Process Design Library Workflow – \*Exact Same for A and H screen

1. Load NGS reads (raw\_data/design\_lib\_Data1.rData) & sequence metadata (raw\_data/barcodes.rds)
2. Merge NGS reads with sequence metadata by each sequences unique barcode
3. Add pseudo-count of 0.5 reads to each time point
4. For each sequence indicate which of the 5 bioreps have starting (day0) NGS reads < 5. If a sequence does not have 2+ bioreps with enough day0 NGS reads its growth slope will not be calculated.
5. For each sequence aggregate all 5 bioreps into one overall **aggregate count** for each timepoint (day0🡪day4).
6. Indicate for each sequence at what time point do **aggregate counts** drop < 3 reads and stay < 3 reads at the next time point.
   1. If this is the case (ex: day3 = 2.8 and day4 = 2.7) then regression will only go to day3
7. Normalize **aggregate counts** by dividing each aggregate count by the total number of NGS reads for that day.
8. Remove sequences previously indicated to not have enough bioreps from analysis
9. Compute log2 fold change from day0 to each subsequent time point
10. Run rlm from MASS package. Output coefficient, converged, and mse.
11. Use stop codons to assign a binary functional-nonfunctional cutoff.
12. Realign slopes from rlm to stop codon cutoff. So that functional sequences all have positive slopes, and non-functional sequences all have negative slopes.

\*\*Additional filtering I have not done.

1. Filter out sequences with mse > 3.5 (based on semi-random library cutoff)
2. Filter out sequences that did not converge in rlm