How to do normalized counts for genes of interest.

?????? Have to check if they just used the 0.1 um datasets.

?????Could express as just frequencies of genes per 100,000 bp for each size fraction, as well as normalized to recA copies.

Howard et al. 2006

* Used the Sargasso Sea reads, ALOHA (pooled photic zone), ALOHA (pooled deep water) and Sapelo Island coastal.
* Used Ecoli recA as a blast query vs the whole datasets.
* Calculated % of cells with DMDA as counts\_dmdA\*100/recA

Howard et al. 2008

\* Used the GOS samples and looked for frequencies of dmdA, dddD, dddL per 100,000 reads.

\* Also calculated percentage of cells with each dmdA type.

\* Calculated the percentage of cells with genes of interest including dmdA, dddD, dddL and RUBISCO using several single copy genes to normalized including recA, rpoB, gyrA, dnaK, tufA and atpD.

\* Note, the way they detect dmdA is to blastp < e-20, remove duplicates (really like clustering for OTUs), those hits were checked against COG to see if they were in the correct COG group, then they were made into a tree using to see if they clustered with dmdA or the GcvT outgroup. Those in a dmdA clade with bootstrap > 70 were considered a dmdA A,B,C,D or E clade or unclassified clade if they did not group > 70.

\*single copy genes in GOS were retrieved by blastp of Ecoli K12 substrain MG1655 as queries < e-20. To account for differences in the lengths of the single copy genes, length of recA was divided by the length of the other genes, and that value was multiplied by the single copy gene count. (probably a bit overkill)

Percentage of cells with dmdA is then countDMDA\*100/average length normalised single copy gene count.

Todd et al. 2009

\*Did the same as Howard et al. 2008 but added in dddP and showed dddP containing cells as a percentage of recA.

\*DddP accepted if <e-80, and the others if <e-20.