For all given sub-regions across all genes we will estimate three tallies (and in parallel each of their corresponding 'sum of their mutation rates'):

a) the tally of (and the sum of mutation rates for) all simulated variants resulting in nonsense (stop gain/loss) predicted effects

b) the sum of (and the sum of mutation rates for) all simulated variants resulting in missense predicted effects

c) the sum of (and the sum of mutation rates for) all simulated variants resulting in synonymous predicted effects

Goal is to then use the tallies (or sum of mutation rates) to calculate the simple 'expected ratio of non-synonymous variants', defined as: (missense rate + nonsense rate) / sum(all three rates).

As an example, here is GNB1 coordinates:

GNB1 1 (1718768..1718878,1720490..1720710,1721832..1722037,1724682..1724752,1735856..1736022,1737912..1737979,1747193..1747303,1749274..1749316,1756834..1756894) 1059

and the corresponding domains within it:

GNB1:238121:238121\_0 1 (1718773..1718876,1720492..1720708,1721834..1722035,1724684..1724750,1735858..1736020,1737914..1737977,1747195..1747256) 879

GNB1:-:-\_0 1 (1718770..1718772) 3

GNB1:-:-\_1 1 (1747257..1747301,1749276..1749314,1756836..1756892) 141

essentially, based on earlier evaluations, the corresponding % expected non-synonymous simulated variants (tally-based) in GNB1 should correspond to ~77.6%. In fact, the overwhelming majority of genes should be around the 70-75% range.

Based on the mutation rate of GNB1, the corresponding expected percentile should be approximately

(0.0000133100645665 + 0.00000111310616569)/ (0.00000628882828764 + 0.0000133100645665 + 0.00000111310616569) = 0.696367874408502 (aka 69.64%)

where:

0.0000133100645665 = ~ sum of missense mutation rate

0.00000111310616569 = ~ sum of stop gain/loss mutation rate

0.00000628882828764  = ~ sum of synonymous mutation rate

Note, these are all just approximates - not expecting you to get exact same! but hope is the eventual percentage should be the very similar.. P.S. I've attached the percentage expected to be achieved for a longer list of genes. also see histogram as a guide for this list. essentially, based on histogram one could even probably get away with just assigning all regions with expected non-syn mutation rate of 0.7 and pretty much I would expect they'd get similar results..! Obviously, we want to try and do better so i think appropriate to account for exact context, but just stating that the window of variability for majority of genes isn't that great..

**In brief**: we essentially:

1) simulate all possible mutations in a protein-coding region and then

2) annotate them, and then

3.1) just count the tallies within the region and

3.2) in parallel sum the corresponding mutation rates within the three tallies per gene.