Beagle 5.5填充实操：核对参考面板、按染色体运行与验收

确认目标和参考面板位点完全协调后，再分染色体运行 Beagle 5.5；记录Java堆、图谱、日志，并检查样本、位点覆盖和填充指标。

Beagle 5.5 可对基因型分相，并在参考面板协助下填补目标样本缺失或未分型的标记。它不是基因型似然输入工具；如果手头是低深度测序的 GL/PL，而非 VCF 中的 GT，需要先确认方法是否匹配。下面以已质控的目标 VCF、已分相参考面板和遗传图谱为例，逐条染色体运行，并保留日志与验证记录。

先核对参考面板和目标位点

Beagle 要求参考面板和目标文件中对应标记的 CHROM、POS、REF、ALT 完全一致，且顺序一致；参考基因型需已分相且无缺失。只把染色体名改成带 chr 前缀、或单纯排序，并不能修复参考/替代等位基因方向不一致。先统一参考基因组版本、染色体命名与变异规范化；需要时用 conform-gt 协调等位基因，再检查日志中的保留与排除数量。若提供参考，目标中不在参考面板里的位点可能不会出现在输出中。

项目	运行前检查	常见误区
目标 VCF	含 GT、样本顺序正确、BGZF 压缩且可索引	把 PL/GL 当作 Beagle 5.5 输入
参考面板	相同组装、匹配 CHROM/POS/REF/ALT 与顺序，基因型已分相且非缺失	只按位置相同就当作等位基因匹配
遗传图谱	组装版本及染色体对应，PLINK map 使用 cM	无图谱时默认按 1 cM/Mb，不一定适合局部重组率
资源	Java 可用堆、作业线程和输出磁盘配额	把 -Xmx 设成主机全部内存

每条染色体独立运行

以下假设染色体 1、2 的目标、参考和 PLINK 格式图谱已经准备好，路径需按实际数据替换。8g、2 线程和 seed 仅为演示值。Beagle 用 -Xmx 设置 Java 最大堆，仍需给 JVM、系统和其他进程留出余量；out 是文件名前缀，不是目录。先用小而具代表性的染色体验证输入与资源，再扩大运行。

```
set -euo pipefail
jar=/opt/beagle/beagle.27Feb25.75f.jar
test -s "$jar"
mkdir beagle-run1
java -version > beagle-run1/java.txt 2>&1
sha256sum "$jar" > beagle-run1/jar.sha256
for chr in 1 2; do
  gt="/data/target.chr${chr}.vcf.gz"
  ref="/ref/panel/ref.chr${chr}.vcf.gz"
  map="/ref/maps/chr${chr}.map"
  for f in "$gt" "$ref" "$map"; do test -s "$f"; done
  d="beagle-run1/chr${chr}"; mkdir "$d"
  bcftools query -l "$gt" > "$d/target.samples"
  /usr/bin/time -v -o "$d/resources.txt" \
    java -Xmx8g -jar "$jar" gt="$gt" ref="$ref" \
    map="$map" out="$d/chr${chr}" nthreads=2 seed=2026
  test -s "$d/chr${chr}.vcf.gz" && test -s "$d/chr${chr}.log"
  bcftools quickcheck -v "$d/chr${chr}.vcf.gz"
  bcftools query -l "$d/chr${chr}.vcf.gz" > "$d/output.samples"
  diff -u "$d/target.samples" "$d/output.samples"
  bcftools index -t "$d/chr${chr}.vcf.gz"
done
```

检查输出、覆盖和资源记录

逐条查看 Beagle 日志中的软件版本、参数、线程、计算时间及警告；确认输出样本清单与目标一致，并用 bcftools quickcheck 和索引确认 VCF 文件结构可读。再比较输入、参考和输出的变异数、坐标范围及交集，重点检查目标标记是否因未匹配参考而被排除。若要合并染色体文件，先固定染色体自然顺序，确认每个文件的样本顺序一致且区间不重叠，再用 bcftools concat；不要直接依赖字符排序把 chr10 放在 chr2 前面。

Beagle 输出的 GT 已填补且分相；参考面板填充时 INFO 中可包含 DR2 和 AF。DR2 是估计的信息指标，不是已知真实基因型上的准确率。若分析依赖剂量，保留并检查 FORMAT/DS，不要只导出硬调用 GT。条件允许时，使用独立的高质量真实基因型做遮蔽验证，按等位基因频率、区域和样本报告准确率；避免让目标样本或近亲泄漏进验证面板。

每条染色体记录输入哈希、目标与参考样本数、变异数、软件/JAR 版本、遗传图谱版本、Java 堆、线程、运行时间、峰值内存、日志和输出路径。Beagle 手册指出缩短 window 或 window-markers 可减少内存，但窗口和迭代会影响计算与结果；不要为赶进度盲目改默认值。先测代表性染色体，再决定是否串行排队或在内存余量充足时有限并发。若研究输入是基因型似然、组装不匹配或参考不适用，应先解决方法与数据适配，单纯增加内存不会修复结果。

参考资料

Beagle 5.5 官方手册：输入、参数、输出与资源 官方文档
Beagle 5.5 官方版本、下载与示例脚本 官方文档
BCFtools 官方手册：VCF检查、索引与合并 官方文档