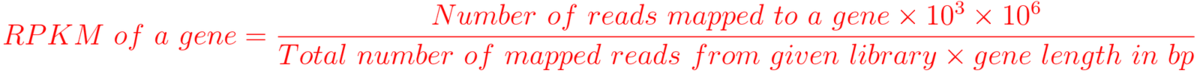
#### **RPKM (Reads Per Kilobase of transcript per Million mapped reads)**

- 추후에 FPKM 설명에서 FPKM과 RPKM에 차이에 대해서 설명하겠지만 과거에 single-end RNA-seq이 주로 생산되었을 때 단순히 read counts in transcripts, gene length, total mapped reads 만을 가지고 계산한 값이다.



- 먼저 total mapped reads로 나누는 이유는 sample에 따라서 total depth가 다른 것을 normalize해주기 위함이다.

sample A는 1,000,000개의 reads가 있고 sample B에는 2,000,000개의 reads가 있을 때, 같은 유전자에 같은 갯수의 reads가 붙었다고하면 sample A가 더 발현량이 높다고 계산하기 위함이다.

(sample B가 PCR cycle이 한 번 더 진행되었기 때문에 전체 depth가 높은 것이라고 가정한 것이다. 단순히 reads count가 같다고 해서 발현량이 같은 것이 아니다!)

- gene length로 나누는 이유는 샘플간의 비교가 아닌 유전자 간의 비교를 위함이다.

gene A는 길이가 1000고 gene B는 길이가 2000일 때 위와 마찬가지로 같은 갯수의 reads가 붙었다면 gene A의 발현량이 더 높다고 추측할 수 있다.

(길이가 2배이고 발현량이 같다면 gene B가 두 배 많은 reads를 생산 되었어야 한다!)

#### **FPKM (Fragments Per Kilobase of transcripts per Million mapped reads)**

- FPKM과 RPKM의 차이는 paired-end로 생산된 RNA-seq에서 나타난다. paired-end는 하나의 reads에서 두 개의 fragments가 나온다고 생각하면 된다. 즉 left fragements와 right fragments가 각각 계산되어 RPKM의 대략 두 배의 값을 가지게 된다.(paired-end의 두 fragments가 항상 같이 맵핑되는 것은 아니기 때문에 정확하게 두 배 일 수는 없다.)

용어가 헷갈릴 수 있는데 일반적으로 paired-end는 reads를 두 개 가지고 있다고 말하는 경우도 있다. 하지만 여기서는 각각을 fragments로 계산하고 있음을 유의해야 한다.

#### **TPM (Transcripts Per Million)**

- TPM에서는 total mapped reads를 사용하지 않으며 transcripts level에서의 값이 계산된다. gene level로 계산하고 싶으면 TPM값들을 모두 더하면 된다.

계산하는 방법은 순서가 중요한데

1. 각각의 transcipt에 대해 mapped reads / transcripts length / killobase 로 값을 구한다. (=normalized transcripts expression)

2. 1번에서 구한 값들을 모두 더한뒤 1,000,000으로 나눈다. (=scaling factor)

3. 각각 1번에서 구한 값들을 2번에서 구한 값으로 나눈다. (=TPM)

아래 reference에 있는 논문에 따르면 RPKM을 사용하는 것 보다 TPM을 사용하는 것이 더 정확한 발현 값을 구할 수 있다고 한다.