**Explanation of functional score and DESEq2 use in SGE**

The BAP1 SGE functional score is a growth rate, produced with a log-linear regression with an appropriate error model (Negative Binomial a.k.a. Gamma-Poisson distribution for count processes with over-dispersion).

The functional score when hypothesis-tested asks:

"Is there a significant linear correlation between time and normalised log read count?"

* SGE count data (discrete positive numbers) with over-dispersion, can be well modelled by a negative binomial distribution.
* A GLM to fit data as a negative binomial has two key parameters: the mean and dispersion. Variability between replicate raw count data about the mean is routinely modelled using dispersion estimates. However, in experimental designs that have few, but reasonable, numbers of replicates (n=2 or 3), dispersion estimates will be highly variable for each variant (over-dispersed), and accurate estimation of dispersion is critical for statistical inferences about variant change over time. If dispersion estimates are used directly, they will be highly noisy, compromising subsequent statistical tests. An appropriate error model needs to be employed on a log-linear regression. DESeq2 satisfies this requirement as it accurately estimates dispersion and applies it to the GLM.
* LFC estimates in which time is considered as a continuous variable are computed as a log-linear regression through DESeq2. This is because the GLM of DESeq2 includes the requisite exponential function to link the probability distribution parameter (the central parameter of negative binomial distribution) and the linear regression term (the linear predictor, change in log read counts over time).
* The link can be understood mathematically as follows:

Assuming that for an oligo and time , the observed cell count is given by a negative binomial random variable , where is the mean of the distribution and is the dispersion coefficient. Assuming that at two different time stamps and the expected count follows log linear growth model, thus:

Where is decay time parameter, and are expected cell counts at time and , respectively. Then we can put these assumptions into one-to-one correspondence with DESeq2 (equations ‘1’ and ‘2’ in Love *et al*, 2014) such that:

and

Where and are precisions and given by and , respectively.

* The ‘functional score’ for a variant is therefore the Log2-Fold change in count abundance per unit time over Days 4 to 21 (inclusive of D7, 10, 14), which is a growth rate computed through log-linear regression. In addition, corresponding standard error values for each LFC estimate are produced, which allow for the creation of functional classifications.