Thoughts on multiplexed amplicon read depth

Suppose we have 1000 multiplexed individuals, each with 8 chromosomes, each with 200 amplicons in a perfect PCR reaction that results in an equimolar concentration of each individual amplicon. The total number of (potentially) unique amplicons is therefore 1,600,000 at a ratio of 1 : 1 : 1… Of these amplicons, only a proportion, X, contains useful information; we do not expect SNPs across all sub-genomes for all amplicons. As a PCR reaction allows for (theoretically) limitless amplification of amplicons, we assume that the concentration of all of these unique amplicons is very high.

Suppose we have a perfect sequencing machine, for which we wish to minimise the number of reads, but remain confident that we have coverage of every useful amplicon, with some p value, p. In other words, we want to be confident that the probability of failing to sample every single useful amplicon at least once (or some number of times) after n total reads is low. We assume that the probability of drawing any individual is independent of drawing any other individual and that the individual is drawn from a probability mass function 1 : 1 : 1…

To solve this problem computationally by simulation, we define X specific amplicons. We then draw individuals sequentially from the probability mass function until all X amplicons have been picked, counting the number of trials needed. We repeat this sequence a large, L, number of times, and numerically order the number of trials needed. Providing L >> 1/p, then the [1-(p\*L)]th number of trials becomes an estimate for n (i.e. total number of reads) needed to draw all X amplicons with a probability > p

Practically, this is probably a very large number and we would accept some missing amplicons in our reads, by replacing X with x, where x is a smaller number of any elements of X.