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

true

GM is related to Information and Shannon's entropy in HTS datasets

Shannon's entropy has a volume or size

Information is a fundamental property of all measured systems. Shannon defined the information properties of discrete probability vectors and launched the field of information theory for communications. Famously, for any probability vector, Shannon set the total entropy as the inverse of the sum of the probability weighted logarithm of the probabilities. Less well known is that an unweighted average entropy was also defined. In the context of compositional data (CoDa) this measure is the inverse of the logarithm of the geometric mean of the probability vector. Thus, information theory and compositional data analysis intersect through the geometric mean of a probability vector. Here I show that the information theoretic interpretation can help us understand the scale of a system as defined by Nixon and Silverman, and that this interpretation can help in the interpretation of highly asymmetric datasets. I will show the Asymptotic Equipartition Property of information can be defined as a volume for discrete distributions, and how that volume relates to the scale of a system. Finally, I will show how Shannon's entropy can substitute for the geometric mean in common CoDa operations and how that alters the interpretation of the results.

We can think about scale from an information theoretic point of view as a measure of how much information, or total uncertainty, is encoded in a particular sample (1, 2). In the geometric interpretation of information theory used in quantum information theory, formally described as the Asymptotic Equipartition Property of information (3, 4), entropy can be interpreted as the volume occupied by a probability distribution relative to the maximum total entropy. See chapters 4 of the PhD thesis of Lecamwasam (5) for a more complete explanation of this.

For notational simplicity assume we have a single discrete random variable to represent a probability distribution with d elements; i.e. $X = \mathbf{p}_{i=(1\ldots d)}$. In information theory, the elemental amount of information or surprisal for p_i is the inverse of the logarithm of the elemental probability, $-log_2(p_i)$(6). This measure is often called self-information.

The total entropy of the system $H(X)$ is the weighted sum of the elemental information;

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

$H(X)$ corresponds to the amount of information needed in order to reproduce X. We can also calculate the expected amount of information for each observation in the random variable, and this is the mean of the elemental probability. This measure is also called the sample average of the information (4);

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

The linkage between compositional analysis, scale inference and information theory comes when we realize that the logarithm of the geometric mean calculated in base(2) is:

$$l2G(X) = \log_2 G(X) = \frac{1}{d}\sum_{i=1}^{d} \log_2 p_i$$

;

We see that $h(X) = -l2G(x)$. Furthermore, $l2G$ is used as the basis for the centred log-ratio transform and is the starting point for scale-based inference:

$$CLR(X) = log2(p_i) - l2G(X) = log2(p_i) + h(X)$$

Thus we see that the geometric mean used in the centred log ratio (CLR), often used for Compositional Data Analysis (CoDa) (7) is related to entropy or H. Indeed, we can rearrange the CLR formula to show that it can be interpreted as computing the difference between the elemental information and the mean information content:

$$CLR(X) = -log2(p_i) - (-l2G(X)) = -(-log2(p_i) - h(X))$$

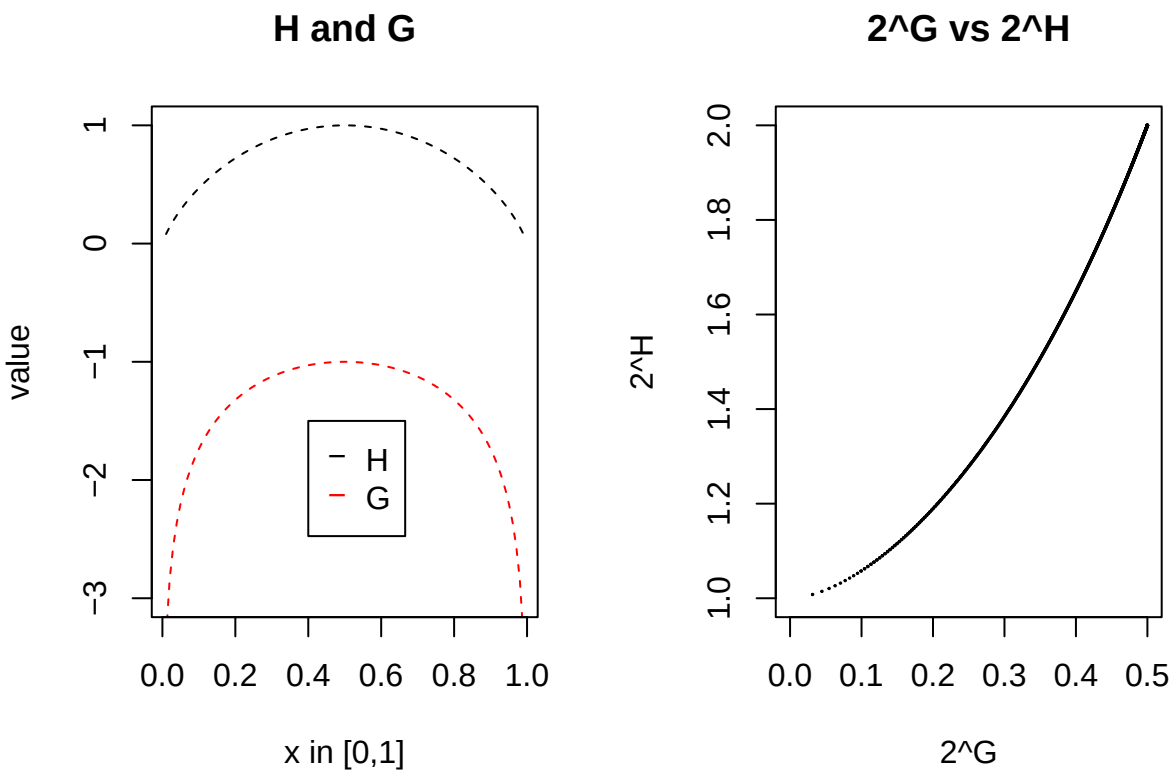
As defined in (8), the scale is the inverse of $l2G$, which is $h(X)$. Thus, one way we can understand scale is that it is measuring the total complexity of the system, and this is expected to increase with absolute size in most cases. Moreover, $H(X)$ and $G(X)$ share similar shapes in the continuum between 0-1 for a bivariate distribution as shown below:

```
par(mfrow=c(1,2))
curve(mf.G, from=1e-2, to =.99,  col='red', lty=2, ylim=c(-3,1),
   xlab="x in [0,1]", ylab="value", main="H and G")
curve(mf.H, from=1e-2, to =.99,  col='black', lty=2, add=T )

legend(.4,-1.5, legend=c('H','G'), col=c('black','red'), pch="-")

vals <- seq(from=.001, to=0.99, by=0.001)

plot(mf.2G(vals),mf.2H(vals), xlim=c(0,0.5), ylim=c(1,2), pch=19, cex=0.1,
  xlab="2^G", ylab="2^H", main="2^G vs 2^H")
```



The difference being that entropy is constructed to have a value of 0 at the margins because Shannon defined $0log(0) = 0$ while the geometric mean approaches negative infinity.

Now let's think about the idea of entropy as a volume which allows us to identify what amount of the available entropy space a given observation fills. The following is taken and modified from (5), to which you should refer for a more fulsome discussion, and it is covered in Chapter 2 of Wilde(4) and Chapter 3 of Cover and Thomas (3). If we start with a four part system $X1 = [A, C, G, T]$ where the frequencies are equally and identically distributed, then $p_A = p_C = p_G = p_T = \frac{1}{4}$. $H(X1) = -1 * 4 * (\frac{1}{4} * log2(\frac{1}{4})) = 2$. This is the maximum entropy possible. We can obtain the "volume" of $X1$ by exponentiating $H1$ using the same base as was used to calculate the entropy; $V1 = 2^{H1} = 4$. This is the same as the number of letters in the system; so the volume needed to explain the system is 4 units (in this case bits). But what happens in another system, $X2$ where A occurs with a much higher probability, say 0.7, and the other three are distributed equiprobably amongst the remainder with a probability of 0.1; i.e. $p_C = p_G = p_T = 0.1$. In this case $H2 = -1 * ((\frac{7}{10} * log2(\frac{7}{10})) + (3 * (\frac{1}{10} * log2(\frac{1}{10})))) = 1.358$. Here the volume of $X2 = 2^{H2} = 2.56$; meaning that less than the maximum volume is taken up by the information. Here $X2$ consumes about 64% of the volume of system $X1$. Thus, the entropic volume is a measure of the total complexity or the scales of the two systems.

But what happens if we consider the geometric mean instead of the entropy? In the example above, $l2G(X1) = (4 * log2(0.25))/4$ = -2, and $l2G(X2) = (log2(\frac{7}{10}) + 3 * log2(\frac{1}{10}))/4 = -2.62$. Exponentiating gives us values of 0.25 and 0.16 suggesting that the $l2G$ measure includes some estimate of size. Comparing the size of H vs $2^(l2G)$ suggests that both measures contain related information. Thus we can understand that scale is related to the information volume of a system, which in turn is related to the size of the system.

Empirically, we can see that $G(\mathbf{Y}_n^{\|})$ is strongly correlated with Shannon's Entropy $H(\mathbf{Y}_n^{\|})$ as expected from the discussion above, and that this difference converges to a constant as the number of entries in the probability vector increases regardless of the distribution, although different distributions converge at different rates. For example, if we plot the relationship between H and G as a function of the length of the probability vector we can see a direct inverse relationship.

```
## (Intercept)
```

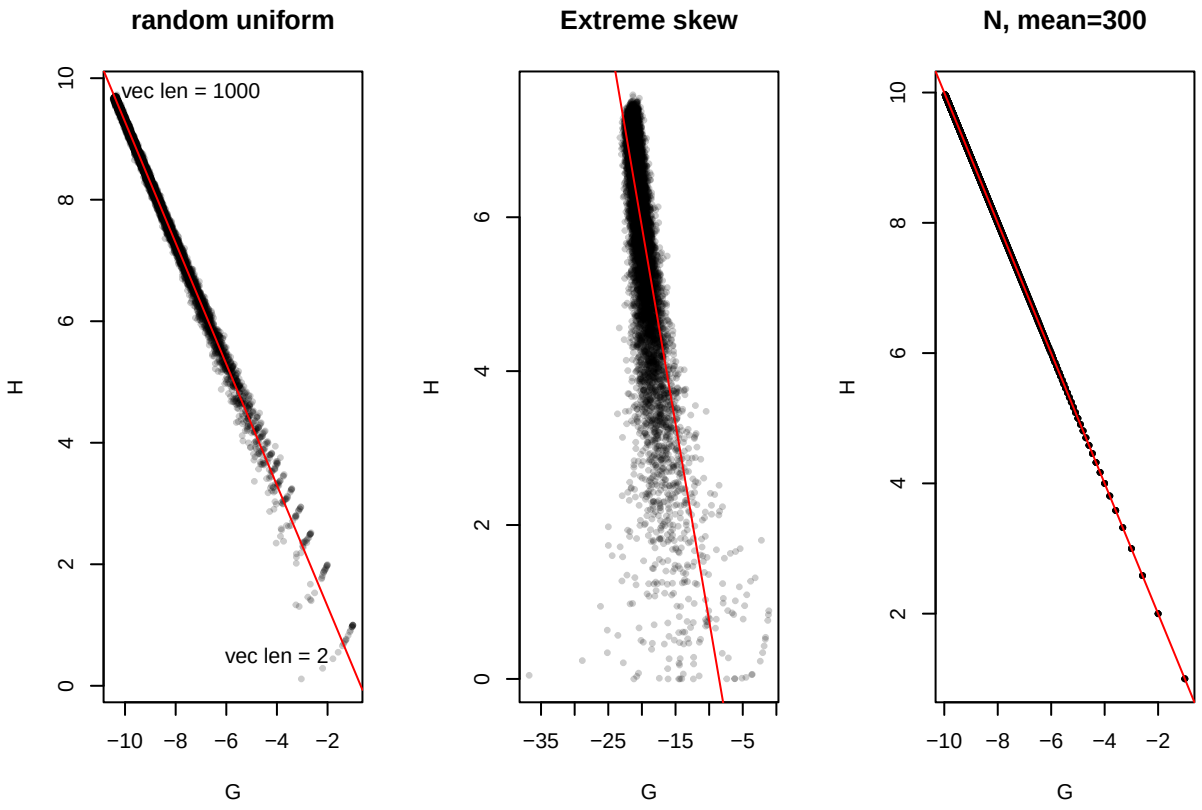


Figure 1: Association between entropy (H) and geometric mean (G) as a function of vector length. Twenty random vectors were constructed for each length between 2 and 1000 in increments of 2 for each of the random distributions in the legend; N = Normal, U = Uniform, B = Beta. The bottom right of each plot represents vectors of length 2, and the top left represents the vector of lenght 500. The maximum value of H increases as the vector length increases, and the maximum value of G decreases in lock-step. Each random distribution has an obviously distinct relationship between the two measures. For the purposes of high throughput sequencing the Beta distribution is most similar to that seen in the majority of instances.

```
##  -0.6860878
```

```
## (Intercept)
##   -4.359547
```

```
## (Intercept)
## -0.00139989
```

When we plot the relationship for any individual probability vector, we see that there is an direct relationship between the entropy and the log of the geometric mean, but that this relationship strongly depends on the underlying distribution of the probability distribution X.

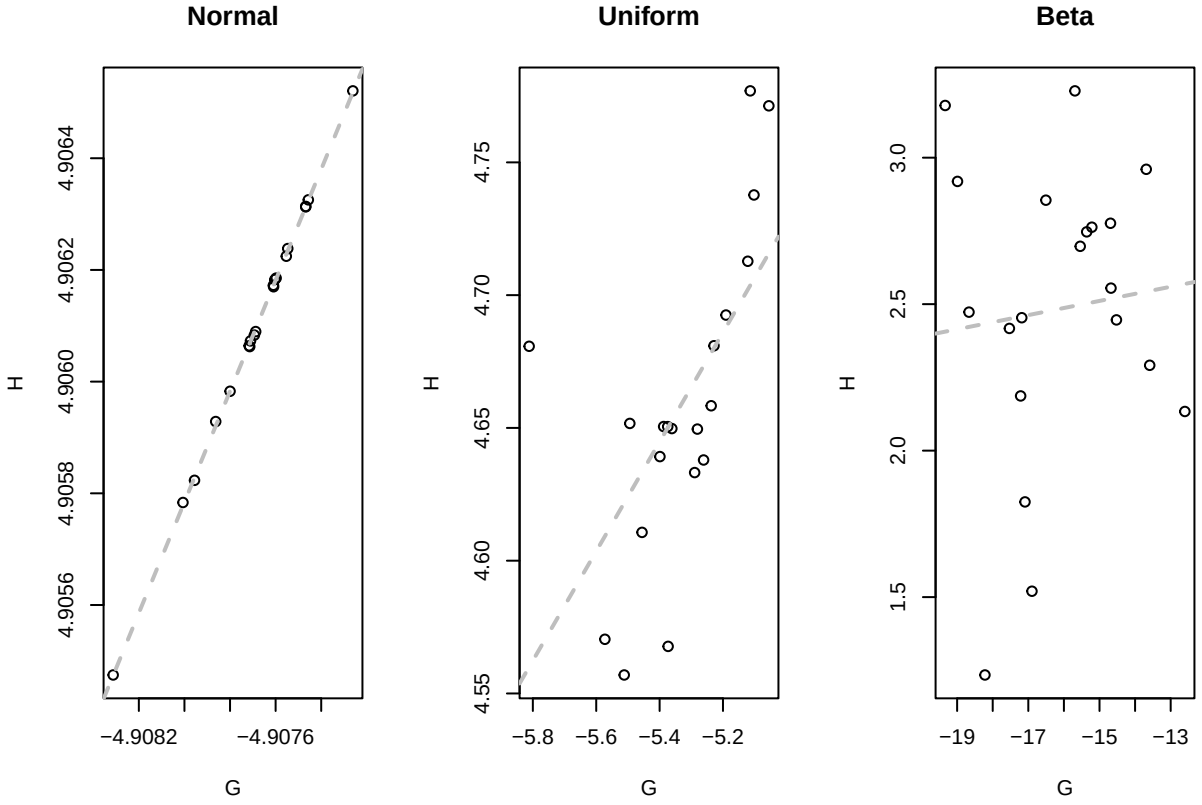


Figure 2: Plot of the association between H and G at a vector length of 30. The relationship between H and G is inverse, and the strength of that association depends on the distribution. The N distribution shows a very strong assocation, while the Beta distribution is less well defined. Associations are shown for a vector length of 30.

In real data, shown in Supplemental Figure 3 and Table 1 the correspondence is not as predictable, likely because the real data is a more complex distribution than any of the idealized distributions. Thus, these two measures have different behaviours with different distributions of p_i. In the case of a uniform distribution both $H(\mathbf{Y}_n^{\|})$ and $G(\mathbf{Y}_n^{\|})$ are maximal when $p(x)$ is equally and identically distributed. Thus, we expect that they are positively correlated here. In a Normal or a skewed distribution, we also observe 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 $H(\mathbf{Y}_n^{\|})$ could ascribe some uncertainty to the large number of low probability events, while $G(\mathbf{Y}_n^{\|})$ 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.

Intuitively, systems with different scales will contain different amounts of information and so we would expect $W_n^{\perp} \sim H_n$. As the scale of a system as defined by Nixon et al. (8) is inversely related to G, this means that scale is directly proportional to the information content and entropy of the data.

Below I show that we can replace G with H in the calculations performed by ALDEx2 without loss of utility.

Recall the underlying system is described by a $D \times N$ matrix of counts $\mathbf{W}$ decomposed into the proportions for the n^{th} sample $\mathbf{W}_n^{\parallel}$ (or the equivalent probability distribution $\mathbf{p}(w_n)$), and its scale $\mathbf{W}_n^{\perp}$, such that $\mathbf{W} = \mathbf{W}^{\parallel}\mathbf{W}^{\perp}$. Sequencing returns counts which are related to the underlying proportion; i.e., $\mathbf{Y}_n^{\parallel} \sim \mathbf{W}_n^{\parallel}$

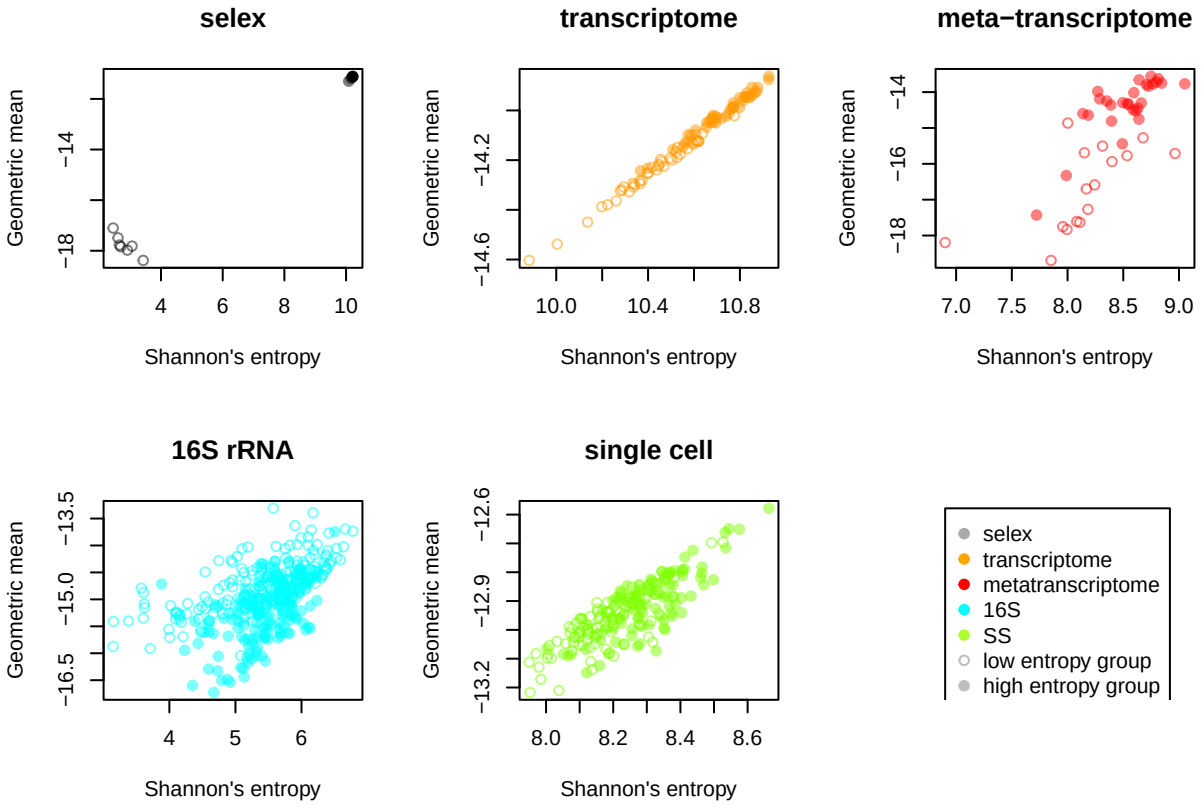


Figure 3: Plot of Shannon's entropy (H) vs geometric mean (G) for each sample in different datasets. The groups that each sample belong to are highlighted as filled or open circles. Each group in each dataset has different entropy with the groups in the selex and metatranscriptome datasets being highly distinct.

The table below summarizes the mean values for, and the correlation between, G and H (cor) and the sparsity defined as the proportion of features with less than 1 count per sample (spar) for each association in each group of samples:

Dataset	group	$\overline{G}$	$\overline{H}$	cor	spar
Selex	control	-11.2	10.2	0.99	0
"	selected	-17.8	2.8	-0.88	0.802
yeast	snf2 ko	-14.0	10.7	0.99	0.004
"	WT	-14.2	10.4	0.99	0.007
Meta	H	-18.8	8.6	0.78	0.451
"	BV	-18.2	8.9	0.79	0.238
16S	Pup	-14.7	5.4	0.68	0.079
"	Cent	-15.2	5.4	0.53	0.251
SS	A	-13.0	8.2	0.83	0.978
"	B	-12.9	8.3	0.80	0.977

1. Shannon,C.E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

2. Jaynes,E.T. and Bretthorst,G.L. (2003) Probability theory: The logic of science Cambridge University Press, Cambridge, UK.

3. Cover,T.M. and Thomas,J.A. (1991) Elements of information theory Wiley, New York.

4. Wilde,M.M. (2017) Quantum information theory 2nd ed. Cambridge University Press.

5. Lecamwasam,R. (2021) Investigations of metrology in optomechanics and quantum information theory.

6. Reza,F.M. (1994) An introduction to information theory Courier Corporation.

7. Aitchison,J. (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**, 139–160.

8. Nixon,M.P., Letourneau,J., David,L.A., Lazar,N.A., Mukherjee,S. and Silverman,J.D. (2023) Scale reliant inference.