GLIMPSE2分块填充实操：生成区间、相位计算与连接验收

按GLIMPSE2流程生成带缓冲的区块、预拆参考面板、逐块填充并ligate；核对样本、重叠边界和资源后再扩大运行。

GLIMPSE2 面向低深度测序的基因型填充和相位推断，可从 BAM/CRAM 读取 reads，也可读取基因型似然。它将染色体切成带缓冲区的窗口，逐块计算后再连接相位。本文按 BAM 输入演示一个常染色体；若只有少量高置信硬调用，不能把它当作等价低深度 reads/似然输入。先确认样本、参考面板与基因组版本一致，再启动整批任务。

先准备参考和染色体区间

参考面板需有索引，遗传图谱应与组装和染色体命名对应。先按项目质控移除目标样本及不应参与填充的近亲，避免信息泄漏；确认参考中的单倍型质量、目标 BAM 的排序与索引，并校验 CRAM 所需的同版本 FASTA。GLIMPSE2_chunk 根据参考位点和遗传图谱生成区间表：每行包含块编号、染色体、含缓冲输入区间及无缓冲输出区间。缓冲使相邻块有重叠，后续 ligate 才能衔接相位。

阶段	输入/产物	关键核对
chunk	参考位点、遗传图谱、区间表	坐标命名与组装一致，保留原始行序
split_reference	每块二进制参考	同时传入 input-region 与 output-region
phase	每块参考、BAM 或 GL	一块一输出，记录失败码和资源
ligate	全部成功的同染色体块	确认相邻区间重叠，不混入旧结果

先跑一条染色体

下面示例使用 chr22、一个样本和单线程，路径需换成实际文件。参考 BCF 是有基因型的面板，chunk 工具可用参考或目标变异位点文件生成区间；地图提供 cM 位置。输出前缀应放在全新目录。确认块清单字段后再运行循环；不要把 `*.bcf` 的字母序当作任务清单，也不要把失败块从列表中悄悄删掉。

```
set -euo pipefail
bam=/data/S01.bam
ref=/ref/panel.chr22.bcf
map=/ref/chr22.b38.gmap.gz
mkdir glimpse-run1
out="$PWD/glimpse-run1"
mkdir "$out/ref"
samtools quickcheck -v "$bam"
samtools idxstats "$bam" > "$out/bam.idxstats.txt"
bcftools index -s "$ref" > "$out/ref.contigs.txt"
GLIMPSE2_chunk --input "$ref" --map "$map" --region chr22 \
  --sequential --output "$out/chunks.txt" --log "$out/chunk.log"
: > "$out/ligate.list"
while read -r id chr irg org rest; do
  region=${irg#*:}; start=${region%-*}; end=${region#*-}
  /usr/bin/time -v -o "$out/split-$id.time.txt" \
    GLIMPSE2_split_reference --reference "$ref" --map "$map" \
    --input-region "$irg" --output-region "$org" \
    --output "$out/ref/panel" --threads 1 \
    --log "$out/split-$id.log"
  panel="$out/ref/panel_${chr}_${start}_${end}.bin"
  test -s "$panel"
  result="$out/chunk-$id.bcf"
  /usr/bin/time -v -o "$out/phase-$id.time.txt" GLIMPSE2_phase \
    --bam-file "$bam" --ind-name S01 --reference "$panel" \
    --threads 1 --seed 2026 --output "$result" \
    --log "$out/phase-$id.log"
  test -s "$result"
  bcftools index "$result"
  printf '%s\n' "$result" >> "$out/ligate.list"
done < "$out/chunks.txt"
test -s "$out/ligate.list"
GLIMPSE2_ligate --input "$out/ligate.list" \
  --output "$out/chr22.bcf" --threads 1 --log "$out/ligate.log"
bcftools index "$out/chr22.bcf"
bcftools stats "$out/chr22.bcf" > "$out/chr22.stats.txt"
bcftools query -l "$out/chr22.bcf" > "$out/samples.txt"
```

区间清单每行的 input-region 含缓冲、output-region 不含缓冲；二进制参考文件按 input-region 命名，循环时应保留 chunk 工具输出的编号和坐标，不能自行重排或重新切块。上例只适用于常染色体二倍体；性染色体需按项目的倍性、伪常染色体区和 GLIMPSE 样本倍性文件规则另行处理，不能直接套用 chr22 参数。

连接后做完整性和准确性检查

检查每个 chunk 是否都有非空 BCF、索引和成功日志，再确认 ligate 清单只含本次目标染色体、每个块恰好一次。GLIMPSE2_ligate 会按位置排序输入块，但相邻块必须重叠才能对齐相位；日志若提示缺少重叠，应回看 chunk 清单和块输出，不能手工删除重复位点来掩盖问题。最终 BCF 用 bcftools quickcheck、索引、样本清单、位点数和坐标范围核验，并查看输出是否含预期的 GT、DS、GP 字段。

结构正确并不表示填充准确。条件允许时，用独立高质量基因型做遮蔽验证，按频率、区域和样本报告一致率；参考面板代表性不足或低覆盖区域的结果应单独标记。记录软件版本与二进制校验、参考和图谱版本、BAM/CRAM 清单、chunk 表、每块日志、峰值内存、运行时间及失败原因。按最慢区块和实测峰值估算批次资源；并发数受磁盘、内存和计算配额限制，不能按平均耗时线性外推。单条染色体验收后再扩展。

参考资料

GLIMPSE2 官方入门流程：参考准备、分块、填充和 ligate 官方文档
GLIMPSE2 chunk：输入位点与缓冲区间 官方文档
GLIMPSE2 split_reference：区间和二进制面板 官方文档
GLIMPSE2 phase：BAM、GL、线程与输出字段 官方文档
GLIMPSE2 ligate：重叠区块连接规则 官方文档