vg Giraffe短读长比对：索引配套与样本验收

核对GBZ、minimizer、zipcodes和dist是否来自同一图索引发布，再用一个双端样本运行Giraffe并检查GAM与资源。

vg Giraffe部署的重点是索引配套：单独看到一个GBZ文件不足以开始比对。短读长工作流通常需要同一图发布的GBZ、short-read minimizer、zipcodes和dist；本例不适用于长读长索引与参数；跨图混用同名文件会造成错误或隐蔽的比对差异。本文以配对FASTQ为例，先登记索引来源，再跑一份代表样本。正式分析前还应确认图的参考组装和变异面板适用于研究样本。若自行建图，保存所用FASTA与VCF清单及构建命令；相位信息和样本面板会影响图路径，未相位输入与相位图的解释能力不同。不要只凭图的文件名或映射率判断其适合特定课题。

索引文件	用途	验收重点
giraffe.gbz	图序列与单倍型路径	同一图构建
shortread.withzip.min	短读长种子查找	与图和模式一致
shortread.zipcodes	种子位置提示	对应minimizer
dist	图距离信息	与GBZ节点匹配

登记索引和FASTQ

若用官方vg autoindex --workflow giraffe生成索引，保留完整命令、输入FASTA/VCF指纹、vg版本和输出清单；从外部取得索引则记录发布页及校验值。把四类文件放入只读的版本目录，不能只按文件名判断它们属于同一次构建。不同vg版本可能采用不同索引布局；如接收旧版分离的GBWT/图文件，应查阅对应版本教程，不要把旧、新文件拼接使用。R1和R2需来自同一样本且配对关系一致；交错FASTQ应使用对应选项，不照搬本例。

先跑一个双端样本

以下环境变量需替换为实际文件，THREADS填可用的正整数，OUT必须是新目录。显式传入所有索引，避免程序自动猜测文件或临时尝试建索引。GAM是图坐标比对结果，不是线性参考BAM。

```
set -euo pipefail
: "${GBZ:?■■giraffe.gbz}"
: "${MIN:?■■shortread.withzip.min}"
: "${ZIP:?■■■■shortread.zipcodes}"
: "${DIST:?■■■■dist}"
: "${R1:?■■FASTQ■■■}"
: "${R2:?■■FASTQ■■■}"
: "${THREADS:?■■■■■■■■■}"
: "${OUT:?■■■■■■■■}"
case "$THREADS" in ""|*[!0-9]*) exit 2;; esac
test "$THREADS" -ge 1
for f in "$GBZ" "$MIN" "$ZIP" "$DIST" "$R1" "$R2"; do test -s "$f"; done
test ! -e "$OUT" && test ! -L "$OUT"
mkdir "$OUT"
vg version > "$OUT/version.txt" 2>&1
sha256sum "$GBZ" "$MIN" "$ZIP" "$DIST" > "$OUT/indexes.sha256"
/usr/bin/time -v -o "$OUT/resources.txt" \
  vg giraffe -Z "$GBZ" -m "$MIN" -z "$ZIP" -d "$DIST" \
  -f "$R1" -f "$R2" -t "$THREADS" \
  > "$OUT/mapped.gam" 2> "$OUT/mapping.log"
test -s "$OUT/mapped.gam"
vg stats -a "$OUT/mapped.gam" > "$OUT/alignment-stats.txt" 2> "$OUT/stats.log"
test -s "$OUT/alignment-stats.txt"
```

检查输出，再安排批量运行

先看GAM是否可读、统计中的总比对数和配对统计是否与输入规模相符、日志是否有索引警告，并抽查mapping质量。将统计结果与FASTQ读数和配对质控相互核对；文件非空只说明有输出，不表示每条reads都成功比对。比对率没有通用合格线，要结合读长、样本类型和图面板覆盖解释；即使比对率高，也不能证明图的样本面板适用或变异分型准确。若下游需要BAM，按该图的参考路径执行投影，并另外检查BAM头、坐标和丢失的信息；不能把GAM改后缀当成BAM。

索引加载、FASTQ解压和图遍历会分别占用内存、CPU和I/O。记录FASTQ大小、索引版本与指纹、线程、耗时、峰值RSS、GAM大小和退出码；通过代表样本后再小规模增加并发。工作盘预留输出和失败重跑空间。索引发布记录中还应有文件大小、散列、vg版本、构建参数和参考路径；索引迁移后复核散列，避免复制中断导致任务在运行时才报错。若运行失败，先区分索引不匹配、读段配对、路径权限、磁盘和内存问题；索引缺失不能靠增加内存解决。每个样本使用独立目录，保留本次日志和资源记录，便于复核。

参考资料

vg官方短读长Giraffe教程：索引清单及运行方式 官方文档
Giraffe官方最佳实践：显式提供索引 官方文档