RPKMとTPM

参考：

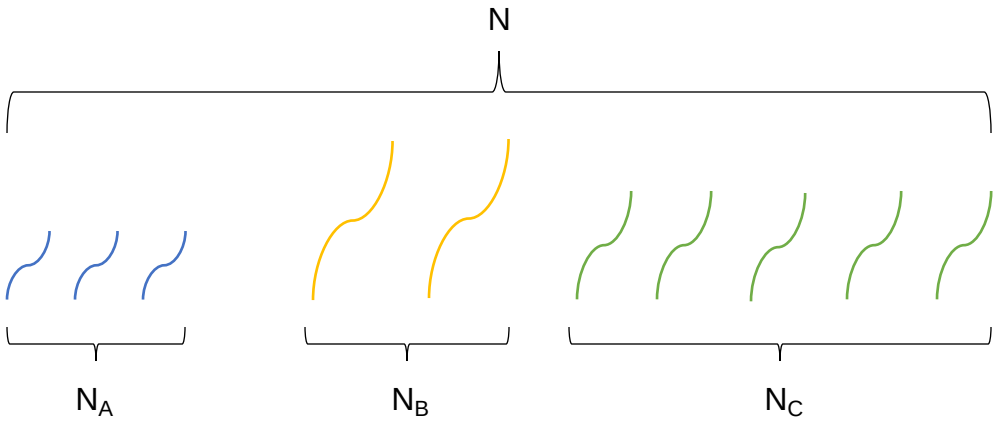
Wagner, Günter P., Koryu Kin, and Vincent J. Lynch.

"Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples."

*Theory in biosciences*131.4 (2012): 281-285.

RNA-seqの（理想的な）正規化

遺伝子集合を G := {geneA, geneB, geneC, …} とする



理想的なmRNAの「存在量」は（モル濃度のように）発現しているmRNAの個数の、全mRNA中の割合

$$abundance_i = \frac{N_i}{N} = \frac{N_i}{\sum_{i \in G} N_i}$$

このとき、すべての遺伝子で存在量を平均すると、

$$average\_abundance_i = \frac{1}{|G|} \sum_{i \in G} \frac{N_i}{\sum_{i \in G} N_i} = \frac{1}{|G|}$$

となって、（リファレンスのゲノムが同じである限り）サンプルによらず一定の値となる。

RPKM正規化

r_i … 遺伝子 i にマッピングされたリード数。 i ∈ G
L_i … 遺伝子 i の長さ（bp）

$$RPKM_i = r_i \times \frac{10^3}{L_i} \times \frac{10^6}{\sum_{i \in G} r_i} = \frac{10^9 r_i}{L_i \sum_{i \in G} r_i}$$

RPKMは、遺伝子の長さによる因子と、トータルリード数（$\sum_{i \in G} r_i$）で補正する。
この値はどの程度、遺伝子の存在量を反映しているのか？
遺伝子セット全体の平均RPKMを計算してみると、

$$average\_RPKM_i = \frac{1}{|G|} \sum_{i \in G} \frac{10^9 r_i}{L_i \sum_{i \in G} r_i} = \frac{10^9}{|G| \sum_{i \in G} r_i} \sum_{i \in G} \frac{r_i}{L_i}$$

この値は、サンプルによってバラバラ。
（定数部分はいいが、$\sum_{i \in G} r_i$やr/Lはサンプルによって違う）

なぜこうなってしまうのか？
問題点は、トータルリード数の補正をする際に、単純に$\sum_{i \in G} r_i$で割っているから。
トータルのリード数はけっして、N （前ページスライド、実験に用いたmRNA全体の個数）と比例しない。
長いmRNAからはリードがシーケンスされやすいのだから、発現しているmRNA全体の長さの分布によって、同じ個数のmRNAから得られるトータルリード数は異なる。
（N個のmRNAをRNA-seq実験に使ったサンプルであっても、長い遺伝子が多く発現しているサンプルではトータルリード数が多く、短い遺伝子が多く発現していたらトータルリード数は少なくなる。真に補正すべきはNなのに、トータルリード数で補正するとずれてしまう）

RPKMは、長さ補正の項はよかったが、トータルリード数補正で長さの分布を考慮していなかった。それを解決するのがTPM

TPM正規化

r_i … 遺伝子 i にマッピングされたリード数。 i ∈ G
L_i … 遺伝子 i の長さ（bp）

$$TPM_i = r_i \times \frac{1}{L_i} \times \frac{10^6}{\sum_{i \in G} \frac{r_i}{L_i}} = \frac{10^6 r_i}{L_i \sum_{i \in G} \frac{r_i}{L_i}}$$

トータルリード数で割り算するのではなく、まず遺伝子の長さで調整した値を計算し、
その和で補正する。
遺伝子セット全体の平均TPMを計算してみると、

$$average\_TPM_i = \frac{1}{|G|} \sum_{i \in G} \frac{10^6 r_i}{L_i \sum_{i \in G} \frac{r_i}{L_i}} = \frac{10^6}{|G|}$$

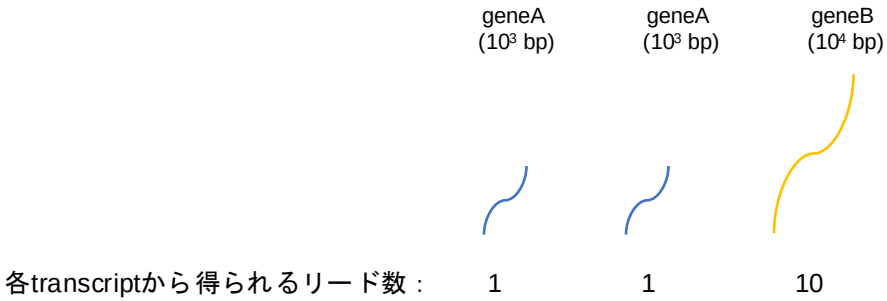
この値は（リファレンスのゲノムが同じである限り）サンプルによらず一定の値となる。
=> サンプル間で比較する際に、より適切な値（遺伝子存在量）となっている

極端な例

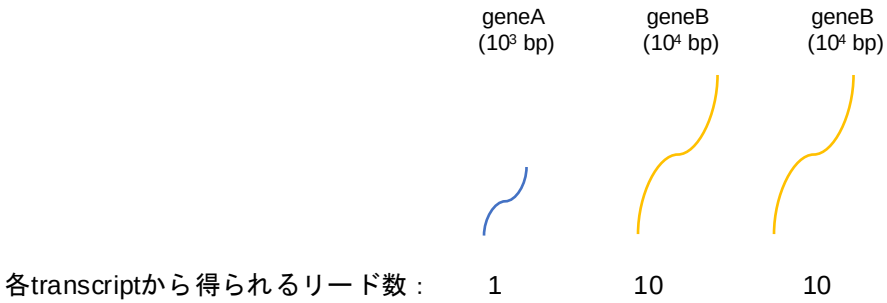
geneA（長さ：1kbp）とgeneB（長さ：10kbp）の２つしかない系を考える。
シーケンスは、1kbpあたり1リード得られる、とする。

２つの異なる状態のサンプルでRNA-seq実験をする。

１）geneAが2個発現、geneBが1個発現（得られるトータルリード数は12）



２）geneAが1個発現、geneBが2個発現（得られるトータルリード数は21）



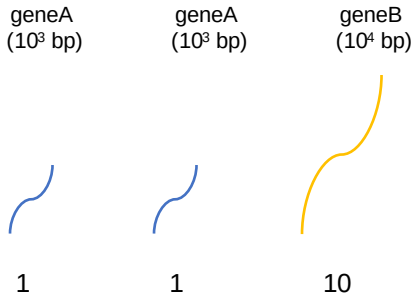
極端な例

各サンプルのRPKMを計算する。

１）geneAが2個発現、geneBが1個発現（得られるトータルリード数は12）

$$RPKM_A = 2 \times \frac{10^3}{10^3} \times \frac{10^6}{12} = \frac{2}{12} \times 10^6$$

$$RPKM_B = 10 \times \frac{10^3}{10^4} \times \frac{10^6}{12} = \frac{1}{12} \times 10^6$$

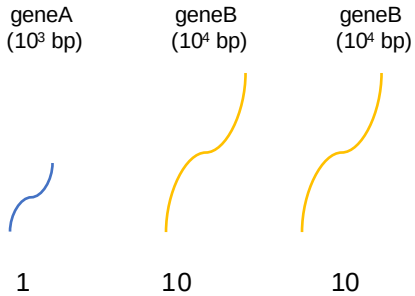


このサンプルの中だけで見れば、たしかにAはBの２倍の値になっていて、よく補正できてそうに思える。

２）geneAが1個発現、geneBが2個発現（得られるトータルリード数は21）

$$RPKM_A = 1 \times \frac{10^3}{10^3} \times \frac{10^6}{21} = \frac{1}{21} \times 10^6$$

$$RPKM_B = 20 \times \frac{10^3}{10^4} \times \frac{10^6}{21} = \frac{2}{21} \times 10^6$$



このサンプルの中だけで見れば、たしかにAはBの半分の値になっていて、よく補正できてそうに思える。

しかし、実験１と実験２のサンプル間を比較すると、
Aのfold changeは、(1/21) / (2/12) = 0.29… 、Bのfold changeは、(2/21) / (1/12) = 1.14…
となり、サンプル間での量的変化のスケールが実際の変動を全然反映していない。

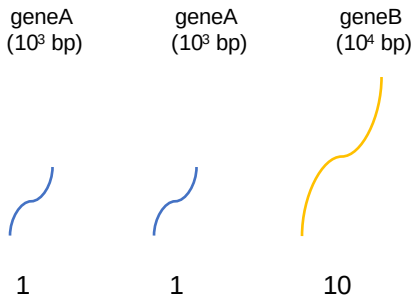
極端な例

各サンプルのTPMを計算する。

１）geneAが2個発現、geneBが1個発現（得られるトータルリード数は12）

$$TPM_A = 2 \times \frac{1}{10^3} \times \frac{10^6}{\left(\frac{2}{10^3} + \frac{10}{10^4}\right)} = \frac{2}{3} \times 10^6$$

$$TPM_B = 10 \times \frac{1}{10^4} \times \frac{10^6}{\left(\frac{2}{10^3} + \frac{10}{10^4}\right)} = \frac{1}{3} \times 10^6$$

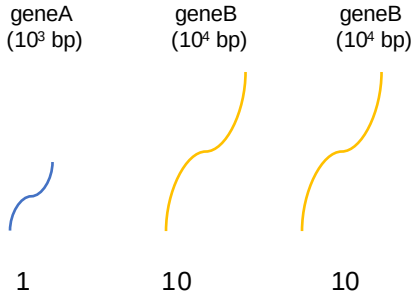


まさに、TPMの名前の通り（Transcripts per million）の値となっている。

２）geneAが1個発現、geneBが2個発現（得られるトータルリード数は21）

$$TPM_A = 1 \times \frac{1}{10^3} \times \frac{10^6}{\left(\frac{1}{10^3} + \frac{20}{10^4}\right)} = \frac{1}{3} \times 10^6$$

$$TPM_B = 20 \times \frac{1}{10^4} \times \frac{10^6}{\left(\frac{1}{10^3} + \frac{20}{10^4}\right)} = \frac{2}{3} \times 10^6$$



TPMは、サンプル内での相対的な比較が正確であるだけでなく、
サンプル間で比較した際にも正確な量的変動を捉えていて、遺伝子間でその変動量を比較できる