Using *UIDcollision\_VarPCR.r*

Jenna VanLiere Canzoniero, M.D., M.S.

Johns Hopkins University Department of Medicine

Last updated August 17, 2016

1. Description

The file UIDcollision\_VarPCR.r contains the primary function UIDerrorVarPCR, as well as supporting function estP2same and ecol. UIDerrorVarPCR simulates the results of a next-generation sequencing run performed with barcoding and processed into consensus sequences by family. UID assignment influences the success rate of PCR according to specified parameters. estP2same estimates the probability that at least 2 sample molecules have the same UID. ecol calculates the expected number of samples that will have non-unique UIDs.

1. Usage

UIDerrorVarPCR(nsamp, mutFreq, nUID, seed=1, conFrac=0.95, PCRcyc=25, PCRsuc=c(0.25, 0.5), nb, minMembers=2)

1. Arguments

|  |  |
| --- | --- |
| nsamp | Number of sample molecules, must be a positive integer |
| mutFreq | Actual underlying frequency of mutation in data, note that if the mutFreq x nsamp does not results in an integer number of mutants, then the number of mutant molecules will be rounded according to rules of round() |
| nUID | Number of possible UIDs, must be a positive integer |
| seed | Random generator seed, default 1 |
| conFrac | Fraction of molecules within each UID family that must agree in order for a consensus sequence to be called. Default 0.95 (95%) |
| PCRcyc | Number of PCR cycles, default 25 |
| PCRsuc | Vector that contains the low and high probability of success for each PCR cycle, default (0.25, 0.5). Only the minimum and maximum values of an inputted vector are used. |
| nb | Of the UIDs that bind to sample molecules, number of unique UIDs with higher success rate, can be any positive integer, should be less than nsamp |
| minMembers | Minimum number of members of UID family in order for UID consensus to be called |

1. Details

UIDerrorVarPCR simulates binding of UID molecules to sample molecules as sampling with replacement. Only 1 locus of mutation is simulated. Of the UIDs that got bound to sample molecules, the first nb UIDs are chosen to “better”, i.e. have higher PCR success rate. All UID-sample combined molecules where the UID is “better” are then amplified using binomial distribution with the higher PCR success rate. UID-sample combined molecules where the UID is not “better” are then amplified using binomial distribution with the lower PCR success rate.

Please note that the r file contains an example at the bottom of how to set up the simulation. You should comment this out before sourcing the file. Please see in-code notes for further details.

1. Value

UIDerrorVarPCR returns a numerical vector with the following components in order:

|  |  |
| --- | --- |
| emutFreq1 | Mutation frequency estimated from the data assuming that families without a consensus call are removed from the denominator. |
| err1 | Error of the emutFreq1 compared to mutFreq |
| emutFreq2 | Mutation frequency estimated from the data assuming that all families, including those without a consensus call, are included in the denominator. |
| err2 | Error of emutFreq2 compared to mutFreq |
| ncfreq | Fraction of UID families that did not produce a consensus sequence |
| uUIDs | Number of unique UIDs used |
| nreads | Number of reads expected for sequencing based on success of PCR amplification |