BLAST自建库教程：序列清单、makeblastdb索引与抽样检索验收

完整操作教程 · 参考固定 · 往返恢复验收

自有核酸序列规模增大后，可把FASTA整理成受控的本地BLAST库。建库之前先核序列类型、唯一ID和授权来源；建库之后再做抽样检索与完整性检查。

运行边界：合成FASTA只验证命令结构；正式使用前核实序列来源、授权、版本和BLAST+参数。

先确认自建库与下载公共库的区别

如果目标序列已经属于NCBI等机构维护的公共数据库，通常应评估下载其预格式化数据库，而不必重复从原始FASTA构建。自建库适用于项目自己的序列、实验室整理的参考集或需要固定版本的子集。先确认检索对象是核酸还是蛋白质、查询使用blastn还是blastp、序列是否有共享和再分发许可。数据库构建本身不验证序列来源、注释正确性或物种归属；这些质量责任仍由数据整理者承担。

输入文件应记录来源、版本/发布日期、筛选条件、序列数、总碱基数、文件大小和SHA256。多个FASTA合并前，先判断重复序列是保留、去重还是按来源分别建库；不能只凭序列完全相同就删除，因为不同accession和样本可能对应不同的追溯信息。合并后保留原始文件和合并脚本，避免后续无法解释库中某条记录来自哪里。大型库往往包含多个索引文件，磁盘规划不能只看原始FASTA大小，还要预留临时空间、版本并存空间和备份。

整理FASTA标识符，避免可检索性和追溯性问题

当需要按序列ID取回命中记录时，FASTA定义行应有唯一、稳定且无空格的标识符，并按NCBI规则启用`-parse_seqids`。NCBI说明标识符应紧跟在“>”后，空格会结束ID；竖线字符在NCBI ID解析中有特殊规则，不应随意拼接。建议把原始标题保存在单独的TSV映射表中：规范ID、原始标题、来源文件、来源记录ID、序列长度、版本和筛选标记。这样既便于BLAST输出引用规范ID，也可以回溯到原始记录。

建库前扫描空序列、非法字符、重复ID、非ACGTN字符比例异常和过短序列。核酸序列含IUPAC模糊碱基可能是合理的，不要一律删除；应依照研究目标和软件接受范围决定。不同来源可能对软掩蔽、大小写和低复杂度序列采用不同处理，建库参数须写入manifest。若启用了分类学taxid，还需准备与ID对应的映射文件及适用的taxonomy数据；不能把物种名称直接当成taxid。

检查项	建议证据	发现问题时
序列类型与来源	来源清单、许可、日期	先确认可使用范围
FASTA ID	唯一ID与标题映射表	保留原始ID并生成稳定规范ID
建库参数	BLAST+版本和完整命令	在新版本目录重建，不覆盖旧库
检索抽查	已知序列小查询及命中ID	检查索引、序列类型和解析参数

分版本构建，不覆盖可用数据库

先在新目录构建候选版本，再通过符号链接或配置项切换当前库。数据库目录名宜包含日期或版本号，并在目录外保留manifest。构建过程中记录makeblastdb完整命令、stdout/stderr、退出状态、输入摘要、数据库类型、序列数和索引文件清单。不要在BLAST任务仍运行时覆盖原库文件，否则运行中的进程可能读到不完整或混合版本。切换前停掉或等待旧任务，切换后再用少量已知查询验证。

下方为小型演示命令。示例序列是合成序列，不代表任何物种。正式运行时，把输入换成已获授权的核酸FASTA，使用数据库专用目录并确认空间配额。`-blastdb_version 5`是当前示例参数，实际版本选项需与安装的BLAST+版本及下游工具兼容；先查看本机`makeblastdb -help`，不要盲目照抄在线命令。

```
set -euo pipefail
blastn -version
mkdir -p demo-db-v1
makeblastdb -in synthetic.fna -dbtype nucl -parse_seqids   -blastdb_version 5 -title
"synthetic demo"   -out demo-db-v1/sequences
blastdbcheck -db demo-db-v1/sequences
blastdbcmd -db demo-db-v1/sequences -entry demo_seq_1 -out recovered.fna
blastn -query synthetic-query.fna -db demo-db-v1/sequences   -outfmt '6 qseqid sseqid
pident length evalue bitscore'   -max_target_seqs 5 -out demo-db-v1/sample.tsv
sha256sum synthetic.fna synthetic-query.fna demo-db-v1/* > demo-db-v1/SHA256SUMS
```

输出文件清单应覆盖所有数据库卷和元数据文件。不要用shell通配符把`SHA256SUMS`自己也加入哈希后再覆盖生成；更稳妥的做法是对输入摘要和数据库清单分别生成manifest。示例中hash命令仅演示收集文件摘要，不适用于目录本身；遇到子目录、软链接和权限文件时，要明确递归规则。数据库版本切换后，保留上一版直到验证完成且旧作业结束。

用已知查询和ID取回验证建库

验证分为“能打开”“能取回”“能检索”三步。`blastdbcheck`可以检查数据库结构；`blastdbcmd`按ID取回序列，能发现ID解析、输入映射和数据库卷问题；最后用一条已知查询运行blastn，检查输出字段、命中序列ID和预期方向。若查询来自建库中的同一条序列，应注意自比对几乎必然命中自身，只能验证管线工作，不能评价检索灵敏度或生物学有效性。还要准备一条不在库中的阴性查询，观察它是否意外出现低复杂度命中。

抽样结果要回溯到FASTA原记录和来源表。检查输出中的sseqid是否可在manifest中找到，使用blastdbcmd取回后比对序列摘要是否一致。对多卷数据库确认每个卷都在清单中，不能只检查第一个文件。若使用掩蔽、taxid或alias文件，这些都属于数据库版本的一部分，应和索引一并归档。

资源规划：分别估计CPU、内存、临时盘与长期盘

建库时间受序列总碱基数、重复程度、参数、CPU、磁盘IO和并发影响，不能只按序列条数线性外推。先用代表性子集测量建库耗时、最大RSS、临时空间和成库大小，再按全量输入估算并留出不确定余量。批量blastn查询与一次建库是不同阶段：建库通常是一次性成本，查询可能反复占用CPU和IO。多个课题组共享数据库时，可以把只读数据库放在经管理员批准的公共目录，但缓存和临时输出应按项目隔离。

如无法在本机建库，先准备软件版本、输入规模、数据库类型、预计查询量、并发数和可用存储信息，再讨论计算环境。租用CPU服务器能提供运行环境与资源，但并不自动包含序列注释、数据库授权确认或结果解释服务。涉及敏感样本或受限序列时，先确认数据是否允许离开单位环境。

解决坏输入和重复标识，不靠建库阶段报错猜测

FASTA的第一个字符应是序列标题行的“>”，但文件头可能有BOM、空白行或Windows换行。建议先用一个轻量解析程序逐条统计ID、序列长度和非法字符，再生成规范化文件。对于重复ID，可建立稳定的新ID，例如来源前缀加原始编号，同时把原始标题和原始文件摘要映射到TSV；不要无提示地给重复项加计数后缀，因为这可能让原记录对应关系丢失。对于空序列、极短序列和含大量N的记录，输出待人工处理清单，不要静默删除。

如果输入由多个实验批次合并，每条序列应保留来源批次、样本或构建版本等追溯字段。将元数据与序列文件解耦也能让数据库更新更透明：新版本可比较新增、删除和修改的ID。若需taxonomy功能，必须使用来源一致的taxid映射和taxonomy files。NCBI的identifier解析有保留格式，任意使用竖线可能导致误解析；对外部accession先按其官方格式处理，别把显示名称直接拼成NCBI样式ID。

建库失败时先保存命令、退出码、stderr、BLAST+版本和目录剩余空间。常见原因包括输入文件格式、重复ID、核酸/蛋白类型选错、输出路径不可写、索引文件系统限制和临时空间不足。不要在失败后只删掉某几个疑似索引文件再续跑；在全新候选目录中重建，防止目录同时包含两次不同构建的文件。大型输入可先抽一小部分做验证，但要确保抽样保留不同

长度、字符组成和来源类别，不能只挑最简单的记录。

检索验收要覆盖ID、方向和阴性情况

准备至少三种小查询：库内已知序列用于验证可命中，库内序列的片段用于验证局部比对，确认不在库中的序列用于观察背景命中。每条查询保存SHA256、预期方向和判断标准。BLAST输出中核对qseqid、sseqid、pident、alignment length、E-value和bitscore，并用blastdbcmd回取命中序列进行ID映射。库内完全相同序列的自命中不能证明参数适合真实任务；还需按研究目标确定查询长度、阈值、低复杂度过滤、链方向和最大命中数。

如果目标是近缘菌株或高相似序列，重复区域会产生多个强命中；若目标是短片段，默认参数可能不适合。应在小样本上对照已知阳性、阴性和人工构造的边界例子，记录参数变化如何影响结果，再决定批量检索设置。数据库搜索结果是序列相似性证据，不能单独推出功能、致病性或实验验证结论。输出报告要保留搜索软件版本、数据库版本与参数，后续解释还要结合注释来源和实验背景。

如果同时维护不同物种或核酸与蛋白库，目录结构和配置文件要显式区分，不要让同名数据库前缀指向错误索引。批量脚本启动时应检查目标库的类型、manifest版本和完整文件清单。别在公共共享目录直接覆盖`current`链接；先建新版本、验证、停止依赖旧库的任务后再切换。发生回滚时保留失败版本日志以便复现，而不是把目录直接删除。

最后将输入清单、规范ID映射、构建命令、BLAST版本、索引文件清单、抽样测试、资源记录和维护责任人一起交接。数据库更新需要重新构建并生成新版本，不要把新增序列直接拼到旧索引旁边。若提供给多个课题组使用，还需说明来源许可、适用物种、版本日期、更新策略、存储位置和联系责任人，避免使用者将自建库当作权威公共参考数据库。

可维护的数据库还需要版本切换和更新策略

对经常更新的内部序列库，定义版本发布条件：输入来源清单齐全、ID冲突解决、构建成功、完整性检查通过、抽样查询符合预期、更新差异可解释。发布目录最好只读，构建进程在单独的暂存目录写入，验收后再由管理员更新配置指向新版本。回滚时保留旧目录和版本说明，不要删除旧索引后才发现新库的taxid映射不兼容。每次更新用差异清单说明新增和移除的记录数、序列总长度变化及规则变化。

查询程序也要将数据库版本写进结果表头或伴随manifest。相同的query和参数，如果库更新了，命中集合变化并不一定意味着程序错误。若分析结论依赖某个特定版本，应在方法和补充材料中准确记录数据库构建日期、来源与参数。公共数据库的官方预格式化库也会定期更新；下载前核对目标版本、体积、校验和以及维护方提供的更新说明。超大公共库下载还涉及网络、磁盘、许可证和维护周期，不能按自建小库的成本估算。

为保障并发检索，测量代表性query的CPU、RSS、磁盘读吞吐和并发影响。多个任务同时扫描同一大库时，瓶颈可能从CPU转为IO；简单增加CPU核数未必缩短时间。管理员可以测试只读共享库的缓存命中、节点本地缓存和查询并发限制。记录测试机器、BLAST版本、库大小和命令，避免将单机小样本耗时直接外推到全量批处理。若数据库太大，应考虑分库、按分类筛选或使用已维护参考库，但筛选规则会改变可搜索空间，需要在方法中明示。

配套练习包

本PDF对应的ZIP包含本文代码、主题模板或合成数据、README和SHA256校验文件。运行前核对包内清单，并按文章说明在独立目录验证。

参考资料

NCBI BLAST 命令行手册：本地序列建库 · 来源资料

NCBI：从自有序列构建 BLAST 数据库 · 来源资料

NCBI：BLAST 数据库完整性检查 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。