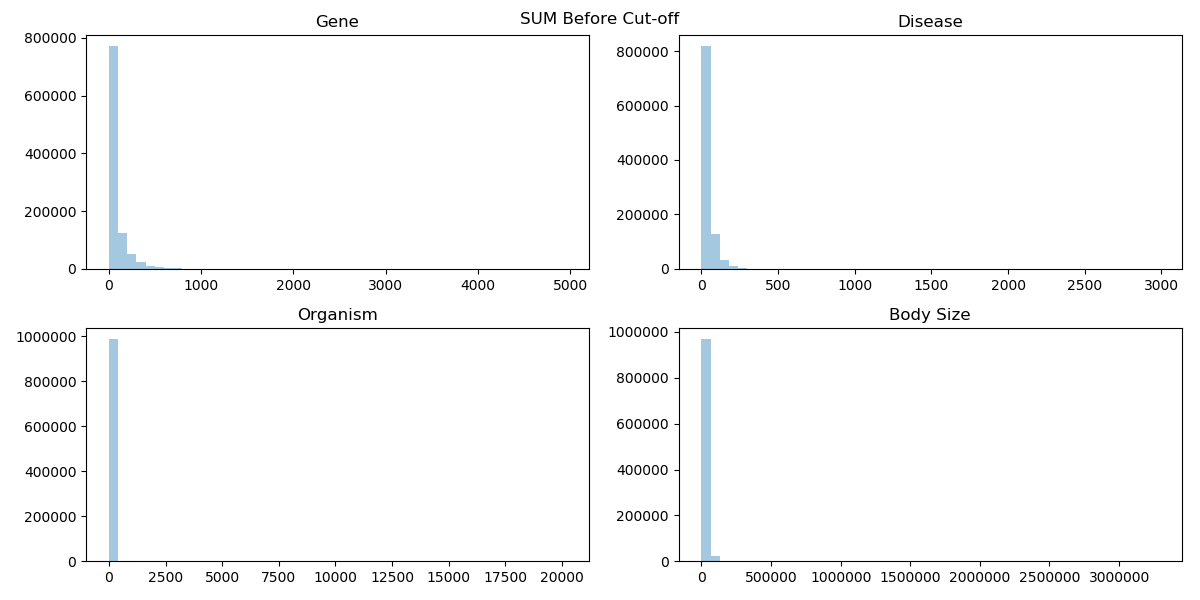
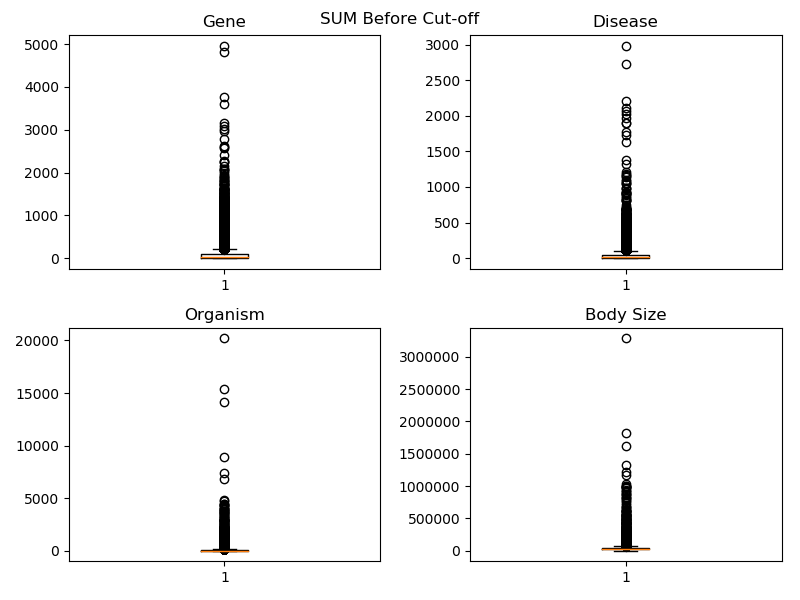
## Training set dev analysis

#### Distributions before any cut-off (considering number of all entity mentions in each type e.g. [p53, p54, p53] counts as 3):

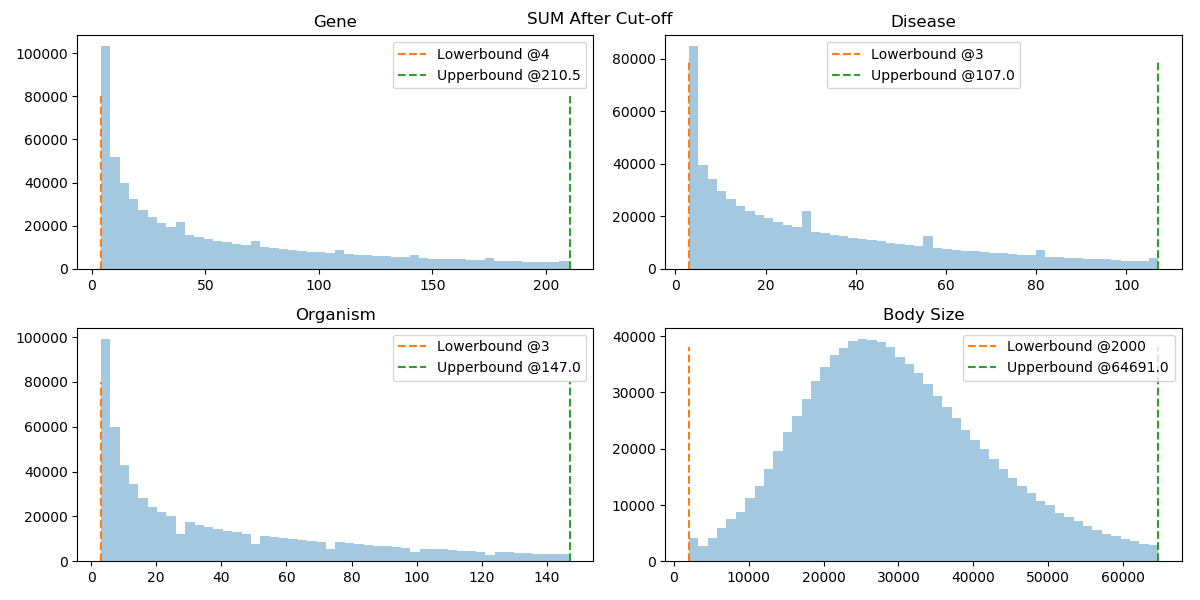
Long tails are observed, means potential outliers with high numbers. Most articles have lower frequency of entity mentions.

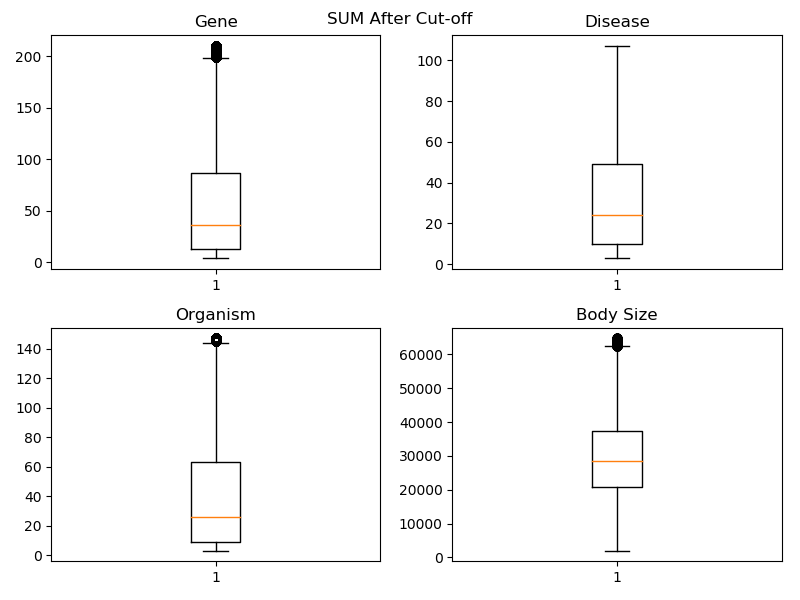




#### Distribution after the cut-off (considering number of all entity mentions in each type e.g. [p53, p54, p53] counts as 3):

By working out the numbers at 25% and 75% percentiles, any number less the 25% percentile (as lower bound) or greater than 75% percentile + 1.5xIQR (interquartile range) is removed. E.g. If 25% percentile is 2 and 75% percentile is 10, any number less than 2 and greater than 10+1.5x(10-2) is removed.



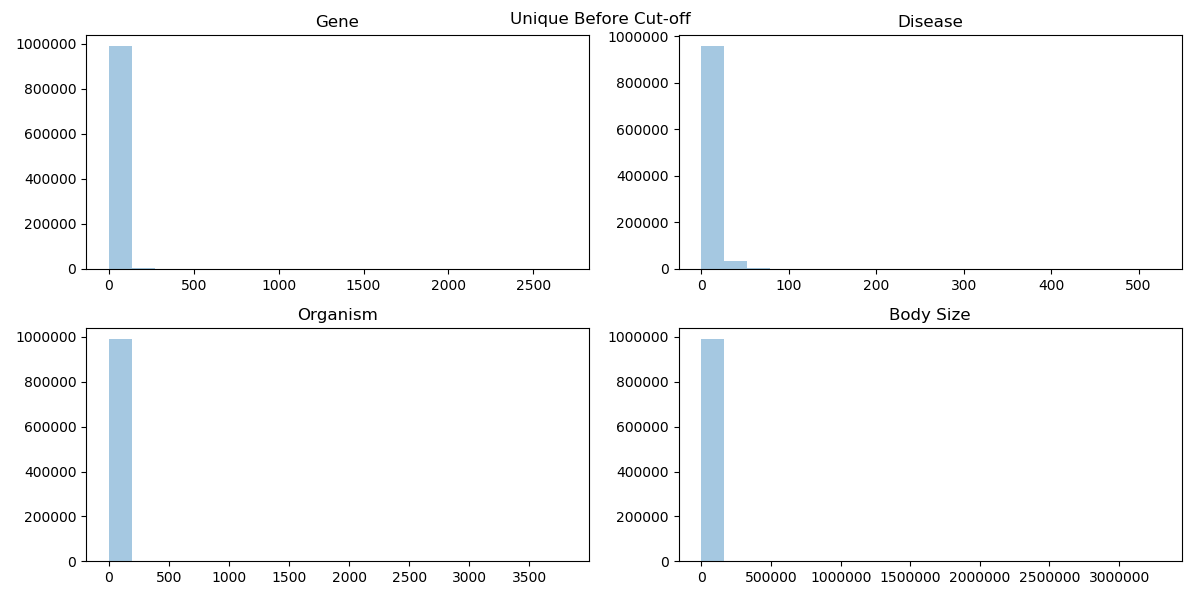


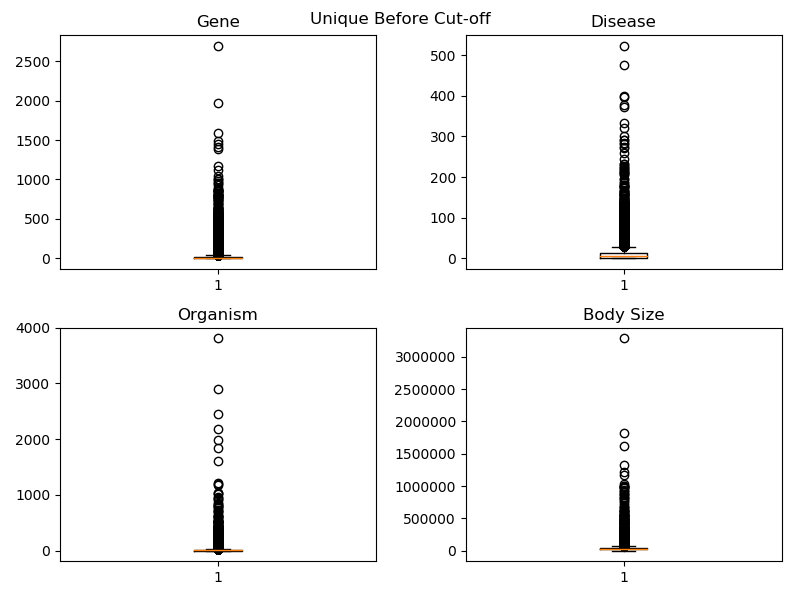
Except the body size, all of them are not normal distribution. The frequency decreases when number of entities increases.

From box plots, data is still skewed. 50% articles mentioned around 10-80 gene names (with duplicates), 10-45 disease names (with duplicates) and 10-60 Organism names (with duplicates).

Distribution before any cut-off (only considering the number of distinct entity names in each type e.g. [p53,p54,p53] counts as 2):

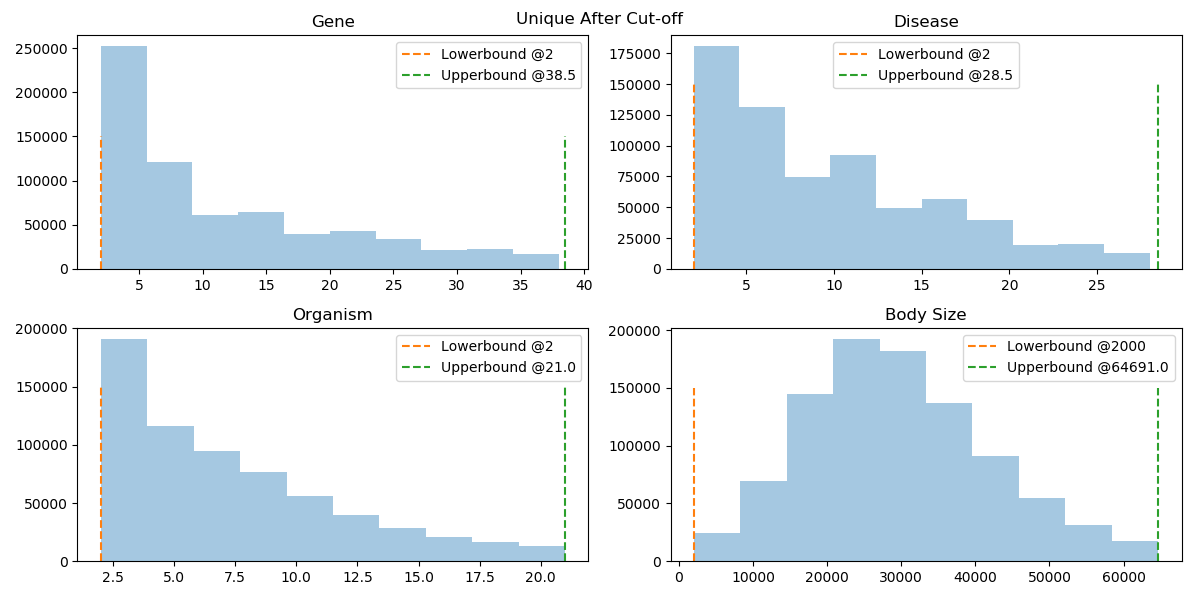
Data is very skewed and most articles have less than 200 entity mentioned without duplicates.

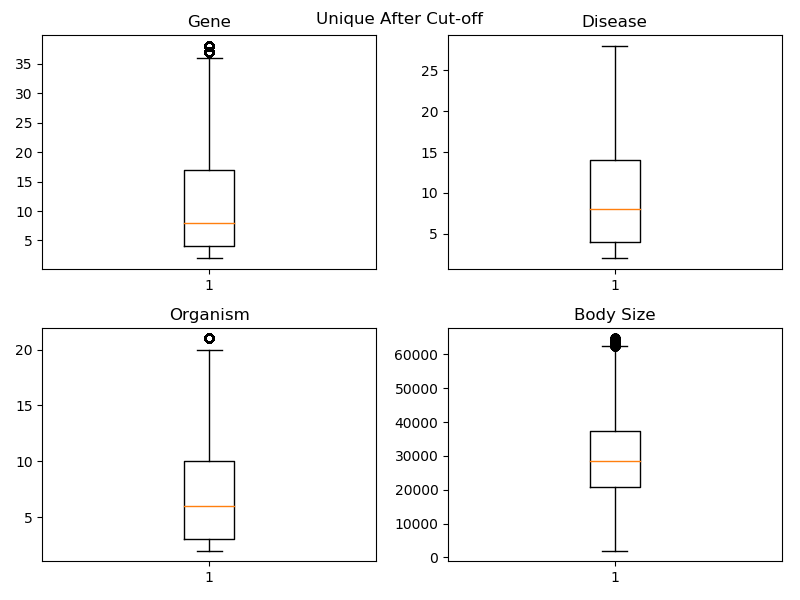




#### Distribution after cut-off (only considering the number of distinct entity names in each type e.g. [p53,p54,p53] counts as 2):

After cut-off at lower and upper bound, they are not normal distribution. Interestingly, 50% articles only have 5-15 mentions (without duplicate) of each entity type.



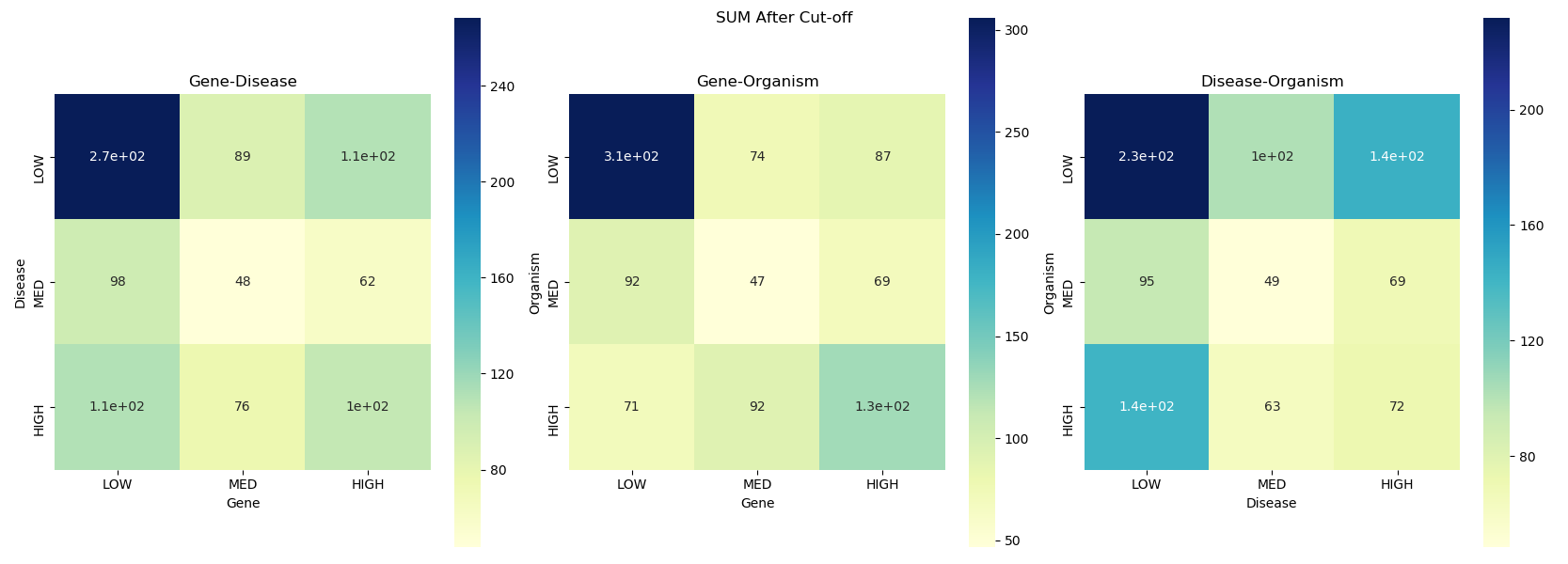


### 

### Co-occurrence of each two types:

For easy look-up, numbers are in 1000 e.g. 48 means 48,000.

Low, Med and High are separated by 33% and 66% percentile of cut-off data.



Similar to Vid’s spreadsheet, LOW-LOW-LOW is always the highest density.