Sample swap

Jonathan Bloom
jbloom@broadinstitute.org
June 20, 2014

Let θ and φ denote the haplotypes of samples producing sequencing data sets x and y, respectively. Let s be a Bernoulli random variable with $s = 1$ if these samples are from distinct individuals. We call the event $s = 1$ a *swap*, and in this note we compute its posterior probability $p(s = 1 \,|\, x, y)$. In the context of sample validation, x is low-throughput data from a sample of interest and y is micro-array or high-throughput data from a sample intended to be from the same individual. If $p(s = 1 \,|\, x, y)$ is non-trivial, then one should toss the data and start again.

By Bayes rule, the posterior probability of a swap is given by

$$p(s = 1 \,|\, x, y) = \frac{p(x, y \,|\, s = 1)\, p(s = 1)}{p(x, y \,|\, s = 1)\, p(s = 1) + p(x, y \,|\, s = 0)\, p(s = 0)} \tag{1}$$

Equivalently, the posterior odds of a swap is the product of the Bayes factor (likelihood ratio) and prior odds:

$$\frac{p(s = 1 \,|\, x, y)}{p(s = 0 \,|\, x, y)} = \frac{p(x, y \,|\, s = 1)\, p(s = 1)}{p(x, y \,|\, s = 0)\, p(s = 0)} \tag{2}$$

In particular, if sample swap rarely occurs then the posterior log odds of a swap is well-approximated by

$$\log(L_1) - \log(L_0) + \log(S)$$

where $L_i = p(x, y \,|\, s = i)$ and S is the prior probability of a swap.

To compute these, the following functions must be empirically estimated:

- the prior $p(s)$, or equivalently the prior probability of a swap.
- the haplotype distribution $p(\theta)$ in the source population.
- the likelihood function $p(x \,|\, \theta)$ up to a scaling factor $c(x)$.
- the likelihood function $p(y \,|\, \varphi)$ up to a scaling factor $c(y)$.

Then by (1), it suffices to express $p(x, y \mid s)$ in terms of $p(\theta)$, $p(x \mid \theta)$, and $p(y \mid \varphi)$. To do this, we assume that distinct individuals are independently drawn from the population and that samples from the same individual have the same haplotype:

$$p(\theta, \varphi \mid s) = \begin{cases} p(\theta)\, p(\varphi) & \text{if } s = 1, \\ p(\theta) & \text{if } \theta = \varphi, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Next we have

$$\begin{aligned} p(x, y \mid \theta, \varphi) &= \frac{p(x, y \mid \theta, \varphi)}{p(y \mid \theta, \varphi)} p(y \mid \theta, \varphi) \\ &= p(x \mid \theta, \varphi, y)\, p(y \mid \theta, \varphi) && (4) \\ &= p(x \mid \theta)\, p(y \mid \varphi). && (5) \end{aligned}$$

where (4) applies the definition of conditional probability and (5) uses that x is conditionally independent of φ and y given θ, and y is conditionally independent of θ given φ. Therefore

$$\begin{aligned} p(x, y \mid s) &= \sum_{\theta,\varphi} p(x, y \mid \theta, \varphi, s)\, p(\theta, \varphi \mid s) && (6) \\ &= \sum_{\theta,\varphi} p(x, y \mid \theta, \varphi)\, p(\theta, \varphi \mid s) && (7) \\ &= \sum_{\theta,\varphi} p(x \mid \theta)\, p(y \mid \varphi)\, p(\theta, \varphi \mid s) && (8) \\ &= \begin{cases} \sum_{\theta} p(x \mid \theta)\, p(\theta) \, \sum_{\varphi} p(y \mid \varphi)\, p(\varphi) & \text{if } s = 1, \\ \sum_{\theta=\varphi} p(\theta)\, p(y \mid \varphi)\, p(\theta) & \text{if } s = 0. \end{cases} && (9) \end{aligned}$$

Here (6) is the law of total probability, (7) uses that x and y are conditionally independent of s given θ and φ, (8) applies (5), and (9) applies (3). Substituting (9) into (2), we conclude that the posterior odds of a swap is:

$$\boxed{\frac{\sum_{\theta} p(x \mid \theta)\, p(\theta) \, \sum_{\varphi} p(y \mid \varphi)\, p(\varphi)}{\sum_{\theta=\varphi} p(x \mid \theta)\, p(y \mid \varphi)\, p(\theta)} \cdot \frac{p(s = 1)}{p(s = 0)}}. \tag{10}$$

Computing (2) is easiest when θ, φ, x, and y are understood as tuples indexed by loci such that:

- the haplotypes at distinct loci are independent, i.e. $p(\theta) = \prod_i p(\theta_i)$.
- x_i and θ_j are independent for $i \neq j$, i.e. $p(x_i \,|\, \theta) = p(x_i \,|\, \theta_i)$.
- y_i and φ_j are independent for $i \neq j$, i.e. $p(y_i \,|\, \varphi) = p(y_i \,|\, \varphi_i)$.

In this case, by a generalization of the argument in (5) we have

$$
\begin{aligned}
p(x \,|\, \theta) &= \prod_i p(x_i \,|\, \theta_i) \\
p(y \,|\, \varphi) &= \prod_i p(y_i \,|\, \varphi_i).
\end{aligned}
$$

Substituting these expressions into (10) and re-distributing yields

$$
\boxed{\prod_i \left(\frac{\sum_{\theta_i} p(x_i \,|\, \theta_i)\, p(\theta_i) \, \sum_{\varphi_i} p(y_i \,|\, \varphi_i)\, p(\varphi_i)}{\sum_{\theta_i = \varphi_i} p(x_i \,|\, \theta_i)\, p(y_i \,|\, \varphi_i)\, p(\theta_i)} \right) \cdot \frac{p(s=1)}{p(s=0)}}.
$$