

# Estimating heterozygosity and LOH frequencies, accounting for genotyping failure

The purpose of this document is to describe the statistical procedure to estimate heterozygosity and LOH frequencies of microsatellites from genotype data, accounting for the facts that a) many genotypes fail (missing data); and b) genotype failure is correlated with both heterozygosity rates and LOH rates (the missing data is non-ignorable).

In general, suppose that we have  $M$  microsatellites genotyped in  $N$  cancer patients. These genotypes are performed in two compartments for each patient: germline and tumor. We now define a set of terms as they will be used in this document. Each of the  $MN$  patient microsatellites may be categorized in two ways. First, the individual is either *heterozygous* or *homozygous* at the microsatellite in the **germline**. Second, either the individual underwent *loss-of-heterozygosity (LOH)*, *retention-of-heterozygosity (ROH)*, or was *non-informative due to homozygosity in the germline* at this microsatellite. Note that, for many of the  $MN$  patient microsatellites, the heterozygous/homozygous status and/or the LOH/ROH/non-informative status may be obscured because of genotype failure.

We now define some notation to be used in this document. For individual  $i$  at microsatellite  $j$  (where  $i = 1, \dots, N$  and  $j = 1, \dots, M$ ) let

$$h_{ij} = \begin{cases} 1 & \text{if } i \text{ heterozygous at } j \\ 0 & \text{if } i \text{ homozygous at } j \end{cases}$$

We point out again that  $h_{ij}$  is not always observed. Similarly, let

$$l_{ij} = \begin{cases} 1 & \text{if } i \text{ undergoes LOH at } j \\ 0 & \text{otherwise} \end{cases}$$

As with  $h_{ij}$ ,  $l_{ij}$  is not always observed. Finally, let

$$z_{ij} = \begin{cases} 1 & \text{if genotype does not fail} \\ 0 & \text{if genotype fails} \end{cases}$$

In this case, note that  $z_{ij}$  **is always observed**, since genotyping failure of individual  $i$  at microsatellite  $j$  forces  $z_{ij} = 0$ .

Recall that the goal is to estimate heterozygosity frequency  $\beta_j$  and LOH frequency  $\gamma_j$  of each microsatellite  $j = 1, \dots, M$ . Natural estimates for each

of these parameters would be  $\frac{1}{N} \sum_{i=1}^N h_{ij}$ , and  $\frac{1}{N} \sum_{i=1}^N l_{ij}$ , but (as mentioned above) some of these quantities are missing. Therefore, in order to modify these estimates, we take advantage of the fact that missingness is tied to heterozygosity and LOH status. We introduce four new parameters, defined as:

$$\lambda_1 = \text{P}[\text{genotype does not fail} \mid \text{germline heterozygote}] \quad (1)$$

$$\lambda_2 = \text{P}[\text{genotype does not fail} \mid \text{germline homozygote}] \quad (2)$$

$$\nu_1 = \text{P}[\text{genotype does not fail} \mid \text{LOH}] \quad (3)$$

$$\nu_2 = \text{P}[\text{genotype does not fail} \mid \text{ROH}], \quad (4)$$

where  $\text{P}[\cdot]$  denotes the probability of the event.

Standard probability theory implies that, for any microsatellite  $j$ ,

$$\lambda_1 = \text{P}[\text{genotype does not fail and germline heterozygote}] / \beta_j$$

$$\lambda_2 = \text{P}[\text{genotype does not fail and germline homozygote}] / (1 - \beta_j)$$

The numerators of these expression can be estimated for each microsatellite  $j$ , in an unbiased manner, by  $\frac{1}{N} \sum_{i=1}^N h_{ij} z_{ij}$  and  $\frac{1}{N} \sum_{i=1}^N (1 - h_{ij}) z_{ij}$ . Note that, although  $h_{ij}$  is sometimes not observed, we may always determine the value of  $h_{ij} z_{ij}$ , since  $z_{ij} = 0$  by definition whenever  $h_{ij}$  is unobserved. Therefore, given estimates  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  of  $\lambda_1$  and  $\lambda_2$ , we may estimate each  $\beta_j$  as  $\frac{1}{N\hat{\lambda}_1} \sum_{i=1}^N h_{ij} z_{ij}$  or  $1 - \frac{1}{N\hat{\lambda}_2} \sum_{i=1}^N (1 - h_{ij}) z_{ij}$  by solving equations (1) or (2) above for  $\beta_j$ . Our estimate  $\hat{\beta}_j$  for  $\beta_j$  is the mean of these two,

$$\hat{\beta}_j = \left( \frac{1}{N\hat{\lambda}_1} \sum_{i=1}^N h_{ij} z_{ij} + 1 - \frac{1}{N\hat{\lambda}_2} \sum_{i=1}^N (1 - h_{ij}) z_{ij} \right) / 2 \quad (5)$$

Similarly, equations (1) or (2) may be used to estimate  $\lambda_1$  and  $\lambda_2$  for any estimate of  $\beta_j$ . Our estimate is obtained by averaging across all microsatellites  $j$ , yielding

$$\hat{\lambda}_1 = \frac{1}{M} \sum_{j=1}^M \left( \frac{1}{N\hat{\beta}_j} \sum_{i=1}^N h_{ij} z_{ij} \right) \quad (6)$$

$$\hat{\lambda}_2 = \frac{1}{M} \sum_{j=1}^M \left( \frac{1}{N(1 - \hat{\beta}_j)} \sum_{i=1}^N (1 - h_{ij}) z_{ij} \right) \quad (7)$$

Our iterative approach to produce final estimates  $\hat{\lambda}_1$ ,  $\hat{\lambda}_2$ , and the  $\hat{\beta}_j$  then proceeds as follows. First, since by definition  $\lambda_1$  is larger than the probability that the genotype does not fail and the germline is a heterozygote, we initialize

$$\hat{\lambda}_1 = \max_j \left( \frac{1}{N} \sum_{i=1}^N h_{ij} z_{ij} \right).$$

For similar reasons, we initialize

$$\hat{\lambda}_2 = \max_j \left( \frac{1}{N} \sum_{i=1}^N (1 - h_{ij}) z_{ij} \right).$$

Next we: (a) use equation (5) to estimate the  $\beta_j$ , and then (b) use equations (6) and (7) to re-estimate  $\lambda_1$  and  $\lambda_2$ . Steps (a) and (b) are iterated until all estimates converge.

We estimate the  $\gamma_j$  using an analogous iterative procedure, this time using  $l_{ij}$ ,  $\nu_1$ , and  $\nu_2$ .