Advanced NGS Analysis (Day 1)
Session II

Lyda Hill department of Bioinformatics 2022 Nanocourse Series

Date & Time: June 27-28: 9AM-5PM (NG3.202)

Course Instructors: Bo Li, Daehwan Kim, Christopher Chaney, & **Micah Thornton**

UTSouthwestern
Medical Center

Day 2: RNA-Seq Analysis Using Pseudo/Quasi-Alignment and Expectation Maximization (Kallisto, Salmon, H2Q).

UTSouthwestern Medical Center

What you will learn in this Session (2 Parts)

- *Some Theoretical Considerations:*
  - *What is Pseudo/Quasi-Alignment*
    - *What is Alignment*
  - *What is Expectation Maximization*
  - *What is Expectation Maximization for Gene Transcript Quantification*
  - *What is the resolution of Genetic Transcript Data?*

- *Some Practical Considerations:*
  - *What is Kallisto*
  - *What is Salmon*
  - *What is H2Q*

UTSouthwestern
Medical Center

Part 1: Pseudo and Quasi Alignment & Quantification Resolution

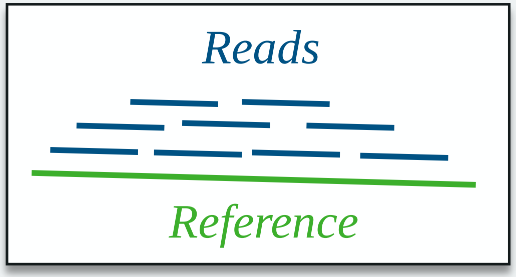
RNA-Seq Experiments

- RNA-Seq experiments are fundamentally distinct from DNA-seq experiments, and seeks to answer a different set of questions.
- Usually we are seeking to determine whether the level of expression of a particular gene is related to a phenotypic characteristic of interest.

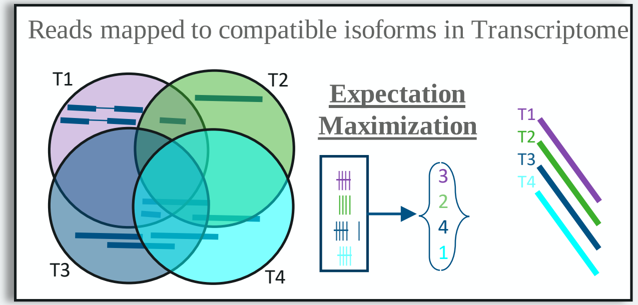
Pseudo/Quasi Alignment in RNA Experiments

- Sometimes the *exact* position of a sequencing read is not of critical import.
  - There are a few approaches for resolving the *approximate* location of a read.
  - Procedures work by determining the subset of *transcript isoforms* compatible with a read.
  - Two such approaches are known as:
    - Pseudo-Alignment
      - The Approach used by **Kallisto.**
      - Uses the De Brujin ('Deh-Broine') graph procedure.
    - Quasi-Alignment
      - The Approach used by **Salmon.**
      - Uses a K-mer Hash table and Suffix Array.

Typical 'DNA-Seq Like' Experiment



Typical 'RNA-Seq Like' Experiment



Recall that in most typical sequencing experiments we are dealing with a large collection of shorter subsequences called ***reads***, which we attempt to map to a larger sequence known as the ***reference***.

Resources – Kallisto (Pseudo-alignment)

1. https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html (Higher Level Overview pseudo alignment)
2. https://www.youtube.com/watch?v=f-ecmECK7lw (Video Describing how To Build The De Brujin graph)
3. https://www.nature.com/articles/nbt.2023 (Nature Primer on Using De Brujin Graphs for Genomic Alignments).

Resources – Salmon (Quasi-alignment)

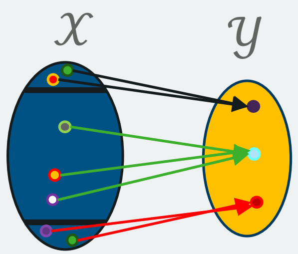
1. https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon-flipped/lessons/08_quasi_alignment_salmon.html (Higher Level Overview Quasi-Alignment)
2. https://academic.oup.com/bioinformatics/article/32/12/i192/2288985?login=true (RapMap Paper and Description).

UTSouthwestern Medical Center

Part 2: Expectation Maximization & Gene Transcript Quantification

Expectation Maximization (in general) – Incomplete Data & A Restricted Case

Many-to-one relationship



- Two general uses include:
  - determination of maximum likelihood estimates for parameters when missing data is present and
  - estimation of missing or otherwise incomplete data.

- In general, suppose that we would like to observe the values, $x_1, x_2, \ldots x_n$, to determine something about the parameters of the random variable X which has sample space $\mathcal{X}$ as shown (top right).
  - However, we are only able to observe, $y_1, y_2, \ldots y_n$, valuations of the random variable Y which has sample space $\mathcal{Y}$ onto which there exists a many-to-one mapping from $\mathcal{X}$.
    - In other words, there are multiple values possible to observe in $\mathcal{X}$ corresponding to the same value in $\mathcal{Y}$.

- Suppose, at first, that the distribution of $\boldsymbol{X}$ (note boldface indicates that $\boldsymbol{X}$ could be a vector quantity) is one of the exponential family of distributions generally denoted,

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = b(\boldsymbol{x})e^{(\boldsymbol{\theta t}(\boldsymbol{x})^{\boldsymbol{T}})}a(\boldsymbol{\theta})^{-\mathbf{1}}$$

$\boldsymbol{\theta}$ is a parameter [column]-vector (of size r).
$\boldsymbol{t}(\boldsymbol{x})^{\boldsymbol{T}}$ is the sufficient statistic [row]-vector (of size r).
$a(\cdot), b(\cdot)$, are any arbitrary function.
e is the natural number.

See section II of the Dempster, Laird, Rubin paper mentioned below for more details about natural parameters.

These Expectation Maximization Notes Draw Heavily from “Maximum Likelihood from Incomplete Data via the EM Algorithm” by Dempster Rubin and Laird (https://www.jstor.org/stable/2984875)

Expectation Maximization (in general) – The Algorithm

- The “simple characterization” of the EM algorithm according to Dempster, Laird, and Rubin (DLR77) is:

(1) With $\theta^{(p)}$ indicating the estimate of θ at the p^{th} step of the algorithm, estimate the complete-data sufficient statistics $\boldsymbol{t}(\boldsymbol{x})$ by finding

$$\boldsymbol{t}^{(p)} = E\left(\boldsymbol{t}(\boldsymbol{x}) \middle| \boldsymbol{y}, \boldsymbol{\theta}^{(p)}\right).$$

(2) Perform maximum likelihood estimation to determine $\theta^{(p+1)}$ from $\boldsymbol{t}^{(p)}$,

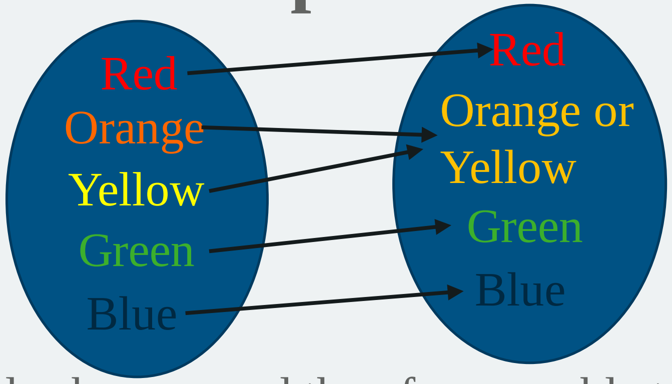
$$\boldsymbol{E}(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\theta}) = \boldsymbol{t}^{(p)}.$$

- Proof of convergence to the maximum likelihood value is given the DLR77, as are details regarding further generalizations of the expectation maximization algorithm.
- The algorithm is broadly applicable in many cases, and not all of the applications have been discovered yet.

UTSouthwestern Medical Center

These Expectation Maximization Notes Draw Heavily from “Maximum Likelihood from Incomplete Data via the EM Algorithm” by Dempster Rubin and Laird (https://www.jstor.org/stable/2984875)

Expectation Maximization (in general) – A Multinomial Example



- Suppose that there are marbles of five colors in a bag.
  - Red marbles are denoted by ‘R’
  - Orange marbles are denoted by ‘O’
  - Yellow marbles by ‘Y’
  - Green marbles by ‘G’
  - Blue marbles by ‘B’
- Now, you personally cannot tell a difference between the orange and the yellow marbles by eye, and therefore are able to produce counts of four categories of marbles only (that is: “Red”, “Orange or Yellow”, “Green”, and “Blue”).

[EXAMPLE]

- Suppose it is known ahead of time that the proportions of the *actual* colors of each of the marbles are related via an unknown parameter π, such that for the unobservable true color of an arbitrarily selected marble i, denoted c_i (true color) given below induces a distribution on the observable o_i (observed color) follows this distribution:

$$P\left(c_i = \begin{pmatrix} \text{Red} \\ \text{Orange} \\ \text{Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}\right) = \begin{pmatrix} (1-\pi)/4 \\ \pi/4 \\ 1/2 \\ (1-\pi)/4 \\ \pi/4 \end{pmatrix} \Rightarrow P\left(o_i = \begin{pmatrix} \text{Red} \\ \text{Orange or Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}\right) = \begin{pmatrix} (1-\pi)/4 \\ 1/2+\pi/4 \\ (1-\pi)/4 \\ \pi/4 \end{pmatrix}$$

UTSouthwestern Medical Center

These Expectation Maximization Notes Draw Heavily from “Maximum Likelihood from Incomplete Data via the EM Algorithm” by Dempster Rubin and Laird (https://www.jstor.org/stable/2984875)

Expectation Maximization (in general) – A Multinomial Example (Continued)

Suppose that we observe 197 marbles, and arrive at the following counts:

$$\begin{pmatrix} \text{R-Red: 18} \\ \text{OY-Orange or Yellow:125} \\ \text{G-Green: 20} \\ \text{B – Blue: 34} \end{pmatrix}$$

$$P\left(c_i = \begin{pmatrix} \text{Red} \\ \text{Orange} \\ \text{Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}\right) = \begin{pmatrix} (1-\pi)/4 \\ \pi/4 \\ 1/2 \\ (1-\pi)/4 \\ \pi/4 \end{pmatrix} \Rightarrow P\left(c_i = \begin{pmatrix} \text{Red} \\ \text{Orange or Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}\right) = \begin{pmatrix} (1-\pi)/4 \\ 1/2+\pi/4 \\ (1-\pi)/4 \\ \pi/4 \end{pmatrix}$$

$$x_j = \sum_{i=1}^{197} \mathbb{1}(c_i \equiv j) \qquad j \in \begin{pmatrix} \text{Red} \\ \text{Orange} \\ \text{Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}$$

$$y_t = \sum_{i=1}^{197} \mathbb{1}(o_i \equiv t) \qquad t \in \begin{pmatrix} \text{Red} \\ \text{Orange or Yellow} \\ \text{Green} \\ \text{Blue} \end{pmatrix}$$

- Let the *actual* color counts be denoted by the values $(x_1, x_2, x_3, x_4, x_5)$ such that x_1 corresponds to the count of marbles which were actually red, x_2 to those which were Orange, and so on…
- Let the observed color counts be denoted by the values (y_1, y_2, y_3, y_4) which are given in this example as (18,125,20,34).
- Furthermore, it is known that $y_2 = x_2 + x_3$.

- The Likelihood on π for the full data can be expressed as:

$$f(\boldsymbol{x}|\pi) = \frac{\left(\sum_{i=1}^{5} x_i\right)!}{\prod_{i=1}^{5}(x_i!)} \cdot \left(\frac{1-\pi}{4}\right)^{x_1} \cdot \left(\frac{\pi}{4}\right)^{x_2} \cdot \left(\frac{1}{2}\right)^{x_3} \cdot \left(\frac{1-\pi}{4}\right)^{x_4} \cdot \left(\frac{\pi}{4}\right)^{x_5}$$

- The *coarsened/incomplete* Likelihood on π for the full data can be expressed as:

$$g(\boldsymbol{y}|\pi) = \frac{\left(\sum_{i=1}^{4} y_i\right)!}{\prod_{i=1}^{4}(y_i!)} \cdot \left(1-\frac{\pi}{4}\right)^{y_1} \cdot \left(\frac{1}{2}+\frac{\pi}{4}\right)^{y_2} \cdot \left(1-\frac{\pi}{4}\right)^{y_3} \cdot \left(\frac{\pi}{4}\right)^{y_4}$$

UTSouthwestern Medical Center

DLR77 (https://www.jstor.org/stable/2984875)

Expectation Maximization (in general) – A Multinomial Example (E-Step)

- Clearly, due to the fact that a marble cannot *actually* be two colors simultaneously, there is no probability that any marble is *truly* both orange and yellow at the same time, therefore we may express the probability that a marble is orange or yellow as follows:

$$P(o_i = (\text{Orange or Yellow})) = P\left(c_i \in \binom{\text{Orange}}{\text{Yellow}}\right) = P(c_i = \text{Orange}) + P(c_i = \text{Yellow}) - P(c_i = \text{Yellow} \;\&\; c_i = \text{Orange})$$

$$= P(c_i = \text{Orange}) + P(c_i = \text{Yellow}) - 0 \;= \frac{\pi}{4} + \frac{1}{2}$$

- From here we can derive the expression for the maximum likelihood estimates of the unobserved counts for orange and yellow marbles (x_2, x_3) in terms of the observed count of "orange or yellow" marbles (y_2).

$$P(c_i = \text{Orange} \,|o_i = (\text{Orange or Yellow})) = \frac{P(o_i = (\text{Orange or Yellow}) \;\&\; c_i = \text{Orange})}{P(o_i = (\text{Orange or Yellow}))} = \frac{P(c_i = \text{Orange})}{P\left(o_i = (\text{Orange or Yellow})\right)} = \frac{\frac{\pi}{4}}{\frac{\pi}{4} + \frac{1}{2}}$$

$$P(c_i = \text{Yellow} \,|o_i = (\text{Orange or Yellow})) = \frac{\frac{1}{2}}{\frac{\pi}{4} + \frac{1}{2}}$$

Therefore the conditional expectation of x_2 and x_3 are:

$$E(x_2|y_2) = y_2 \frac{\frac{\pi}{4}}{\frac{\pi}{4} + \frac{1}{2}} \quad \text{- and -} \quad E(x_3|y_2) = y_2 \frac{\frac{1}{2}}{\frac{\pi}{4} + \frac{1}{2}}$$

Suppose that we observe 197 marbles, and arrive at the following counts:

$$\begin{pmatrix} \text{R-Red: 18} \\ \text{OY-Orange or Yellow:125} \\ \text{G-Green: 20} \\ \text{B – Blue: 34} \end{pmatrix}$$

These Expectation Maximization Notes Draw Heavily from "Maximum Likelihood from Incomplete Data via the EM Algorithm" by Dempster Rubin and Laird (https://www.jstor.org/stable/2984875)

UTSouthwestern Medical Center

Expectation Maximization (in general) – A Multinomial Example (M-Step)

- Recall that the full likelihood for the multinomial distribution was given by:

$$f(\boldsymbol{x}|\pi) = \frac{\left(\sum_{i=1}^{5} x_i\right)!}{\prod_{i=1}^{5}(x_i!)} \cdot \left(\frac{1-\pi}{4}\right)^{x_1} \cdot \left(\frac{\pi}{4}\right)^{x_2} \cdot \left(\frac{1}{2}\right)^{x_3} \cdot \left(\frac{1-\pi}{4}\right)^{x_4} \cdot \left(\frac{\pi}{4}\right)^{x_5}$$

$$\Rightarrow \log L(\pi|\boldsymbol{x}) = \log \frac{\left(\sum_{i=1}^{5} x_i\right)!}{\prod_{i=1}^{5}(x_i!)} + x_1 \log\left(\frac{(1-\pi)}{4}\right) + x_2 \log\left(\frac{\pi}{4}\right) + x_3 \log\left(\frac{1}{2}\right) + x_4 \log\left(\frac{(1-\pi)}{4}\right) + x_5 \log\left(\frac{\pi}{4}\right)$$

- In this example, only x_2 and x_3 are unobservable, the rest are known:

$$\Rightarrow \frac{\partial \log L(\pi|\boldsymbol{x})}{\partial\, \pi} = x_1\left(\frac{4}{1-\pi}\right)\left(-\frac{1}{4}\right) + x_2\left(\frac{4}{\pi}\right)\left(\frac{1}{4}\right) + x_4\left(\frac{4}{1-\pi}\right)\left(-\frac{1}{4}\right) + x_5\left(\frac{4}{\pi}\right)\left(\frac{1}{4}\right) = \frac{x_1}{\pi-1} + \frac{x_2}{\pi} + \frac{x_4}{\pi-1} + \frac{x_5}{\pi} =$$

$$\frac{(x_1\pi + x_2(\pi-1) + x_4\pi + x_5(\pi-1))}{(\pi^2-\pi)} = \frac{(x_1+x_4)\pi + (x_2+x_5)(\pi-1)}{\pi^2-\pi}$$

Suppose that we observe 197 marbles, and arrive at the following counts:

$$\frac{\partial \log L(\pi|\boldsymbol{x})}{\partial\, \hat{\pi}} = 0 \Rightarrow -\frac{x_1+x_4}{x_2+x_5} = \frac{\hat{\pi}}{\hat{\pi}-1} \Rightarrow 1 + \frac{x_2+x_5}{x_1+x_4} = \frac{1}{\hat{\pi}}$$

$$\left(\begin{array}{c} \text{R-Red: 18} \\ \text{OY-Orange or Yellow:125} \\ \text{G-Green: 20} \\ \text{B – Blue: 34} \end{array}\right)$$

$$\therefore\ \hat{\pi} = \frac{1}{1+\frac{x_2+x_5}{x_1+x_4}} = \frac{1}{1+\frac{x_2+34}{38}}$$

UTSouthwestern Medical Center

Expectation Maximization (in general) – A Multinomial Example (Iteration)

- Taking the conditional expectations for the computation of x_2 and x_3 will depend on a particular estimation of π, an initial estimate ($\pi^{(0)}$) must be supplied to the algorithm to start the procedure, then conditional expectations for the missing (coarsened) data at the p^{th} step (where $p \in \{1,2,\ldots\}$) is given by:

[E – Step]

$$E_{(p)}(x_2|y_2) = y_2 \frac{\frac{\pi^{(p-1)}}{4}}{\frac{\pi^{(p-1)}}{4}+\frac{1}{2}} \quad \text{- and -} \quad E_{(p)}(x_3|y_2) = y_2 \frac{\frac{1}{2}}{\frac{\pi^{(p-1)}}{4}+\frac{1}{2}}$$

[M – Step]

$$\widehat{\pi^{(p)}} = \frac{1}{1+\frac{E_{(p)}(x_2|y_2)+34}{38}}$$

- Convergence Criteria:
  - Generally we use relative convergence criteria (when the change in the parameters from step p to step $p+1$ falls below a relative tolerance ε_R) to determine when to stop iterating, for instance, the iteration will continue until:

[Convergence]

$$\left(\frac{1}{1+\frac{E_{(p)}(x_2|y_2)+34}{38}} - \frac{1}{1+\frac{E_{(p-1)}(x_2|y_2)+34}{38}}\right)^2 \leq \varepsilon_R$$

UTSouthwestern Medical Center

Expectation Maximization (Genetic Abundance Estimation)

UTSouthwestern
Medical Center