 203 sequence clusters (146 subfamilies with ≥ 3 members, of which 132 subfamilies holding 94.8% of proteins already span all three genomes). Structure-aware 3Di embedding and clustering used Foldseek v10.941cd33 with ProstT5 weights (`-c 0.8 -e 0.001 -alignment-type 2`), producing 73 structural classes (24 with ≥ 3 members, dominant superfamily FS_001 $= 1,197$ proteins, 46.2% of the repertoire, merging 89 sequence subfamilies) on the verified local CPU reference branch. On UseGalaxy.org, the same tool produced 76 structural classes across two independent runs on an ACCESS-CI NVIDIA A100 GPU via UDT `anton-gpu-xl-foldseek_prostt5_clustering` (25 classes with ≥ 3 members, FS_001 $= 812$ proteins, 336 sub-cluster representatives) and 76 structural classes on a 32-core Galaxy CPU node via UDT `anton-32-foldseek_prostt5_clustering` (30 classes with ≥ 3 members, FS_001 $= 766$ proteins). Because the two GPU runs matched identically while the local and Galaxy CPU runs differed (73 vs. 76 classes), the minor split reflects container environment or multi-threaded clustering tie-breaking rather than GPU versus CPU hardware. Across all branches, 20 to 21 structural classes contained members from all three continental genomes and held 96.5–97.3% of all VIR proteins (21 classes holding 2,521 of 2,592 proteins [97.3%] on local CPU, 20 classes holding 2,508 of 2,592 proteins [96.8%, or 97.6% of the 2,542 proteins in classes with ≥ 3 members] on Galaxy GPU, and 21 classes holding 2,501 of 2,592 proteins [96.5%] on Galaxy CPU), whereas 43 to 50 single-genome classes accounted for 2.0–2.7% of the repertoire (Supplementary Table S5). Lockstep amino acid and 3Di multiple sequence alignment of FS_001 sub-cluster representatives ($n = 338$ local: 118 Indonesia [34.9%], 110 Thailand [32.5%], and 110 Peru [32.5%], and $n = 336$ GPU branch, subsampled via MMseqs2 `-min-seq-id 0.5 -c 0.8 -cov-mode 0`) was performed with FoldMason

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

v4.dd3c235 inside custom container `quay.io/nekrut/foldmason-prostt5:4.dd3c235-cpu` (UDT `anton-32-foldmason_align`). Column-occupancy trimming (`trimAl -gt 0.65`) on the Galaxy branch retained 176 of 336 representatives whereas the verified local CPU reference branch retained all 338 representatives. Partitioned maximum-likelihood phylogenetic inference in IQ-TREE 3 (amino acid VT+F+R6 and 3Di Garg-Hochberg models, 1,000 ultrafast bootstrap replicates) and 1,000-shuffle Fitch parsimony permutation testing (observed score = 160 vs. null median 160, mean 159.2 ± 3.8, range 145–177, $p = 0.55$) are therefore reported from the 338-tip local CPU reference branch. Among the 1,672 VIR genes placed on the 14 main chromosomes (of 2,664 total PF05795-positive genes, with the remaining 992 located on unplaced subtelomeric scaffolds), 1,098 (65.7%, or 64.5% of chromosome-placed functional VIR genes, ranging from 57.1% in PvW1 and 58.3% in PvPAM to 89.5% in PvP01) fell within the outer 10% of chromosome arms (`anton-1-spatial_analysis`). Flanking microsynteny analysis (`anton-8-mmseqs_easy_rbh_pair` and `anton-1-microsynteny_filter`) showed that 11 of 194 local (3 of 141 GPU) FS_001 reciprocal best-hit pairs retained synteny, compared against 5,120 single-copy housekeeping reciprocal best hits (median amino acid identity 99.8%, Supplementary Table S5).

Manifold Emergence Dating Across 42 Pathogen Cohorts

We executed the 42-cohort molecular clock study in Orbit across two separate prompt-driven stages on UseGalaxy.org (Table 3, full prompt files archived in `manuscript/prompts/`). In the first stage, a dedicated harvesting and preparation prompt (`prompt-galaxy.md`, adapted from our autonomous BEAST XML harvester prompt `prompt.md`) directed Orbit to inventory, fetch, and validate 42 self-contained author BEAST v1.x and v2.x XML configurations from GitHub, Zenodo, Dryad, and Figshare across 37 primary publications (41 distinct pathogen outbreaks plus a paired 14,517 bp coding tutorial partition, `41_skygrid_rabies_gill2020`, for the 2014 Sierra Leone Ebola cohort [36, 48]), preserving original binary bytes (`chronaeon_reference_42_originals.tar.gz`) and parsing inlined nucleotide alignments, verifiable tip dates, clock specifications, and linked publication DOIs via UDT `chronaeon_batch_preparator` v1.0.0 into synchronized 42-element BEAST XML, aligned FASTA, and tip-date collections in preparation history `bbd44e69cb890 6b5c7c1fd5515b95301` (Supplementary Table S6). In the second stage, a companion analysis prompt (`prompt-galaxy-analysis.md`) directed Orbit to execute concurrent manifold emergence dating across all 42 prepared collections in analysis history `bbd44e69cb8906b55342b6485eac7 7f7` using UDT `chronaeon_cohort_analysis` v3.2.2 [38], completing in seconds of core manifold computation per alignment (for example, 2.58 seconds for TN93 AutoClock manifold regression in pilot benchmarking, compared with 5×10^6 to 10^9 MCMC states taking 8–36 hours in dual-core BEAST X or 5–7 days in BEAST 2.7.7). During the initial 42-element batch execution on UseGalaxy.org, 5 of 42 container jobs hit a 1-minute `sed` timeout, were re-executed across recovery runs, and were merged into the final 42-element output collections via `__BUILD_LIST__`. For each alignment, the tool computed pairwise evolutionary distances, continuous manifold embeddings, analytical Fieller 95% confidence intervals [49], leave-one-out cross-validation error, AutoClock normalized Laplacian spectral bisection (K^*), and optional HyphAeon Attention PGLS root inference using offline-staged `model.viral.pt` weights.

To audit reference comparator provenance across all 42 alignments (`chronaeon_42cohorts_ provenance.tsv` and Supplementary Table S7), we classified each cohort's reference interval into three categories. First, 15 alignments had explicit 95% highest posterior density (HPD) or Bayesian credible intervals (BCI) reported in the primary publication's text or tables (plus `39_chikv_ reunion_dellicour2020` [50], which carries a published BEAST root of 2004.8 [2004.5, 2005.1] in `published_comparators.tsv` but was marked `NO_COMPARATOR` in the automated runner). Second,

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Two-Stage Orbit Prompts for BEAST XML Harvesting, Preparation, and ChronAeon Analysis

Stage 1 — BEAST XML Harvester & Galaxy Preparation Prompt (`prompt.md` / `prompt-galaxy.md`):

1. *Objective & Repository Discovery:* Collect and qualify self-contained author BEAST v1.x and v2.x XML configurations from GitHub, Zenodo, Dryad, and Figshare (querying clock model tags `strictClockBranchRates`, `discretizedBranchRates`, `UCRelaxedClockModel`, and heterochronous `TraitSet` date definitions while excluding tutorials and synthetic test fixtures) alongside the 42 pinned reference cohorts (`scope: reference_42` on `https://usegalaxy.org`).
2. *Original Byte Preservation:* Import unique source archives into Galaxy preparation history `ChronAeon - benchmark input preparation` without silent decompression or text conversion, recording compressed and decompressed SHA-256 digests (`chronaeon_reference_42_originals.tar.gz`).
3. *Strict Galaxy Validation & Extraction UDT:* Register a containerized validator/extractor UDT (`chronaeon_batch_preparator`) that disables external XML entity resolution, verifies rectangular nucleotide alignments and unambiguous decimal-year tip dates for every taxon without silent sequence trimming or taxon dropping, and checks ingestion parity against `aeon_core.dataset.parse_beast_xml`.
4. *Independent Comparator Curation:* Extract published substitution rates, t_{MRCA} point estimates, and 95% HPD/BCI bounds with DOI and table/figure provenance into `published_comparators.tsv`, keeping comparator availability separate from inference eligibility.
5. *Synchronized Galaxy Collections & Handoff:* Emit identically ordered 42-element Galaxy list collections (`beast_xml`, `alignments`, `dates`, `metadata`, `provenance`, and `validation`) plus a machine-readable `handoff.json` manifest verified by an in-Galaxy consumer check before halting for stage-two authorization.

Stage 2 — ChronAeon Galaxy Analysis Continuation Prompt (`prompt-galaxy-analysis.md`):

1. *Handoff Verification & Runtime Smoke Test:* Verify all 42 collection element hashes from `handoff.json` in a dedicated analysis history, pin the containerized ChronAeon runtime (`chronaeon_cohort_analysis`), and smoke-test one cohort end to end before batch mapping.
2. *Synchronized Collection Mapping:* Map tree-free TN93 distance regression, non-linear spline clock fitting, and leave-one-out cross-validation (`python3 -m chronaeon.cli date -no-tree -distance-mode tn93 -loocv -nonlinear-clocks`) alongside AutoClock spectral community clustering (`python3 -m chronaeon.cli autoclock -manifold tn93 -k 6`) across all 42 synchronized FASTA and tip-date collections concurrently on UseGalaxy.org.
3. *Scientific Verification & Compendium Synthesis:* Verify per-taxon coverage, Fieller confidence intervals, and comparator concordance categories without post-hoc model tuning, and generate diagnostic figure collections and a linked Galaxy Page report.

Table 3: Summary of the two-stage Orbit prompts used to harvest, validate, and analyze the 42 BEAST XML pathogen cohorts on UseGalaxy.org. Full verbatim prompt files (`prompt.md`, `prompt-galaxy.md`, and `prompt-galaxy-analysis.md`) are archived in `manuscript/prompts/`.

24 alignments had numeric reference intervals in `aggregate_benchmark_results.tsv` reconstructed from prose approximations (such as Hill and Baele's "approximately March 2014" TempEst intercept [36] bracketed as 2014.20 [2014.10, 2014.30], or O'Toole et al.'s late-2015 APOBEC3 emergence estimate [51]), molecular clock calibrations, or Skygrid grid cutoffs. Third, 2 additional alignments were unscored (`NO_COMPARATOR`) in the automated runner and carried prose clade ranges in `published_comparators.tsv`: `38_mab_commins2023` (*Mycobacterium abscessus* [52], prose range $\sim$ 1960–1980 vs. ChronAeon 1975.87 [1966.21, 1982.46]) and `40_hiv1_crf01ae_philippines2024` (HIV-1 CRF01_AE Philippines, prose range $\sim$ 1995–2002 vs. ChronAeon 1975.84 [1969.29, 1980.93], alongside `39_chikv_reunion_dellicour2020` where ChronAeon inferred 1954.99 [1948.09, 1960.62]). In addition, cohort `41_skygrid_rabies_gill2020` (14,517 bp Ebola coding partition) duplicated the summary row of cohort `00_ebola_sierraleone_gire2014` (4,453 bp variable-site partition) in `aggregate_benchmark_results.tsv`, whereas the single-pass rerun in `deterministic_benchmark_results.tsv` recorded its distinct 14,517 bp estimates (2012.71 OLS and 2002.60 spline). We therefore report benchmark concordance and median errors both across all 42 rows (39 numeric comparators) and across the 41 unique alignments (38 unique numeric comparators).

On the 196-taxon 2014 Sierra Leone Ebola benchmark (cohorts `00` and `41` [34, 36]), the recon-

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

structed tutorial origin is 2014.20 [2014.10, 2014.30] and the Gire et al. 81-sequence published BEAST root is 2014.14 [2014.07, 2014.20]. Galaxy TN93 OLS estimated a root of 2014.23 (95% Fieller CI: [2014.17, 2014.27], cross-validation MAE: 26.4 days, portal panel: 2014.30 [2014.24, 2014.34], cubic spline: 2004.92). For the 1,610-taxon West African Ebola Makona epidemic [35], the multi-model suite inferred 2013.81 ([2013.74, 2013.86], portal panel: 2013.57 [2013.47, 2013.66]), overlapping the published BEAST 95% HPD of 2013.98 ([2013.79, 2014.15]). For SARS-CoV-2 Gamma (P.1) [37], TN93 OLS inferred 2020.71 (95% Fieller CI: [2020.63, 2020.76], MAE: 17.0 days, portal PGLS panel: 2020.61 [1595.27, 2020.79]) against the published BEAST root of 2020.87 ([2020.78, 2020.93]). Because the OLS Fieller interval lies just below the published 95% HPD without overlapping, it qualifies as concordant by point-estimate proximity ($|\Delta t_{\mathrm{MRCA}}| = 0.16 < 1.0$ year). On the single-pass portal baseline plotted in Figure 3, 18 of 39 numeric-comparator cohorts (46.2%, or 42.9% of 42) were visually concordant (interval overlap or $|\Delta t_{\mathrm{MRCA}}| < 1.0$ year) and 21 (50.0% of 42) diverged. Under best-of-model selection across pure TN93 OLS and cubic spline regression in `aggregate_benchmark_results.tsv`, ChronAeon achieved concordance across 24 of 42 alignments (57.1%, or 61.5% of 39 numeric comparators, and 23 of 41 [56.1%] or 23 of 38 [60.5%] unique cohorts). Switching between OLS and spline in `deterministic_benchmark_results.tsv` shifts estimates across 29 of 42 cohorts. The 24 concordant rows comprise 13 alignments (12 unique, including Zika Angola `08` at $|\Delta t_{\mathrm{MRCA}}| = 0.79$ years: 2005.71 [2000.71, 2007.45] vs. 2006.50 [2005.80, 2007.20]) with $|\Delta t_{\mathrm{MRCA}}| < 1.0$ year and 11 cohorts qualifying by interval overlap alone. Interval-overlapping cohorts included Caribbean DENV-1 `04` (1846.46 [1830.56, 1859.76] compared with BEAST 1880.98 [1844.11, 1901.82]) [53], DENV polyprotein `06` (1969.33 [1966.02, 1972.11] vs. 1965.00 [1958.00, 1972.00]), and DENV-1 `07` (1940.17 [1892.32, 1956.84] vs. 1952.00 [1945.00, 1960.00]). Additional interval-overlapping cohorts were Influenza H3N2 `12` (1966.63 [1965.56, 1967.63] vs. 1968.00 [1967.50, 1968.50]) [54], Influenza H5N1 `13` (1993.14 [1992.09, 1993.84] vs. 1994.50 [1993.80, 1995.80]), Caribbean DENV-2 `05` (1842.02 [1592.75, 1957.09]), and *Y. pestis* second pandemic `29` (1141.78 [−383.82, 1335.47]) [55]. Restricted to the 15 cohorts with explicit published 95% HPDs/BCIs in the primary text, concordance was 10 of 15 (66.7%, or 10 of 16 [62.5%] including `39_chikv_reunion_dellicour2020`). Pairing pairwise-distance regression with HyphAeon Attention PGLS (`model.viral.pt`) resolved structured cohort `30_mpox_clade_ib_burundi2025` (2024.29 [2024.19, 2024.36] vs. reference 2023.95 [2023.67, 2024.22], $|\Delta t_{\mathrm{MRCA}}| = 0.34$ years, where unguided spline extrapolation gave 1967.29). Attention PGLS raised multi-model concordance to 25 of 42 alignments (59.5%, or 64.1% of 39, and 24 of 41 [58.5%] or 24 of 38 [63.2%] unique cohorts). Across all 39 numeric-comparator rows, median $|\Delta t_{\mathrm{MRCA}}|$ was 7.19 years for pure pairwise-distance regression and 6.41 years with Mpox PGLS (8.32 and 6.80 years across the 38 unique comparators). Across the 19 post-1980 comparator rows, median $|\Delta t_{\mathrm{MRCA}}|$ was 0.51 years for pure pairwise-distance regression and 0.40 years with Mpox PGLS (0.65 and 0.46 years across the 18 unique post-1980 cohorts). Among the 14 discordant comparator cohorts, several reflected ancient roots compressed by purifying selection and saturation, such as Lassa (`18_lassa_andersen_suchard2020`) [43,44], *Mycobacterium tuberculosis* (`35-36`) [56,57], and Measles/Rinderpest (`37_measles_1912_dux2020`, unbounded Fieller interval $[-\infty, 1870.69]$ [49] and portal interval [944.09, 1380.77] vs. 528 BCE [−528 CE] root [45]). The remainder were multi-lineage alignments where whole-alignment regression dated a deep species split instead of a focal regional introduction, including Zika Pacific/Fiji (`14`), Avian Influenza H7/N7 (`19-20`), Caribbean DENV-3/4 (`23-24`), Foot-and-Mouth Disease serotype A (`25`), Mpox Clade I DRC (`27`) [58], RSV-A (`31`), and African Swine Fever (`34`) (Supplementary Tables S6 and S7).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Data and Code Availability

Primary sequencing datasets and genome assemblies analyzed in this study are available from public repositories. Bacterial RNA-seq runs (`ERR14660412`, `ERR14660413`, and `ERR14660420`–`ERR14660422`) are under ENA BioProject `PRJEB86181`, and the *B. cereus* ATCC 10987 reference assembly (`GCF_000008005.1` / `GCA_000008005.1`) is under NCBI BioProject `PRJNA74`. Fungal, vector chromatin, and vector single-cell datasets are available under NCBI BioProject `PRJNA482512` (*A. fumigatus* Af293 `ASM265v1`), NCBI GEO accession `GSE152924` (`SRR12064732`–`SRR12064735`), and NCBI SRA runs `SRR26591148`–`SRR26591149` (*A. aegypti* `GCF_002204515.2`). Spatial and phylogenetic benchmark archives are available at Zenodo record `8259942` (TISSUE Dataset 15) and on UseGalaxy.org (42 BEAST XML benchmarks from https://github.com/veg/chronaeon and https://github.com/veg/HyphAeon, Supplementary Tables S6 and S7). Galaxy history, workflow, collection, and Page links for each investigation are provided in Supplementary Tables S3–S6.

All core platform software components developed in this work are open-source and publicly available:

- Galaxy MCP Server: https://github.com/galaxyproject/galaxy-mcp.
- Orbit Workbench and Loom Runtime: https://github.com/galaxyproject/loom.
- Galaxy Workflow Foundry: https://github.com/galaxyproject/foundry.
- BRC-Analytics Portal: https://brc-analytics.org (repository: https://github.com/galaxyproject/brc-analytics).
- Galactic Skills and Developer Guides: https://github.com/galaxyproject/galaxy-skills.
- Universal Agentic Plugins: https://github.com/galaxyproject/agentic-plugins.

Manuscript figure scripts, in-repository figure source tables (`manuscript/figures/data/`), benchmark summary tables, and companion case-study repositories (https://github.com/nekrut/orbit-paper, https://github.com/nekrut/spatTrOrbit, https://github.com/nekrut/phist, https://github.com/veg/chronaeon, and the BRC reanalysis repositories in Supplementary Table S3) are archived on GitHub and UseGalaxy.org (with any histories, workflows, or companion repositories under review access controls made publicly accessible upon publication).

Acknowledgments

We thank Anup Kumar, Ivo Leist, Wolfgang Maier, Daniela Schneider, and Paul Zierep of the University of Freiburg (UseGalaxy.eu) for helpful discussions, technical contributions, and feedback on the manuscript, and David Y. Hancock and the Jetstream2 team at Indiana University for supporting the cloud infrastructure where User-Defined Tools execute. This work was supported by National Institutes of Health grants U24AI183870 (NIAID) and U24HG006620 (NHGRI), and by National Science Foundation awards 2419522 and 2138585. This work used the Jetstream2 cloud resource at Indiana University [59] (supported by U.S. National Science Foundation award OAC-2005506), Stampede3, Lonestar6, and Ranch at the Texas Advanced Computing Center (TACC), Bridges-2 at the Pittsburgh Supercomputing Center, Expanse at the San Diego Supercomputer Center, and Rockfish at Johns Hopkins University through allocation MCB140147 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [60], which is

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The authors acknowledge Indiana University and the Jetstream2 team (https://jetstream-cloud.org) and the Texas Advanced Computing Center (TACC, http://www.tacc.utexas.edu) at The University of Texas at Austin for providing high-performance computing, cloud, and storage resources that have contributed to the research results reported within this paper, alongside computational resources provided by UseGalaxy.org and UseGalaxy.eu.

References

[1] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023.

[2] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024.

[3] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, July 2025.

[4] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026.

[5] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv*, page 2511.02824, 2025.

[6] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical AI agent. *bioRxiv*, page 2025.05.30.656746, 2025.

[7] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, May 2026.

[8] The Galaxy Community, Enis Afgan, Najat Amoukou, Rafael Andrade Buono, Mihail Anton, Christophe Antoniewski, Patrick Austin, Ahmed Hamid Awan, Rolf Backofen, Wendi Anne Bacon, Dannon Baker, Arthur Barreau, Bérénice Batut, Matthias Bernt, Daniel Blankenberg, Anthony Bretaudeau, Catherine Joan Bromhead, Richard Burhans, Melissa Burke, Scott Cain, Danielle Callan, Martin Čech, María Chavero-Díez, Haswane Chei, Ying Chen, Diana Chiang, John M Chilton, Tyler James Collins, Frederik Coppens, Nate Coraor, Charles Coulombe, Michael R Crusoe, Fabio Cumbo, Triskell Cumunel, Michael D'Silva, Armin Dadras, Céline

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Dalle, Khai Dang, John Y Davis, Paul De Geest, Willem de Koning, Giuseppe Defazio, Boris Depoortere, Katherine Do, José Manuel Domínguez Begines, Bert Droesbeke, Erick Nicolas Duarte, Enol Fernández-del Castillo, Jeremy Fix, Giulio Formenti, Julian Frey, Melanie Christine Föll, Carol Gauthier, Kilian Gerberding, Franck Giacomoni, Jeremy Goecks, Nuwan A Goonasekera, Nadia Goué, Timothy J Griffin, Björn Andreas Grüning, Aysam Guerler, Yann Guitton, Ove Johan Ragnar Gustafsson, Kristína Gömöryová, Magdalena Harakalova, Helge Hecht, Alireza Heidari, Florian Heyl, Jennifer Hillman-Jackson, Saskia Hiltemann, Mina Hojat Ansari, Hans-Rudolf Hotz, Cameron John Hyde, Pierre-Étienne Jacques, Pratik Dilip Jagtap, Jayadev Joshi, Marie Jossé, Khaled Moahmmad Ahmad Jumah, Arash Kadkhodaei Elyaderani, Katarzyna Kamieniecka, Teja Kattenborn, Markus Konkol, Leonid Kostrykin, Natalie Kucher, Anup Kumar, Mira Kuntz, Bradley William Langhorst, Delphine Lariviere, Yvan Le Bras, Gildas Le Corguillé, Jean Le Cras, Justin Lee, Jan Leendertse, Hugo Lefeuvre, Brane Leskošek, Leandro Miguel Liborio, Romane Libouban, Marisa Loach, Jose David Lopez Tabernero, Lucille Lopez-Delisle, Daniel Lusk, Alexandru Mahmoud, Molène Mahé, Wolfgang Maier, Igor Makunin, Kirsty McCaffrey, Jack A Medico, Subina Mehta, Hailiang Mei, Mirela Minkova, Saim Momin, Paulo Cilas Morais Lyra Junior, Teresa Müller, Amirhossein Naghsh Nilchi, Tannistha Nandi, Engy Mohamed Taha Nasr, Anton Nekrutenko, Tiffanie Nelson, Johannes Nussbaum, Asime James Oba, Łukasz Opioła, Kevin Payet, Melanie Petera, Polina V Polunina, Sergei Pond, Krzysztof Poterlowicz, Gareth Robert Price, Junhao Qiu, Helena Rasche, Bryan Raubenolt, Tristan N Reynolds, Dave Rogers, Karl Rohr, Gabriel Ferreira Saudade, Michelle Terese Savage, Volodymyr Savchenko, Michael C Schatz, Isabelle Schmitz, Daniela Schneider, Pauline Seguineau, Beatriz Serrano-Solano, Clea Siguret, Patrik Smeds, Marco Sollitto, Nicola Soranzo, Sanjay Kumar Srikakulam, Lieven Sterck, Nikolaos Strepis, Andrew Stubbs, Keith Suderman, Anna Syme, Marco Antonio Tangaro, Reyhaneh Tavakoli Koopaei, Jonathan Andrew Tedds, Mehmet Tekman, Wai Cheng (Mike) Thang, Anil S Thanki, Michael Uhl, Janusch Vajna-Jehle, Marius van den Beek, Deepti Varshney, Nikolay Alexandrov Vazov, Jennifer Vessio, Pavankumar Videm, Tomas Vondrak, Reid Wagner, Gregory R Watson, Ralf J M Weber, Natalie Whitaker-Allen, Federico Zambelli, Paul Zierep, and Rand Zoabi. Galaxy for accessible, reproducible, and collaborative data analyses: 2026 update. *Nucleic Acids Research*, 54(W1):W105–W116, June 2026.

[9] Felipe da Veiga Leprevost, Björn A. Grüning, Saulo Alves Aflitos, Hannes L. Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, Mingze Bai, Rafael C. Jimenez, Timo Sachsenberg, Julianus Pfeuffer, Roberto Vera Alvarez, Johannes Griss, Alexey I. Nesvizhskii, and Yasset Perez-Riverol. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16):2580–2582, 2017.

[10] Intergalactic Utilities Commission. Galaxy Intergalactic Utilities Commission (IUC) tool repository. https://github.com/galaxyproject/tools-iuc, 2026. GitHub repository.

[11] Intergalactic Workflow Commission. Galaxy Intergalactic Workflow Commission (IWC). https://iwc.galaxyproject.org, 2026. Workflow registry and GitHub repository (https://github.com/galaxyproject/iwc).

[12] Bérénice Batut, Saskia Hiltemann, Andrea Bagnacani, Dannon Baker, Vikas Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Brillet-Guéguen, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A. Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Delphine Larivière, Ross Lazarus, Yvan Le Bras, Hervé Ménager, Anton Nekrutenko, Tobias Rausch,

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Helena Rasche, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, James Taylor, and Daniel Blankenberg. Community-driven data analysis training for biology. *Cell Systems*, 6(6):752–758.e1, 2018.

[13] Saskia Hiltemann, Helena Rasche, Simon Gladman, Hans-Rudolf Hotz, Delphine Larivière, Daniel Blankenberg, Pratik D. Jagtap, Thomas Wollmann, Anthony Bretaudeau, Nadia Goué, Timothy J. Griffin, Coline Royaux, Yvan Le Bras, Subina Mehta, Anna Syme, Frederik Coppens, Bert Droesbeke, Nicola Soranzo, Wendi Bacon, Fotis Psomopoulos, Cristóbal Gallardo-Alba, John Davis, Melanie Christine Föll, Matthias Fahrner, Maria A. Doyle, Beatriz Serrano-Solano, Anne Fouilloux, Peter van Heusden, Wolfgang Maier, Dave Clements, Florian Heyl, Björn Grüning, Bérénice Batut, and Galaxy Training Network. Galaxy Training: A powerful framework for teaching! *PLOS Computational Biology*, 19(1):e1010752, 2023.

[14] Galaxy Project. Galaxy workflow Foundry. https://github.com/galaxyproject/foundry, 2026. GitHub repository.

[15] Kateřina Vyklická, Jiří Kučera, Vojtěch Bartoň, Jan Böhm, Roman Remínek, Petra Kubásová, Kateřina Pašková, Zdeněk Glatz, Jan Lochman, Filip Růžička, and Petra Bořilová Linhartová. Effect of carbohydrate substrates on growth and enterotoxin gene expression in *Bacillus cereus* (pacificus). *Scientific Reports*, 16(1):1917, 2025.

[16] Galaxy Project. Galaxy MCP server. https://github.com/galaxyproject/galaxy-mcp, 2026. GitHub repository.

[17] Galaxy Project. Galactic skills. https://github.com/galaxyproject/galaxy-skills, 2026. GitHub repository.

[18] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[19] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026.

[20] Rayan Chikhi, Téo Lemane, Raphaël Loll-Krippleber, Mercè Montoliu-Nerin, Brice Raffestin, Antonio Pedro Camargo, Carson J. Miller, Mateus Bernabe Fiamenghi, Daniel Paiva Agustinho, Sina Majidian, Greg Autric, Maxime Hugues, Junkyoung Lee, Roland Faure, Kristen D. Curry, Jorge A. Moura de Sousa, Eduardo P. C. Rocha, David Koslicki, Paul Medvedev, Purav Gupta, Jessica Shen, Alejandro Morales-Tapia, Kate Sihuta, Peter J. Roy, Grant W. Brown, Robert C. Edgar, Anton Korobeynikov, Martin Steinegger, Caleb A. Lareau, Pierre Peterlongo, and Artem Babaian. Logan: Planetary-scale genome assembly surveys life's diversity. *bioRxiv*, page 2024.07.30.605881, 2024.

[21] Wei Shen, John A. Lees, and Zamin Iqbal. Efficient sequence alignment against millions of prokaryotic genomes with LexicMap. *Nature Biotechnology*, 44(8):1383–1390, 2026.

[22] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[23] Galaxy Project. User-Defined tools (UDT) authoring specification. https://docs.galaxyproject.org/en/release_26.1/dev/user_defined_tools_authoring.html, 2026. Galaxy Release 26.1 Documentation.

[24] Jiacheng Miao, Joe R. Davis, Yaohui Zhang, Jonathan K. Pritchard, and James Zou. Reimagining research papers as interactive and reliable AI agents. *Nature*, 2026. Advance online publication; PMID 42749808.

[25] Eric D. Sun, Rong Ma, Paloma Navarro Negredo, Anne Brunet, and James Zou. TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *Nature Methods*, 21(3):444–454, 2024.

[26] Margriet W. J. Hokken, Jan Zoll, Jordy P. M. Coolen, Bas J. Zwaan, Paul E. Verweij, and Willem J. G. Melchers. Phenotypic plasticity and the evolution of azole resistance in *Aspergillus fumigatus*; an expression profile of clinical isolates upon exposure to itraconazole. *BMC Genomics*, 20(1):28, 2019.

[27] José L. Ruiz, Lisa C. Ranford-Cartwright, and Elena Gómez-Díaz. The regulatory genome of the malaria vector *Anopheles gambiae*: integrating chromatin accessibility and gene expression. *NAR Genomics and Bioinformatics*, 3(1):lqaa113, 2021.

[28] Shunlong Wang, Ying Huang, Fei Wang, Qian Han, Nanjie Ren, Xiaoyu Wang, Yingjun Cui, Zhiming Yuan, and Han Xia. A cell atlas of the adult female *Aedes aegypti* midgut revealed by single-cell RNA sequencing. *Scientific Data*, 11:587, 2024.

[29] Jane M. Carlton, John H. Adams, Joana C. Silva, Shelby L. Bidwell, Hernan Lorenzi, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature*, 455(7214):757–763, 2008.

[30] Sarah Auburn, Ulrike Böhme, Sascha Steinbiss, Hidayat Trimarsanto, Jessica Hostetler, Mandy J. Sanders, Qi Gao, Francois Nosten, Christopher I. Newbold, Matthew Berriman, Ric N. Price, and Thomas D. Otto. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Research*, 1:4, 2016.

[31] Angela M. Minassian, Yrene Themistocleous, Sarah E. Silk, Jordan R. Barrett, Alison Kemp, Doris Quinkert, Carolyn M. Nielsen, Nick J. Edwards, Thomas A. Rawlinson, Fernando Ramos Lopez, et al. Controlled human malaria infection with a clonal *Plasmodium vivax* isolate for blood-stage vaccine testing. *JCI Insight*, 6(23):e152465, 2021.

[32] Katlijn De Meulenaere, Bart Cuypers, Dionicia Gamboa, Kris Laukens, and Anna Rosanas-Urgell. A new Plasmodium vivax reference genome for South American isolates. *BMC Genomics*, 24(1):606, 2023.

[33] Thomas E. Harrison, Adam J. Reid, Deirdre Cunningham, Jean Langhorne, and Matthew K. Higgins. Structure of the Plasmodium-interspersed repeat proteins of the malaria parasite. *Proceedings of the National Academy of Sciences*, 117(50):32098–32104, 2020.

[34] Stephen K. Gire, Augustine Goba, Kristian G. Andersen, Rachel S. G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[35] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Guy Baele, Nuno R. Faria, Daniel J. Park, Jason T. Ladner, Armando Arias, Danny Asogun, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544(7650):309–315, 2017.

[36] Verity Hill and Guy Baele. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Molecular Biology and Evolution*, 36(11):2620–2628, 2019.

[37] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, Flavia C. S. Sales, Iwona Hawryluk, John T. McCrone, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372(6544):815–821, 2021.

[38] Sergei L. Kosakovsky Pond, Steven Weaver, Danielle Callan, Jordan D. Zehr, Alexander G. Lucaci, Hannah Verdonk, Avery Selberg, Gallean Brown, Maria Chikina, Nathan L. Clark, Kateryna D. Makova, Darren P. Martin, and Anton Nekrutenko. HyphAeon: Attention on evolution across deep time transforms comparative genomics. *bioRxiv*, page 2026.09.06.749597, 2026. Preprint; ChronAeon molecular-clock module in https://github.com/veg/chronaeon and https://github.com/veg/HyphAeon.

[39] Galaxy Project and BRC-Analytics Consortium. BRC-Analytics: Bioinformatics Resource Center analytics portal. https://brc-analytics.org, 2026. Web portal and GitHub repository (https://github.com/galaxyproject/brc-analytics).

[40] F. J. López, M. Bernabeu, C. Fernandez-Becerra, and H. A. del Portillo. A new computational approach redefines the subtelomeric vir superfamily of Plasmodium vivax. *BMC Genomics*, 14:8, 2013.

[41] Andrew P. Jackson, Deirdre A. Cunningham, Lin Lin, Naomi Mara Claro de Oliveira, Séverine C. Chevalley-Maurel, et al. Essential function reflected in the phylodynamics of a multigene family – the pir genes of malaria parasites. *eLife Reviewed Preprint*, page RP111505, 2026.

[42] Marc A. Suchard, Philippe Lemey, Guy Baele, Daniel L. Ayres, Alexei J. Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016, 2018.

[43] Kristian G. Andersen, B. Jesse Shapiro, Christian B. Matranga, Rachel Sealfon, Aaron E. Lin, Lina M. Moses, Onikepe A. Folarin, Augustine Goba, Ikponmwosa Odia, Philomena E. Ehiane, et al. Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell*, 162(4):738–750, 2015.

[44] Guy Baele, Mandev S. Gill, Philippe Lemey, and Marc A. Suchard. Fast, scalable and accurate model selection for phylodynamic inference. *Systematic Biology*, 69(5):953–965, 2020.

[45] Ariane Düx, Sebastian Lequime, Livia V. Patrono, Bram Vrancken, Sengül Boral, Jan F. Gogarten, Antonia Hilbig, David Horst, Kevin Merkel, Baptiste Prepoint, et al. Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science*, 368(6497):1367–1370, 2020.

[46] Morluto. REA: Reverse Engineer Anything. https://github.com/morluto/rea, 2026. GitHub repository.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[47] Simon Bray, John Chilton, Matthias Bernt, Nicola Soranzo, Marius van den Beek, Bérénice Batut, Helena Rasche, Martin Čech, Peter J. A. Cock, Björn Grüning, and Anton Nekrutenko. The Planemo toolkit for developing, deploying, and executing scientific data analyses in Galaxy and beyond. *Genome Research*, 33(2):261–268, 2023.

[48] Mandev S. Gill, Philippe Lemey, Marc A. Suchard, Andrew Rambaut, and Guy Baele. Online bayesian phylodynamic inference in BEAST with application to epidemic reconstruction. *Molecular Biology and Evolution*, 37(6):1832–1842, 2020.

[49] Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

[50] Simon Dellicour, Guy Baele, Gytis Dudas, Nuno R. Faria, Oliver G. Pybus, Marc A. Suchard, Andrew Rambaut, and Philippe Lemey. Phylodynamic assessment of intervention strategies for the west african Ebola virus outbreak and Chikungunya epidemic. *Molecular Biology and Evolution*, 37(10):2993–3006, 2020.

[51] Áine O'Toole, Richard A. Neher, Nnaemeka Ndodo, Vítor Borges, Ben Gannon, João Paulo Gomes, Natalie Groves, Daniel J. King, David Lau, John T. McCrone, et al. APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016. *Science*, 382(6670):595–600, 2023.

[52] Neil Commins, Matthew R. Sullivan, Kira McGowen, Evan Koch, Eric J. Rubin, and Maha Farhat. Phylogenomic analysis of *Mycobacterium abscessus* clinical isolates. *Nature Microbiology*, 8(8):1495–1507, 2023.

[53] Emma Taylor-Salmon, Verity Hill, Lauren M. Paul, Robert T. Koch, Mallery I. Breban, et al. Travel surveillance uncovers dengue virus dynamics and introductions in the Caribbean. *Nature Communications*, 15(1):3508, 2024.

[54] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257, 2014.

[55] Maria A. Spyrou, Marcel Keller, Rezeda I. Tukhbatova, Christiana L. Scheib, Elizabeth A. Nelson, Aida Andrades Valtueña, Gunnar U. Neumann, Don Walker, Amelie Alterauge, Niamh Carty, et al. Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. *Nature Communications*, 10(1):4470, 2019.

[56] Iñaki Comas, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E. Holt, Midori Kato-Maeda, Julian Parkhill, Bijaya Malla, Stefan Berg, Guy Thwaites, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics*, 45(10):1176–1182, 2013.

[57] Kirsten I. Bos, Kelly M. Harkins, Alexander Herbig, Mireia Coscolla, Nico Weber, Iñaki Comas, Stephen A. Forrest, Josephine M. Bryant, Simon R. Harris, Verena J. Schuenemann, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514(7523):494–497, 2014.

[58] Livia V. Patrono, Kátia Pléh, Liran Samuni, Markus Ulrich, Caroline Röthemeier, Andreas Sachse, Sonja Muschter, Andreas Nitsche, Emmanuel Couacy-Hymann, Christophe Boesch, et al. Monkeypox virus emergence in wild chimpanzees reveals distinct clinical outcomes and viral diversity. *Nature Microbiology*, 5(7):955–965, 2020.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

[59] David Y. Hancock, Jeremy Fischer, John Michael Lowe, Winona Snapp-Childs, Marlon Pierce, Suresh Marru, J. Eric Coulter, Matthew Vaughn, Brian Beck, Nirav Merchant, Edwin Skidmore, and Gwen Jacobs. Jetstream2: Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced Research Computing (PEARC '21)*, pages 1–8, New York, NY, USA, 2021. ACM.

[60] Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing (PEARC '23)*, pages 173–176, New York, NY, USA, 2023. ACM.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Information

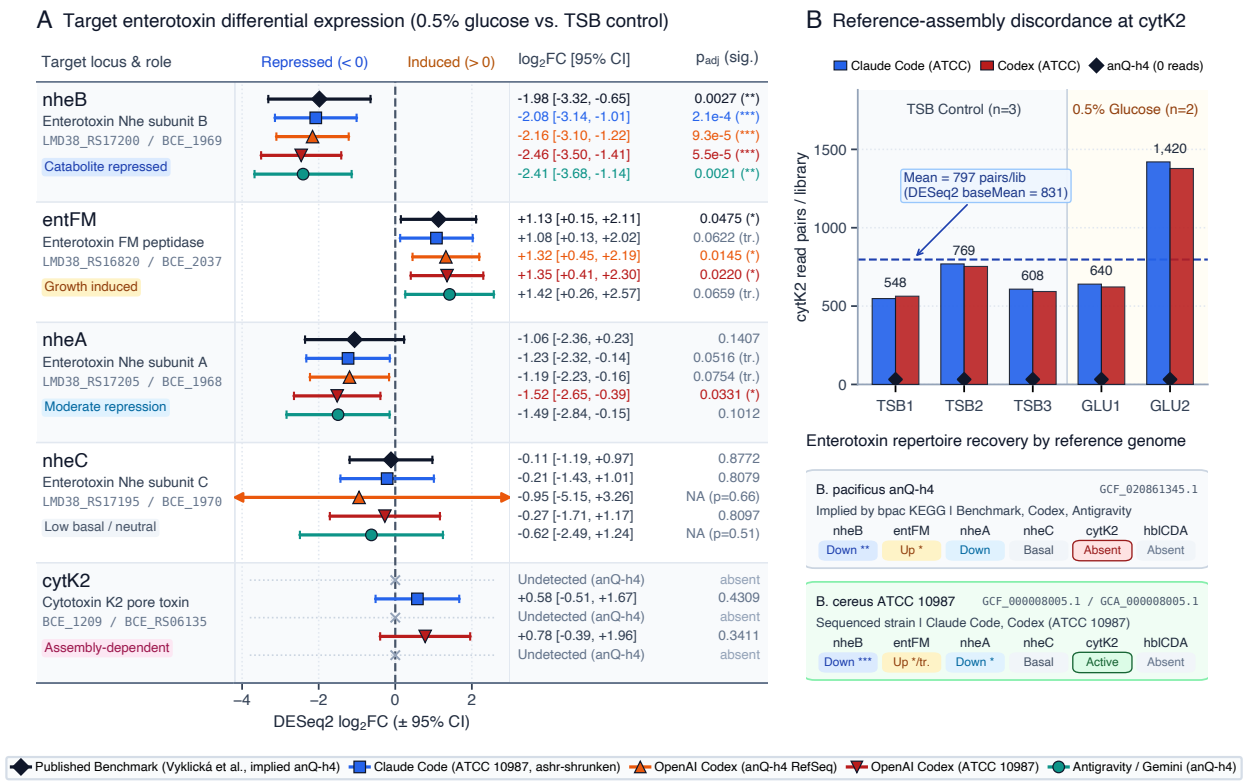


Figure S1: Local agent runs recover headline enterotoxin signals but diverge on reference genomes and leave unversioned glue code. A, Differential expression estimates ($\log_2$ FC with 95% confidence intervals) and adjusted p-values for target enterotoxin genes (*nheB*, *entFM*, *nheA*, *nheC*, and *cytK2*) compared against the published study (black diamonds, Vyklická et al. [15]). OpenAI Codex (orange and rust across two reference assemblies) and Google Gemini via Antigravity (teal) report unshrunken DESeq2 Wald estimates, whereas Claude Code (blue, `nf-core/differentialabundance`) reports `ashr`-shrunken estimates (for *nheC* on *B. pacificus*, where independent filtering sets p_{adj} to `NA`, the unadjusted p-value is shown). All three agents confirmed strong repression of *nheB* and induction of *entFM*. **B**, Reference-assembly mismatch at the cytotoxin K (*cytK2*) locus. Aligning against the sequenced isolate *B. cereus* ATCC 10987 in Claude Code and Codex recovered active baseline *cytK2* transcription across all libraries (548–1,420 read pairs per library, mean 797, DESeq2 baseMean = 831, $p_{\text{adj}} = 0.43$). Aligning against *B. pacificus* anQ-h4 missed *cytK2* entirely because that assembly lacks the locus.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

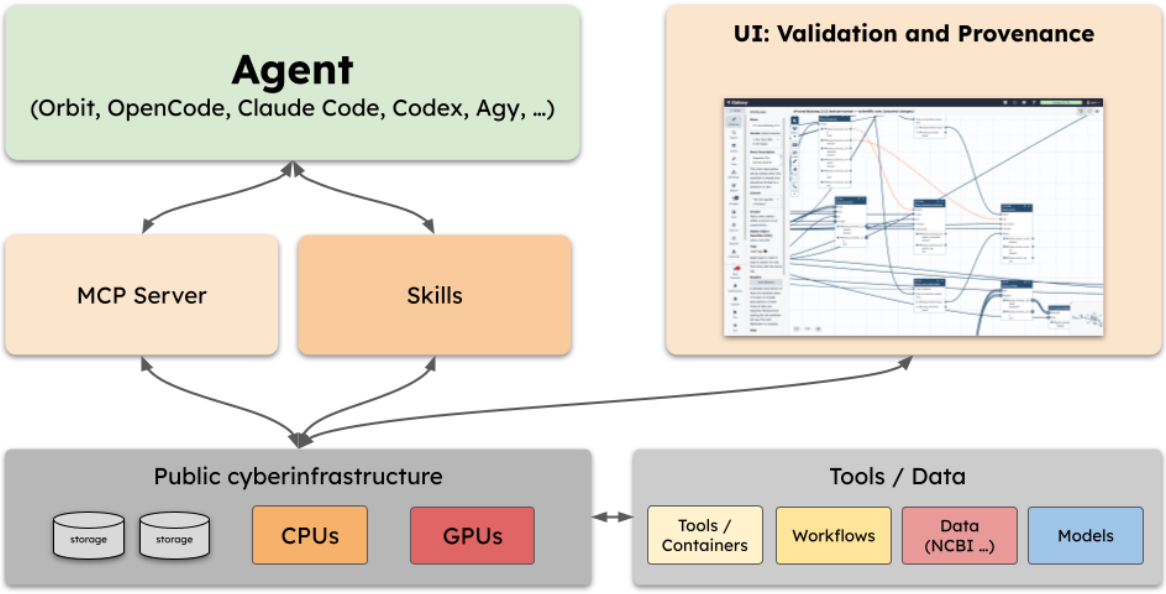


Figure S2: Architecture for verifiable agentic science on public infrastructure. Agent (top left): Researchers use Orbit or connect any generic coding agent (such as OpenCode, Claude Code, OpenAI Codex, or Antigravity CLI). **MCP Server and Skills** (middle left): Curated domain skills supply validated bioinformatics protocols, while the Galaxy MCP server translates agent actions into typed API calls in place of unconstrained shell commands. **UI: Validation and Provenance** (top right): Galaxy's visual workflow editor and history interface provide pre-execution schema validation and an inspectable provenance record for human steering and verification. **Public cyberinfrastructure and Tools/Data** (bottom): Galaxy schedules containerized tools, peer-reviewed workflows, public datasets (NCBI, ENA, BRC-Analytics), and large models across CPU, GPU, and storage allocations on national supercomputing infrastructure.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Figure S3: Workflow and execution provenance on free public infrastructure catches pipeline errors and reproduces bacterial RNA-seq profiles. **A**, Pre-execution workflow validation on UseGalaxy.org. When an agent connected via the Galaxy MCP server inserts an incompatible standalone `txiimport` step between a Kallisto output collection and DESeq2, the workflow editor flags the type mismatch in orange. **B**, Validated workflow after repair. Passing the transcript-to-gene table to DESeq2's built-in `txiimport` mode resolves all workflow connections in green. **C**, Live provenance and interactive Galaxy History Notebook. Jobs execute on free public ACCESS-CI supercomputing allocations while completed datasets populate the right-hand history panel and the center notebook links narrative text to verified history dataset IDs. **D**, Biological replication of Vyklická et al. [15] across the *B. cereus* ATCC 10987 coding transcriptome. Supplementation with 0.5% glucose represses enterotoxin subunits *nheB* (5.05-fold down, $\log_2\mathrm{FC} = -2.335$, $p_{\mathrm{adj}} = 2.99 \times 10^{-3}$) and *nheA* (2.80-fold down, $\log_2\mathrm{FC} = -1.484$, raw $p = 0.0274$, $p_{\mathrm{adj}} = 0.106$), leaves component *nheC* unchanged ($\log_2\mathrm{FC} = -0.158$, $p_{\mathrm{adj}} = 0.907$), and induces *entFM* (2.70-fold up, $\log_2\mathrm{FC} = +1.432$, raw $p = 0.0158$, $p_{\mathrm{adj}} = 0.0730$). Across all 5,250 expressed coding loci (of 5,472 total), Kallisto pseudoalignment and HISAT2 alignment agree closely (Pearson $r = 0.9622$, $R^2 = 0.9257$).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

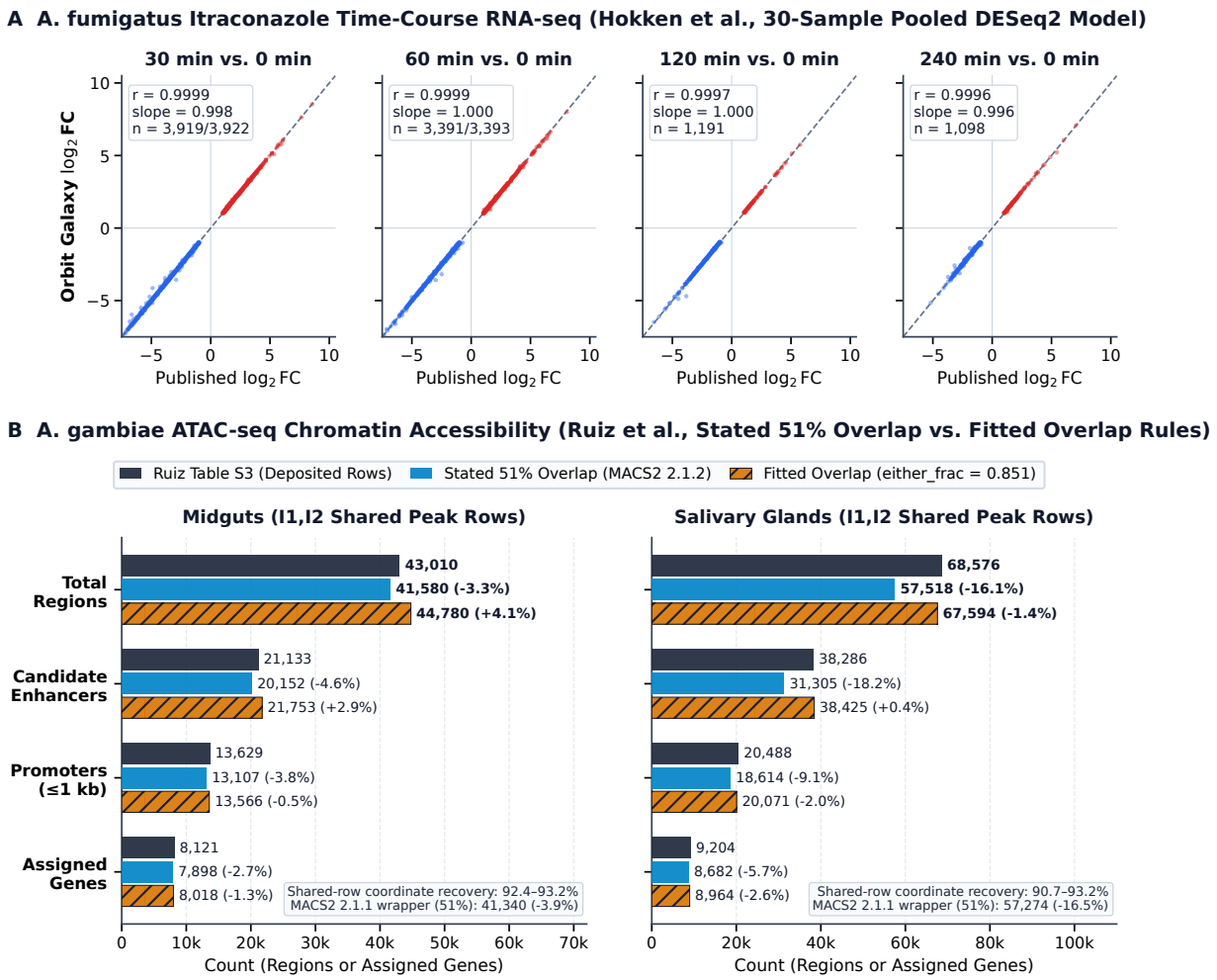


Figure S4: Reproducing pathogen and vector functional genomics studies on UseGalaxy.org. A, Transcriptomic response of the pathogenic mold *Aspergillus fumigatus* to the antifungal drug itraconazole across 30 libraries (Hokken et al. [26]). A pooled 30-sample DESeq2 model compiled via Workflow Foundry reproduces published $\log_2$ fold changes across all four post-exposure timepoints ($r = 0.9996$–0.9999, slopes 0.996–1.000). Published differentially expressed gene counts match within 0.3–1.3% ($n = 3,919$ of $3,922$ at 30 min, $3,391$ of $3,393$ at 60 min, $1,191$ at 120 min, and $1,098$ at 240 min). **B**, Chromatin accessibility in *Anopheles gambiae* malaria mosquito midguts and salivary glands during *Plasmodium falciparum* infection (Ruiz et al. [27]). Independent MACS2 peak calling recovers 90.7–93.2% of the 111,586 replicate-shared genomic regions in the authors' deposited table across all four libraries (81.8–92.2% across all 193,005 deposited peaks). Provenance-tracked overlap rules show how unmerged per-replicate rows and overlap filtering account for the headline region, promoter, and enhancer counts derived from Ruiz Table S3 within 0.4–2.9% (0.4–4.6% across tracks).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

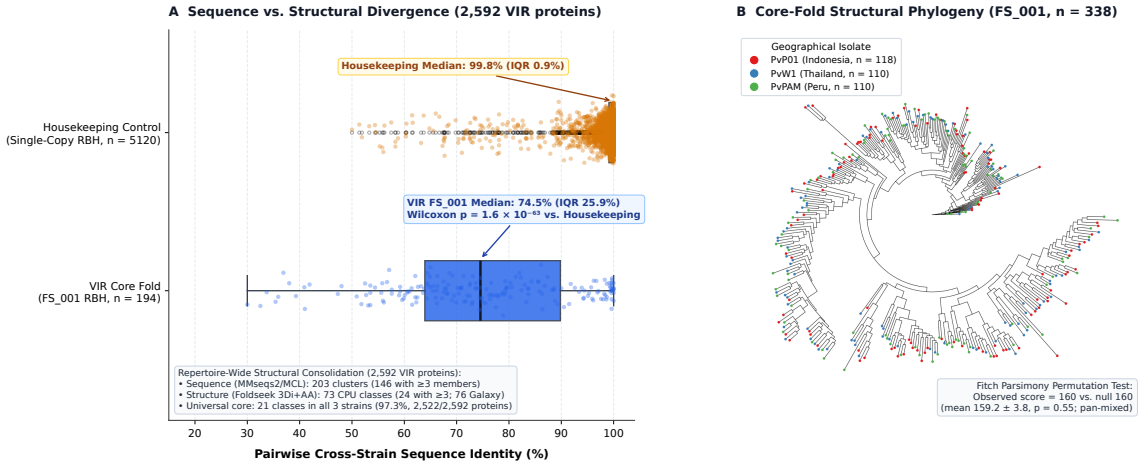


Figure S5: Structure-aware family reconstruction and lineage sharing of *Plasmodium vivax* VIR surface antigens across continental genomes. A, Sequence twilight zone and structural unification of over-split subfamilies. Horizontal box-and-strip plots display pairwise amino acid sequence identity between PvW1 (Thailand) and PvPAM (Peru) for conserved single-copy housekeeping genes (median 99.8%, IQR 0.9%, amber, $n = 5,120$ pairs) versus VIR pairs within dominant superfamily FS_001 (median 74.5%, IQR 25.9%, Wilcoxon $p = 1.6 \times 10^{-63}$, blue, $n = 194$ pairs). Sequence-only Markov clustering at 30% minimum sequence identity splits the 2,592 VIR proteins across Indonesia (PvP01), Thailand (PvW1), and Peru (PvPAM) into 203 sequence clusters (146 subfamilies with ≥ 3 members). Structure-aware clustering using ProstT5 and Foldseek 3Di representations collapses this repertoire into 73–76 structural classes (24–30 with ≥ 3 members) across local and UseGalaxy.org execution environments, where the dominant superfamily FS_001 unifies 89 sequence subfamilies on the local reference branch (see Methods). **B**, Circular maximum-likelihood structural phylogeny inferred by IQ-TREE 3 from FoldMason alignments of 338 FS_001 representatives on the local reference branch across Indonesia (PvP01, red, 34.9%), Thailand (PvW1, blue, 32.5%), and Peru (PvPAM, green, 32.5%) using partitioned amino acid and 3Di structural models. Continental lineages are interleaved across the tree, and a tip-label parsimony permutation test shows no detectable geographic clustering (observed score = 160 vs. 1,000 random shuffles with null median 160, $p = 0.55$).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Metric	Google Gemini 3.8 Flash (Antigravity)	Anthropic Claude Opus 5.5 (Claude Code)	OpenAI Codex / GPT-6 Astra (Codex CLI)
Agent Runtime	Antigravity CLI (`agy`) driving ad-hoc Bash and R scripts	Claude Code (`cc`) driving Nextflow workflows via Docker	Codex CLI (`codex exec`) driving modular Bash runner
Software Stack	Bowtie2 v2.5.5, featureCounts v2.1.1, DESeq2 v1.50.2, fastp v1.4.0	nf-core/rnaseq 3.14.0 (STAR, Salmon), nf-core/differentialabundance 1.4.0	HISAT2 v2.2.1, featureCounts, DESeq2, fastp v0.24.0, FastQC
Thread & CPU Strategy	Greedy 16-thread allocation across download, alignment, and quantification	Automated Docker scheduling across 24 host CPU cores	Strict budget: `OMP/MKL=1`, explicit slices (10 align, 3 sort, 2 QC, 4 count)
Peak Memory (RAM)	2–4 GB (standard heap allocation)	16–30 GB (STAR 2-pass indexing and alignment in Docker)	<6 GB (explicit buffer limit: `samtools sort -m 512M`)
Disk Footprint	15–20 GB (immediate inline cleanup of raw SRA and FASTQ files)	>100 GB scratch disk (49 GB Nextflow cache and >40 GB QC sorting files)	<20 GB (sequential processing, pruned FASTQs tracked in ledger)
Dependency Model	Shared host Conda environment (`bacillus_rnaseq`)	Local Conda environment plus Docker OCI container layers	Project-isolated cache (`.cache/`) with explicit package URLs
Custom Scripts Generated	3 scripts (1 Bash, 1 Python, 1 R, 302 LOC)	9 scripts (7 Bash, 1 Python, 1 R, 190 LOC)	14 scripts (6 Bash, 7 Python, 1 R, 774 LOC)
Reference Strategy	Single reference (*B. pacificus* anQ-h4 RefSeq)	Single reference (*B. cereus* ATCC 10987 GenBank/RefSeq)	Dual-reference sensitivity (*B. pacificus* anQ-h4 and ATCC 10987)
Execution Friction	None (completed without intervention)	Host disk exhaustion (required manual container deletion and skipping QC)	None (preventative buffer limits and resumption caching)

Table S1: Resource allocation, software environments, and compute footprints across three autonomous agent execution runs. All three runs re-analyzed the primary glucose versus control contrast from the *Bacillus cereus* transcriptomic benchmark (BioProject PRJEB86181) on an identical host machine (Ubuntu Linux workstation with 24 CPU cores, 62 GB RAM, and NVMe SSD storage). While all three platforms recovered the core biological signal of enterotoxin regulation, their resource management paradigms differed sharply. Gemini utilized a lightweight, greedy Unix shell pipeline with immediate file cleanup. Claude executed a heavy containerized Nextflow workflow that consumed over 100 GB of scratch disk and caused disk exhaustion until secondary quality control tools were disabled. Codex applied strict computational budgeting. It capped memory buffers and thread allocations per stream while maintaining an explicit ledger of purged intermediate files. None of the three runs produced persistent provenance records that external researchers could inspect or re-execute without rebuilding the local execution environment.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Agent Runtime	Custom Scripts	Script Breakdown and Roles	Execution Paradigm and Portability
Google Gemini 3.8 Flash (Antigravity CLI, `agy`)	**3 scripts** (302 LOC) 1 Bash 1 Python 1 R	`download_fastqs.py` (45 LOC): parallel ENA FTP fetch. `run_pipeline.sh` (93 LOC): Bowtie2, fastp, featureCounts. `run_deseq2.R` (164 LOC): DESeq2 Wald testing, PCA, plots.	Monolithic shell execution with hard-coded paths. Performs immediate cleanup of raw files but lacks checkpoint resumption or portable provenance.
Anthropic Claude Opus 5.5 (Claude Code, `cc`)	**9 scripts** (190 LOC) 7 Bash 1 Python 1 R	**Driver:** `run_all.sh` (12 LOC), `env.sh` (2 LOC). **Ref and Data:** `00_reference.sh` (16 LOC), `01_download.sh` (22 LOC), `02_make_gtf.py` (28 LOC). **Workflows:** `03_rnaseq.sh` (15 LOC), `04_featurecounts.sh` (40 LOC), `05_differentialabundance.sh` (7 LOC), `06_target_genes.R` (48 LOC).	Wraps heavy containerized Nextflow pipelines with custom glue scripts. Managed disk cache resumption but experienced host storage exhaustion.
OpenAI Codex (GPT-6 Astra, Codex CLI)	**14 scripts** (774 LOC) 6 Bash 7 Python 1 R	**Driver and Env:** `run_all.sh` (17 LOC), `env.sh` (16 LOC), `status.py` (18 LOC). **Data and Ref:** `fetch_metadata.py` (46 LOC), `download_sra.py` (42 LOC), `download_reads.py` (56 LOC), `convert_reads.sh` (21 LOC), `prepare_reference.py` (60 LOC). **Compute:** `process_sample.sh` (58 LOC), `process_all.sh` (8 LOC), `diagnose_mapping.sh` (15 LOC), `count_and_qc.py` (105 LOC), `differential_expression.R` (137 LOC), `make_report.py` (175 LOC).	Modular Unix pipeline with defensive thread and buffer throttling. Logs pruned intermediates in a custom ledger but cannot be shared without rebuilding the local environment.
All Runtimes Combined	**26 scripts** (1,266 LOC)	14 Bash, 9 Python, 3 R	Zero shared scripts, zero common workflow definitions, and zero portable provenance across all three platforms.

Table S2: Summary of custom scripts generated by autonomous agents during empirical reanalysis. Number of custom scripts, programming languages, code volumes, and functional scopes generated by each runtime to replicate the *Bacillus cereus* RNA-seq benchmark. Gemini implemented a minimal three-script pipeline driving standard command-line utilities. Claude built nine wrapper and glue scripts to drive containerized nf-core workflows and parse downstream abundances. Codex constructed a modular fourteen-script suite featuring automated metadata queries, dual-reference mapping, explicit buffer throttling, and HTML dashboard synthesis. None of the twenty-six scripts were portable or auditable beyond the specific workstation where they were generated.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Study / System	Identifier / Link	Description
Vyklická et al. (2025) *Bacillus cereus* RNA-seq (PRJEB86181)	bbd44e69cb8906b59549f5524d3906f3 bbd44e69cb8906b5d979c5818e07f6ad ea7670a8c3fa3562 93fcddc692e0986f	Complete 5-library prokaryotic RNA-seq replication history on *B. cereus* ATCC 10987 (raw CDS HID 66, deduplicated CDS HID 106, `tx2gene.tsv` HID 148, Kallisto collection HID 175, DESeq2 results HID 188, and JupyterLab comparison notebook HID 192), initial HISAT2/featureCounts comparison history (HIDs 1–64), validated four-step Kallisto–DESeq2 integrated `tximport` workflow, and published interactive Galaxy Page.
Hokken et al. (2019) *Aspergillus fumigatus* RNA-seq (PRJNA482512)	bbd44e69cb8906b5a5cd735af575e8f5 b3638784a5d2aa43 b4346d545fb2b437	Complete 30-sample time-course history (pairwise DESeq2 fits HIDs 953–959, pooled 30-sample `~timepoint` model HIDs 961 and 4423), 60-step Galaxy workflow compiled via Workflow Foundry, workflow invocation, and three custom UDTs (`hokken_go_enrichment`, `hokken_venn_comparison`, and `hokken_revigo_summary`).
Ruiz et al. (2021) *Anopheles gambiae* ATAC-seq & RNA-seq (GSE152924)	bbd44e69cb8906b5e87b8ac73fb252a5 bbd44e69cb8906b57c33c7b295ddf477 bbd44e69cb8906b51645d526a9bd48d6 32f57ede003070c3, 6c3ed747eba652a6	Primary ATAC-seq peak-calling, two-sided 51% overlap, and `either_frac` cross-fit history, raw ATAC QC history, companion RNA-seq quantification history, two published ATAC-seq workflows, and five custom UDTs (`ruiz_tn5_fragment_extractor`, `ruiz_macs2_212_callpeak`, `ruiz_reproducible_ths`, `ruiz_either_frac_shared_peaks`, and `ruiz_either_frac_crossfit`).
Wang et al. (2024) *Aedes aegypti* scRNA-seq (SRR26591148-49)	https://usegalaxy.org/u/scottcain/p/wang-2024-aaegypti-midgut-scrna-seq-reanalysis https://usegalaxy.org/u/scottcain/h/wang-aaegypti-scrna-seq-veupathdb-annotation-rerun	Published Galaxy Page report and companion history containing STARsolo matrices, DropletUtils cell calls, custom `MT_`-prefixed GTF, Scanpy AnnData objects, and sample-blocked edgeR pseudobulk tables.

Table S3: Computational artifacts from the bacterial, fungal, and vector genomics reanalysis studies. Galaxy histories, workflows, User-Defined Tool definitions, and interactive Galaxy Pages archived on UseGalaxy.org.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Artifact Type	**Identifier / Link**	**Description**
Primary History	bbd44e69cb8906b596a6dfeaa66d7b35	Reproduction of published *Acta2* 1,000-cell tutorial subset
Full Dataset History	bbd44e69cb8906b5e19946d9d8770575	Full 33-gene spatial evaluation on Dataset 15 (public)
Single-Gene Workflow	fde4525ee8a0c99f	Executable workflow for target gene prediction
Full Evaluation Workflow	63012fffa29eb7bb	Published 10-fold cross-validation workflow across 33 genes
Workflow Invocation	7e806872876f7fb0	Terminal execution trace for *Acta2* prediction
User-Defined Tools	`tissue_spatial_prediction` v1.2.0, `tissue_full_evaluation` v1.0.0	Containerized TISSUE UDT specifications
Interactive Pages	14ebc62e2d1afb49, 787090c6bb87743a	Research notebook and evaluation report on UseGalaxy.org
Code Repository	https://github.com/nekrut/spatTrOrbit	Scripts, local references, and workflow templates

Table S4: Computational artifacts from the spatial transcriptomics reproduction study. Galaxy histories, workflows, UDT definitions, and execution traces archived on UseGalaxy.org and GitHub.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Artifact Type	Identifier / Link	Description
Primary Galaxy History	bbd44e69cb8906b5c0a92f1d9b046fdd	Multi-step phylogenomics execution and UDT routing trace on UseGalaxy.org
GPU & CPU Clustering UDTs	`anton-gpu-xl-foldseek_prostt5_clustering`, `anton-32-foldseek_prostt5_clustering`	Foldseek + ProstT5 3Di embedding routed to NVIDIA A100 GPU and 32-core CPU nodes
Cluster Extraction UDT	`anton-1-foldseek_cluster_extract`	Superfamily member extraction from Foldseek clustering tables
Structural Alignment UDT	`anton-32-foldmason_align`	32-core FoldMason lockstep AA + 3Di MSA in custom ProstT5 container
Pseudogene Rescue UDT	`anton-1-gff_pseudogene_translator`	Six-frame GFF3 pseudogene translation for PF05795 HMM screening
Orthology & Synteny UDTs	`anton-1-step7_prepare_fastas`, `anton-8-mmseqs_easy_rbh_pair`, `anton-1-microsynteny_filter`	Proteome partitioning, MMseqs2 reciprocal best hits, and flanking microsynteny scoring
Spatial & Phylo Prep UDTs	`anton-1-spatial_analysis`, `anton-1-iqtree_structural_prep`	Subtelomeric enrichment, label permutation tests, and partitioned IQ-TREE matrix preparation
Custom Container Image	`quay.io/nekrut/foldmason-prostt5:4.dd3c235-cpu`	Purpose-built ProstT5-enabled FoldMason container
Code Repository	https://github.com/nekrut/phist	Pipeline command ledger, 10 UDT specifications, and local reference outputs

Table S5: Computational artifacts from the *Plasmodium vivax* phylogenomics study. Galaxy histories, ten User-Defined Tool definitions, custom container images, and verification suites archived on UseGalaxy.org, Quay.io, and GitHub.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Artifact Type	Identifier / Link	Description
Preparation History	bbd44e69cb8906b5c7c1fd5515b95301	Input validation and collection preparation for 42 cohorts
Analysis History	bbd44e69cb8906b55342b6485eac77f7	End-to-end tree-free dating and diagnostics execution
Originals Archive	f9cad7b01a472135e65a92a49dc98cbe	Public binary archive of 42 author BEAST XML files
Batch Preparator UDT	`chronaeon_batch_preparator` v1.0.0	XML parser, FASTA extractor, and date validator
Cohort Analysis UDT	`chronaeon_cohort_analysis` v3.2.2	Tree-free dating, spline clock fitting, and LOOCV validation
Input Collections	e2f9a561d7c485b5 (XML), 25b287f1e09c05f2 (FASTA), 73f4b0a869dae7a4 (dates)	Synchronized 42-element BEAST XML, aligned FASTA, and tip-date collections
Output Collections	eacb6fa1da5e5cbe, c2c3f23ed16816c8	Dating JSON and diagnostic plot collections for 42 cohorts
Interactive Page	feee8a2c33d4f9e1	Published compendium evaluation report and concordance dashboard
Code, Prompts & Provenance	https://github.com/veg/chronaeon, `manuscript/prompts/`, `manuscript/figures/data/chronaeon/`	Verbatim Orbit harvesting/analysis prompts (`prompt.md`, `prompt-galaxy.md`, `prompt-galaxy-analysis.md`), benchmark TSVs, and comparator provenance table (`chronaeon_42cohorts_provenance.tsv`)

Table S6: Computational artifacts from the 42-pathogen molecular clock benchmark study. Galaxy histories, User-Defined Tool definitions, dataset collections, and published interactive reports archived on UseGalaxy.org and GitHub.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.26.761864; this version posted October 28, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Comparator Provenance Class	**Alignments**	**Cohort Identifiers (00–41)**	**Concordance & Median Error Summary**
Explicit Published 95% HPD / BCI (Primary paper text or table reports root median and 95% HPD/BCI directly)	15 scored (+1 unscored in runner)	`00` (Gire et al. 2014 81-seq root), `01`, `03`, `08`, `09`, `12`, `14`, `15`, `22`, `27`, `28`, `31`, `32`, `34`, `37` (plus `39_chikv_reunion_dellicour2020`: 2004.8 [2004.5, 2005.1] in `published_comparators.tsv`)	**10 / 15 (66.7%)** concordant under tree-free OLS/Spline (or **10 / 16 [62.5%]** including `39_chikv_reunion_dellicour2020`). Includes 9 acute outbreaks with $\lvert\Delta t_{\mathrm{MRCA}}\rvert < 1.0$ yr (`00`, `01`, `03`, `08`, `09`, `15`, `22`, `28`, `32`) and Influenza H3N2 (`12`, interval overlap).
Reconstructed Reference Interval (Interval bracketed from prose origin statement, clock rate, or Skygrid cutoff)	24 scored (23 unique)	`02`, `04`, `05`, `06`, `07`, `10`, `11`, `13`, `16`, `17`, `18`, `19`, `20`, `21`, `23`, `24`, `25`, `26`, `29`, `30`, `33`, `35`, `36`, `41` (`41_skygrid_rabies_gill2020`, paired 14.5 kb coding partition of `00`)	**14 / 24 (58.3%)** concordant under tree-free OLS/Spline (**13 / 23 [56.5%]** unique), rising to **15 / 24 (62.5%)** (**14 / 23 [60.9%]** unique) with HyphAeon Attention PGLS on Mpox Clade Ib (`30`).
All Numeric Comparators Combined (`aggregate_benchmark_results.tsv`)	39 scored (38 unique)	Cohorts `00`–`37` plus `41`	**Tree-free OLS/Spline:** 24 / 39 (61.5%, or 57.1% of 42) raw, 23 / 38 (60.5%, or 56.1% of 41) unique (13 with $\lvert\Delta t_{\mathrm{MRCA}}\rvert < 1.0$ yr [12 unique] + 11 by interval overlap). Median $\lvert\Delta t_{\mathrm{MRCA}}\rvert = 7.19$ yr overall (8.32 yr unique) and 0.51 yr across 19 post-1980 cohorts (0.65 yr across 18 unique). **With Mpox PGLS:** 25 / 39 (64.1%, or 59.5% of 42) raw, 24 / 38 (63.2%, or 58.5% of 41) unique. Median $\lvert\Delta t_{\mathrm{MRCA}}\rvert = 6.41$ yr overall (6.80 yr unique) and 0.40 yr post-1980 (0.46 yr unique).
Unscored in Automated Runner (`NO_COMPARATOR` in `aggregate_benchmark_results.tsv`)	3 unscored	`38_mab_commins2023` (*M. abscessus*, prose $\sim$ 1960–1980 vs. 1975.87 [1966.21, 1982.46]), `39_chikv_reunion_dellicour2020` (CHIKV Reunion, published 2004.8 [2004.5, 2005.1] vs. 1954.99 [1948.09, 1960.62]), `40_hiv1_crf01ae_philippines2024` (HIV-1 CRF01_AE, prose $\sim$ 1995–2002 vs. 1975.84 [1969.29, 1980.93])	Left unscored in `aggregate_benchmark_results.tsv`. Against the auxiliary entries in `published_comparators.tsv`, `38_mab_commins2023` falls inside its prose clade range ($\sim$ 1960–1980), whereas `39` and `40` infer pre-epidemic ancestral roots earlier than the focal outbreak introductions.

Table S7: Provenance classification and concordance breakdown of reference comparators across the 42 ChronAeon pathogen benchmark alignments. Per-cohort estimates, Fieller 95% confidence intervals, comparator sources, and single-pass versus multi-model values are archived in `manuscript/figures/data/chronaeon/chronaeon_42cohorts_provenance.tsv`.