Given is the crude gene expression profile for 4 cell types involved in human plasma cell differentiation. The process includes 4 stages of development starting from memory B cells (MBC) to pre-plasmablasts (prePB), plasmablasts (PB) and plasma cells (PC).

Alex, a Scientist wants to understand the biological processes differentiating the 4 stages of plasma cell differentiation and designs an experiment to capture the bulk or total RNA-profile of 3 replicates of each cell type across the 4 stages through RNA-sequencing. Please help Alex understand the data generated from the experiment and the analysis steps to get to the identification of biological processes that differentiate the 4 cell types.

Part I: Data exploration and normalization

Q1: Alex wants to first have a feel of the data generated and wants to look at the size and composition of the data:

i) How many total genes have been captured in the experiment for the 4 different cell types? [HINT: Count the unique GENE_IDs]

ii) Alex knows that there are some gene names that are identical for a subset of the genes even when they have different gene IDs. Identify those and remove them from the data:

a) How many duplicate genes are present [HINT: Count the duplicated GENE_NAMEs]

b) List the top 5 gene types that have the maximum count of duplicated genes names