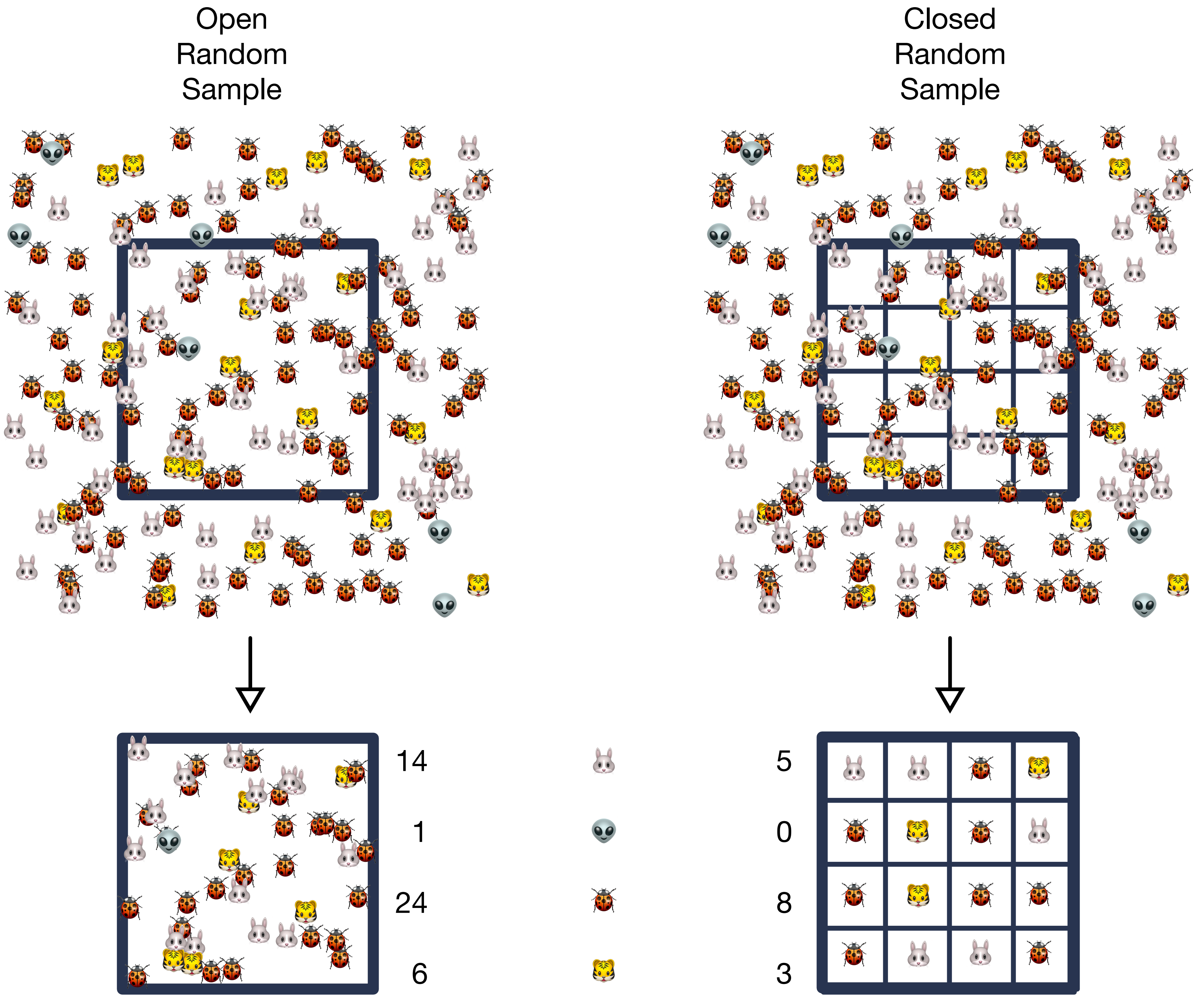
Interaction between data normalization and distance metrics in high−throughput sequencing data

Greg Gloor
Biochemistry, U. Western Ontario
https://github.com/ggloor/compositions/presentations
ggloor@uwo.ca @gbgloor

Motivation

- ecological methods spreading to other domains of HTS
- 'maybe sometimes ecological metrics are useful'
- test method designed for probability vectors
- want (need) feedback

Sequencing is probabilistic

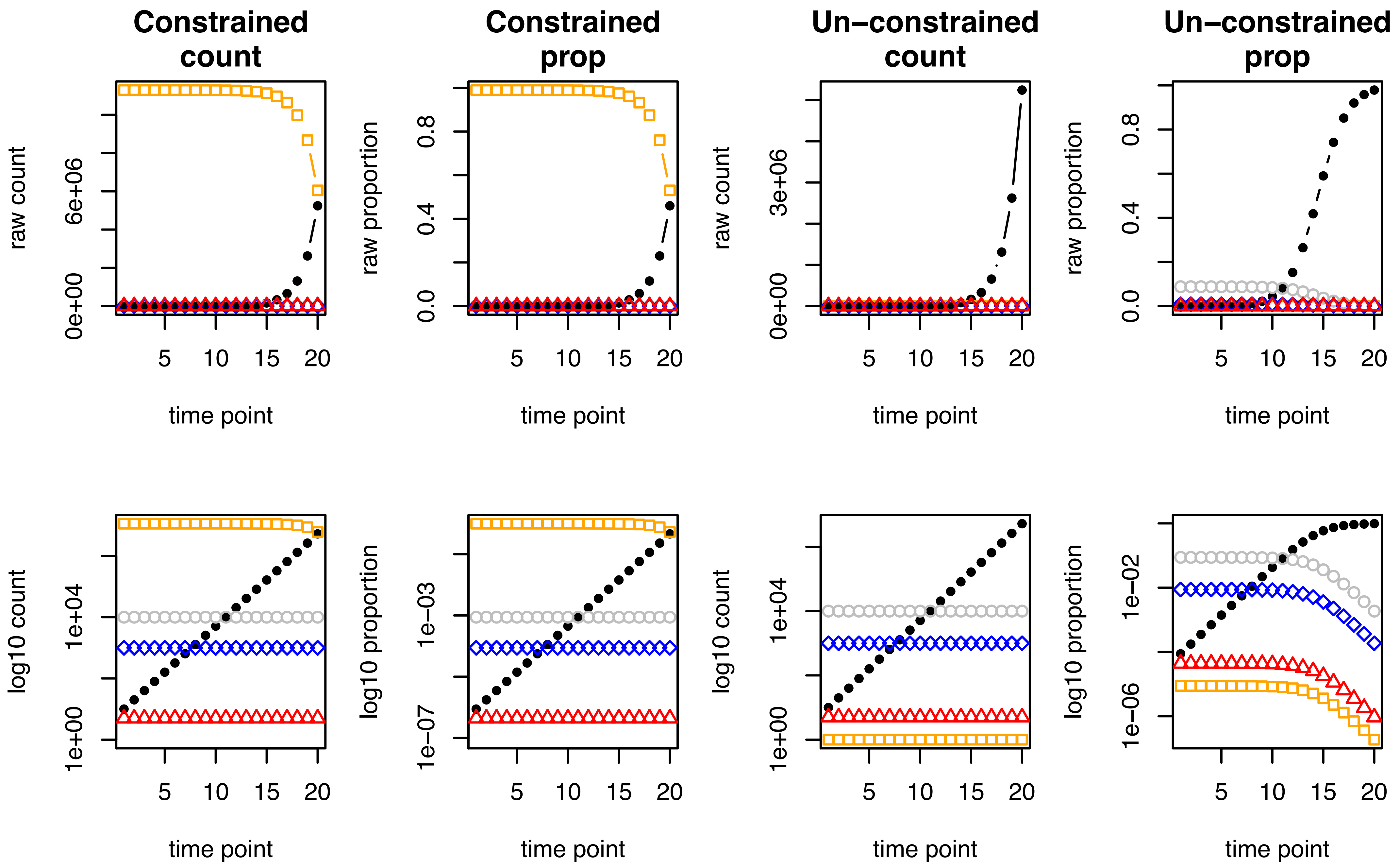


Open Random Sample

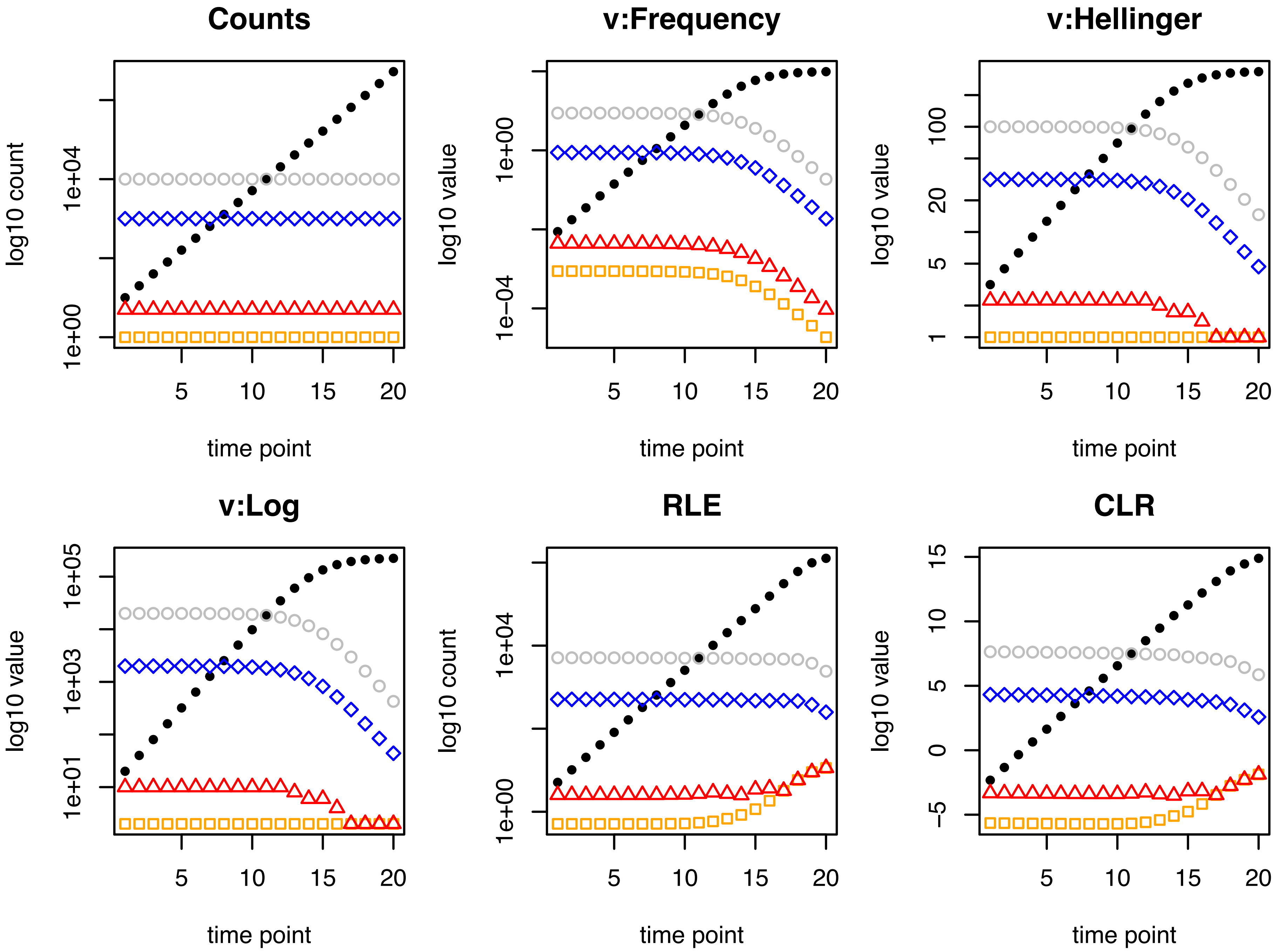
Closed Random Sample

Open Random Sample		Closed Random Sample
14	🐰	5
1	👽	0
24	🐞	8
6	🐯	3

Environments may be constrained or free



Some common transformations



Early transcriptome normalizations

$$rAB_i = \frac{s_i}{\sum \overrightarrow{\mathbf{s}}}$$

- also called Total Sum Scaling. Often log-transformed

$$RPKM_i = \frac{K \cdot s_i}{\sum \overrightarrow{\mathbf{s}} \cdot L_i}$$

- unclosed perturbation of original data. 1 RPKM=1 transcript is C2C12 cells, 3 RPKM = 1 transcript in liver cell line (Mortazavi et al. 2008)

$$TPM_i = \frac{RPKM_i}{\sum RPKM} \cdot K$$

- closed RPKM multiple by a constant (Li et al. 2010)

None of these are scale invariant

Scaling normalization methods

- Concept here is that counts per sample can be normalized to a per-sample midpoint, and that such a normalization can approximate the numbers of features in the underlying environment
- Popular (pervasive) in transcriptome, microbiome, metagenome
- Assume that the total count in the environment is the same, and that most features do not vary (constrained environment)
- Methods differ in how they choose the midpoint
    - trimmed mean of M values (TMM, edgeR)
    - cumulative sum scaling (CSS, metaGenomeSeq)
    - relative log expression (RLE, DESeq)

Calculating RLE

$\vec{\mathbf{g}} = \mathrm{G}\vec{f}_i$

- feature-wise geometric mean

$\vec{\mathbf{r}}_j = \frac{\vec{\mathbf{s}}_j}{\vec{\mathbf{g}}}$

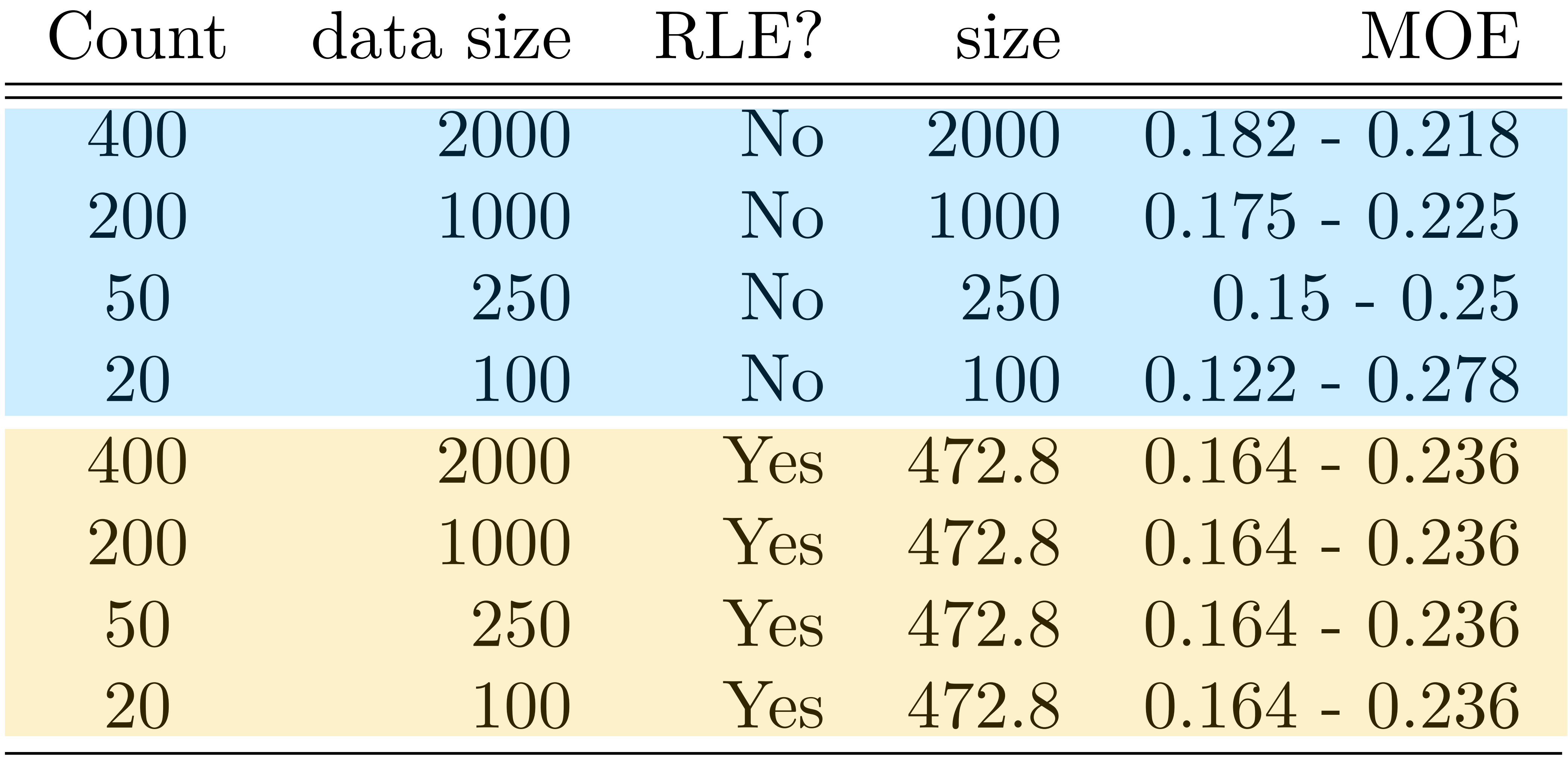
- sample count divided by previous

$\vec{\mathbf{d}}_j = \frac{\vec{\mathbf{s}}_j}{\widetilde{\vec{\mathbf{r}}_j}}$

- sample could divided by median of previous sample-wise

Feature	$\vec{\mathbf{s}}_1$	$\vec{\mathbf{s}}_2$	$\vec{\mathbf{g}}$	$\vec{\mathbf{r}}_1$	$\vec{\mathbf{r}}_2$	$\vec{\mathbf{d}}_1$	$\vec{\mathbf{d}}_2$
F1	1500	1000	1224.7	1.22	**0.81**	1219.5	1234.6
F2	25	15	19.4	1.29	0.77	20.3	18.5
F3	1000	500	707.1	1.41	0.71	813.0	617.3
F4	75	50	61.2	**1.23**	0.82	61.0	61.7
F5	500	1500	866.0	0.58	1.73	406.5	1851.9

We gain or lose apparent information

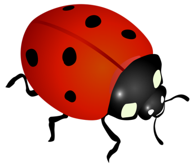


Count	data size	RLE?	size	MOE
400	2000	No	2000	0.182 - 0.218
200	1000	No	1000	0.175 - 0.225
50	250	No	250	0.15 - 0.25
20	100	No	100	0.122 - 0.278
400	2000	Yes	472.8	0.164 - 0.236
200	1000	Yes	472.8	0.164 - 0.236
50	250	Yes	472.8	0.164 - 0.236
20	100	Yes	472.8	0.164 - 0.236

A simple test dataset - unconstrained

50 samples of random Normal data, enforced minimum 0.1

- 50 +/- 25

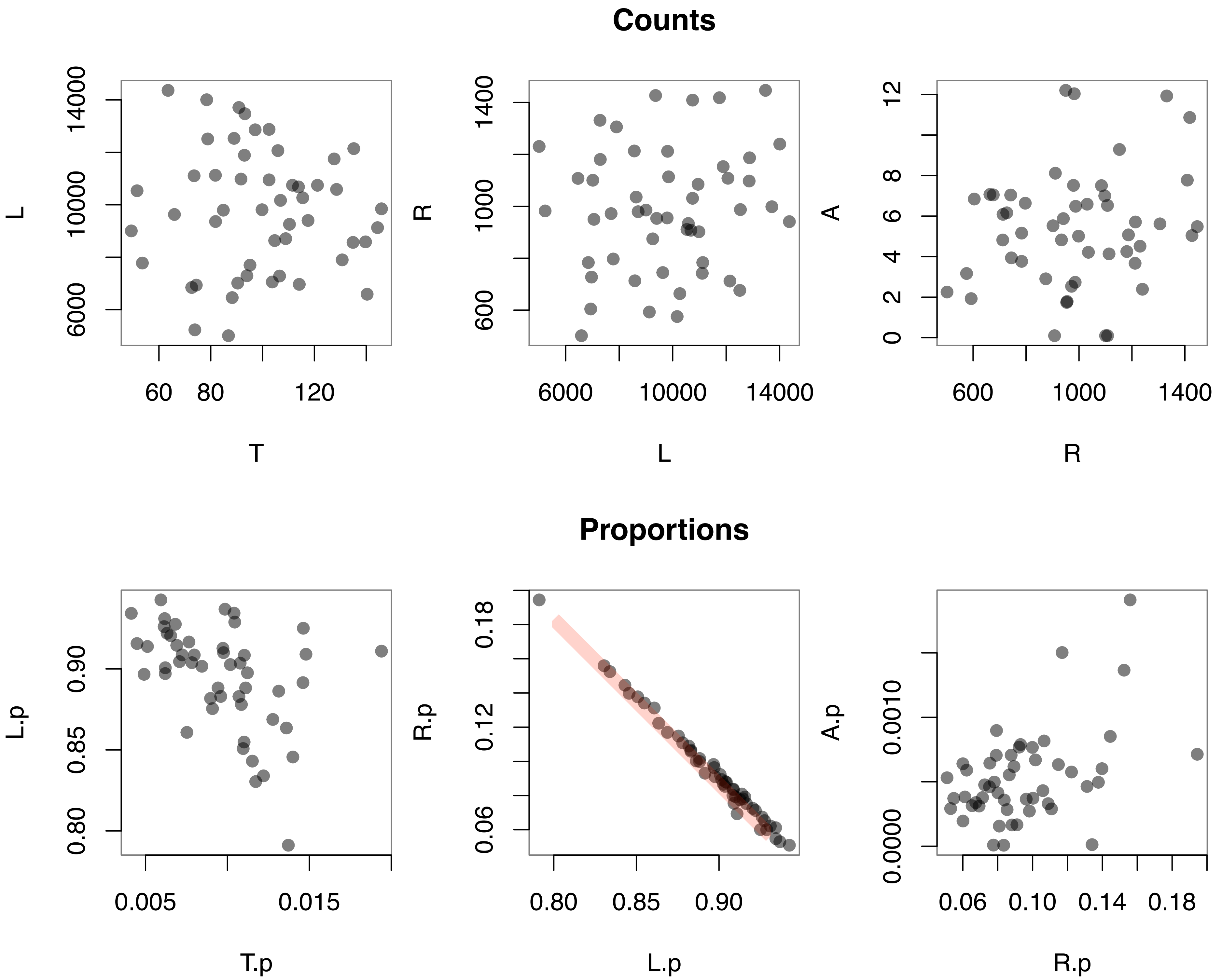


- 10000 +/- 2500

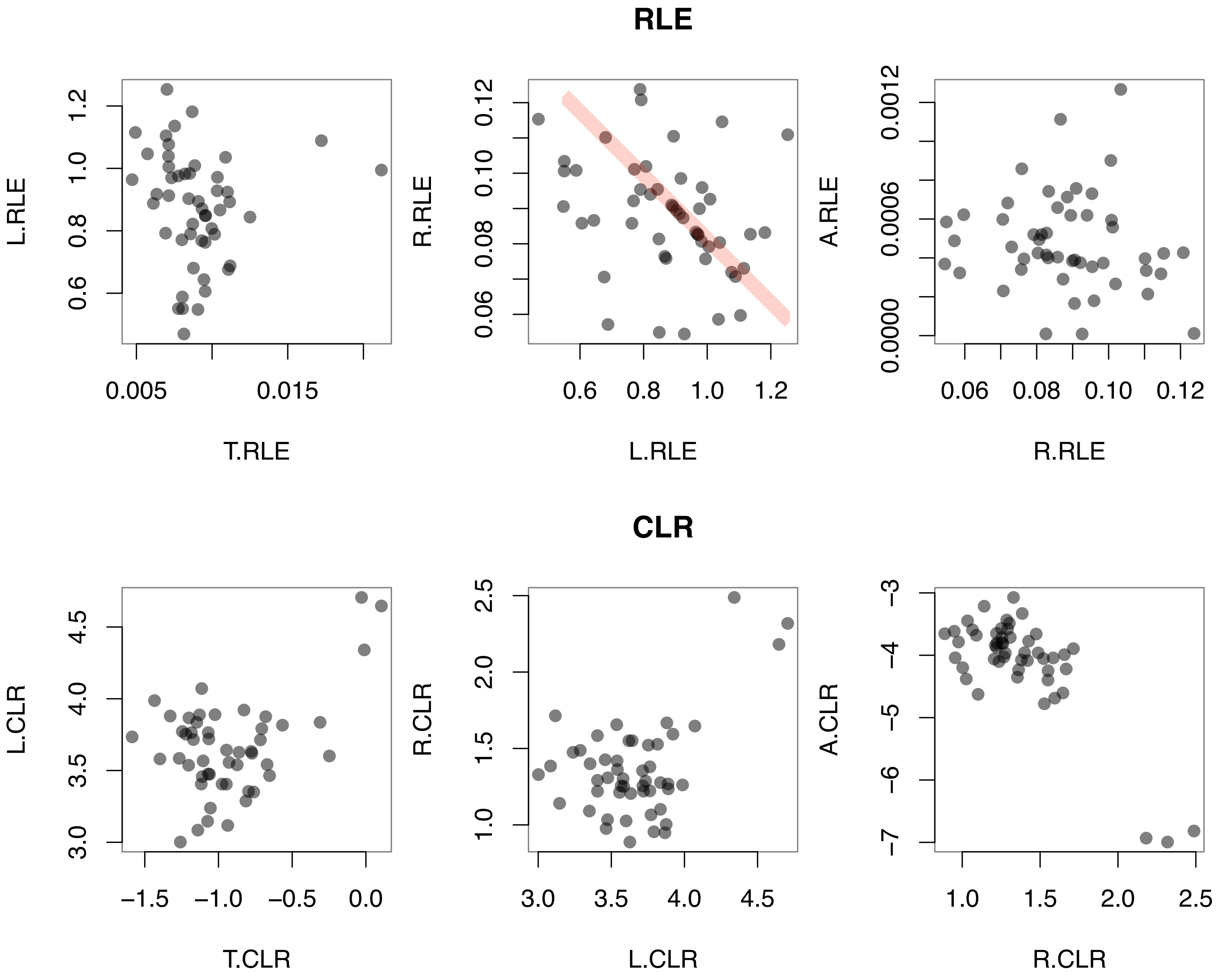
- 1000 +/- 250

- 5 +/- 2.5

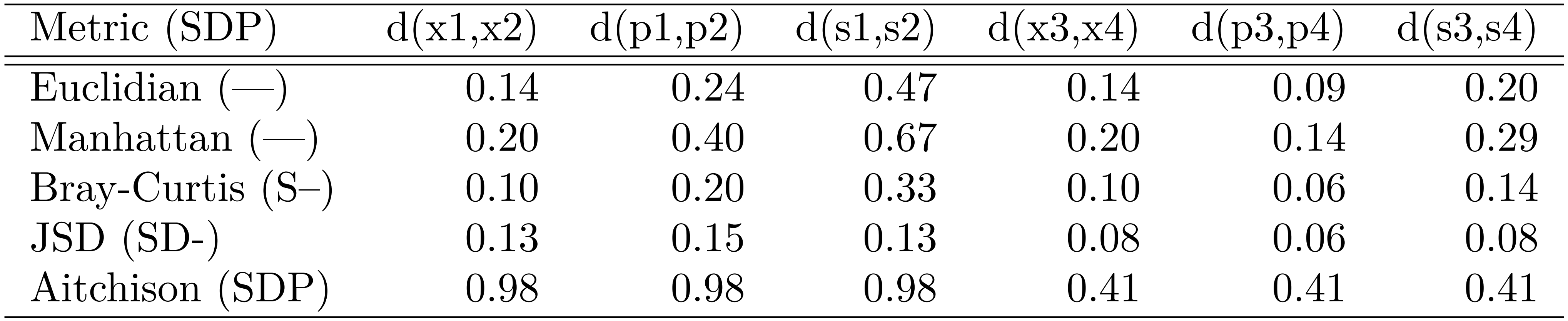
counts vs. proportions (etc)



RLE and CLR


RLE
L.RLE
0.6
0.8
1.0
1.2
0.005
0.015
T.RLE
R.RLE
0.06
0.08
0.10
0.12
0.6
0.8
1.0
1.2
L.RLE
A.RLE
0.0000
0.0006
0.0012
0.06
0.08
0.10
0.12
R.RLE
CLR
L.CLR
3.0
3.5
4.0
4.5
−1.5
−1.0
−0.5
0.0
T.CLR
R.CLR
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
L.CLR
A.CLR
−7
−6
−5
−4
−3
1.0
1.5
2.0
2.5
R.CLR

Distances



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

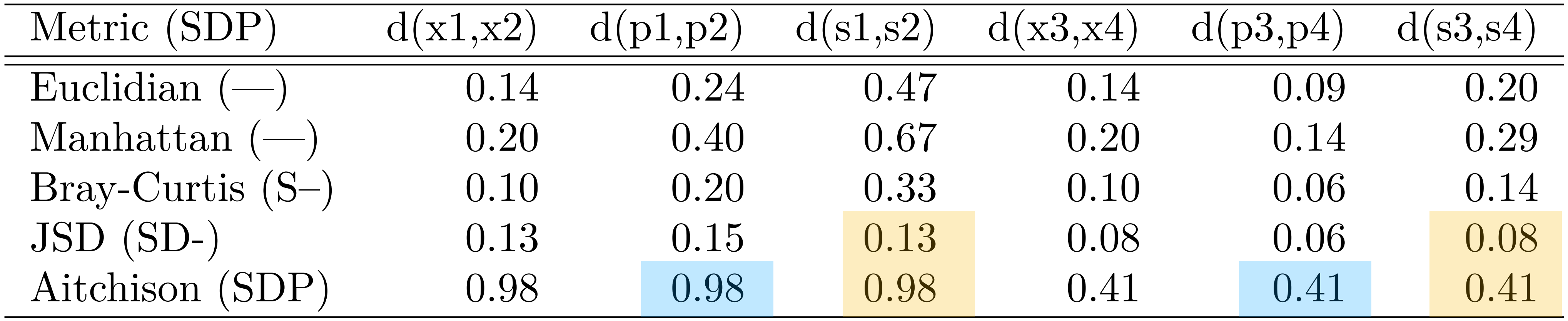
Martín-Fernández et al. 1998

- Bray-Curtis dissimilarity (or symmetrized as a distance) is a normalized Manhattan distance

- Jensen-Shannon Distance is a symmetric version of the Kulback-Leibler divergence metric widely used to compare probability vectors (Enterotypes: Arumugam Nature 2011)

- Aitchison is the Euclidan distance of the CLR

Distances



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

Martín-Fernández et al. 1998

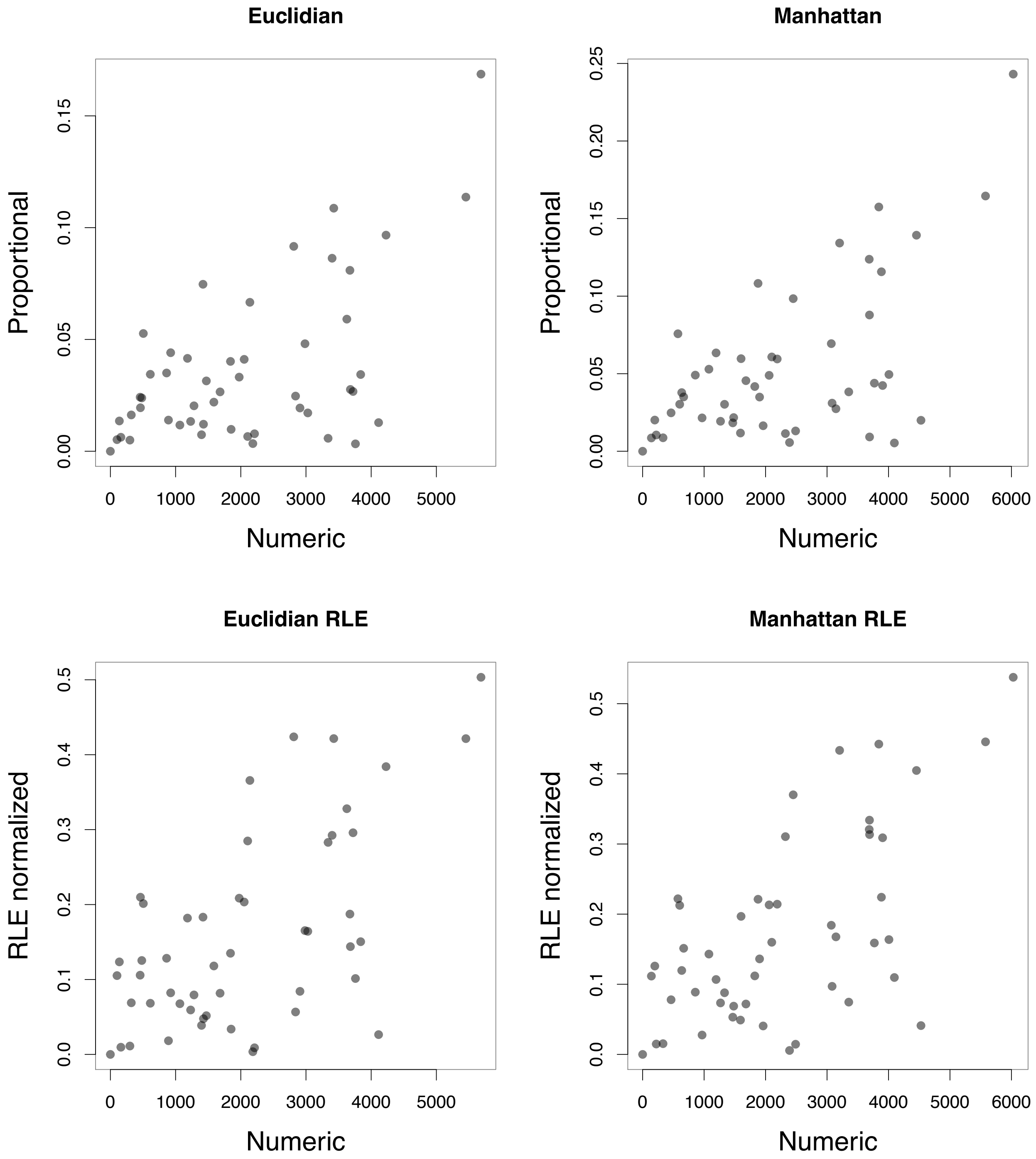
Perturbation invariance

Subcompositional dominance

Remember, we care about the environment, not the data after sequencing!

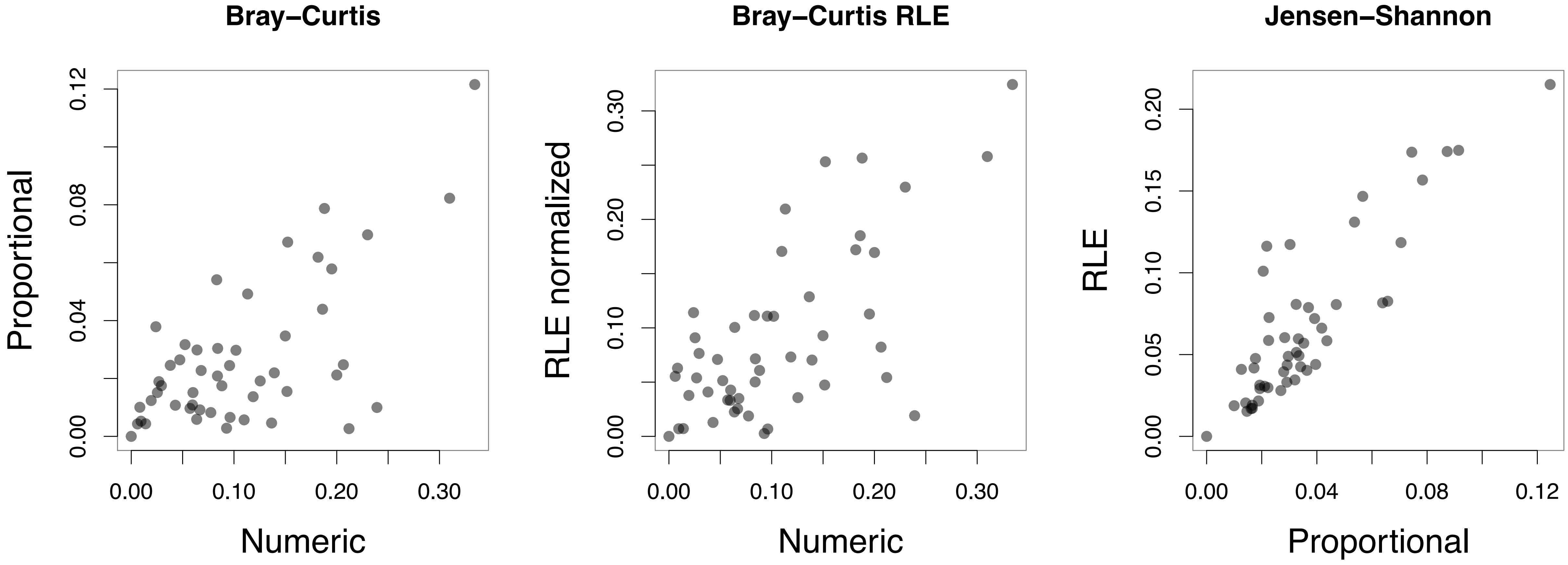
Distances and normalizations

RLE normalization fixes the problem - right?



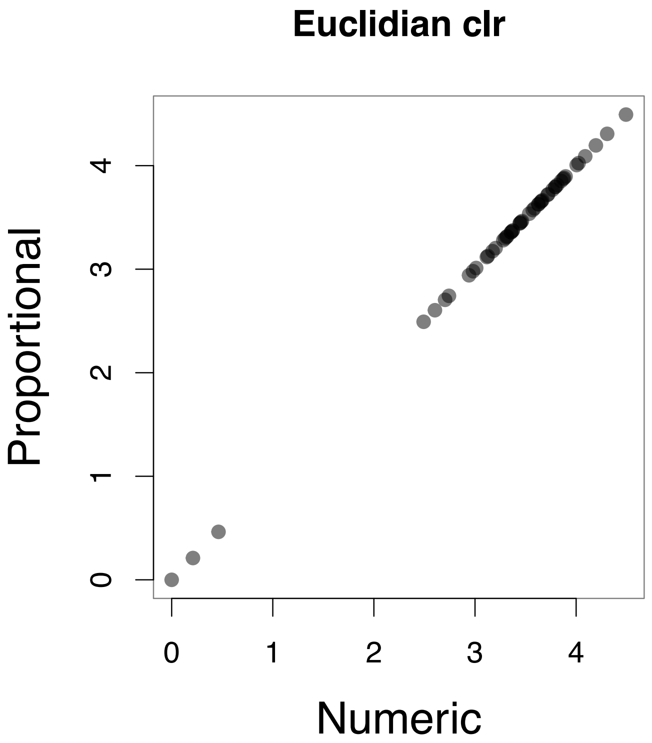
Distances and normalizations

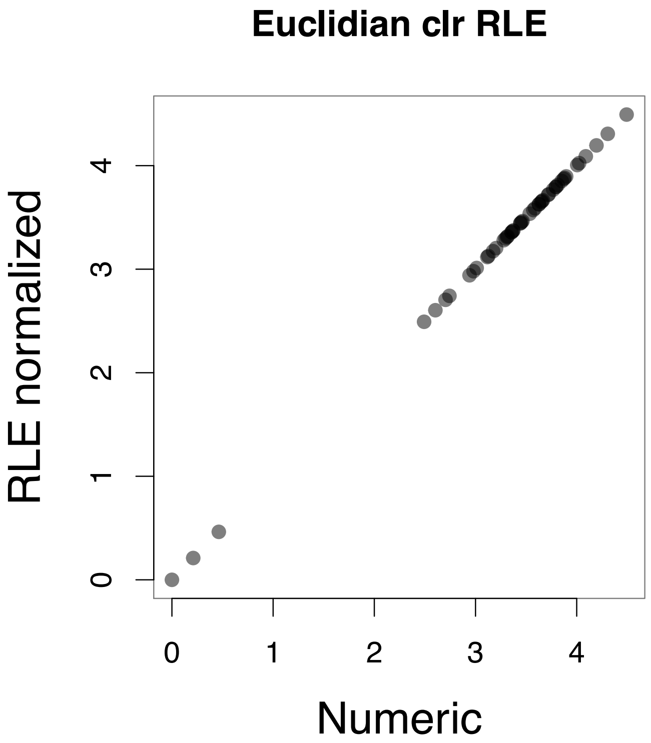
RLE fixes the problem with the right metric - right?



Distances and normalizations

We need to change the problem





Summary

- HTS is compositional
- The environment can only be safely modelled as an open environment
- Sequencing data should tell us about the environment, not just the post-sequencing data
- Only compositionally-appropriate distances tell us about the environment