Supplement: Beyond compositionally in high throughput sequencing; estimating the importance of scale in data analysis with ALDEx2

Greg Gloor, Michelle Pistner Nixon, Justin Silverman *[1]

[1] Dep't of Biochemistry, University of Western Ontario, Penn State

*ggloor@uwo.ca

21 August 2023

Abstract

Introduction to scale simulation and FDR correction with ALDEx2.

Package

ALDEx2 1.33.1

Contents

1 GM can correlate with Shannon's entropy

We can also think about this relationship from an information theoretic point of view. Empirically in most datasets the geometric mean of $Y_n^{\parallel}$ is strongly correlated with Shannon's Information H (supplemental figure XX) suggesting that some knowledge of W is contained in the post-sequencing data that is not strictly compositional. Indeed, the logarithm of the geometric mean G and H can be understood from an information theoretic point of view to be an unweighted G or a weighted measure H of 'surprisal' in the dataset (supplement). Intuitively, underlying systems with different scales will contain different amounts of information and so we would expect $W_n^{\perp} \sim H_n$.

Entropy H and the geometric mean G (or its logarithm $\log_2 G$) are not simple to relate algebraically, but they can be understood in terms of what they are measuring if their description is rephrased in a common language. Recall have a $D \times N$ matrix of counts W decomposed into the proportions for the n^{th} sample $W_n^{\parallel}$ (or the equivalent probability distribution $p(w_n)$), and its scale $W_n^{\perp}$.

For notational simplicity assume a single discrete random variable X with a probability distribution $p(x)$ over $1 \ldots d$ features. The entropy $H(X)$ in bits is:

$$H(X) = -\sum_{i=1}^{d} p_i \log_2 p_i$$

and for the same distribution log2 of the geometric mean G is:

$$\log_2 G = \frac{1}{d}\sum_{i=1}^{d} \log_2 p_i$$

As defined here $H(X)$ is a weighted total of the uncertainty or 'surprisal' contained in $p(x)$, and relates to the amount of information we would need to have in order to reproduce $p(x)$. Conversely, G is an unweighted average of the same distribution. However, G is used as the denominator to calculate the centred log-ratio normalization (clr):

$$clr = \log_2(p_i) - \log_2 G$$

over $i = 1 \ldots d$. This compares each p_i with the geometric mean G. Thus G can be interpreted as a measure of difference for how far each p_i is from G, or in information theoretic terms as an unweighted measure of the mean surprisal for the distribution.

Thus, we can thus understand $H(X)$ as a measure of the total weighed surprisal and G as a measure of the average unweighted surprisal for $p(x)$. The total and the mean surprisal are related by the number of terms in $p(x)$ and by the weighting factor for each term.

These two measures are expected to have different behaviours in different distributions of p_i. In the case of a uniform distribution both $H(X)$ and G are maximal since $p(x)$ is equally and identically distributed. Thus, we expect that they are positively correlated here. In a Normal or a skewed distribution, we also anticipate a positive correlation because both are affected in the same direction by outlier values. In very sparse datasets, the two measures could become uncoupled because $H(X)$ could ascribe some uncertainty to the large number of low probability events, while G would tend to be very small. Here these two measures could be either uncorrelated or exhibit negative correlation. We can see this distributional behaviour in different datasets.

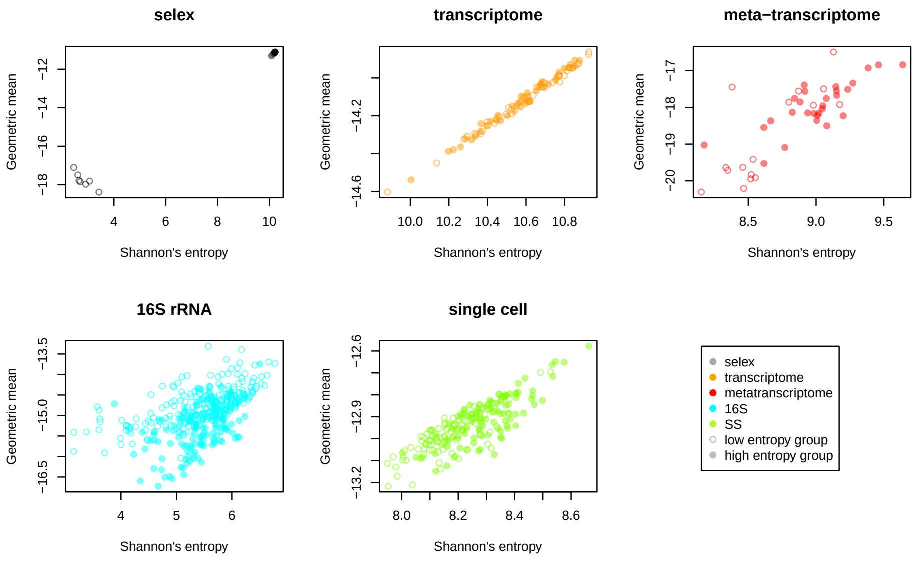
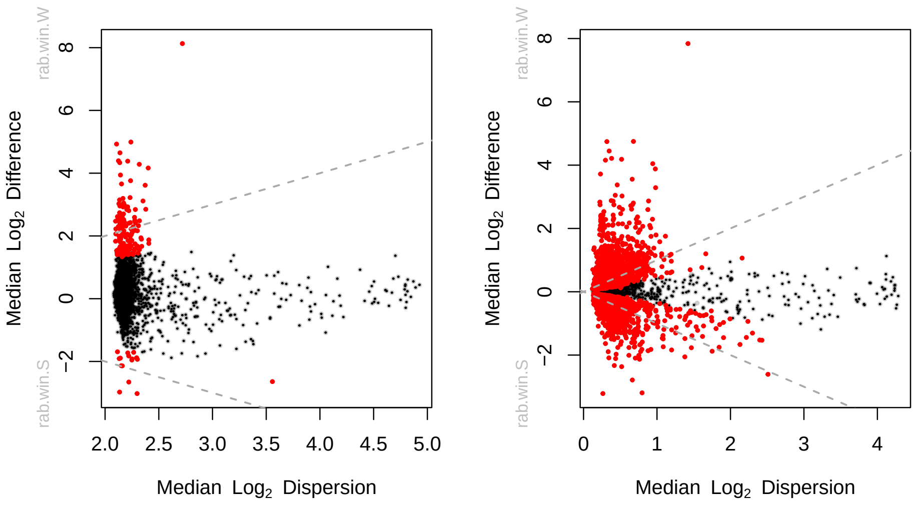


Figure 1: **Plot of Shannon's entropy (H) vs geometric mean (G) for each sample in different datasets** The groups that each sample belong to are highlighted as filled or open circles. Each group in each dataset has different entropy with the groups in the selex and metatranscriptome datasets being highly distinct.

```
rep.Inf <- function(x){ x[x == -Inf] <- min(x[x > -Inf]) -1 }

# check correlation of H and G for different distributions
sel.prop <- apply(selex, 2, function(x) x/sum(x))
yst.prop <- apply(yst, 2, function(x) x/sum(x))

L.yst.prop <- log2(yst.prop)
L.sel.prop <- log2(sel.prop)

L.yst <- apply(L.yst.prop, 2, rep.Inf)

L.sel <- apply(L.sel.prop, 2, rep.Inf)
```

```
load('analysis/x.s.mu.all.Rda')
load('analysis/x.all.Rda')

par(mfrow=c(1,2))
aldex.plot(x.s.mu.all)
aldex.plot(x.all)
```


rab.win.W
rab.win.S
Median Log$_2$ Difference
8
6
4
2
0
−2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Median Log$_2$ Dispersion
rab.win.W
rab.win.S
Median Log$_2$ Difference
8
6
4
2
0
−2
0
1
2
3
4
Median Log$_2$ Dispersion