IMPAQT Isoform Abundance Estimation

MISO: https://www.nature.com/articles/nmeth.1528#Sec13

MISO Supps: https://static-content.springer.com/esm/art%3A10.1038%2Fnmeth.1528/MediaObjects/41592\_2010\_BFnmeth1528\_MOESM151\_ESM.pdf

QuantSeq: https://www.nature.com/articles/nmeth.f.376

Ok so the issue is that I was saying P(R|GC(I)) impacts estimation and not actually what it represents which is it’s update of read assignment. This is dumb or at least not necessarily our problem. GC bias should probably be used to estimate the prior.

P( ***Ψ*** | ***R*** ) = P( ***R*** | ***Ψ*** ) \* P( ***Ψ*** )

------------------------

P( ***R*** )

P( ***R*** | ***Ψ*** ) = P( ***R*** | ***I*** ) \* P( ***I*** | ***Ψ*** )

P( ***Ψ*** | ***R*** ) = P( ***R*** | ***I*** ) \* P( ***I*** | ***Ψ*** ) \* P( ***Ψ*** )

-------------------------------------

P( ***R*** )

P( ***R*** | ***I*** ) = Some function of it’s distance to the read

P( ***I*** | ***Ψ*** ) = Basically the relative expression of the clusters updated for GC bias (will have to model GC bias for this)

P( ***Ψ*** ) = Uniform Distribution

Let ***I*** represent the abundance of a transcript isoform and ***R*** represent a read potentially belonging to the isoform ***I***. Intuitively, we know that the true abundance of ***I*** depends upon the probability that ***R*** originated from that isoform. So, from Bayes Rule we get…

P( ***I*** | ***R*** ) = P( ***R*** | ***I*** ) \* P( ***I*** )

----------------------

P( ***R*** )

Our understanding of the underlying process by which reads are generated from a particular isoform in TAGseq experiments tells that the likelihood of ***R*** given ***I*** can be described as ***R***’s dependence on two random variables representing biases introduced during sequencing, namely random priming along the 3’ UTR for a given isoform, which we will denote as ***D(I),*** and the under-representation of sequences with greater proportions of G and C nucleotides, which we will denote as ***GC(I)***. So, our likelihood term can be rewritten as…

P( ***R*** | ***I*** ) = P( ***R*** | ***GC(I)*** ∩ ***D(I)*** )

Given the rules for joint probabilities, we can rewrite our updated likelihood term as the joint probability of ***R***, ***GC(I)***, and ***D(I)***.

P( ***R*** | ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** ∩ ***GC(I)*** ∩ ***D(I)*** )

------------------------------

P( ***GC(I)*** ) \* P( ***D(I)*** )

However, given that ***R*** is conditional on ***GC(I)*** and ***D(I)*** but ***GC(I)*** and ***D(I)*** can be reasonably assumed to be independent, the joint probability of the three events can be rewritten as...

P( ***R*** ∩ ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** | ***GC(I***) ) \* P( ***GC(I)*** ) \* P( ***R*** | ***D(I)*** ) \* P( ***D(I)*** )

Substituting this term back into our joint probability formula, our definition of the likelihood of ***R*** given ***GC(I)*** and ***D(I)*** reduces to…

P( ***R*** | ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** | ***GC(I***) ) \* P( ***R*** | ***D(I)*** )

Finally, we can rewrite our definition of the posterior probability as…

P( ***I*** | ***R*** ) = P( ***R*** | ***GC(I***) ) \* P( ***R*** | ***D(I)*** ) \* P( ***I*** )

--------------------------------------------

P( ***R*** )

This part is potentially not right, give it some more thought.

P( ***R*** | ***GC(I***) ) = modeling the GC content of the isoform (cluster of reads) as binomial distribution, what is the likelihood of getting a GC count of the read given the distribution of GC content in the cluster.

So potentially, GC is modelled as number of bases that are modeled as G or C (successes). When getting probabilty of successes, we scale the GC% of read by length of cluster to get propotion equivalent.

GC(I) ~ Bin(N,P)

N = Length of cluster

P = Probability of getting G or C (% GC)

P( ***R*** | ***D(I)*** ) = the likelihood of the read given it’s distance to the bounds of the cluster will be defined as the distance to the bounds of the cluster divided by the sum of the distance to all clusters subtracted from one divided again by the number of clusters - 1. Formula below:

[ 1 - ( d / ∑d) ] / [ NumClusters – 1 ]

P( ***I*** ) = prior distribution of isoform abundnance that is initially estimated as the ratio of core points in that cluster to the sum of all core points in every isoform. Formula below:

Ci / ∑Cj

R

GC(I)

D(I)