Last modified: December 2019; link change, file name change
Last reviewed: December 2019

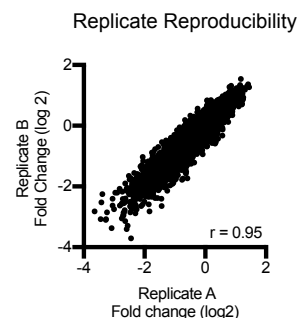# Protocol: Introduction to Pooled Screen Analysis

Using the **conditions** file that you provided and a **reference** file that contains the sequences of the perturbations in the library, we deconvolute the sequencing reads to generate a matrix of read counts that is then provided back to you in the **counts** file. The data are then transformed in the <u>log-norm</u> file by the following formula:

Log_2(Reads from an individual perturbation ÷ Total reads in PCR well * 10^6 + 1)

In addition to the scores file, we also produce a **quality** file that produces a high level overview of the performance of the sequencing run. In particular, pay attention to the total number of reads and matching reads at the top of the file. In general, there should be between 120 – 180 million total reads (sometimes more), and the number of matching reads should be 60 – 90% of the total. You can then also examine each individual well to see the performance of samples across the PCR plate, and identify aberrant samples that may need to be excluded from further analysis.

In some cases, if examination of the quality file merits it, we will produce an **unexpected** file. These are sequences that are found to be abundant in the sample but are not perturbations listed in the reference file. There are always some unexpected sequences (5 – 20%, due to errors in oligonucleotide synthesis during library production, PCR, and Illumina sequencing). But we have seen cases where people have a) used the wrong library; b) contaminated their genomic DNA samples with abundant plasmid DNAs in their labs; c) spiked-in positive controls; d) other. The use of the unexpected file can help to troubleshoot such problems.

All of these files should be used to determine the technical performance of the screen, such as examining the reproducibility of technical and biological replicates. How exactly one does that is beyond the scope of this introduction, but needless to say, it requires thought and is experiment-specific. Once that is done, compare samples and calculate the change in abundance of perturbations from one sample to another, working with the log-normalized data (thus, add / subtract to compare samples, not multiply / divide) to generate the log-fold-change (LFC) values.

The next step of the analysis is to go from a perturbation-level analysis of the data to a gene-level analysis, that is, combine the information from multiple perturbations targeting a gene. Many libraries have perturbations that map to multiple genes (for example, highly homologous genes) and the **chip** file contains the information for this many-to-many mapping problem. Additionally, as the annotation of the genome continues to evolve, perturbations that once mapped to a gene may no longer do so; conversely sometimes perturbations acquire a new target. Thus, it is important to use a chip file that was generated within the past few months (the date of the chip file is encoded in the file name). There can be multiple versions of chip files that vary in their assumptions of how perturbations are mapped to genes, and a full discussion of these differences is beyond the scope of this introduction.

For CRISPR screens, the default chip file ignores the first 3 nucleotides of the sgRNA, as they contribute little specificity to Cas9. Likewise, for RNAi screens, the default chip file ignores the final 2 nts of the sense strand RNA, as these are removed via Dicer processing. As a result, a small minority of genes with highly homologous sequences in the genome will have dozens of perturbations targeting them.

There are seemingly an innumerable number of ways of coming up with a gene-level score, and you do not need to try all of them. It is also important to resist the view that various software packages that can execute this step are black boxes that don't require your understanding.

One such approach we developed, called STARS, can be downloaded here:
http://portals.broadinstitute.org/gpp/public/software/stars

Another useful (and similar in concept) package is MaGECK:
https://sourceforge.net/projects/mageck/

A simple analysis we have started to use involves the generation of volcano plots, as we believe this is a nice way of displaying primary screening data:
https://portals.broadinstitute.org/gpp/public/analysis-tools/crispr-gene-scoring.
Here, the x-axis is the average $\log_2$-fold-change (LFC) of all the perturbations that map to a gene, and thus gives a sense of magnitude. The y-axis is the average $-\log_{10}$ p-value for all the guides targeting a gene. There are multiple ways of calculating such p-values, and the method we use is the hypergeometric distribution without replacement based on the rank order of the LFC of the perturbations; this is equivalent to a one-sided Fisher's exact test. Note, however, that the list can be ranked by LFC either ascending or descending, and the resulting calculations for each gene will not be equivalent. We choose to resolve this problem by calculating the average $-\log_{10}$ p-value both ways and reporting the more significant one. It is important to note that one is testing multiple hypotheses in these calculations, and thus one cannot simply call all genes with p-values < 0.05 as hits. There are multiple methods for correcting for this; additionally, control perturbations and/or shuffling the data can be useful for determining an empirical false discovery rate.