Sample swap

Jonathan Bloom jbloom@broadinstitute.org
and
Yossi Farjoun farjoun@broadinstitute.org
June 20, 2014

Let θ and φ denote the diploid haplotypes (that is AA, AB or BB) of samples from which sequencing data sets x and y, respectively, have been generated. 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 sequencing data from a sample of interest and y is micro-array or sequencing data from a sample intended to be from the same individual. If $p(s = 1 \,|\, x, y)$ is non-trivial, then one should flag the suspect data and investigate.

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 a 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

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

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

$$p(x \,|\, \theta) = \prod_i p(x_i \,|\, \theta_i)$$
$$p(y \,|\, \varphi) = \prod_i p(y_i \,|\, \varphi_i).$$

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

$$\boxed{\prod_i \left(\frac{\sum_{\theta_i} p(x_i \,|\, \theta_i)\, p(\theta_i) \;\sum_{\varphi_i} p(y_i \,|\, \varphi_i)\, p(\varphi_i)}{\sum_{\theta_i = \varphi_i} p(x_i \,|\, \theta_i)\, p(y_i \,|\, \varphi_i)\, p(\theta_i)} \right) \cdot \frac{p(s=1)}{p(s=0)}} \tag{11}$$

1 Haplotype Likelihoods

In this section we describe how the haplotype likelihood $p(x|\theta)$ can be computed for various kinds of data.

1.1 Sequence data

For sequence data, we assume that the data come from a single individual (i.e. not contaminated, see subsection below) and that there is no reference bias (can correct for that, but it's rarely needed). Sequence data arrives in the form of reads. We assume that evidence for haplotype $h_i \in \{A, B\}$ with probability of error $e_i \in (0, 1)$ are given at a certain haplotype block. We further assume that said evidence is independent, for example, reads have been duplicate marked, and close SNPs from the same read-pair are not used twice. Then we can write:

$$p(h, e|\theta) = \prod_{i=0}^{n} p(h_i, e_i|\theta) \tag{12}$$

The likelihood of a single datum $p(h_i, e_i|\theta)$ is expressed by

$$p(h_i, e_i|\theta) = \begin{cases} I_B(h_i)e_i + I_A(h_i)(1 - e_i) & \theta = AA \\ 0.5 & \theta = AB \\ I_A(h_i)e_i + I_B(h_i)(1 - e_i) & \theta = BB \end{cases} \tag{13}$$

Where I_x is the indicator function of x and the assumption is that an error will cause a switch in the interpretation of the haplotype between A and B. This assumes that we throw away non-conformant haplotypes, and ignores the possibility of a non-conformant haplotype erroneously looking conformant. (By conformant we mean either A or B.)

This LOD calculation is implemented in Picard's CheckFingerprints and CrosscheckFingerprints.

1.2 Contaminated Samples

At times, one knows that data from a particular sample is contaminated at a known level (that level can be estimated using VarifyBamID, or ContEst, for example). However, the identity of the contaminator is unknown. In this section we describe how the diploid Haplotype likelihood of the contaminator can be (sometimes) extracted from the data. For the calculation we will need the prior on the haplotypes, $p(\theta)$, which can be calculated from the haplotype frequency by assuming Hardy-Weinberg equilibrium.

We assume that the underlying samples have (unknown) haplotypes θ and ϕ, mixed with proportion $1-c$ and c, respectively. The data collected from this mixture, is x. Then we can write:

$$p(x|c,\theta) = \sum_{\phi} p(x|c,\theta,\phi)p(\phi) \tag{14}$$

This enables us to discover (and thus extract) $p(x|c,\theta)$ which can then be used in (11).

Picard's ExtractContaminantFingerprint implements this and writes a new VCF with the likelihoods of the contaminant. This VCF can then be further used to compare the contaminant (or the contaminated sample) to another sample.

1.3 LoH samples

When a sample comes from a tumor there is the possibility that it has undergone a loss of heterozygosity (LoH) where large sections of chromosomes are lost (whole arms, and sometimes one copy of a whole chromosome can be lost). This makes all the heterozygous haplotypes (from the germline) in that region of the chromosome seem homozygous since the only evidence

comes from the remaining copy. In this case, the standard calculation of $p(x|\theta)$ will yield an incorrect likelihood for the genotypes of the normal sample. Furthermore, since the LoH event is correlated among the different sites, our assumption of independence is incorrect. We need to be able to infer the genotype posterior of the normal sample given the data from the tumor, $p(G_n|D_t)$, so that we can compare genotypes of the individual.

We assume that a heterozygous site can become homozygous at a probability p_{loh}. Thus we can write a transition probability:

$$T = \begin{pmatrix} 1 & 0 & 0 \\ p_{loh}/2 & 1 - p_{loh} & p_{loh}/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{15}$$

and now

$$p(g_t = j|g_n = i) = T^i_j \tag{16}$$

where g_t is the tumor's genotype and g_n is the normal's genotype.

We are looking for $p(G_n|D_t)$ and so we observe:

$$\begin{aligned} p(D_t|g_n) &= \sum_{g_t} p(D_t, g_t|g_n) \\ &= \sum_{g_t} p(D_t|g_n, g_t)p(g_t|g_n) \\ &= \sum_{g_t} p(D_t|g_t)T^{g_t}_{g_n} \end{aligned}$$

Which gives us the likelihood of the tumor's data in the context of the normal's genotype.

Thus, in (10) above, if x' is data from the tumor with genotype θ' and whose normal has genotype θ, we get that the posterior odds of a swap is:

$$\boxed{\frac{\sum_\theta \sum_{\theta'} p(x' \mid \theta') T^{\theta'}_\theta \, p(\theta) \, \sum_\varphi p(y \mid \varphi) \, p(\varphi)}{\sum_{\theta=\varphi} \sum_{\theta'} p(x' \mid \theta') T^{\theta'}_\theta \, p(y \mid \varphi) \, p(\theta)} \cdot \frac{p(s=1)}{p(s=0)}} \tag{17}$$

This is implemented in CrosscheckFingerprints and emitted in the TUMOR_NORMAL_LOD and NORMAL_TUMOR_LOD columns in the fingerprinting report.