STRmix™

Variation in STRmix™ versions 2.6, 2.7 and 2.8 regarding calculation of varNOC *LR*s

11 December 2020

Summary

This is relevant to users of the Variable Number of Contributors (varNOC) function only. When varNOC was first implemented in STRmix™ V2.6.0, the size of the hyper-rectangle was set as a fraction (M%) of the total iterations. The algorithm used to set the size of the hyper-rectangle mistallied the number of iterations within this volume resulting in the volume and hence M being only approximate. It therefore contains only approximately M% of the total iterations. This was detected during the ongoing development of STRmix™. This discrepancy applies to versions 2.8 and earlier (to V2.6.0). The effect on the varNOC *LR* is marginal and is less than the effect of run to run variation.

Specific Mechanism

During calculation of the hyper-rectangle, the algorithm finds the iteration with the highest likelihood and terms this the mode. Starting from the mode and with decreasing likelihood the algorithm should tally the residence time in each state until it reaches M x total iterations. However it actually tallies the number of iterations since last accept which is the residence time for the previous state. It accrues these iterations until they tally M x total iterations. The extrema of each variable in this set define the boundaries of the hyper-rectangle. Since the residence time for the previous state is used, the count goes either too far or not far enough. Hence if the hyper-rectangle is expected to include the top M% of likelihoods it actually includes a percentage close to M, but not exact.

It is difficult to visualise if this will be systematically larger or smaller than M. The high likelihoods should on average have high residence times since this is the finding of the Metropolis-Hastings algorithm. The residence time for the previous state for the count may, on average, be smaller. If this is so, then a hyper-rectangle expected to include M% of highest likelihoods will in fact include a larger fraction and therefore have a larger volume.

The consequences of changing the size of the hyper-rectangle have previously been studied. In McGovern et al. [1] supplementary material Figure S4 the size of the hyper-rectangle was varied significantly $M \in \{0.01, 0.025, 0.05\}$. The varNOC *LR*s produced under these different conditions were plotted on an x,y scatter plot using STRmix™ V2.7. This is

reproduced below. These plots demonstrate that varying the size of the hyper-rectangle considerably does not have a large difference on the varNOC *LR.*

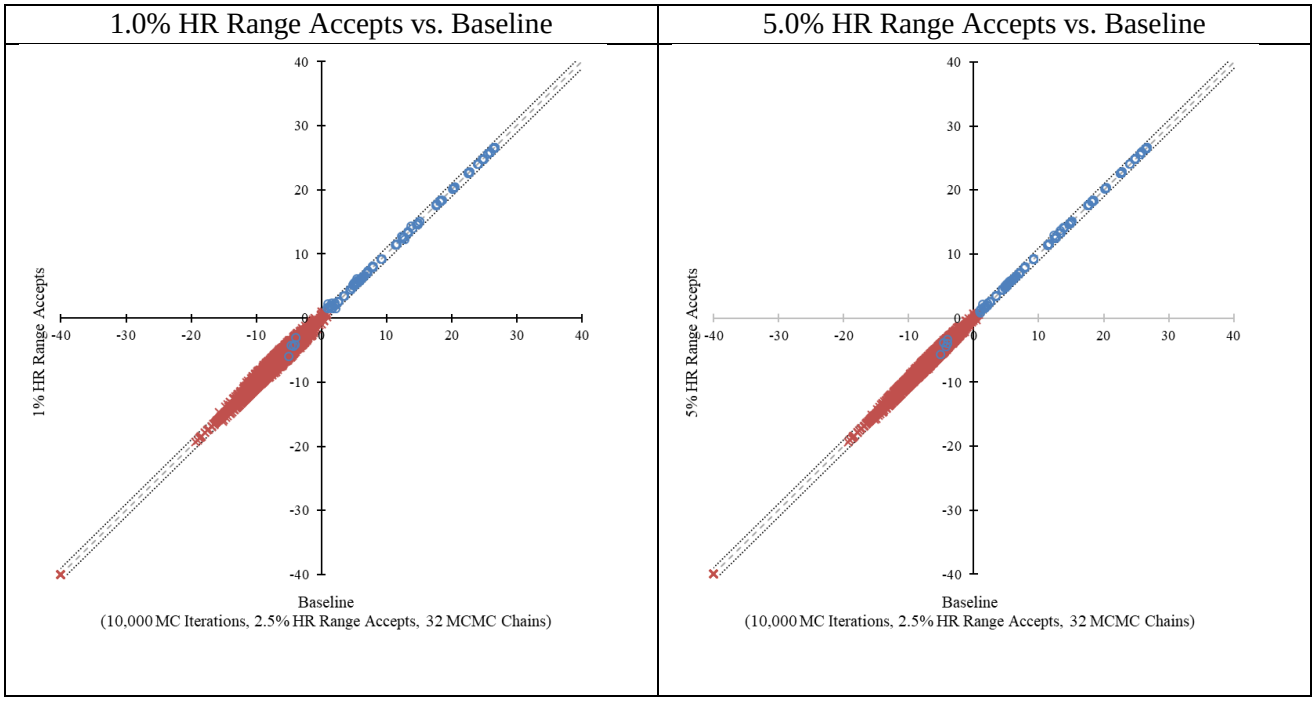

Figure 1: Reproduction of Figure S4 from McGovern et al. [1]. Plots of the $\log_{10}(LR)$s assigned for each database individual across the five replicates interpreted using varNoC run parameters of 10,000 MC iterations, 32 MCMC chains and 1.0% or 5.0% HR range accepts versus the baseline varNoC run parameters of 10,000 MC iterations, 32 MCMC chains and 2.5% HR range accepts. This demonstrates the changes in the *LR* due to the HR range parameter. Exclusions (*LR* = 0) are plotted as log10 (*LR*) = -40. Blue circles are *LR*s assigned to known contributors and red crosses are *LR*s assigned to known non-contributors. If the calculated *LR*s were identical between two sets of parameters, the comparison of *LR*s will lie along the dashed $x=y$. line. The dotted lines represent the bounds for a one-order of magnitude difference.

Effect on the varNOC *LR* calculation:

A variety of mixtures (some GlobalFiler™ from the PROVEDIt dataset [2], some multikit samples from NIST MIX13 [3] and some Fusion samples) were examined using the varNOC functionality of STRmix™ where the assigned contributor ranges contained examples of both under- and over-estimation. For each combination of STRmix™ version and seed tested, 39,300 stratifed varNOC *LR*s were produced by comparing against true and false donors. These were calculated in the commerically released STRmix™ V2.8 using two different seeds. This allowed an assessment of the run to run variability. Using the same run parameters and seed as one of the previous two sets, the samples were also run in a test version of STRmix™ V2.8 where the hyper-rectangle calculation fix had been applied. This allowed an assessment of the variability in the varNOC *LR* between the released version of STRmix™ V2.8 and the test version. The results appear in Figure 2 below.

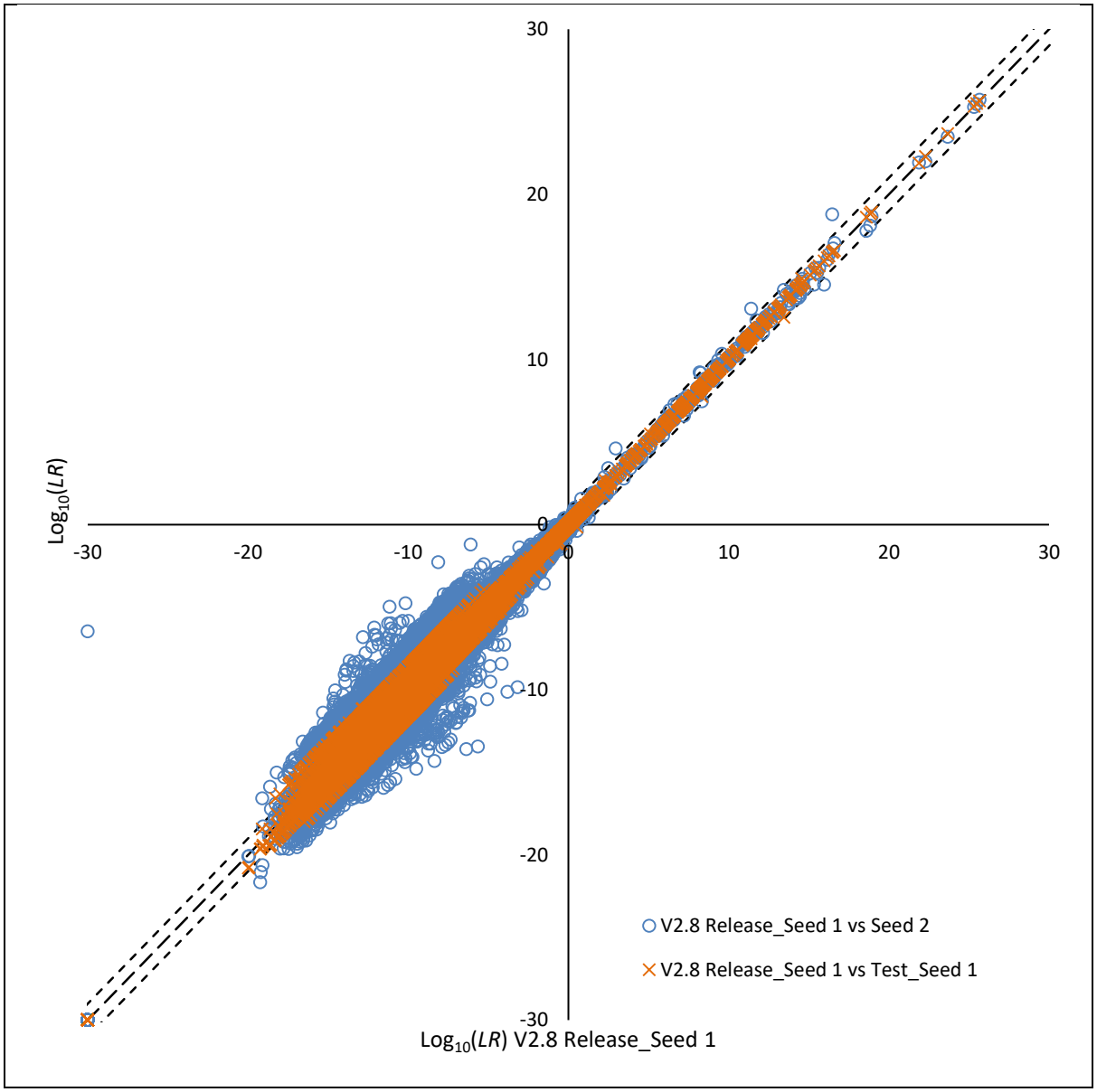

Figure 2: Plot of varNOC $\log_{10}(LR)$s run in i) the released version of STRmix™ V2.8 with two different seeds (blue circles) and ii) run with the same seed in the released version of STRmix™ V2.8 and a test version with the hyper-rectangle fix applied (orange crosses). Exclusions (*LR* = 0) are plotted as log10 (*LR*) = -30. If the calculated *LR*s were identical between two run conditions, the comparison of *LR*s will lie along the dashed $x=y$. line. The dotted lines represent the bounds for a one-order of magnitude difference.

These results indicate that the effect of the change to the hyper-rectangle bounds on the varNOC *LR* is minimal and more variability is observed due to run to run variation.

References

[1] C. McGovern, K. Cheng, H. Kelly, A. Ciecko, D. Taylor, J.S. Buckleton, J.-A. Bright, Performance of a method for weighting a range in the number of contributors in probabilistic genotyping, Forensic Sci. Int. Genet. 48 (2020).

[2] L.E. Alfonse, A.D. Garrett, D.S. Lun, K.R. Duffy, C.M. Grgicak, A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt, Forensic Sci. Int. Genet. 32 (2018) 62-70.
[3] J.S. Buckleton, J.-A. Bright, K. Cheng, B. Budowle, M.D. Coble, NIST Interlaboratory Studies Involving DNA Mixtures (MIX13): A modern analysis, Forensic Sci. Int. Genet. 37 (2018) 172-179.