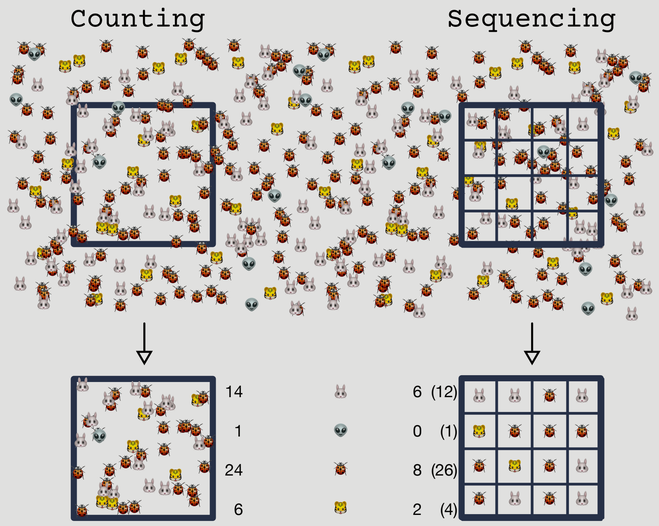
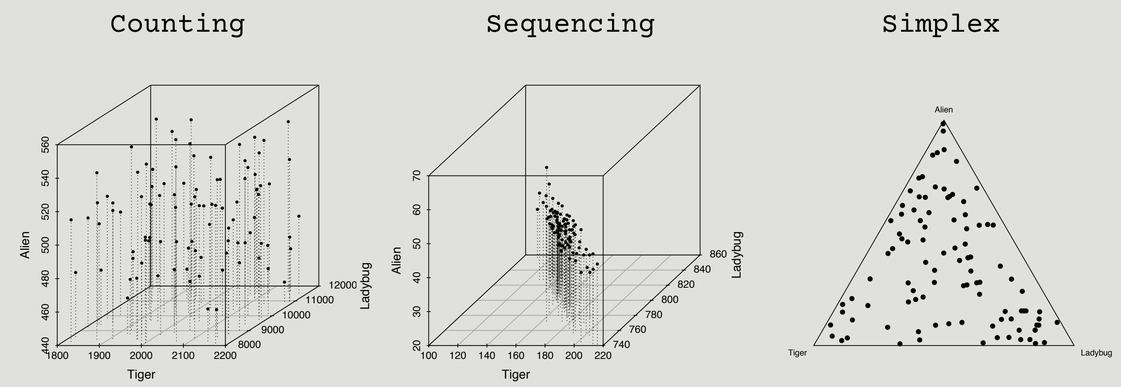
Data transforms and distance metrics on the simplex: not what they seem

Greg Gloor
Department of Biochemistry
University of Western Ontario

High throughput sequencing (HTS) data are probabilistic samples of the environment, that can be manipulated using the rules and tools of compositional data analysis. That HTS is difficult to analyze is well known, leading to multiple methods for 'count normalization' and the use of multiple distance metrics. This situation is similar to other domains prior to their understanding the problems with compositional data. I will use simple datasets to show that count normalization methods in use in the microbiome and RNA-Seq literature are irrelevant or misleading when analyzing HTS data. I will go on to show that distance metrics in common use are not informative about the distances in the underlying environment. Only the log-ratio family of transformations coupled with the Aitchison distance are guaranteed to be meaningful for these data.

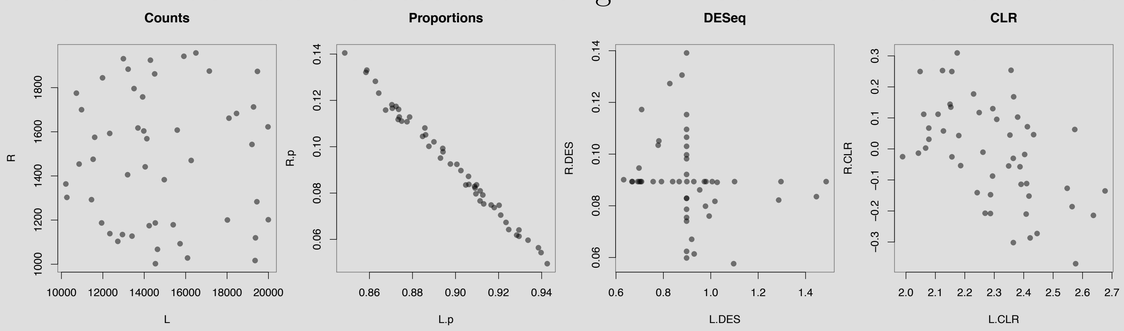


Sequencing returns an integer representation of the probability of observing a feature conditioned on the sample frequencies and the sequencing depth.



Sequencing data have one fewer dimensions than the number of features. These data exist on an D-1 dimensional plane termed a Simplex. Data on a Simplex cannot be recovered to the volume without additional information. Data normalization does not do what we hope it does: normalization does not restore the data to the volume.

Normalization does not recover the original information in the count data.



Given a count vector:
$\vec{\mathbf{s}} = [s_1, s_2, s_3, \ldots s_D]$

Relative abundance:
$$rAB_i = \frac{s_i}{\sum \vec{\mathbf{s}}}$$

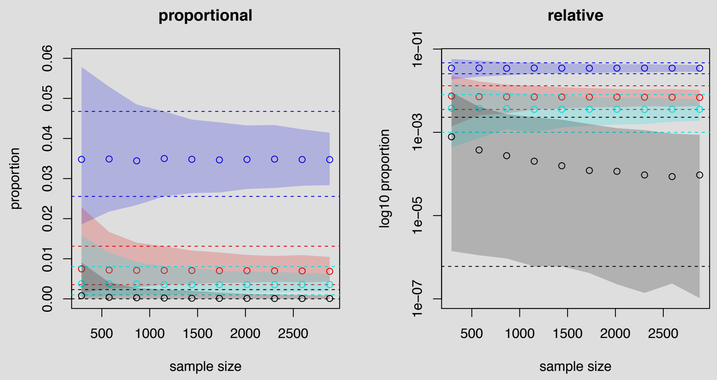
DESeq normalization:

$$\vec{\mathbf{g}} = \mathrm{g}\vec{\mathbf{f}}i-$$
$$\vec{\mathbf{r}}_j = \vec{\mathbf{s}}_j / \vec{\mathbf{g}}$$
$$\vec{\mathbf{d}}_j = \vec{\mathbf{s}}_j / Md(\vec{\mathbf{r}}_j)$$

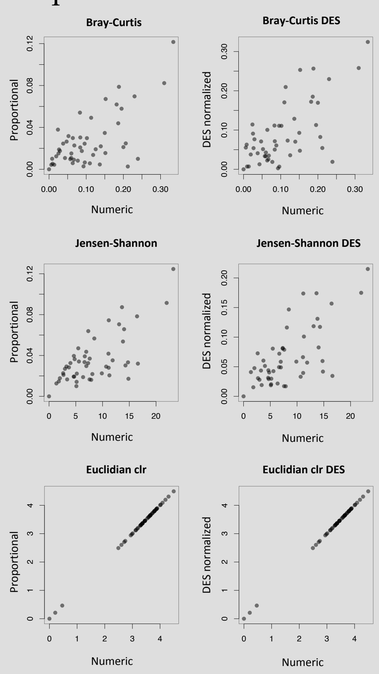
log-ratio transformation:

$$\vec{\mathbf{s}}_{clr} = [log(\frac{s_1}{g\vec{\mathbf{s}}}), log(\frac{s_2}{g\vec{\mathbf{s}}}) \ldots log(\frac{s_D}{g\vec{\mathbf{s}}})]$$
$$g\vec{\mathbf{s}} = \sqrt[D]{x_1 \cdot x_2 \cdot \ldots \cdot x_D}$$

DESeq normalization obscures the precision of the measurement estimate.



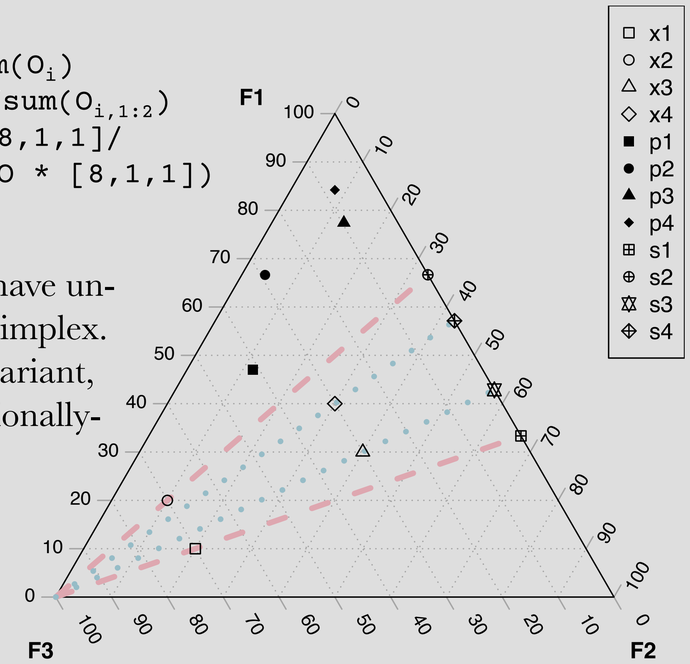
Aitchison distances preserve information



Distances between samples should be the same in the the underlying environment and in the post-sequencing data. A simple test dataset of four samples with three features to test if the distances are scale invariant, subcompositionally dominant, and perturbation (rotation) invariant.

```
O   F1 F2 F3
O1   1  2  7
O2   2  1  7
O3   3  4  3
O4   4  3  3
```

$x = O_i / \mathrm{sum}(O_i)$
$s = O_{i,1:2} / \mathrm{sum}(O_{i,1:2})$
$p = O * [8,1,1] / (\mathrm{sum}(O * [8,1,1])$



Widely used distance metrics have unpredictable properties on the simplex. BC, JSD and AD are scale invariant, JSD and AD are subcompositionally-dominant, but only the AD is resistant to perturbation or rotation of the data on the Simplex.

Metric (SDP)	d(x1,x2)	d(p1,p2)	d(s1,s2)	d(x3,x4)	d(p3,p4)	d(s3,s4)
Euclidian (—)	0.14	0.24	0.47	0.14	0.09	0.20
Manhattan (—)	0.20	0.40	0.67	0.20	0.14	0.29
Bray-Curtis (S–)	0.10	0.20	0.33	0.10	0.06	0.14
JSD (SD-)	0.13	0.15	0.13	0.08	0.06	0.08
Aitchison (SDP)	0.98	0.98	0.98	0.41	0.41	0.41