Anthropic Bio 与湿实验室

从 ART 论文到科研闭环的深度调研

一项生物学发现，如何连接基础模型、科研软件、真实实验与药物研发。

研究截止：2026 年 9 月 25 日｜增补版

本报告以用户提供的 40 页论文为中心，逐节覆盖正文、全部主图与补充图、方法和参考文献；结合官方公告、公开招聘、研究者履历、Dario Amodei 的原始论述及外部报道，评估 Anthropic 生命科学布局的实际进度。

本次增补重点：2021 年 MarsHill 旧论文与新颖性争议、Nathan Frey 的药物研发方向、Dario 在 2026 年的新表述、实验运营岗位、药企合作，以及可核查的办公地产记录。新增内容见第 34-42 节。

核心判断

Anthropic 已经建立由人类科学家操作的生物实验室，并在把实验反馈接入 AI 科研流程。ART 是一个值得继续验证的新系统线索，也是一例可追溯的自主探索；它尚未成为功能得到证明的基因编辑工具。这两件事应当同时成立。

阅读方式

前半部是中文研究报告，后附完整英文论文。报告中的“原文 p.X”始终指原论文页码；“[Sxx]”对应文末可点击的外部来源。原文的 103 条参考文献完整保留在附录中。

这是一份独立研究与分析，不是 Anthropic 官方译本。凡涉及未公开的实验室门牌、团队规模、预算、内部汇报关系与商业化承诺，均不以推测代替事实。

阅读地图

按问题阅读，或沿论文证据链顺序阅读。页码为本文件的 PDF 页码。

先判断方向

结论、时间线、团队、Dario 的主张 / 第 3-6 页

再看实验室

Bio 工作线、实验分工、SF 位置线索 / 第 7-9 页

读懂全部论文

框架、搜索、ART、RNA、伙伴、可重复性、评测与成本 / 第 10-21 页

核查图与方法

五张主图、七张补图、全部方法模块 / 第 22-27 页

评估战略意义

外界反馈、限制、能力优势与待验证问题 / 第 28-31 页

查来源与原文

全文覆盖表与首批来源 / 第 32-35 页

增补调查

新颖性争议、团队与 Dario 更新、运营、地址与合作 / 第 36-44 页

证据层级	阅读规则
[P] 原论文	逐页核对；作者记录仍需与独立复现区分。
[Sxx] 外部来源	文末列原始入口、用途及访问限制。
分析 / 待证	本报告推论或未知项均明确标识。
第 45-84 页	完整 40 页英文论文，保持原页码。

01 / 先读这八个结论

问题	结论与证据强度
Anthropic 是否真有 wet lab?	有。官方确认自建实验室，实验由人类科学家完成；地点公开到 Bay Area。[S01]
是否已发现“下一个 CRISPR”？	没有这样的证据。发现的是与逆转录酶相邻的重复阵列、非编码 RNA 和伙伴蛋白组合；编辑功能尚未验证。[P]
这次 AI 的突破在哪里?	在开放式搜索中选择跟进一条被部分否定的线索，直接阅读 DNA 后识别异常，并主动检查新颖性。[P]
是否稳定可重复?	同一框架随后重跑十次，均未再次识别阵列。定向给出 DNA 后，强模型却能高概率识别。能力与可靠调用能力不同。[P]
“950 个 agent”意味着什么？	949 次会话，最高并发 58；21.5 小时是计算探索的墙钟时间，不含此后的全部实验与论文工作。[P]
Bio 是否只是做药?	公开证据涉及基础发现、药物研发、模型生物/化学训练和 Claude Science。应按多条工作线理解。[S01,S07-S11]
Dario 的判断是不是科学界共识?	他的“压缩百年进步”是有前提的技术预测。公开反馈支持继续研究，但不足以构成功能、临床或时间表的共识。[S03-S05,S16-S18]
SF 到底哪里?	目前没有找到把某个公开门牌与该实验室直接对应的可靠记录。SF 招聘地点不等于实验台地址；Dogpatch 线索归属未证实。[S12,S13,S23]

本报告的分析判断：更值得跟踪的是“从计算候选到可验证实验、再反馈模型”的周转能力。一个功能未知的分子系统不能证明整套模式已经成熟；一套能持续产出、排除错误并积累实验反馈的流程，却可能形成重要能力。

证据标签：事实＝文献或原始材料直接记录；公司判断＝当事方表达；分析＝本报告推论；待证＝公开证据仍不足。“论文记录”不自动等于独立复现。

02 / 时间线：不是一周内突然长出的实验室

时间	可核查事件	应如何理解
2024 年 10 月	Dario 发布《Machines of Loving Grace》。[S03]	以足够强 AI 出现为前提，提出大幅压缩生物医学进步时间的愿景。
2026 年春	官方称基础生物学研究组在此时成立。[S01]	这是研究组成立时间；不能精确等同租约签署、装修或实验室启用日。
2026 年 4 月 3 日	Newcomer 报道收购 Coefficient Bio，交易对价据报超过 4 亿美元、以股票为主。[S06]	收购价格是媒体据文件报道的数字，不是湿实验室建设预算。
2026 年 5 月 7 日	SynBioBeta 公开记录 Eric 谈到招募生物学家、建设湿实验室与加速药物研发。[S07]	9 月的广泛报道不是此前完全没有公开线索。
2026 年 6 月 19 日	John Jumper 宣布将离开 Google DeepMind、转向 Anthropic 的消息被报道。[S24,S25]	“宣布将加入”与“已到岗并领导 ART 项目”不能混用。
2026 年 6 月 30 日	Claude Science 发布活动。[S08]	把科研工具和工作流产品化，与内部自建实验室并行。
2026 年 8 月 18 日	公开蛋白质设计及分析化学结果。[S09]	蛋白结合实验由外部机构执行，证据类型与 ART 论文不同。
2026 年 8 月 27 日	发布 Model Hardware Standard 研究预览。[S11]	开始解决模型连接实验硬件的接口问题。
2026 年 9 月 17 日	公开生物分子模型优化结果，并开放生命科学验证计划公测。[S10,S22]	计算效率、科研准入与外部实验合作同时推进。
2026 年 9 月 18 日	Reuters 报道自有生物实验室及药物研发布局。[S02]	采访提供业务边界与外部合作信息。
2026 年 9 月 23 日	ART 论文及研究组介绍公开。[P,S01]	首次集中展示这支基础生物学团队与发现案例。
9 月 24 日至截止日	科研媒体与研究者开始讨论意义、可重复性和宣传尺度。[S16-S18]	发表时间很短，独立复现和同行审议仍来不及形成。

03 / 团队：可以画能力关系，不能伪造组织架构

官方能够确认的上层结构是 Life Sciences 组织中包含基础生物学发现，以及药物研发、Claude 生物与化学训练相关团队。公开活动称 Eric Kauderer-Abrams 为 Head of Life Sciences。[S01,S08] 以下是人物和能力连接，不代表正式汇报线。

人物/群体	公开依据	有用的判断边界
Eric Kauderer-Abrams	生命科学负责人；ART 六位作者之一。[P,S08]	同时连接研究与业务，但未披露每个实验项目的分工。
Matthew G. Durrant	ART 通讯作者；个人履历记录 Stanford 遗传学博士、IGI、Arc、Stylus Medicine、FutureHouse 等经历。[S14]	有从生物机制发现、工具转化到 AI agent 的连续经验；不能据此推断整个 Bio 的负责人。
Nicholas T. Perry	ART 通讯作者；Arc Hsu Lab 校友页列为转往 Anthropic。[P,S15]	研究团队与前沿基因工具实验室有直接人才联系。
Januka S. Athukorala ge	ART 作者；同样出现在 Hsu Lab 转往 Anthropic 的校友记录中。[P,S15]	支持团队具备生物机制研究积累。
Peter H. Yoon、Emmanuel Ameisen	ART 作者。[P]	论文未给出逐人贡献声明；不凭作者顺序编造实验/工程分工。
Alexander / Alec Tarashansky	Durrant 的公开帖子称 Claude Science 开发由其带领。[S26]	产品开发线索，不能自动归入 ART 实验组。
Coefficient Bio 团队	收购报道及其进入 Eric 团队的描述。[S06]	加强计算生物学与药物研发能力；不说明他们就是某处实验室的租户。
Nathan C. Frey	本人现行网站称其领导 Anthropic 药物发现与开发研究。[S28]	这一范围比“整个 Life Sciences 负责人”更具体；详见增补第 35 节。
John Jumper	离职并拟加入的报道；另有 garden leave 报道。[S24,S25]	截至材料所示，不能直接算作 ART 已在岗作者或实验室主管。

分析：这支队伍的信号不在“六个作者所以只有六个人”，而在知识组合：能提出分子生物学问题的人、能判断序列与结构证据的人、能把研究拆成持续运行任务的人。现有材料不足以给出总人数、预算或完整管理层级。

04 / Dario 的主张，和“共识”的距离

他长期押注的逻辑。2024 年文章把强 AI 想象为大量高水平研究者并行工作；认为生物学的一些瓶颈在于发现关键知识与工具，因此可能在强 AI 出现后的 5-10 年，压缩原本约 50-100 年的进步。他也承认实验耗时、顺序依赖和现实约束。这不是从文章发布日期开始计时的保证。[S03]

这次 ART 发布后的表达。Dario 在公开长帖中承认用途和意义仍不确定、实验仍由人完成，同时用数学能力进步类比生物学，预期未来几年可能加速。原帖直读受限，本报告只采用可交叉核对的公开转载，不把它当作额外实验数据。[S04,S05]

主张	ART 实际提供了什么	仍缺少什么
AI 能贡献新科学线索	记录显示 agent 自主跟进并发现未被提示的重复结构。	广泛领域、预注册任务上的持续产出与独立审查。
并行智能能加速研究	21.5 小时完成一次大规模计算探索。	与同资源人类/传统流程的受控比较；完整实验周期。
模型会发现可转化的新工具	ART 具有令人感兴趣的组合结构。	底物、酶活、机制、可编程性和工程性能。
实验瓶颈将被自动化缓解	自有实验室存在；硬件接口项目存在。	ART 全流程自动执行的展示；通用异常处理能力。
生物学会像数学一样快速提升	给出一个可讨论的早期案例。	能够外推到复杂疾病、动物研究和临床结局的证据。

可以称为较低争议的判断：模型有潜力提高信息处理与部分科研工作效率；关键生物学结论仍需要实验。不能称为已经形成的共识：ART 将成为基因编辑工具、几年的临床进步可以确定压缩到数月、整个生物学将沿同一条指数曲线发展。

这些区分不是否认愿景，而是把愿景变成可以逐步检验的命题。上表第三列是本报告依据论文边界给出的分析，并非对科学界做过系统问卷后的“共识测量”。

增补：2026 年 9 月的新文章中，Dario 已直接使用“未来 5-10 年”治愈大多数重大疾病的表述。不能只用 2024 年文章的条件化时间表概括他现在的公开预测；详见第 37 节。[S31]

05 / Bio 布局：四条工作线与一个反馈回路

工作线	公开观察	分析：为什么要做
基础发现	ART：从公开序列中挖掘未知系统。[P]	获得可追踪的开放式科研任务，检验模型如何发现异常。
药物研发	Eric 公开讨论缩短研发周期；另有纽约 Life Sciences Operator Lead 招聘。[S07,S21]	把模型能力与具体研发项目、运营判断结合。
科学模型与计算	优化结构预测、设计和基因组模型；组织外部设计验证。[S10]	降低科学工具的使用门槛与计算成本。
科研产品与访问	Claude Science、科研支持、Life Sciences Verification Program。[S08,S20,S22]	把内部工作方式逐步变成可分发的能力和产品。

论文之外的一项更直接功能证据。8 月公开的蛋白设计研究涉及 15 个靶标，其中 14 个得到结合命中；不同设置的设计级命中率约 22%-35%，由 Adaptyv Bio 与 Twist Bioscience 制备、测试。它证明的是特定设计任务中得到结合结果，不能直接等同药效、可开发性或临床成功。[S09]

计算成本也在被主动压低。9 月官方报告称四周内优化 30 多个专用模型，平均约四倍加速；简化后的设计流程用远少于此前的 GPU 时间，达到相近的计算评分。这是计算指标与工程结果，不能把相近 ipSAE 评分写成新的等量湿实验成功率。[S10]

本报告的分析：组合起来更像“研究系统＋模型改进＋科研产品”。内部实验既能验证生物学假说，也能暴露 agent 在任务选择、数据读取、工具使用和结论判断上的失败。这些失败记录可能进入评估、训练或产品设计；具体训练数据规模、实际使用方式及增益尚未披露。

两项科学家招聘明确要求把实验或计算工作中的发现、评估和失败反馈给模型/产品团队。[S12,S13] 这比“公司买了一个实验室”更能说明其潜在组织价值，但目前仍不能量化这个反馈回路的回报。

06 / 湿实验室到底做了什么

已确认范围。官方称实验室位于湾区，属于常见分子生物学实验室范围，使用 BSL-1 / BSL-2 条件；不在该实验室研究可感染人类的病原体，实验工作由人类科学家完成。[S01] 这些是公司公开说明，不是本报告实地检查结论。

ART 的真实实验边界。论文新做的湿实验，是把 SA1 相关构建体放入大肠杆菌表达，做小 RNA 测序，确认阵列能够产生有特定边界的 RNA。文章还重新分析了他人公开的噬菌体感染转录组。两者不能合并写成“Anthropic 已在自己的实验室重现自然感染并证明功能”。[P, 原文 pp.8-10、35-36]

环节	谁/什么执行	这篇论文展示到哪里
研究方向与研究简报	人类作者	设定 RT 新系统这一搜索范围。
最初计算探索	Claude Code + Mythos 5 的 agent 框架	按记录自主搜索、跟进、识别阵列、撰写报告。
后续分析	作者在 Claude Science 中指导 Claude	家族定义、系统发育、RNA、结构及评测等。
分子构建与实验执行	人类科学家	表达和 RNA 测序；没有证明编辑能力。
仪器自动控制	非本研究的已展示能力	MHS 是相关接口项目，不能据此补写 ART 自动实验流程。

为何还不是“机器人科学家”？官方解释，探索性分子生物学经常临时改变方法，现阶段不容易直接套用通用自动化。[S01] 这与 MHS 的存在并不矛盾：后者解决软件如何发现和调用硬件能力，不能消除样品准备、设备异常、实验设计和结果歧义。[S11；后半句为分析]

业务边界。Reuters 的采访称其既使用自有设施，也有外部伙伴，并将重点放在临床前工作；公司当时表示不打算自己进行商业临床试验、把药物带到市场。[S02] 这是当前表述，不应推导为永久不涉及任何候选药物或知识产权。

07 / SF 位置线索：能确认到哪一层

结论：公开证据确认“湾区”，尚未确认门牌。“在 SF 的哪里”目前不能给出可靠街道地址，也不能默认一定在旧金山市行政边界内。

公开线索	它支持什么	它不能支持什么	判断
官方使用 Bay Area	实验室的区域位置。[S01]	Mission Bay、Dogpatch、South SF 等具体片区。	强
论文作者统一署 Anthropic, San Francisco, CA	作者的机构归属和城市署名。[P]	所有实验都在这个城市内完成。	强事实，弱定位
实验科学家与计算科学家职位写 San Francisco	岗位的招聘城市，且涉及真实实验工作。[S12,S13]	租户名称、建筑、楼层或仪器所在地点。	中
5 月 SynBioBeta 已公开谈湿实验室	建设/使用实验能力的时间线早于 9 月集中报道。[S07]	实验室确切开门日期或地址。	中
部分前成员来自 Arc / IGI 等湾区机构	人才网络与科研经验。[S14,S15]	实验仍借用这些机构场地。	不构成定位
社交帖子在 Dogpatch 招湿实验人员，提及 RL 环境	确有相似主题的第三方招聘帖子。[S23]	发帖方就是 Anthropic；两条新闻属于同一实验室。	归属未证实
公司总部、注册地址或邮件地址	行政或办公地址，若另经核实。	实验室地址。	不能替代

对 Dogpatch 线索的处理。已看到 Andrew Ho 名下的相关公开帖子及转载，内容提到湿实验人员、驻场与 RL 工作。但未找到把这条招聘与 Anthropic 连起来的可靠当事方说明。因此本报告保留它作为“已检查但未成立的关联”，不给它一个虚假的概率。

检索范围与局限。核查了官方公告、招聘、作者履历、新闻与可检索的商业地产线索；没有找到明确把某一门牌绑定到该实验室的公开记录。这不是“所有许可数据库都已查完”，也不是“地址不存在”。租约可能由关联实体或服务商承接，公开网页也可能尚未索引。

什么证据可以升级判断？当事方明确披露；地址一致的实验室岗位或场地介绍；能同时对应主体与用途的公开租赁/许可记录。只有邻近科研园区、总部地址或模糊照片，都不足以定址。

08 / 论文是什么，为什么会选逆转录酶

论文：Autonomous AI agents discover reverse transcriptases with tandem repeat arrays。作者 Peter H. Yoon、Januka S. Athukoralage、Emmanuel Ameisen、Eric Kauderer-Abrams、Nicholas T. Perry、Matthew G. Durrant；2026 年 9 月公开。[P]

摘要的完整逻辑。作者让通用语言模型 agent 搜索大量宏基因组数据中的逆转录酶新系统。agent 发现一类主要存在于巨型噬菌体中的 RT，旁边有串联重复阵列与伙伴蛋白；作者把它称为 ART。后续分析显示阵列会表达 RNA，RT 有特殊 N 端结构，伙伴蛋白分成不同类型。作者同时研究模型如何发现重复、为什么常常错过它，并承认该系统的具体功能仍未知。

术语	给非生物学研究者的解释	本研究为什么关心
逆转录酶 RT	以 RNA 为模板合成 DNA 的酶类。	天然系统中常与特殊 RNA 和其他蛋白协作，可能产生新机制。
宏基因组	从环境等样本得到的混合遗传序列。	大量尚未命名、培养或深入研究的分子系统藏在其中。
contig	拼接得到的一段连续 DNA 序列。	它可能断在基因或阵列中间，缺失邻域不等于自然界没有邻域。
retron	RT 与特殊非编码 RNA 等构成的一类系统。	为“去看 RT 上游 RNA”提供先验。
噬菌体 / jumbo phage	感染细菌的病毒 / 基因组较大的噬菌体。	ART 主要出现在这一背景，不能直接套用人类疾病机制。
重复与间隔序列	相似短片段重复出现，中间隔着不同序列。	布局让人联想到 CRISPR，但相似布局不决定相同功能。
伙伴蛋白	与 RT 在基因组上相关联的其他蛋白。	可能共同工作，必须与偶然相邻区分。

引言的出发点。自然界已经给过人类限制酶、PCR 相关酶、CRISPR 等工具。规模化序列数据使搜索空间很大，但“哪里值得继续看”需要判断。这篇论文试图把开放式探索和专用生物信息工具连起来，而不是训练一个只输出某种序列标签的分类器。

09 / 自主研究框架：四种角色，五道阶段门

运行边界。每个 agent 是配备 Mythos 5 的 Claude Code 会话；使用 60 核 CPU、192 GiB 内存的沙箱，最高并发 58。沙箱没有本地 GPU，但可提交外部 GPU 队列，实际运行了 19 个 A100 / L4 作业。因此“整个研究不用 GPU”是错误读法。[P, 原文 p.28]

角色	职责	为什么重要
Worker	制定计划、运行工具、提交结果与代码数据。	负责执行与产生新跟进线索。
Supervisor	审阅计划、总结和改动，接受或要求修改。	让中间结果经过第二次检查。
Curator	将完成任务的发现写入共享知识库。	降低不同会话互相重复或丢失上下文。
Editor	审阅最终研究报告。	提高报告质量，但不等于外部科学审稿。

阶段顺序：输入整理 → 数据库扫描 → RT 分类 → 邻域普查 → 深入研究。上一步通过脚本检查后才能开启下一阶段。论文流程图把建 profile、搜索和整理拆开画，因而视觉步骤数与这五个阶段不完全相同。

119 个任务中 49 个至少修改过一次；计划、结果、审阅和脚本进入共同可读的版本记录。跟进任务先排队，再由框架放行或拒绝；最终没有可派发任务时结束。

自主不等于零人工基础设施。作者提供研究简报、数据库接口、工具和约 140 份技能指南。agent 实际查阅 15 份，其中七份针对宏基因组研究：独立出现次数、操纵子信号、新候选判定、新颖性反证、未知家族注释、机制及反事实、邻域作图。[P]

分析：这是一套带研究方法先验和审查机制的工作环境。应比较“模型＋工具＋数据＋框架”的整体表现；不能把全部进展归因于裸模型，也不能因为方法由人写过就否认运行期间的自主决策。

10 / 搜索漏斗：从 19.39 亿簇到少数候选

层次	论文记录	解读
原始来源	约 1,100 万 biosamples；Logan、ENA、JGI、NCBI 等。	不是 1,100 万种独立生物。
清理后的蛋白	约 158 亿条。	经过完整性、长度、复杂度等过滤。
搜索空间	1,939,242,578 个 50% 相似度蛋白簇。	“19 亿”是簇数，不是 agent 逐字读过的 DNA 条数。
HMM 初筛	52 个 profile，3,391,714 个 RT 候选。	6 个 Pfam、45 个 myRT、1 个文献构建 profile。
保留并聚类	203,381 条序列 → 198,290 个 RT 簇。	过滤碎片和弱匹配后再聚类。
分类	35,168 已知；25,737 暂称 novel；137,385 unplaced。	“novel”是计算分组，不是 25,737 项验证后的新发现。
邻域采样	选 7,308 个锚点；取回 7,238 个，合计 10,983 个位点。	没有对所有 RT 做同深度调查。
邻域蛋白	79,680 条编码序列；3,564 个重复出现的家族。	85.9% 的邻域窗口被 contig 末端截断。
晋级	46 个过筛，其中 13 个控制；其余 33 个中选 16 个，跟进再增 1 个。	后续仍需逐个排除已知系统和假相关。
最后产物	三项新伙伴关联、三条新 RT 谱系；共 19 份报告。	不同计数口径，不应相加成六项已验证分子功能。

已知类别数目。retron 11,517；UG 10,759；group II intron 5,461；DGR 3,857；CRISPR 相关 2,067；group II-like 827；Abi 680。

伙伴晋级原则。需要谱系内保守、独立样本出现，以及相较随机邻居更贴近 RT 的基因排列信号；还检查方向与间隔。原始筛选对不同 RT 类别有要求，这会影响单一谱系内真实伙伴的可见性。

分析：这是有策略的巨大空间抽样，不是“穷尽全球 DNA”。大量未分类簇、截断序列和采样门槛意味着假阴性难以估计。

11 / 最有价值的过程：一条被否定的线索

第一步：从错误的伙伴关联出发。一个未知基因离 RT 仅 11 bp，进入候选。深入分析发现它其实是巨型噬菌体的非病毒颗粒 RNA 聚合酶 β 样亚基；相邻更可能反映基因组排列，而不是新的专属 RT 伙伴。[P, 原文 pp.5-6、补图 2]

第二步：否定关联，但保留异常。worker 仍觉得“巨型噬菌体中一个独立、retron-like 的 RT”值得追。supervisor 提出进一步检查 RT 上游是否有 retron 式非编码 RNA。任务没有提到“重复”或“阵列”。

第三步：换到信息更完整的亲缘位点。最初位点上游可用序列中位数仅 27 bp，检索到的近缘位点中位数约 940 bp。一个 L0050 位点来自 228,907 bp 的 contig，worker 读入其 2,900 nt 上游 DNA。

第四步：读序列时发现模式。下一条记录中，模型描述了串联重复。此前没有运行既有重复识别程序，读取和识别之间没有工具调用。随后才编写脚本计数，识别到约 16 nt 的重复、14 个拷贝及约 100-200 nt 的不同间隔区；另一位点也出现相似布局。

第五步：主动尝试推翻新颖性。agent 检查它是不是已知 CRISPR、retron、DGR 或其他 RT 系统；检查间隔区是否像编码序列、能否形成稳定 RNA，以及 RT 与聚合酶能否独立出现。

agent 报告中的检查	记录结果	边界
42 个间隔序列查蛋白匹配	未找到符合阈值的匹配。	不等于证明全部无编码功能。
40 个间隔区的 RNA 折叠对照	29 个比保持二核苷酸组成的打乱序列更稳定。	计算预测，不是实验结构。
伙伴出现	76 个位点中 52 个带伙伴。	来自这一阶段家族口径，不能与后续 95 位点直接混算。
聚合酶与 RT 是否绑定	39 个相关噬菌体中 36 个不含 RT。	削弱“必需聚合酶亚系统”的解释。

作者说明上述部分比例取自 agent 报告，后续没有重新计算。后来的系统性家族分析由作者指导完成。

12 / ART 家族：真正新的是组合结构

定义结果。作者后续建立含 95 个代表的 ART 家族，代表间按约 90% 序列相似度去冗余。其中 28 个位点能明确界定邻近重复阵列；不少序列的上游不完整，不能把其余 67 个一律算成“没有阵列”。[P, 原文 pp.6-10、32-34]

组成	观测	证据状态
RT 核心	与已知逆转录酶核心结构相似；覆盖催化区域的 93 个成员均保留 YxDD 类基序，另两条为截断。	支持 RT 身份；未直接测酶活。
额外 N 端	典型约 180 aa，明显长于多数参考 RT 的短 N 端。	序列/结构分析支持；其具体作用待实验。
上游阵列	约 0.3-4.1 kb，3-21 个重复拷贝；重复 15-49 nt，常有约 15 nt 回文核心。	基因组序列证据。
间隔序列	常见约 120-220 nt；阵列内彼此不同、间距较规则。	与通常约 30 nt 的 CRISPR spacer 明显不同。
伙伴蛋白	存在三类主要伙伴，常与 RT 相邻。	共同出现和结构预测支持关联，不等于已测物理结合。

与 NART 的比较。另有一支与 ART 相近、同样带较长 N 端的 RT，作者称 NART。其可评估位点未检测到 ART 式阵列，也缺少相应伙伴；某些位点与噬菌体聚合酶相邻。长 N 端本身不足以定义 ART。

进化关系。774 个 RT 域的系统发育分析支持 ART 与 NART 为相邻分支，并与 retron 相关谱系相区别。不同分析用不同参考集、去冗余程度和可评估子集，因此图上可能出现 71、95、98、103 等数目；这些不是同一个分母的矛盾结果。

关键解读。底层噬菌体 RT 之前已被观察过。新增知识是把 RT、特殊 N 端、重复阵列及伙伴组织成一个可研究的新系统家族，并补上表达与比较证据。科学新颖性无需等于“从未有人见过任何组成部件”。

13 / RNA 证据：论文最实在的实验进展

自然感染数据来自公开旧数据。作者重新分析噬菌体 SA1 感染 Staphylococcus lentus 的转录组，BioProject PRJNA836150。数据包括三个培养重复、0/5/15/55 分钟四个时间点，共 12 个文库。[P, 原文 pp.8-10、35-36；补图 6]

感染后阵列区域强烈转录；在 15 分钟时，其 RNA 可占噬菌体来源 RNA 的约 8%。这是“噬菌体转录本”的比例，不能写成整个宿主细胞 RNA 的 8%。阵列出现几个边界较明确的 RNA 物种，而不是只有模糊背景读段。

Anthropic 新做的验证。将相关序列放入 E. coli DH10B，比较天然位点构建体与异源启动子构建体，进行小 RNA 测序，看到对应 RNA。综合两类数据界定的三种主要 RNA 长度为 165、201、166 nt：RNA 1 的边界来自异源表达文库，RNA 2、3 来自自然感染数据。预测结构中，重复序列有助于形成茎结构。

可以支持	不能据此声称
阵列在自然感染背景中被转录。	它一定通过哪一种加工酶成熟。
在异源细菌表达背景也能出现特定 RNA。	这些 RNA 一定是 ART 的逆转录底物。
某些边界和长度具有一致性。	已经发现 DNA 合成产物或测定催化动力学。
系统值得继续做机制研究。	它具备 RNA 引导的基因编辑或治疗功能。

一个容易漏掉的设计限制。新构建体实验每个条件只有一个培养样本。它与自然感染数据中的三个培养重复不是同一实验，不能互相借用重复数。这不抹去表达观测，但限制了变异估计与稳健性判断。

本报告的分析：这一步把“序列布局看起来有趣”推进到“确实存在 RNA 表达产物”。要跨向分子机器的功能证明，仍需直接连接 RT、RNA、伙伴与特定产物/表型。论文没有完成这一整条因果链。

14 / 伙伴蛋白与机制假说

类型	主要特征	已知程度
I	在 95 个 ART 代表中有 59 个对应位点；蛋白约 600 aa，预测有 GNAT-like 结构。	没有证明它是乙酰转移酶。
II	约 270 aa、以 α 螺旋为主；在相应葡萄球菌噬菌体分支出现。	与 I 型无明确序列/结构同源。
III	约 170 aa、螺旋结构；部分环境序列中出现，另伴随第二个基因。	功能未知；不能仅凭位置指定酶学作用。

共同出现。作者反向检索各伙伴家族，发现它们倾向于在 ART 背景出现，降低“普通邻居”的解释。不同伙伴类型与 RT N 端分组也有对应关系。[P, 原文 pp.10-11、补图 7]

结构预测。AlphaFold2 multimer / Boltz-2 的模型提示伙伴可能接触 RT N 端。部分界面置信度约 ipTM 0.6，属于值得跟进的模型信号；不是纯化后的结合实验、显微结构或酶活依赖性证据。

GNAT 名字容易误导。补图 7 专门比较辅酶结合区域：预测伙伴与已知 GNAT 的关键位点/空间不完全相容，存在位阻和基序缺失。论文因此没有证明 I 型通过乙酰化发挥作用。

共进化的谨慎读法。379 对比较中，伙伴相似度与 N 端相似度的 Spearman ρ≈0.35，与 RT 核心相似度的 ρ≈0.55。这些成对点并不独立。不能把图改写成“N 端比核心更强地共进化”，也不能从相关直接推出接触机制。

作者提出的工作假说。同一 RT 和伙伴可能配合多个不同的阵列 RNA，形成具有不同作用的复合物；retron 等系统为此提供类比。这个想法解释了“为什么要保存多个不同间隔序列”，但目前不知道触发条件、底物、产物或对噬菌体的好处。

本页机制描述均为论文提出或模型预测。最有价值的下一步是建立可反驳的功能模型，而不是给未知伙伴提前贴一个确定酶名。

15 / 到底与 CRISPR 有多像?

维度	ART 论文中的证据	与 CRISPR 类比的边界
基因组布局	短重复与不同间隔区交替，邻近蛋白基因。	布局相似可生成假说，不能决定功能。
主要酶	逆转录酶样核心，带额外 N 端。	不等同 Cas 核酸酶。
间隔区	常约 120-220 nt；相关噬菌体间顺序较保守。	与典型短 spacer 的快速获取/丢失模式有区别。
RNA	已观察阵列转录及特定 RNA 边界。	没有证明 RNA 指定外源靶标。
蛋白伴侣	三类未知伙伴，预测可能参与复合物。	附近未见典型 Cas 装置。
可编程性	暂无直接实验证据。	不能宣称已能改变靶点、编辑基因。
应用	作者与公司提出潜在工具可能。	与可用研究试剂、治疗方案相距多个验证阶段。

不是“功能未知，所以没有科学价值”。界定一个反复出现、拥有一致结构特征和表达产物的系统，可以构成有价值的发现。很多机制研究从这一层起步。评价其科学贡献，应问证据是否排除了偶然布局与注释错误，而不是要求第一篇文章就完成临床转化。

也不是“曾经 CRISPR 起初未知，所以 ART 必然伟大”。这是幸存者偏差：许多有趣系统后来没有成为通用工具。需要机制和工程性能逐步缩小不确定性。

同期研究。论文讨论了 Minerva 工作：用基因组语言模型挖掘原核非编码元件及相互作用，在不同的 RT 背景（UG27）发现相关阵列现象；参考文献 56 列出该预印本。[P,S19] 两项工作说明这个问题有多种切入方式，不等于发现完全相同，更不构成抄袭判断。

分析：通用模型的强项可能是把文献、工具、异常和下一步问题连起来；专用基因组/结构模型提供另一类密集信号。论文没有证明前者取代后者，实际研究过程恰好大量使用后者。

16 / 可重复性：最应该出现在标题旁的限制

作者又运行了十次同类自主探索，十次都没有识别 ART 阵列。大多数完成邻域普查的运行接触过 ART 位点，其中两次进一步研究了该谱系，但没有一次去读取决定性的 RT 上游 DNA。[P, 原文 pp.10、13、38]

这是论文主动报告的重要负结果。它指向的瓶颈是探索路径：模型可能拥有识别模式的能力，但在广阔任务里未必把注意力放到正确对象上。

证据	能推断什么	不能推断什么
最初一次成功	系统具备产生这类发现的能力。	同样输入下一次必然成功。
随后十次漏掉	当前流程对该发现不稳健。	整套系统在所有科学任务的成功率就是 0%。
定向提供 DNA 时较高识别率	识别能力与检索/注意力选择可分开。	大搜索空间里的发现率同样高。
识别日志检索	检查 3,084 个任务记录、5,632 个会话。	不存在任何其他有价值发现。

不要计算一个伪精确的“9.1% 成功率”。最初案例是成功后选中的研究对象；重跑关注是否重现同一发现。它不是预注册、随机抽取的十一项不同科学任务。1/11 可以描述这组运行事实，不能当作通用科研成功率或直接计算发现成本。

日志检查本身也有范围。作者围绕首轮归入家族的 130 个 RT 与 171 个 contig 标识检索日志。此方法能追踪已知路径，但对采用不同标识、未落入首轮集合的潜在线索不保证完全覆盖。

分析：下一代系统应把“什么时候必须看原始数据”变成可评估策略，而不是简单增加 agent 数量。更广的搜索、更多报告或更高 token 消耗，都不自动等于更可靠的科学观察。

17 / 3,500 次定向评测：工具为什么会让它漏看

作者用七个 Claude 模型、五种输入/工具条件，每条件 100 次，合计 3,500 次尝试；每次输出上限一百万 token。[P, 原文 pp.10-13、38-39]

级别	输入与能力
1	上下文内给两条 RT 蛋白序列。
2	上下文内给两个位点的 DNA。
3	通过文件提供全部 96 个位点，配离线生物信息工具。
4	再增加 1,131 个结构，以及 Foldseek、PyMOL、US-align 等。
5	再开放文献、网络与软件安装。

核心结果。最强四个模型在 DNA 直接位于上下文时，至少 90% 尝试识别阵列；同类信息转成文件并增加工具后，某些条件下降到 32%。工具增强了伙伴和结构研究，却可能把注意力带离非编码原始序列。

文件条件下，强模型的 39% 尝试从未连续读取至少 200 nt；这往往连一个完整重复单元都看不全。读取更长 DNA 与识别率提高相关：汇总观测从约 29% 升到最高约 76%，Mythos 5 的某些分组可到 96%。但这些是行为分组相关，不是随机干预后的因果效应。

评价对象。十项二元主张覆盖上游非编码候选/阵列/变体、RT 分类与结构，以及伙伴类型与结构；由 Mythos 5 按固定规则判定。模型包括 Opus 5.5、Mythos 5.1、Mythos 5、Opus 5，以及 Opus 4.6、Opus 4.8、Sonnet 5。

三个限制。第一，只比较 Claude 家族，没有跨厂商结论。第二，同系列模型承担评分，未见全面独立人工盲评。第三，题目是从已经知道答案的 ART 案例倒构出来的，适合分解能力，不等同真实新发现 benchmark。

分析：企业部署里“工具更多”可能同时带来收益和路径偏差。应观测原始数据覆盖、关键证据遗漏、反证质量，而非只看是否能连续调用大量工具。

18 / 可解释性：看到了重复，不等于理解了机制

论文用两类分析判断模型是否直接利用重复序列信息。[P, 原文 pp.12-14、39-40]

专用基因组模型。在同一段 L0050 序列上分析 Evo 2 7B 与 gLM2 650M 的预测。前者按下一个碱基预测，遇到若干重复后信号更清楚；后者可以双向使用上下文，对第一份重复也能显示信息。把每份重复内部打乱后，对应模式减弱。

通用模型的内部特征。作者用原发现日志重新构建输入，在同检查点、另一宿主环境中重放。分析一个层的稀疏字典特征，逐 token 只记录激活最高的 128 个特征；未记录的值按零处理。

先用合成重复序列挑出 12 个候选特征，再看 L0050 后选出两个展示。一个偏向一般逐字重复，读到约第三份后减弱；另一个在重复模式中持续出现。逐份打乱重复后，两个特征在 14 份中的 12 份、14 份上分别被压低。

观察	合理含义	需要避免的扩大解释
真实重复处激活高	模型内部确实表征了相关重复模式。	因此知道这套分子的自然功能。
打乱后信号降低	激活与序列顺序结构有关。	因此这个特征导致了科学发现。
数字、字母重复也能激活	可能是一般符号/重复检测能力。	这是专门的 DNA 生物学模块。
重放原日志	可以检查当时输入与模型表示的关系。	完全复原了生产会话的所有条件。

方法限制。原系统提示与工具定义没有保留，重放时用占位替代；只有一个层；特征选择含观察后的筛选；未通过特征消融证明因果性。这些限制不否定图形，但会限制“模型如何科学思考”的结论强度。

分析：最强证据是“模型确实看到了文本中的规律并继续追踪它”；不是“我们已经定位了自主科学发现的神经机制”。

19 / 资源账本：215.6M tokens 不能直接换算美元

指标	论文数值	解释
墙钟时间	21.5 小时	一次自主计算探索的经过时间。
总会话	949	1 启动＋414 worker＋375 supervisor＋107 curator＋52 editor。
累计 agent-hours	76.9，约写作 77	包含会话内工具执行；其中 worker 63.7。
任务状态	107 完成、10 分流拒绝、2 停滞	119 总任务；停滞原因涉及多次修改/检查未过。
工具活动	7,578 shell 命令、696 数据库查询、131 文献搜索、61 web 请求	反映系统调用规模，不等于独立科学检验数。
报告评比	19 份、342 次有序两两比较	ART 排第三，36 局中赢 32 局。
未缓存输入	11.3M tokens	只是输入成本的一部分。
输出	14.9M tokens	不是全部 215.6M 都按输出价格计费。
写缓存	189.5M tokens	论文合计口径包含该项。
读缓存	未计入 215.6M	总量和对应收费没有在此给齐。

三个 token 分项是取整数，表面相加为 215.7M，与作者汇总 215.6M 的微小差异来自报告精度；不应据此指控计数矛盾。

为什么不能给一个可信的总成本？缺少全部读缓存量、内部/外部模型计费口径、数据库建设与服务成本、完整 GPU 队列成本、实验成本和人工时间。把总 token 乘一个公开输出单价，会系统性混淆计费类别。

分析建议：真正有用的单位经济指标。分别测“每个经专家确认的新假说成本”“每个实验通过的机制结论成本”“每次完整迭代耗时”，并把失败实验、探索失败、数据库维护摊进去。最终药物的成本还包含后续开发与临床，不能从这张计算账本外推。

报告排序器使用 impact / novelty / soundness / actionability 权重 0.35 / 0.30 / 0.25 / 0.10；低可靠性评分自动判负，再以 Bradley-Terry 排序。这是框架内部的优先级工具，不是市场价值评估或同行认可度。

20 / 全部主图导读

原图 / 页码	图中在回答什么	正确的阅读重点
Figure 1 / p.4	大规模 RT 搜索及自主任务如何组织，出现了哪些研究分支。	图区分原始任务与 agent 追加任务；ART 是深挖分支中的发现，不是输入就指定的终点。
Figure 2 / p.7	ART 的基因组布局、家族、RT/N 端特征与阵列结构。	先看代表位点和重复间距，再看各类 RT 比较；不能只截一张类似 CRISPR 的示意图。
Figure 3 / p.9	SA1 感染中的 RNA 表达、RNA 结构、伙伴类型、预测复合物及功能假说。	A-E 是感染背景与 RNA 证据，F-I 是伙伴和机制模型；虚线箭头表示假说步骤。
Figure 4 / p.11	五级输入条件下，七个模型对十项主张的识别，以及读取 DNA 与识别阵列的关系。	定向评测不是开放式发现率；相关分组与置信区间要一起看。
Figure 5 / p.12	专用模型与通用模型对重复序列的信号。	区分预测似然、稀疏特征激活与实际因果机制，不能只凭激活热图宣称“理解生命”。

图不能脱离正文读。图中不同小面板使用的集合不总相同。比如家族 95 个代表、系统发育去冗余子集、可评估上游窗口、天然感染文库、人工构建体实验，各自对应不同分母。

三个常见误读。第一，把结构预测图当作实测复合物结构。第二，把 transcript coverage 当成 RNA 具有指定生物功能。第三，把直接给 DNA 的识别率当作从数十亿序列里自主发现同一系统的概率。

完整高清主图保留在后附原论文中；本报告不重绘坐标或重新估算图中点值。正文数据以作者明确给出的数值为准。

核对提示。图 3 集中展示伙伴与假说；图 4 集中展示模型评测；新构建体实验的条件还必须查方法页。后面的覆盖表给出原文页面入口，避免仅凭单个示意图得出全论文结论。

21 / 补图 1-2：负结果是这篇论文的一部分

补图 1（原文 p.15）列出 17 个晋级伙伴家族的最终去向。保留的“新关联”仍是报告层面的关联成立，不是每一个都完成了实验功能验证。

类别	家族	最终解释
伪影或旁观者	Peptidase M50	邻近误判为 RT 的基因。
伪影或旁观者	Unannotated #164	RT 基因的片段。
伪影或旁观者	PPK2	重复测序菌株造成独立性问题。
伪影或旁观者	Unannotated #506	巨型噬菌体 RNA 聚合酶亚基；由此分支引出 ART。
已知系统	DnaG-type primase	UG6 已知 primase。
已知系统	DUF6088	已知 AbiEi 抗毒素。
已知系统	Unannotated #79	已知 CRISPR 核酸酶 Csx3。
一般邻域成分	YhcG nuclease；MUG113 GIY-YIG	一般防御核酸酶；另一个防御系统。
一般邻域成分	DUF2971；BRCT	防御岛邻居；其他防御基因。
一般邻域成分	MscS channel；DUF2726；MoxR ATPase	一般膜通道、邻居或 ATPase。
保留的新关联	DEDDh exonuclease	与 UG6 RT 关联，图示 11 个盒式基因单元。
保留的新关联	ThiF adenylase	与 retron / CRISPR RT 关联，图示 retron 55 个位点。
保留的新关联	vWA / DUF2201	与 retron RT 关联；含 MoxR-vWA 模块，图示 44 位点。

同图还展示 RT-DEDDh 融合谱系（31 条 contig）及其他新谱系背景。正文所说的三条新 RT 谱系是 ART、TPR 融合 retron RT、RT-DEDDh 融合；与“三项新伙伴关联”不是同一列表。

补图 2（p.16）还原从未知邻居、聚合酶误关联、近缘位点长上游区域，到直接识别重复的路径；并画出额外阵列和 NART 对照。其价值在于记录发现如何转弯，也暴露了成功对某次观察路径的依赖。

22 / 补图 3-7：把支持与限制一起读

补图 / 原页	内容覆盖	研究含义与边界
S3 / p.17	95 个 ART 成员树；774 个 RT 域的关系；多个位点示意。	支持家族不是单一序列偶然现象。上游不足 1.5 kb 的成员标为未评估，不能当作阴性。
S4 / p.18	YxDD 催化基序；与已知 retron、DGR、group II intron RT 的结构叠合；16 个已解结构的比较。	支持 RT 核心保守；结构相似和保留基序仍不等于测得催化。
S5 / p.19	ART / NART / 近缘 retron 的 N 端长度、序列与结构比较；阵列、长非编码区、伙伴和聚合酶的分布。	“长 N 端”不是唯一判据。低置信结构及不同边界定义需要注意。
S6 / p.20	SA1 感染 5、15、55 分钟 RNA 覆盖；阵列与 RT/伙伴转录量；七个 II 型阵列 RNA 预测折叠。	加强真实表达背景；预测折叠不是 RNA 实验结构。
S7 / p.21	三类伙伴全库上下文、预测结构、GNAT 位点、长度、N 端与伙伴共变。	伙伴关联证据较丰富；乙酰转移酶功能和直接结合仍未证实。

S5 为什么会看到“ART 103”而不总是 95？若比较涉及系统发育、扩展参考集或可预测结构的集合，作者使用不同子集。读图时应以该面板的集合和方法定义为准，而不能把所有数值套入主家族 95 的分母。

S7 为什么值得单独看？它主动展示 GNAT-like 结构与典型乙酰辅酶 A 结合不吻合的地方，避免把一个结构名称变成未经验证的功能。对投资或产品叙事而言，保留这种否定信息尤其重要。

补充材料范围。用户提供的文件含七张补图、103 条参考文献和十三页方法。方法引用了 Supplementary Note 1 研究简报，但该 Note 未附在这份 40 页文件内。本报告不能把没有取得的简报全文、全部原始日志或代码仓库宣称为已审阅。

23 / 方法全覆盖 I：数据、搜索与研究记录

方法模块 / 原页	做法摘要	对结论的影响
自主框架 / 28	五阶段门、任务计划与审阅、共享知识库、版本记录。	能追溯研究过程，但方法描述不等于完整可执行框架已公开。
工具与技能 / 28	序列、结构、RNA、文献工具；19 个外部 GPU 作业；15 份被查阅技能。	发现来自通用模型与专门基础设施协作。
数据库 / 28-29	基因预测、过滤；90%、70%、50% 分级聚类；代表与 contig、坐标、样本关联。	数据预处理和可追溯接口是先决条件。
RT 普查 / 29	52 个 HMM；完整度、类别长度和催化基序相关过滤；三层分类。	阈值先对参考序列校准；仍可能遗漏高度发散系统。
邻域采样 / 29-30	按类别设上限；部分全采、部分随机或分层；两侧最多各 10 kb。	类别覆盖不均、短 contig 严重影响可见性。
伙伴打分 / 30	保守性、独立出现、相对位置置换检验及方向/间隔约束。	通过统计门槛不等于确认蛋白相互作用。
深入研究 / 30	重做标准、对齐、查注释与文献。	“confirmed”在此指报告维持新关联判断。
资源核算 / 30	会话时长、任务与 token 归属；跟进来源有重复分类及人工解分歧。	需理解缓存与工具时长的统计口径。
报告排名 / 30-31	342 次有序比较，固定权重、可靠性低分淘汰、Bradley-Terry。	可整理研究优先级，不是外部评价。
发现日志 / 31	定位首次识别事件，检查前序工具输入、重数所称重复。	支持未被“重复”提示且直接读序列识别。
agent 的阵列分析 / 31	家族搜索、重复计数、蛋白匹配、RNA 折叠打乱对照、文献反证。	部分结果直接采自 agent 报告，没有另行重算。
人类指导 / 31	首轮 campaign 后由作者在 Claude Science 指导后续代码与分析。	必须与最初自主探索分开归因。

这是方法的研究性解读，不是逐字译文。完整参数、软件版本及引文见原文附录；本报告保留会改变证据判断的关键设置。

24 / 方法全覆盖 II：如何建立分子系统证据

方法模块 / 原页	做法摘要	应注意的限制
家族定义 / 31-32	从 MarsHill RT 搜索，同源图聚合、补充簇成员、90% 去冗余并建树，得到 95 个代表。	家族边界结合伙伴和上游非编码区确定；不是完全无先验的天然分类。
重复检测与边界 / 32-33	上游 k-mer 扫描、规则间距、打乱序列对照；再确定重复长度、拷贝及间隔。	检测依赖窗口和阈值；一个三拷贝位点通过额外精确短词规则保留。
参考 RT 集 / 33	957 条用于 N 端等比较；774 条用于 RT 系统发育。	不要混用不同集的比例。
系统发育 / 33	MAFFT 对齐、trimAl 保留 254 列、IQ-TREE；模型选择与两种支持度。	短的保守核心提供关系信息，不直接提供机制。
N 端定义 / 33	通常取 RVT_1 域之前；部分图在域边界前再移 40 aa。	同一术语在指定图中边界不同。
催化基序 / 33	定位 YxDD 类模式，比较频率和信息量。	基序保守是功能线索，不是活性测量。
结构预测、比较及 N 端序列 / 34	预测 RT 和 N 端，与已知结构比较；计算最近邻的核心与 N 端相似度。	置信度、结构边界及归一化影响数值。
谱系位点特征 / 34	仅在可评估位点计数阵列、长非编码区、伙伴和聚合酶。	不同分母不可直接比较缺失比例。
SA1/MarsHill 与阵列比较 / 34-35	逐份标注重复；比较两个噬菌体位点、间隔序列与 RT；剔除近重复阵列。	以核心词对齐会使该词处保守性按定义很高。
RNA 边界及折叠 / 36	按覆盖阈值定义 RNA，再预测最小自由能结构。	边界受规则影响；结构仅为预测。
公开转录组 / 35	重新比对 SA1 感染 RNA-seq，检查覆盖与丰度。	原实验是外部既有数据。
表达与小 RNA 测序 / 35-36	两种构建背景，表达后确定小 RNA 边界和长度。	每个构建体一个培养样本；功能链条仍不完整。

分析：这些方法把“发现一段重复”推进成一个有家族边界、有对照、有表达证据的研究对象。它们主要回答系统是否存在、是否一致、是否值得做机制实验；仍没有回答其产物是什么、如何影响细胞或噬菌体。

25 / 方法全覆盖 III：伙伴、模型与评测

模块 / 原页	方法要点	结论边界
伙伴基因与分类 / 36-37	根据基因邻域、序列、结构与谱系，把主要伙伴分成三类。	相邻和共现不等于直接作用。
伙伴反向检索 / 37	在整个数据库搜索伙伴，分类其邻近 RT 或其他背景。	可支持专属性；受基因预测、contig 完整性影响。
复合物预测 / 37-38	AlphaFold2 multimer / Boltz-2 预测伙伴与 RT 的相互位置，含错配对照。	界面置信度有限，缺少实测结合。
GNAT-like 位点比较 / 38	与已知 GNAT 结构、辅酶空间、相关基序比较。	不支持简单指定乙酰转移酶活性。
N 端与伙伴比较 / 37-38	序列相似度、聚类及 379 对相关性。	共享谱系导致点间依赖；不作因果解释。
十次重跑 / 38	同简报与框架，检查已知 ART 标识对应的任务、会话。	未重现阵列识别；并非全部发现能力的总体估计。
五级 benchmark / 38-39	七模型各 100 次、文件/上下文/工具条件递增，固定主张评分。	只覆盖 Claude；同系列模型评分；针对既知案例构造。
DNA 读取统计 / 39	从工具输出和上下文判断读到的连续序列长度，再关联识别结果。	相关不证明“强制读取”能获得同样幅度增益。
基因组模型分数 / 39	Evo 2 与 gLM2 对重复及打乱序列的预测分析。	反映序列规律，不证明识别系统功能。
稀疏特征 / 39-40	合成示例筛选、日志重放、单层 top-128 激活、逐份打乱对照。	系统/工具上下文替代，特征后选，无因果消融。

可复核性的实际状态。文中给出大量方法和引用，足以理解分析设计；但本次取得的文件并未包含被引用的研究简报全文、全部 949 个会话原始记录、完整数据库快照或每个中间产物。本报告没有独立重跑这些规模化分析。

需要分开三件事：作者描述了怎么做；我们检查了描述与结果是否自洽；第三方用可取得的数据与代码复现了结果。当前报告主要完成前两者，不能冒充第三者。

26 / 外界反馈：赞同、质疑与来源权重

来源	公开反馈	应给多大权重
Feng Zhang，经官方公告转述	认为 RNA 重复阵列与 RT 的关联值得进一步研究，肯定 AI 参与发现的潜力。[S01]	领域专家正面判断；他也列在论文早期反馈致谢中，不是独立复现实验。
The Scientist	报道研究组与发现，并提示功能尚待确认、科研资源和访问差异。[S16]	专业媒体的背景与问题意识；不是同行评议报告。
Ravid Shwartz-Ziv，经 Decrypt 引述	质疑将意义未知的发现包装成重大突破，尤其对 PhD 水平等叙事持保留。[S17]	关于证据与宣传尺度的批评；不等于完成生物学反证。
Owen Lewis，经同篇报道引述	更重视科研过程的意义，同时保留对分子用途的判断。[S17]	说明乐观可以落在工作流，而不必预判 ART 功能。
Gautam Parab 的独立解读	聚焦首轮成功、十次重跑遗漏，以及模型是否实际读取 DNA。[S18]	有可对照原文的批评；个人评论，不应包装成机构共识。
Bloomberg / Robert Langreth	可读取导语称部分专家认为重要性被夸大。[S27]	正文未完整取得，不据此编造专家姓名、详细论据或一致结论。

反馈的时间窗口很短。论文 9 月 23 日公开，本报告截止 9 月 25 日。此时可见的是早期评价，不是数月后的复现、会议讨论或正式审稿结论。对“业内都震惊”“生物学家普遍不认可”这类说法，现有来源都不够。

本报告对两边观点的判断。积极方有理由重视可追溯的自主观察和新家族线索；谨慎方有理由要求酶活、机制和稳定复现。二者并不互斥。真正越界的是把“系统值得研究”直接跳成“新一代治疗编辑器已经诞生”。

容易被新闻省略的数字：最高并发 58、随后十次未重现、95 个家族代表但 28 个明确阵列、异源表达每个构建体一个培养样本。它们直接影响对自主性、稳健性和实验强度的判断。

27 / 论文的十个关键局限

局限	为什么会影响解读
1. 功能未知	没有直接建立 RNA → RT/伙伴 → 特定 DNA 产物/表型的因果链。
2. 首轮发现路径不稳	十次重跑未识别阵列，表明能力调用具有偶然性。
3. 无匹配人类基线	不能严谨声称整个项目比相同资源的人类团队快多少倍。
4. 无完整经济账	token 分类不齐；数据库、GPU、人工、实验和失败摊销未形成总成本。
5. 数据与抽样偏差	contig 截断、类别采样、同源搜索阈值会影响发现范围。
6. 结构主要是预测	RT-伙伴接触和 GNAT-like 功能联想需要实验证据。
7. 小 RNA 实验重复有限	新构建体每个一个培养样本，不能把公开感染数据重复数借来。
8. 评测内生	题目源自已发现系统，只比 Claude，由 Mythos 评分。
9. 部分分析后选	解释性特征经筛选展示，重放也未完全保留原环境。
10. 开放复现材料不完整	本次文件未含完整简报、日志、数据库快照与中间数据包。

同时应保留它的优点。作者公开十次遗漏；明确功能未知；区分初始自主发现和后续人类指导；讨论 GNAT 推断的冲突；报告实际工具、资源和较细方法。透明报告限制本身提高可评估性。

分析：适合给这项工作两条独立评价轴。一条是“新分子系统的证据有多完整”，目前到家族、表达及机制线索；另一条是“自主科研流程有多可靠”，目前证明了可能性，也暴露出路径选择的脆弱性。不要把一条轴上的进展替代另一条。

本文未通过外部实验室重复 ART 的功能，因为原论文也没有提供该功能的完成性结论。任何“第三方已验证新编辑器”的描述，都应另找对应原始实验论文。

28 / 战略含义：哪些能力可能形成优势

以下为本报告的分析，不是 Anthropic 已披露的商业计划。

第一，数据的可用性比“拥有公开数据”更稀缺。原始序列很多人能下载；把蛋白、坐标、邻域、结构、样本和文献连成可查询体系，并让 agent 用得正确，需要持续工程。优势若存在，可能落在整理质量、检索速度和错误反馈，而不是对公共序列的排他占有。

第二，湿实验的价值在于给出外部约束。模型很容易生成合理故事，真实样品会否定其中很多。实验室如果能迅速把失败变成清晰的训练/评估信号，价值会超过单次新闻发现。不过生物实验的阴性常有多种解释，奖励并不像代码单元测试那样干净。

第三，科研软件是接触真实工作流的入口。产品能帮助更多科学家调用工具，也可能提供关于困难任务的反馈。用户数据是否进入训练、如何隔离机构机密，是需要按实际合同核查的问题；不能从产品存在推断客户数据已被拿来训练。

第四，专用模型未必被通用模型替代。本论文与蛋白设计项目都使用结构、序列和 RNA 工具。通用模型可能成为规划与整合层，专用模型仍承担高密度科学计算。评估竞争关系，应拆分基础模型、工作流、科学模型、实验服务和药物资产。

第五，内部药物项目会带来合作边界问题。同一家公司既为药企供工具，也自己探索分子，客户会关心知识产权与竞争。公开“目前不自办临床”的表述有助澄清一部分边界，但并不回答发现期的所有利益问题。[S02；本段后半为分析]

应避免的估值跳步：从一次 ART 发现直接推算药物管线价值；把据报收购价当 Bio 年预算；把 21 小时当整个研发周期；把一次内部演示的上限表现当客户日常可获得的能力。

29 / 下一步该盯什么，什么会改变判断

当前判断	能提升信心的公开结果	会使判断转弱的结果
ART 是有价值的机制候选	直接酶活、RNA/伙伴依赖、特定产物及自然功能证据。	长期只能观察到布局与表达，关键机制实验反复不成立。
自主探索具有潜力但不稳	预先定义任务集、独立专家盲评、多次运行的发现质量分布。	产出主要依赖事后挑选一条最好路径。
实验回路可能形成能力优势	公开多轮“预测→实验→修改”的增益与失败记录。	实验只用于发布前验收，很少改变模型或下一轮策略。
软件可能成为科研入口	真实团队的持续使用、可复核效率提升、明确数据边界。	演示可行但日常流程需大量隐形人工。
成本可能下降	按实验验证结果计的全成本、通量和周转时间持续改善。	算力便宜了，人工整合和失败实验仍主导成本。

对当事方最值得问的十个问题

1. 自有实验室的正式主体、启用时间、团队人数与公开地址是什么？
2. ART 研究和药物项目各自占多少人力，是否共用实验设施？
3. 哪些实验由内部完成，哪些由外部合作方完成？
4. 单次计算 campaign 到首个有解释力的实验结果，典型需要多久？
5. 如何区分生物学阴性、执行失败和模型提出的错误假说？
6. 实验反馈究竟进入训练、评估、技能说明还是产品设计，增益如何测？
7. 会否公开研究简报、完整日志、代码和可复现数据子集？
8. 能否展示跨项目、跨运行的科学质量分布，而非只展示最佳案例？
9. 哪些成果计划申请专利，如何与药企客户划分发现及使用权？
10. 哪些新证据会让团队停止把 ART 当作潜在可编程工具推进？

这些是研究和商业判断的待核问题，不代表已存在问题或负面事实。本报告没有联系当事人、未进行访谈，也未把尚未发生的跟踪工作写成已完成调查。

30 / 全文覆盖清单与阅读入口

原论文页码	实际内容	本报告对应
p.1	标题、作者、摘要	08，以及 01 的结论。
p.2	引言、问题背景	08、15。
pp.3-6	自主框架、RT 搜索与发现路径；主图 1	09-11、19-20。
pp.6-10	ART 家族、阵列、表达与伙伴；主图 2-3	12-15、20。
pp.10-13	定向模型评测与重复识别；主图 4-5	16-18、20。
pp.13-14	讨论、能力和机制边界、同期工作	15-18、27-28。
pp.15-21	全部七张补图	21-22；相关正文逐项解读。
pp.22-27	103 条参考文献、致谢	原文附录完整保留；重点文献脉络见下。
pp.28-40	全部 Materials and Methods	23-25；关键设计限制分散于正文。

参考文献脉络。论文不是凭空建立生物知识：它依赖宏基因组数据与去冗余工具；retron、DGR、CRISPR 相关 RT 和防御系统研究；基因邻域和 RNA 分析；AlphaFold 类结构工具；自主科学 agent 与假说评比方法；以及 Evo 2、gLM2 和 Minerva 等基因组研究。完整引文仍以原文的 103 项为准，本报告没有假装独立审读其中每篇。

致谢的含义。Feng Zhang、Brian Hie、David Li、Garyk Brixi 等被列为早期反馈者。致谢可证明交流线索，不能证明这些人分别验证了每项结果，也不等于期刊审稿。

内容完整性声明。本报告逐节解析了所取得的全部 40 页，并把原 PDF 全部附后；没有把它改写为逐句中文译本。文中引用但未包含的 Supplementary Note 1、外部完整代码/日志、未取得的媒体付费正文，不在“已完整审读”的范围内。

来源使用规则。原论文记为 [P]；外部材料记为 [S01-S37]，其中 S28-S37 为本次增补。公司材料用于确认其说法和公开记录，人物言论用于说明观点，论文用于科学事实；本报告的推断已明确标识。

31 / 外部来源目录 I

所有网页核查截至 2026-09-25；日期为来源署名日期，招聘与个人页可能持续更新。点击标题可访问原始页面。

[P] Yoon 等，Autonomous AI agents discover reverse transcriptases with tandem repeat arrays，40 页，2026-09。用户提供 PDF；全文见本文件附录。SHA-256：d43c2fa66aa1d255aba37a0499390bcd86e5b2cafa1f8d249d728e998a4b3b27。

[S01] Anthropic，Claude discovers a novel enzyme system with CRISPR-like repeats，2026-09-23。官方研究组、实验室与论文说明。

www.anthropic.com/news/claude-discovers-novel-enzyme-system

[S02] Reuters，Anthropic quietly sets up biology lab as it ramps AI drug program，2026-09-18。公司采访与业务边界；价格等非官方数字另行标识。

www.reuters.com/world/anthropic-quietly-sets-up-biology-lab-it-ramps-ai-drug-program-2026-09-18/

[S03] Dario Amodei，Machines of Loving Grace，2024-10。愿景与前提的原始来源。

darioamodei.com/essay/machines-of-loving-grace

[S04] Dario Amodei，ART 发布后的 X 长帖，2026-09-23。原帖链接；直读受限，正文与 S05 公开转载核对。

x.com/DarioAmodei/status/2102831170299834652

[S05] Reddit，Anthropic CEO Dario Amodei: AI Biology Will Go From Weak To Superhuman Within A Few Years，2026-09-23。用作原帖公开转载入口，不作为生物学证据。

www.reddit.com/r/singularity/comments/1wog3z8/anthropic_ceo_dario_amodei_ai_biology_will_go/

[S06] Newcomer，Anthropic Buys Stealth Dimension-Backed Coefficient Bio，2026-04-03。收购与交易对价的原始媒体报道，金额未作公司确认值使用。

www.newcomer.co/p/anthropic-buys-stealth-dimension

[S07] SynBioBeta，Anthropic Is Hiring Biologists, Building Wet Labs, and Betting Big on Drug Discovery，2026-05-07。活动主办方对 Eric 发言的记录。

www.synbiobeta.com/read/anthropic-is-hiring-biologists-building-wet-labs-and-betting-big-on-drug-discovery

[S08] Anthropic，The Briefing: AI for Science，2026-06-30 活动页。产品发布与公开职务。

www.anthropic.com/events/the-briefing-ai-for-science

[S09] Anthropic，How Claude is accelerating protein design and analytical chemistry，2026-08-18。蛋白设计的实验结果与任务范围。

www.anthropic.com/research/Claude-accelerates-protein-design

32 / 外部来源目录 II

[S10] Anthropic，How Claude is uplifting biomolecular modeling，2026-09-17。模型优化、计算指标与外部实验竞赛。
www.anthropic.com/research/claude-uplifts-biomolecular-modeling

[S11] Anthropic，Model Hardware Standard research preview，2026-08-27。硬件接口工作；不是 ART 全自动实验的证据。
www.anthropic.com/news/model-hardware-standard-research-preview

[S12] Anthropic 招聘，Computational Scientist, Life Sciences，San Francisco。关于数据库、agent 与模型反馈的岗位说明。
job-boards.greenhouse.io/anthropic/jobs/5357739008

[S13] Anthropic 招聘，Experimental Scientist, Life Sciences，San Francisco。关于实验、快速周转与模型反馈的岗位说明。
job-boards.greenhouse.io/anthropic/jobs/5285250008

[S14] Matthew Durrant，About me。本人公开履历。
mattdurrant.me/aboutme.html

[S15] Arc Institute，Hsu Lab。校友去向公开记录。
arcinstitute.org/labs/hsulab

[S16] The Scientist，Anthropic’s Secretive AI-Powered Wet Lab Breaks Cover-and Makes First Discovery，2026-09。科研媒体对新研究组和发现的报道。
www.the-scientist.com/anthropic-s-secretive-ai-powered-wet-lab-breaks-cover-and-makes-first-discovery-75037

[S17] Decrypt，Anthropic Says Claude Found Something Big in DNA. It Doesn't Know What，2026-09。包含研究者/评论者的外部反应；本报告不把评论当复现。
decrypt.co/379128/anthropic-claude-found-something-big-dna-doesnt-know-what

[S18] Gautam Parab，Claude Found the Enzyme System in One Campaign Out of Eleven，2026-09。个人原文解读，重点为重复运行和数据读取。
gautamparab.com/writing/one-campaign-in-eleven/

[S19] Li 等，Coevolutionary mining of prokaryotic non-coding elements with a genome language model，bioRxiv，2026-09-22，DOI 10.64898/2026.09.22.753630。同期 Minerva 工作；本报告的具体比较以 ART 论文讨论为主，未完整取得该预印本。
www.biorxiv.org/content/10.64898/2026.09.22.753630v1

33 / 外部来源目录 III

[S20] Anthropic，Expanding support for scientists，2026-08-27。科研使用支持与 Claude Science 背景。
www.anthropic.com/news/expanding-support-for-scientists

[S21] Anthropic 招聘，Life Sciences Operator Lead，New York。药物研发与运营方向的公开岗位。
job-boards.greenhouse.io/anthropic/jobs/5357746008

[S22] Anthropic，Life Sciences Verification Program，2026-09-17。科学研究访问与公测信息。
www.anthropic.com/news/life-sciences-verification-program

[S23] Andrew Ho，公开 X 招聘帖，状态 ID 2098193901815521702。Dogpatch / 湿实验 / RL 线索；归属未与 Anthropic 建立，本文只作排除性记录。原帖及公开镜像入口如下。
x.com/andrewho03/status/2098193901815521702

x.noodl3.net/andrewho03/status/2098193901815521702

[S24] Reuters，US scientist John Jumper to leave Google DeepMind for Anthropic，2026-06-19。拟加入的报道。
www.reuters.com/technology/us-scientist-john-jumper-leave-google-deepmind-anthropic-2026-06-19/

[S25] Business Insider，Google / AI talent wars 相关报道，2026-06。关于 garden leave 的报道；不据此指定到岗日期。
www.businessinsider.com/google-ai-talent-wars-anthropic-jumper-shazeer-karpathy-openai-2026-6

[S26] Matthew Durrant，Introducing Claude Science: now in beta，公开 LinkedIn 帖子。关于 Eric 和 Alec Tarashansky 的工作线索。
www.linkedin.com/posts/matthew-durrant_introducing-claude-science-now-in-beta-activity-7478253727192109057-kNW6

[S27] Bloomberg / Robert Langreth，Anthropic Biology Discovery Draws Cautious Notes From Scientists，2026-09-24 21:00 UTC。只取得标题及导语，未把付费正文中可能存在的内容写成已读。
news.bloomberglaw.com/artificial-intelligence/anthropic-biology-discovery-draws-cautious-notes-from-scientists

附录说明

本目录后为新增的第 34-42 节，再附用户提供论文的完整 40 页英文原文，保持原页码、图表和参考文献。正文可按第 30 节的覆盖表定位。附录并不新增任何未取得的补充文件。

本报告中的未知项仍为未知：实验室门牌、准确启用日、总人数、年度预算、完整内部汇报线、ART 的自然功能与可编程编辑性能。

34 / 增补：哪些判断得到推进

本次继续调查没有改变“ART 功能尚未确认”这一结论，但补上了几块会影响战略与新颖性判断的证据。

问题	上一版的边界	本次新增证据
团队分工	知道 Coefficient Bio 被收购，但对其工作方向描述较宽。	Nathan Frey 本人明确负责药物发现与开发研究；提出让一位熟练 biotech 运营者与多个 Claude 运行药物项目。[S28]
旧研究贡献	知道底层 RT 此前被观察过。	直接核对 2021 年原论文，确认它还描述了较长 N 端、上游非编码区和 ncRNA 假说。[S32]
是否遗漏引用	尚未展开社交平台中的优先权指责。	2026 年论文参考文献 61 确实列出 Korn 等旧研究；“完全没有引用”的指责不成立。[P]
Dario 时间表	重点解释 2024 年文章的条件化愿景。	2026 年 9 月文章的“未来 5-10 年”表述更直接、更强，应单独记录。[S31]
实验室的组织建设	已有科学家岗位及官方介绍。	实验运营招聘涉及采购、设备、样本追踪、场地与 CRO，显示其在建设完整运营能力。[S33]
药企合作的落点	已有一般性产品与合作背景。	Novo 公告把早期目标写为具体 R&D 工作流、科学问题与软件开发，并强调治理和人工监督。[S34]
位置线索	能确认到湾区。	补查到 Howard Street 办公租赁记录，但没有把这些办公室与 wet lab 建立联系。[S35,S36]

更新后的分析。Bio 布局可以更清楚地分成三个层面：发现新分子系统、让药物项目更高效地运转、把能力交付给外部科学团队。人才、任务、验证方式和商业目标在这三个层面并不相同。

仍未解决的问题。未找到可靠的湿实验室门牌；未取得完整团队人数、预算和汇报图；未发现 ART 已完成催化或编辑功能验证的新原始论文。新增证据没有填上这些空白，报告也不会将办公地址或愿景改写成答案。

35 / Nathan Frey：从分子设计到药物项目运营

当事人给出的定位。Nathan Frey 现行个人网站明确称自己领导 Anthropic 的 drug discovery and development research，此前共同创立 Coefficient Bio，并曾在 Genentech 带领研究组。他用“pipeline-in-a-person”描述团队方向：一位熟练 biotech 运营者，加上多个 Claude，创建并运行药物项目。[S28]

这个表述补上了上一版只写“收购计算生物学人才”遗漏的一层：至少从负责人公开愿景看，目标涉及药物项目的组织与执行，不局限于训练一个蛋白模型。

角色/能力	可以确认	不能确认
Eric Kauderer-Abrams	公开职务为 Head of Life Sciences。[S08]	所有子团队具体管理层级和资源分配。
Nathan Frey	本人称领导药物发现与开发研究。[S28]	他就是整个 Life Sciences 的最高负责人，或 ART 湿实验室的直接主管。
Matthew Durrant / Nicholas Perry	ART 通讯作者，具备相关基础研究背景。[P,S14,S15]	仅凭作者顺序确定行政级别。
Samuel Stanton	Coefficient Bio 共同创始人；本人旧网站描述 ML、抗体设计与 lab-in-the-loop 兴趣。[S06,S29]	旧网站仍写 Genentech，不能当作今天的任职与汇报线。
Claude Science 开发	Durrant 的公开帖子将开发领导角色归于 Alec Tarashansky。[S26]	与上述所有工作线一一对应的正式部门图。

分析：这对 Agent Infra 的含义。一个药物项目不仅需要生成候选，还要协调证据、实验、供应商、时间与资源。若 agent 能接手这些跨工具工作，价值不只来自模型一次提出更好的分子，也来自减少信息传递与重复决策的摩擦。

应当怎样检验，而不是直接相信？比较同类项目的决策等待时间、实验重做比例、外包交付质量、证据追踪完整性和阶段推进成本。这里提出的是评估框架；Frey 的网站没有公布这些指标，也没有证明“一人可以替代完整药企”。

人才线索的含义。从已有研究经验走向 agent 驱动的项目运营，说明药物线可能重视“如何做正确的下一步”与“如何调动现有工具”。这仍不足以推断未来药物资产归属、合作条款或内部训练方法。

36 / 2021 与 2026：新颖性争议的逐项核对

Frey 的公开发布帖下，有评论者指出这套系统五年前已被描述，并要求引用旧作者。[S30] 本报告据此回到 Korn 等 2021 年原文，而不把评论直接当结论。[S32]

具体主张	2021 年论文	2026 年 ART 论文
MarsHill 中存在 RT	已描述 gp206：501 aa、retron-like RT 域。	承认此前已有相关 RT 记录。
特殊 N 端	已描述约 200 aa、功能未知的 N 端。	扩大到家族、结构、谱系与伙伴关系的系统比较。
上游非编码区	已描述约 1,200 bp 区域，并提出可能含 retroelement ncRNA。	界定重复阵列、拷贝、间隔及跨位点规律。
自然功能	提出包括防御在内的可能解释，保持未知。	提出多 RNA / RT / 伙伴的假说，功能仍未知。
专属伙伴	旧文未识别出其所讨论的已知效应伙伴。	提出三类 ART 伙伴及其关联与预测结构。
RNA 实际表达	相关旧文段落没有给出本文这类阵列 RNA 证据。	增加公开感染转录组分析与异源表达小 RNA 测序。
引用旧工作	不适用。	参考文献 61 明确列出 Korn 等 2021 年论文。

旧研究的三项具体事实，来自原文讨论 gp206 的段落；2026 年新增项来自用户提供论文。这里没有声称读完了 2021 年全部补充数据，也没有把“相关段落没给出某结果”扩大成穷尽整个领域的优先权裁决。

批评成立的部分：若传播口径让人以为 Claude 首次看见这条 RT、首次注意到较长 N 端，或首次想到上游可能有 RNA，就掩盖了旧研究的重要贡献。新颖性应定位在更具体的证据组合上。

批评不成立或尚不足的部分：“论文完全未引用旧工作”与文献表不符；“旧文已经完成全部 ART 系统描述”也不能仅凭 RT、N 端与非编码区三点推出。对正式抄袭或知识产权问题，本次科学内容比较并不作法律结论。

最准确的贡献描述。agent 从既有公开序列和知识背景出发，识别并跟进此前未被这样系统表征的阵列组合；随后人类指导的研究补上家族、伙伴和表达证据。其增量真实与否，应在这些具体层面被复核。

37 / Dario 在 2026 年说得更强了

需要补全的时间线。2024 年《Machines of Loving Grace》以足够强 AI 的出现为重要前提。2026 年 9 月《We Must Pace the Frontier》开篇则直接表达：他相信 AI 可能在未来 5-10 年治愈大多数重大疾病。同文又主张放慢前沿能力提升，使安全措施跟上。[S03,S31]

9 月 16 日 Novo 合作公告中，他再次使用“把一个世纪的生物医学进步压缩到十年”的愿景。[S34] 因此，仅用早期文章中谨慎解释过的起算点，去概括其所有最新言论，会低估后来表述的强度。

命题	应如何归类	可用于检验的证据
未来 5-10 年治愈大多数重大疾病	Dario 的技术与医学预测。	应先定义疾病集合、“治愈”标准、起算点与可及性，再跟踪实际结局。
让科学家得到更强模型会缩短研发时间	合作方的目标与待检验假说。	同类任务、完整周期、真实实验与项目结果的比较。
前沿能力需要与安全措施同步	治理立场和公司承诺。	独立评估的权限、覆盖范围、公开结果和纠偏记录。
ART 说明医学加速已经实现	不能由本论文直接成立。	ART 尚需功能验证，更未涉及疾病疗效或临床可及性。

分析：这些话可以同时存在，但会产生张力。加速研究应用与放慢某些前沿能力提升不是形式逻辑上的直接矛盾；实际问题是访问范围、模型能力和实验速度如何分配，以及可验证的收益是否足以支撑公开时间表。

“共识”更适合分三层。第一层是可观察事实：模型已能辅助部分计算和实验研究。第二层是路线判断：更多智能和更好的工具可能提高研发效率。第三层是强时间预测：大多数重大疾病将在几年内被治愈。ART 的证据能支持讨论前两层，不能把第三层升级成科学界共识。

这不是对医学前景给出反向保证。它只是要求同样严格地对待乐观与悲观预测：提出可以被证伪的定义，并让实际证据更新判断。

38 / 运营招聘透露的实验室建设阶段

一个已关闭的 Research Operations Lead, Biology 招聘页面，比泛泛的“我们有实验室”披露了更多运营线索。岗位位于 San Francisco，要求接手已开始建立的采购、库存、设备、安全与场地系统，并随团队增长扩展；职责还包括预算、供应商与 CRO。[S33]

岗位明确涉及	可以支持的判断	不能倒推出的事实
设备维护、校准、排期	需要有人持续维护真实实验能力。	具体仪器数量、品牌或当前利用率。
样本追踪、库存、SOP	实验过程需要可追踪的组织系统。	这些系统已经完全集成到 Claude。
装修、搬迁、设备安装协调	岗位预期要处理设施与扩张问题。	当前正在搬迁，或已经签下某一处地址。
预算与资本设备采购	实验工作有专门运营与资本开支流程。	年度预算、已花费金额或实验室估值。
CRO 与服务商管理	内外部实验能力可能需要统一协调。	每项实验已经外包给哪家机构。

职位状态也必须记录。截至本次访问，LinkedIn 显示不再接受申请，原 Greenhouse 链接跳回职位列表。这只能说明公开招聘入口状态，不能确定已经招到人，也不能据此判断实验室扩张停止。

分析：湿实验反馈如何变成 agent 的有效学习信号？至少要把一个结果与任务假设、样本身份、实验条件、设备状态、执行记录和后续解释连起来。否则“失败”可能是模型假说错，也可能是样品、操作或测量问题；混在一起会误导评估。

对 RL / eval 的判断。实验室能够产生真实反馈，但这不意味着它天然是可规模化的 RL 环境。奖励归因、反馈延迟、并发能力和复现实验成本，都会限制训练吞吐。招聘支持“正在建立承载这些工作的运营能力”，尚不支持“已经证明生物实验 RL 可大规模扩展”。

与 ART 的十次探索重跑结合看，较迫切的工作可能包括选择该看什么数据、排除伪阳性以及记录完整证据链；这些改进并不都需要机器人或新的专用模型。

39 / 地址追查：办公楼找到了，实验室仍未定址

补查商业地产报告后，可以列出明确与 Anthropic 有联系的公开办公租赁记录。以下均不能标注为湿实验室地址。

公开地点	可核查记录	与 wet lab 的关系
500 Howard Street	Cushman & Wakefield 的 2023 Q4 办公市场报告列 Anthropic 转租 230,325 平方英尺。[S36]	支持历史办公租赁，不证明实验用途。
300 Howard Street	同机构 2026 Q1 报告列 480,000 平方英尺新租约，正文称未来总部。[S35]	未来总部/办公室，不代表实验室已在这里运营。
400 Howard Street	同报告列 104,336 平方英尺新租约，并注记将计入修订季度数据。[S35]	办公需求线索；不与其他报道中的 505 Howard 自动合并。
1675 Owens Street，Mission Bay Conference Center	AICamp 的 AWS AI Builder Lab 活动场地；Anthropic 出现在合作方名单。[S37]	这里的 Lab 是开发者实践活动，不是生物实验室证据。
Dogpatch	此前第三方湿实验/RL 招聘帖。[S23]	仍未建立 Anthropic 归属。

为何搜索容易误导？办公市场报告同页列 OpenAI 的 Mission Bay 租约与 Anthropic 的 Howard Street 租约；关键词片段会把城市片区、公司和“lab”混在一起。会议的 hands-on lab、科研园区名字、投资组合公司地址，也可能制造类似误关联。

政府记录检索的真实边界。本次补搜 SF 政府与规划域名，并尝试读取一个搜索命中的 DataSF 数据集。数据接口未成功返回可核查的单条记录，因此没有把摘要里的“laboratory”和“Anthropic”拼成同一个项目，也没有声称已经完成全库许可排查。

目前最可靠的定位结论仍是湾区。科学家与运营岗位增强了 SF 招聘基地的证据，但没有证明实验台必在 SF 市界内。未获得主体、地址、用途三者一致的记录前，不能进一步压缩到 Howard Street、Mission Bay、Dogpatch 或 South San Francisco。

这张表的价值是明确排除“查到了公司门牌＝找到了湿实验室”的推断。若用于安排访问，办公楼与实验室都仍应通过当事方确认。

40 / 从 Novo 合作和用户反馈看落地边界

合作原文比“AI 联合制药”标题更具体。Novo 9 月 16 日公告称，双方围绕其科学家和计算团队提出的研发问题开发方案，初期测试 Claude Science 的特定工作流，也推进软件开发，并保留数据治理和人工监督。[S34]

这份公告确认了合作及目标，没有公布一项由 Claude 独立发现、已获临床成功的新药。它也不能证明 Novo 的数据与 Anthropic 自有实验室之间采用了何种未披露的训练或共享机制。

工作模式	验收对象	应看什么指标（本报告分析）
科研工具交付	科学家是否更容易完成日常任务。	任务质量、证据完整度、等待时间、用户留存。
项目共同开发	是否解决具体研发困难。	阶段目标达成、实验通过率、失败原因和完整周期。
内部 agent 研发	模型是否稳定做出有价值的下一步。	重复运行分布、独立盲评、成本和人工干预。
药物资产转化	候选是否能安全有效地进入后续开发。	功能、可开发性、临床与监管结果；不能由软件指标替代。

外部反馈还包括“用不到”的问题。Frey 发布帖下一名自称从事 RNA 与蛋白设计的博士研究者抱怨，最新模型的限制妨碍其正常研究。[S30] 这是一条具体使用者陈述，不是经过统计的普遍故障，也没有核实其账号、请求或资格状态。

分析：能力、准入、日常可用性是三件事。内部演示能做、某些机构获准使用、普通科学家在真实工作中稳定用上，不能互相替代。已有验证计划也不自动说明所有受阻用户立即可以解决问题。[S22]

更新后的商业判断。目前最有证据的方向，是把科研与项目运营中的可分解任务逐步产品化，并利用内外部实验做验证。是否会形成比专门 AI4S 团队更强的优势，要看长期实验反馈、科学判断、任务可靠性与合作治理，而不能只看通用模型品牌或单次发现的声量。

本页不估算未披露合同金额，也不把药企采用与临床成功建立未经验证的因果关系。

41 / 增补来源 IV

[S28] Nathan C. Frey，个人网站，核查于 2026-09-25。本人对 Anthropic 药物发现与开发研究角色、pipeline-in-a-person 方向及此前经历的说明。旧 GitHub Pages 站点已指向此处。

ncfrey.com/

[S29] Samuel Stanton，个人网站及 Bio 页面，页面版权标记 2023。用于历史研究兴趣与 Prescient Design 背景，不将其过时任职文字视为当前状态。

samuelstanton.github.io/

samuelstanton.github.io/bio.html

[S30] Nathan Frey，宣布 Anthropic 分子生物学实验室和首次发现的 LinkedIn 帖子及公开评论，2026-09。评论中包括旧研究优先权质疑和研究访问受阻的个体陈述。评论者自述身份未另行验证；科学结论以原论文核查为准。

www.linkedin.com/posts/ncfrey_weve-set-up-a-molecular-biology-lab-at-anthropic-activity-7508597023206526976-DIln

[S31] Dario Amodei，We Must Pace the Frontier，2026-09。原文开头包含“未来 5-10 年”医学预测；正文提出放慢能力发展并强化评估。用于记录他的观点，不作为医学预测已经成立的证据。

darioamodei.com/post/we-must-pace-the-frontier

[S32] Abby M. Korn、Andrew E. Hillhouse、Lichang Sun、Jason J. Gill，Comparative Genomics of Three Novel Jumbo Bacteriophages Infecting Staphylococcus aureus，Journal of Virology，2021，DOI 10.1128/JVI.02391-20。重点核查 Results and Discussion 中讨论 MarsHill gp206 的段落；2026 ART 论文参考文献 61 也引用此文。

journals.asm.org/doi/10.1128/jvi.02391-20

新颖性核查原则。旧论文证明哪些内容已被描述，新论文证明增加了哪些内容；社交评论用于发现争议，不能替代科学原始材料。对评论中的未证实动机、法律指责与人物身份，本报告不作扩张。

42 / 增补来源 V

[S33] Anthropic，Research Operations Lead, Biology，公开 LinkedIn 职位页。核查时显示不再接受申请；可读取职责、SF 招聘地点与运营要求。原 Greenhouse 链接已跳回职位列表。

www.linkedin.com/jobs/view/research-operations-lead-biology-at-anthropic-4437345132

job-boards.greenhouse.io/anthropic/jobs/5304242008

[S34] Novo Nordisk A/S，Novo and Anthropic will collaborate to advance drug discovery with Claude，2026-09-16，官方新闻稿，经 GlobeNewswire 发布。用于合作范围、初期目标、Dario 发言及治理说明。

www.globenewswire.com/news-release/2026/09/16/3363106/0/en/novo-and-anthropic-will-collaborate-to-advance-drug-discovery-with-claude.html

[S35] Cushman & Wakefield，San Francisco MarketBeat Office Q1 2026，4 页。第 2 页正文及第 3 页交易表列出 300 / 400 Howard Street 的办公租赁；没有将其标为生物实验室。

assets.cushmanwakefield.com/-/media/cw/marketbeat-pdfs/2026/q1/us-reports/office/san-francisco_americas_marketbeat_office_q12026.pdf?rev=f4082512d6c0402782606fbf8876769c

[S36] Cushman & Wakefield，San Francisco MarketBeat Office Q4 2023，4 页。第 3 页办公租赁表列 500 Howard Street、Anthropic、230,325 平方英尺。

assets.cushmanwakefield.com/-/media/cw/marketbeat-pdfs/2023/q4/us-reports/office/san-francisco_americas_marketbeat_office_q4-2023_v3.pdf?rev=734f04b7c43f43e3833b9f713798f69c

[S37] AICamp，AWS AI Builder Lab (San Francisco) - Clash of Agents，2026-02-13 活动页。用于排除 Mission Bay Conference Center 与 wet lab 的误关联。

www.aicamp.ai/event/eventdetails/W2026021308

原论文附录

下一页起为完整 40 页英文论文，保留原始图表、页码和参考文献。增补版合计 44 页中文报告与 40 页原文；本次新增 9 页内容，并同步更新了前文团队与 Dario 时间表。

增补未改变的证据边界：ART 功能与基因编辑能力尚待验证；实验室公开位置仍只确认到湾区，办公楼的门牌不作为实验室地址。

Autonomous AI agents discover reverse transcriptases with tandem repeat arrays

Peter H. Yoon[1], Januka S. Athukoralage[1], Emmanuel Ameisen[1], Eric Kauderer-Abrams[1], Nicholas T. Perry[1,†], Matthew G. Durrant[1,†]

[1]Anthropic, San Francisco, CA, USA

Abstract

Metagenomic databases hold vast quantities of unexplored genetic material that can seed tools for biotechnology and medicine, but discovery is limited by our ability to examine sequence data at scale. Here, we show that large language models (LLMs) in an autonomous harness can alleviate this bottleneck and surface novel molecular systems. We deployed Claude Code instances to survey reverse transcriptase (RT) loci across 1.9 billion protein clusters. Among the top candidates were array-associated RTs (ART), a novel family of jumbo-phage RTs coupled with an array of ~200-nucleotide units and a dedicated partner gene. Session transcripts show that Claude identified the ART family by directly examining its DNA sequences and recognizing the atypical repeats, a behavior attributable to specific Mythos 5 internal signals that respond to repeated DNA. ART arrays are highly expressed and appear as discrete units during *Staphylococcus* phage infection, suggesting an RT system directed by a repertoire of distinct RNAs. These findings indicate that LLMs can autonomously detect anomalies and drive analyses to initiate biological discoveries.

[†]Correspondence: nicholasp@anthropic.com, mdurrant@anthropic.com

Introduction

Sequence databases have become the largest and fastest-growing record of biological diversity (1–4). Mining them has yielded industrial enzymes, new antibiotics, and genome-engineering tools (5–7). Many of these discoveries began with a researcher inspecting sequences and recognizing atypical features (8–10). Driven by metagenomic sequencing, these databases now double every four years and hold billions of protein sequences, increasing not only the diversity available for discovery but also the volume of data that must be examined (2, 3, 11, 12).

Systematic study of databases of this size is possible only through computational pipelines, which retrieve homologs of a known system's components and sort the hits into classes by a set of predefined features (13). Novelty in such pipelines is prescribed in advance, and anomalies that fall outside those features go undetected. A human expert examining the data can attend to features beyond the predefined set and flag anomalies, which is why even systematic surveys end in expert curation (7, 13, 14). Such curation does not scale with the data and remains a bottleneck in mining these resources.

Large language models (LLMs) can now write and run code, operate analysis tools and databases, and complete long tasks with little supervision (15–18). LLMs are automating parts of software engineering and data analysis (19, 20), and accelerating biological research, where multi-agent systems have originated drug candidates, protein designs, gene-editing experiments, and mechanistic hypotheses that were confirmed at the bench (21–25). More recently, autonomous agents have planned and executed chemical syntheses, machine-learning studies, and biological data analyses end to end, from a stated objective to a written report (26–30). We reasoned that genome mining, an iterative process of collecting, classifying, and investigating sequences, could be automated in the same way with our latest frontier models. We further asked whether LLMs that reason over the data might also notice anomalies beyond the features a pipeline prescribes, supplementing manual curation.

As a test case we chose reverse transcriptases (RTs), a widespread enzyme family that canonically copies RNA into DNA, an activity that underpins biotechnologies from cDNA synthesis to genome engineering (31–34). In prokaryotes, RTs are often co-expressed with a partner protein and non-coding RNAs (ncRNAs) encoded in the same locus (35, 36). In retrons, for example, the RT copies a dedicated ncRNA into single-stranded DNA that functions with an effector protein in antiphage defense (37, 38). These effectors are diverse, including an N-glycosidase that hydrolyzes NAD^+ in Retron-Eco1 and a membrane protein predicted to disrupt the membrane in Retron-Eco6 (37, 39, 40). New RT systems that act with novel ncRNAs and partner genes continue to be reported (36, 41–47), suggesting that more remain among RT loci that have yet to be examined.

Here, we present an autonomous genome-mining campaign in which Claude agents surveyed RTs across 1.9 billion metagenomic protein clusters. The campaign was structured so that agents could pursue their own observations beyond a predefined pipeline. One such observation led to the discovery of the array-associated RT (ART) family, when an agent reading the DNA beside a phage RT noticed an unannotated array of tandem repeats. ART belongs to no described RT class and is typified by its non-coding DNA (ncDNA) repeats reminiscent of CRISPR arrays, an architecture that may serve to diversify the function of a single enzyme. This work demonstrates that AI systems can notice anomalies in primary biological data

and surface them as a human expert would, easing the curation bottleneck in genome mining at a large scale.

Results

Agentic genome mining with Claude

We built an agentic harness, software that coordinates LLM agents, to carry out a genome-mining campaign based on a brief that outlines research goals (**Fig. 1A**). The research brief is first converted into a series of analysis stages that represent broad steps, such as surveying a database for homologs. Each stage is carried out as smaller tasks, such as a sequence profile search, in which one Claude Code agent acting as the worker proposes a plan and executes it while a second agent acts as the supervisor to review the plan and results. The supervisor can open new tasks for new workers in response to observations made during the work, allowing the campaign to extend itself as findings accumulate. All plans, results, and reviews are written to a shared record, so that the harness operates as a multi-agent system with a common memory. The campaign ends when the task queue is exhausted, and its findings are delivered to human reviewers as written reports.

We wrote a research brief that asked the agents to identify novel RT systems on the basis of new partner-gene associations (Supplementary Note 1). Running with Claude Mythos 5, the agents assembled their own sequence profile hidden Markov models (HMMs) and reference sets (Methods), recovered ~200,000 RT clusters, and classified them into nine distinct RT classes (**Fig. 1B**). They then sampled ~11,000 RT loci spanning every class and scored the 3,564 protein families that recurred in their neighborhoods as candidate partners. Sixteen families passed the agent-defined selection criteria and a review of their annotations, and were each investigated by a dedicated worker (**Fig. 1B**). The agents opened 98 new tasks from observations made along the way, one of which led to the promotion of a seventeenth candidate (**Fig. 1C**). The full campaign comprised 119 tasks and 949 agent sessions, amounting to 77 agent-hours and 215.6 million tokens over 21.5 hours of wall-clock time, without human intervention (**Supp. Fig. 1A**).

Of the 17 candidate partner families, only three were confirmed as previously unreported RT associations. The other 14 were rejected or set aside as annotation artifacts, parts of previously described systems, or general neighborhood residents rather than dedicated partners (**Supp. Fig. 1, B and C**). Beyond the partner-gene associations the census scored, workers also flagged unusual features of individual RTs and their flanking DNA for follow-up by new workers. Three of the flagged RTs were identified as new RT lineages (**Fig. 1D**; **Supp. Fig. 1C**). In total the campaign concluded with 19 reports, comprising 16 for the 17 candidate families (two were described in a single report) and three for the new RT lineages.

An unexpected observation led agents to array-associated RTs

We turned our focus to the reports that ranked highly according to a tournament-style evaluation of the 19 reports (Methods). One report described a new lineage of RTs, largely confined to jumbo-phage genomes, whose defining feature lay in the ncDNA upstream of the RT (**Fig. 1D**). Since the mission was directed at partner protein-coding genes rather than ncDNA features, we examined the session transcripts to de-

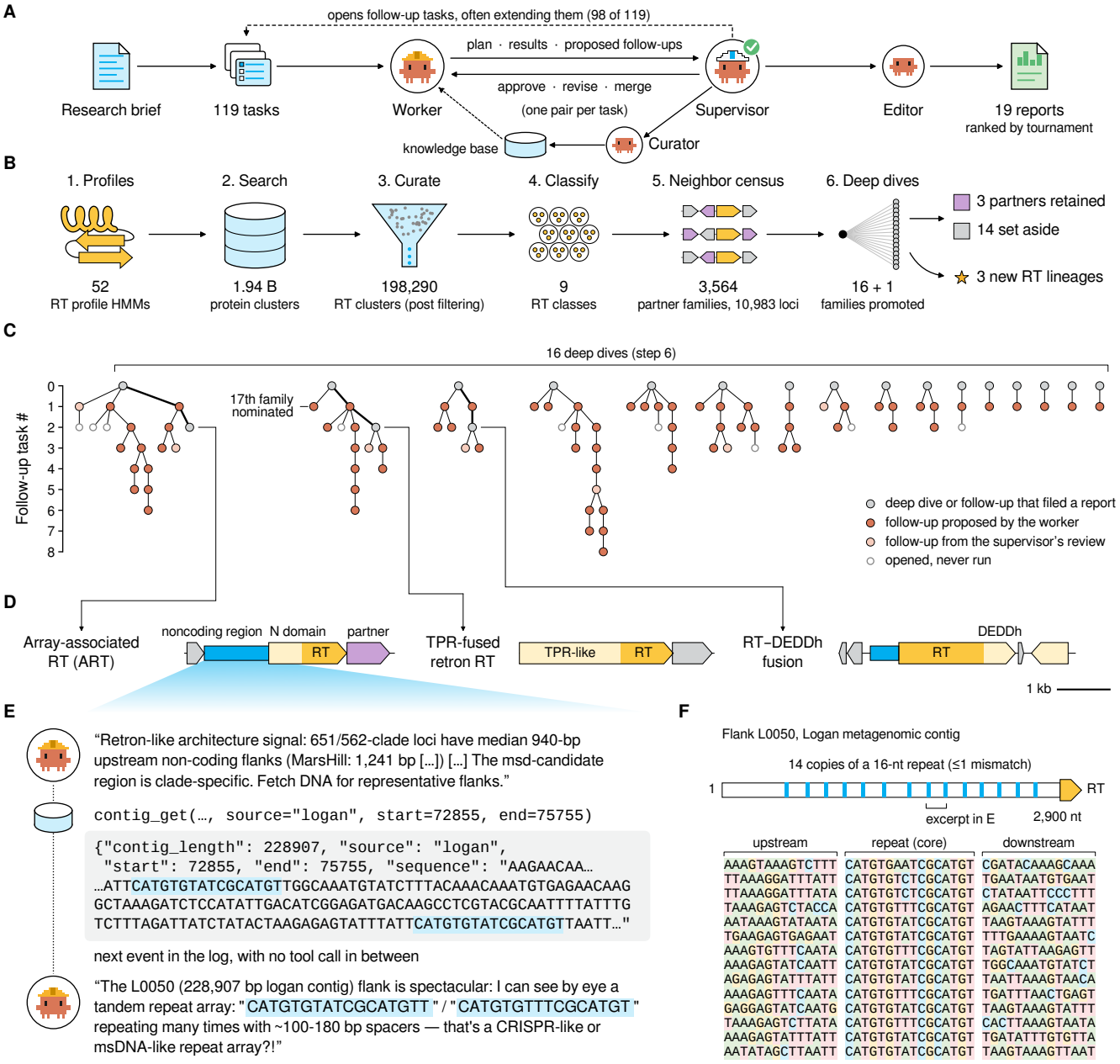


Figure 1. Agentic discovery of novel RT lineages. (**A**) Schematic of the harness. Instances of Claude act as workers to plan and run tasks, supervisors to review them and open follow-up tasks (98 of 119 tasks), curators to record findings in a shared knowledge base, and editors to review reports. (**B**) Pipeline for survey of RTs and their partner proteins. (**C**) Task trees for each of 16 deep dives on RT families nominated in (B). The 17th family was promoted by a follow-up task. Bold paths lead to the first report on each new RT lineage. (**D**) Representative loci of three newly described RT lineages, to scale. (**E**) Worker agent transcripts showing loading and interpretation of DNA in context identifying a repeat motif. Blue, repeat copies. (**F**) Schematic of locus examined by worker in (E), with 14 repeats and surrounding context.

termine how the agents had identified this RT family (**Supp. Fig. 2A**). The agents initially selected this candidate because it co-occurred with a nearby phage gene, the β-like subunit of the jumbo-phage non-virion RNA polymerase. A worker agent later rejected this gene as a spurious association, but chose to queue a follow-up task to study the RT more closely. From the reasoning trace:

> So the adjacency is lineage-specific gene order, and the family's identity = phage core transcription machinery. The "retron accessory" interpretation dies. […] But ALSO a genuinely reportable secondary finding: AR9-like jumbo phages carry a free-standing retron-like RT […]
>
> *Worker agent, task t0010.*

Retrons typically act in phage defense, so the agent considered the presence of the retron in the jumbo-phage genome to be noteworthy. Because retron RTs act with an ncRNA encoded upstream, the agent's supervisor asked the next worker to check the 5′ flanks of these RTs for such an RNA. The RTs beside the polymerase gene had only short non-coding regions upstream (median 27 bp), too short to encode an RNA, whereas their closest relatives had long ones (median 940 bp). The worker read the upstream sequence of these relatives directly into its context, which led to the following reasoning trace (**Fig. 1E**; **Supp. Fig. 2B**):

> The L0050 (228,907 bp logan contig) flank is spectacular: I can see by eye a **tandem repeat array**: "CATGTGTATCGCATGTT" / "CATGTGTTTCGCATGT" repeating many times with ~100-180 bp spacers — that's a CRISPR-like or msDNA-like repeat array?! Look: "TACATGTGTATCGCATGTTAAGTAAAGTATT..." repeated. Pattern `CATGTG-T/A-TCGCATGT` — ~17-bp repeat, spacers ~110-170 bp. This looks like a **DGR-like VR array? No — more like a CRISPR mini-array or a retron-like repeat region**. […] Similarly L0020 (jgi 12,964) flank shows repeats: "ACTTGTAAGAATTTCGCAAGTTA" recurring — "TACTTGTA[AT]GAATTTCGCAAGTT" multiple times with spacers. Same architecture, different repeat sequence!
>
> *Worker agent, task t0062. L0050 and L0020 are the worker's labels for the two metagenomic flanks.*

The worker directly read this tandem repeat array pattern as a potentially undescribed RT feature. The subsequent reasoning trace shows evidence that the worker then questioned the novelty of this system, comparing it with known RT systems:

> Actually wait — reconsider. Repeat+spacer arrays immediately 5' of an RT in phages... This is EXACTLY the architecture of **"retron-like tandem arrays" in the recently described 'DRT9'?** I recall 2024-25 papers on phage/bacterial RTs that polymerize from repeat arrays, e.g., **"ApeA"?** Hmm, and there's the Zhang lab's 2024 "**programmable RNA-guided...**" no. The honest path: characterize quantitatively, then run the novelty kill-test literature search with precise descriptors […].
>
> *Worker agent, task t0062.*

The worker subsequently analyzed the non-coding regions and confirmed they were different from any known non-coding elements, including CRISPR arrays (**Supp. Fig. 2C**). A repeat-counting script written by the worker found one locus with 14 copies of a 16-nt repeat in the region, separated by unique spacers of 100 to 200 nt (**Fig. 1F**). A literature search later confirmed that the array did not fit any literature reported

feature, and the agents filed a report for human review. We named this lineage of jumbo-phage RTs array-associated RTs (ART) after the array that defines it.

Array-associated reverse transcriptases (ART) are a new family of RTs

We next determined in interactive Claude Science sessions whether the array and its RT define a coherent family beyond the loci the agents had examined (Methods). Profile HMM searches of our metagenomic database and public genomes identified 95 distinct RT clusters (90% identity) in cultured jumbo phages and predicted viral contigs, 28 of which carried a detectable repeat array upstream of the RT (**Supp. Fig. 3A**; Methods). Phylogenetic analysis placed the RTs in a clade next to retrons, in line with the agents' initial classification (**Fig. 2A**; **Supp. Fig. 3B**). In these loci, the array lies upstream of the RT, a feature missing from other RT systems, and a dedicated partner gene lies directly downstream, both of which the agents described in their report (**Fig. 2B**; **Supp. Fig. 3C**).

The RT itself has an unusually long N-terminus, with about 180 residues preceding the polymerase domain where other RTs typically carry about 50 or fewer (**Fig. 2, C and D**). The polymerase domain retains the catalytic YxDD motif in all 93 members whose sequences span it (two are truncated before it) and superimposes on solved retron and diversity-generating retroelement (DGR) RT structures (**Supp. Fig. 4, A to C**). In contrast, the N-terminal extension of the RTs shows at most weak similarity to any known sequence family or fold, and is the most variable region of the protein among members (**Fig. 2, E and F**; **Supp. Fig. 5A**). Nevertheless, it is present in every member of the family and constitutes the second hallmark of ART RTs (**Supp. Fig. 5B**). The polymerase-adjacent RTs that first led the agents to ART form its sister clade (**Fig. 2, A and B**). These RTs also have a long N-terminus (**Supp. Fig. 5C**), yet none of their 17 assessable loci carries a repeat array and none encodes an ART partner (**Supp. Fig. 5, D and E**). We therefore refer to them as non-array-associated RTs (NART). Their N-terminus adopts a single predicted fold distinct from those of ART (**Supp. Fig. 5, F to H**).

We next investigated the non-coding arrays that typify ART systems. The 28 arrays adjacent to an RT span 0.3 to 4.1 kb and consist of 3 to 21 copies of a short repeat at near-constant spacing (**Fig. 2, G and H**). Repeats are 15 to 49 nt long, each with a palindromic core of ~15 nt and less conserved edges, and are shared within but not across ART clades (**Fig. 2G**). The spacers between repeats are 120 to 220 nt long and unrelated to one another within an array (**Fig. 2, H to K**). Between related phages such as MarsHill and SA1, spacers are retained in order and diverge at a similar rate to the RT gene (**Fig. 2F**). Across ARTs, conservation of repeats is generally higher than conservation of spacer sequences. These features contrast with CRISPR arrays, where near-identical repeats with sharp boundaries alternate with spacers of only ~30 nt that are gained and lost between related strains. Critically, no *cas* genes occur near any ART locus, suggesting that the arrays represent a new type of non-coding repeat element.

ART repeat arrays are highly expressed during phage infection

To learn more about the array, we prompted a Claude Science session to search for further information on ART loci. The session autonomously identified and retrieved RNA reads from a published infection time-course study of *Staphylococcus* phage SA1 (48), which contains an ART system (**Fig. 3, A and B**). This dataset revealed that the ART repeat arrays are highly expressed at early, mid, and late stages of

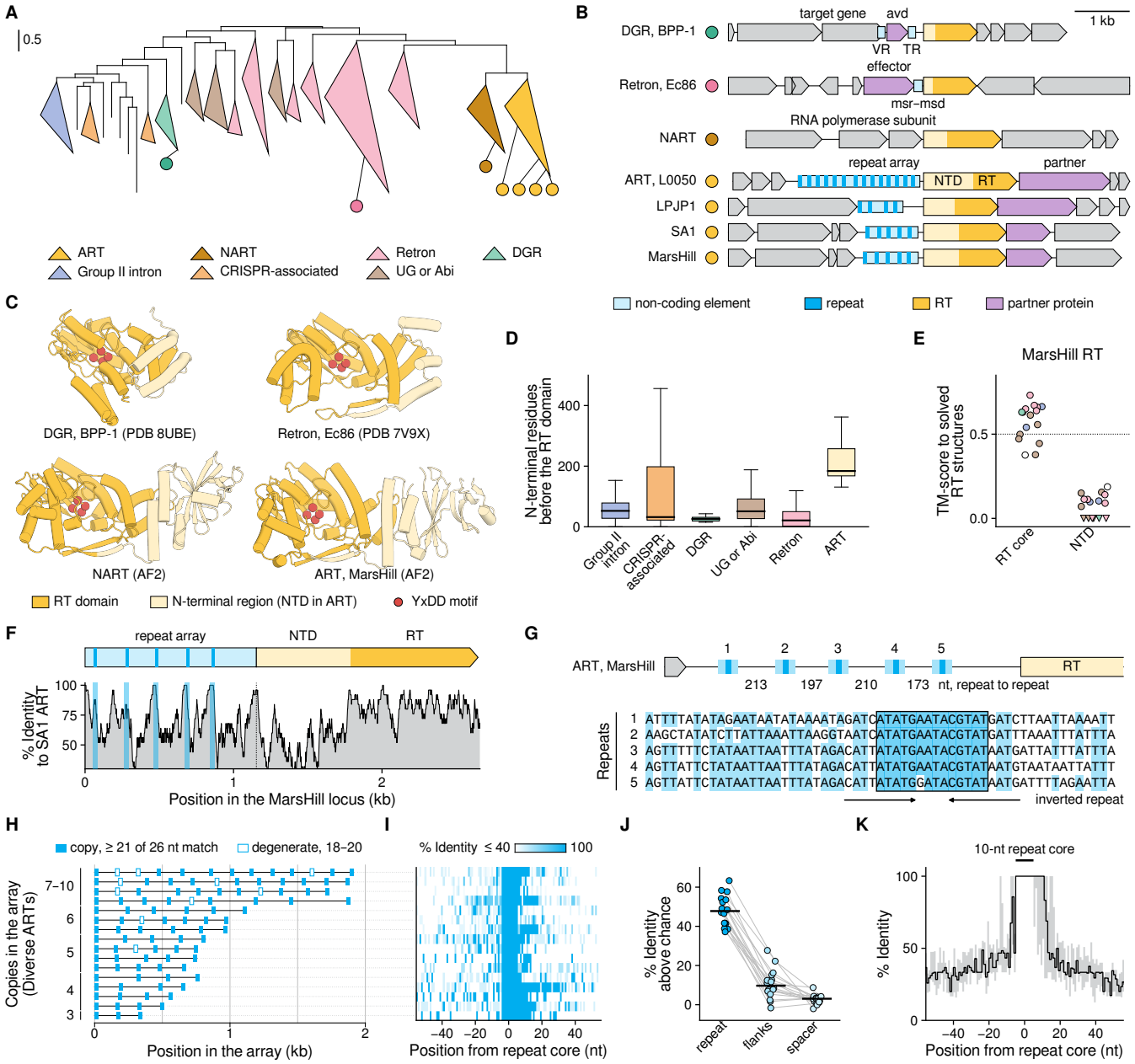


Figure 2. Features of array-associated reverse transcriptase loci. (**A**) Maximum-likelihood tree of 774 RT domains, midpoint-rooted and collapsed by class. (**B**) Representative loci of the RT families in (A), to scale. NTD, N-terminal domain. (**C**) Solved and predicted (AlphaFold2) structures of related RTs. (**D**) Number of N-terminal residues ahead of the RT domain of representative RT systems. Boxes, quartiles; whiskers, 1.5 times the interquartile range. (**E**) Foldseek TM-scores of the MarsHill RT core and NTD against solved RT structures (n = 16; triangles, no alignment reported). (**F**) Sequence identity between the SA1 and MarsHill ART loci (25-nt sliding window). Blue, repeats. (**G**) Repeats of the MarsHill ART, shown with flanking conserved regions. (**H**) Pattern of repeat arrays across 16 representative ART loci. Degenerate repeats are shown. (**I**) Identity between array units in (H), centered on conserved repeat core. (**J**) Sequence identity of repeats, repeat flanks, and spacers within the same array, above that of shuffled sequences. Lines join one array; bars, medians. (**K**) Sequence identity around repeat core across ART systems (n = 16 arrays; line, median; band, interquartile range).

infection (**Fig. 3C**; **Supp. Fig. 6, A and B**). Remarkably, at 15 minutes post infection, the array-derived RNAs account for up to 8% of phage RNAs, making them among the most abundant phage transcripts (**Fig. 3C**). Within the array, the transcripts resolve into shorter species whose boundaries are reproducible across replicates (**Fig. 3D; Supp. Fig. 6A**).

To examine the array RNA independently of phage infection, we expressed the SA1 ART system on plasmids in *Escherichia coli* and performed small-RNA sequencing. Across both a native locus and a heterologous-promoter construct, we observed discrete short RNAs from the array similar to those formed during infection (**Fig. 3D**). Folding predictions for array units placed the repeat in a base-paired stem regardless of how each unit was cut from the array, consistent with the repeat itself containing a short inverted repeat (e.g., a palindrome) (**Fig. 2G**; **Fig. 3E**; **Supp. Fig. 6C**). These results suggest that a single ART array gives rise to several distinct short ncRNAs with the potential to form secondary structures.

These RNAs raised the possibility that ART is a retron-like system that operates with a bank of distinct RNAs. This prompted us to investigate the ART partner genes, since the partner protein executes the system's function in retrons. We found three unrelated partner families that each occur nearly exclusively beside ART family RTs, and classified ART systems accordingly (**Fig. 3, F and G**; **Supp. Fig. 7A**). Type I, found at 59 of 95 loci across the *Listeria* phage and environmental lineages, encodes a protein of about 600 residues with two tandem GCN5-related N-acetyltransferase (GNAT)-like folds (**Supp. Fig. 7, B to D**). Type II, confined to the *Staphylococcus* phage clade, encodes an all-helical protein of about 270 residues, and type III encodes a smaller helical protein of about 170 residues in one environmental sub-lineage; neither appears to have clear homologs in protein domain or structural databases (**Supp. Fig. 7, B and D**). The three partner families share no detectable similarity in sequence or predicted structure, suggesting that ART family RTs have been paired with unrelated partners more than once. Structure predictions of cognate type I and II pairs converged with moderate confidence (interface predicted template modeling [ipTM] score ~0.6) on an interface between the N-terminal extension of the RT and the partner (**Fig. 3H**). The N-terminal extension adopts different predicted folds in the two lineages, raising the possibility that it acts as a partner-binding module that co-evolves with its partner (**Supp. Fig. 5, F to H**; **Supp. Fig. 7, E and F**).

Taken together, ART loci share the three components of a retron (an ncRNA, an RT, and a dedicated partner) but carry a bank of distinct RNAs where retrons carry one (**Fig. 3I**). In retrons the RT and the single-stranded DNA that it produces (msDNA) hold the effector in an inactive complex until phage infection or damage to the msDNA releases it, killing the cell (37, 38). We hypothesize that ART may work on a similar principle, with each unit RNA assembling its own complex with the RT and partner, so that one enzyme and effector pair yields several complexes that differ only in their RNA (**Fig. 3I**). The differences between these RNAs may in turn allow the system to respond to different triggers, given that phage proteins trigger retrons by directly perturbing the msDNA (38).

The most capable models recognize the ART array when they read its DNA

We next asked whether the discovery of ART was reproducible within our harness, so we ran the same campaign ten more times. Nearly every campaign that completed the census sampled ART loci, and in two campaigns, workers investigated the lineage as follow-up. However, none read the DNA upstream of the

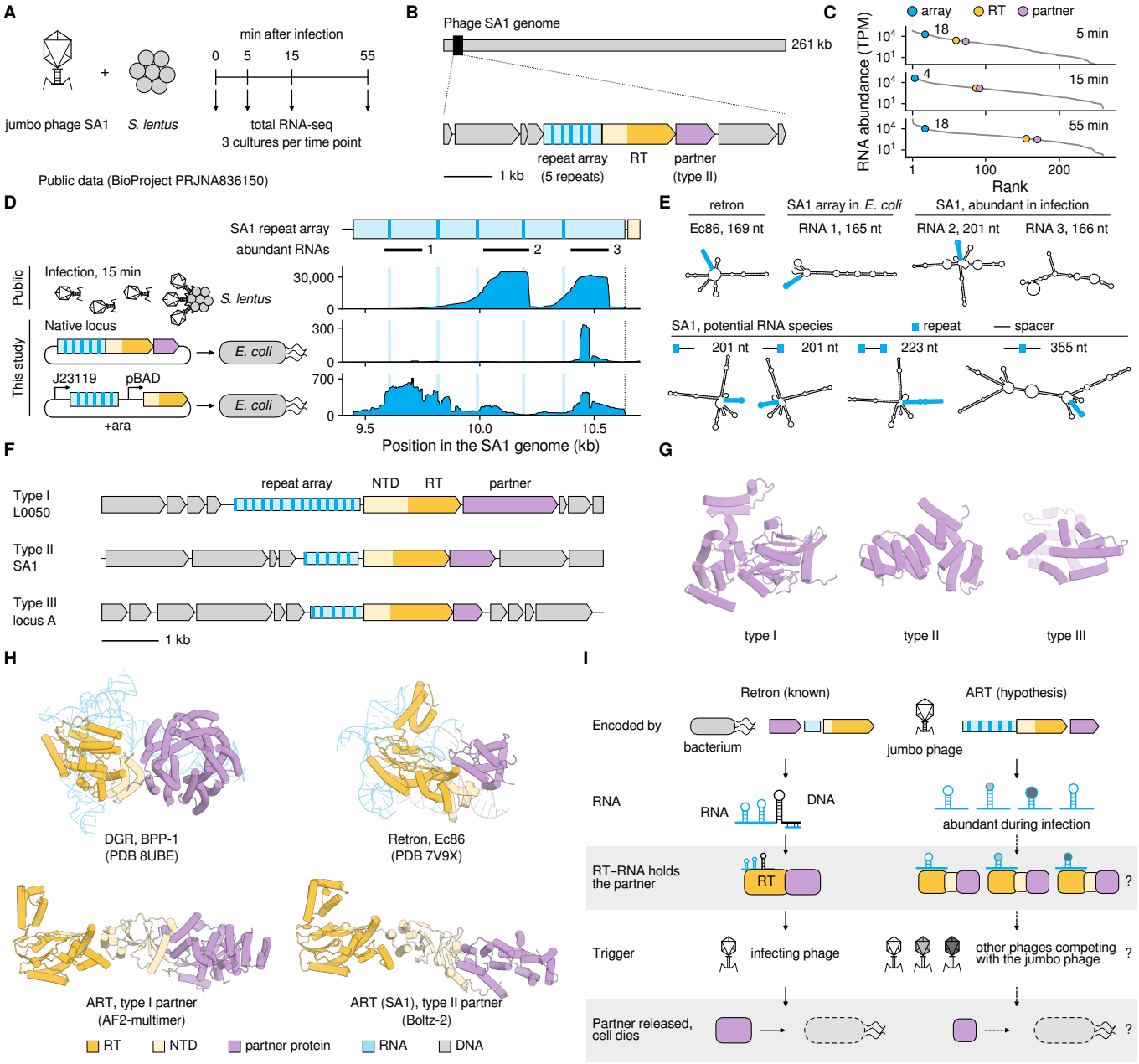


Figure 3. ART arrays are transcribed into abundant RNAs, and ART family RTs co-occur with dedicated partner proteins. (**A**) Schematic of the SA1 phage infection time-course experiment in *S. lentus*, performed with RNA-seq for BioProject PRJNA836150. (**B**) Schematic of phage SA1 genome and resident ART locus. (**C**) SA1 transcript abundance during infection. TPM, transcripts per million. Numbers, rank of the array RNA. (**D**) Alignments of RNA-seq traces to SA1 ART array. (**E**) Secondary structures of Ec86 retron RNA compared to putative SA1 RNA repeats. (**F**) Representative ART loci with divergent partner proteins. (**G**) Predicted structures of the three types of observed ART partner proteins (59, 7, and 3 lineages; AlphaFold2-multimer for type I, Boltz-2 for types II and III). (**H**) Structures of RT-partner protein interactions compared to putative co-folds of ART family RT and partners. (**I**) Putative model of ART biological function compared to characterized retron function. Dashed arrows, hypothesized steps.

RTs, and the array was missed in every rerun. We attribute these results to the broad protein search space of RTs and the non-deterministic behavior of the harness.

We therefore built a set of fixed-input benchmarks in which a model is given ART sequences directly in context or in a filesystem with tools, and scored on its ability to analyze and accurately describe the system (**Fig. 4A**; Methods). Every model was tasked with characterizing the ART system from the input provided and reporting what it deemed important. The resulting report was scored by a judge model (Mythos 5) against ten curated features that we deemed characteristic of the ART family (**Fig. 4B**). Testing seven Claude models, we found a clear gap between the four most capable models (Opus 5.5, Mythos 5.1, Mythos 5, and Opus 5) and the other three (Opus 4.6, Opus 4.8, and Sonnet 5) (**Fig. 4C**). This suggests that the original search campaign, run on Mythos 5, benefited from these improved capabilities.

We then broke down our benchmarks to look at how the most capable models performed at characterizing the three main components of ART: the array, the RT, and the partner (**Fig. 4, D and E**). We noticed that the RT was accurately described by the most capable models, even with just the protein sequences of two RTs provided. Comparatively, more tools and information helped the models better describe the partner gene. Surprisingly, more information and tooling inhibited the models' ability to accurately describe the array, as was accomplished in the original search campaign. Given only the loci in context, the most capable models accurately described the array in at least 90% of attempts, compared to as low as 32% (Opus 5, level 4) with tooling.

We hypothesized that this lower performance stems from these models failing to read the attached sequences into their context, so we inspected transcripts for the amount of DNA each attempt read (**Fig. 4F**). With files, 39% of their attempts never read a contiguous stretch of 200 nt or more and therefore never saw more than about one repeat unit. When a stretch of at least 200 nt was read, recognition was 16 to 32 percentage points higher for each of these models (**Fig. 4G**). As more DNA sequence was read into the context of the model, recognition of the array rose from 29% to as high as 76% of attempts for the four models pooled, and as high as 96% for Mythos 5 (**Fig. 4H**). These results indicate that direct observation of the DNA by the model is critical for recognizing the array, as it was in the original campaign.

Mythos 5 internal signals respond to DNA repeats

Genomic language models have been shown to learn representations of repeated DNA (**Fig. 5A**). These models can be used to detect repeat motifs in DNA (10), their embeddings can distinguish repeat families (49, 50), and a signal that responds to the spacers of CRISPR repeat arrays has been isolated in Evo 2 (49). We wanted to determine if Claude models have similarly learned to represent and recognize DNA repeats like those found in ART arrays (**Fig. 5B**). To ask how the model's internal activity responds as it reads DNA, we used the transcript of the original discovery session to decompose that activity into interpretable signals (**Fig. 1, A and E; Fig. 5C**).

Since genomic language models can detect repeats, we ran Evo 2 and gLM2 on the same L0050 ART locus Mythos 5 explored during the initial campaign (**Fig. 1F**). The log-likelihood profiles of the genomic language models consistently identified the repeats with high confidence (**Fig. 5D**). These profiles were centered on the ART repeat and lost when the nucleotides of each copy were shuffled in place (**Fig. 5, E and F**). gLM2, which reads the sequence on both sides of a base, predicted every copy including the first.

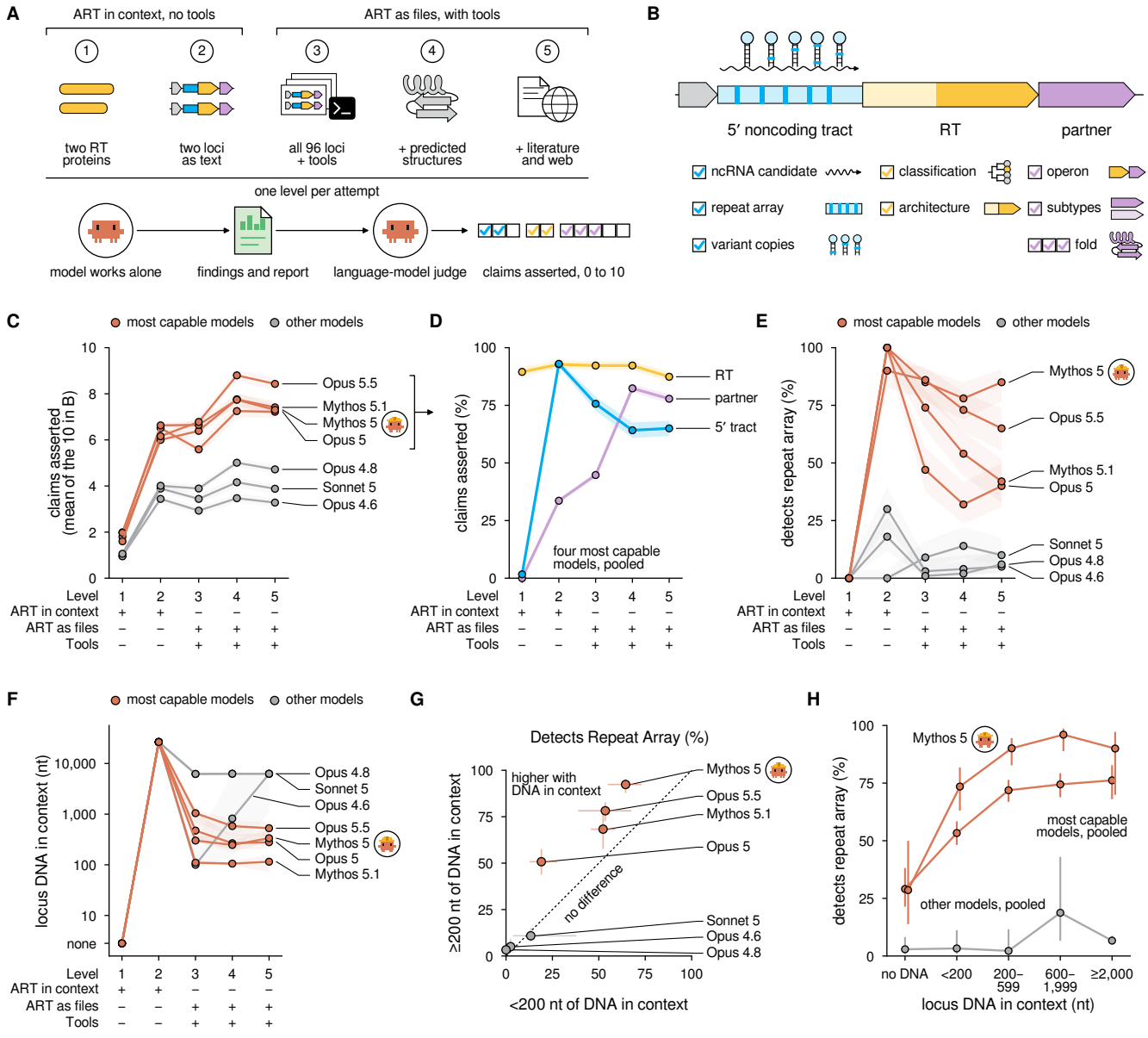


Figure 4. The most capable models recognize the ART array, and recognition depends on reading DNA into context. (**A**) The fixed-input benchmark. Each attempt receives one of five levels of information, with ART sequences in the model's context and no tools (levels 1 and 2) or as files with analysis tools (levels 3 to 5). The model works alone, and a language-model judge scores its findings and report on ten claims. (**B**) The ten claims, grouped by the part of the locus they concern. Each box is one claim. (**C**) Mean number of rubric claims asserted at each level by seven models (n = 100 attempts per model and level). Orange, the four most capable models. Gray, the other three. Icon, the model of the original campaign. Shading, 95% confidence intervals. (**D**) Claims asserted about each part of the locus as a percentage of the total number of claims, for the four most capable models pooled. (**E**) Attempts that recognized the repeat array. (**F**) Median longest contiguous stretch of locus DNA in the model's context. (**G**) Recognition of the array at levels 3 to 5 in attempts with at least 200 nt of contiguous locus DNA in context, against attempts with less. Dashed line, no difference. (**H**) Attempts at levels 3 to 5 that reported the repeat array, by the longest contiguous stretch of locus DNA in the model's context, for the four most capable models pooled (n = 110, 360, 338, 266, and 126 attempts per bin) and the other three pooled (n = 103, 61, 45, 16, and 675). Bars, 95% confidence intervals.

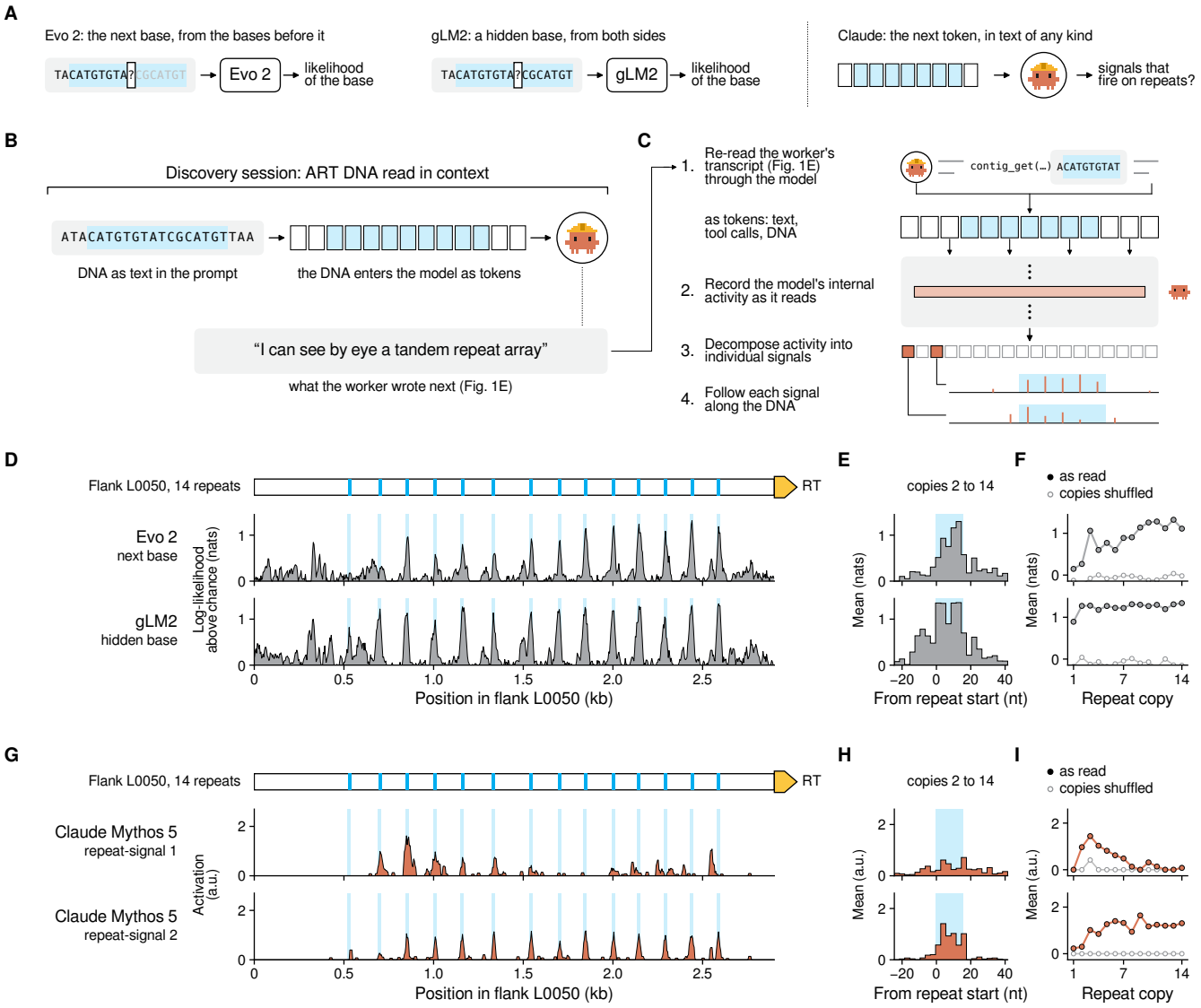


Figure 5. Internal Mythos 5 signals respond to repeats as the sequence is read. (**A**) How each model reads DNA. Evo 2 predicts the next base, gLM2 a hidden base, and Claude the next token of any text. (**B**) In the discovery session the ART DNA was in the prompt and entered Claude as tokens. Quote, what the worker wrote next (**Fig. 1E**). (**C**) Readout of internal signals. The worker transcript of **Fig. 1E** was passed through Mythos 5 again, and its activity at one layer was decomposed into individual signals. (**D**) Genomic models along flank L0050 (**Fig. 1F**). Log-likelihood of the true base above chance (16-nt running mean). Evo 2 was scored through Evo Designer. (**E**) Aligned on the repeat, mean of copies 2 to 14 in 3-nt bins. (**F**) Mean per copy, as read and with each copy shuffled in place. (**G to I**) As D to F, for two of the 12 Mythos 5 signals preselected on synthetic repeats, repeat-signal 1 and repeat-signal 2 (activation; G, 16-nt running mean). Neither is specific to DNA.

Evo 2, which reads only the preceding sequence, predicted the first two copies poorly and improved along the array.

Two Mythos 5 signals (repeat-signal 1 and repeat-signal 2) responded to the array in a similar fashion to the models trained only on biological sequences (**Fig. 5, G to I**). Both were silent before the array, and only repeat-signal 2 weakly fired at the first repeat copy. Repeat-signal 1 peaked at the third copy and faded along the array, with activity scattered across the copies and flanking sequences (**Fig. 5, G and I**). In contrast, repeat-signal 2 strengthened by the third copy and remained steady thereafter, with activity largely confined to the repeat and the nucleotide that follows it (**Fig. 5H**). Shuffling each copy in place silenced repeat-signal 1 on 12 of 14 copies and repeat-signal 2 on all 14 copies (**Fig. 5I**). These results indicate that the repeat triggered a signal inside the model as the ART array sequence was read. This was immediately followed by natural language reasoning that named the sequence a ”tandem repeat array,” which enabled the initial identification of ARTs.

Discussion

Many scientific discoveries originated when a scientist looked closely at the data and recognized something unusual as worth pursuing. Here, these steps were taken by AI agents working together without human supervision. The array that defines ART was recognized after the model read the raw DNA sequence into its context. Both our benchmarks and the Mythos 5 internal signals that respond to repeated DNA suggest that this ”genomic vision” directly enabled the discovery. Paired with a harness that let agents pursue and refute their own hypotheses, this capability allowed an AI system to take the first steps of a biological discovery on its own.

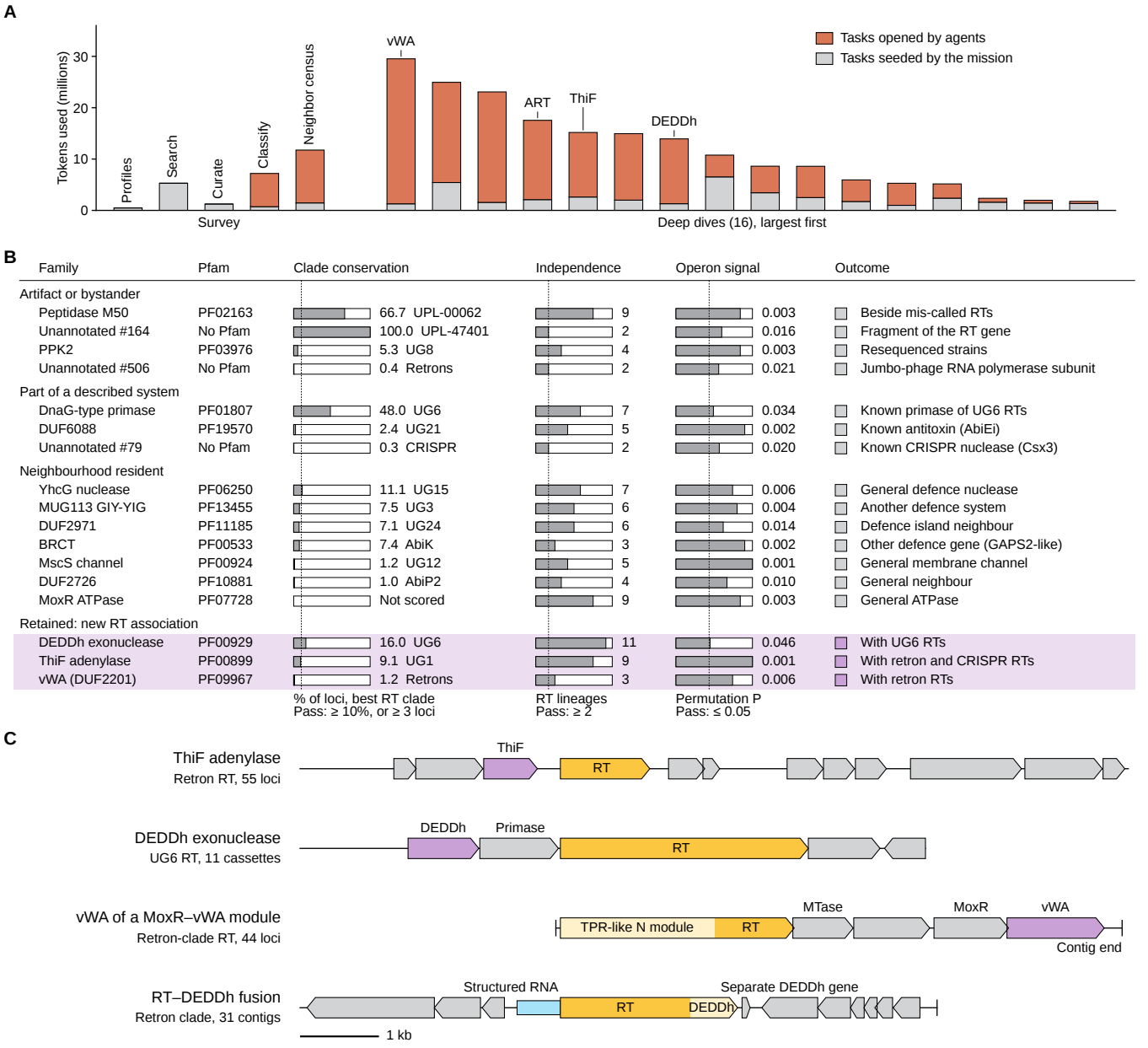
These observations come at a time of rapid progress by AI systems in mathematics. This is exemplified by LLM-derived solutions to long-standing open problems (51–53). Whereas a mathematician can work entirely from first principles, a biologist often must test their ideas in the laboratory (54, 55). Much can be inferred from sequence data alone, but our findings on ART await experimental characterization. ART is a new phage-encoded member of the growing list of prokaryotic RT systems that are typified by ncDNA repeats that bear striking resemblance to CRISPR arrays. Arrays of ncRNAs were recently discovered beside the unrelated UG27 RTs using a bespoke genome language model developed to mine non-coding elements (56), suggesting that this architecture arose more than once in nature. Here, the ART array was recognized by a general-purpose model directly reading the DNA instead. Unlike CRISPR spacers, the ART units are conserved between related phages, and they are transcribed in large excess over the enzyme, suggesting one enzyme supplied with a repertoire of templates or baits (37, 38, 41, 42). Beyond these observations, we have not shown that the RT is active or that the unit RNAs are its substrates. Whether the RT and its partner interact, and what the system does for the phage, are currently unknown.

Awareness of anomaly is proposed to be at the start of scientific discovery (57), and an anomaly exists only against an expectation. For statisticians that expectation is an explicit model of the data, whereas for biologists that expectation is often derived from contextual and tacit knowledge (58, 59). In earlier attempts to automate scientific discovery, both the expectation and the judgment of novelty were supplied by human experts (54, 55, 60). Recent AI-powered methods allow agents to screen their own hypotheses for novelty, but a scientist still sets the expectation and makes the final judgment (21, 23, 24, 27–30). Here,

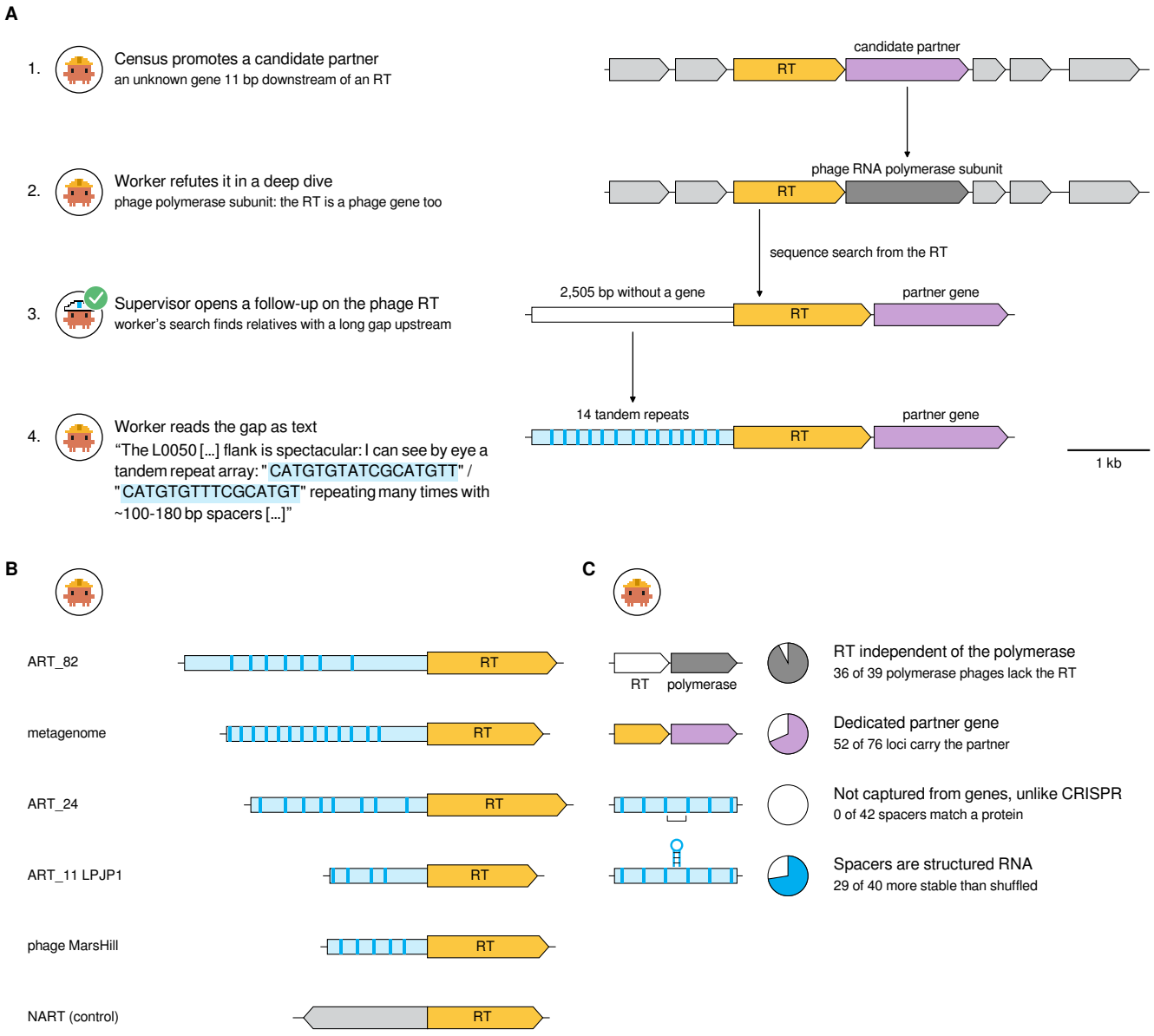
the agents supplied both the expectation, drawn from their own knowledge rather than their task, and the judgment that the atypical observations merited further pursuit.

Agents that read primary data directly can provide the judgment that limits discovery wherever data have outgrown the attention of human scientists. The MarsHill ART locus is an example, as its original genome report identified the RT and proposed a 5′ ncRNA, yet described neither the repeats nor the partner gene (61). Our results show that agentic harness configurations and model capabilities can dramatically affect research outcomes and successful discovery. We expect that as models improve, human scientists will be enabled to explore a wider range of ideas, focus on experimental design and validation, and discover many more systems like ART by using AI systems to make connections between the vast stores of human-produced scientific knowledge.

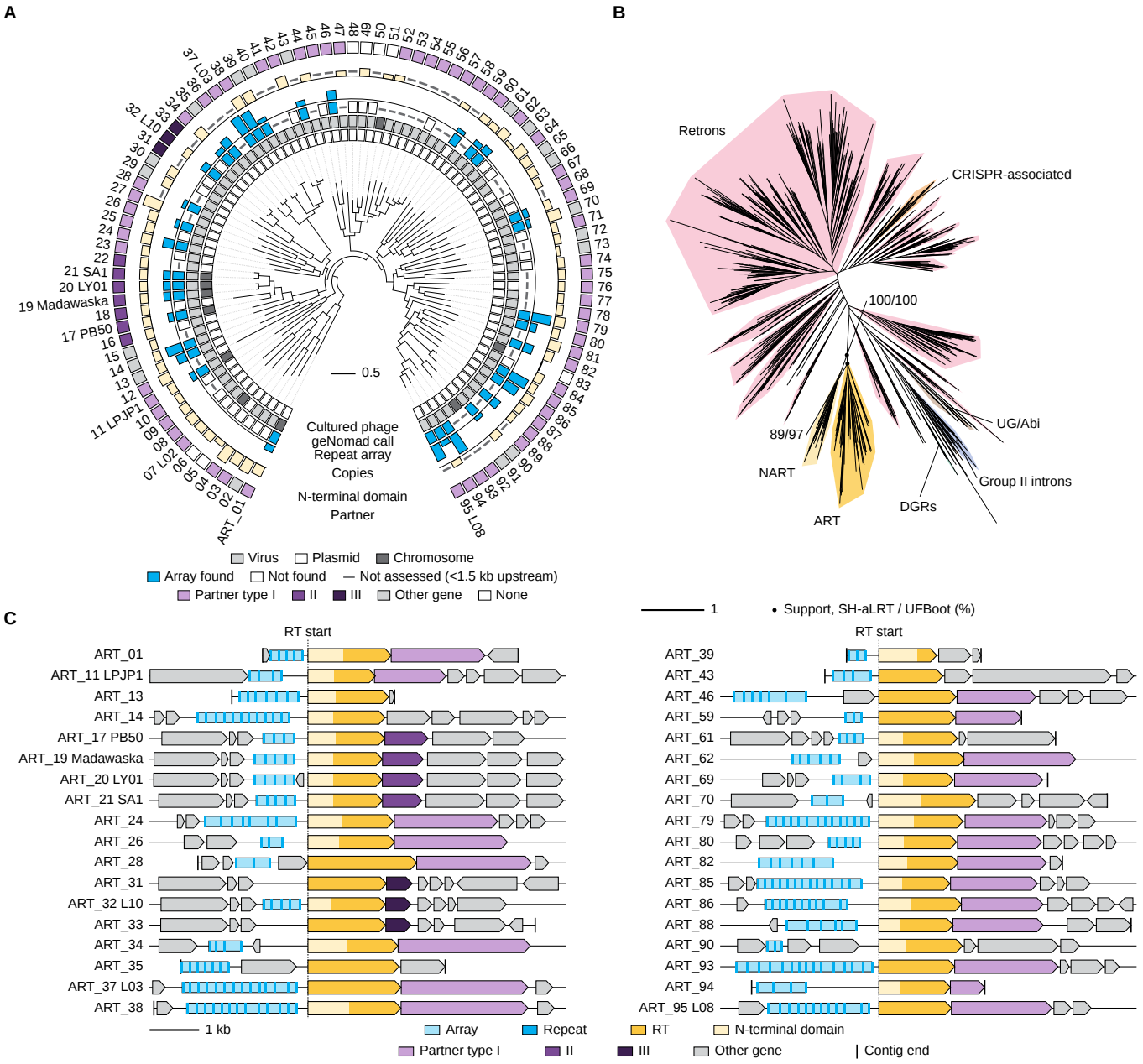
Supplementary Figures



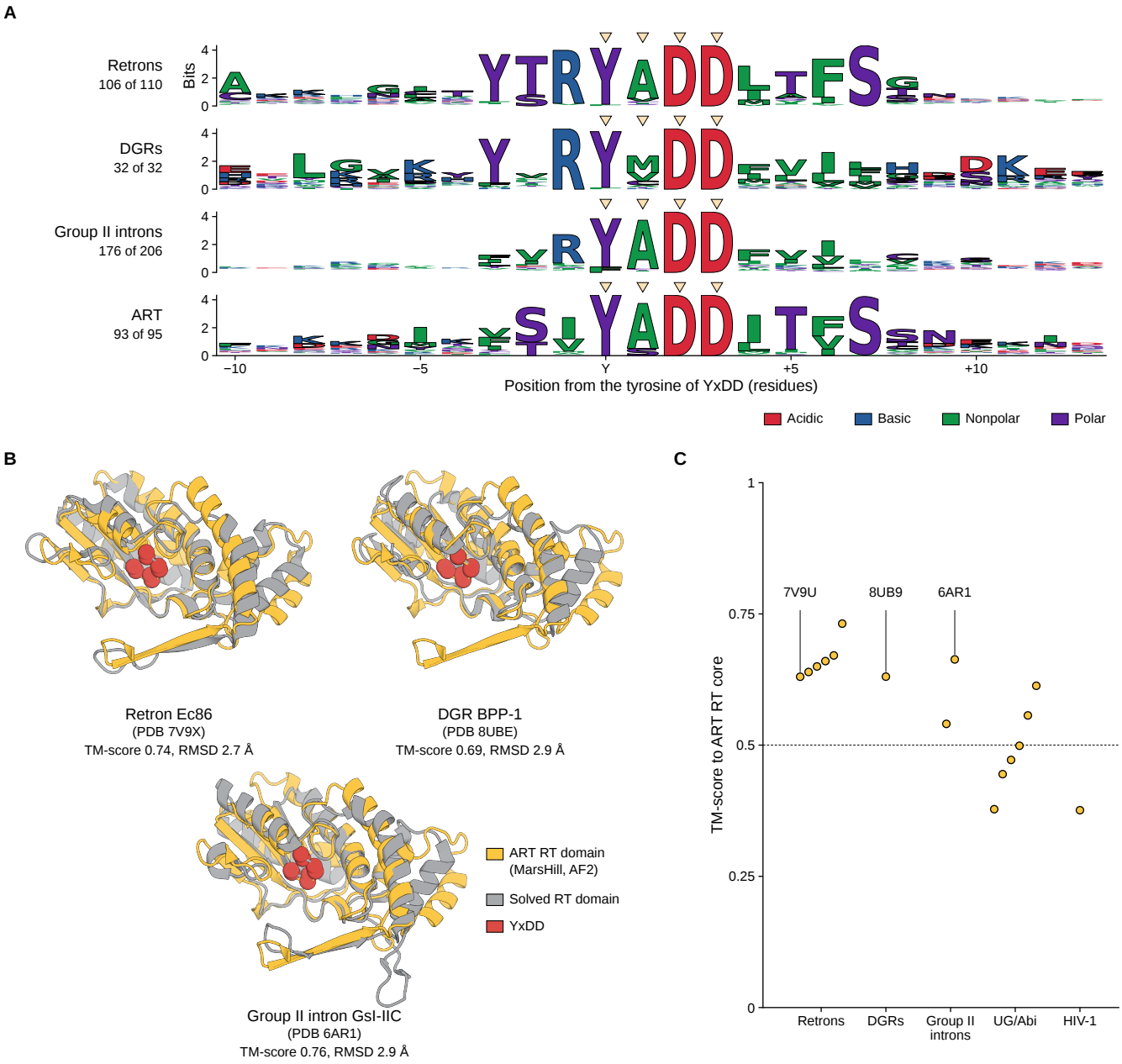
Supp. Fig. 1. Outcomes of the promoted partner families and other new RT lineages. (**A**) Tokens used by each survey step of **Fig. 1B** and by each of the 16 deep dives together with all follow-up tasks that descend from it. (**B**) The 17 candidate partner families promoted from the neighbor census, grouped by final outcome, with their scores on the three promotion filters. Dotted lines, pass thresholds. (**C**) Representative loci of the three retained partner families and of the RT-DEDDh fusion.



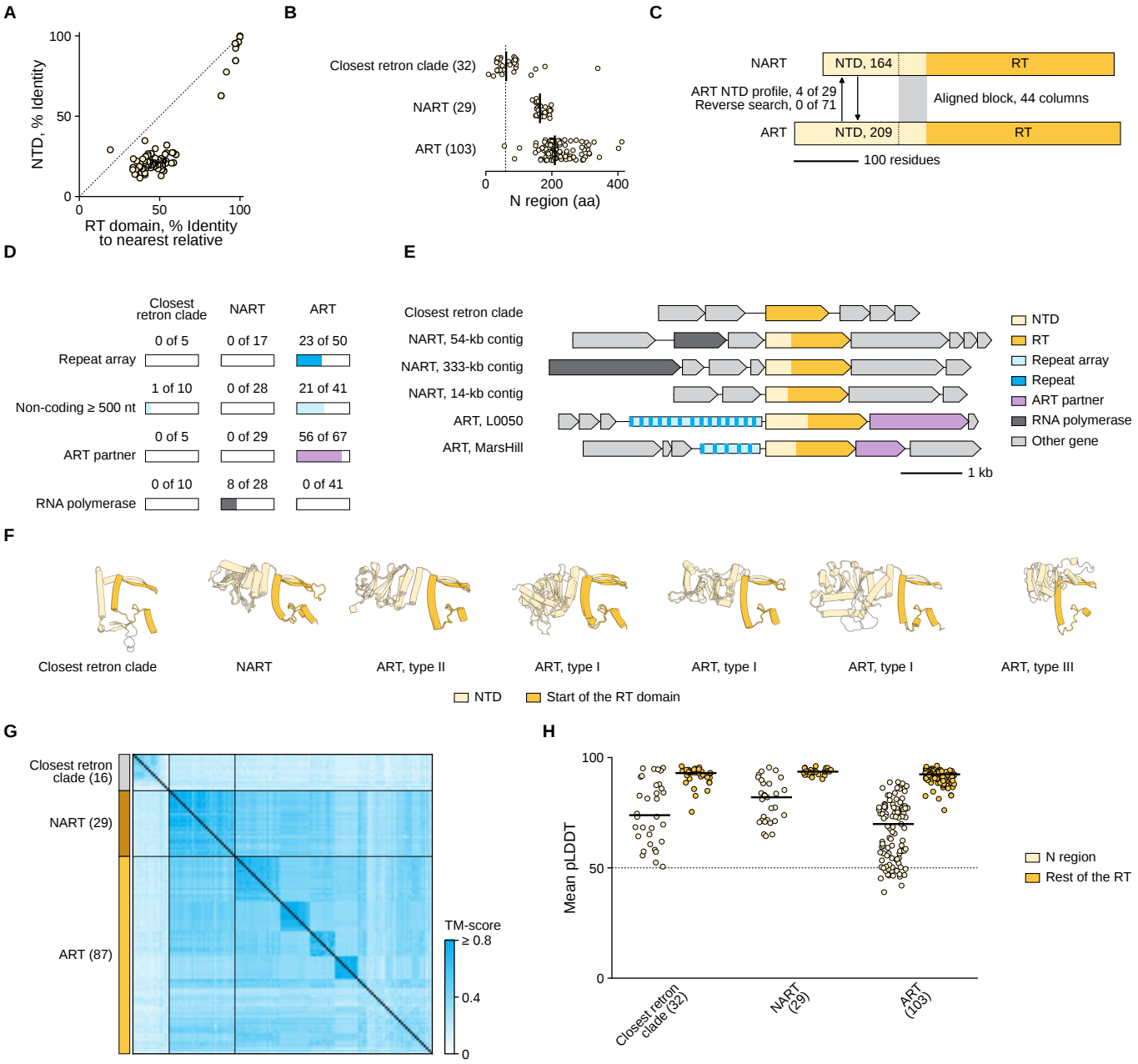
Supp. Fig. 2. Path of the agents to the ART locus. (**A**) The four steps from a census candidate to the ART repeat array locus. (**B**) Upstream regions of further relatives retrieved by the worker, each with its own repeat array, and a control RT from the polymerase-linked lineage without one. (**C**) Tests run by the agents to refute the novelty of ARTs.



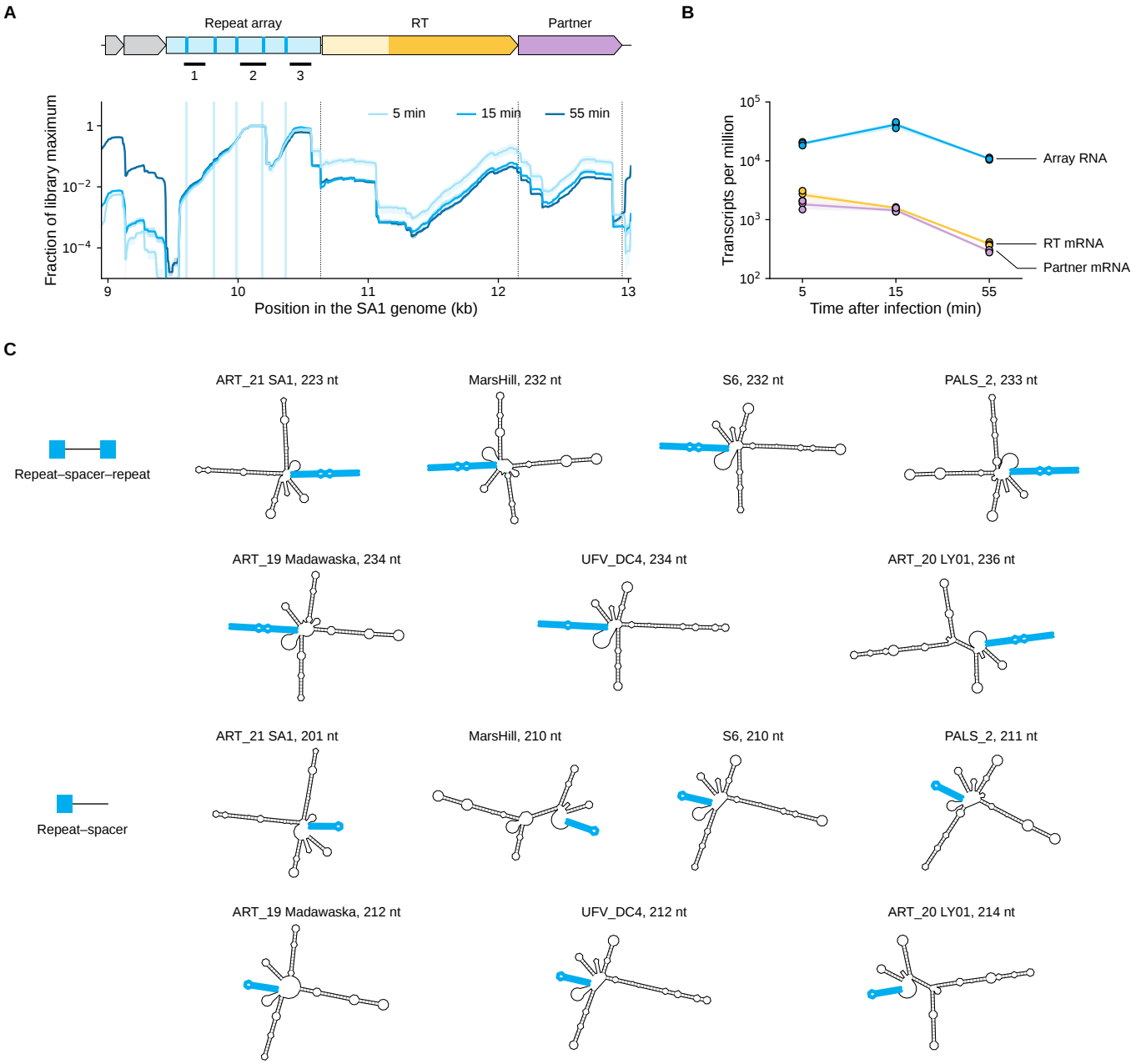
Supp. Fig. 3. The ART family and its place among RTs. (**A**) Tree of the 95 ART family members. Dashes, not assessed (less than 1.5 kb of contig upstream of the RT) or not measured. (**B**) Unrooted maximum-likelihood tree of 774 RT domains with RT classes shaded. (**C**) Schematics of diverse ART loci.



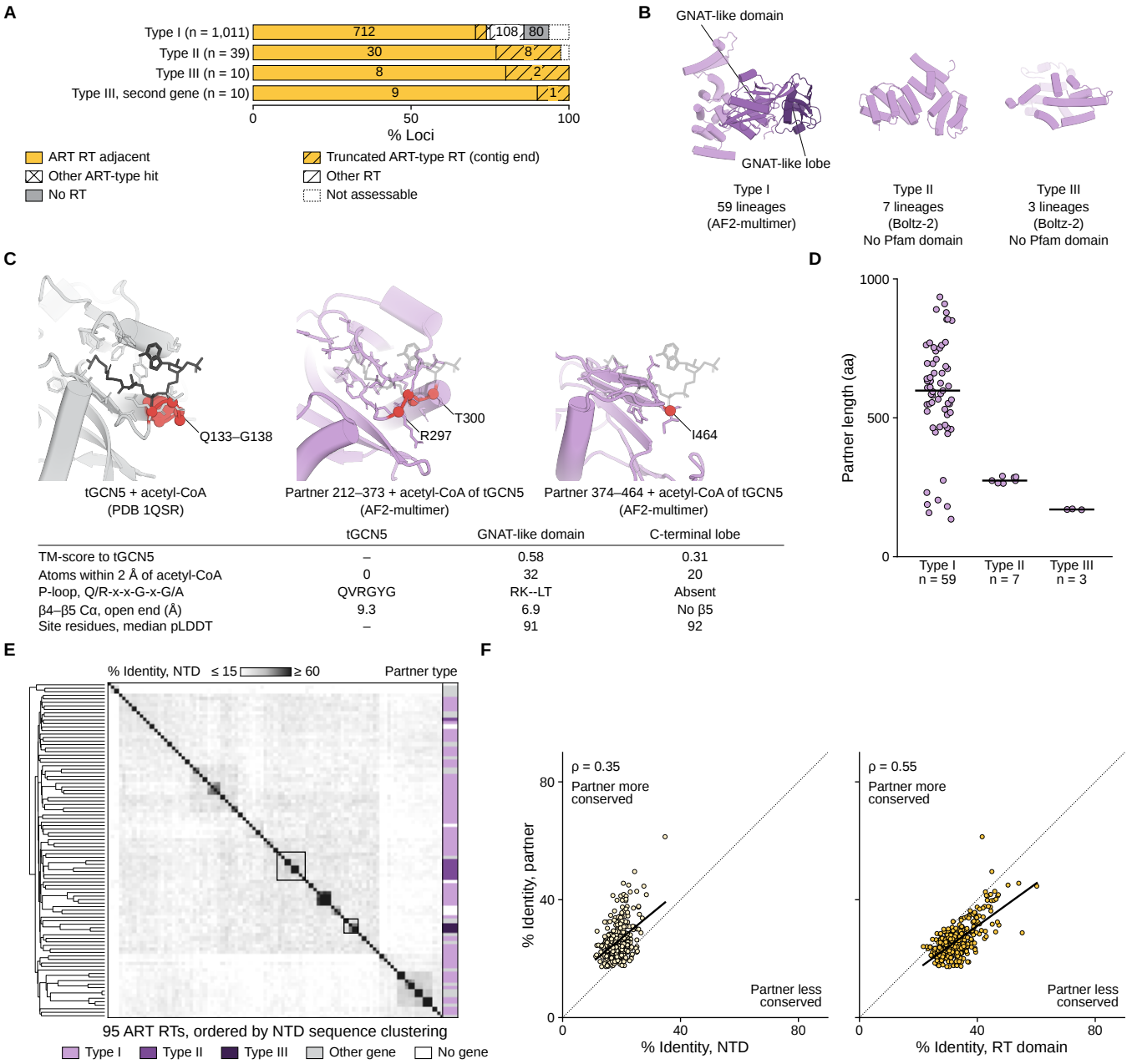
Supp. Fig. 4. The ART family RT domain carries the catalytic motif and fold of characterized RTs. (**A**) Sequence logos of the region around the catalytic YxDD motif in retron, DGR, group II intron, and ART family RTs. (**B**) AlphaFold2 model of the MarsHill ART family RT domain (gold) superposed on solved RT domains (gray). Red spheres, Cα atoms of YxDD. (**C**) TM-scores of the ART family RT core against 16 solved RT structures, by class.



Supp. Fig. 5. The N-terminal domain (NTD) of ART and NART family RTs. (**A**) Identity of the NTD against identity of the RT domain, both to the locus with the most similar RT domain. Dotted line, y = x. (**B**) Length of the N-terminal region of retron, ART, and NART RTs. (**C**) Comparison of median-length ART and NART RTs. (**D**) Loci with each feature among those assessable, for retron, ART, and NART RTs. (**E**) Example loci, to scale and aligned at the RT start. (**F**) AlphaFold2 models of the N-terminal region (cream) and RT-domain start (gold) for representative retron, NART, and ART RTs, with ART examples for each partner type (three structural clusters for type I). (**G**) Pairwise TM-scores between the 132 N-terminal regions of at least 60 residues with mean pLDDT of 50 or more, normalized by the longer region and ordered by group and structural cluster. (**H**) Mean pLDDT of the N-terminal region and of the rest of the RT in each model, by clade. Bars, medians; dotted line, 50.



Supp. Fig. 6. Array RNAs of phage SA1 during infection and predicted folds of ART array units. (**A**) Sense-strand RNA-seq coverage of the SA1 ART locus at 5, 15, and 55 min after infection (BioProject PRJNA836150). (**B**) Abundance of the array RNA, the RT mRNA, and the partner mRNA. TPM, transcripts per million. (**C**) Predicted minimum-free-energy structures, at one scale, of one repeat-spacer-repeat and one repeat-spacer unit from each of seven type II ART arrays. Repeats in blue.



Supp. Fig. 7. ART partner types and the RT N-terminal domain (NTD). (**A**) Database-wide occurrences of each partner family, by genomic context. (**B**) Predicted structure of one partner of each type. (**C**) Acetyl-CoA site of the GNAT acetyltransferase tGCN5 (PDB 1QSR; gray) compared to GNAT-like structures of ART partner proteins. (**D**) Partner length by type. Bars, medians. (**E**) Pairwise sequence identity of the NTDs of the 95 ART family RTs, ordered by hierarchical clustering. (**F**) Identity of the partner against identity of the NTD (left) and of the RT domain (right) for all 379 pairs of RTs that share a partner type.

References

(1) K. Katz et al., The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Research 50(D1), D387–D390 (2022). doi:10.1093/nar/gkab1053

(2) M. Karasikov et al., Efficient and accurate search in petabase-scale sequence repositories. Nature 647(8091), 1036–1044 (2025). doi:10.1038/s41586-025-09603-w

(3) L. Richardson et al., MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Research 51(D1), D753–D759 (2023). doi:10.1093/nar/gkac1080

(4) R. Chikhi et al., Logan: Planetary-scale genome assembly surveys life's diversity. bioRxiv [Preprint] (2024). doi:10.1101/2024.07.30.605881

(5) M. Ferrer et al., Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. Microbial Biotechnology 9(1), 22–34 (2016). doi:10.1111/1751-7915.12309

(6) M. H. Medema, T. de Rond, B. S. Moore, Mining genomes to illuminate the specialized chemistry of life. Nature Reviews Genetics 22(9), 553–571 (2021). doi:10.1038/s41576-021-00363-7

(7) K. S. Makarova et al., An updated evolutionary classification of CRISPR–Cas systems including rare variants. Nature Microbiology 10(12), 3346–3361 (2025). doi:10.1038/s41564-025-02180-8

(8) Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of Bacteriology 169(12), 5429–5433 (1987). doi:10.1128/jb.169.12.5429-5433.1987

(9) M. G. Durrant et al., Bridge RNAs direct programmable recombination of target and donor DNA. Nature 630(8018), 984–993 (2024). doi:10.1038/s41586-024-07552-4

(10) P. H. Yoon et al., A noncontiguous code for RNA-guided DNA recognition at the origin of CRISPR-Cas. Science 393(6817), 1230–1235 (2026). doi:10.1126/science.aei0498

(11) G. A. Pavlopoulos et al., Unraveling the functional dark matter through global metagenomics. Nature 622(7983), 594–602 (2023). doi:10.1038/s41586-023-06583-7

(12) Á. Rodríguez del Río et al., Functional and evolutionary significance of unknown genes from uncultivated taxa. Nature 626(7998), 377–384 (2024). doi:10.1038/s41586-023-06955-z

(13) H. Altae-Tran et al., Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. Science 382(6673), eadi1910 (2023). doi:10.1126/science.adi1910

(14) S. Doron et al., Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359(6379), eaar4120 (2018). doi:10.1126/science.aar4120

(15) S. Yao et al., ReAct: Synergizing reasoning and acting in language models, in The Eleventh International Conference on Learning Representations (ICLR, 2023).

(16) T. Schick et al., Toolformer: Language models can teach themselves to use tools, in Advances in Neural Information Processing Systems 36 (NeurIPS, 2023), pp. 68539–68551. doi:10.52202/075280-2997

(17) C. E. Jimenez et al., SWE-bench: Can language models resolve real-world GitHub issues?, in The Twelfth International Conference on Learning Representations (ICLR, 2024).

(18) T. Kwa et al., Measuring AI ability to complete long software tasks, in Advances in Neural Information Processing Systems 38 (NeurIPS, 2025), pp. 102382–102435. doi:10.52202/085713-3086

(19) J. Yang et al., SWE-agent: Agent-computer interfaces enable automated software engineering, in Advances in Neural Information Processing Systems 37 (NeurIPS, 2024), pp. 50528–50652. doi:10.52202/079017-1601

(20) J. S. Chan et al., MLE-bench: Evaluating machine learning agents on machine learning engineering, in The Thirteenth International Conference on Learning Representations (ICLR, 2025).

(21) J. Gottweis et al., Accelerating scientific discovery with Co-Scientist. Nature 655(8122), 487–496 (2026). doi:10.1038/s41586-026-10644-y

(22) J. R. Penadés et al., AI mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution. Cell 188(23), 6654–6665.e2 (2025). doi:10.1016/j.cell.2025.08.018

(23) K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. Nature 646(8085), 716–723 (2025). doi:10.1038/s41586-025-09442-9

(24) A. E. Ghareeb et al., A multi-agent system for automating scientific discovery. Nature 655(8122), 497–505 (2026). doi:10.1038/s41586-026-10652-y

(25) Y. Qu et al., CRISPR-GPT for agentic automation of gene-editing experiments. Nature Biomedical Engineering 10(2), 245–258 (2026). doi:10.1038/s41551-025-01463-z

(26) D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models. Nature 624(7992), 570–578 (2023). doi:10.1038/s41586-023-06792-0

(27) C. Lu et al., Towards end-to-end automation of AI research. Nature 651(8107), 914–919 (2026). doi:10.1038/s41586-026-10265-5

(28) K. Huang et al., Autonomous biomedical research with an artificial intelligence agent. Science 393(6813), eadz4351 (2026). doi:10.1126/science.adz4351

(29) S. Alber et al., CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. Nature Methods 23(4), 749–759 (2026). doi:10.1038/s41592-026-03029-6

(30) L. Mitchener et al., Kosmos: An AI scientist for autonomous discovery. arXiv:2511.02824 [Preprint] (2025).

(31) A. V. Anzalone et al., Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 576(7785), 149–157 (2019). doi:10.1038/s41586-019-1711-4

(32) M. G. Schubert et al., High-throughput functional variant screens via in vivo production of single-stranded DNA. Proceedings of the National Academy of Sciences 118(18), e2018181118 (2021). doi:10.1073/pnas.2018181118

(33) S. C. Lopez, K. D. Crawford, S. K. Lear, S. Bhattarai-Kline, S. L. Shipman, Precise genome editing across kingdoms of life using retron-derived DNA. Nature Chemical Biology 18(2), 199–206 (2022). doi:10.1038/s41589-021-00927-y

(34) X. Zhang et al., Harnessing eukaryotic retroelement proteins for transgene insertion into human safe-harbor loci. Nature Biotechnology 43(1), 42–51 (2025). doi:10.1038/s41587-024-02137-y

(35) S. Zimmerly, L. Wu, An unexplored diversity of reverse transcriptases in bacteria. Microbiology Spectrum 3(2), MDNA3-0058-2014 (2015). doi:10.1128/microbiolspec.mdna3-0058-2014

(36) L. Gao et al., Diverse enzymatic activities mediate antiviral immunity in prokaryotes. Science 369(6507), 1077–1084 (2020). doi:10.1126/science.aba0372

(37) A. Millman et al., Bacterial retrons function in anti-phage defense. Cell 183(6), 1551–1561.e12 (2020). doi:10.1016/j.cell.2020.09.065

(38) J. Bobonis et al., Bacterial retrons encode phage-defending tripartite toxin–antitoxin systems. Nature 609(7925), 144–150 (2022). doi:10.1038/s41586-022-05091-4

(39) M. R. Mestre, A. González-Delgado, L. I. Gutiérrez-Rus, F. Martínez-Abarca, N. Toro, Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. Nucleic Acids Research 48(22), 12632–12647 (2020). doi:10.1093/nar/gkaa1149

(40) A. Carabias et al., Retron-Eco1 assembles NAD+-hydrolyzing filaments that provide immunity against bacteriophages. Molecular Cell 84(11), 2185–2202.e12 (2024). doi:10.1016/j.molcel.2024.05.001

(41) S. Tang et al., De novo gene synthesis by an antiviral reverse transcriptase. Science 386(6717), eadq0876 (2024). doi:10.1126/science.adq0876

(42) M. E. Wilkinson, D. Li, A. Gao, R. K. Macrae, F. Zhang, Phage-triggered reverse transcription assembles a toxic repetitive gene from a noncoding RNA. Science 386(6717), eadq3977 (2024). doi:10.1126/science.adq3977

(43) M. R. Mestre et al., UG/Abi: a highly diverse family of prokaryotic reverse transcriptases associated with defense functions. Nucleic Acids Research 50(11), 6084–6101 (2022). doi:10.1093/nar/gkac467

(44) S. Tang et al., Protein-primed homopolymer synthesis by an antiviral reverse transcriptase. Nature 643(8074), 1352–1362 (2025). doi:10.1038/s41586-025-09179-5

(45) X.-Y. Song et al., Bacterial reverse transcriptase synthesizes long poly(A)-rich cDNA for antiphage defense. Science 388(6753), eads4639 (2025). doi:10.1126/science.ads4639

(46) P. Deng, H. Lee, C. Armijo, H. Wang, A. Gao, Protein-templated synthesis of dinucleotide repeat DNA by an antiphage reverse transcriptase. Science 392(6804), 1274–1281 (2026). doi:10.1126/science.aed1656

(47) M. Wang et al., Coordinated RNA- and protein-templated synthesis of double-stranded DNA by a dual reverse transcriptase immune system. Cell 189(16), 4997–5011.e11 (2026). doi:10.1016/j.cell.2026.07.012

(48) B. Zhang, J. Xu, X. He, Y. Tong, H. Ren, Interactions between jumbo phage SA1 and Staphylococcus: A global transcriptomic analysis. Microorganisms 10(8), 1590 (2022). doi:10.3390/microorganisms10081590

(49) G. Brixi et al., Genome modelling and design across all domains of life with Evo 2. Nature 652(8112), 1349–1361 (2026). doi:10.1038/s41586-026-10176-5

(50) A. Cornman et al., The OMG dataset: An open metagenomic corpus for mixed-modality genomic language modeling, in The Thirteenth International Conference on Learning Representations (ICLR, 2025).

(51) T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations. Nature 625(7995), 476–482 (2024). doi:10.1038/s41586-023-06747-5

(52) B. Romera-Paredes et al., Mathematical discoveries from program search with large language models. Nature 625(7995), 468–475 (2024). doi:10.1038/s41586-023-06924-6

(53) OpenAI, On the Navier–Stokes Millennium Prize Problem (8 September 2026). https://openai.com/index/navier-stokes-solution/ [company report with the manuscript ”Finite Time Blowup for Navier–Stokes” and a Lean formalization, not peer reviewed, accessed 19 September 2026]

(54) R. D. King et al., The automation of science. Science 324(5923), 85–89 (2009). doi:10.1126/science.1165620

(55) H. Kitano, Nobel Turing Challenge: Creating the engine for scientific discovery. npj Systems Biology and Applications 7(1), 29 (2021). doi:10.1038/s41540-021-00189-3

(56) D. B. Li et al., Coevolutionary mining of prokaryotic non-coding elements with a genome language model. bioRxiv [Preprint] (2026). doi:10.64898/2026.09.22.753630

(57) T. S. Kuhn, The Structure of Scientific Revolutions (University of Chicago Press, 1996; first published 1962). doi:10.7208/chicago/9780226458106.001.0001

(58) V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey. ACM Computing Surveys 41(3), 1–58 (2009). doi:10.1145/1541880.1541882

(59) A. Silberschatz, A. Tuzhilin, What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering 8(6), 970–974 (1996). doi:10.1109/69.553165

(60) R. E. Valdés-Pérez, Principles of human-computer collaboration for knowledge discovery in science. Artificial Intelligence 107(2), 335–346 (1999). doi:10.1016/s0004-3702(98)00116-7

(61) A. M. Korn, A. E. Hillhouse, L. Sun, J. J. Gill, Comparative genomics of three novel jumbo bacteriophages infecting Staphylococcus aureus. Journal of Virology 95(19), e02391-20 (2021). doi:10.1128/jvi.02391-20

(62) Z. Lin et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637), 1123–1130 (2023). doi:10.1126/science.ade2574

(63) M. Mirdita et al., ColabFold: making protein folding accessible to all. Nature Methods 19(6), 679–682 (2022). doi:10.1038/s41592-022-01488-1

(64) S. R. Eddy, Accelerated profile HMM searches. PLoS Computational Biology 7(10), e1002195 (2011). doi:10.1371/journal.pcbi.1002195

(65) M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology 35(11), 1026–1028 (2017). doi:10.1038/nbt.3988

(66) C. Camacho et al., BLAST+: architecture and applications. BMC Bioinformatics 10(1), 421 (2009). doi:10.1186/1471-2105-10-421

(67) K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30(4), 772–780 (2013). doi:10.1093/molbev/mst010

(68) M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS ONE 5(3), e9490 (2010). doi:10.1371/journal.pone.0009490

(69) W. Shen, B. Sipos, L. Zhao, SeqKit2: A Swiss army knife for sequence and alignment processing. iMeta 3(3), e191 (2024). doi:10.1002/imt2.191

(70) J. Shaw, Y. W. Yu, Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nature Methods 20(11), 1661–1665 (2023). doi:10.1038/s41592-023-02018-3

(71) A. P. Camargo et al., Identification of mobile genetic elements with geNomad. Nature Biotechnology 42(8), 1303–1312 (2024). doi:10.1038/s41587-023-01953-y

(72) E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22), 2933–2935 (2013). doi:10.1093/bioinformatics/btt509

(73) R. Lorenz et al., ViennaRNA Package 2.0. Algorithms for Molecular Biology 6(1), 26 (2011). doi:10.1186/1748-7188-6-26

(74) M. Šošić, M. Šikić, Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. Bioinformatics 33(9), 1394–1395 (2017). doi:10.1093/bioinformatics/btw753

(75) Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33(7), 2302–2309 (2005). doi:10.1093/nar/gki524

(76) M. van Kempen et al., Fast and accurate protein structure search with Foldseek. Nature Biotechnology 42(2), 243–246 (2024). doi:10.1038/s41587-023-01773-0

(77) D. Hyatt et al., Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1), 119 (2010). doi:10.1186/1471-2105-11-119

(78) M. C. Frith, A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Research 39(4), e23 (2011). doi:10.1093/nar/gkq1212

(79) B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 18(4), 366–368 (2021). doi:10.1038/s41592-021-01101-x

(80) B. Buchfink, H. Ashkenazy, K. Reuter, J. A. Kennedy, H.-G. Drost, Sensitive clustering of protein sequences at tree-of-life scale using DIAMOND DeepClust. bioRxiv [Preprint] (2023). doi:10.1101/2023.01.24.525373

(81) J. Mistry et al., Pfam: The protein families database in 2021. Nucleic Acids Research 49(D1), D412–D419 (2021). doi:10.1093/nar/gkaa913

(82) F. Sharifi, Y. Ye, Identification and classification of reverse transcriptases in bacterial genomes and metagenomes. Nucleic Acids Research 50(5), e29 (2022). doi:10.1093/nar/gkab1207

(83) N. Toro, R. Nisa-Martínez, Comprehensive phylogenetic analysis of bacterial reverse transcriptases. PLoS ONE 9(11), e114083 (2014). doi:10.1371/journal.pone.0114083

(84) C. Si, D. Yang, T. Hashimoto, Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers, in The Thirteenth International Conference on Learning Representations (ICLR, 2025).

(85) S. Deorowicz, A. Debudaj-Grabysz, A. Gudyś, FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Scientific Reports 6(1), 33964 (2016). doi:10.1038/srep33964

(86) S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15), 1972–1973 (2009). doi:10.1093/bioinformatics/btp348

(87) T. K. F. Wong et al., IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. Molecular Biology and Evolution 43(5), msag117 (2026). doi:10.1093/molbev/msag117

(88) S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods 14(6), 587–589 (2017). doi:10.1038/nmeth.4285

(89) D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution 35(2), 518–522 (2018). doi:10.1093/molbev/msx281

(90) S. Guindon et al., New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology 59(3), 307–321 (2010). doi:10.1093/sysbio/syq010

(91) J. Jumper et al., Highly accurate protein structure prediction with AlphaFold. Nature 596(7873), 583–589 (2021). doi:10.1038/s41586-021-03819-2

(92) S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34(17), i884–i890 (2018). doi:10.1093/bioinformatics/bty560

(93) B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nature Methods 9(4), 357–359 (2012). doi:10.1038/nmeth.1923

(94) P. Danecek et al., Twelve years of SAMtools and BCFtools. GigaScience 10(2), giab008 (2021). doi:10.1093/gigascience/giab008

(95) M. B. Stead et al., RNAsnap™: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. Nucleic Acids Research 40(20), e156 (2012). doi:10.1093/nar/gks680

(96) P. H. Yoon et al., Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. Science 385(6708), 538–543 (2024). doi:10.1126/science.adq0553

(97) M. Steinegger et al., HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20(1), 473 (2019). doi:10.1186/s12859-019-3019-7

(98) R. Evans et al., Protein complex prediction with AlphaFold-Multimer. bioRxiv [Preprint] (2021). doi:10.1101/2021.10.04.463034

(99) S. Passaro et al., Boltz-2: Towards accurate and efficient binding affinity prediction. bioRxiv [Preprint] (2025). doi:10.1101/2025.06.14.659707

(100) W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12), 2577–2637 (1983). doi:10.1002/bip.360221211

(101) C. Zhang, M. Shine, A. M. Pyle, Y. Zhang, US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nature Methods 19(9), 1109–1115 (2022). doi:10.1038/s41592-022-01585-1

(102) T. Bricken et al., Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html

(103) A. Templeton et al., Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Acknowledgments

We thank F. Zhang, B. Hie, D. Li, and G. Brixi for reviewing an early copy of the manuscript and providing feedback. We thank the Anthropic life sciences team and the broader Anthropic community for building the models and tools that enabled this study.

Materials and Methods

Autonomous research harness. The campaign was run with an autonomous research harness in which every agent was an instance of Claude Code that was configured with Claude Mythos 5. The harness permitted up to 58 concurrent sessions, and agents ran in a sandbox with 60 CPU cores, 192 GiB of memory, and no GPU. A launch agent encoded the five stages of the research brief (Supplementary Note 1) as a chain, with each stage closed by scripted completion checks. The stages were 1) input assembly, 2) database sweep, 3) RT classification, 4) neighborhood census, and 5) deep dives. No task of a later stage was opened until the preceding stage was completed. Each task was defined by a task brief and assigned to a worker agent, which wrote and executed a plan before submitting a written summary of the results with the code and data files. A supervisor agent reviewed the plan, the summary, and the edited files, and either accepted them or returned them for revision. Of the 119 tasks, 49 were revised at least once. Follow-up tasks proposed by agents during the work entered a triage queue, from which the harness released or rejected with a written reason. After each completed task, a curator agent read the task brief and the worker's summary and entered the findings into a shared knowledge base. Relevant entries were included in the prompts of later workers and supervisors. Reports were reviewed by an editor agent before filing. Plans, results, reviews, and scripts were committed to a version-controlled record readable by every agent. A campaign ended when no task remained that could be dispatched.

Tools and skills available to the agents. Agents acted through the tools of Claude Code (a shell, file access, and web search). The harness added connectors to a metagenomic protein database, to literature search and full-text retrieval services, to the shared knowledge base, and to a queue for GPU jobs. Structure predictions were dispatched to NVIDIA A100 and L4 GPUs (19 jobs) and used either ESMFold (esmfold_v1) (62) or ColabFold 1.6.1 with AlphaFold2-ptm models (63). The sandbox provided Python with Biopython, NumPy, pandas, SciPy, and PyMOL, and agents installed further software as required. For sequence and RNA analysis, the agents ran HMMER 3.3.2 (64), MMseqs2 release 17-b804f (65), BLAST+ 2.16.0 (66), MAFFT 7.490 (67), FastTree 2.1.11 (68), SeqKit 2.8.2 (69), skani 0.3.2 (70), geNomad 1.12.0 (71), Infernal 1.1.5 (72), ViennaRNA 2.7.2 (73), and edlib 1.3.9 (74). For structure analysis, they ran TM-align (75) through tmtools 0.3.0 and the Foldseek web server (76). They also queried NCBI, UniProt, InterPro, the European Nucleotide Archive (ENA), the AlphaFold Protein Structure Database, the Protein Data Bank, and Rfam. In the retained session logs the agents issued 7,578 shell commands, 696 database queries, 131 literature searches, and 61 web requests. The sandbox also held a library of about 140 skills (written method guides) that workers consulted 15. Seven of these were guides for metagenomic surveys. They covered 1) counting the independent occurrences of a gene family, 2) scoring operonic arrangement against the background of the window, 3) adjudicating candidate families, 4) testing novelty claims, 5) annotating uncharacterized protein families, 6) proposing mechanisms together with the counterfactual evidence, and 7) drawing gene-neighborhood figures. The other libraries covered database use, literature access, GPU job dispatch, report editing, and the use of bioinformatic tools.

Metagenomic sequence database. The database used in this study held predicted proteins from public metagenomic and genomic assemblies that spanned approximately 11 million biosamples, mainly from Logan (4), ENA, the Joint Genome Institute (JGI), and NCBI. Genes were predicted with prodigal-gv 2.10.0 (71, 77) under default parameters in unannotated assemblies of Logan, ENA, and NCBI, and the source gene calls were retained for annotated assemblies of NCBI and JGI. Each protein was linked to its contig,

coordinates, strand, and biosample. Proteins were deduplicated, and were excluded if they had incomplete coding sequences, non-standard amino acids, more than 8,000 residues, degenerate k-mer repeats, or at least 50% low-complexity sequence as masked by tantan (78). The remaining proteins (approximately 15.8 billion) were clustered with DIAMOND (79, 80). Linear-time clustering was applied at 90% and then 70% identity with at least 80% coverage of the shorter sequence and was followed by cascaded clustering at 50% identity with at least 80% mutual coverage. Each cluster was represented by the member that was closest to the 80th percentile of length in the cluster. This yielded 1,939,242,578 clusters at 50% identity, whose representatives (hereafter, database representatives) formed the search space of the campaign and are the protein clusters that are counted in the main text. Agents accessed the database through a programmatic interface that returned the members of a cluster, protein and nucleotide sequences, gene neighborhoods, biosample records, and precomputed Pfam-A annotations (81). The interface also ran MMseqs2 searches (65) either against the database representatives or against a reference set of 365 million proteins (UniProt and NCBI nr, clustered at 90% identity).

RT census by the agents. 52 profile hidden Markov models (HMMs) were assembled from six Pfam profiles (RVT_1, RVT_2, RVT_3, RVT_N, GIIM, and Intron_maturas2) (81), the 45 class-specific profiles of myRT (82), and one profile that was built from an alignment from Toro (83). All database representatives were searched with hmmsearch (--noali -Z 8 --domZ 8 -E 0.01 --domE 0.01), which identified 3,391,714 candidate RT sequences. Candidates were retained if the RT core was complete (profile coverage ≥ 0.75) and the length reached a class-specific minimum (225 to 346 residues). The remainder were discarded as fragments (n = 2,016,377) or weak hits (n = 1,171,334; bit score < 25, coverage < 0.35, and no YxDD catalytic motif), or excluded for other reasons (n = 622). These selection thresholds were fixed beforehand on the five reference RTs that were named in the brief (Ec86, LtrA, BPP-1 Brt, AbiK, and an RT-Cas1 fusion) and on 2,339 labeled myRT sequences, which retained 99.0% of full-length labeled RTs. The 203,381 retained sequences were clustered with MMseqs2 easy-cluster (--min-seq-id 0.5 -c 0.8 --cov-mode 0) into 198,290 RT clusters. Clusters were then classified in three tiers. The first tier used 38 lineage-specific HMMs that were built from myRT (mafft --auto, hmmbuild) (67), with a class-specific bit-score floor and a margin of at least 10 bits over every other class (precision 0.997 in five-fold cross-validation). The second used MMseqs2 searches against labeled RTs (bit score ≥ 150, query coverage ≥ 0.5, identity ≥ 30%). The third used a FastTree tree (-lg -gamma) (68) of 2,241 RT domains that were aligned to RVT_1, which required agreement among the five nearest labeled leaves. This placed 35,168 clusters in the seven known classes, namely retron (11,517), unknown group (UG; 10,759), group II intron (5,461), diversity-generating retroelement (DGR; 3,857), CRISPR-associated (2,067), group II-like (827), and abortive-infection (Abi; 680). A further 25,737 clusters formed a loosely organized group on the tree apart from any labeled member and were treated as an eighth class of “novel” RTs. The 137,385 scattered clusters that could not be placed formed the ninth class of “unplaced” RTs.

Sampling of RT neighborhoods. The worker that was assigned to the census defined a sampling cap for each RT class and drew 7,308 RT clusters as anchors. The abortive-infection (n = 680) and group II-like (n = 827) classes were sampled exhaustively, and the DGR and CRISPR-associated classes were sampled at random (1,000 clusters each). The UG class (700 clusters) was sampled in proportion to clade size, with clades of fewer than 30 clusters taken whole. Every retron cluster and every group II intron cluster carried a single clade label, so these two classes were sampled at random (1,000 retron clusters, with one control cluster added, and 500 group II intron clusters). The novel class was sampled at 60 clusters from each of its

16 clades, and the unplaced class was sampled with one representative of each of its 440 largest lineages plus 200 random singletons. At these caps, the worker estimated that a family had to occur at 0.3 to 0.6% of the loci of a class to reach three independent occurrences. For each anchor, up to three loci were retrieved from each of four source collections (Logan, ENA, JGI, and NCBI) with 10 kb of flanking sequence on each side. Loci were recovered for 7,238 of the 7,308 anchors (10,983 loci with 79,680 coding sequences), and 85.9% of the windows were truncated by a contig end.

Scoring of partner families. Neighboring proteins were grouped into families either by their best Pfam-A match or, where none was found, by MMseqs2 clustering. The 3,564 families that recurred across loci were evaluated against three criteria (promotion filters). The worker that was assigned to the census had set the thresholds of these criteria before the census was run. First, the family had to be conserved within at least one RT clade. Second, its occurrences had to be independent, meaning that they were found in at least three clusters at 90% identity, in at least three biosamples, and beside RTs of at least two different classes. For example, families that were found only beside retrons could not meet this criterion, because all retrons carried one clade label. Third, genes of the family had to lie nearer to the RT than randomly drawn genes of the same loci ($P \leq 0.05$). They also had to lie either on the strand of the RT or within 100 bp of the adjacent gene. Known partners of RT systems and general residents of mobile elements were designated as controls beforehand. Forty-six families met all three criteria, and 13 of these were controls that included Avd, Cas1, and Cas2. Of the other 33, the worker promoted 16 after a review of their annotations. A seventeenth family was promoted later by a follow-up task.

Deep dives. In each deep dive, the worker re-applied the three criteria to the members of the promoted family, examined alignments, and searched annotations and the literature for prior descriptions of the association. A family was counted as confirmed when its report upheld it as a new RT partner. Otherwise it was counted as rejected or set aside.

Campaign accounting. The research campaign comprised 119 tasks. Of these, 107 were completed, 10 were rejected at triage, and 2 were stalled (one after ten revisions and one after ten failed completion checks). One task assembled the input data, 96 performed analyses, and 22 wrote or revised a report. Five tasks were seeded from the stages of the brief, 16 were deep dives that were seeded from the promoted families, and 98 were follow-up tasks. The agents ran 949 sessions (the launch session and 414 worker, 375 supervisor, 107 curator, and 52 editor sessions). Agent-hours were calculated as the summed duration of all agent sessions, including the execution of tools within sessions (76.9 agent-hours, 63.7 of them in worker sessions). Token use was calculated as the sum over all sessions of uncached input tokens (11.3 million), output tokens (14.9 million), and tokens that were written to the prompt cache (189.5 million), which gave 215.6 million tokens. Tokens that were read from the prompt cache were excluded. To attribute token use to the stages of the campaign, the tokens of each session were assigned to its task, and each follow-up task was assigned to the survey stage or deep dive from which it descended in the task record. To determine whether each follow-up task originated with a worker or a supervisor, a Claude instance was given the brief and worker summary of the parent task and the brief of the follow-up task. It classified each follow-up task as either proposed by the worker or originating in the supervisor's review. The classification was run twice, and the six discordant calls were resolved by reading the task records.

Ranking of reports. The 19 reports of the campaign were ranked in a tournament of pairwise comparisons, a design that was used previously to rank machine-generated hypotheses (21, 24, 84). Every ordered

pair of reports was compared once (342 games). In each game, a language-model judge (Mythos 5) read a fixed rubric and two reports under anonymized file names before scoring each report from 1 to 5 for impact, novelty, soundness, and actionability (weights 0.35, 0.30, 0.25, and 0.10). The report with the higher weighted score won, and a report with a soundness score of 2 or less lost automatically. Ratings were fitted to all games with a Bradley–Terry model. The ART report ranked third (32 wins in 36 games).

Transcript analysis of the discovery. Claude Code records each session as a log of reasoning blocks, messages, tool calls, and tool results, which along with messages make up the reasoning trace. The log of task t0062, in which a worker recognized the ART array, was parsed into an ordered list of events. Recognition was placed at the first reasoning block or message that described a repeat array, and the tool results just before it were taken as what the worker had read. The brief of this task, which was written by the supervisor agent of the preceding task, requested a test for a retron-type ncRNA upstream of the RT. Neither this task brief nor the research brief mentioned repeats or arrays. The two retrievals before recognition returned sequence without annotation, and no tool call came between them and recognition. No earlier tool call contained a repeat analysis, and no established repeat finder was run in the session. Every repeat that the worker named was recounted in the retrieved sequences. The first repeat pattern that the worker wrote (CATGTGT[AT]TCGCATGT) was recounted on both strands of the 2,900-nt sequence that the worker had retrieved, with at most 1 mismatch. This sequence is the flank of locus L0050, the worker's label for the locus at which it first recognized an array. Quotations from the session logs are verbatim, including the agents' emphasis marks, with omissions marked by bracketed ellipses and shading added. Loci that the agents examined were drawn to scale from the gene tables that were printed in the logs. Repeat positions were obtained by re-scanning the logged sequences with the worker's repeat core, with at most 2 mismatches allowed. For phage MarsHill, the GenBank record (MW248466.1) was scanned.

Analyses that the agents ran on the array. The agents assembled the RT family from MMseqs2 searches and took as members the RTs that were linked to one another by hits above a score threshold. In each upstream region the most frequent 14-mer was taken as the repeat, its copies were located with up to two mismatches, and the sequences between neighboring copies were taken as spacers. For each of the 42 spacers from six arrays, the longest stop-free peptide was searched against a reference set of 365 million proteins. The spacers of two cultured phages were each aligned to the genome of the other phage with edlib. The folding energy of 40 spacers was predicted with ViennaRNA and compared with that of 30 shuffled sequences of the same dinucleotide composition. Novelty was tested against the annotations of family members and the literature on retrons, defense-associated RTs, AbiK, and the CRISPR arrays of jumbo phages. The fractions that are given for these analyses (spacers without a protein match at $E \leq 0.001$, and spacers with $z < -2$ against the shuffled sequences) were taken from the agents' report and were not recomputed.

Human-directed analyses. All analyses that were performed after the campaign were carried out in interactive Claude Science sessions, in which the authors directed the analysis and Claude wrote and ran the code.

Family definition. Homologs of the MarsHill RT (GenBank QQM14740.1) were collected from GenBank with BLASTP (66) and from our metagenomic database with a profile HMM. The profile was built from 12 of the GenBank phage RTs (MAFFT --auto, hmmbuild). Ten of these came from *Staphylococcus* phages and two from *Listeria* phage LPJP1. The profile was searched against all database representatives with

hmmsearch (HMMER 3.4; E ≤ 1×10^{-5}), which returned 9,100 proteins. Hits of 350 to 900 residues (n = 4,379) were pooled with the 12 phage RTs and 2,019 reference RTs of the myRT classification (82). The pooled set was compared all against all with DIAMOND blastp (--very-sensitive, E ≤ 1×10^{-5}) (79), and proteins were linked at bit scores of at least 140. The connected component that contained the MarsHill RT held 117 proteins and no reference RT. All members of the database clusters that were represented in this component were then retrieved, and the 706 further members of at least 400 residues were added to the 117. The resulting 823 full-length RTs were clustered with MMseqs2 easy-cluster (--min-seq-id 0.9 -c 0.8) into 230 clusters, and the member on the longest contig was taken as the representative of each cluster. Representatives were aligned with FAMSA (85), and a tree was inferred with FastTree under the LG model (68). The ART family was defined as the smallest clade of this tree that contains every locus with both a type I partner gene and a non-coding region of at least 500 nt upstream of the RT. Partner types are defined under Partner genes and partner types. This clade held 95 representatives, which were numbered ART_01 to ART_95 in tree order. Members that were encoded on GenBank phage genomes were classed as cultured phages. Contigs were classified as virus, plasmid, or chromosome with geNomad 1.12.0 (database release 1.9, end-to-end workflow, default settings) (71), and gene calls on database contigs were taken from the database annotation.

Repeat-array detection. Arrays upstream of RTs were handled in three steps, with each step described under its own heading. They were 1) detected with a k-mer scan, 2) delimited and measured with a procedure that does not fix the repeat length, and 3) annotated copy by copy for sequence comparisons between arrays. The scan searched up to 3,000 nt upstream of the RT start codon for short words that recurred at least three times at regular spacing (100 to 450 nt, start to start) and was run with two word settings. For the 95 members, the 20 most frequent 14-nt words of the window served as seeds (at least three distinct bases, no homopolymer of six or more). Copies were non-overlapping windows within two mismatches of a seed. An array was called when the longest run of regularly spaced copies, R, was at least 3 and exceeded the longest such run in 100 mononucleotide shuffles of the window. No array was called when a longer run of copies that were spaced under 100 nt apart was present. R was taken as the copy number of the array. One locus with R = 3 that did not exceed its shuffles was retained because an exact 12-nt word recurred three times at such spacing. For the tips of the RT phylogeny, which include retron loci, an array was called either from annotated repeat copies (see Array annotation and comparison) or from an exact 12-nt word that recurred three times at such spacing. Under both settings, loci with less than 1,500 nt of contig upstream of the RT and no array were recorded as not assessed. Neither setting called an array at any NART locus.

Array delimitation and measurement. Up to 6,000 nt upstream of the RT was rescanned in the 95 members and in 8 further GenBank phage ART loci with every recurring 10-nt word as a seed. Copies were collected at up to one mismatch per seed. The longest chain of copies at near-constant spacing (60 to 600 nt, 30% tolerance, single skipped copies allowed) was retained for scoring. The score of a chain was the number of copies minus one, multiplied by the information content of its conserved block relative to the base composition of the region. A chain was retained when its score exceeded the best chain score in each of 200 shuffles of the region, in both a 3,000-nt and a 6,000-nt window. Shuffles permuted nucleotides within consecutive 50-nt blocks to preserve local base composition. Chains that scored less than twice the best shuffle score were retested against 2,000 shuffles. A position weight matrix of the repeat was then built, and matches that scored above the maximum of 200 shuffled regions were counted as copies. The repeat

was taken as the contiguous run of columns in which at least 80% of copies carry the consensus base, with one lapse tolerated. The repeat length was measured both on ungapped copies and on a MAFFT alignment of the copies. Chains whose spacings were all multiples of three and whose copies lay within an annotated gene were set aside as coding repeats. An array was called adjacent to the RT when no annotated gene of 300 nt or more lay between its last copy and the RT. Each locus was also scanned with the position weight matrix of every other array to group arrays that share a repeat. Copy numbers, spacings, repeat and spacer lengths, and array spans that are given in the text are from this procedure, which delimited arrays adjacent to the RT at 28 of the 95 loci and detected none at NART loci.

Reference RT sets. Two reference sets of RTs were assembled. The 957-sequence set, which was used to compare N-terminal lengths across RT classes, comprised 633 labeled RTs from published classifications (39, 43, 82), 294 classified representatives from our database, and 30 ART RTs. Each lineage of the published classifications was clustered at 50% identity with MMseqs2. Labeled references were then sampled with a quota per lineage that was proportional to the square root of its size. The 294 database representatives were drawn from one of 64 equal partitions of the database. They had been assigned to a class by their best-scoring profile in a panel of RT class profiles ($E \leq 1 \times 10^{-10}$, profile coverage ≥ 0.60) and were clustered at 40% identity. The 774-sequence set, which was used for the RT phylogeny, comprised 98 ART and NART RTs, 617 retron RTs, and 59 group II intron, DGR, UG/Abi, and CRISPR-associated RTs. Of the retron RTs, 325 came from published classifications (39, 82) and 292 were retron-classified representatives from our database. The published retron RTs spanned retron clades 1 to 11 and the retron-like RTs of these classifications, and included the 13 references that were most similar to MarsHill RT homologs in a DIAMOND search (79).

Phylogeny. RT-domain sequences of the 774-sequence set (taken as the Pfam RVT_1 envelope (81) extended by 40 residues on each side) were aligned with MAFFT L-INS-i (67) and trimmed with trimAl (-gappyout) (86), which left 254 columns. A maximum-likelihood tree was inferred with IQ-TREE 3.1.2 (87) under the model that ModelFinder (88) selected by BIC (Q.pfam+F+R6), with 1,000 ultrafast bootstrap (89) and 1,000 SH-aLRT (90) replicates, and was midpoint-rooted. In this tree the ART and NART RTs form the two child clades of the node that unites them (71 and 29 tips, which include two database representatives that had been sampled as retrons). For display, every maximal clade with at least 90% of tips from one RT class was collapsed.

N-terminal region. The N-terminal region of an RT was defined as the residues that precede the envelope of its first Pfam RVT_1 (PF00078) domain (hmmsearch, HMMER 3.4; independent $E \leq 1 \times 10^{-3}$). This boundary was used throughout except in Fig. 2D and Supp. Figs. 3A, 5A, and 5E, where the region ends 40 residues before the envelope, at the start of the RT-domain sequence that was used for phylogeny. N-terminal lengths were compared across the classes of the 957-sequence set (group II intron and group II-like, n = 226; CRISPR-associated, 125; DGR, 32; UG or Abi, 434; retron, 110; ART, 30).

Catalytic-motif logos. The YxDD motif was located with the pattern [YF].DD in the 95 ART members and in RVT_1 domain sequences from the 957-sequence set. These domain sequences came from the retron (n = 110), DGR (n = 32), and group II intron (n = 206) representatives. Windows of 10 residues on either side of the motif were stacked without gaps. Letter heights are residue frequency multiplied by column information content ($\log_2 20$ minus Shannon entropy).

Structure prediction. AlphaFold2 (91) models were built with ColabFold 1.6.1 (63) for the 95 ART members, the RTs of the 8 further GenBank phage ART loci, 29 NART RTs, and 32 retron RTs that are closely related to them. Settings were alphafold2_ptm, 5 models, 3 recycles, no templates, and no relaxation, and the top-ranked model was used. Each MSA combined the ColabFold MMseqs2 server alignment (UniRef30 2302, ColabFold environmental database 202108) with an alignment of related RTs that was mapped onto the query sequence. The MarsHill RT model that was used for the structural comparisons was built with the same ColabFold version from the server alignment alone (3 models, 3 recycles; top-ranked model used).

Structural comparisons. Retron Ec86 (PDB 7V9X, chain B), DGR BPP-1 (PDB 8UBE, chain A), and the NART model were superposed on the MarsHill model over the RT domain in PyMOL. N-terminal residues with a per-residue model confidence (pLDDT) below 50 were not drawn. The RVT_1 ranges of Ec86, BPP-1, and the GsI-IIC group II intron RT (PDB 6AR1, chain A) were superposed on the MarsHill RT domain with TM-align (75). Structural similarity was reported as the TM-score that was normalized by the length of the MarsHill RT domain. The MarsHill RT core (residues 211 to 501) and N-terminal region (residues 1 to 210) were each searched against a panel of 16 solved RT structures with Foldseek 10.941cd33 (easy-search --exhaustive-search 1 -e 10) (76). The panel comprised PDB 7V9U, 8QBK, 8QBL, 8QBM, 9NNB, 9P4J, 8UB9, 5HHL, 6AR1, 9C0I, 9IOA, 7R06, 7R07, 7R08, 8OZ7, and 1RTD. It was chosen by hand to cover solved RTs of retrons, a DGR, group II introns, DRT2, DRT9, and abortive-infection systems, with HIV-1 RT as an outgroup. The highest query-normalized TM-score over the chains of an entry was taken as the score of that entry. To cluster N-terminal folds, N-terminal regions of at least 60 residues with mean pLDDT of at least 50 were compared all against all with TM-align on N-terminal coordinates only (132 of 164 models; 87 ART, 29 NART, 16 retron). The TM-score that was normalized by the longer region was used. Structural clusters were defined by average linkage on 1 − TM-score, cut at 0.5.

N-terminal sequence comparisons. To compare the divergence of N-terminal regions with that of RT domains, each locus was paired with the locus whose RT domain was most similar. Identities of the RT domains and of the N-terminal regions were computed from global alignments (BLOSUM62, gap open −11, extension −1) as identical positions over the length of the longer sequence.

Locus features by clade. Four locus features were recorded for each tip of the ART (n = 71), NART (n = 29), and nearest retron (n = 13) clades of the RT phylogeny and were counted only where they could be assessed. Repeat arrays were called as described under Repeat-array detection. A non-coding tract was recorded when at least 500 nt separated the RT from the preceding gene. An ART partner was recorded when the protein that was encoded directly downstream on the RT strand matched the partner protein of an ART locus (phmmer, $E \leq 1 \times 10^{-3}$). An RNA polymerase was recorded when an upstream gene on the RT strand carried a Pfam RNA-polymerase domain.

Comparison of the SA1 and MarsHill loci. Loci were taken from the GenBank records of phages MarsHill (MW248466.1) (61) and SA1 (MW218148.1) (48). Repeat copies were located as matches to the 14-nt core ATATGAATACGTAT with at most one mismatch. The MarsHill array region, from 60 nt before the first core to the RT start codon, was aligned globally to that of SA1 (match 2, mismatch −1, gap open −4, extension −0.5). The RT genes were aligned as codons, with a global protein alignment (BLOSUM62) as the guide. Per-nucleotide identity was smoothed with a centered 25-nt running mean. The five MarsHill copies were aligned without gaps at the core with 29 nt on either side, and positions with the same base in at least four copies were classed as conserved.

Array annotation and comparison. For sequence comparisons between arrays, repeat copies were annotated in ART loci that hold at least three exact copies of a locus-specific 10-nt repeat word, with each copy within 600 nt of the next. The word of each locus was taken from a scan of up to 3,000 nt upstream of the RT for 10-nt words that recurred in tandem (at least three exact copies). The periodicity of the word was 141 to 317 nt in the 24 loci with such a word. A 26-nt consensus (6 nt before to 20 nt after the start of the word) was built from the exact copies. The strand was then rescanned, and windows that matched at 21 or more of 26 positions were taken as copies. Gaps of at least 1.6 times the usual spacing of an array were then searched, within 40 nt of each expected position, for degenerate copies that matched at 18 to 20 positions. Of the 24 annotated arrays, 16 were analyzed. Five *Staphylococcus* phage arrays that are closely related to those of MarsHill and SA1 were omitted, and three loci were excluded (no usual spacing definable, too few exact copies, or RT outside the ART clade). To profile conservation along the array unit, windows from 60 nt before to 160 nt after the start of each exact copy were stacked without gaps. Identity at each position was the mean over all pairs of copies and is 100% at the 10-nt word by construction. To compare repeats, flanks, and spacers between arrays, each copy was divided into the 26-nt repeat, its flanks (20 nt on either side), and the spacer between the flanks of consecutive copies. Repeats and flanks were compared without gaps, and spacers were aligned globally (match 1, mismatch −1, gap open −3, extension −1). Chance identity was the same measure with one sequence of each pair shuffled (median of 20 mononucleotide shuffles).

RNA-seq analysis of published SA1 infection data. RNA-seq reads of SA1 infection of *Staphylococcus lentus* (BioProject PRJNA836150 (48); 12 paired-end, strand-specific libraries from three cultures sampled before infection and at 5, 15, and 55 min after infection) were downloaded from ENA. Reads were trimmed with fastp 1.3.6 (minimum length 30 nt) (92) and aligned with Bowtie 2 v2.5.5 (--very-sensitive -X 1000 --no-unal) (93) to the SA1 genome (MW218148.1) together with the *S. lentus* chromosome (NZ_CP059679.1). Alignments were sorted and indexed with SAMtools 1.24 (94). Properly paired alignments with MAPQ of at least 10 and a template of at most 1,500 nt were retained as fragments. Each fragment was counted once on the strand of read 2 (the sense read) and assigned to the feature that contained its midpoint. Features were the 258 annotated SA1 coding sequences and the array RNA, which was defined as the whole non-coding region upstream of the RT (positions 9,448 to 10,646). Transcripts per million (TPM) were computed from sense fragments per kilobase over these 259 features. TPM was averaged over the three cultures of a time point, and features were ranked by this mean. The share of phage RNA was calculated as the number of array-RNA fragments divided by all fragments that aligned to the SA1 genome in that library. Coverage was calculated as the number of fragments that spanned each position per million phage fragments. Array coverage was averaged over the three 15-min cultures and, for comparison of replicates, was scaled to the maximum of each library over the array.

ART expression plasmids and cultures. Two plasmids that carry parts of the SA1 ART locus (GenBank MW218148.1) were assembled by Golden Gate cloning and verified by whole-plasmid sequencing. The native-locus plasmid carries positions 9,448 to 12,951 (non-coding region, RT, and partner genes). This insert was synthesized as five fragments and assembled downstream of a $P_{LtetO-1}$ promoter on a pSC101-origin vector that encodes TetR and ampicillin resistance. The clone that was used carries one substitution in the non-coding region (C to T at position 9,936). The heterologous-promoter plasmid carries the 1,187-nt non-coding region behind the J23119 promoter and a synthesized RT gene behind the *araBAD* promoter, on a pMB1-origin vector that encodes AraC and kanamycin resistance. Its RT gene was recoded to encode the 505-residue protein that starts at position 10,635, 12 nt upstream of the annotated start codon, and its non-

coding region was amplified by PCR from a native-locus clone. Plasmids were carried in *E. coli* DH10B, and one culture per construct was grown at 37 °C in Terrific Broth with 100 µg/mL carbenicillin (native-locus plasmid) or 50 µg/mL kanamycin (heterologous-promoter plasmid). The native-locus culture was grown without anhydrotetracycline to an OD_{600} of about 0.4 and harvested 2 h later. The heterologous-promoter culture was diluted 20-fold from an overnight culture into medium with 0.2% L-arabinose and harvested after about 4 h.

Small RNA sequencing. RNA was extracted with hot formamide as previously described (95, 96). In brief, cell pellets were resuspended in 95% formamide with 18 mM EDTA at 65 °C for 5 min. The cleared extract was purified on RNA Clean & Concentrator-5 columns (Zymo Research), which retain RNAs of 17 nt and longer, with in-column DNase I digestion. No size selection was applied. Ribosomal RNA was depleted from 500 ng of RNA with the NEBNext rRNA Depletion Kit (Bacteria) (New England Biolabs). RNA ends were repaired with Quick CIP and then with T4 polynucleotide kinase, first without and then with ATP (New England Biolabs), and RNA was recovered on SPRI beads. Libraries were prepared with the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs) following the manufacturer's protocol, except that adaptors and the reverse-transcription primer were diluted twofold. Libraries were cleaned with 1.5 volumes of SPRI beads. Pooled libraries were sequenced on a NextSeq 2000 (Illumina) with 2 × 151-nt reads and 20% PhiX. Reads were demultiplexed and adapter-trimmed with BCL Convert 4.2.11 (Illumina) and filtered with fastp 1.3.6 (adapter trimming off, poly-G trimming, minimum length 12 nt). Filtered reads were aligned with Bowtie 2 v2.5.5 (--end-to-end -D 20 -R 3 -N 0 -L 12 -i S,1,0.50 -X 1000) to the plasmid of each library together with the *E. coli* DH10B genome (NC_010473.1) and phiX174. Alignments were sorted with SAMtools 1.24 (94). Fragments were taken from the start of read 1 to the end of read 2 for proper pairs and from the read itself otherwise, with read 1 in the sense orientation. Secondary and supplementary alignments, fragments longer than 1,000 nt, and alignments with MAPQ below 2 were discarded. Coverage was reported as raw sense-strand fragment counts in SA1 genome coordinates.

Array RNA boundaries. Array RNAs were called as contiguous runs of at least 40 nt at or above half of the maximum coverage over the array. RNA 1 was called in the *E. coli* library that expressed the array from the J23119 promoter, and RNAs 2 and 3 were called on the mean coverage of the three 15-min infection cultures.

RNA secondary-structure prediction. Minimum-free-energy structures were predicted with RNAfold of ViennaRNA 2.7.2 (73) at 37 °C with default parameters and drawn at one scale within a panel. The Ec86 ncRNA was the 169-nt interval between the outer inverted repeats of retron Ec86 (M24408.1, positions 373 to 541), and SA1 RNAs 1 to 3 were the intervals that were called from coverage. Array units were cut from the arrays of seven *Staphylococcus* phages, SA1 (MW218148.1), MarsHill (MW248466.1), Madawaska (MW349129.1), LY01 (OR836606.1), S6 (LC680885.1), PALS_2 (MN091626.1), and UFV_DC4 (MZ779063.1). A repeat was defined as the 22 nt from 4 nt before to 17 nt after the start of the 14-nt core, and a spacer was defined as the sequence between consecutive repeats. Units of the SA1 array were built from repeats 3 and 4 and spacers 3 and 4. For each of the seven arrays, two units were folded. One was formed by repeat 3, spacer 3, and repeat 4, and the other was formed by repeat 3 and spacer 3.

Partner genes and partner types. The partner of each of the 95 ART RTs was taken as the first gene downstream of the RT. Proteins that were encoded in the 230 representative loci were clustered with MMseqs2 cluster (--cluster-mode 1 -e 0.01 -s 7.5 -c 0.5), and the cluster alignments were compared all

against all with HHblits (2 iterations, E ≤ 0.01) (97). Clusters whose profiles matched with at least 90% probability over at least half of both profiles were merged into families. The three families that recurred as partners of ART RTs were designated types I to III, and type III loci carry a second recurring family directly downstream of the partner. Similarity between types was assessed on predicted structures of eight partners (six of type II, one of type I, and one of type III; see RT-partner co-folding and partner models). The eight were compared all against all both by sequence (phmmer, HMMER 3.4) and by structure (TM-align, after removal of residues with pLDDT ≤ 50). Members of each family were also searched against the profile HMMs of the other families. Thirty-six partner sequences of the three types were searched against Pfam-A (81) with hmmsearch at gathering thresholds (--cut_ga), with the search-space size set to the number of models.

Database-wide occurrence of partner families. A profile HMM was built for each partner family (MAFFT L-INS-i, hmmbuild) from 46 type I partners, 13 type II partners, and the 8 and 6 database members of the two type III genes. Each profile was searched with hmmsearch against all database representatives. A representative was assigned to a family when it scored at or above the lowest-scoring known member (leave-one-out score for families of at least 10 members, self-score otherwise). The cluster of each assigned representative was expanded to all member proteins, and every locus was retrieved as the partner gene ±6 kb (1,011 type I and 39 type II loci, and 10 loci for each type III gene). Annotated proteins and six-frame ORFs of at least 40 residues in each window were searched against a panel of 221 RT profiles (hmmsearch, $E \le 1 \times 10^{-5}$). Loci were then assigned to the first matching category of six. An ART RT was called adjacent when an RT that scored at least 207 bits on the ART profile was co-oriented with the partner and lay one to three genes upstream. A truncated ART-type RT was such an RT that scored below the threshold and abutted the contig end. The remaining categories were another ART-type hit, another RT, no RT (no hit and at least 1,800 nt of contig upstream of the partner), and not assessable (no hit and less than 1,800 nt upstream).

RT-partner co-folding and partner models. Type I RT-partner pairs from five environmental loci and one shuffled non-cognate pair were co-folded with AlphaFold2-multimer v3 (98) in ColabFold 1.6.1 on locus-paired MSAs (3 models per pair, 3 recycles, 1 seed, no templates). Each MSA opened with a paired block. To build this block, the RT and partner sequences of type I loci whose partner starts within 60 nt of the RT stop were aligned to the query pair (MAFFT --auto) and joined by locus. Unpaired blocks of MarsHill RT homologs and of type I partners followed (MAFFT --add --keeplength), and no server alignment was used. In the shuffled pair, the paired rows joined RTs and partners of different loci. Type II pairs of phages SA1, MarsHill, LY01, and PB50 and of two metagenomic loci were co-folded with two engines. AlphaFold2-multimer v3 was run in ColabFold 1.6.1 (5 models × 3 seeds, 3 recycles, no templates), and Boltz-2 v2.2.1 (99) was run with 3 recycling steps and 5 diffusion samples × 2 seeds. One type I pair served as a positive control, and the SA1 RT with a type I partner served as a negative control. Both engines received the same MSAs. These MSAs combined 1) a paired block of same-locus RT and partner sequences from the other loci of that partner type, 2) unpaired ART family RT and partner sequences, and 3) ColabFold MMseqs2 server hits (UniRef30 2302, environmental database 202108). The interface confidence score (ipTM) is the value that each engine reported, which was summarized per pair and engine as the median over all models. The type I partner and complex displayed in the figures are from the top-ranked AlphaFold2-multimer model of the ART_57 pair, and the type II partner and complex are from a Boltz-2 model of the SA1 pair. The type III partner is the most confident Boltz-2 model of a co-fold of the two type III genes of one locus. Models

were drawn with residues of pLDDT below 50 hidden, and for comparison with solved RT complexes the RT of each co-fold was superposed on the RT of PDB 8UBE.

Comparison of the type I partner with a GNAT acetyltransferase. The ART_57 type I partner model was compared with *Tetrahymena* GCN5 (tGCN5), a GCN5-related N-acetyltransferase (GNAT), bound to acetyl-CoA (PDB 1QSR, chain A). GNAT-like regions were located by TM-align of tGCN5 against the whole chain and then against the chain with the first hit removed. The first hit defined the GNAT-like domain (residues 212 to 373), and the remaining C-terminal residues (374 to 464) were analyzed as a second region. Each region was superposed on tGCN5 with TM-align (TM-scores normalized by tGCN5), and the acetyl-CoA of 1QSR was kept at its crystallographic position. For each region, partner heavy atoms within 2.0 Å of acetyl-CoA were counted, and residues with any atom within 4.0 Å were recorded with their median pLDDT. The residues that aligned to the P-loop of tGCN5 (residues 133 to 138) were identified, and the Cα-Cα distance between strands β4 and β5 at their open end was measured. Strands were assigned on 1QSR with DSSP (100).

N-terminal region and partner divergence. The N-terminal regions of the 95 ART RTs were aligned pairwise (global, BLOSUM62, gap open −10, extension −0.5). Identity was calculated as the number of identical aligned residues over the length of the longer sequence, and RTs were ordered by average-linkage clustering of 1 − identity. Partner divergence was related to RT divergence on a subset of ART tips of the RT phylogeny (30 RTs). These were the tips whose downstream gene lay on the same strand and was a type I, II, or III partner, and whose N-terminal region was at least 100 residues. For all 379 pairs of these RTs that share a partner type, the N-terminal regions, RT domains, and partners were aligned, and their identities were calculated in the same way. Spearman correlations and least-squares lines are shown, and no P value is given because pairs that share an RT are not independent.

Replicate campaigns. The campaign was launched ten more times with the same harness and brief. All task records (n = 3,084), reports, and session transcripts (n = 5,632) of the replicates were searched for identifiers from the original campaign. These were the identifiers of the RTs (n = 130) and contigs (n = 171) that the agents of the original campaign had assigned to the RT family, and the names and accessions of cultured phages that encode such RTs. Sessions that named one of these contigs (n = 11) were parsed event by event for tool results that contained a contiguous DNA string of at least 200 nt and for remarks on repeats. An ART locus on a contig outside this identifier set would not be found by this search.

Fixed-input benchmark. Each attempt was an independent run of one model that received the same task statement and one of five levels of input. The task statement asked the model to characterize the RT and to ”predict its function: report every feature you judge important, the evidence for it, how confident you are, and what experiment would test your main prediction.” It stated that the loci had been drawn from a large sequence collection because each encodes a member of one related group of RTs, and that nothing else was known. Each attempt ended with a report and a structured submission of findings, in which each finding consisted of a claim, its evidence, and a confidence between 0 and 1. The inputs were 96 ART loci (1,093 to 26,554 bp) with anonymized identifiers, gene calls, and Pfam matches, and 1,131 predicted protein structures. At level 1 the prompt contained two RT protein sequences, and at level 2 it contained the two corresponding loci as text. Neither level provided tools. At level 3 no sequence was in the prompt, and all 96 loci were supplied as files in an offline environment with gene calling, HMMER, BLAST+, MMseqs2, MAFFT, FastTree, ViennaRNA, and Python. Level 4 added a predicted structure for

every open reading frame, Foldseek with a local reference structure database, PyMOL, and US-align (101). Level 5 added literature search, web access, and the installation of software. The seven Claude models that are named in the Results were each run for 100 attempts at every level (3,500 attempts). Each attempt was run in isolation in Claude Code with a maximum output budget of 1 million tokens.

Benchmark grading. Submissions were graded on the ten binary claims that are listed in Fig. 4B, which the authors selected as characteristic of the ART family. Three claims concerned the 5′ non-coding tract, two concerned the RT, and five concerned the partner. Grading was performed by a language-model judge (Mythos 5) that was presented only with the submission and the rubric. A programmatic keyword screen identified the findings that were relevant to each claim for the judge and did not contribute to the grade. The judge graded each claim independently. A claim was counted only when the submission asserted it as a conclusion and not as one of several hedged possibilities, and no credit was given for over-claims that were specified in the rubric. The score of an attempt was the number of asserted claims (0 to 10). An attempt was counted as having recognized the repeat array when the judge graded the repeat-array claim as asserted.

Genomic language models. The 2,900-nt flank of locus L0050, as returned to the worker in task t0062, was evaluated with gLM2 (50) and Evo 2 (49). The gLM2 model (gLM2_650M) received the whole flank as one input. Each nucleotide was masked in turn, and the probability of the true base was taken from a softmax over the four nucleotides at the masked position. Evo 2 (7B model) was accessed through the Evo Designer web interface, and the probability of each base given the preceding sequence was taken from the returned logits, which were renormalized over A, C, G, and T. Values are the natural-log probability of the true base minus ln 0.25, so that zero corresponds to a uniform guess among the four bases. As a control, the nucleotides of each of the 14 copies (16 nt each), and of one degenerate copy upstream of the first, were permuted in place with the rest of the flank unchanged, and both models were run again. Tracks are centered 16-nt running means. Aligned profiles are means over copies 2 to 14 in 3-nt bins from 24 nt before to 41 nt after the start of the repeat, and per-copy values are means over the 16 nt of a copy.

Mythos 5 internal signals during the discovery session. The transcript of the session of task t0062 was reconstructed from the session log, up to and including the reasoning block in which the model first reports the array, and passed once through a separately hosted copy of the same Claude Mythos 5 checkpoint (saved version of the model) that ran the session (Fig. 5, G to I). The worker's system prompt and tool definitions are not in the log and were left out (placeholder tool definitions were used). The model's activity at one layer (one of its successive processing stages) was decomposed with a sparse dictionary (102, 103): a set of recurring activity patterns learned without supervision from the model's own activity, each with an automatic description generated from the texts on which it is strongest. An internal signal is the strength of one pattern at one token (the model's unit of text). Only the 128 strongest signals are recorded at each token, so a zero means no strong response. The model reads text from left to right, and its activity at a token depends only on that token and the text before it, so a response at copy 3 cannot depend on later copies. Each nucleotide takes the value of the token that contains it, a copy's value is the mean over its 16 nucleotides, and a nucleotide counts as responding if its token has a recorded value.

Candidate signals were selected before flank L0050 (Fig. 1F) was examined: the 12 with the largest differences among those that responded to synthetic repeats, that is, repeats planted in computer-generated random sequences (to qualify, a signal had to be higher on the third and later copies than on a twin se-

quence in which every copy was shuffled, and higher than between the copies, in at least 9 of 10 sequence pairs). The two shown were chosen from these 12 after the flank had been examined: the signal labeled automatically as ”tokens repeating an earlier pattern/text verbatim” (repeat-signal 1), and the signal labeled as ”letters in repeated DNA/protein sequence motifs, and garbled text” (repeat-signal 2). Two controls support the text. With each of the 14 copies shuffled in place in the same transcript, repeat-signal 2 was 0.00 on all 14 copies, against per-copy means of 0.85 to 1.62 on copies 3 to 14 in the flank as the model retrieved it. On repeats planted in random strings of letters and digits (10 pairs), repeat-signal 1 was 0.48 on the fourth and later copies against 0.05 on the shuffled twins, and repeat-signal 2 was 0.75 against 0.01 (each higher in 10 of 10 pairs); by the same measure repeat-signal 2 was 0.69 against 0.00 on 10 further computer-generated DNA pairs that played no part in the selection (the two sets differ in more than the alphabet, so the two values are not compared with each other). Both signals therefore respond to repeats that are not DNA; neither is specific to DNA. Tracks, aligned profiles, and per-copy values are displayed as in Fig. 5, D to F.