Overview of **DESeq2**'s Differential Expression Analysis

The DESeq2 package from BioConductor performs the following operations on a matrix of raw counts:

1. **Library Normalization**:

   Two properties that can affect the number of reads for a given gene are the **sequencing depth** (i.e. the number of times a base was covered, e.g. due to generation of more fragments in the PCR step of library preparation) and the **gene length** (where larger genes will have more reads mapping to them). These both increase the read counts of a given sample; since they are only due to experimental methodology and not actual biological difference, they must be corrected before comparing genes within and across samples.

	SAMPLE #1	**SAMPLE #2**
GENE	635 reads	635 reads
G-1	30	235
G-2	24	188
G-3	0	0
G-4	563	0
G-5	5	39
G-6	13	102

   In addition to different library sizes, there can be different **library compositions;** e.g. when a gene G-3 is expressed in lower counts in a sample #2, the reads (normally attributed to gene G-3 in other samples like #1) will be distributed to other genes in sample #2, leading to higher counts of other genes in that sample, even though the only truly differentially expressed gene in #2 is G-3 (refer to the table).

   In order to correct these, a **scaling factor** is calculated using the following steps:

   - Find the log (base *e* by default) of all read counts.

- Find the average of these log values for each gene across all available samples (this is the geometric mean, which is resistant to outliers).
- Filter out genes with infinity values in any of the samples (these genes had zero counts; this allows the scaling factors to be calculated from "housekeeping" genes)
- Subtract this average for each gene from each log value for that gene.
- Calculate the median value for each sample.
- Convert this median back to a number by exponentiation to the base used for the log. These are now our scaling factors for each sample.

We can now divide the original count values for each sample by its corresponding scale factor.

In summary, this scaling focuses on the genes expressed on all samples and smooths over the outlier read counts by considering the log values. In addition, the median mitigates the effect of genes that "soak up" a lot of reads, emphasizing on moderately expressed genes.

2. **Independent Filtering:**

As the ratio of the samples drawn from the same distribution to those drawn from different distributions increases, the ratio of *true positives* kept by the FDR (Benjamini-Hochberg method) decreases. That is, even though FDR can limit the number of *false positives,* it will increase the number of *false negatives* as "bogus tests" (such as uninformative read counts of low count genes) are added to the analysis.

One of the factors contributing to this issue are the genes with low read counts, thus filtering these genes with an empirically determined cutoff would decrease the number of false negatives.

- DESeq2 first calculates the significance scores (p-values) for all genes; these are *raw* p-values.
- Then, it calculates a metric called **CPM** (counts per million) for each gene, by:
    - Dividing the total read count of each sample by 1 million, and
    - Dividing the read counts of each sample by this number.
- It then calculates different CPM cutoffs for **significant genes** across all percentiles; that is, it calculates the CPM cutoff value that e.g. would filter 20% of the **significant genes** for having a CPM value below the cutoff. Note that it filters genes by their average across **all** samples, i.e. the gene is dropped only when the average counts of the gene is below the CPM cutoff.
- Since we have the cutoffs for each percentile, DESeq2 first smooths the values by fitting a curve, and then calculates the standard deviation between the raw values and this curve. The cutoff is then the peak of this curve minus the standard deviation. That is, the first percentile that is in the *noise range* of the peak is the CPM cutoff.

If none of the raw values goes above the cutoff, then no filtering is done.