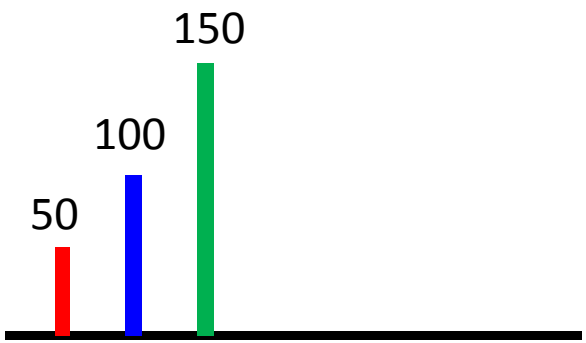
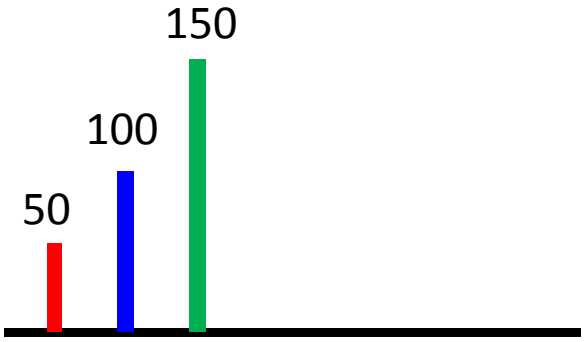
Thought experiment (I)

Suppose

- Two RNA populations (samples): A and B
- The same three genes expressed in both samples
- Numbers indicate number of transcripts / cell



Condition A



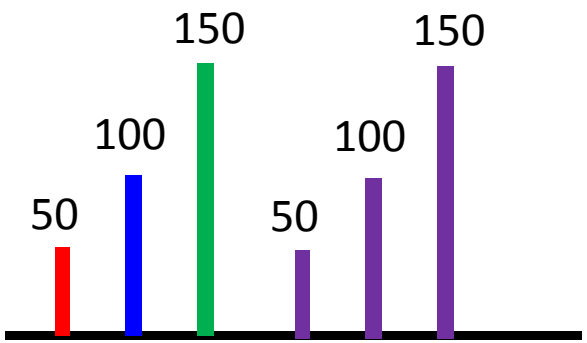
Condition B

No differential expression of these genes

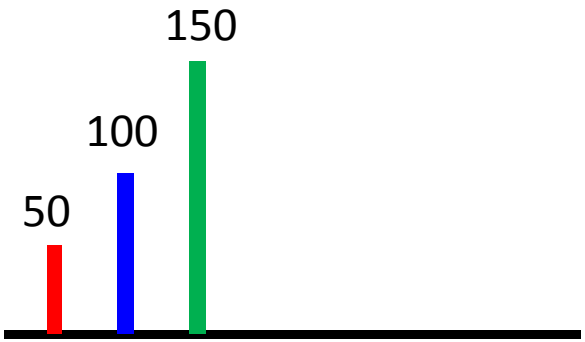
Thought experiment (II)

Suppose

- Two RNA populations (samples): A and B
- The same 3 genes expressed in both samples
- Numbers indicate number of transcripts / cell
- Now condition A has 3 additional genes not in B with equal number and expression



Condition A



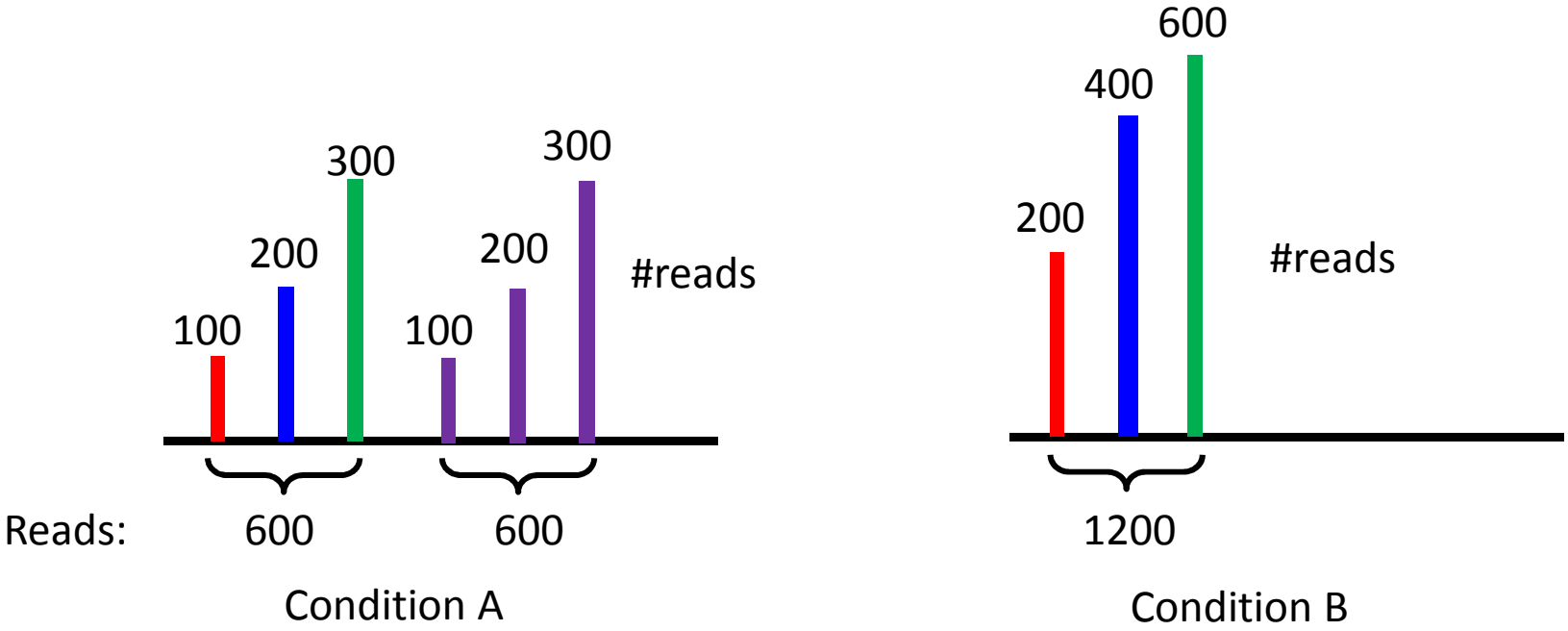
Condition B

Still no differential expression of first three genes. However, RNA production in A is twice the one in B.

Thought experiment (III)

Suppose we sequence both samples with the same depth (1200 reads)

These reads get 'distributed' over the expressed genes



(1) Correct normalization factor would adjust condition A by a factor of two

(2) Proportion of reads attributed to a gene in one library depends on expression properties of whole sample ➔ If a sample has larger RNA output , RNA-seq will undersample many genes

RPKM would fail in this example

$$\text{RPKM} = \frac{\#\text{ mapped reads} \times 10^6}{\text{total number of mapped reads}}$$

(assuming transcript lengths are the same)

In this example:

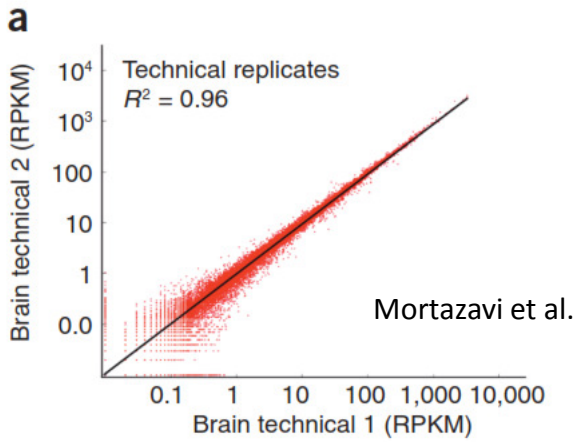
Condition A, first (red) gene: $$RPKM = \frac{100 \times 10^6}{1200} = 83333$$

Condition B, first (red) gene: $$RPKM = \frac{200 \times 10^6}{1200} = 166666$$

RPKM normalization would result in differential expression: we did not take total RNA production into account

When does RKPM fail?

- If samples have largely different RNA production
  - Many unique genes and/or highly expressed genes
  - If many genes in one sample have a very high expression compared to the other samples
- If RNA sample is contaminated
  - Reads that represent the contamination will take away reads from the true sample, thus dropping the number of reads of interest.
- If you can assume that your samples are ‘comparable’ then RKPM is OK
  - e.g., technical replicates



Mortazavi et al.