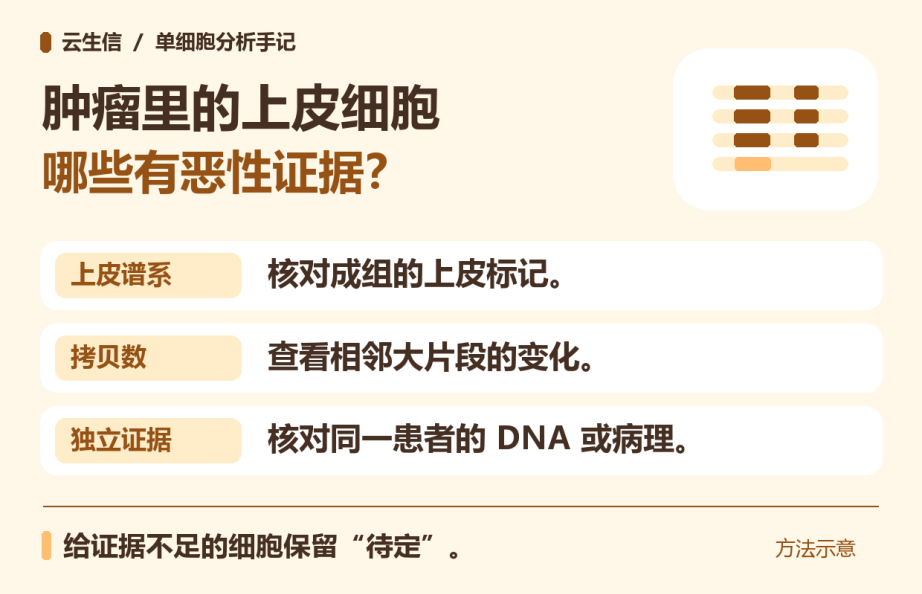
恶性上皮候选的参考复核

从示例数据跑到第一张结果表

对同一份样本的两套参考设置，逐细胞核对软件分类，再列出参考敏感、缺少预测和需要独立证据的条码。教学输入是构造的预测表，练习从已有 CopyKAT 输出开始。



准备 R 环境

仅使用 R 自带的软件包。 示例已在 R 4.3.3 运行。

完整解压资料包，在 run.R 所在文件夹打开终端，运行以下命令。

```
Rscript run.R example results
```

若使用 RStudio，将工作目录设为同一文件夹，再在控制台输入下面这一行。

```
source("run.R", encoding="UTF-8")
```

打开 results 文件夹查看输出。expected 文件夹中有示例的参考结果，示例数字均来自教学模拟。

逐步复算示例

在示例已经跑完的同一个资料包目录中打开 R。下面两段代码可依次粘贴，也可以执行包内的“读表练习.R”。

读第一张表

```
x <- read.delim("results/cell_review.tsv")
epi <- subset(x, lineage == "epithelial")
with(epi, table(call_auto, call_reference, useNA = "ifany"))
subset(epi, sample_id == "S1" & cell_id == "C02",
       c(cell_id, call_auto, call_reference, review))
```

S1 的 C02 在自动参考中为 aneuploid，在指定参考中为 diploid；它进入 reference_sensitive，优先回查参考、表达覆盖和染色体片段图。两个样本都故意使用 C02 这个条码，按样本和条码联合匹配才不会错位。

用第二项结果核对

```
todo <- read.delim("results/manual_review.tsv")
with(todo, table(sample_id, review))
subset(todo, sample_id == "S1",
       c(cell_id, call_auto, call_reference, review))
```

S1 的 C04 接近二倍体且有预先记录的独立支持，C06 缺少第二次预测，二者进入不同的复核队列。样本级病理只能说明组织来源，不能直接给某个条码的 independent_support 填 1。

应能复现的三个观察

- S1 / C01 为 cnv_supported_candidate。
- S1 / C02 为 reference_sensitive，C04 为 near_diploid_review。
- S1 / C06 的第二次预测缺失，归入 insufficient。

cnv_supported_candidate 表示两套参考下均有非整倍体预测支持。接近二倍体、缺失预测或参考敏感都要单独解释。这个脚本不会从原始 RNA 推断拷贝数，也不会输出临床诊断。

换成自己的数据

逐样本完成实际 CopyKAT 分析。从每次运行的结果对象提取 prediction（cell.names、copykat.pred），分别加上 sample_id，再合并为两张预测表。输入独立证据时只标记有具体细胞或可归属细胞群支持的情形，并在实验记录中保留依据。

先对齐输入字段

cells.tsv	一行一个细胞，含 sample_id、cell_id、lineage 和 independent_support。仅 epithelial 进入上皮判读。
auto.tsv	自动参考预测。含 sample_id、cell.names、copykat.pred，保留 aneuploid、diploid、not.defined 或缺值。
reference.tsv	指定正常参考所得的另一套预测。字段与 auto.tsv 相同；须沿用原样本编号及同一细胞条码。

fit_auto 与 fit_reference 对应同一样本用两种参考运行的 CopyKAT 返回对象。多样本时先逐份附上 sample_id，再用 rbind 分别合并所有自动与指定参考表。

```
source("prepare.R", encoding="UTF-8")
auto <- tag_prediction(fit_auto$prediction, "S1")
ref <- tag_prediction(fit_reference$prediction, "S1")
write.table(auto, "auto.tsv", sep="\t", quote=FALSE, row.names=FALSE)
write.table(ref, "reference.tsv", sep="\t", quote=FALSE, row.names=FALSE)
```

将输入放入 my-data，补齐其余示例字段和设置，再运行 Rscript run.R my-data my-results。

输出文件怎样读

cell_review.tsv	原始分类、两套预测、外部证据标记和复核原因。
sample_summary.tsv / manual_review.tsv	按样本计数和需要人工回看的条码。
reference-agreement.pdf	上皮细胞两次判读的交叉计数图。

方法资料 / CopyKAT 作者说明

换参考以后，哪些上皮细胞改变了 CNV 判定?

两种参考设置给出同样的标签，需先看究竟有多少条码在两边都获得了可比较的分类。把缺失、未分类和发生翻转的条码拆开，逐样本核对，再交给病理或基因组数据复核。

```
Rscript advanced.R example results epithelial
```

末尾参数指定谱系，教学输入可填 epithelial、immune 或 stromal。先在每个样本内限制谱系，只有在两次均判为 diploid 或 aneuploid 的条码才进入一致率分母。运行主脚本时需要已经准备好的两套 CopyKAT 结果。

S1 · 上皮谱系 · 两套分类

上皮条码	可比较	一致	翻转
6	4	3	1
无法比较	2	一致率	75%

模拟样本 S1 有 6 个上皮条码。两套结果都给出可比较分类的有 4 个，其中 3 个一致，所以一致率为 3/4。另 2 个无法比较。如果把那两个算进分母，会把覆盖问题误写成参考引起的翻转。

从结果表复算

```
a <- read.delim("results/agreement-by-sample.tsv")
subset(a, sample_id == "S1")
x <- read.delim("results/changed-calls.tsv")
subset(x, sample_id == "S1", c(cell_id, call_auto,
                               call_reference, review))
```

comparable 是可比较的双侧分类数。auto_only_aneuploid 和 ref_only_aneuploid 分别统计两种翻转方向。changed-calls.tsv 留下原始样本和条码，便于返回表达图复查。

在 changed-calls.tsv 找到 S1 的 C02，自动参考下为 aneuploid，指定参考下为 diploid。再查看 C04 的独立证据，它虽然在两次均为 diploid，仍保留近二倍体复核状态。两种参考的一致率无法代替恶性判定的准确率。