5オ2回

◎ RNA-Seq の実験デザインについて

◍ replicate

RNA-Seq で知りたいこと ⇒ 異なる条件下で(RNA)遺伝子発現がどう変化するか
自分の想定する条件 & その他の変動

(当たり前だけど) 同じ群内のデータを複数個使った方が正確に判定できる.

replicate : "同一群内" から取ってきたデータ

— Technical replicates : "同一"のサンプルから取ってきたデータ
ENCODE の定義 "different library preparations from the same RNA sample"

— Biological replicates : (生物学的な多様性を捉えたような) 異なるけど"同じ群"とするサンプルのデータ
ENCODE の定義 "RNA from an independent growth of cells/tissue"
しかし, 何を"多様性の素"と考え, 何を"ノイズの素"と考えるかは実験ごとに変わる

ex.) 同じ細胞から異なる Kit で用意した Run = Technical
一個一個の細胞 = Biological
同じ組織から回収したサンプルを = Technical ⇐ 平均的に"均一"と考えるからか…?
異なる分類(分離)方法で回収したもの

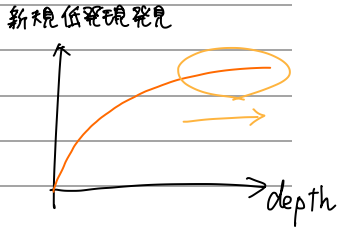
性別・集団・細胞株・臓器 … = Biological

◍ Sequencing depth

Sequencing depth (Coverage) : ゲノムの各塩基が何回読まれるか

$$\frac{\text{リード長} \times \text{リード数}}{\text{1倍体ゲノムサイズ}}$$ the Lander-Waterman equation

もちろん, クロマチン構造なり AT含率で変わるけどね



Transcriptome においては, 元の発現量に依存して変化する.
しかし事前に"最低限必要なリード"を知るのは不可能!!
⇒ ガイドラインや size, error 率, 問題ごとに適度に考える必要性

おやみに sequencing depth を増やしても, コストが増えるだけで, 低コピーの転写物をより検出できる訳ではない

Biological replicates を増やすと, 再現性は上がるが発現量の高いもののみ検出することとなる
トレードオフ (コスト的意味)

② 正規化

各サンプルのリード数を揃える ⇒ 比較が可能となる

・基本：CPM (count per million)
総リード数を 100万 (1e6) に揃える

・1つの遺伝子の発現量を複数サンプル間で比較したい ⇒ 発現変動遺伝子

TMM：非発現遺伝子の発現量が同じになるようにする.

Y_i^g：リードカウント
N_i：総リード数

(i) 各遺伝子の M と A を求める.

$$M_{12}^g = \log \frac{Y_1^g}{N_1} - \log \frac{Y_2^g}{N_2}$$

$$A_{12}^g = \left(\log \frac{Y_1^g}{N_1} + \log \frac{Y_2^g}{N_2}\right)/2$$

(ii) 全M値のうち 30%〜70% を，全A値のうち 5〜95%を用いる（非発現変動と考える）

(iii) 正規化係数を求める

edgeR

ライブラリー j を基準として

$$n_{ij} = \frac{\sum_g w_{ij}^g M_{ij}^g}{\sum_g w_{ij}^g}$$

$$w_{ij}^g = \left(\frac{1}{Y_i^g} - \frac{1}{N_i}\right) + \left(\frac{1}{Y_j^g} - \frac{1}{N_j}\right)$$

漸近分散の逆数 (delta method)

どっちも欠損値に弱い！

DESeq 2

RLE：size factor による正規化

$$S_j = \text{median}\left(\frac{C_j^g}{(\prod_n^m C_n^g)^{1/m}}\right) \quad \Rightarrow \quad q_{ij}^g = C_j^g / S_j$$

分子：算術平均
分母：幾何平均

・複数遺伝子間

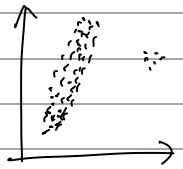
RPKM/FPKM

TPM

⑥ 統計的検定

◍ Spearman 相関係数

普通の相関係数（Pearson）は母集団に正規分布を仮定している
今回のデータでは"外れ値"が結構あって影響されやすい



Pearson = 0.67
Spearman = 0.84

Non parametric な方法 ⇒ Spearman's rank correlation coefficient
順番のデータさえあれば良い（データを順位に変換する）

$$\rho = 1 - 6\Sigma D^2/(N^3-N)$$

順位差　　点（ペア）の数