This short document explains the important difference between `bed2diffs-v1` and `bed2diffs-v2`.

1 bed2diffs-v1

Here the average difference between two samples, i and j, is given by

$$D_{ij} = \frac{1}{\left|M_{ij}\right|} \sum_{m \in M_{ij}} \left(z_{im} - z_{jm}\right)^2, \tag{1}$$

where M_{ij} is the set of SNPs where both i and j are called, and z_{im}, z_{jm} are the genotypes of individuals i, j at marker m.

In practice, this computation is not guaranteed to produce an Euclidean distance matrix – a nonnegative matrix with 0s on the diagonal, with exactly one positive eigenvalue [Gower, 1982] – especially, if genotypes are not missing at random.

2 bed2diffs-v2

Alternatively,

$$D_{ij} = \frac{1}{\left|M_{tot}\right|} \sum_{m \in M_{tot}} \left(z^*_{im} - z^*_{jm}\right)^2, \tag{2}$$

where M_{tot} is the set of all markers and

$$z^*_{im} = \begin{cases} z_{im} & \text{if } z_{im} \text{ is called,} \\ \bar{z}_m & \text{otherwise,} \end{cases} \tag{3}$$

where $\bar{z}_m$ is the average genotype at marker m.

This would produce an Euclidean distance matrix. Setting z_{im} to the observed average at the marker m is similar to the "imputation" often used before principal component analysis [Price et al., 2006].

References

[Gower, 1982] Gower, J. C. (1982). Euclidean distance geometry. *Math. Sci.*, 7:1–14.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38:904–909.