CollectHsMetrics教程：核对bait、target与覆盖指标

以WES样本为例，区分捕获探针区与评估靶区，按参考字典转换区间并检查CollectHsMetrics输出。

CollectHsMetrics用于评估捕获测序表现。运行前要区分bait（捕获探针区）和target（项目希望统计的区域）；把同一BED随意填入两项，会改变指标定义。本文以已完成质控的人类WES BAM为例，演示区间转换、指标输出和结果核查。数值用于评估实验与分析，不直接代表变异检测质量。

输入	运行前核对	留档
BAM	排序、样本和参考字典	校验、版本、上游过滤
bait BED	对应探针和试剂版本	来源、坐标和指纹
target BED	项目评估的实际靶区	范围、边界和序列字典

使用同一参考字典转换区间

BAM、FASTA、字典和BED必须使用匹配的组装及contig名称。BED起点为0且末端不包含，Picard interval_list从1起且末端包含；用BedToIntervalList正式转换，不要手工改坐标。转换后抽查序列名和边界，并确认bait、target分别符合试剂和项目定义。

运行并保留逐靶区结果

下面的MAPQ 20、BQ 20和Java堆8 GB是演示设置，应按项目分析约定和实际可用内存调整。PER_TARGET_COVERAGE便于定位低覆盖区，可能产生较大文件；只需要summary时不要额外开启逐碱基输出。

```
set -euo pipefail
bam=/data/S1.marked.bam
ref=/refs/genome.fa
dict=/refs/genome.dict
baits=/refs/panel.baits.bed
targets=/refs/panel.targets.bed
for f in "$bam" "$ref" "$ref.fai" "$dict" "$baits" "$targets"; do test -s "$f"; done
samtools quickcheck -v "$bam"
mkdir hsmetrics-S1-001
out="$PWD/hsmetrics-S1-001"
gatk --version > "$out/version.txt" 2>&1
gatk BedToIntervalList -I "$baits" --SEQUENCE_DICTIONARY "$dict" \
  -O "$out/baits.interval_list" > "$out/baits.log" 2>&1
gatk BedToIntervalList -I "$targets" --SEQUENCE_DICTIONARY "$dict" \
  -O "$out/targets.interval_list" > "$out/targets.log" 2>&1
test -s "$out/baits.interval_list"; test -s "$out/targets.interval_list"
/usr/bin/time -v -o "$out/resources.txt" \
  gatk --java-options "-Xmx8g" CollectHsMetrics -I "$bam" \
  -O "$out/hs.metrics.txt" -R "$ref" \
  --BAIT_INTERVALS "$out/baits.interval_list" \
  --TARGET_INTERVALS "$out/targets.interval_list" \
  --PER_TARGET_COVERAGE "$out/per_target.tsv" \
  --MINIMUM_MAPPING_QUALITY 20 --MINIMUM_BASE_QUALITY 20 \
  > "$out/run.log" 2>&1
test -s "$out/hs.metrics.txt"; test -s "$out/per_target.tsv"
```

按定义解释指标，不只看平均深度

先查看metrics文件中的BAIT_TERRITORY和TARGET_TERRITORY是否符合预期，再结合测序量、比对和过滤损失解读覆盖。MEAN_TARGET_COVERAGE是目标区平均深度，PCT_TARGET_BASES_20X是达到20倍覆盖的目标碱基比例；平均值可能被少数高深度区域拉高，所以两项需一起看。跨试剂或版本比较时，bait和target分母可能不同，应按捕获方案分组，不把不同区间集合的汇总值放在同一排名里。PCT类字段通常以小数比例表示，例如0.8对应80%。Fold 80 base penalty反映覆盖均匀性与达到目标深度的代价，需按Picard定义解释。

先检查per-target表的目标数、区间名称和零覆盖记录，确认与转换后的target interval相符；不要把目标数不同的样本直接按原始行号逐行比较。将per-target结果与低覆盖区域、探针设计和样本质控对应起来。on-target异常时先查bait/target、参考版本、污染与比对；总测序量足够但target coverage偏低时，再查捕获均匀性、重复标记、MAPQ和目标区间过滤。不要仅为了提高指标而改BED或过滤阈值。若异常集中在个别样本，比较其测序量、参考头和target覆盖；全批相似异常则优先复查区间版本与运行参数。

批量时记录资源和差异

每个样本使用独立输出目录，并记录输入校验值、参考和区间文件指纹、GATK版本、阈值、summary、逐靶区表、日志、运行时间、峰值内存及文件大小。先测代表样本再安排并发；每次用样本ID关联summary、per-target文件和原始区间清单，避免表格排序后错配。不同批次应固定bait/target定义、GATK版本和过滤条件。若改变MAPQ/BQ阈值，结果应作为另一套分析口径报告。失败时分辨输入/坐标字典错误、工具运行失败和真实低覆盖，不覆盖已审核输出；需要比较样本时同时报告各自QC和实验批次，便于区分实验与设置变化，也方便团队复核结果。

参考资料

GATK Picard CollectHsMetrics官方参数与指标 官方文档
GATK Picard BedToIntervalList：BED坐标与转换 官方文档