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 reads 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.

***PROBLEM, You can’t simply use their counts of genes of interest as they use a different way of detecting them. You will have to try out your way on their datasets and see if you get a similar answer.***

Sharma et al. 2008

* Metagenomic sequences from 51 sample sites from GOS were downloaded, translated into all 6 frames and used as BLASTP database. Rhodopsins from a broad phylogenetic origins were used as query sequences with E-value e-10.
* The same strategy was used to retrieve the recA sequences. Query recA were Acetobater (Q08327), Chlorobium, Prochlorococcus, Frankia, Thermus, Moorella.
* Rad A was used to normalise for Archaeal frequencies.
* GOS samples were grouped according to the categories: ocean (23), coastal (22), estuarine (GS011) and freshwater (Lake Gatun GS020) and lagoon (Punta Cormorant GS033). Frequency of rhodopsin per category was given by count of rhodopsin BLASTP hits divided by sum of recA hits from that category.
* They found the proportion of genomes with rhodopsins were: 63% open ocean, 67% coastal, 62% estuaries, 35% Lake Gatun, 36% Punta Cormorant.
* They found that the actinorhodopsins of Lake Gatun (LG1 and LG2) and Punta Cormorant (PCL1) were more abundant in the non-marine environments. These are non-marine types.

Sabehi et al. 2005

* They cloned large metagenomic bits into BACS (up to 170 kb) from Mediterranean and Red Sea photic zone and PCRed for PR.
* 0.52% of clones had a PR, which assiming a genomes size of 2 Mb and PR is single copy, this suggests 13% of bacteria have a PR.

Raina et al. 2010

* Took dmdA, dddD, dddL and dddP as queries vs 152 metagenomes (San Diego State University Center for Universal Microbial Sequencing (SCUMS) database) of marine and terrestrial ecosystems. 26/152 had DMSP degrading enzymes.
* dmdA was more abundant than Ddds and also present in marine viriomes.
* Porites compressa coral metagenomes had no DMSP deg enzymes at all.
* dddL and dddD associated with open ocean viromes and saltern environments.
* dddP is the marine ddd.
* Shows a figure of the relative abundances of the DMSP degradation genes in the different environments.