

CLOTHO功能介绍

AI驱动的生物信息学研究与分析平台

产品能力说明 · 2026年7月

面向科研机构、生物医药团队、临床研究人员及具有基础生物学知识的研究者。

目录

1. CLOTHO是什么
2. CLOTHO可以帮助客户完成什么
3. 核心产品能力
4. 生物信息学与组学分析能力
5. 跨组学、临床与药物研究能力
6. 交付成果
7. 适用边界与科学原则
8. 典型应用场景

1. CLOTHO是什么

CLOTHO是一个AI驱动的生物信息学分析平台。研究者提供数据、实验设计与科学问题后，CLOTHO可以组织数据检查、质量控制、比对或定量、统计建模、结果解释、可视化和综合报告，形成可追溯的分析链路。

CLOTHO不是只回答问题的聊天机器人，也不是只调用固定脚本的流程平台。它支持三条对等路径：

1. 标准分析流程：对成熟、高频的分析任务使用经过组织的方法和交付结构。
2. 按需自定义分析：针对特殊实验设计、个性化统计问题和标准流程之后的深入分析。
3. 可复用流程搭建：把反复执行的分析固化为可重复运行的工作流。

它的目标是把研究者从大量工具选择、数据衔接、代码实现和结果整理工作中解放出来，同时保留科学决策、过程透明度和结果复核能力。

2. CLOTHO可以帮助客户完成什么

2.1 从原始数据到分析结论

- 检查输入文件、样本命名、分组、批次、缺失值和元数据匹配。
- 根据数据类型选择合适的质控、比对、定量和统计方法。
- 执行标准流程或针对研究问题编写自定义分析。
- 生成图、表、候选列表、中间结果与综合分析报告。
- 解释指标的生物学意义、证据强度和分析限制。

2.2 从科学问题到可执行方案

- 梳理研究假设、主要比较、混杂因素和验证路径。
- 在实验开始前进行样本量、统计功效、批次分配和多重检验规划。
- 判断当前数据能回答什么、不能回答什么，还缺少哪些信息。
- 把复杂任务拆成可确认、可追踪的分析阶段。

2.3 从分析结果到科研交付

- 生成论文或汇报可用的图表和结果表。
- 形成贯通“方法—结果—解释—结论—局限”的报告。
- 根据汇报目标和受众制作结构化PPT，组织封面、研究背景、方法、关键结果、讨论和结论。
- 制作报告或演示文稿封面，以及科研示意图、流程图、机制图和与主题匹配的配图。
- 支持报告逐段评论、AI修改和不同版本比较。
- 导出DOCX、完整Notebook和自选结果ZIP包。
- 将已有项目的输出复用为新项目输入，串联多个研究阶段。

2.4 文献与证据工作

- 检索、筛选并核实相关论文。
- 围绕指定主题撰写文献综述，按问题或证据主题组织内容，而不是简单罗列论文。

- 整理研究现状、关键证据、争议、证据局限与研究空白。
- 为结果解释、靶点优先级和适应症研究补充可核查证据。
- 无法核实的引用会明确说明，不以记忆代替验证。

3. 核心产品能力

3.1 项目化研究空间

每个科学问题可以建立独立项目，分别保存项目输入、分析输出、对话、报告和版本历史。项目支持搜索、重命名和状态提示，适合同时管理多个研究任务。

3.2 数据资产与跨项目复用

- 云数据库可保存个人数据，并提供只读共享数据入口。
- 支持单文件、多文件和整个文件夹上传。
- 数据可以链接到不同项目，减少重复存储。
- 其他项目的输出可直接链接为当前项目输入，适合串联分析。
- 链接保留单一数据来源；源内容更新后，下游可以读取最新内容。

3.3 对话式分析与文件引用

用户可以用自然语言描述研究问题，并在对话中引用输入或输出文件、上传附件或直接粘贴文件。CLOTHO可以在完成一轮工作后给出可点击的下一步建议；部署开放多个模型档位时，还可按回合选择模型档位。

3.4 报告原生的分析体验

分析报告不是最后才生成的静态附件，而是随分析推进更新的主要工作界面：

- 报告包含组织化文字、图、表和可选择显示的分析代码。
- 新增或修改内容可通过右侧改动标记快速定位。
- 用户可以在具体段落、表格或图表旁留下评论。
- 多条评论可以一次性提交给AI，触发报告修改和必要的重新计算。

3.5 对话回退、分支与报告比较

- 可以回到某条历史消息之前重新提问。
- 原来的后续对话保留为历史分支，可再次切回。
- 版本图展示对话分支、运行状态、时间和模型档位。
- 可以选择两个或更多版本，将对应报告并排比较。

3.6 多格式结果预览

CLOTHO可直接预览图片、PDF、HTML、Markdown、JSON、代码、日志、CSV/TSV/Parquet表格、VCF/BED/GFF/GTF等生物信息表格、FASTA序列以及PDB/mmCIF三维结构。大表格使用受控预览，并保留完整文件下载入口。

3.7 中英文与多端使用

- 界面支持中文和英文。

- 每个项目独立选择Agent回复和报告语言。
- 提供桌面端和移动端布局。

4. 生物信息学与组学分析能力

4.1 转录组与RNA

Bulk RNA-seq

- FASTQ质控、STAR比对、featureCounts计数或外部count matrix导入。
- DESeq2、edgeR、limma-voom差异表达，支持多因素、交互、配对和批次协变量。
- ORA/GSEA、GSVA/ssGSEA、WGCNA、免疫浸润、TF活性、时间序列和生存关联。

可变剪接

- rMATS-turbo与leafcutter互补检测SE、A5SS、A3SS、MXE、RI等事件。
- ΔPSI、异构体切换、差异转录本使用、NMD与结构域后果。
- 新剪接连接、隐性外显子、微外显子、exon和sashimi图。

Small RNA/miRNA-seq

- 接头和长度质控、miRNA定量与差异、isomiR、A-to-I编辑和新miRNA预测。
- 靶基因证据整合、miRNA—mRNA负相关网络及其他小RNA分类。

Ribo-seq/翻译组

- RPF预处理、P-site校准、三核苷酸周期性和metagene QC。
- ORF/uORF、新翻译区域、翻译效率、差异翻译、核糖体停滞和密码子占用。

长读长转录本/Iso-Seq

- PacBio Iso-Seq、ONT cDNA与direct RNA全长转录本发现。
- SQANTI3质量分类、已知与新异构体、差异异构体使用、异构体切换和NMD。

罕见病RNA离群分析

- 单患者或病例组相对对照队列的异常剪接、异常表达和单等位表达。
- FRASER、OUTRIDER、隐性外显子、深内含子异常及组织匹配证据。

4.2 单细胞与空间组学

单细胞转录组

- 双胞胎、环境RNA、批次整合和整合质量诊断。
- 聚类、marker、自动与参考注释、pseudobulk差异、细胞组成和亚群深挖。
- PAGA/DPT轨迹、RNA velocity、CellRank命运概率、pySCENIC和inferCNV。

细胞通讯

- CellChat、CellPhoneDB、LIANA等配体—受体推断。
- 信号通路聚合、发送者/接收者、网络中心性和条件比较。

scATAC/Multiome

- fragments导入、每细胞QC、doublet、谱嵌入和pseudobulk峰检测。
- gene activity、chromVAR、motif、峰—基因连接及RNA + ATAC整合。

空间转录组与空间蛋白组

- 空间域、空间高变基因、Moran统计、解卷积和空间通讯。
- CODEX/IMC/MIBI表型分型、空间邻域、共现、niche和组织互作。

Perturb-seq与谱系追踪

- sgRNA→细胞匹配、扰动QC、每扰动差异和机制富集。
- CRISPR barcode、慢病毒barcode或mtDNA标记的谱系树、克隆扩张和命运耦合。

单细胞药物靶点发现

- 按细胞类型构建疾病信号，整合通路、细胞通讯、人类遗传学和可成药性证据。

4.3 表观基因组与染色质

- ATAC-seq：峰检测、TSS enrichment、FRiP、差异可及性、motif和TOBIAS足迹。
- ChIP-seq：窄峰/宽峰、IDR、DiffBind、motif、超级增强子和spike-in校准。
- CUT&Tag/CUT&RUN：低起始量质控、SEACR/MACS3、IgG和E. coli spike-in。
- 甲基化：WGBS/RRBS/beta matrix、DMC/DMR、注释、年龄和细胞组成。
- Hi-C/三维基因组：contact matrix、compartment、TAD、loop、APA、P(s)和差异互作。
- ChIP-Atlas公共调控挖掘：TF富集、靶基因、上游调控因子和表达方向一致性。

4.4 基因组、变异与组装

- WGS/WES胚系变异：BWA-MEM2、GATK HaplotypeCaller、联合分型、注释和候选筛选。
- 体细胞变异：Mutect2、tumor-normal/tumor-only、PoN、污染、方向偏倚、TMB和驱动基因。
- 突变签名：COSMIC SBS/DBS/ID/CN签名提取或refit，分析APOBEC、HRD、MMR等过程。
- CNV：CNVkit分段、整数拷贝数、焦点/广域事件、FGA、复发CNV和GISTIC风格分析。
- 胚系CNV解释：ACMG/ClinGen剂量敏感性、频率、de novo和家系证据分级。
- 基因组注释：原核/真核结构注释、重复屏蔽、功能和ncRNA注释、BUSCO/QUAST。
- 从头组装：Illumina、ONT、PacBio及hybrid组装、抛光、Hi-C支架化和污染检查。
- 长读长DNA：SV、SNV/indel、从头组装及5mC/6mA甲基化。

4.5 群体与统计遗传

- GWAS质控、PLINK2/REGENIE关联、Manhattan和QQ图。
- LD clumping、SuSiE fine-mapping、credible set和PIP。
- 共定位、LDSC遗传力与遗传相关、MTAG、TWAS和meta-analysis。
- 孟德尔随机化及敏感性分析。
- PRS构建、验证、风险百分位、绝对风险和跨族群可迁移性。

4.6 蛋白质组与代谢组

- 蛋白质组：MaxQuant/DIA-NN导入、归一化、缺失处理、差异蛋白、PPI和biomarker。
- PTM蛋白组：磷酸化、乙酰化、泛素化、糖基化位点差异、motif和KSEA激酶活性。
- 代谢组：特征表QC、PQN归一化、差异代谢物、通路和表型关联。
- 代谢物鉴定：加合物/同位素归组、精确质量、保留时间、MS/MS谱库和计算碎裂，输出MSI置信等级。

4.7 微生物与宏基因组

- 16S/Amplicon的ASV/OTU、 α/β 多样性、PERMANOVA、差异丰度、功能预测和共现网络。
- Shotgun宏基因组的宿主去除、Kraken2/Bracken、HUMAnN3、菌株、AMR、组装和MAG分箱。

4.8 免疫组库、免疫信息学与CRISPR

- TCR/BCR clonotype、克隆扩张、多样性、V/J偏好、CDR3、样本重叠、SHM和克隆谱系。
- HLA分型、MHC-I/II结合、新抗原、T/B细胞表位、免疫原性及个体化疫苗候选。
- pooled CRISPR screen的MAGeCK、BAGEL2、drugZ、基因必需性和通路解释。

4.9 多组学、机器学习与时间序列

- MOFA/MOFA+、跨组学相关和联合因子。
- DIABLO监督式多组学、Similarity Network Fusion和患者分层。
- PCA/UMAP、层次聚类、k-means、HDBSCAN、GMM及稳定性评估。
- LASSO/Elastic Net biomarker面板、嵌套交叉验证、AUC和校准。
- CoxNet、随机生存森林、时间依赖AUC、C-index和Brier score。
- 纵向患者轨迹、疾病进展速度和结局验证。
- 昼夜节律、周期发现、相位/振幅及动态调控网络。

5. 跨组学、临床与药物研究能力

5.1 通用高级方法

- 功能富集、leading-edge、富集网络和GO语义去冗余。
- 批次校正、协变量、配对、混合设计和多重检验。
- 生存、纵向、分类、回归、风险预测和模型解释。
- WGCNA、PPI、GRN、社区、中心性和差异网络。
- 自定义统计检验、相关分析和专属可视化。

5.2 临床与药物研究

- 生存、竞争风险、诊断/预后标志物和外部验证。
- EHR/表型组、PheWAS、共病网络和患者分群。
- 遗传风险、ACMG解释思路、CNV分级和肿瘤免疫指标。
- 靶点发现与优先级、适应症档案、流行病学、标准治疗和临床试验证据。

- 按需支持化学信息学、虚拟筛选、ADMET/QSAR、蛋白结构预测与设计、分子动力学和CRISPR编辑设计。

5.3 其他按需方法支持

CLOTHO的方法参考范围还覆盖CLIP-seq、m6A/表观转录组、RNA结构、基因型填补与单倍型定相、引物设计、系统发育、比较基因组、eDNA、病原暴发追踪、流式细胞术、液体活检、药物基因组学、临床试验统计、公共数据集发现以及Nextflow/Snakemake等流程工程。

这些能力不全是一键流程，实际执行取决于输入数据、计算资源、软件许可、参考数据库和外部服务可用性。

6. 交付成果

CLOTHO通常交付：

1. 分步结果目录：保留QC、中间结果和最终结果。
2. 分析图与科研配图：交付统计图、组合图、流程图、机制示意图及其他主题配图，可按论文或汇报用途调整风格。
3. 表：差异结果、统计指标、候选排序和质量报告。
4. 分析报告与文献综述：说明方法、结果、结论和限制，或围绕指定主题综合经过核实的文献证据。
5. DOCX报告：适合传阅、批注和二次编辑。
6. PPT演示文稿：按汇报目标组织封面、内容结构、关键图表、结论和下一步，交付可继续编辑的PPTX文件。
7. 完整Notebook：包含可交付的分析代码、结果和图表。
8. ZIP交付包：按用户选择打包输出文件和报告格式。

7. 适用边界与科学原则

- 关键分组、样本对应、物种、参考版本或临床结局缺失时，应先补充信息。
- 输入冲突、维度不一致、严重混杂或方法不适用时，应明确停下说明。
- 不以降低阈值换取“显著结果”；零显著结果同样可能是有效结论。
- 不静默丢弃样本、填补关键值或替换成科学含义不同的方法。
- 文献和外部证据需要核实；无法确认时应说明不确定性。
- 生成的封面和科研配图需要由用户复核科学含义、标签准确性和最终使用场景。
- 需要联网、外部数据库、许可软件或大量计算资源时，应提前说明条件。
- 临床相关结果用于科研与决策辅助，不构成诊断或治疗建议。

8. 典型应用场景

场景一：从RNA-seq到论文图表

上传FASTQ或count matrix和样本表，完成质控、差异表达、富集、WGCNA及报告；研究者在报告中评论图表和结论，AI根据意见修改后导出DOCX与结果包。

场景二：单细胞疾病机制与靶点发现

完成单细胞QC、注释、pseudobulk差异、轨迹和细胞通讯，再整合人类遗传学与可成药性证据，对候选靶点排序。

场景三：多个项目串联

第一个项目完成原始组学处理，第二个项目链接其输出做富集和网络分析，第三个项目继续构建biomarker或临床模型，避免反复复制大文件。

场景四：分析方案纠偏

发现某轮分析的问题后回退到对应消息，保留原分支并提出新的设计；完成后把两个版本报告并排比较。

场景五：临床与适应症研究

将组学、临床结局、文献、流行病学和试验证据组织为生存模型、风险分层、靶点优先级或适应症档案。

场景六：从文献综述到汇报材料

围绕指定课题检索并核实文献，综合领域共识、争议和研究空白；再根据汇报对象提炼叙事结构，制作综述文稿、PPT封面、内容页和配套科研图示。