RPKM (Reads Per Kilobase Million) is a common normalization method used in bioinformatics to measure gene expression levels in high-throughput sequencing experiments, such as RNA-seq. RPKM is used to quantify gene expression while accounting for variations in both the total number of reads and the length of the gene.

*RPKM = (Reads mapped to gene) / (Gene length in kilobases \* Total mapped reads in millions)*

**Reads mapped to gene:** The number of sequencings reads that are aligned to a specific gene.

**Gene length in kilobases:** The length of the gene in kilobases (1 kilobase = 1000 base pairs).

**Total mapped reads in millions:** The total number of reads that are successfully aligned to the entire transcriptome or genome, expressed in millions.

The RPKM value provides a normalized measurement of gene expression that accounts for both the gene's length and the total sequencing depth of the experiment. RPKM values can be useful for comparing gene expression levels between different samples or conditions.

Constructing a dataset using RPKM (or its variant, FPKM) values for gene expression can be effective for certain types of analyses, particularly when working with high-throughput sequencing data such as RNA-seq. However, there are several considerations and potential challenges to keep in mind:

**Normalization and Comparison:** RPKM/FPKM values provide a measure of gene expression normalized by the gene's length and the total number of reads. This makes them suitable for comparing expression levels across genes and samples. However, keep in mind that these values are not directly comparable between different samples or experiments due to variations in library preparation, sequencing depth, and other technical factors. Additional normalization methods may be necessary for accurate comparisons.

**Differential Expression Analysis:** RPKM/FPKM values can be used for differential expression analysis to identify genes that are differentially expressed between different conditions. However, specialized statistical methods are required to account for the variability in sequencing data and to control for false positives.

**Data Transformation:** Depending on the downstream analysis, you might need to perform data transformation, such as logarithmic transformation, to achieve a more symmetrical distribution of expression values and meet the assumptions of certain analysis methods.

**Batch Effects:** If your data comes from multiple experiments or sources, batch effects can introduce bias in the analysis. Proper batch correction methods might be needed to remove these effects.

**Feature Selection and Dimensionality Reduction:** High-throughput data can be high-dimensional. Careful feature selection and dimensionality reduction techniques can be helpful in reducing noise and focusing on relevant genes.

**Exploratory Data Analysis:** Visualizing your data through plots, heatmaps, and clustering can help you identify patterns and outliers.

In summary, constructing a dataset using RPKM/FPKM values can be effective for gene expression analysis, but it's important to consider the experimental design, normalization, statistical analysis, and biological context. It's often a good practice to collaborate with experts in bioinformatics or genomics to ensure the appropriate methods are chosen and applied correctly to your specific research question.

New Dataset

**Unstranded:** This term refers to the type of library preparation used in sequencing. In an unstranded library, the sequencing process doesn't differentiate between the two strands of DNA or RNA. This means that the directionality of the original RNA molecules is not preserved during library construction. Unstranded libraries are often used for simple gene expression quantification, but they can't distinguish between sense and antisense transcripts.

**Stranded First:** In a stranded library, the directionality of the original RNA molecule is preserved. Stranded libraries can distinguish between sense and antisense transcripts. The "stranded\_first" column likely contains gene expression values obtained from the first strand of a stranded library.

**Stranded Second:** Like "stranded\_first," the "stranded\_second" column likely contains gene expression values obtained from the second strand of a stranded library.

**TPM (Transcripts Per Million):** TPM is a normalized measure of gene expression that accounts for the length of the gene and the total number of reads in the dataset. It represents the number of RNA-seq fragments mapping to a gene per million total fragments in the library. TPM values allow for comparison of gene expression levels across samples.

**FPKM (Fragments Per Kilobase Million):** FPKM is similar to TPM but also takes into account the length of the gene in kilobases. FPKM values represent the number of RNA-seq fragments mapping to a gene per kilobase of transcript length per million total fragments in the library.

**FPKM UQ (Upper Quartile):** FPKM UQ is an alternative to FPKM that uses the upper quartile normalization. It calculates the expression of a gene relative to the upper 25% most highly expressed genes, which can help mitigate the influence of highly expressed genes on the overall normalization.